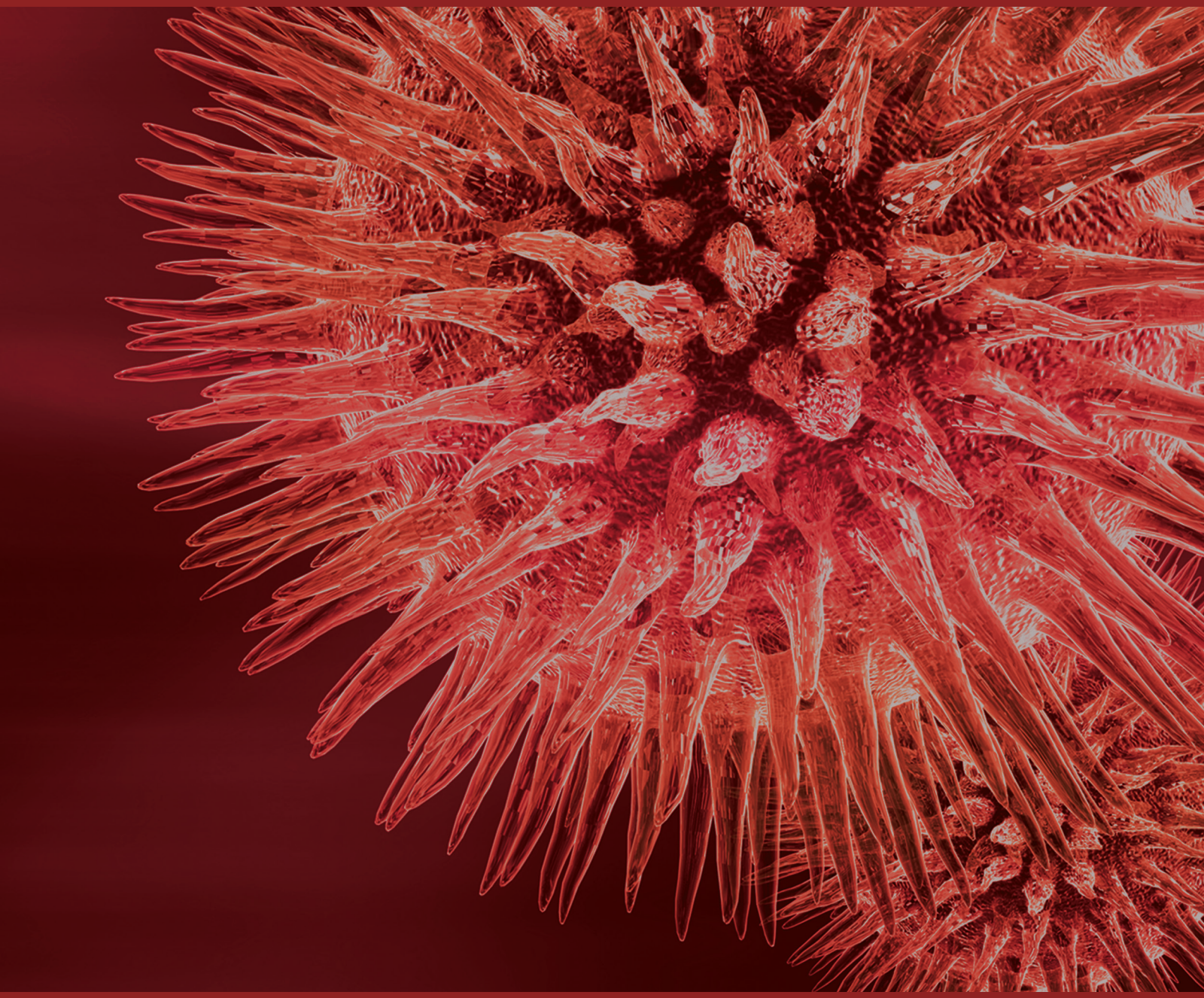
BioMed Research International

Integrated Analysis of Multiscale Large-Scale Biological Data for Investigating Human Disease

Guest Editors: Tao Huang, Lei Chen, Mingyue Zheng, and Jiangning Song



Integrated Analysis of Multiscale Large-Scale Biological Data for Investigating Human Disease

BioMed Research International

Integrated Analysis of Multiscale Large-Scale Biological Data for Investigating Human Disease

Guest Editors: Tao Huang, Lei Chen, Mingyue Zheng, and Jiangning Song

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 760765, 2 pages
http://dx.doi.org/10.1155/2015/760765

Editorial

Integrated Analysis of Multiscale Large-Scale Biological Data for Investigating Human Disease

Tao Huang,[1] Lei Chen,[2] Mingyue Zheng,[3] and Jiangning Song[4]

[1]*Mount Sinai School of Medicine, New York, NY 10029, USA*
[2]*Shanghai Maritime University, Shanghai 201306, China*
[3]*Chinese Academy of Sciences, Shanghai 201203, China*
[4]*Monash University, Clayton, VIC 3800, Australia*

Correspondence should be addressed to Tao Huang; tohuangtao@126.com

Received 1 January 2015; Accepted 1 January 2015

In recent years, more and more omics data is generated. Even for the same samples, multiple levels of omics data can be measured in large scale. These multiscale and large-scale data could help in revealing the biological basis of complex diseases and optimizing the therapeutic strategies. Analysis of such data is very challenging since the data is inaccessible in the past and few methods are developed. In this special issue, we presented 17 novel studies about the analysis method of such complex data and their applications to interesting medical and biological questions.

Y. Jiang et al. proposed a method to identify novel thyroid cancer-related genes and chemicals using shortest path algorithm. Some of the identified genes are crucial to the tumorigenesis and development of thyroid cancer. This method can be generalized to identify genes for other complex diseases.

W. Kong et al. constructed the common and brain region specific subnetworks of Alzheimer's disease. The identified common subnetworks across six brain regions suggested that inflammation of the brain nerves is one of the critical factors of Alzheimer's disease and calcium imbalance may link several causative factors of Alzheimer's disease.

M. F. Rogers et al. studied supervised interactive network inference using multiple kernel learning. The proposed method was composed of cautious classification and data cleaning, where cautious classification was used to increase the accuracy by restricting predictions to high-confidence instances, whereas data cleaning was used to mitigate the influence of mislabeled training instances.

J. Che and M. Shin proposed a meta-analysis strategy for gene prioritization that integratively employs three different genetic resources: gene expression data, single nucleotide polymorphism (SNP) genotype data, and expression quantitative trait loci (eQTL) data. The strategy for gene prioritization showed its superiority to conventional methods in discovering significant disease-related genes with several types of resources, while making good use of potential complementarities among available genetic resources.

J. E. Vargas et al. used a network flow approach to predict protein targets and flavonoid backbones to treat respiratory syncytial virus (RSV) infection. They identified 26 flavonoids and 5 compounds through topological analysis of chemical-protein and protein-protein interaction network. Some mechanisms of action of early RSV infection were uncovered.

Y. Dong et al. reported a support vector machine (SVM) model for predicting new interactions between the human papillomavirus 16 (HPV16) and other proteins. The analysis of protein-protein interactions indicated that HPV16 enlarged its scope of influence by interacting with human proteins as much as possible, and these interactions alter a broad array of cell cycle progression.

Y. Zhang et al. proposed a computational approach, integrating Ping-Pong algorithm and multiobjective genetic algorithm, to identify subtype-specific miRNA-mRNA functional regulatory modules. And this method was applied to investigate subtype-specific miRNA-mRNA functional regulatory modules of multiple myeloma.

D. Li et al. identified novel breast cancer subtype-specific biomarkers by integrating genomics analysis of DNA copy number aberrations and miRNA-mRNA dual expression profiling. The predicted correlation between BRCA1 and miR-143/miR-145 was validated by experiments and miR-143/miR-145 were promising novel biomarkers for breast cancer subtyping.

Z.-H. You et al. proposed a method to detect large-scale protein-protein interactions by integrating big biosensing data with computational model. The model was based on extreme learning machine (ELM) combined with a novel representation of protein sequence descriptor. The accuracy of their method was 84.8% while the sensitivity and specificity were 84.08% and 85.53%, respectively. It outperformed support vector machine (SVM).

J. Zhang et al. developed an effective method for distinguishing age-related macular degeneration (AMD) related genes using gene ontology (GO) and KEGG enrichment scores. 720 GO terms and 11 KEGG pathways were found to be important for predicting AMD related genes. These GO terms and KEGG pathways could help understand the underlying mechanisms of AMD.

X. Li et al. conducted a genome-wide association study of CNVs in two large-scale lung cancer datasets: the Environment And Genetics in Lung cancer Etiology (EAGLE) and the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial datasets. With a combined analysis of the association accordance between the two datasets, they identified 167 risk SNP loci and 22 CNVs associated with lung cancer and linked them with recombination hotspots.

X. Wang et al. evaluated the use of the KDIGO staging system for the prediction of prognosis in patients with septic acute kidney injury in intensive care units in Beijing, China, via a multicenter clinical study. Six independent risk factors for mortality were identified, which may help in making early and accurate diagnosis and adopting preventive and therapeutic interventions that could reduce mortality rates in the patients.

Y.-F. Gao et al. proposed a graphic method to identify novel glioma related genes. The known glioma related genes were mapped onto a weighted protein-protein interaction network and the genes that link the known glioma related genes on the network were considered as candidate novel genes. The candidate genes were further filtered by permutation test. Some of the final novel glioma related genes were supported by latest literatures.

L. He et al. combined the statistical algorithm with the gene-pathway bipartite networks to generate the reliable lists of cancer-related DEGs and constructed the models, which can be used for predicting the prognosis of three types of cancers, namely, breast cancer, acute myeloma leukemia, and glioblastoma.

L. Zhu et al. investigated the genome-wide gene expression in *Gleditsia sinensis* with transcriptome sequencing which generated 58583 unigenes. In these genes, 31385 unigenes were annotated. What is more, a PPI network was constructed and used to predict new stress resistance genes, in order to provide a platform for future functional genomic studies.

N. Zhang et al. proposed a computational method to identify influenza A/H7N9 virus infection-related human genes from shortest paths in a virus-human protein interaction network. Finally, 20 human genes were screened out which could be the most significant, providing guidelines for further experimental validation.

L. Guo et al. performed an integrated analysis of miRNA, lncRNA, and mRNA expression profiles in human HepG2 and L02 cells. They found that isomiR repertoires and expression patterns might contribute to tumorigenesis through different biological roles. The cross-talk between different RNA molecules could help reveal the complex mechanisms underlying tumorigenesis.

How to analyze the multiscale large-scale biological data is one of the most important questions in postgenomic era. It is even more important than generating these data. It is the battlefield for computational scientists, the bridge between biotechnology and clinical applications, and the stepping stone for translational medicine. The collaborations between scientists from different backgrounds are essential since no one can fully understand and completely solve such complex question by himself. The complexities of biological system amaze people and inspire people to investigate with all means.

Tao Huang
Lei Chen
Mingyue Zheng
Jiangning Song

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 890381, 10 pages
http://dx.doi.org/10.1155/2015/890381

Research Article

Improving the Understanding of Pathogenesis of Human Papillomavirus 16 via Mapping Protein-Protein Interaction Network

Yongcheng Dong,[1] Qifan Kuang,[2] Xu Dai,[2] Rong Li,[3] Yiming Wu,[2] Weijia Leng,[2] Yizhou Li,[2] and Menglong Li[2]

[1]*College of Life Sciences, Sichuan University, Chengdu 610064, China*
[2]*College of Chemistry, Sichuan University, Chengdu 610064, China*
[3]*College of Computer Science, Sichuan University, Chengdu 610064, China*

Correspondence should be addressed to Yizhou Li; liyizhou_415@163.com and Menglong Li; liml@scu.edu.cn

Received 19 July 2014; Revised 27 August 2014; Accepted 1 September 2014

Academic Editor: Mingyue Zheng

The human papillomavirus 16 (HPV16) has high risk to lead various cancers and afflictions, especially, the cervical cancer. Therefore, investigating the pathogenesis of HPV16 is very important for public health. Protein-protein interaction (PPI) network between HPV16 and human was used as a measure to improve our understanding of its pathogenesis. By adopting sequence and topological features, a support vector machine (SVM) model was built to predict new interactions between HPV16 and human proteins. All interactions were comprehensively investigated and analyzed. The analysis indicated that HPV16 enlarged its scope of influence by interacting with human proteins as much as possible. These interactions alter a broad array of cell cycle progression. Furthermore, not only was HPV16 highly prone to interact with hub proteins and bottleneck proteins, but also it could effectively affect a breadth of signaling pathways. In addition, we found that the HPV16 evolved into high carcinogenicity on the condition that its own reproduction had been ensured. Meanwhile, this work will contribute to providing potential new targets for antiviral therapeutics and help experimental research in the future.

1. Introduction

Human papillomavirus (HPV) has been tantamount to cervical cancer which ranked as the third most common cancer and the fourth most common cause of cancer death, but its actual footprint is much bigger [1, 2]. Persistent infection with mucosal HPV types, especially with HPV16, can also lead to the form of penile, vulvar, vaginal, anal, and oropharyngeal cancer, recurrent respiratory papillomatosis, and certain head afflictions [3, 4]. Furthermore, some data show that the actual number of cases of anal and oropharyngeal cancers is increasing and may have already exceeded (or will soon exceed) that of cervical cancer. HPVs were divided into five different genera: Alpha, Beta, Gamma, Mu, and Nu [5, 6]. HPVs were also classified as cutaneous or mucosal according to their tropism. There are both cutaneous and mucosal HPV for Alphapapillomavirus. Other genera are cutaneous. In addition, 12 mucosal HPVs (HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59) were classified as high-risk (HR) HPV types by the International Agency for Research on Cancer (IARC) in 2009 [7, 8]. More than 96.6% of cervical cancer is caused by HR HPVs, while about 54.4% is caused by HPV16. In all HPV-positive noncervical cancers, HPV16 is also the most common HPV type detected. The HPV16 encodes eight proteins: E1, E2, E4, E5, E6, E7, L1, and L2 [9, 10]. These proteins are classified as adaptive proteins which have high carcinogenicity (E5, E6, and E7) and core set (E1, E2, L2, and L1). The E4 protein is embedded within the E2 protein [11].

HPV16 appears to be extraordinary: how can such a small amount of proteins do so much [12]? Protein-protein interaction (PPI) network is a feasible strategy to improve our understanding of its pathogenesis. Several human-pathogen

interaction networks have been reported, such as *Plasmodium falciparum*, *Yersinia pestis*, hepatitis C virus (HCV), and Epstein-Barr virus (EBV) [13–16]. Dyer et al. integrated and compared publicly available human-pathogen PPIs from 190 different pathogens to provide a global view of pathogenesis strategies [17]. Unfortunately, it is very limited that PPI pairs between HPV16 and human are obtained by experiment. Therefore, computational methods to predict PPIs have an important role [18]. The SVM with 217-dimensional vector was employed to predict the interactions of HPV16 and HPV18 proteins with human proteins by Cui et al. at the same time [19]. But it is easy to lead overfitting for small sample. In this paper, a new method was employed to represent protein sequence. A support vector machine (SVM) model with sequence and topological features was built to predict new interactions between HPV16 and human proteins. Subsequently, all interactions were filtered and further analyzed by some strategies.

2. Methods

2.1. Data Sources. We collected human PPIs from large-scale high-throughput screens [20–22] and several interaction databases [23–26], which contained 193,801 interactions among 13,306 proteins. The Pathway Interaction Database (PID) is a growing collection of human signaling and regulatory pathways curated from peer-reviewed literature [27]. As a source of reliable information we extracted about 224 different pathways from the PID. Then the interactions between HPV16 and human proteins were extracted from IntAct [28], APID [29], and VirHostNet [30]. After removing redundancy, a total of 174 interactions were identified and used as positive training set (see Table S1 in the supplementary material available online at http://dx.doi.org/10.1155/2015/890381).

We collected 254 new nonredundant interaction pairs from the literature (see Table S2 in the supplementary material). Finally, the 254 interaction pairs were used as positive test set. It should be noted that whether it was positive training set or positive test set, the interactions were centered on E6 protein and E7 protein because of experimental biases.

2.2. Choosing of Negative Set. As a 2-class classification, both positive set and negative set are needed [31]. Positive set is interacting pairs and negative set is noninteracting pairs. Unfortunately, the noninteracting pairs were not readily available. In the absence of negative set, the following strategy was adopted to choose negative set. This strategy was based on such an assumption that proteins locating different subcellular localizations do not interact [32]. First, the all human proteins of human PPI network were grouped into eight subsets based on the eight main types of localization—cytoplasm, nucleus, mitochondrion, endoplasmic reticulum, golgi apparatus, peroxisome, cytoplasm&nucleus and secreted. Then we totaled subsets of human proteins which were targeted by a kind of HPV16 protein denoted as h. Therefore, other proteins that did not appear in those subsets were made as candidates who did not interact with h. Finally, the same amounts of proteins with targeted human proteins of h were randomly picked as negative set of h. For example, eight human proteins targeted by E5 protein occupied cytoplasm subset and nucleus subset in positive training set; thus, other human proteins which did not appear in those two subsets were made as candidates and eight proteins of candidates were randomly picked as negative training set of E5 protein.

2.3. Feature Extraction. The sequence compositions of the protein pair and the topological features of corresponding human protein were employed to represent protein interaction between HPV16 and human.

In accordance with Shen et al. [33], a protein sequence was represented by three consecutive amino acids. On account of limited sample, however, another class of amino acids was used to reduce the dimension of the vector space of feature vectors. Based on the chemical nature of the side chain of the amino acid, twenty amino acids were classified into five categories: {GAVLIMP}, {STCNQ}, {KRH}, {ED}, and {FYW}. The third category and the fourth category were incorporated into one category, and four categories were considered in total. So there are $4 \times 4 \times 4 = 64$ possible amino acid combinations. The frequency of a combination k in a protein i was defined as $f_{ik} = n_{ik} / \sum_{l=1}^{64} n_{il}$, where n_{ik} was the occurrences of combination k in protein i. An interaction between a HPV protein i and a human protein j was represented by their frequency difference, $d_{ij} = f_i - f_j$. The parameter d_{ij} was normalized by

$$n_{ijk} = 2 \times \frac{d_{ijk} - \min\left\{d_{ij1}, d_{ij2}, \ldots, d_{ij64}\right\}}{\max\left\{d_{ij1}, d_{ij2}, \ldots, d_{ij64}\right\} - \min\left\{d_{ij1}, d_{ij2}, \ldots, d_{ij64}\right\}} - 1, \quad (1)$$

where d_{ijk} is the frequency difference of the kth combination. The numerical value of n_{ijk} ranges from −1 to 1.

Besides the standardized frequency difference, degree and betweenness of the human proteins were also used as features. Ultimately, a 66-dimensional vector was built to represent each protein pair. Each interaction was labeled +1 and noninteraction was labeled −1.

The classification model for predicting PPIs was based on support vector machine (SVM) using LIBSVM [34] with the radial basis function (RBF).

There are three differences between our representation and that of Cui et al. [19]. First, twenty amino acids were classified into six classes by Cui et al.: {IVLM}, {FYW}, {HKR}, {DE}, {QNTP}, and {ACGS}. So there are $6 \times 6 \times 6 = 216$ possible amino acid combinations. Second, standardization was done by

$$d_i = \left\{e^{(f_i - \min\{f_1, f_2, \ldots, f_{216}\}) / (\max\{f_1, f_2, \ldots, f_{216}\} - \min\{f_1, f_2, \ldots, f_{216}\})}\right\} - 1. \quad (2)$$

Third, a feature element was used to represent the types of virus proteins and was included in a feature vector.

2.4. Tissue Specificity Filtering. To ensure utmost biological relevance, tissue specificity filtering was adopted. It has been known that HPVs infect epithelial cells in oral mucosa or skin [6]. In addition, HPVs also lead to recurrent respiratory papillomatosis, head afflictions, and cancers of the cervix uteri, vulva, anus, and oropharynx (including base of the tongue and tonsils) and interact with basal cell and the immune system [3, 35]. We extracted proteins in those cells, tissues, and systems from HPRD [26]. Finally, interactions were filtered by selecting interaction pairs which only contain those proteins.

2.5. Enrichment and Pathway Participation Coefficient. The two parameters have been described by Wuchty et al. [13, 36] in detail. But for the sake of completeness, we would describe the two parameters in brief.

Proteins were grouped according to their degree in integrated human PPI network. Each group where each protein has at least k interactions was represented by $N_{\geq k}$. In each group the number of human proteins that were targeted by HPV16, $N_{t,\geq k}$, was calculated. As a null hypothesis, we randomly sampled protein set from the integrated human PPI network and then calculated corresponding number of targeted proteins, $N^r_{t,\geq k}$. Finally the enrichment of targeted proteins was defined as $E_{t,\geq k} = N_{t,\geq k}/N^r_{t,\geq k}$. In addition to degree, the same calculation was performed for betweenness. It was noted that $E > 1$ points to an enrichment and vice versa.

For each protein i that was involved in pathways and the integrated human PPI network, the corresponding pathway participation coefficient (PPC) in the total set of pathways P was defined as $\mathrm{PPC}_i = \sum_{p\in P}[|\Gamma(i) \in p| / \sum_{p\in P} \Gamma(i) \in p|]^2$, where $\Gamma(i) \in p$ was the set of interaction partners of i in the pathway p. If a protein predominantly interacted with partners that were members of the same pathway, PPC tended toward 1. Otherwise PPC tended to 0.

2.6. GO Term Enrichment. The Gene Ontology (GO) is a hierarchically organized, controlled vocabulary to consistently describe and annotate gene products [37]. GO term enrichment was performed using the DAVID Functional Annotation Chart tool [38, 39]. GO terms are controlled vocabularies that form a directed acyclic graph (DAG), whereby individual terms are represented as nodes connected to more specific nodes by directed edges, such that each term is a more specific child of one or more parents. Therefore, to avoid very general and uninformative GO terms, only GO level 5 terms were considered. The P values were corrected for multiple testing using the Bonferroni procedure and transformed by taking the $-\log_{10}$ for easier visualization [40, 41].

3. Results and Discussion

We extracted 174 interactions between HPV16 and human proteins and integrated a human PPI network including 193,801 interactions. A flowchart of the whole experiment is shown in Figure 1.

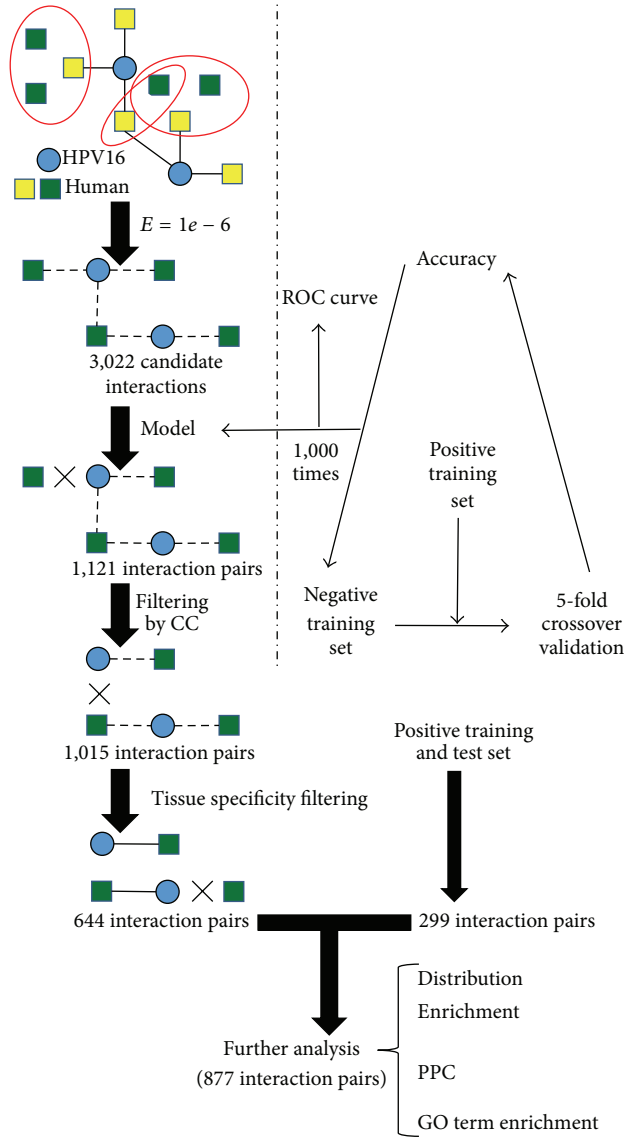


Figure 1: Flowchart to integrate and analyze PPI network between HPV16 and human proteins. A candidate interaction was found, if the human protein had homologs in the human PPI network. This method provided 3,022 candidate interactions. An SVM model was employed to evaluate candidate interactions and 1,121 interactions were left. Subsequently, these interactions were filtered if human proteins with targeted human proteins had the same as cellular component. 1,015 interactions were obtained; positive training set and test set were further filtered by tissue specificity. Finally, 877 interactions were obtained and analyzed. Solid lines delineate validated interactions between virus and human proteins, and dotted lines delineate candidate interactions which would be validated. Homologous proteins are surrounded by ellipse.

3.1. Choosing of Negative Training Set and Evaluating of Model. The 174 interactions between virus and human proteins were used as positive training set. The selection of negative training set was fundamental to the reliability of the prediction model [33]. Based on a rational assumption, the negative training set was chosen (see Methods section). The SVM

with 5-fold cross-validation was employed to optimize the parameters and check the reliability of randomly selected negative training set. Repeating such random trials 1,000 times and calculating average accuracy (81.3±1.3%), we chose a result approaching average accuracy to build model and plot ROC curve (Figure 2) which allowed for a true positive rate TPR = 74.71%, a false positive rate FPR = 8.62%, and area under the curve AUC = 0.8627. Other results were dotted clouds. It was demonstrated that the method of choosing negative training set was significantly reliable and robust.

To evaluate expansibility of the model, a positive test set was collected. Negative test set was selected by the same method with choosing negative training set. Repeating trials 1,000 times, this model, on average, achieved an accuracy AC = 80.0 ± 1.8%, TPR = 78.7%, and FPR = 18.2 ± 3.6%. For comparison, we tested the method of Cui et al. on whole modeling and evaluating. Our method outperformed the method of Cui et al., which, on average, achieved AC = 57.25± 1.5%, TPR = 63.4%, and FPR = 47.9 ± 3.1%.

3.2. Inferring and Filtering of Candidate Interactions. To find candidate interactions, we ran BLAST with the known targeted human proteins as query sequences against the human proteins in integrated human PPI network. Specifically, we considered a pair of proteins with homology if their E-value was $<10^{-6}$. A candidate interaction was detected between a HPV16 protein and homologous protein of targeted human protein. The final set contained 3022 candidate interactions between 8 virus and 1,950 human proteins.

The model built by SVM was applied to evaluate candidate interactions. The 1,121 interactions between 8 virus and 701 human proteins were finally obtained. The 701 human proteins were refined further by selecting human proteins that have the same GO cellular component terms with homologous human proteins from the positive training set. 1,015 interactions were obtained by this refinement. To ensure utmost biological relevance for the 1,015 interactions, tissue specificity filtering was adopted (see Methods section). Filtering interactions provided 644 interactions between 8 HPV16 proteins and 405 human proteins. For simplicity of reference, the filtering result was named as predicted set. Meanwhile, positive set including training set and test set was also filtered by tissue specificity. Finally, all filtering results were combined, providing a total of 877 interactions between 8 virus and 603 human proteins. This set was called as all set.

3.3. Distribution of Targeted Human Proteins Based on Host-Virus Interaction. Now we paid more attention to the all set. The frequency of human proteins that interacted with the same number of viral proteins was calculated. We observed that most human proteins (69.52%) merely interacted with a virus protein in Figure 3(a). The positive training set and the predicted set were addressed by the same calculation method, and their results illustrated similar trend with all set. It suggested that HPV16 interacted with human proteins as much as possible to enlarge its scope of influence by its limited proteins. In order to provide all necessary cellular proteins required for viral replication, the virus has to keep its host cell in cycle [42]. At the molecular level, virus proteins interact with many key cell cycle regulatory proteins, including cyclin-dependent kinase (CDK), cyclin-dependent kinase inhibitors, and cyclin proteins (Figure 3(b)). Among them, CDK2 and CDK7 are the most prominent. The two proteins simultaneously interact with five virus proteins in all set and the five virus proteins are L2, E4, E5, E6, and E7. Combination of CDK2 and some cyclins regulates G_1/S transition. CDK7 is both a CDK-activating kinase (CAK), which is able to phosphorylate and activate CDK1, CDK2, CDK4, and CDK6 within the activation segment (T-loop) [43–46], and an essential component of the transcription factor TFIIH, which phosphorylates the C-terminal domain (CTD) at Ser 5 of the largest subunit of Pol II [47–49]. These interactions, together with other proteins that bind to HPV16, alter a broad array of cell cycle progression; for example, they block cellular proliferation by causing cell cycle arrest in S-phase [12, 50, 51]. The myosin light chain kinase (MLCK) is also targeted by five virus proteins. It has been proven that MLCK plays a role in the regulation of epithelial cell survival [52] and modulates hypotonicity-induced Ca^{2+} entry and Cl^- channel activity in human cervical cancer cells [53]. In addition, HPV16 may be similar to arrest defective-1 that controls tumor cell behavior by MLCK [54].

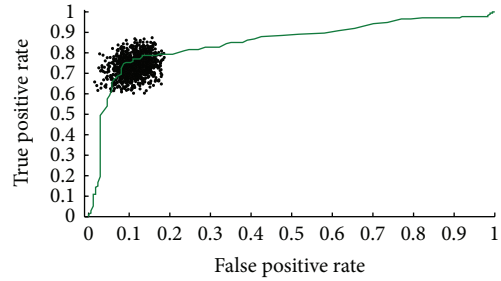
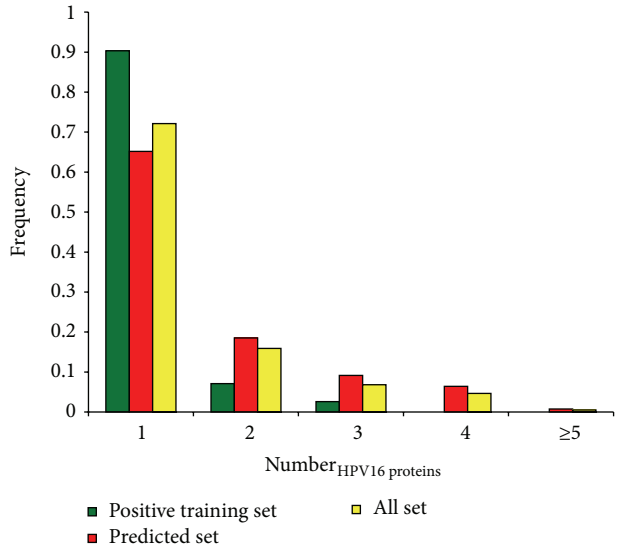
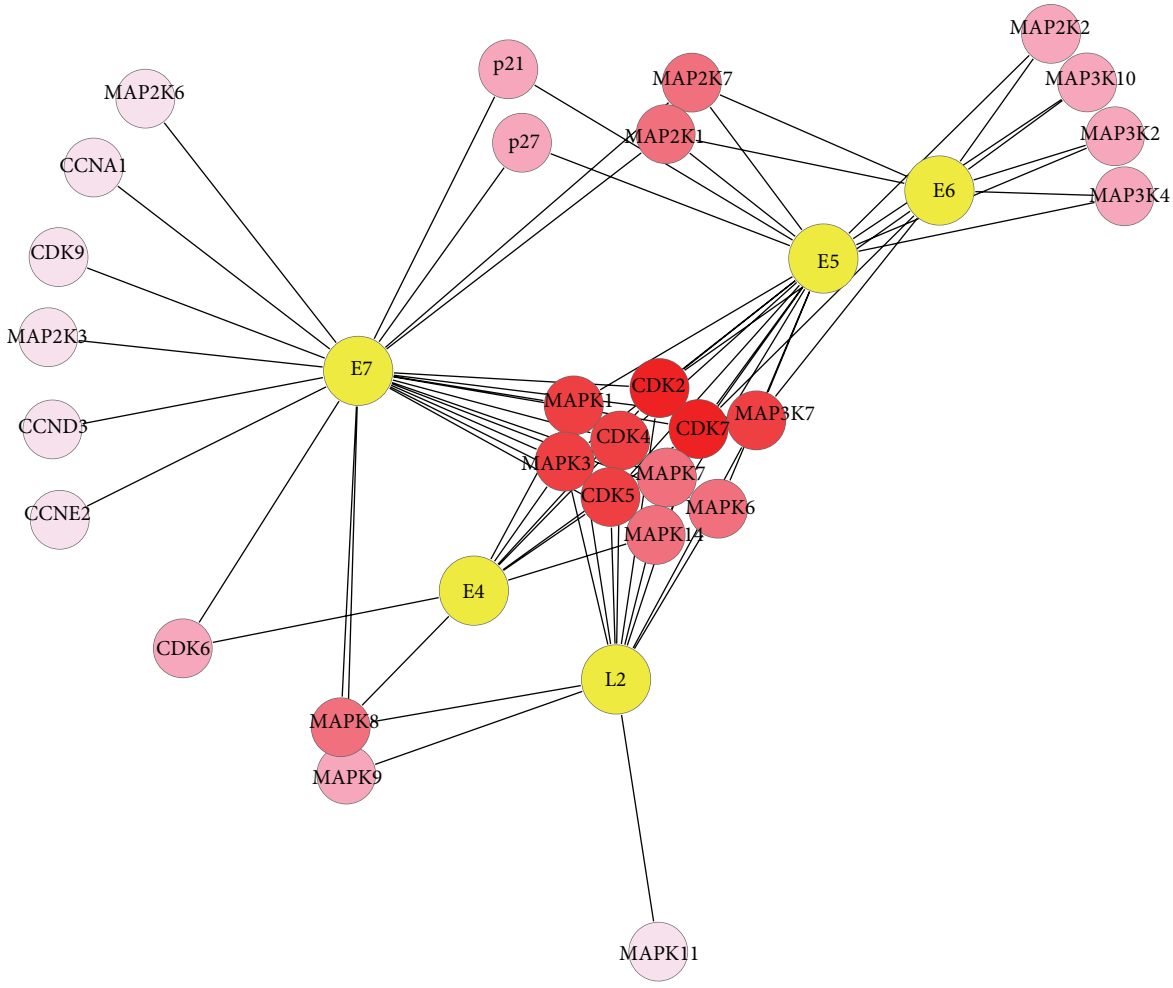


Figure 2: ROC curve of training set. Negative training sets were repeatedly chosen 1,000 times. Applying SVM with 5-fold cross-validation the training sets allowed for a true positive rate TPR = 74.71%, false positive rate FPR = 8.62%, and area under the curve AUC = 0.8627.

3.4. Statistical Implications of Targeted Host Proteins Based on Human PPI Network. We calculated the enrichment of targeted human proteins as a function of the degree of human proteins (see Methods section). With an average over 1,000 randomizations, we observed that whether it was all set, predicted set, or positive training set, HPV16 preferred to interact with hub proteins (proteins interacting with a large number of partners) in the integrated human PPI network (Figure 4(a)). Subsequently, we calculated the enrichment of targeted proteins as a function of the betweenness and consistent trend has shown that bottleneck proteins (proteins that are central to many paths in the network) were more affected by virus (Figure 4(b)). Testing the significance that HPV16 tended to interact with hub and bottleneck proteins, we used Fisher's exact test, allowing us to find a statistically significant tendency that HPV16 is indeed highly

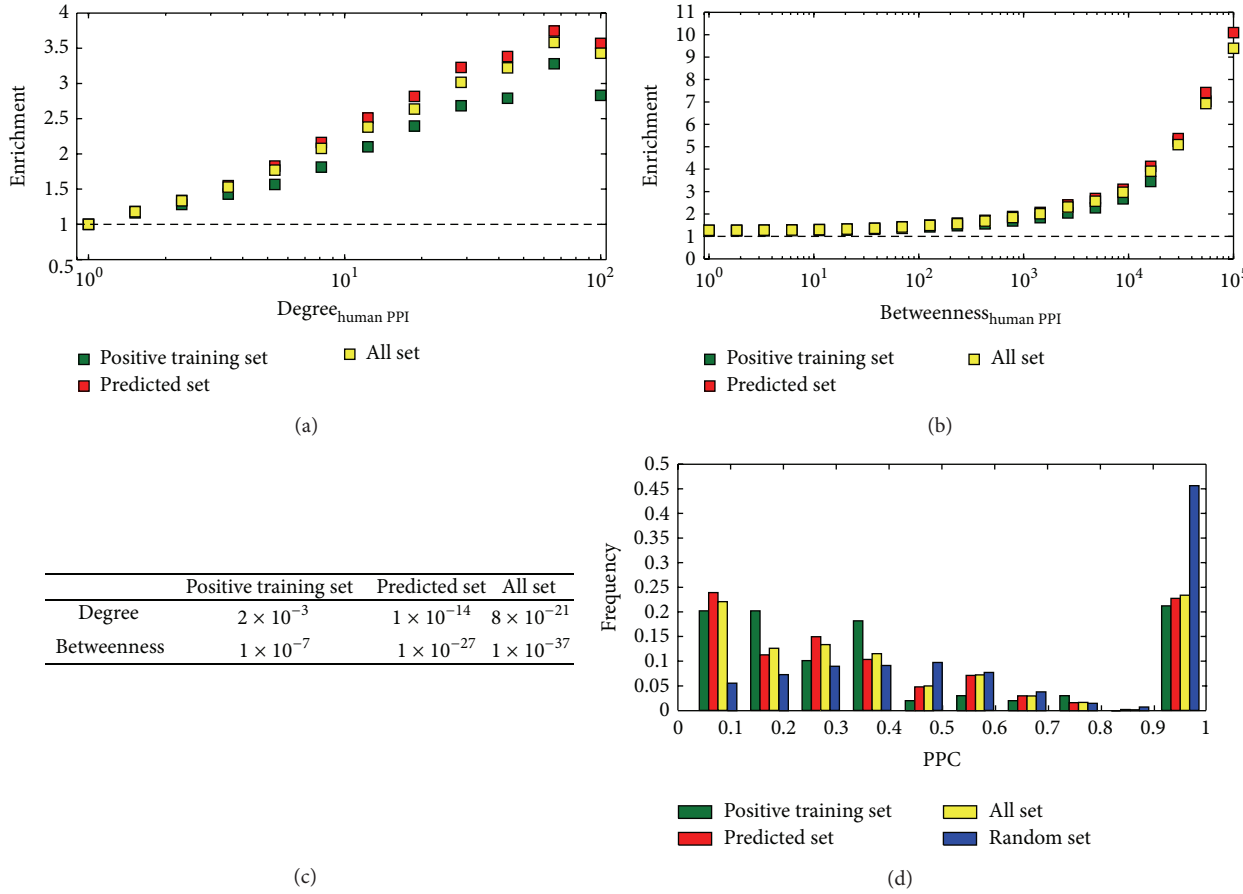


(a)



(b)

Figure 3: Characteristics of PPI network between HPV16 and human. (a) Whether it was training positive set, predicted set, or all set, a majority of host proteins interacted with a small amount of virus proteins. (b) A network between five virus proteins and some human proteins about cell cycle and phosphorylation cascade. The more virus proteins human protein is targeted by, the darker the node color is.



(c)

	Positive training set	Predicted set	All set
Degree	2×10^{-3}	1×10^{-14}	8×10^{-21}
Betweenness	1×10^{-7}	1×10^{-27}	1×10^{-37}

FIGURE 4: Characteristics of targeted human proteins. (a) The enrichment of targeted human proteins as a function of their degree was calculated. It indicated that hub proteins appeared to be primarily targeted. (b) Analogously, HPV16 tended to interact with bottleneck proteins. (c) P values of Fisher's exact tests indicated that HPV16 is highly prone to interact with hub proteins and bottleneck proteins. (d) Considering all set, most proteins have low pathway participation coefficients, which indicated that HPV16 reached into a breadth of signaling pathways. Such a result was shown by positive training set and predicted set.

prone to interact with hub proteins and bottleneck proteins (Figure 4(c)).

We speculated that virus interacted with human proteins as much as possible while tending to influence many signaling pathways to mediate the infection. PPC was adopted to measure this tendency (see Methods section). Focusing on the positive training set, we observed that most human proteins occurred in a variety of pathways through its interaction partners in integrated human PPI network (Figure 4(d)). The predicted set and the all set showed more enforced maxima around low values of PPC. As a comparison, we randomly selected a subset of equal size with human proteins in all set from integrated human PPI network and repeated 1,000 times to calculate average value of PPC. Ignoring the last bar, we found that the random set obeyed the normal distribution, but the all set was linear relationship. Such results strongly indicated that the HPV16 effectively affected a breadth of signaling pathways [13, 55, 56].

3.5. Functional Analysis of Targeted Host Proteins. GO term enrichment was employed to perform the comprehensive functional analysis for human proteins of the all set. The main advantage of this approach is that we can make use of term-term relationships, in which joint terms may contain unique biological meaning for a given study [57].

For all targeted human proteins, significant enrichment was observed in the processes of phosphorylation, metabolism, signaling, cell death and apoptosis, gene expression, and positive or negative regulation terms (Figure 5(a)). This observation was also reflected on the functions which include kinase activity, receptor activity, promoter, DNA binding, and so on (Figure 5(b)). MAPK is a particularly important component in protein kinase phosphorylation cascade. It can enter the nucleus and phosphorylate serine/threonine residues of substrate proteins which contain transcription factors of regulating the cell cycle and cell differentiation. Notably, viral proteins strongly interacted with members of the MAPK family (MAPK1, 3, 6, 7, 8, 9, 11, and 14). Besides MAPK family, partial members of MAP2K and MAP3K family were also targeted (Figure 3(b)). HPV16 controls phosphorylation cascade so that cell behaviors including cell proliferation and differentiation, cell survival, and apoptosis are broken.

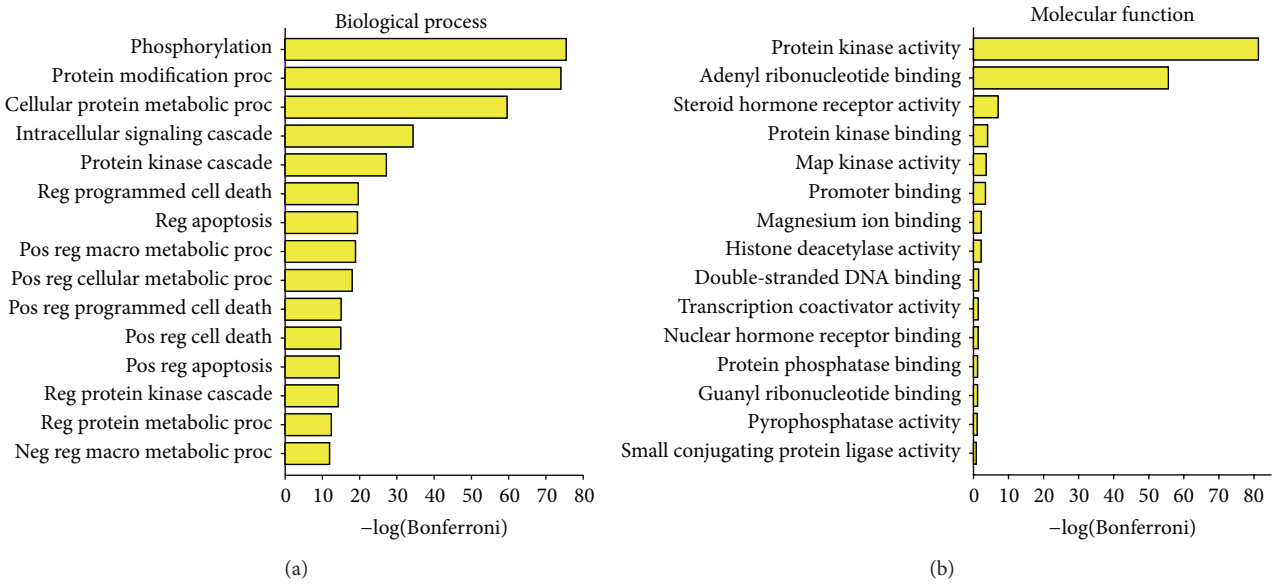


Figure 5: GO term enrichment of all targeted human proteins. (a) Enriched GO biological process terms. (b) Enriched GO molecular function terms. Here only fifteen most significant terms are shown. Bonferroni collected P values were transformed by $-\log_{10}$. The following abbreviations are used: "reg" is "regulation of," "pos" is "positive," "neg" is "negative," "proc" is "process," "macro" is "macromolecule," and "bsyn" is "biosynthetic."

Five proteins (E1, E2 (and E4), L1, and L2) are encoded by all known PVs. There is a hypothesis that the ancestral papillomavirus did not contain adaptive proteins and only need the core set to meet the basic requirements of a viral infection [11]. In the process of evolution, HPV16 produced all of the adaptive proteins. It was surprising that the top four of biological process enrichment of all adaptive proteins were the same as core set's top four, and then processes involving apoptosis and death were enriched for core set (Figures 6(a) and 6(c)). This showed that HPV16 would evolve carcinogenicity, but only on the condition that its own reproduction had been ensured. The E4 protein has the functions of adaptive proteins and core set (Figure 6) but prefers the latter. In other words, as a part of the proteins encoded by all known PVs E4 must first guarantee viral reproduction and then together with adaptive proteins enhance the carcinogenicity of HPV16.

4. Conclusions

Significant challenges currently impair experiments to get a more complete map of interactions between HPV16 and human proteins, facilitating computational methods to detect potential interactions. Sequence features are popular because of its simplicity and availability. SVM has been shown to perform well in multiple areas including detecting remote protein homologies, evaluating microarray expression data, and checking new interactions [33, 58, 59]. On the basis of facts above we predicted new interactions between HPV16 and human proteins. The predicted set and other known interactions were integrated and filtered, providing a total of 877 interactions between 8 virus and 603 human proteins. According to the interactions between the virus and human proteins, we plotted the distribution of targeted host proteins. The distribution showed that the virus enlarged its scope of influence by interacting with host proteins as much as possible. HPV16 alters a broad array of cell cycle progression by a number of PPIs. Utilizing integrated human PPI network the enrichment of targeted host proteins as a function of their degree or betweenness was calculated. Results suggested that HPV16 was highly prone to interact with hub proteins and bottleneck proteins, perhaps because these proteins control critical processes in the human cell [17]. PPC was used as a measure of diversity. In the light of their distributions, targeted human proteins effectively mediated the diversity of influenced signaling pathways which helps virus mediate the infection. GO term enrichment was utilized to perform the comprehensive functional analysis. We found that cell behaviors of host cell were broken; the HPV16 produced many other functions by evolution, but it was based on the premise that its own reproduction has been guaranteed.

The integration and analysis of virus-host interactions boosts our knowledge about the function of HPV16 proteins and relations between virus and human proteins. These results improve our understanding of HPV16 pathogenesis and provide potential new targets for interfering with either HPV16 or human at key points in the infection. Our results may point to important areas of research to guide further experimental studies.

Conflict of Interests

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

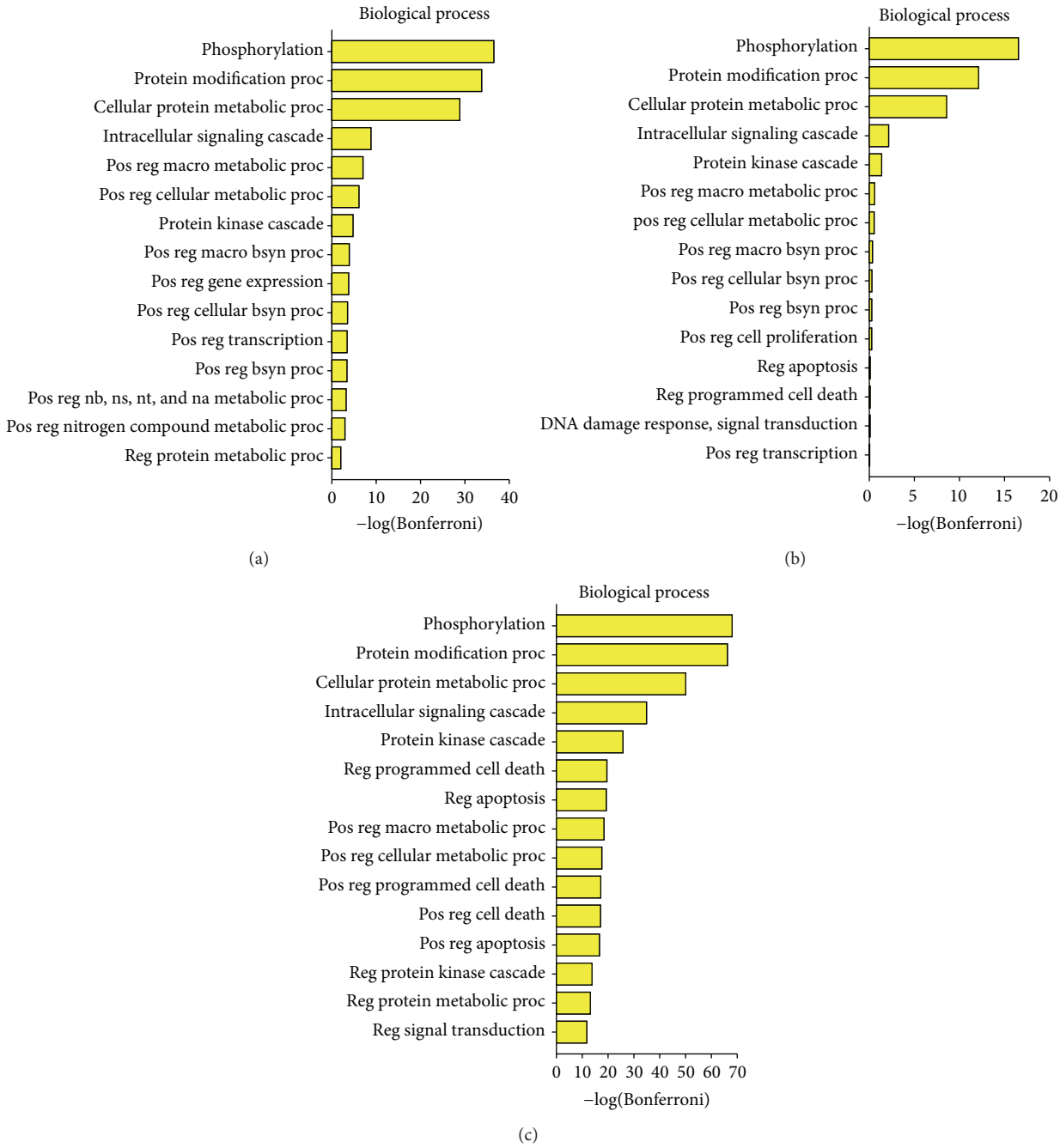


Figure 6: Significantly enriched GO biological process terms. (a) Enriched results of E1, E2, L2, and L1. (b) Enriched results of E4. (c) Enriched results of E5, E6, and E7. Here only fifteen most significant terms are shown. Bonferroni collected P values were transformed by $-\log_{10}$.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (21375095) and the National Natural Science Foundation of China (21305096).

References

[1] J. M. Crow, "HPV: the global burden," *Nature*, vol. 488, no. 7413, pp. S2–S3, 2012.

[2] M. Arbyn, X. Castellsagué, S. de sanjosé et al., "Worldwide burden of cervical cancer in 2008," *Annals of Oncology*, vol. 22, no. 12, Article ID mdr015, pp. 2675–2686, 2011.

[3] F. X. Bosch, T. R. Broker, D. Forman et al., "Comprehensive control of human papillomavirus infections and related diseases," *Vaccine*, vol. 31, no. 5, pp. F1–F31, 2013.

[4] M. L. Tornesello, L. Buonaguro, P. Giorgi-Rossi, and F. M. Buonaguro, "Viral and cellular biomarkers in the diagnosis of cervical intraepithelial neoplasia and cancer," *BioMed Research International*, vol. 2013, Article ID 519619, 10 pages, 2013.

[5] H.-U. Bernard, R. D. Burk, Z. Chen, K. van Doorslaer, H. Z. Hausen, and E.-M. de Villiers, "Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments," *Virology*, vol. 401, no. 1, pp. 70–79, 2010.

[6] D. Bzhalava, P. Guan, S. Franceschi, J. Dillner, and G. Clifford, "A systematic review of the prevalence of mucosal and cutaneous human papillomavirus types," *Virology*, vol. 445, no. 1-2, pp. 224–231, 2013.

[7] M. Schiffman, G. Clifford, and F. M. Buonaguro, "Classification of weakly carcinogenic human papillomavirus types: addressing the limits of epidemiology at the borderline," *Infectious Agents and Cancer*, vol. 4, no. 1, article 8, 2009.

[8] Humans IWGotEoCRt, "Biological agents. Volume 100 B. A review of human carcinogens," *IARC monographs on the evaluation of carcinogenic risks to humans/World Health Organization, International Agency for Research on Cancer. Part B*, vol. 100, pp. 1–441, 2012.

[9] C. N. Hansen, L. Nielsen, and B. Norrild, "Activities of E7 promoters in the human papillomavirus type 16 genome during cell differentiation," *Virus Research*, vol. 150, no. 1-2, pp. 34–42, 2010.

[10] J. W. Wang and R. B. S. Roden, "L2, the minor capsid protein of papillomavirus," *Virology*, vol. 445, no. 1-2, pp. 175–186, 2013.

[11] K. van Doorslaer, "Evolution of the Papillomaviridae," *Virology*, vol. 445, no. 1-2, pp. 11–20, 2013.

[12] S. B. Vande Pol and A. J. Klingelhutz, "Papillomavirus E6 oncoproteins," *Virology*, vol. 445, no. 1-2, pp. 115–137, 2013.

[13] S. Wuchty, "Computational prediction of Host-Parasite protein interactions between *P. falciparum* and *H. sapiens*," *PLoS ONE*, vol. 6, no. 11, Article ID e26960, 2011.

[14] H. Yang, Y. Ke, J. Wang et al., "Insight into bacterial virulence mechanisms against host immune response via the *Yersinia pestis*-human protein-protein interaction network," *Infection and Immunity*, vol. 79, no. 11, pp. 4413–4424, 2011.

[15] M. A. Calderwood, K. Venkatesan, L. Xing et al., "Epstein-Barr virus and virus human protein interaction maps," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 18, pp. 7606–7611, 2007.

[16] B. de Chassey, V. Navratil, L. Tafforeau et al., "Hepatitis C virus infection protein network," *Molecular Systems Biology*, vol. 4, no. 1, article 230, 2008.

[17] M. D. Dyer, T. M. Murali, and B. W. Sobral, "The landscape of human proteins interacting with viruses and other pathogens," *PLoS Pathogens*, vol. 4, no. 2, 2008.

[18] S. J. Wodak and R. Méndez, "Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications," *Current Opinion in Structural Biology*, vol. 14, no. 2, pp. 242–249, 2004.

[19] G. Cui, C. Fang, and K. Han, "Prediction of protein-protein interactions between viruses and human by an SVM model," *BMC Bioinformatics*, vol. 13, supplement 7, article S5, 2012.

[20] R. M. Ewing, P. Chu, F. Elisma et al., "Large-scale mapping of human protein-protein interactions by mass spectrometry," *Molecular Systems Biology*, vol. 3, article 89, 2007.

[21] J.-F. Rual, K. Venkatesan, T. Hao et al., "Towards a proteome-scale map of the human protein-protein interaction network," *Nature*, vol. 437, no. 7062, pp. 1173–1178, 2005.

[22] U. Stelzl, U. Worm, M. Lalowski et al., "A human protein-protein interaction network: a resource for annotating the proteome," *Cell*, vol. 122, no. 6, pp. 957–968, 2005.

[23] D. Croft, G. O'Kelly, G. Wu et al., "Reactome: a database of reactions, pathways and biological processes," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D691–D697, 2011.

[24] L. Licata, L. Briganti, D. Peluso et al., "MINT, the molecular interaction database: 2012 update," *Nucleic Acids Research*, vol. 40, no. D1, pp. D857–D861, 2012.

[25] S. Kerrien, B. Aranda, L. Breuza et al., "The IntAct molecular interaction database in 2012," *Nucleic Acids Research*, vol. 40, no. 1, pp. D841–D846, 2012.

[26] T. S. K. Prasad, R. Goel, K. Kandasamy et al., "Human protein reference database—2009 update," *Nucleic Acids Research*, vol. 37, no. 1, pp. D767–D772, 2009.

[27] C. F. Schaefer, K. Anthony, S. Krupa et al., "PID: the pathway interaction database," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D674–D679, 2009.

[28] S. Orchard, M. Ammari, B. Aranda et al., "The MIntAct project—intAct as a common curation platform for 11 molecular interaction databases," *Nucleic Acids Research*, vol. 42, no. 1, pp. D358–D363, 2014.

[29] C. Prieto and J. de Las Rivas, "APID: agile protein interaction DataAnalyzer," *Nucleic Acids Research*, vol. 34, pp. W298–W302, 2006.

[30] V. Navratil, B. de chassey, L. Meyniel et al., "VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks," *Nucleic Acids Research*, vol. 37, no. 1, pp. D661–D668, 2009.

[31] S. Mei, "Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins," *PLoS ONE*, vol. 8, no. 11, Article ID e79606, 2013.

[32] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences," *Nucleic Acids Research*, vol. 36, no. 9, pp. 3025–3030, 2008.

[33] J. Shen, J. Zhang, X. Luo et al., "Predicting protein-protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 11, pp. 4337–4341, 2007.

[34] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[35] E. Lukesova, J. Boucek, E. Rotnaglova et al., "High level of tregs is a positive prognostic marker in patients with HPV-positive oral and oropharyngeal squamous cell carcinomas," *BioMed Research International*, vol. 2014, Article ID 303929, 11 pages, 2014.

[36] S. Wuchty, G. Siwo, and M. T. Ferdig, "Viral organization of human proteins," *PLoS ONE*, vol. 5, no. 8, Article ID e11796, 2010.

[37] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[38] G. Dennis Jr., B. T. Sherman, D. A. Hosack et al., "DAVID: database for annotation, visualization, and integrated discovery," *Genome Biology*, vol. 4, no. 5, 2003.

[39] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[40] J. M. Doolittle and S. M. Gomez, "Mapping protein interactions between dengue virus and its human and insect hosts," *PLoS Neglected Tropical Diseases*, vol. 5, no. 2, article e954, 2011.

[41] J. M. Doolittle and S. M. Gomez, "Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens," *Virology Journal*, vol. 7, article 82, 2010.

[42] M. Habig, H. Smola, V. S. Dole, R. Derynck, H. Pfister, and S. Smola-Hess, "E7 proteins from high- and low-risk human papillomaviruses bind to TGF-β-regulated Smad proteins and inhibit their transcriptional activity," *Archives of Virology*, vol. 151, no. 10, pp. 1961–1972, 2006.

[43] S. Larochelle, J. Pandur, R. P. Fisher, H. K. Salz, and B. Suter, "Cdk7 is essential for mitosis and for in vivo Cdk-activating kinase activity," *Genes and Development*, vol. 12, no. 3, pp. 370–381, 1998.

[44] M. M. Schachter and R. P. Fisher, "The CDK-activating kinase Cdk7: taking yes for an answer," *Cell Cycle*, vol. 12, no. 20, pp. 3239–3240, 2013.

[45] X. Bisteau, S. Paternot, B. Colleoni et al., "CDK4 T172 phosphorylation is central in a CDK7-dependent bidirectional CDK4/CDK2 interplay mediated by p21 phosphorylation at the restriction point," *PLoS Genetics*, vol. 9, no. 5, Article ID e1003546, 2013.

[46] M. M. Schachter, K. A. Merrick, S. Larochelle et al., "A Cdk7-Cdk4 T-loop phosphorylation cascade promotes G1 progression," *Molecular Cell*, vol. 50, no. 2, pp. 250–260, 2013.

[47] J. W. Harper and S. J. Elledge, "The role of Cdk7 in CAK function, a retro-retrospective," *Genes & Development*, vol. 12, no. 3, pp. 285–289, 1998.

[48] R. P. Fisher, "Secrets of a double agent: CDK7 in cell-cycle control and transcription," *Journal of Cell Science*, vol. 118, no. 22, pp. 5171–5180, 2005.

[49] J. Holcakova, P. Muller, P. Tomasec et al., "Inhibition of post-transcriptional RNA processing by CDK inhibitors and its implication in anti-viral therapy," *PLoS ONE*, vol. 9, no. 2, Article ID e89228, 2014.

[50] M. Bergvall, T. Melendy, and J. Archambault, "The E1 proteins," *Virology*, vol. 445, no. 1-2, pp. 35–56, 2013.

[51] J. Doorbar, "The E4 protein; structure, function and patterns of expression," *Virology*, vol. 445, no. 1-2, pp. 80–98, 2013.

[52] L. E. Connell and D. M. Helfman, "Myosin light chain kinase plays a role in the regulation of epithelial cell survival," *Journal of Cell Science*, vol. 119, no. 11, pp. 2269–2281, 2006.

[53] M.-R. Shen, P. Furla, C.-Y. Chou, and C. J. Ellory, "Myosin light chain kinase modulates hypotonicity-induced Ca^{2+} entry and Cl- channel activity in human cervical cancer cells," *Pflugers Archiv European Journal of Physiology*, vol. 444, no. 1-2, pp. 276–285, 2002.

[54] D. H. Shin, Y.-S. Chun, K.-H. Lee, H.-W. Shin, and J.-W. Park, "Arrest defective-1 controls tumor cell behavior by acetylating myosin light chain kinase," *PLoS ONE*, vol. 4, no. 10, Article ID e7451, 2009.

[55] B. Kaczkowski, M. Rossing, D. K. Andersen et al., "Integrative analyses reveal novel strategies in HPV11,–16 and −45 early infection," *Scientific Reports*, vol. 2, article 515, 2012.

[56] C. Pérez-Plasencia, G. Vázquez-Ortiz, R. López-Romero, P. Piña-Sanchez, J. Moreno, and M. Salcedo, "Genome wide expression analysis in HPV16 cervical cancer: identification of altered metabolic pathways," *Infectious Agents and Cancer*, vol. 2, no. 1, article 16, 2007.

[57] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.

[58] T. Jaakkola, M. Diekhans, and D. Haussler, "Using the Fisher kernel method to detect remote protein homologies," *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB '99)*, pp. 149–158, 1999.

[59] M. P. S. Brown, W. N. Grundy, D. Lin et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 1, pp. 262–267, 2000.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 746970, 17 pages
http://dx.doi.org/10.1155/2015/746970

Research Article

Identification of Novel Breast Cancer Subtype-Specific Biomarkers by Integrating Genomics Analysis of DNA Copy Number Aberrations and miRNA-mRNA Dual Expression Profiling

Dongguo Li,[1] Hong Xia,[1] Zhen-ya Li,[2] Lin Hua,[1] and Lin Li[1]

[1]*Institute of Biomedical Engineering, Capital Medical University, Beijing 100069, China*
[2]*Institute of Basic Medical Science, Peking Union Medical College, Qinghua University, No. 5 Dong Dan San Tiao, Beijing 100005, China*

Correspondence should be addressed to Lin Hua; hualin7750@139.com and Lin Li; lil@ccmu.edu.cn

Received 16 July 2014; Revised 15 September 2014; Accepted 22 September 2014

Academic Editor: Jiangning Song

Breast cancer is a heterogeneous disease with well-defined molecular subtypes. Currently, comparative genomic hybridization arrays (aCGH) techniques have been developed rapidly, and recent evidences in studies of breast cancer suggest that tumors within gene expression subtypes share similar DNA copy number aberrations (CNA) which can be used to further subdivide subtypes. Moreover, subtype-specific miRNA expression profiles are also proposed as novel signatures for breast cancer classification. The identification of mRNA or miRNA expression-based breast cancer subtypes is considered an instructive means of prognosis. Here, we conducted an integrated analysis based on copy number aberrations data and miRNA-mRNA dual expression profiling data to identify breast cancer subtype-specific biomarkers. Interestingly, we found a group of genes residing in subtype-specific CNA regions that also display the corresponding changes in mRNAs levels and their target miRNAs' expression. Among them, the predicted direct correlation of BRCA1-miR-143-miR-145 pairs was selected for experimental validation. The study results indicated that BRCA1 positively regulates miR-143-miR-145 expression and miR-143-miR-145 can serve as promising novel biomarkers for breast cancer subtyping. In our integrated genomics analysis and experimental validation, a new frame to predict candidate biomarkers of breast cancer subtype is provided and offers assistance in order to understand the potential disease etiology of the breast cancer subtypes.

1. Introduction

Luminal-A and basal-like subtypes are two major breast cancer subtypes and have shown significant differences in terms of incidence, risk factors, baseline prognosis, age at diagnosis, and response to treatment [1]. Luminal-A breast cancers that express estrogen receptors (ERs) and/or progesterone receptors (PRs) and are negative for human epidermal growth factor receptor 2 (HER2) expression respond well to endocrine therapy and have a generally favorable prognosis [2]. In contrast, the basal-like subtype is of particular clinical focus due to its high frequency, lack of effective targeted therapies, poor baseline prognosis, and tendency to affect younger women [3]. Therefore, the identification of breast cancer subtype-specific biomarkers will help provide biological value for the breast cancer clinical trials and therapy strategies.

Recently, more and more evidence showed that an increasing number of genomic aberrations were observed in the progression from normal sample to tumour sample [4]. Array comparative genome hybridization (aCGH) studies of tumor copy number states have demonstrated that tumors with similar gene expression subtypes may also share similar DNA copy number aberrations (CNA) [5, 6] which can be used to further subdivide expression classes. In the practice, some early aCGH studies on breast cancer found that the highly amplified genes were overexpressed and the highly

overexpressed genes were amplified. For example, Pollack et al. found that 62% of highly amplified genes show moderately or highly elevated expression, and DNA copy number influences gene expression across a wide range of DNA copy number alterations [7]. By analyzing the linear and nonlinear relationship between gene copy number and expression, Solvang et al. reveal distinct molecular pathways in breast cancer [8]. Recently, CNA coupled with gene expression has been explored for cancer drivers. For example, Akavia et al. developed a computational framework that integrates chromosomal copy number and gene expression data for identifying known drivers of melanoma and predicted multiple novel tumor dependencies [9]. In addition, cancer subtype-specific biomarkers identification has become the important topic which have also demonstrated diverse prognostic power. For example, Liu et al.'s study revealed that gene expression profiles and clinical features show different prognostic power for the five breast cancer subtypes, and gene expression data of the normal-like subgroup contains more valuable prognostic information and survival associated contexts than the other subtypes [10]. Román-Pérez et al. documented the presence of two distinct subtypes of microenvironment, with active versus inactive cancer-adjacent extratumoral microenvironment influencing the aggressiveness and outcome of ER positive human breast cancers [11]. Therefore, identifying the genes that contribute to the instability of cancer phenotype or cancer subtype by integrating DNA copy number aberrations and gene expression would be useful as a clinical predictor of therapeutic response.

Currently, people have studied the microRNA-gene comodule in cancer patients by integrating multiple genomic data including dual microRNA-gene expression and predicted miRNA-gene interactions. For example, Zhang et al. proposed an effective data integration framework using a multiple nonnegative matrix factorization framework and simultaneously integrated additional network in a regularized manner to identify microRNA-gene regulatory modules associated with ovarian cancer [12, 13]. In particular, it is worth noting that jointly analyzing multiple data types will help enhance the understanding of the role of biomarkers in breast cancer pathogenesis and progression [14]. In our previous studies, we prioritized some breast cancer subtype related biomarkers by analyzing miRNA-mRNA dual expression profiling data [15, 16]. At this time, in our extended study, we conducted an integrative data analysis by combining copy number aberrations and miRNA-mRNA dual expression profiling data to further identify breast cancer subtype-specific biomarkers. Our methods were aimed at the discovery of the interconnected regulatory miRNAs-mRNA pairs as well as the identification of important subtype-specific (luminal-A and basal-like) mRNAs or miRNAs. The simultaneous use of miRNA expression microarray, gene expression microarray, array-CGH, and miRNA-mRNA target relationships can give a more comprehensive understanding of the whole genome of the cancerous cells. Our results have shown that the proposed approach is able to produce meaningful gene regulatory networks that are highly relevant to the biological conditions of the data sets. We found that gene sets residing in subtype-specific CNA regions also display the corresponding changes in mRNAs levels and their counterpart miRNAs expressions. BRCA1-miR-143-miR-145 pairs were selected for further experimental validation and showed positive correlation in breast cancer subtypes.

2. Materials and Methods

2.1. Data Source. We introduced three data sets into our integrated data analysis. The first dataset included 180 breast cancer subtype samples along with their copy number data. Due to our focus on two distinct breast cancer subtypes (luminal-A and basal-like) in the present study, we selected all of 52 luminal-A samples and all of 40 basal-like samples from this dataset to perform our analysis. All of these samples include Illumina 109,000 SNP marker DNA copy number data. The data are available from the Gene Expression Omnibus series GSE10893 [17]. The second dataset included the mRNA expression levels of 23,256 genes, which are also available from GSE10893. We then selected the mRNA levels of 241 samples (158 luminal-A subtypes and 83 basal-like subtypes). The third dataset reported by Enerly et al. [14] is a miRNA-mRNA dual expression profiling dataset (GSE19536). For this dataset, we selected the miRNA and mRNA expression data of 15 basal-like samples and 41 luminal-A samples. The original miRNA microarrays covered 799 miRNAs arisen from the Agilent Technologies. miRNA expression status was scored as present or absent for each gene in each sample by default settings. miRNAs in samples that were run in replicates were considered present if scored in one of the two arrays. Those miRNAs that were detected in less than 10% of the samples were excluded. This filtering resulted in 489 miRNAs considered to be expressed in this set of human breast tumors. SAM (significance analysis of microarrays) method [18] was used to identify statistically significant differential expression of miRNAs and mRNAs which distinguished the reciprocal basal-like and luminal-A breast cancer subtypes. According to a filtering criterion of $P < 0.05$ and false discovery rates (FDR) < 0.1, 201 differentially expressed miRNAs and 8,796 differentially expressed mRNAs for the third dataset were identified (see our previous study [16]). For each identified miRNA, we obtained its target genes from MicroCosm Targets database (http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/), in which the candidate miRNA-target relationships were mostly predicted by miRanda algorithm [19].

2.2. Assessment of Tumor Genomics DNA Copy Number Changes. For the copy number data in the first dataset, we used a sparse Bayesian learning (SBL) model and backward elimination (BE) procedure to find candidate CNA regions. A SBL model was used to find the most likely candidate breakpoints for the copy number state, and the BE procedure was used to remove sequentially the least significant breakpoints estimated by the SBL model, allowing a flexible adjustment of the false discovery rate (FDR) [20]. Therefore, this method provides great flexibility in adjusting the final breakpoint set, and we can obtain a list of breakpoints with lower FDR (<5%) by adjusting the corresponding parameters. We used

gada package of R software (http://www.r-project.org/) to implement this analysis.

2.3. Determining Subtype-Specific CNAs. To determine subtype-specific CNAs, the segment output file which had arisen from the SBL model and BE procedure was converted into an indicator matrix, where, for each sample, each gene's copy state was represented as −1 = loss, 0 = no change, and 1 = gain. The counts of state for luminal-A subtype and basal-like subtype were compared to identify subtype-specific CNAs. We performed a Chi-square test on the subtype for each gene. Genes with $P < 0.05$ were selected as the subtype-specific genes.

2.4. The Expression Change of Genes Residing in Subtype-Specific CNA Regions. Generally, the copy number alteration at gene promoters typically does not alter the coding sequences of genes but contributes to cancer by influencing gene expression [21]. The genes residing in CNA regions are expected to cause the corresponding expression changes. Therefore, genes amplified or deleted as well as overexpressed or underexpressed in a subtype-specific manner are good candidate genes. To determine whether the mRNA levels of the candidate genes residing in subtype-specific CNA regions are correlated with DNA gain or loss, we observed the expression change of these genes between the two subtypes of samples (luminal-A and basal-like). According to the copy state and the expression change of the genes, we divided them into four gene groups: the luminal-A gain group (the significantly high counts of gain and the significantly overexpressed in the luminal-A sample); the luminal-A loss group (the significantly high counts of loss and the significantly underexpressed in the luminal-A sample); the basal-like gain group (the significantly high counts of gain and the significantly overexpressed in the basal-like sample); the basal-like loss group (the significantly high counts of loss and the significantly underexpressed in the basal-like sample). The further analysis will be performed based on these four groups.

2.5. The Reconstruction of miRNA-mRNA Dysregulated Relationships. Although some previous studies have identified cancer-related gene subnetworks using the Bayesian approach, these methods did not take advantage of the existing biological knowledge, such as miRNA-mRNA target information [9]. Therefore, in this study, for each of the four gene groups (luminal-A gain, luminal-A loss, basal-like gain, and basal-like loss) obtained from above analysis, we used the Bayesian network to identify breast cancer subtype related biomarkers and their regulatory relationships. In the process of learning a Bayesian network with more nodes, if the search space is not restricted, all of the possible networks will be formed with the variables and this is time-consuming. To address this issue, we reconstructed the miRNA-mRNA target dysregulation relationships as the prior biological knowledge to construct the Bayesian network. In the present study, we used a correlation coefficient ratio (CCR) defined by a previous study [15] to reflect the contrast of miRNA and mRNA target combination strength in two different breast cancer subtypes (luminal-A versus basal-like). The empirical distribution of |CCR| was obtained by using the permutation tests and a threshold value was considered as a cut-off value at a significant level ($P < 0.05$) to screen out the significant miRNA-mRNA dysregulated relationships. If $|\mathrm{CCR}_{ij}|$ is significant and $P_{ij\text{-}L} < 0.05$, we defined the relationship between miRNA i and target j is luminal-A trend, whereas the relationship is basal-like trend when $|\mathrm{CCR}_{ij}|$ is significant and $P_{ij\text{-}B} < 0.05$, where $P_{ij\text{-}L}$ and $P_{ij\text{-}B}$ are the P values of Pearson correlation coefficients tests for miRNA i and target j in luminal-A samples and basal-like samples, respectively. The detailed method description is illustrated in our previous study [15].

2.6. Construction of Bayesian Networks and Identification of Network Motifs. For each of the four gene groups (luminal-A gain, luminal-A loss, basal-like gain, and basal-like loss), our goal was to discover the interactions between mRNAs and their regulating miRNAs using their dual expression profiling data under the constraint of the reconstructed breast cancer subtype-trend target information. In order to reduce the search space in the Bayesian network learning process, the reconstructed miRNA-mRNA target information based on CCR was used as the initial structure [22]. For the luminal-A gain and luminal-A loss gene groups, the initial miRNA-mRNA target information was a luminal-A trend. In contrast, the initial miRNA-mRNA target information was a basal-like trend for the basal-like gain and basal-like loss gene groups. A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of variables. A Bayesian network over a set of variables is defined as a network structure S, which is a directed acyclic graph (DAG) over X and a set P of local probability distributions. The Bayesian network S encodes the assertions of conditional independence; that is, each variable x_i is independent of its nondescendants, given its parents in S. In the present study, we allow that the nodes fed into the Bayesian network are genes involved in four gene groups and their regulating miRNAs. The conditional likelihood of the variables given their parents is represented in a Bayesian network by using Gaussian conditional densities. Under the assumption of parameter independence, an initial Bayesian network structure S is learned from the training data. From this initial network, greedy search algorithm with random restarts is performed to get the highest score posterior network to avoid local maxima. Finally, an optimized Bayesian network that maximizes the Bayesian factor is obtained using heuristic search of the network space in a specified domain. The Bayesian network learning process was shown in Figure S1 (see Supplementary Material available online at http://dx.doi.org/10.1155/2014/746970). We used BNArray package of R software (http://www.r-project.org/) to construct Bayesian network. Also, we are particularly interested in the network motifs, which are patterns of subgraphs that recur at frequencies much higher than those found in randomized networks [22]. In the present study, network motifs are topological modules which are frequently occurring subgraphs in our integrated regulatory network.

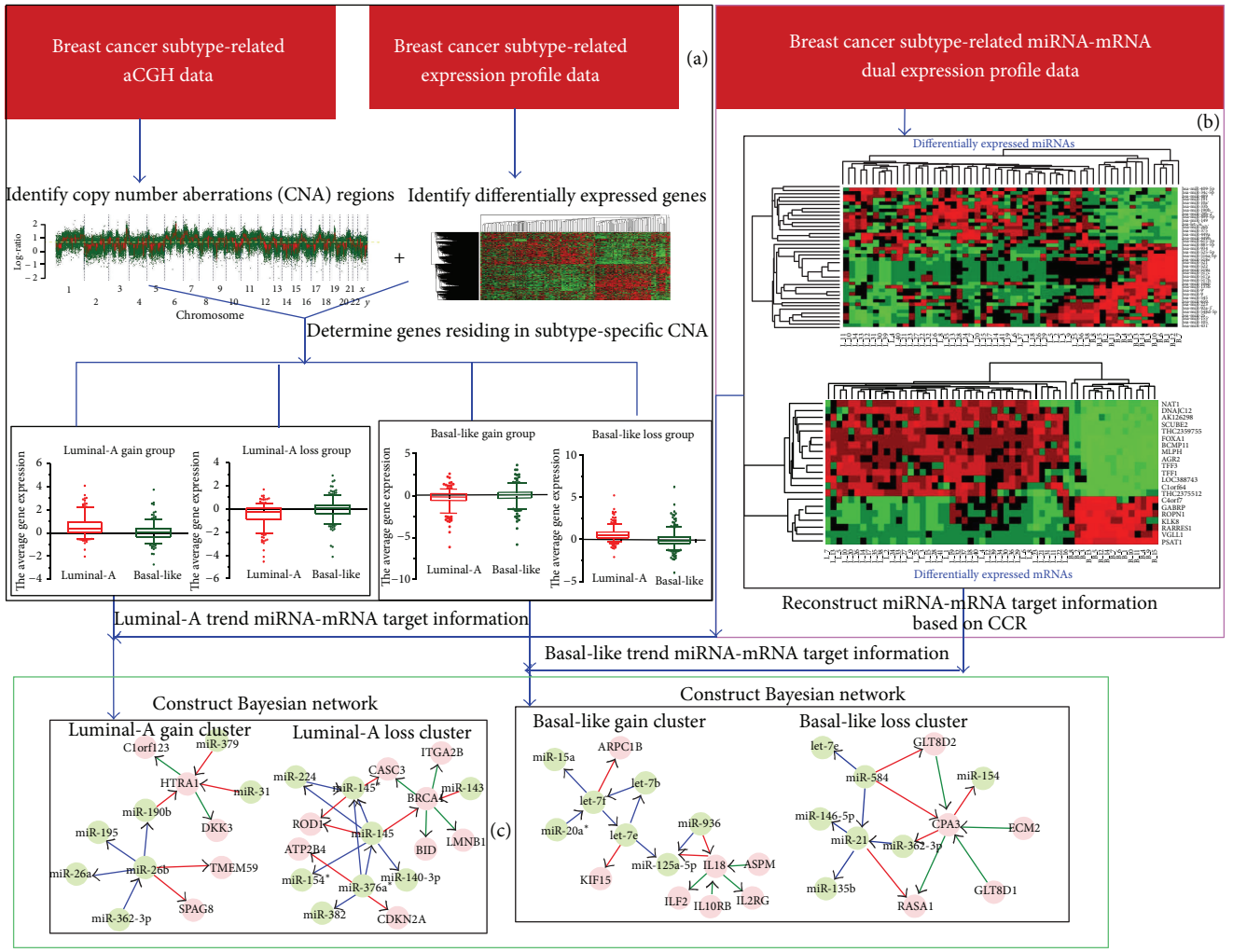


FIGURE 1: The flow chart of our work is as follows. (a) Using breast cancer subtype related aCGH data and breast cancer subtype related expression profile data to determine the genes residing in subtype-specific CNA regions and divide the genes into four gene groups: luminal-A gain group; luminal-A loss group; basal-like gain group; and basal-like loss group. (b) Using breast cancer subtype related miRNA-mRNA dual expression profile data to reconstruct the miRNA-mRNA target information based on the defined correlation coefficient ratio (CCR) values. (c) Construct Bayesian networks and identify network motifs for the four gene groups: for luminal-A gain and luminal-A loss gene groups, the initial target information was a luminal-A trend. In contrast, the initial target information was a basal-like trend for basal-like gain and basal-like loss gene groups.

Each network motif is composed of multiple interaction types which reflect regulatory, signaling, or compensatory pathway mechanisms using the novel network motif finding algorithm-Cyclus3D [23], and the identified network motifs involve at least two biomarkers. Our work flow chart was shown in Figure 1.

2.7. KEGG and CePa Pathway Enrichment Analysis. To further explore each functional gene cluster, we used DAVID (http://david.abcc.ncifcrf.gov/) web tool to perform KEGG pathway enrichment analysis for each of the four gene groups (luminal-A gain, luminal-A loss, basal-like gain, and basal-like loss). A KEGG pathway with a P value of 0.01 adjusted by Benjamini [24] correction was considered to be significant. Considering that the network structure information is necessary for the interpretation of the importance of the pathways, we further used a CePa package [25] to perform the extended enrichment analysis by introducing network centralities as the weight of nodes which have been mapped from differentially expressed genes in pathways. Differing from the traditional overrepresentation analysis methods that find significant pathways without the topological information, the CePa takes into account the pathway structure information so that it can capture new findings that are closely related to the current biological problems [25].

2.8. The Validation of Predicted Regulatory Relationships of BRCA1-miRNAs Pairs by qRT-PCR Detection. To determine whether BRCA1 regulates miRNA processing, we examined the expression levels of primary, precursor, and mature forms of selected miRNAs using quantitative RT-PCR (qRT-PCR) after overexpression or knock-down of BRCA1 in

MCF-7 cells. The major experiment process was described as follows.

(1) Cell Cultures and Transfection. MCF-7 breast tumor cell line was provided by the American Type Culture Collection (ATCC). In our study, MCF-7 were cultured in RPMI 1640 media supplemented with 2 mM L-glutamine (Invitrogen), 20 μg/mL gentamycin (Panpharma), 10% fetal bovine serum (Invitrogen), and 0.04 UI/mL insulin (Novo Nordisk) in a humidified atmosphere at 37°C containing 5% CO_2. MCF-7 cells were transfected with a pCMV6-XL4 vector containing the full-length BRCA1 gene purchased from Origene (Beijing, China) (empty vector as control) or pSilencer siRNA expression vector (Ambion) containing BRCA1 specific shRNA expression cassette (scramble shRNA expression vector as siRNA control). The siRNA target was shown in Table S1. For transfections, cells at 50–60% confluence were incubated with 2 μg of plasmid DNA, using the FuGENE 6 transfection reagent (Roche Molecular Biochemicals, Monza, Italy) according to the manufacturer's instructions. Cells were then selected in G418 (0.4 mg/mL) (Invitrogen Life Technologies, La Jolla, CA, USA). Cell clones that stably expressed G418 and retained growth potential were assayed for BRCA1 expression by qPCR and Western blot assay. qPCR primers for BRCA1 were synthesized as in the following sequence: BRCA1-forward: 5′-TTGTTACAAATCACCCCTCAAGG-3′; BRCA1-reverse: 5′-CCCTGATACTTTTCTGGATGCC-3′. Antibodies for BRCA1 and beta-actin were purchased from Cell Signaling (Danvers, MA, USA).

(2) RNA Extraction, Reverse Transcription, and qRT-PCR Assays. Total RNA was isolated from transfected and control MCF-7 cells with TRIZOL reagent (Invitrogen) according to the manufacturer's protocol. The RNA quality was checked by electrophoresis using a Bioanalyzer 2100 with RNA 6000 Nano LabChip and BioSizing A.02.11 software (Agilent Technologies). For mRNA-, pri-, and pre-miRNAs reverse transcription, 5 μg of total RNA was reverse transcribed in a total volume of 15 μL using the First-Strand DNA Synthesis Kit and performed according to the manufacturer's protocol (Amersham Biosciences). For mature microRNA reverse transcription, 5 μg of total RNA was reverse transcribed in a total volume of 15 μL mix with TaqMan microRNA reverse transcription kit (Applied Biosystems). Reverse transcriptase was thermally inactivated (95°C, 10 min). qRT-PCR assays were performed for determining the expression levels of primary, precursor, and mature miRNAs, as described previously [26, 27]. The qPCR reaction was performed using SYBR Green PCR Master Mix (Applied Biosystems) for pri- and pre-miRNAs. For detection of mature miRNAs, TaqMan MicroRNA assay kit (Applied Biosystems) was used, according to the manufacturer's protocol. The TaqMan reaction was performed with TaqMan Fast Universal PCR Master Mix (Applied Biosystems). Data analysis was performed using the comparative Ct method. Results were normalized to human beta-actin for pri- and pre-miRNAs or human U6 small nuclear RNA (snRNA; hRNU6-1) for mature miRNAs.

(3) In Vivo Monitoring of Pri-miRNA Processing. Plasmid constructs with pri-miRNA at the 3′ untranslated region of firefly luciferase cDNA and BRCA1 or siBRCA1 expression vectors were transfected into MCF-7 cells. Cell extracts were prepared at the 48-hour point after transfection, and the ratio of firefly and *Renilla* luciferase was measured using a Dual-Luciferase Reporter Assay system (Promega). The values were further normalized by using an empty pmirGLO vector and are indicated with standard deviation. The values are presented as the mean ± SD of results of separate experiments and were compared using Student's t-test. Values at $P < 0.05$ were considered to indicate significant differences. All analyses were conducted using JMP IN software, version 5 (SAS).

3. Results

3.1. Subtype-Specific CNAs Were Determined by Assessment of Tumor Genomics DNA Copy Number Changes. By implementing the SBL model and backward elimination (BE) procedure, we obtained the candidate CNA regions. In regard to the luminal-A subtype, the total counts of gains and losses were 3,650 and 3,544, respectively. For the basal-like subtype, the total counts of gains and losses were 4,024 and 3,980, respectively. In particular, the highest counts of CNA regions for a luminal-A subtype sample and a basal-like subtype sample were 663 and 547, respectively (see Figure S2). Moreover, the copy number changes detected high frequency regions (more than 40% across the breast cancer subtype samples) included 8 regions (2 gain regions and 6 loss regions) for the luminal-A subtype and 18 regions (6 gain regions and 12 loss regions) for the basal-like subtype, respectively (see Table 1). Peak incidences were observed in smaller subregions, that is, gains of 10q11-q27 (62.5%) for basal-like subtype and losses of 17q21 (59.60%) for luminal-A subtype. These regions have all been previously shown to be associated with subtype related breast cancers.

In this analysis, the counts of state for subtype were compared to identify the subtype-specific CNAs. To avoid loss, the multiple test corrections were not performed and 4,551 genes with $P < 0.05$ were then gathered for each subtype. To determine whether the mRNA levels of these candidate genes residing in subtype-specific CNA regions were also correlated with DNA gain or loss, we observed the expression change of these genes between the two distinct samples (luminal-A and basal-like). The results indicated that 1,267 genes had the simultaneous subtype-specific CNAs as well as the corresponding expression change. According to the counts state and the expression change of genes, we divided the genes into four gene groups as the method described: luminal-A gain (261 genes), luminal-A loss (297 genes), basal-like gain (298 genes), and basal-like loss (411 genes) (see Figure S3).

3.2. Several Breast Cancer Related Regulatory Network Motifs Were Identified Based on Bayesian Networks and Reconstruction of miRNA-mRNA Dysregulated Relationships. To acquire the prior miRNA-mRNA target information as the initial

Table 1: The detected high frequency CNA regions (more than 40% across samples).

Gains			Losses		
Subtype	Chromosome	Frequency (%)	Subtype	Chromosome	Frequency (%)
Luminal-A	5p15.3-q11.1	46.20	Luminal-A	1p36.3-31.2	55.80
Luminal-A	16p13.3-13.1	61.50	Luminal-A	11p15.5-15.4	42.30
Basal-like	2p25.3-24.2	42.50	Luminal-A	17p13.3-11.2	59.60
Basal-like	6p25.3-21.2	52.50	Luminal-A	18p11.3-q12.1	40.40
Basal-like	9p24.3-22.1	47.50	Luminal-A	19p13.3-13.2	42.30
Basal-like	10p15.3-11.1	62.50	Luminal-A	22q11.1-13.1	50.00
Basal-like	12p13.3-13.1	47.50	Basal-like	1p36.3-35.3	55.00
Basal-like	21q11.1-21.3	60.00	Basal-like	3p26.3-25.2	42.50
			Basal-like	4p16.3-15.3	50.00
			Basal-like	7p22.3-21.3	45.00
			Basal-like	8p23.3-23.1	67.50
			Basal-like	11p15.5-15.2	57.50
			Basal-like	13q11-14.3	57.50
			Basal-like	14q11.2-q13.1	42.50
			Basal-like	16p13-11.2	45.00
			Basal-like	17p13.3-11.2	67.50
			Basal-like	19p13.3-13.1	65.00
			Basal-like	22q11.1-12.1	50.00

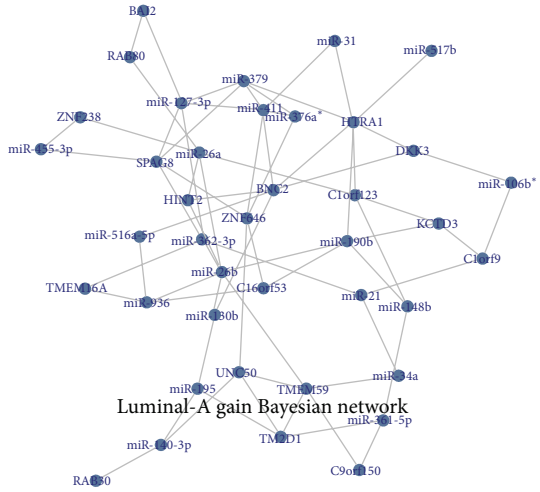
structure of the Bayesian network, we reconstructed the miRNA-mRNA target dysregulated relationships based on the defined CCR. The CCR critical value is of 6.17, and, by using this criterion, 2,659 luminal-A trend and 3,563 basal-like trend miRNA-target dysregulated relationships were identified (see our previous study [15]).

For each of the four gene groups, the nodes fed into the Bayesian network contain the following: (1) luminal-A gain group: 73 nodes involved in miRNA-mRNA luminal-A trend pairs; (2) luminal-A loss group: 81 nodes involved in miRNA-mRNA luminal-A trend pairs; (3) basal-like gain group: 54 nodes involved in miRNA-mRNA basal-like trend pairs; and (4) basal-like loss group: 67 nodes involved in miRNA-mRNA basal-like trend pairs. We utilized their miRNA-mRNA dual expression profiling data to construct the Bayesian networks (see Figures 2(a1)–2(d1) and Table 2). From Table 2, it is evident that the luminal-A loss gene group has a greater betweenness and clustering coefficient when compared to the other three gene groups. That means the genes (miRNAs) involved in the luminal-A loss group are on higher number of shortest paths between partners and tend to interact with each other. In particular, we observed that miR-145 displayed the highest degree in the network. Blenkiron et al. [28] have previously observed higher expression of miR-145 in luminal-A samples, and miR-145 may be one of the miRNAs related to the breast cancer subtype. In addition, the identified let-7e for the basal-like gain group and miR-21 for the basal-like loss group are all approved to be breast cancer subtype related [29, 30]. Furthermore, we used the Cyclus3D algorithm to identify the network motifs that are patterns of subgraphs that recur at frequencies much higher than those found in randomized networks (see Figures 2(a2)–2(d2)). For each of the network motifs, we calculated the Pearson correlations for each of the links. The results showed that all of the links in these network motifs are significant ($P < 0.05$). The average Pearson correlation coefficients were 0.527 for the luminal-A gain group, 0.495 for the luminal-A loss group, 0.658 for the basal-like gain group, and 0.701 for the basal-like loss group, respectively. Although the results generated by our method are compact with only a small number of interactions, some of the identified miRNAs (or mRNAs) by our integrated data analysis are particularly breast cancer related. As an example, in the luminal-A loss related Bayesian network motif, BRCA1 connected multiple significantly underexpressed miRNAs or genes in basal-like samples. It has been reported that breast cancers in BRCA1 mutation carriers frequently have a distinctive basal-like phenotype. A new finding has supported a derivation of the majority of human BRCA1-associated and sporadic basal-like tumors from luminal progenitors rather than from basal stem cells [31]. For another example, in basal-like loss related Bayesian network motif, given the importance of miR-21 in tumorigenesis, Yang et al. found that miR-21 affects the expression of many of its targets through translational inhibition by knocking down the expression of endogenous miR-21 in MCF-7 breast cancer cells [32].

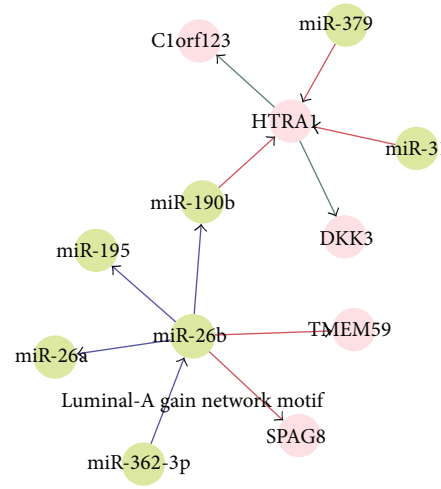
It should be noted that some interactions have been shown to be highly relevant to the subtype of breast cancer, and several of the miRNAs involved in the interactions have been confirmed to be breast cancer subtype related using evidence from previous studies. For example, in the luminal-A gain related Bayesian network motif, we found that HTRA1 was regulated by multiple biomarkers including three miRNAs (miR-379, miR-190b, and miR-31) and two genes (DKK3 and C1orf123), where miR-190b and miR-379 showed

Table 2: The topological properties of Bayesian networks for four gene groups.

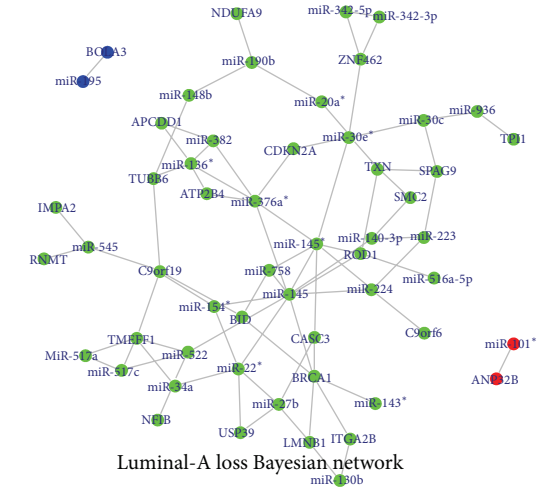
Groups	Avg. degree	Avg. betweenness	Avg. clustering coefficient	Avg. closeness	Avg. number of neighbors	Network density	Network centralization	Characteristic path length	Network diameter	miRNA with highest degree in the network
Luminal-A gain	3.5	38.65	0.073	0.342	3.5	0.090	0.121	2.982	6	miR-26b
Luminal-A loss	2.9	63.24	0.185	0.125	2.9	0.054	0.136	3.724	9	miR-145
Basal-like gain	2.0	48.96	0.080	0.057	2.0	0.037	0.077	4.229	11	let-7e
Basal-like loss	2.4	35.98	0.088	0.039	2.4	0.046	0.092	3.943	11	miR-21



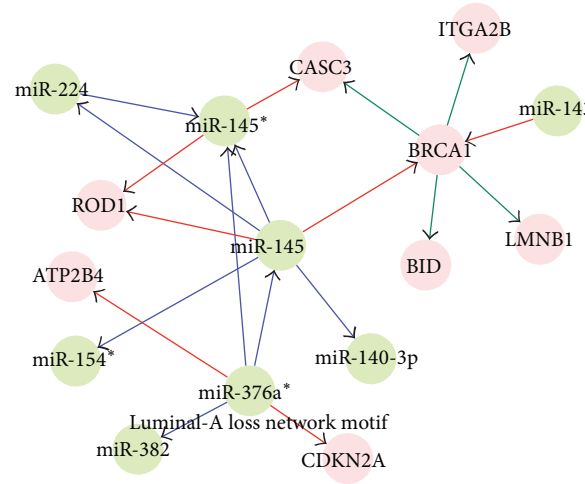
(a1)



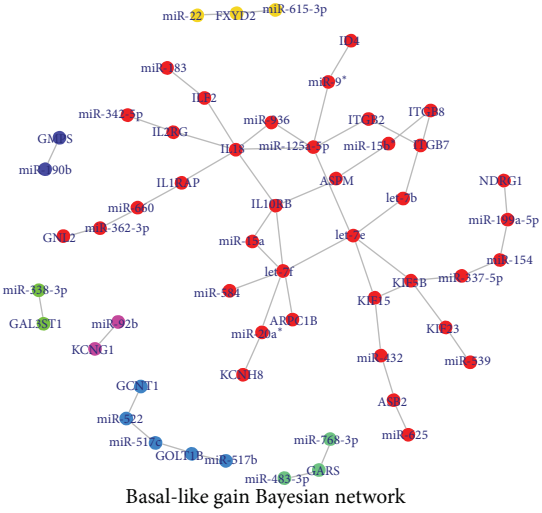
(a2)



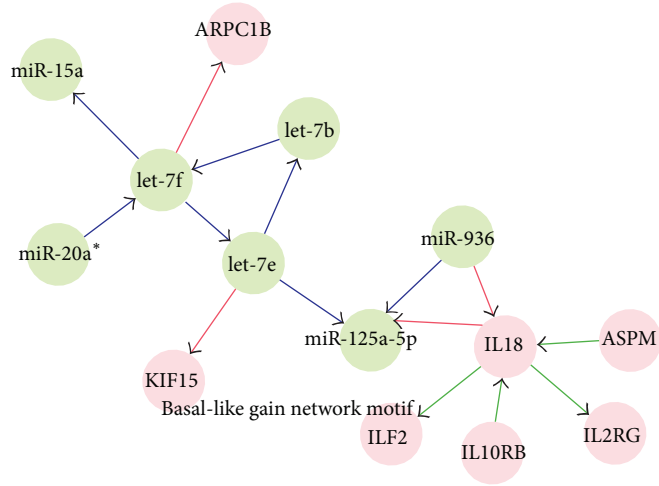
(b1)



(b2)



(c1)



(c2)

Figure 2: Continued.

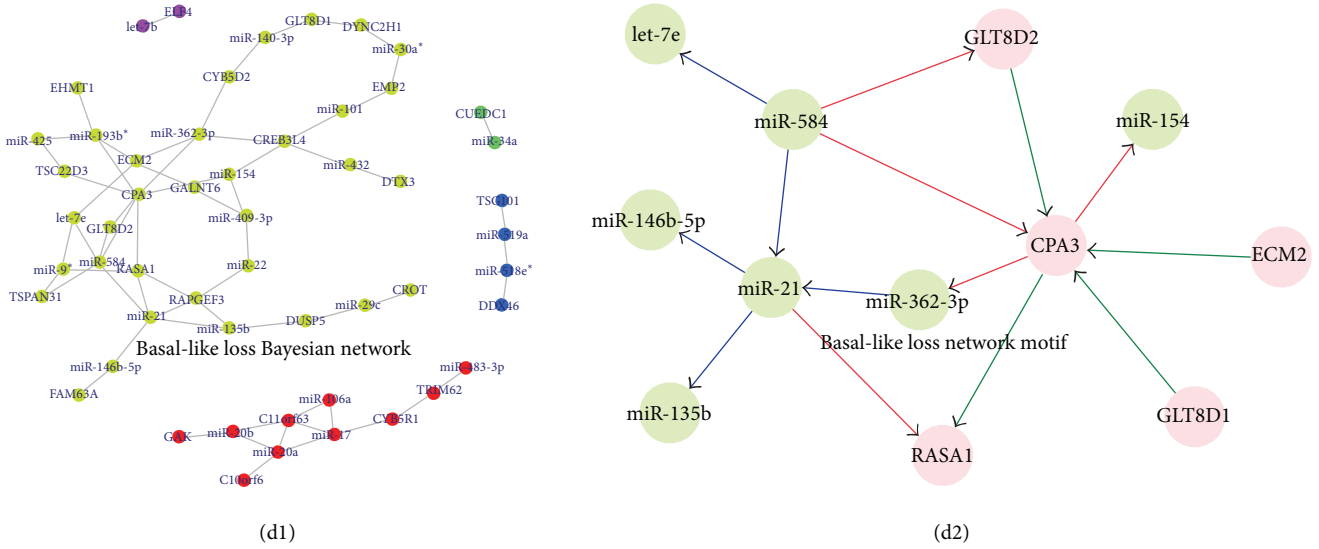


FIGURE 2: Bayesian network and network motifs. (a1) Luminal-A gain Bayesian network; (a2) luminal-A gain network motif; (b1) luminal-A loss Bayesian network; (b2) luminal-A loss network motif; (c1) basal-like gain Bayesian network; (c2) basal-like gain network motif; (d1) basal-like loss Bayesian network; and (d2) basal-like loss network motif. In (a2–d2), each blue line indicates the interaction between two miRNAs; each red line indicates the relationship between miRNA and its target; and each green line indicates the interaction between two genes.

a higher expression in the luminal-A subtype than found in the basal-like subtype, whereas miR-31 presented the contrary situation (see Figure 3(a)). Recent evidences approved that HTRA1 might function as a tumor suppressor by controlling the epithelial-to-mesenchymal transition and might function in chemotherapeutic responsiveness by mediating DNA damage response pathways by characterizing expression in primary breast tissues and the seven human breast epithelial cell lines [33]. In the luminal-A loss related Bayesian network motif, BRCA1 showed the most links with other biomarkers, including two miRNAs (miR-143 and miR-145) and 4 genes (CASC3, ITGA2B, BID, and LMNB1). BRCA1, miR-143, and miR-145 all displayed the higher expression in the luminal-A subtype than in the basal-like subtype (see Figure 3(b)). This suggests that there might be a positive correlation between BRCA1 and miR-143-miR-145. To validate whether BRCA1 regulates these two miRNAs, we examined the expression levels of primary, precursor, and mature forms of miR-143 and miR-145 using quantitative RT-PCR (qRT-PCR) after overexpression or knock-down of a BRCA1 in MCF-7 cells (see Figures 4(a) and 4(b)). Plasmid constructs with pri-miRNA at the 3′ untranslated region of firefly luciferase cDNA (pmirGLO-miR-143, pmirGLO-miR-145) and BRCA1 or siBRCA1 expression vectors were transfected into MCF-7 cells. The primer sequences used for cloning are shown in Table S1. We found that BRCA1 increased precursor and mature miRNAs level of miR-143 and miR-145, though their primary transcripts were found to be with no significant changes (see Figure 4(c)). On the contrary, knock-down of BRCA1 attenuated the expressions of precursor and mature forms of miR-143 and miR-145, whereas their primary transcripts were found to have no significant changes (see Figure 4(d)). Next, to determine whether BRCA1 processes a pri-miRNA substrate, we performed an in vivo pri-miRNA processing monitoring assay of BRCA1 function as previously described [34]. MCF-7 cells were transfected with a luciferase vector construct carrying a segment of pri-miR-143 and pri-miR-145 between the luciferase gene and polyadenylation signal. By using this monitoring system, we observed that BRCA1 overexpression caused a decrease in luciferase activity containing pri-miR-143 and pri-miR-145 sequences (see Figure 4(e)), whereas that activity was increased by knockdown of BRCA1 (see Figure 4(f)). Collectively, these results demonstrate that BRCA1 enhances miR-143 and miR-145 processing of human breast cancer-associated specific miRNAs in vivo.

Besides miR-145, in the basal-like gain related Bayesian network motif, IL18 was found to be regulated by two miRNAs (miR-936 and miR-125a-5p) and four genes (ASPM, ILF2, IL10RB, and IL2RG). We found that two miRNAs displayed the higher expression in the luminal-A subtype than in the basal-like subtype (see Figure 3(c)). Recent studies have shown that the expression of miR-125a-5p is downregulated in human breast cancer and also that a germline mutation in mature miR-125a-5p is closely associated with breast cancer tumorigenesis [35, 36]. In the basal-like loss related Bayesian network motif, CPA3 was regulated by three miRNAs (miR-154, miR-362-3p, and miR-584) and four genes (GLT8D2, GLT8D1, ECM2, and RASA1). Except miR-154 showing the higher expression in the luminal-A subtype than in the basal-like subtype, other two regulating miRNAs (miR-362-3p and miR-584) showed the lower expression in luminal-A subtype samples (see Figure 3(d)). This is supported by the report showing that the expression levels of miR-154 are negatively correlated with Estrogen receptor (ER) positivity in a cohort of early breast cancers [37].

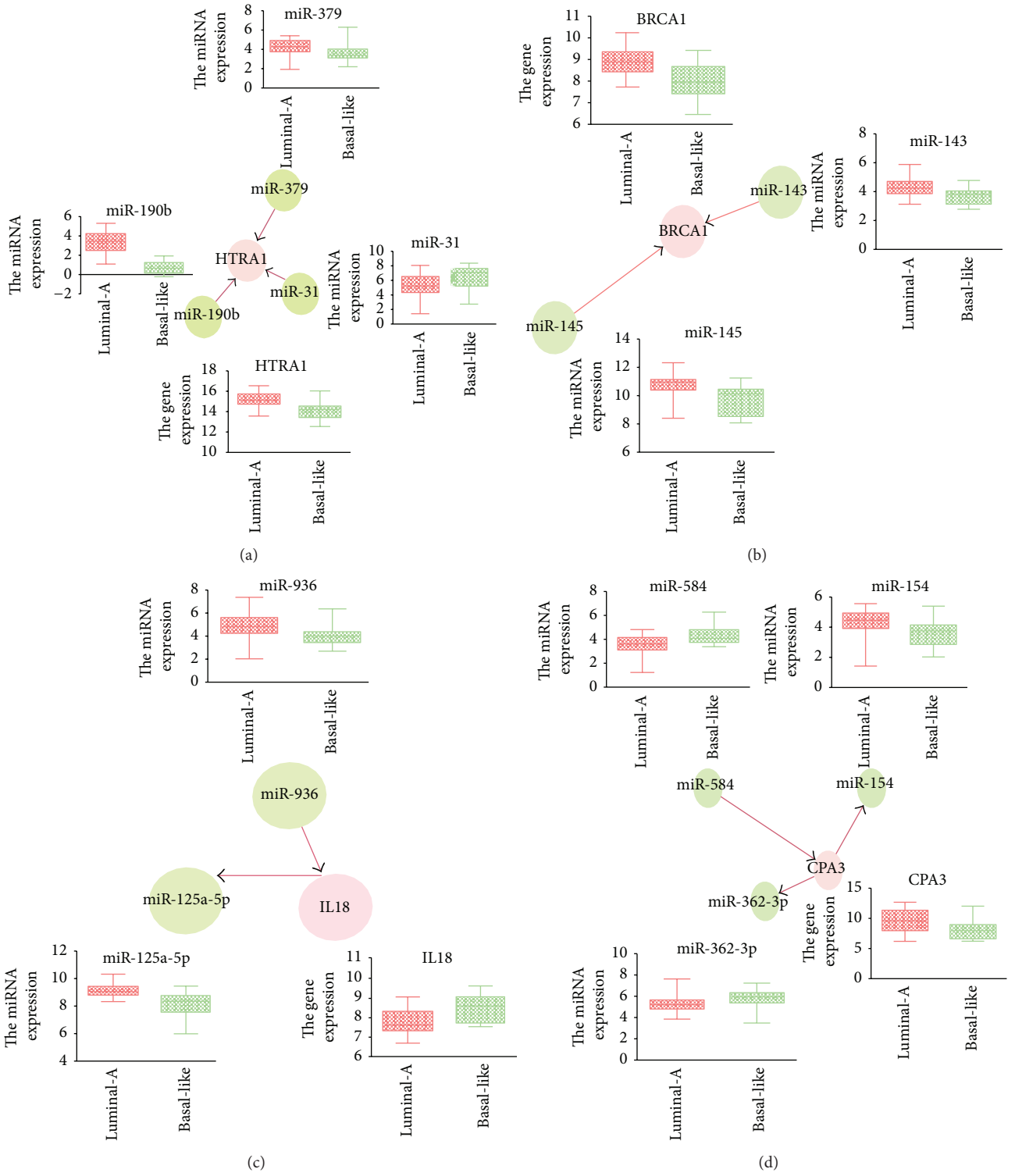


FIGURE 3: Genes targeted by differentially expressed miRNAs in four network motifs. (a) Luminal-A gain network motif; (b) luminal-A loss network motif; (c) basal-like gain network motif; (d) basal-like loss network motif.

3.3. Functional Enrichment Analysis for Four Gene Groups or Network Motifs. For each of the gene groups (luminal-A gain, luminal-A loss, basal-like gain, and basal-like loss), we performed KEGG and CePa enrichment analysis. The results were shown in Table 3. From Table 3, the CePa enrichment analysis results showed that only genes of the luminal-A gain group and basal-like gain group were enriched on the significant pathways. It is interesting here to note that the

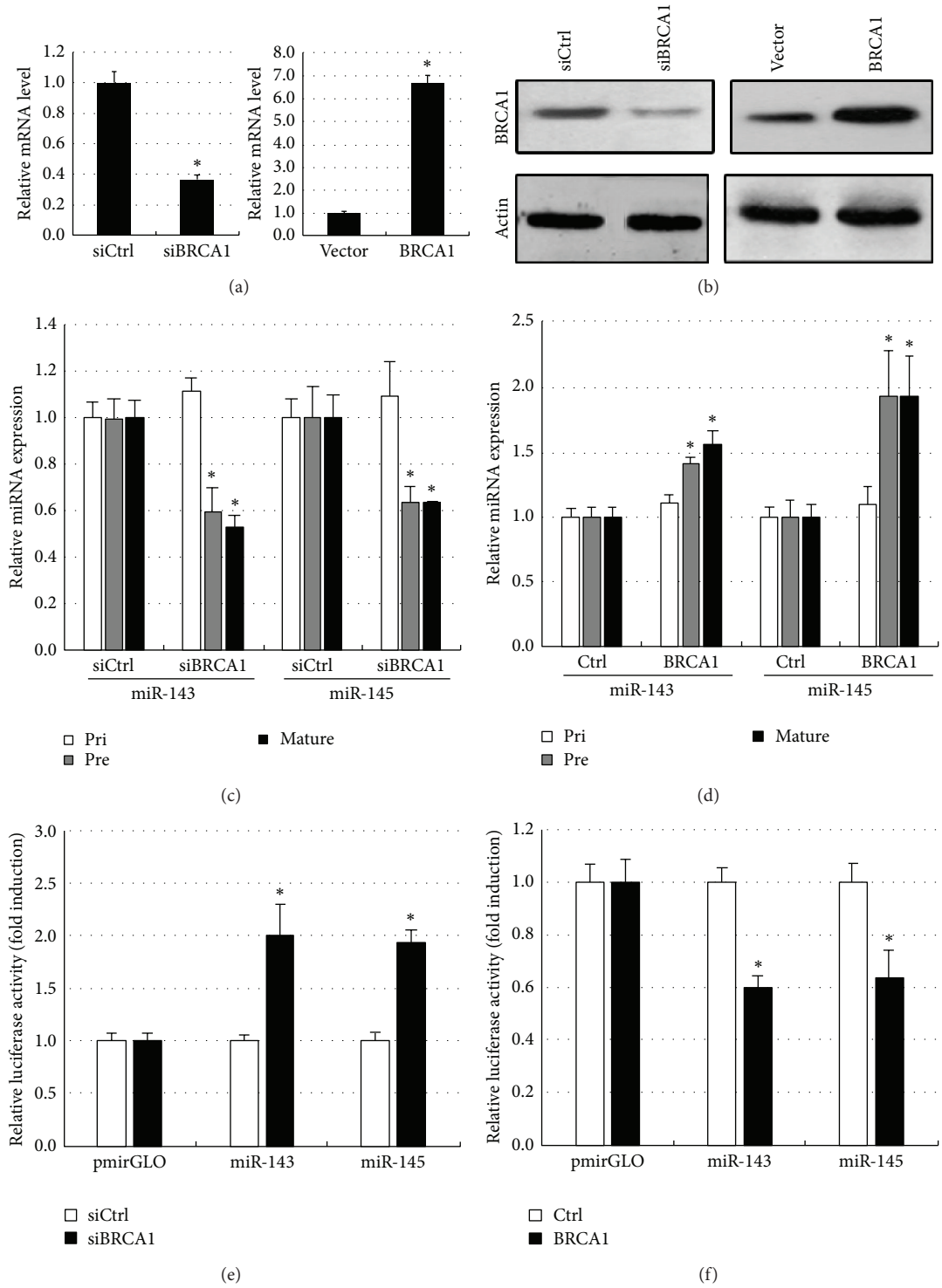


Figure 4: BRCA1 facilitates miR-143 and miR-145 processing. In order to confirm whether BRCA1 facilitates miR-143 and miR-145 processing, BRCA1 overexpression and siBRCA1 expression plasmids were transfected into MCF-7 cells. The empty vector or scramble siRNA expression vector served as controls. (a) BRCA1 mRNA level and (b) protein level were examined by qPCR and Western blot, respectively, in stable transfected cells, normalized by beta-actin ($^*P < 0.05$ as compared with mock control; $n = 3$). Then, the expression levels of the primary (pri), precursor (pre), and mature (mat) forms of the indicated miRNAs were examined in human BRCA1 (c) overexpressed and (d) knock-down MCF-7 cells using qRT-PCR analysis. Pri- and pre-miRNAs were normalized by beta-actin, and mature miRNA was normalized by U6 snRNA ($^*P < 0.05$ as compared with mock control; $n = 3$). Meanwhile, in vivo monitoring assay of pri-miRNA processing in (e) BRCA1 overexpression or (f) knock-down MCF-7 cells carrying miR-143 or miR-145 at the 3′ untranslated region of the luciferase gene. The intensities were normalized by *Renilla* luciferase and are shown as fold induction as compared with an empty pmirGLO vector ($^*P < 0.05$; $n = 3$). Error bars represent standard deviation.

Table 3: The KEGG and CePa pathway enrichment analysis results for the four gene groups.

Gene group	KEGG pathway enrichment analysis	CePa pathway enrichment analysis
Luminal-A gain	hsa05212: pancreatic cancer ($P = 0.0093$)	hsa05212: pancreatic cancer
Luminal-A loss	hsa03440: homologous recombination ($P = 0.0013$)	—
	hsa05200: pathways in cancer ($P = 0.0096$)	—
Basal-like gain	hsa00250: alanine, aspartate, and glutamate metabolism ($P = 0.0052$)	hsa00250: alanine, aspartate, and glutamate metabolism
	hsa04060: cytokine-cytokine receptor interaction ($P = 2.85E-04$)	hsa04060: cytokine-cytokine receptor interaction
	hsa04810: regulation of actin cytoskeleton ($P = 0.0036$)	—
	hsa04630: Jak-STAT signaling pathway ($P = 4.73E-05$)	—
	hsa05130: pathogenic *Escherichia coli* infection ($P = 0.0095$)	—
Basal-like loss	hsa04142: lysosome ($P = 0.0033$)	—
	hsa04512: ECM-receptor interaction ($P = 0.0061$)	—

significantly enriched pathway of the luminal-A gain genes group is pancreatic cancer ($P = 0.0093$). Although the pancreatic cancer pathway appears not to be associated with breast cancer, some reports have determined that mutations in the BRCA2 gene have been implicated in pancreatic cancer susceptibility through studies conducted on high-risk breast and ovarian cancer families. A recent study suggested that BRCA2 mutations could account for 6% of moderate and high-risk pancreatic cancer families [38]. In addition, there were two significantly enriched pathways of the basal-like genes group: alanine, aspartate, and glutamate metabolism pathway ($P = 0.0052$) and cytokine-cytokine receptor interaction pathway ($P = 2.85E-04$). This result indicated that the genes involved in the basal-like group were significantly enriched on the functions related to amino acid metabolism and cytokine-cytokine receptor interaction. In fact, many studies have reported that the genes differentially expressed in breast cancer cells are more inclined to be enriched on cytokine-cytokine receptor interaction pathway [39, 40]. Also, to further explore whether the identified four network motifs are associated with the breast cancer subtype, we used Goeman's global test here to determine their significance. The global test potentially can determine whether the global expression pattern of a group of genes and miRNAs is significantly related to the clinical outcome [41]. The results showed that the identified four network motifs are all strongly associated with the breast cancer subtype ($P = 4.89E-16$ for the luminal-A gain network motif; $P = 1.40E-13$ for the luminal-A loss network motif; $P = 2.14E-14$ for the basal-like gain network motif; and $P = 1.07E-15$ for the basal-like loss network motif; see Figure 5). From Figure 5, we can see that some miRNAs displayed a strong association with a particular breast cancer subtype, such as miR-135b which has been proven to be upregulated in basal-like tumor subtypes [14].

3.4. The Identified Four Network Motifs by Kaplan-Meier (KM) Survival Analysis. To explore whether the global expression pattern of the identified four network motifs or gene groups involved in these network motifs are significantly correlated with survival, we performed a Kaplan-Meier (KM) survival analysis using the mRNA expression levels of the second dataset and the miRNA-mRNA dual expression levels of the third dataset, respectively. In this analysis, each network motif or gene group was classified using K-means clustering ($k = 2$) based on the miRNA or mRNA expression levels into two groups which were defined as either luminal-A trend or basal-like trend according to the proportion of two breast cancer subtype samples. In other words, if the predicted group acquired with K-mean cluster includes a greater number of luminal-A samples than basal-like samples, this group is defined as luminal-A trend and vice versa. We performed the Kaplan-Meier (KM) survival analysis for these network motifs (for the third dataset) or gene groups involved in the network motifs (for the second dataset) and assigned KM P values for each of the sets in order to stratify the patients into survival groups on the basis of the identified sets. For the second dataset in which only mRNA expression values were included, the analysis found that the genes groups extracted from the luminal-A loss network motif and the basal-like gain network motif displayed the significantly different survival rates (log rank $P = 0.0073$ and log rank $P = 0.001$, resp.). Although we cannot see the rest of the two gene groups extracted from the luminal-A gain network motif (log rank $P = 0.0618$) and the basal-like loss network motif (log rank $P = 0.0584$) show the significant survival curve differences, we believe an increased sample size might change these results. Interestingly, for all of the four gene groups, the basal-like trend samples were associated with the worse outcomes and lower survival rates compared to the luminal-A trend samples (see Figure 6). This is supported by the previous study in which the basal-like breast cancer subtype was approved to have a poor prognosis compared with the luminal-A and luminal-B subtypes [42]. However, for the third dataset in which both miRNA and mRNA expression values were included, we did not find any significant association of any of the identified network motifs: log rank $P = 0.6798$ for the luminal-A gain network motif; log rank $P = 0.4074$ for the luminal-A loss network motif; log rank $P = 0.7521$ for the basal-like gain network motif, and log rank $P = 0.5020$ for the basal-like loss network motif, which might be caused by the small sample size of this dataset.

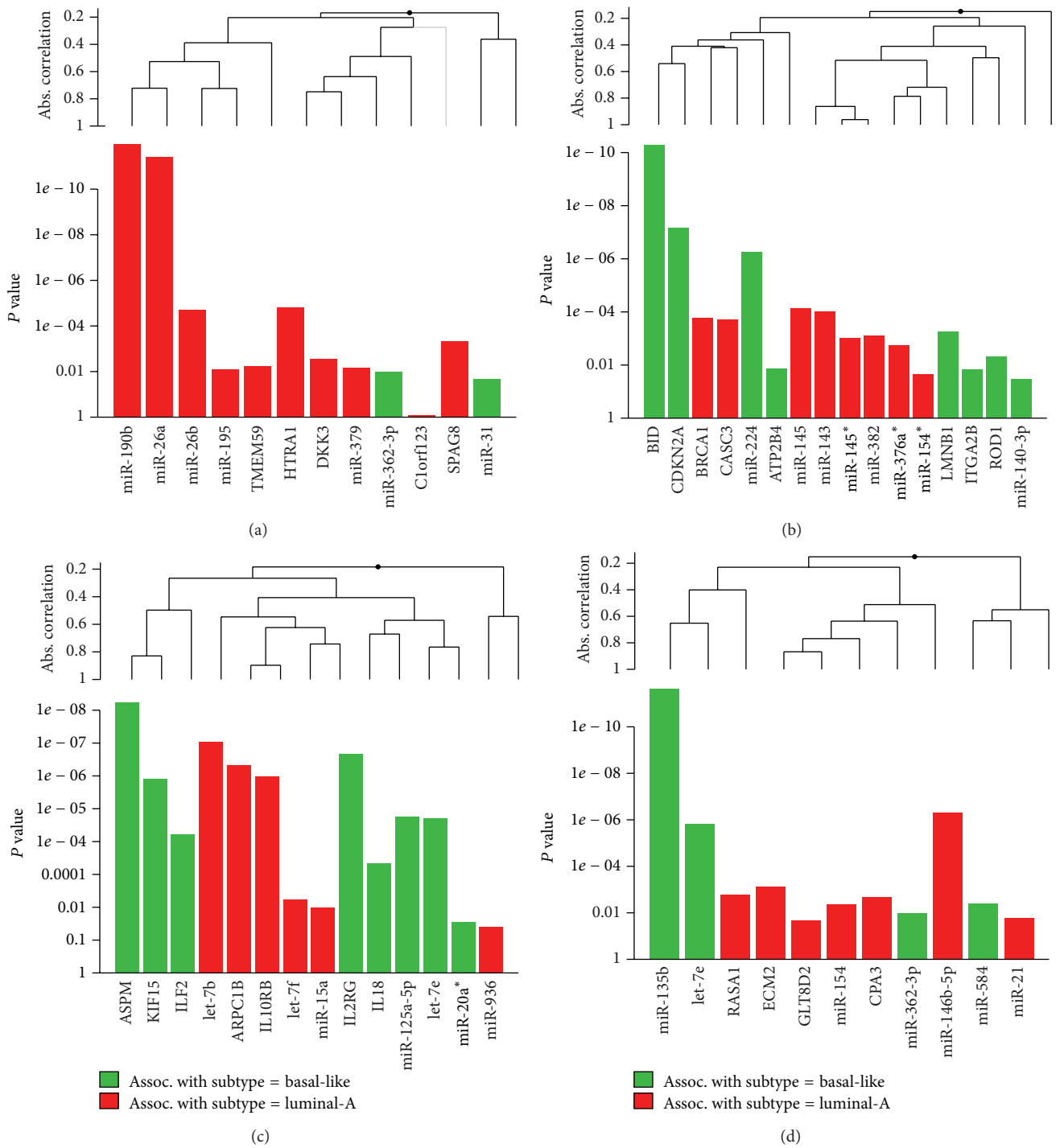


FIGURE 5: Global test for 4 identified network motifs. (a) Luminal-A gain network motif; (b) luminal-A loss network motif; (c) basal-like gain network motif; (d) basal-like loss network motif. This graph is based on the decomposition of the test statistic into the contributions made by each of the genes (or miRNAs) in the alternative hypothesis. The graph illustrated the P values of the tests of individual component genes (or miRNAs) of the alternative. The plotted genes (or miRNAs) are ordered in a hierarchical clustering graph and the clustering method is average linkage.

Moreover, we additionally performed Kaplan-Meier (KM) survival analysis for each of the 25 genes involved in the identified four network motifs. We divided the sample into high expression group when its gene expression value is higher than the median expression. On the contrary, the sample is divided into the low expression group. As a result, 5 genes (SPAG8, KIF15, ILF2, GL8TD1, and ASPM) were found to have a significant log-rank P value in patient stratification (high expression group versus low expression group). We also found that the low average expression of

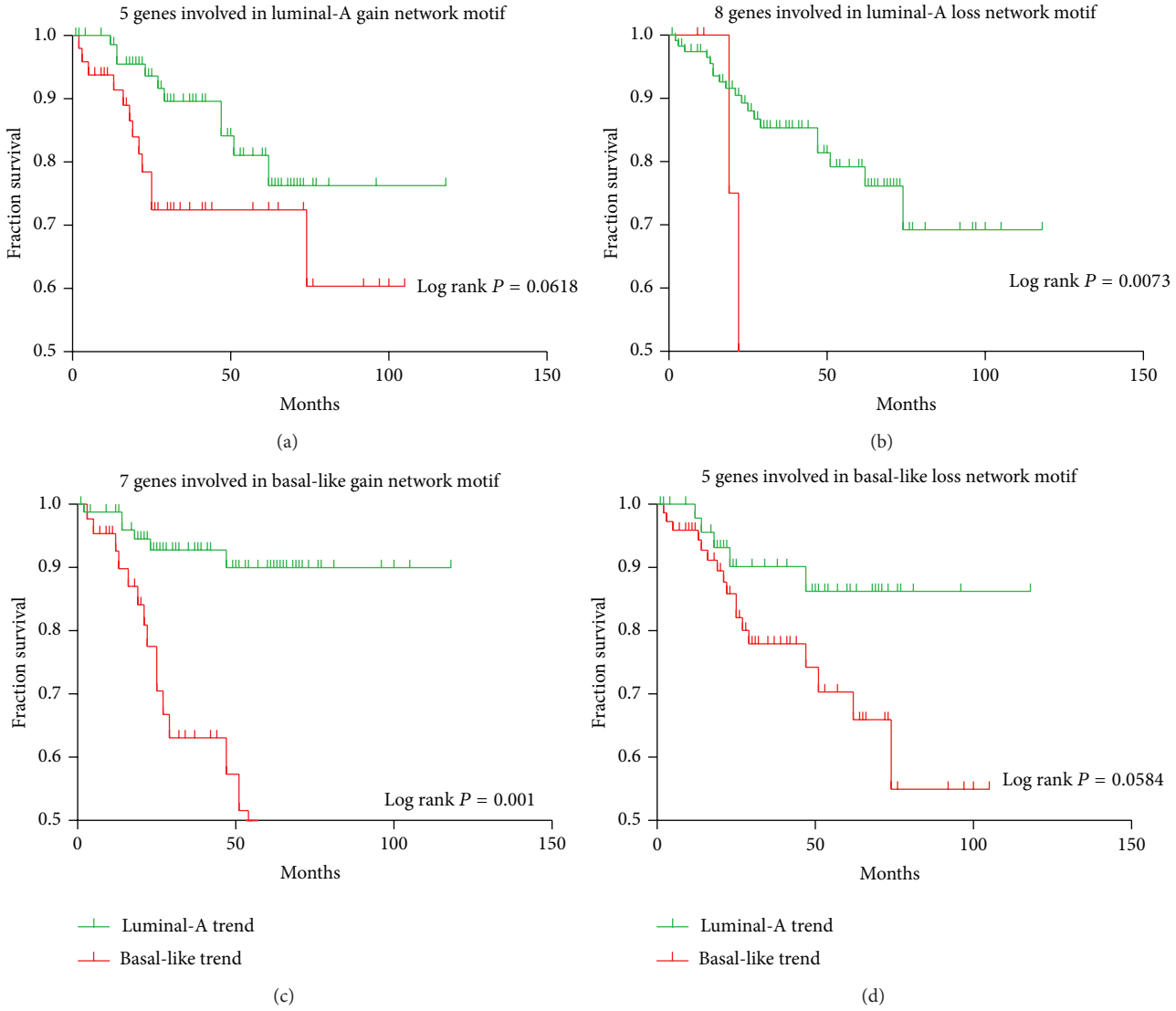


FIGURE 6: Kaplan-Meier (KM) survival analysis for genes involved in the identified network motifs. (a) A group of genes involved in the luminal-A gain network motif; (b) a group of genes involved in the luminal-A loss network motif; (c) a group of genes involved in the basal-like gain network motif; (d) a group of genes involved in the basal-like loss network motif.

SPAG8 and GL8TD1 was associated with the worse outcomes when compared to the high expression (see Figure S4(A) and 4(D)). A contrary trend was observed with KIF15, ILF2, and ASPM (see Figures S4(B), 4(C), and 4(E)). Some of these results can be confirmed by recent studies. For example, a recent study found that ASPM displayed obvious differential expressions in different breast cancer subtypes, and these expression levels were associated with the clinical outcomes [43]. Moreover, our study additionally detected a significant correlation between the expression of the ASPM and CCNB2, and the elevated cytoplasmic CCNB2 protein levels were confirmed to be strongly associated with the short-term disease-specific survival of breast cancer patients [44].

4. Discussions

In the past few years, more and more evidence has shown the relationships between biomarkers and cancers. The significant signatures of mRNA or miRNA expression profiles can be linked to various types of tumors, thereby suggesting that mRNA or miRNA profiling has diagnostic and prognostic potential. As one remarkable feature, increasing number of genomic aberrations has been observed in the progression from normal sample to disease sample. DNA copy number was found to influence gene expression across a wide range of DNA copy number alterations, and it has been reported that at least 12% of all variation in gene expression among breast tumors was directly attributed to variation in gene copy number [8]. The significant CNA of a potential gene should also be reflected in the expression of its mRNA. Previous and recent whole-genome analyses of copy numbers and gene expression have led to the identification of global cellular processes which are underlying malignant transformation and progression of breast cancer subtypes. Some studies have confirmed that the basal-like subtype was the most distinct

in cases with common losses of the regions containing the greatest overall genomic instability. Therefore, application of aCGH allows a direct coupling to the copy number changes with the potential target genes of miRNAs.

In the present study, we provided an integrative data analysis by combining the copy number aberrations and miRNA-mRNA dual expression profiling data to construct regulatory networks from multiple sources of data: copy number data, gene expression profiles of miRNAs and mRNAs, target information based on sequence data, and sample categories. Our study takes into account not only the association between the copy number change and gene expression, but also the association between the expression of miRNAs and their targets. Specifically, we found some breast cancer subtype related genes and miRNAs through the identified network motifs, and these identified biomarkers might be potential targets for the subtype diagnosis and therapy. In the practice, data integration analysis can identify some important biomarkers which cannot be found by many simplistic approaches. Therefore, the clinical subtype-specific driver networks identified through data integration are reproducible and functionally important. Our integrated data analysis can assist in revealing important findings with regard to the underlying molecular mechanisms of breast cancer subtypes. Some discovered interactions and molecular functions have been confirmed by breast cancer documented related study results. In particular, our experimental validation showed the positive correlation for BRCA1-miR-143-miR-145 pairs in breast cancer subtypes. In addition, many of the other discovered biomarkers are of high statistical significance and thus are strong candidates for validation by future experiments.

However, the limitations of our analysis should also be noted. In our analysis, those genes showed copy number abnormalities but did not display the significant over- or underexpression in different subtypes that were excluded. It is important to note that a recent study observed some subtype-specific genes that had no significant CNA, or a relatively poor correlation between CNA and gene expression [45]. Moreover, we only analyzed the predicted direct miRNA-target regulation, due to the computational complexity. Many other relationships such as the protein-protein interaction information (PPI) and the transcription regulation network are not included. On the other hand, the lack of miRNA-mRNA dual expression profiling datasets causes the limitations in the data and these results must be confirmed in the future studies when more miRNA-mRNA dual expression profiling datasets are available. Additionally, it should be noted that, except for the copy number changes and miRNA-mRNA dual expressions, the mutations, protein-protein networks, methylation alterations, and histone modifications can also influence the integrated data analysis results. Therefore more datasets and more biological knowledge are needed to validate the results, and our future study will combine large-scale data from a variety of analyses at the SNP, gene, methylation alteration, and protein levels. This will assist directly towards better understanding of disease pathology. Furthermore, the amplifications and homozygous deletions are relatively small regions, which may be overlooked by CGH techniques. The latest new technique laser microdissection applied for the vast majority of cases will obtain a much higher percentage of cells allowing a more reliable detection of copy number changes, which will be utilized in our future study.

In summary, some of identified biomarkers have been implicated in breast cancer subtypes and might play important roles in breast cancer subtypes since they are considered therapeutic targets. Therefore, the joint analysis of array comparative genomic hybridization (aCGH) copy number data and microarray gene expression data will dissect biological relationships relevant to our understanding of breast cancer subtypes. This may assist in revealing important findings as well as identifying candidate breast cancer subtype related biomarkers by using the constructed biological networks with regard to the underlying molecular mechanisms of breast cancer subtypes.

Conflict of Interests

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

Acknowledgments

This work is supported by Beijing Natural Science Foundation (Grant no. 7132025 and Grant no. 7142015), National Science Foundation of China (Grant no. 31370952), the Science Technology Development Project of Beijing Municipal Commission of Education (SQKM201210025008), and the Excellent Talent Cultivation Project of Beijing (2012D005018000002). This study is also supported by the Open Project Program of Beijing Center of Neural Regeneration and Repair, Beijing Key Laboratory of Fundamental Research on Biomechanics in Clinical Application, China (Grant no. 12345).

References

[1] T. O. Nielsen, J. S. Parker, S. Leung et al., "A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer," *Clinical Cancer Research*, vol. 16, no. 21, pp. 5222–5232, 2010.

[2] N. Uchida, T. Suda, and K. Ishiguro, "Effect of chemotherapy for luminal a breast cancer," *Yonago Acta Medica*, vol. 56, no. 2, pp. 51–56, 2013.

[3] A. Prat, B. Adamo, M. C. U. Cheang, C. K. Anders, L. A. Carey, and C. M. Perou, "Molecular characterization of basal-like and non-basal-like triple-negative breast cancer," *Oncologist*, vol. 18, no. 2, pp. 123–133, 2013.

[4] D. Pinkel and D. G. Albertson, "Array comparative genomic hybridization and its applications in cancer," *Nature Genetics*, vol. 37, no. 6, pp. S11–S17, 2005.

[5] A. Bergamaschi, Y. H. Kim, P. Wang et al., "Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer," *Genes Chromosomes and Cancer*, vol. 45, no. 11, pp. 1033–1040, 2006.

[6] H. G. Russnes, H. K. Vollan, O. C. Lingjaerde et al., "Genomic architecture characterizes tumor progression paths and fate in

breast cancer patients," *Science Translational Medicine*, vol. 2, no. 38, p. 38ra47, 2010.

[7] J. R. Pollack, T. Sørlie, C. M. Perou et al., "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 20, pp. 12963–12968, 2002.

[8] H. K. Solvang, O. C. Lingjærde, A. Frigessi, A.-L. Børresen-Dale, and V. N. Kristensen, "Linear and non-linear dependencies between copy number aberrations and mRNA expression reveal distinct molecular pathways in breast cancer," *BMC Bioinformatics*, vol. 12, article 197, 2011.

[9] U. D. Akavia, O. Litvin, J. Kim et al., "An integrated approach to uncover drivers of cancer," *Cell*, vol. 143, no. 6, pp. 1005–1017, 2010.

[10] Z. Liu, X.-S. Zhang, and S. Zhang, "Breast tumor subgroups reveal diverse clinical prognostic power," *Scientific Reports*, vol. 4, Article ID 04002, 2014.

[11] E. Román-Pérez, P. Casbas-Hernández, J. R. Pirone et al., "Gene expression in extratumoral microenvironment predicts clinical outcome in breast cancer patients," *Breast Cancer Research*, vol. 14, no. 2, article R51, 2012.

[12] S. Zhang, Q. Li, J. Liu, and X. J. Zhou, "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules," *Bioinformatics*, vol. 27, no. 13, pp. i401–i409, 2011.

[13] S. Zhang, C. C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou, "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data," *Nucleic Acids Research*, vol. 40, no. 19, pp. 9379–9391, 2012.

[14] E. Enerly, I. Steinfeld, K. Kleivi et al., "miRNA-mRNA integrated analysis reveals roles for mirnas in primary breast tumors," *PLoS ONE*, vol. 6, no. 2, Article ID e16915, 2011.

[15] L. Hua, P. Zhou, L. Li, H. Liu, and Z. Yang, "Prioritizing breast cancer subtype related miRNAs using miRNA-mRNA dysregulated relationships extracted from their dual expression profiling," *Journal of Theoretical Biology*, vol. 331, pp. 1–11, 2013.

[16] L. Hua, L. Li, and P. Zhou, "Identifying breast cancer subtype related mirnas from two constructed mirnas interaction networks in silico method," *BioMed Research International*, vol. 2013, Article ID 798912, 13 pages, 2013.

[17] V. J. Weigman, H.-H. Chao, A. A. Shabalin et al., "Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival," *Breast Cancer Research and Treatment*, vol. 133, no. 3, pp. 865–880, 2012.

[18] 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.

[19] V. Chandra, R. Girijadevi, A. S. Nair, S. S. Pillai, and R. M. Pillai, "MTar: a computational microRNA target prediction architecture for human transcriptome," *BMC Bioinformatics*, vol. 11, no. 1, article S2, 2010.

[20] R. Pique-Regi, J. Monso-Varona, A. Ortega, R. C. Seeger, T. J. Triche, and S. Asgharzadeh, "Sparse representation and Bayesian detection of genome copy number alterations from microarray data," *Bioinformatics*, vol. 24, no. 3, pp. 309–318, 2008.

[21] X. Shen, S. Li, L. Zhang et al., "An integrated approach to uncover driver genes in breast cancer methylation genomes," *PLoS ONE*, vol. 8, no. 4, Article ID e61214, 2013.

[22] T. D. Le, L. Liu, B. Liu et al., "Inferring microRNA and transcription factor regulatory networks in heterogeneous data," *BMC Bioinformatics*, vol. 14, article 92, 2013.

[23] P. Audenaert, T. V. Parys, F. Brondel et al., "CyClus3D: a cytoscape plugin for clustering network motifs in integrated networks," *Bioinformatics*, vol. 27, no. 11, article 1587, 2011.

[24] Y. Benjamini, "Discovering the false discovery rate," *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 72, no. 4, pp. 405–416, 2010.

[25] Z. Gu and J. Wang, "CePa: an R package for finding significant pathways weighted by multiple network centralities," *Bioinformatics*, vol. 29, no. 5, pp. 658–660, 2013.

[26] B. N. Davis, A. C. Hilyard, G. Lagna, and A. Hata, "SMAD proteins control DROSHA-mediated microRNA maturation," *Nature*, vol. 454, no. 7200, pp. 56–61, 2008.

[27] H. I. Suzuki, K. Yamagata, K. Sugimoto, T. Iwamoto, S. Kato, and K. Miyazono, "Modulation of microRNA processing by p53," *Nature*, vol. 460, no. 7254, pp. 529–533, 2009.

[28] C. Blenkiron, L. D. Goldstein, N. P. Thorne et al., "MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype," *Genome Biology*, vol. 8, no. 10, article R214, 2007.

[29] C. D. May, N. Sphyris, K. W. Evans, S. J. Werden, W. Guo, and S. A. Mani, "Epithelial-mesenchymal transition and cancer stem cells: a dangerously dynamic duo in breast cancer progression," *Breast Cancer Research*, vol. 13, no. 1, article 202, 2011.

[30] V. J. Findlay, "Micrornas and breast cancer," *The Open Cancer Journal*, vol. 3, no. 1, pp. 55–61, 2010.

[31] G. Molyneux, F. C. Geyer, F.-A. Magnay et al., "*BRCA1* basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells," *Cell Stem Cell*, vol. 7, no. 3, pp. 403–417, 2010.

[32] Y. Yang, R. Chaerkady, M. A. Beer, J. T. Mendell, and A. Pandey, "Identification of miR-21 targets in breast cancer cells using a quantitative proteomic approach," *Proteomics*, vol. 9, no. 5, pp. 1374–1384, 2009.

[33] N. Wang, K. A. Eckert, A. R. Zomorrodi et al., "Down-regulation of HtrA1 activates the Epithelial-Mesenchymal transition and ATM DNA damage response pathways," *PLoS ONE*, vol. 7, no. 6, Article ID e39446, 2012.

[34] S. Kawai and A. Amano, "BRCA1 regulates microRNA biogenesis via the DROSHA microprocessor complex," *The Journal of Cell Biology*, vol. 197, no. 2, pp. 201–208, 2012.

[35] N. Nishida, K. Mimori, M. Fabbri et al., "MicroRNA-125a-5p is an independent prognostic factor in gastric cancer and inhibits the proliferation of human gastric cancer cells in combination with trastuzumab," *Clinical Cancer Research*, vol. 17, no. 9, pp. 2725–2733, 2011.

[36] W. Li, R. Duan, F. Kooy, S. L. Sherman, W. Zhou, and P. Jin, "Germline mutation of microRNA-125a is associated with breast cancer," *Journal of Medical Genetics*, vol. 46, no. 5, pp. 358–360, 2009.

[37] H. M. Heneghan, N. Miller, A. J. Lowery, K. J. Sweeney, and M. J. Kerin, "MicroRNAs as novel biomarkers for breast cancer," *Journal of Oncology*, vol. 2010, Article ID 950201, 7 pages, 2010.

[38] F. J. Couch, M. R. Johnson, K. G. Rabe et al., "The prevalence of *BRCA2* mutations in familial pancreatic cancer," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 16, no. 2, pp. 342–346, 2007.

[39] S. Wang, Y.-Q. Xiao, Z.-Q. Liu et al., "Network-guided genetic screening for metastasis-related microRNA-200c in breast cancer," *Tumor*, vol. 33, no. 2, pp. 111–118, 2013.

[40] S. Kim, M. Kon, and C. DeLisi, "Pathway-based classification of cancer subtypes," *Biology Direct*, vol. 7, article 21, 2012.

[41] J. J. Goeman, S. van de Geer, F. de Kort, and H. C. van Houwellingen, "A global test for groups fo genes: testing association with a clinical outcome," *Bioinformatics*, vol. 20, no. 1, pp. 93–99, 2004.

[42] X. R. Yang, M. E. Sherman, D. L. Rimm et al., "Differences in risk factors for breast cancer molecular subtypes in a population-based study," *Cancer Epidemiology Biomarkers and Prevention*, vol. 16, no. 3, pp. 439–443, 2007.

[43] M. Zvelebil, E. Oliemuller, Q. Gao et al., "Embryonic mammary signature subsets are activated in *Brca1*$^{-/-}$ and basal-like breast cancers," *Breast Cancer Research*, vol. 15, no. 2, article R25, 2013.

[44] E. Shubbar, A. Kovács, S. Hajizadeh et al., "Elevated cyclin B2 expression in invasive breast carcinoma is associated with unfavorable clinical outcome," *BMC Cancer*, vol. 13, article 1, 2013.

[45] B. Dutta, L. Pusztai, Y. Qi et al., "A network-based, integrative study to identify core biological pathways that drive breast cancer clinical subtypes," *British Journal of Cancer*, vol. 106, no. 6, pp. 1107–1116, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 301635, 9 pages
http://dx.doi.org/10.1155/2015/301635

Research Article

A Network Flow Approach to Predict Protein Targets and Flavonoid Backbones to Treat Respiratory Syncytial Virus Infection

José Eduardo Vargas,[1] Renato Puga,[2] Joice de Faria Poloni,[3] Luis Fernando Saraiva Macedo Timmers,[4] Barbara Nery Porto,[1] Osmar Norberto de Souza,[4] Diego Bonatto,[3] Paulo Márcio Condessa Pitrez,[1] and Renato Tetelbom Stein[1]

[1] *Centro Infant, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Avenue Ipiranga 6681, 90619-900 Porto Alegre, RS, Brazil*
[2] *Clinical Research Center, Hospital Israelita Albert Einstein (HIAE), São Paulo, Brazil*
[3] *Department of Molecular Biology and Biotechnology, Federal University of Rio Grande do Sul (UFRGS), 90619-900 Porto Alegre, RS, Brazil*
[4] *Faculty of Informatics, Laboratory for Bioinformatics, Modelling & Simulation of Biosystems, Pontifical Catholic University of Rio Grande do Sul (PUCRS), 90619-900 Porto Alegre, RS, Brazil*

Correspondence should be addressed to José Eduardo Vargas; josevargas123@gmail.com

Received 26 July 2014; Accepted 11 September 2014

Academic Editor: Jiangning Song

Background. Respiratory syncytial virus (RSV) infection is the major cause of respiratory disease in lower respiratory tract in infants and young children. Attempts to develop effective vaccines or pharmacological treatments to inhibit RSV infection without undesired effects on human health have been unsuccessful. However, RSV infection has been reported to be affected by flavonoids. The mechanisms underlying viral inhibition induced by these compounds are largely unknown, making the development of new drugs difficult. *Methods.* To understand the mechanisms induced by flavonoids to inhibit RSV infection, a systems pharmacology-based study was performed using microarray data from primary culture of human bronchial cells infected by RSV, together with compound-proteomic interaction data available for *Homo sapiens*. *Results.* After an initial evaluation of 26 flavonoids, 5 compounds (resveratrol, quercetin, myricetin, apigenin, and tricetin) were identified through topological analysis of a major chemical-protein (CP) and protein-protein interacting (PPI) network. In a nonclustered form, these flavonoids regulate directly the activity of two protein bottlenecks involved in inflammation and apoptosis. *Conclusions.* Our findings may potentially help uncovering mechanisms of action of early RSV infection and provide chemical backbones and their protein targets in the difficult quest to develop new effective drugs.

1. Introduction

Respiratory syncytial virus (RSV) is a major cause of lower respiratory tract infection with high level of mortality in children around the world [1–3]. It is estimated that all children by two years of age have been infected by RSV and more than half of them are reinfected [4]. Moreover, RSV pathogenesis is notably associated with an increased airway resistance characterized as wheezing, diagnosed as bronchiolitis [2].

In the 1960 decade, a vaccine trial was performed with unexpected and tragic results [5]. Hence, effective preventive treatment to RSV infection is unavailable, since there is no vaccine against the virus. However, several prototypes are under study [6–9]. The prophylactic therapy with palivizumab, a humanized monoclonal antibody, has been

shown to reduce the number of RSV hospitalizations in preterm infants [10], but the treatment has a very high cost, and it is administered only to children with risk factors for RSV bronchiolitis [11]. Another optional treatment against RVS infection is ribavirin. It is a nucleoside analog that introduces mutations into the RNA viral genome during replication and was previously used routinely for infants hospitalized with RSV. However, it has been associated with undesired side-effects and was not considered an effective treatment [12, 13].

The absence of a vaccine for RSV-induced bronchiolitis and the existence of few antiviral agents against RSV constitute very important problems in pediatric medicine. Thus, the development of novel anti-RSV drugs that can be administered orally or parenteral to children is extremely necessary.

A great variety of viruses have been reported to be inhibited by natural compounds, such as flavonoids [14–16]; however, the molecular mechanisms underlying such effects are largely unclear. In this sense, it is difficult to develop new drugs.

In a search to provide new insights for RSV treatments and to understand the multiples signaling pathways affected by RSV infection, an integrative model based on systems pharmacology predictions has been used. Moreover, this methodology will allow understanding the effect of flavonoid (FLA) compounds against RSV infection, integrating chemical-protein (CP) and protein-protein interaction (PPI) networks.

2. Materials and Methods

2.1. Gene Expression Data from Primary Human Bronchial Epithelial (PHBE) Cells Infected by RSV. The microarray data GSE12144 were downloaded from the Gene Expression Omnibus (GEO) database [http://www.ncbi.nlm.nih.gov/geo/]. Subsequently, a linear model was applied to normalize this data, using Limma package from R/Bioconductor to guarantee maximal statistical stringency [17]. Additionally, a contrast analysis was applied and differentially expressed genes (PHBE mock versus PHBE RSV 24h) were identified by Rank Product with a cutoff P value of ≤ 0.05 [18].

2.2. Selection of Flavonoids. To select flavonoids with potential antiviral effect against pathogenic respiratory agents, a literature mining was performed. Two flavonoids commonly described against respiratory viral infections were selected: quercetin [19–21] and resveratrol [22–24]. Quercetin is found in abundance in onions, apples, broccoli, and berries [25], whereas resveratrol is present in grapes, berries, and peanuts [25].

In order to obtain drug-like compounds, a database-dependent model was applied to calculate the drug-likeness of all compounds similar to resveratrol or quercetin through Tanimoto coefficient (Tc) [37]:

$$\mathrm{Tc} = \frac{\sum_{j=1}^{k} a_j \times b_j}{\left(\sum_{j=1}^{k} a_j^2 + \sum_{j=1}^{k} b_j^2 - \sum_{j=1}^{k} a_j \times b_j\right)}, \quad (1)$$

where "a" is the molecular property of each compound and "b" represents the average molecular properties of the whole compounds in the Drugbank database [http://www.drugbank.ca/]. The Drugbank database is a unique bioinformatic resource that contains 6825 compound data. These chemical compounds are FDA approved drugs or are being evaluated in clinical trials. In our work a criterion of Tc values ≥ 0.611 was used according to suggestion by Drugbank site (data shown in the Table 1).

2.3. Design of CP-PPI Networks. To obtain CP-PPI networks, the metasearch engine STITCH 3.1 [http://stitch.embl.de/] was applied. STITCH software allows visualization of the connections (edge) among different proteins, chemical compounds, and compounds-proteins, where each edge shows a degree of confidence between 0 (lowest confidence) and 1.0 (highest confidence). To this present work, the parameters used were as follows: all prediction methods were enabled, excluding text mining; maximal of 10 interactions by node; degree of confidence, medium (0.400); and a network depth equal to 1. In addition, GeneCard [http://www.genecards.org/] and Pubchem [https://pubchem.ncbi.nlm.nih.gov/] databases were used to search synonymous names of genes and compounds recognizable by STITCH. In sequence, the outcomes obtained through these search engines were analyzed with Cytoscape 2.8.2 [38]. Nonconnected nodes were excluded from the networks.

2.4. Modular Analysis of CPI-PPI Network. ClusterONE was the tool used to discover densely connected and possibly overlapping regions within the Cytoscape network [39]. Dense regions corresponded to protein or compound-protein complexes or parts of them.

ClusterONE identifies subnetworks by the identification of "growing" dense regions out of small seeds guided by a quality function. The quality of a group was evaluated by the number of internal edges divided by the number of edges involving nodes of the group.

2.5. Gene Ontology Analysis. Gene ontology (GO) analysis was determined by biological network gene ontology (BiNGO) software 2.44 [http://chianti.ucsd.edu/cyto_web/plugins/index.php] [40]. The degree of functional enrichment for a given category was assessed (P value ≤ 0.05) by hypergeometric distribution [41] and multiple test correction was applied using the false discovery rate (FDR) algorithm [42], from BiNGO software. Overrepresented biological process categories were obtained after FDR correction, with a significance level of 0.05.

2.6. Centralities Parameters and Topological Analysis. Major network centralities (closeness, betweenness, and node degree) were analyzed with the CP-PPI networks using the Cytoscape plugin CentiScape 2.8.2 [43].

Table 1: List of flavonoid compounds considered to chemical protein-protein network design. Chemical identification (Pubchem), Tanimoto similarity scores, and the antiviral activity of each compound (manually curated from literature).

Compound ID	Pubchem CID	Tanimoto similarity (score)	Antiviral RSV references
A*			
Resveratrol	445154	1	[22, 24, 26–29]
Piceatannol	667639	0.966	UD***
AC1O4D7M	6365297	0.719	UD
Caffeic acid	689043	0.689	UD
Phenol	996	0.687	[30, 31]
HLF	5288545	0.684	UD
Sinapinate	637775	0.635	UD
Ferulic acid	445858	0.622	[32]
Isoferulic acid	736186	0.614	[32]
2MP	7249	0.621	UD
P-coumaric acid	637542	0.611	UD
B**			
Quercetin	5280343	1	[19, 25]
Myricetin	5281672	1	UD
ST059620	5281614	0.959	UD
Kaempferol	5280863	0.946	[25]
Tricetin	5281701	0.884	UD
Apigenin	5280443	0.823	[33]
Oroxylin A	5320315	0.791	[34]
Wogonin	5281703	0.765	[34]
Flavone	10680	0.714	[35]
EMD 21388	128600	0.636	UD
α-Naphthoflavone	11790	0.711	[34]
β-Naphthoflavone	2361	0.711	[35]
Rutin	5280805	0.631	UD
Genistein	5280961	0.618	[36]
DB07032	656936	0.612	UD

A* Group with high similarity to resveratrol.
B** Group with high similarity to quercetin.
UD*** Undescribed in the literature.

Closeness centrality was used to evaluate the shortest path among a random node (protein or chemical compound) and all other nodes [43]:

$$\mathrm{Clo}\,(v) = \frac{1}{\sum w \in v^{\mathrm{dist}(v,w)}}, \tag{2}$$

where the closeness value (Clo(v)) was calculated by computing the shortest path between the node v and all other nodes w found within a network.

The average closeness (Clo) score was calculated by the sum of different closeness scores (Clo_i) divided by the total number of nodes analyzed ($N_{(v)}$):

$$\langle \mathrm{Clo} \rangle = \frac{\sum_i \mathrm{Clo}_i}{N_{(v)}}. \tag{3}$$

Also, the betweenness parameter was taken into account in the analysis. This parameter is a measure equal to the number of shortest paths from a couple of nodes that pass through a different node [43, 44]:

$$\mathrm{Bet}\,(v) = \sum_{s \neq v \neq w \in V} \frac{\sigma_{sw}\,(v)}{\sigma_{sw}}, \tag{4}$$

where σ_{sw} is the total number of the shortest paths from node s to node w and $\sigma_{sw}(v)$ is the number of those paths that pass through the node v.

The average betweenness score (Bet) of the network was calculated using (5), where the sum of different betweenness scores (Bet_i) is divided by the total number of nodes $N_{(v)}$ analyzed:

$$\langle \mathrm{Bet} \rangle = \frac{\sum_i \mathrm{Bet}_i}{N_{(v)}}. \tag{5}$$

The average betweenness score of CP-PPI network was used to obtain responsible nodes of the control of the flow of information in the network. These nodes are called bottlenecks (B) and show higher probability of connections of different modules or biological processes.

Finally, parameter degree was calculated. This parameter is a measure that indicates the number of adjacent nodes (E_i) that are connected to a specific node (v), according to

$$\mathrm{Deg}\,(v) = \sum E_i. \tag{6}$$

The average node degree of a network (Deg) is given by (7), where the sum of different node degree scores (Deg_i) is divided by the total number of nodes $N_{(v)}$ present in the network:

$$\langle \mathrm{Deg} \rangle = \frac{\sum_i \mathrm{Deg}_i}{N_{(v)}}. \tag{7}$$

Nodes with a high node degree score compared to the average are called hubs (H) and are responsible for a central regulatory role in the cell.

In this work, H-B (hub-bottleneck) may correspond to central proteins or FLA compounds that are highly connected to several complexes, while nodes that belong to the NH (non-hub-B) group correspond to proteins or FLA compounds that are important. In order to obtain H-B and NH-nodes, mathematical means (threshold) generated for betweenness and degree parameters were considered.

2.7. Molecular Parameters for the Development of a Potential Drug. All compounds, which were chemically verified by Zinc database [45, 46] were analyzed taking into account the Lipinsky's rule of five (xLogP, molecular weight, number of hydrogen bond acceptors, and donors). Toxicity risks (mutagenic, tumorigenic, irritant, and reproductive effect) were also examined by the Osiris Property Explorer [http://www.organic-chemistry.org/prog/peo].

A diagram of methodological steps used in this work is showed in Figure 1.

3. Results and Discussion

Studies of the FLA effects on viruses only have been performed *in vitro* and *in vivo* but not *in silico* using high-throughput (*omic*) *approaches* and network analysis based on interactome data. This may occur due to the structure of flavonoids, which generally consist of two aromatic rings, each containing at least one hydroxyl group that is connected through a three-carbon "bridge" becoming later part of a heterocyclic ring [47]. These chemical proprieties allow increased permeability across the cellular membrane to interact with multiple intracellular targets [48, 49]. As such, these compounds possess a broad spectrum of biological activities [50, 51], leading to the overrepresentation of many biological pathways, which may not be necessarily linked to antiviral potential. In this sense, systems pharmacology or chemobiology strategies could be employed to define specific targets of flavonoids.

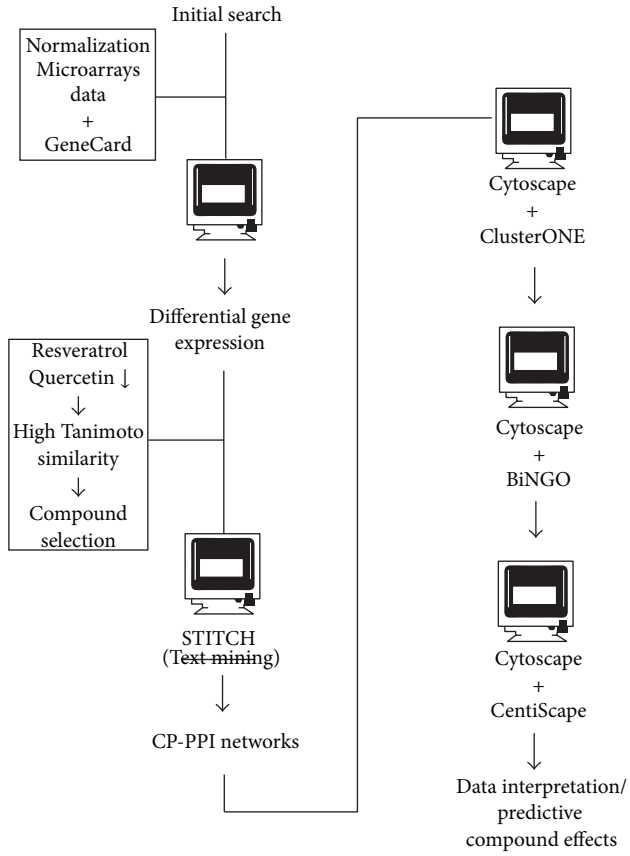


Figure 1: Experimental approach employed to define potential treatments against RSV infection. The interactome data was obtained from microarrays data derived from human bronchial cells infected with RSV. Differential gene expression was considered as initial input for network prospection. Additionally, the natural compounds from flavonoids obtained according to Tanimoto similarity were added to the initial input in STITCH software. The CP-PPI network generated was viewed by Cytoscape and analyzed by ClusterONE in order to identify the major clusters associated. Biological processes found within clusters were retrieved by employing BiNGO plugin. Moreover, to find bottlenecks and hubs, proteins/compounds used CentiScape plugin. Finally, data interpretation was performed based on Zinc database and Osiris Property Explorer.

3.1. Topological Design and Analysis of a Major CP-PPI Network of PHBE Cells Infected by RSV. To focus on RSV antiviral effects of flavonoids, we developed an interatomic network considering 285 genes differentially expressed during RSV infection of PHBE cells and 26 flavonoids compounds (Table 1) as an initial input on STITCH software. As a result of this approach, a major CP-PPI network composed of 57 nodes and 92 edges and integrated by five compound targets with putative antiviral activity was obtained (Figure 2). It is important to note that minor networks without CPI were also detected but were not considered for posterior analysis (Supplementary Figure 1; see Supplementary Material available online at http://dx.doi.org/10.1155/2014/301635).

Network topological features could successfully predict FLA mechanisms of action against RSV infection. In this

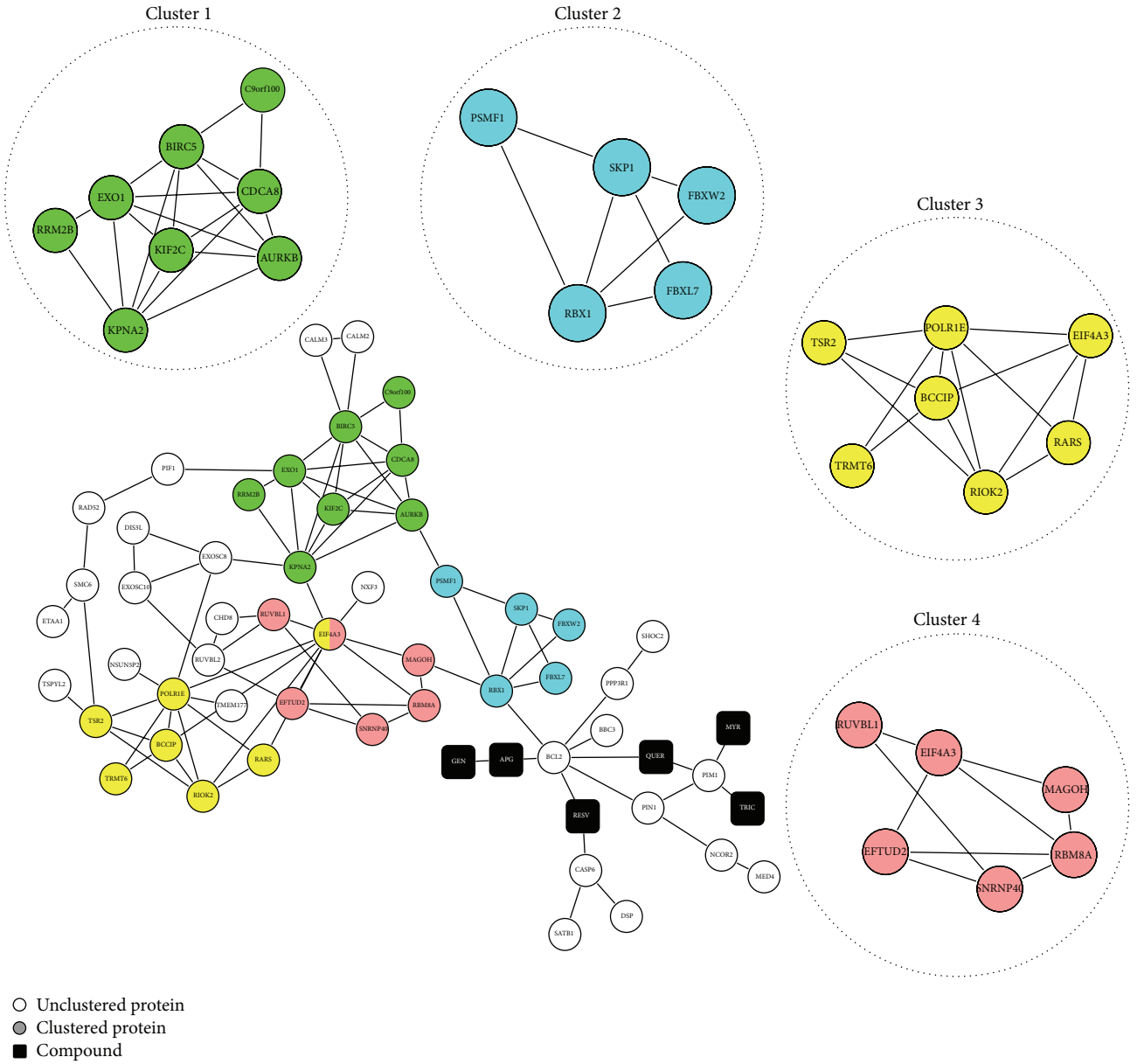


FIGURE 2: ClusterONE analysis of major chemical-protein (CP) and protein-protein interacting (PPI) network. All clustered proteins (composing different subnetworks) and unclustered proteins are represented by nodes of different colors. Chemical compounds are represented by square shape nodes. FLA compounds abbreviations: resveratrol (RESV), apigenin (APG), quercetin (QUER), myricetin (MYR), tricetin (TRIC), and genistein (GEN).

sense, the global organization of clustering in the major network suitable for flavonoid modulation was analyzed. ClusterONE identified four interconnected clusters (Figure 2). Subnetworks of these clusters were created, representing four discrete biological processes, as identified by gene ontology analysis (GO) (Supplementary Table 1): (1) cell cycle phase (corrected P value: 2.33×10^{-6}); (2) ubiquitin-dependent protein catabolic process (corrected P value: 1.61×10^{-5}); (3) nucleic acid metabolic process (corrected P value: 4.68×10^{-4}); and (4) RNA splicing (corrected P value: 1.65×10^{-6}). RSV-host studies have identified these processes that occur upon infection [52–54]. However, all flavonoids and their targets are unclustered in the major CP-PPI network. This shows a compound-target regulation independent of cluster network organization during early RSV infection. An alternative and possible strategy to understand RVS modulation by flavonoids is to predict the best ranking of compound target (high impact on the network) through network connectivity analysis. In this sense, centrality properties were evaluated; however, 11 H-B nodes were identified in the CP-PPI network, represented only by proteins (Figure 3(a), Supplementary Table 2). These same H-B nodes possess high closeness values (Figure 3(b), Supplementary Table 2), suggesting that these nodes may have close communication with others in the

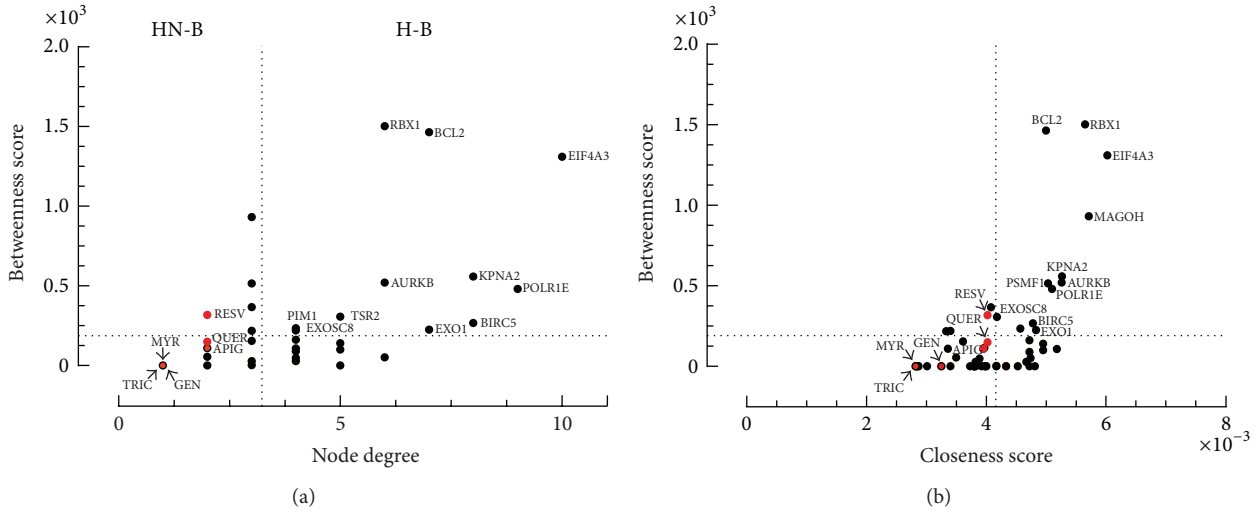


Figure 3: Centrality analysis (a and b) of the major CP-PPI network. Dashed lines represent the threshold value calculated for each centrality. Proteins are represented by black dots, while flavonoid compounds are marked in red. Only proteins or flavonoid compounds with bottleneck scores above the network average are indicated. Legend: hub-bottleneck (H-B); non-hub-bottleneck (NH-B). FLA Compounds abbreviations: resveratrol (RESV); apigenin (APG); quercetin (QUER); myricetin (MYR); tricetin (TRIC), and genistein (GEN).

major network. All flavonoid compounds are H-NB and NH-NB nodes, but these modulate directly 2 H-B proteins (PIM1 and BCL2).

3.1.1. PIM1 and BCL2, as FLA Targets against RSV Infection. PIM1 is a protooncogene which encodes a serine/threonine kinase [55]. This kinase controls cell survival, proliferation, differentiation, and apoptosis [56]. In the context of respiratory diseases, a recent study suggests that PIM1 has a role in the induction of allergic airway responses [57]. Therefore, PIM1 inhibition reduces the development of full spectrum allergen-induced lung inflammatory responses, at least partially through limiting the expansion and actions of CD4+ and CD8 + effector T cells [57]. A similar function for PIM1 has been described in acute RSV infections [58]. PIM1 inhibition attenuates induced RSV reinfection, enhancing airway hyperresponsiveness and activation of the inflammatory cascade. In our analyses, PIM1 showed to be upregulated in comparison with noninfected control (log FC = 0.026) and to interact with three flavonoids (tricetin, myricetin, and quercetin). These compounds are cell-permeable and directly inhibit PIM1 kinase activity [59]. In this sense, these flavonoids are potential inhibitors of RSV-caused inflammation in a target-specific manner, through yet unknown mechanisms. It is important to note that anti-RSV activity of myricetin and tricetin were not tested experimentally and should be further investigated.

On the other hand, our data suggest BLC2 regulation mediated by flavonoids. BCL2 is a regulator of programmed cell death (apoptosis), in part by modulating the release of proapoptotic molecules from mitochondria. For viruses in general (included RSV), apoptotic death of infected cells is a mechanism for reducing virus replication. After 24 h of infection by RSV, several proapoptotic factors of the BCL2 family and caspases 3, 6, 7, 8, 9, and 10 are induced in different epithelial cell lines (primary small airway cells, primary tracheal-bronchial cells, A549, and HEp-2 but not for PHBE) [60]. At the same time, RSV also mediates induction of antiapoptotic factors of the BCL2 family [60], which might account for the delayed induction of apoptosis of RSV-infected cells. This indicates the importance of a complex struggle between apoptotic (host) and antiapoptotic (virus) pathways [60].

In our study, BCL2 was shown to be downregulated in PHBE infected cells in comparison with noninfected controls (log FC = −0.008). We hypothesized that differential expression of this gene may be caused by overexpression of PIM1. In hematopoietic cells, PIM1 kinase acts as a survival factor in cooperation with a regulation of BCL2 [61]. This mechanism should be investigated in RSV infected PHBE.

Furthermore, resveratrol and apigenin control the activity of BCL2 in inducing apoptosis in cancer cells [62, 63], but the effect of these flavonoids has not been explored in PHBE cells or in *in vivo* models for RSV infection. However, these compounds are described as inhibitors of RSV replication *in vitro* (see Table 1).

3.2. In Silico Analysis of FLA Effects on Human Health. We have also predicted potential undesired effects on human health of each of the FLA compounds based on its chemical structures (for more details, see Section 2.7 of Materials and Methods). Our analysis suggests that tricetin may have low risk to human health considering the four main parameters of the analysis (mutagenic, tumorigenic, irritant, and reproductive effectiveness), as shown in Table 2. The other four flavonoids (resveratrol, quercetin, apigenin, and myricetin) may require chemical modification to reduce human health impact but provide versatile chemical backbones for drug development. Biotransformation of flavonoids into drugs is

Table 2: Prediction of effects of FLA compounds based on chemical structure.

Molecules	$x \log P^*$	H-bond acceptors*	H-bond donors*	MV (g/mol)*	Mutagenic**	Tumorigenic**	Irritant**	Reproductive effect**
Resveratrol	2.99	3	3	228.247	High-risk	Low-risk	Low-risk	High-risk
Quercetin	1.68	7	5	302.238	High-risk	Medium-risk	Low-risk	Medium-risk
Apigenin	2.46	5	3	270.24	High-risk	Medium-risk	Low-risk	High-risk
Genistein	2.27	5	3	270.24	High-risk	High-risk	Low-risk	High-risk
Myricetin	1.39	8	6	318.237	High-risk	Low-risk	Low-risk	Low-risk
Tricetin	1.68	5	7	302.238	Low-risk	Low-risk	Low-risk	Low-risk

* All parameters related to Lipinsky's rule of five were obtained from Zinc database.
** All toxicity risks were predicted by Osiris Property Explorer.

the usual approach in the development of anticancer targets [64, 65] but could also be applied in the search of new therapies against RSV.

4. Conclusions

Our model network CPI-PPI identified five target flavonoid compounds: resveratrol, quercetin, tricetin, apigenin, and myricetin. These compounds are suggested as potential candidates in the process of development of novel drugs against early severe RSV infection. Despite these potentially interesting associations, these findings are mainly relying on statistical analysis. Thus, Further experimental testing of these predictions will be required to support the *in silico* data.

Conflict of Interests

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

Acknowledgments

This study was supported by Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (CAPES) Grant no. 02775/09-3. José Eduardo Vargas received postdoctoral fellowships from CAPES-PNPD program.

References

[1] F. W. Denny and F. A. Loda, "Acute respiratory infections are the leading cause of death in children in developing countries," *The American Journal of Tropical Medicine and Hygiene*, vol. 35, no. 1, pp. 1–2, 1986.

[2] C. B. Hall, G. A. Weinberg, M. K. Iwane et al., "The Burden of respiratory syncytial virus infection in young children," *The New England Journal of Medicine*, vol. 360, no. 6, pp. 588–598, 2009.

[3] B. G. Williams, E. Gouws, C. Boschi-Pinto, J. Bryce, and C. Dye, "Estimates of world-wide distribution of child deaths from acute respiratory infections," *The Lancet Infectious Diseases*, vol. 2, no. 1, pp. 25–32, 2002.

[4] W. P. Glezen, L. H. Taber, A. L. Frank, and J. A. Kasel, "Risk of primary infection and reinfection with respiratory syncytial virus," *The American Journal of Diseases of Children*, vol. 140, no. 6, pp. 543–546, 1986.

[5] H. W. Kim, J. G. Canchola, C. D. Brandt et al., "Respiratory syncytial virus disease in infants despite prior administration of antigenic inactivated vaccine," *American Journal of Epidemiology*, vol. 89, no. 4, pp. 422–434, 1969.

[6] J. A. Espinoza, S. M. Bueno, C. A. Riedel, and A. M. Kalergis, "Induction of protective effector immunity to prevent pathogenesis caused by the respiratory syncytial virus. Implications on therapy and vaccine design," *Immunology*, vol. 143, no. 1, pp. 1–12, 2014.

[7] R. A. Karron, B. Thumar, E. Schappell, U. J. Buchholz, and P. L. Collins, "Attenuation of live respiratory syncytial virus vaccines is associated with reductions in levels of nasal cytokines," *The Journal of Infectious Diseases*, vol. 207, no. 11, pp. 1773–1779, 2013.

[8] R. J. Loomis and P. R. Johnson, "Gene-based vaccine approaches for respiratory syncytial virus," *Current Topics in Microbiology and Immunology*, vol. 372, pp. 307–324, 2013.

[9] T. G. Morrison and E. E. Walsh, "Subunit and virus-like particle vaccine approaches for respiratory syncytial virus," *Current Topics in Microbiology and Immunology*, vol. 372, pp. 285–306, 2013.

[10] "Palivizumab, a humanized respiratory syncytial virus monoclonal antibody, reduces hospitalization from respiratory syncytial virus infection in high-risk infants," *Pediatrics*, vol. 102, no. 3, pp. 531–537, 1998.

[11] W. A. Prescott Jr., F. Doloresco, J. Brown, and J. A. Paladino, "Cost effectiveness of respiratory syncytial virus prophylaxis: a critical and systematic review," *PharmacoEconomics*, vol. 28, no. 4, pp. 279–293, 2010.

[12] P. G. Canonico, M. D. Kastello, C. T. Spears, J. R. Brown, E. A. Jackson, and D. E. Jenkins, "Effects of ribavirin on red blood cells," *Toxicology and Applied Pharmacology*, vol. 74, no. 2, pp. 155–162, 1984.

[13] G. Oswald, K. Alzoubi, M. Abed, and F. Lang, "Stimulation of suicidal erythrocyte death by ribavirin," *Basic & Clinical Pharmacology & Toxicology*, vol. 114, no. 4, pp. 311–317, 2013.

[14] A. Cantatore, S. D. Randall, D. Traum, and S. D. Adams, "Effect of black tea extract on herpes simplex virus-1 infection of cultured cells," *BMC Complementary and Alternative Medicine*, vol. 13, article 139, 2013.

[15] J. M. Davis, E. A. Murphy, J. L. McClellan, M. D. Carmichael, and J. D. Gangemi, "Quercetin reduces susceptibility to influenza infection following stressful exercise," *American Journal of Physiology—Regulatory Integrative and Comparative Physiology*, vol. 295, no. 2, pp. R505–R509, 2008.

[16] B. Sritularak, K. Tantrakarnsakul, V. Lipipun, and K. Likhitwitayawuid, "Flavonoids with anti-HSV activity from

the root bark of Artocarpus lakoocha," *Natural Product Communications*, vol. 8, no. 8, pp. 1079–1080, 2013.

[17] G. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pp. 397–420, Springer, Berlin, Germany, 2005.

[18] 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.

[19] T. N. Kaul, E. Middleton Jr., and P. L. Ogra, "Antiviral effect of flavonoids on human viruses," *Journal of Medical Virology*, vol. 15, no. 1, pp. 71–79, 1985.

[20] Y. H. Kim, C. Y. Choi, and Y. Kim, "Covalent modification of the homeodomain-interacting protein kinase 2 (HIPK2) by the ubiquitin-like protein SUMO-1," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 22, pp. 12350–12355, 1999.

[21] M. Thapa, Y. Kim, J. Desper, K.-O. Chang, and D. H. Hua, "Synthesis and antiviral activity of substituted quercetins," *Bioorganic and Medicinal Chemistry Letters*, vol. 22, no. 1, pp. 353–356, 2012.

[22] L. Drago, L. Nicola, F. Ossola, and E. de Vecchi, "*In vitro* antiviral activity of resveratrol against respiratory viruses," *Journal of Chemotherapy*, vol. 20, no. 3, pp. 393–394, 2008.

[23] T. Liu, N. Zang, N. Zhou et al., "Resveratrol inhibits the TRIF-dependent pathway by upregulating sterile alpha and armadillo motif protein, contributing to anti-inflammatory effects after respiratory syncytial virus infection," *Journal of Virology*, vol. 88, no. 8, pp. 4229–4236, 2014.

[24] X. H. Xie, N. Zang, S. M. Li et al., "Resveratrol Inhibits respiratory syncytial virus-induced IL-6 production, decreases viral replication, and downregulates TRIF expression in airway epithelial cells," *Inflammation*, vol. 35, no. 4, pp. 1392–1401, 2012.

[25] R. J. Nijveldt, E. van Nood, D. E. C. van Hoorn, P. G. Boelens, K. van Norren, and P. A. M. van Leeuwen, "Flavonoids: a review of probable mechanisms of action and potential applications," *The American Journal of Clinical Nutrition*, vol. 74, no. 4, pp. 418–425, 2001.

[26] R. P. Garofalo, D. Kolli, and A. Casola, "Respiratory syncytial virus infection: mechanisms of redox control and novel therapeutic opportunities," *Antioxidants and Redox Signaling*, vol. 18, no. 2, pp. 186–217, 2013.

[27] W. D. Guan, Z. F. Yang, N. Liu, S. Qin, F. X. Zhang, and Y. T. Zhu, "[In vitro experimental study on the effect of resveratrol against several kinds of respiroviruses]," *Zhong Yao Cai*, vol. 31, no. 9, pp. 1388–1390, 2008.

[28] J. Li, S. Wang, J. Xu et al., "Regulation trend of resveratrol on TNF-α, IL-1β, IL-6 expressions in bronchoalveolar lavage fluid of RSV-infected BALB/c mice," *Zhongguo Zhong Yao Za Zhi*, vol. 37, no. 10, pp. 1451–1454, 2012.

[29] N. Zang, X. Xie, Y. Deng et al., "Resveratrol-mediated gamma interferon reduction prevents airway inflammation and airway hyperresponsiveness in respiratory syncytial virus-infected immunocompromised mice," *Journal of Virology*, vol. 85, no. 24, pp. 13061–13068, 2011.

[30] J. L. Douglas, M. L. Panis, E. Ho et al., "Inhibition of respiratory syncytial virus fusion by the small molecule VP-14637 via specific interactions with F protein," *Journal of Virology*, vol. 77, no. 9, pp. 5054–5064, 2003.

[31] D. Lai, D. C. Odimegwu, C. Esimone, T. Grunwald, and P. Proksch, "Phenolic compounds with in vitro activity against respiratory syncytial virus from the nigerian lichen Ramalina farinacea," *Planta Medica*, vol. 79, no. 15, pp. 1440–1446, 2013.

[32] S. Sakai, H. Kawamata, T. Kogure et al., "Inhibitory effect of ferulic acid and isoferulic acid on the production of macrophage inflammatory protein-2 in response to respiratory syncytial virus infection in RAW264.7 cells," *Mediators of Inflammation*, vol. 8, no. 3, pp. 173–175, 1999.

[33] Y. Wang, M. Chen, J. Zhang et al., "Flavone C-glycosides from the leaves of Lophatherum gracile and their in vitro antiviral activity," *Planta Medica*, vol. 78, no. 1, pp. 46–51, 2012.

[34] S.-C. Ma, J. Du, P. P.-H. But et al., "Antiviral Chinese medicinal herbs against respiratory syncytial virus," *Journal of Ethnopharmacology*, vol. 79, no. 2, pp. 205–211, 2002.

[35] D. L. Barnard, J. H. Huffman, L. R. Meyerson, and R. W. Sidwell, "Mode of inhibition of respiratory syncytial virus by a plant flavonoid, SP-303," *Chemotherapy*, vol. 39, no. 3, pp. 212–217, 1993.

[36] H. W. M. Rixon, G. Brown, J. T. Murray, and R. J. Sugrue, "The respiratory syncytial virus small hydrophobic protein is phosphorylated via a mitogen-activated protein kinase p38-dependent tyrosine kinase activity during virus infection," *Journal of General Virology*, vol. 86, no. 2, pp. 375–384, 2005.

[37] P. Willett, J. M. Barnard, and G. M. Downs, "Chemical similarity searching," *Journal of Chemical Information and Computer Sciences*, vol. 38, no. 6, pp. 983–996, 1998.

[38] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software Environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[39] T. Nepusz, H. Yu, and A. Paccanaro, "Detecting overlapping protein complexes in protein-protein interaction networks," *Nature Methods*, vol. 9, no. 5, pp. 471–472, 2012.

[40] S. Maere, K. Heymans, and M. Kuiper, "BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks," *Bioinformatics*, vol. 21, no. 16, pp. 3448–3449, 2005.

[41] I. Rivals, L. Personnaz, L. Taing, and M. C. Potier, "Enrichment or depletion of a GO category within a class of genes: which test?" *Bioinformatics*, vol. 23, no. 4, pp. 401–407, 2007.

[42] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, and I. Golani, "Controlling the false discovery rate in behavior genetics research," *Behavioural Brain Research*, vol. 125, no. 1-2, pp. 279–284, 2001.

[43] G. Scardoni, M. Petterlini, and C. Laudanna, "Analyzing biological network parameters with CentiScaPe," *Bioinformatics*, vol. 25, no. 21, pp. 2857–2859, 2009.

[44] M. E. J. Newman, "A measure of betweenness centrality based on random walks," *Social Networks*, vol. 27, no. 1, pp. 39–54, 2005.

[45] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "ZINC: a free tool to discover chemistry for biology," *Journal of Chemical Information and Modeling*, vol. 52, no. 7, pp. 1757–1768, 2012.

[46] J. J. Irwin and B. K. Shoichet, "ZINC—a free database of commercially available compounds for virtual screening," *Journal of Chemical Information and Modeling*, vol. 45, no. 1, pp. 177–182, 2005.

[47] S. J. Flora, "Structural, chemical and biological aspects of antioxidants for strategies against metal and metalloid exposure," *Oxidative Medicine and Cellular Longevity*, vol. 2, no. 4, pp. 191–206, 2009.

[48] J. Brittes, M. Lúcio, C. Nunes, J. L. F. C. Lima, and S. Reis, "Effects of resveratrol on membrane biophysical properties: relevance for its pharmacological effects," *Chemistry and Physics of Lipids*, vol. 163, no. 8, pp. 747–754, 2010.

[49] S. Chaudhuri, A. Banerjee, K. Basu, B. Sengupta, and P. K. Sengupta, "Interaction of flavonoids with red blood cell membrane lipids and proteins: antioxidant and antihemolytic effects," *International Journal of Biological Macromolecules*, vol. 41, no. 1, pp. 42–48, 2007.

[50] A. B. Hendrich, "Flavonoid-membrane interactions: possible consequences for biological effects of some polyphenolic compounds," *Acta Pharmacologica Sinica*, vol. 27, no. 1, pp. 27–40, 2006.

[51] E. Middleton Jr., C. Kandaswami, and T. C. Theoharides, "The effects of plant flavonoids on mammalian cells: implications for inflammation, heart disease, and cancer," *Pharmacological Reviews*, vol. 52, no. 4, pp. 673–751, 2000.

[52] A. Bakre, P. Mitchell, J. K. Coleman et al., "Respiratory syncytial virus modifies microRNAs regulating host genes that affect virus replication," *Journal of General Virology*, vol. 93, part 11, pp. 2346–2356, 2012.

[53] J. Elliott, O. T. Lynch, Y. Suessmuth et al., "Respiratory syncytial virus NS1 protein degrades STAT2 by using the elongin-cullin E3 ligase," *Journal of Virology*, vol. 81, no. 7, pp. 3428–3436, 2007.

[54] W. Wu, D. C. Munday, G. Howell, G. Platt, J. N. Barr, and J. A. Hiscox, "Characterization of the interaction between human respiratory syncytial virus and the cell cycle in continuous cell culture and primary human airway epithelial cells," *Journal of Virology*, vol. 85, no. 19, pp. 10300–10309, 2011.

[55] M. Bachmann and T. Möröy, "The serine/threonine kinase Pim-1," *International Journal of Biochemistry and Cell Biology*, vol. 37, no. 4, pp. 726–730, 2005.

[56] Z. Wang, N. Bhattacharya, M. Weaver et al., "Pim-1: a serine/threonine kinase with a role in cell survival, proliferation, differentiation and tumorigenesis," *Journal of Veterinary Science*, vol. 2, no. 3, pp. 167–179, 2001.

[57] Y. S. Shin, K. Takeda, Y. Shiraishi et al., "Inhibition of Pim1 kinase activation attenuates allergen-induced airway hyperresponsiveness and inflammation," *The American Journal of Respiratory Cell and Molecular Biology*, vol. 46, no. 4, pp. 488–497, 2012.

[58] J. Han, W. Zeng, M. Wang et al., "Inhibition of Pim1 kinase attenuates respiratory syncytial virus (RSV) re-infection-induced enhanced airway hyperresponsiveness (AHR) and inflammation," *The Journal of Allergy and Clinical Immunology*, vol. 131, no. 2, p. AB75, 2013.

[59] S. Holder, M. Lilly, and M. L. Brown, "Comparative molecular field analysis of flavonoid inhibitors of the PIM-1 kinase," *Bioorganic and Medicinal Chemistry*, vol. 15, no. 19, pp. 6463–6473, 2007.

[60] A. Kotelkin, E. A. Prikhod'ko, J. I. Cohen, P. L. Collins, and A. Bukreyev, "Respiratory syncytial virus infection sensitizes cells to apoptosis mediated by tumor necrosis factor-related apoptosis-inducing ligand," *Journal of Virology*, vol. 77, no. 17, pp. 9156–9172, 2003.

[61] M. Lilly, J. Sandholm, J. J. Cooper, P. J. Koskinen, and A. Kraft, "The PIM-1 serine kinase prolongs survival and inhibits apoptosis-related mitochondrial dysfunction in part through a bcl-2-dependent pathway," *Oncogene*, vol. 18, no. 27, pp. 4022–4031, 1999.

[62] S. Ganapathy, Q. Chen, K. P. Singh, S. Shankar, and R. K. Srivastava, "Resveratrol enhances antitumor activity of TRAIL in prostate cancer xenografts through activation of FOXO transcription factor," *PLoS ONE*, vol. 5, no. 12, Article ID e15627, 2010.

[63] Y. Zhu, Y. Mao, H. Chen et al., "Apigenin promotes apoptosis, inhibits invasion and induces cell cycle arrest of T24 human bladder cancer cells," *Cancer Cell International*, vol. 13, no. 1, article 54, 2013.

[64] A. Bartmańska, T. Tronina, J. Popłoński, and E. Huszcza, "Biotransformations of prenylated hop flavonoids for drug discovery and production," *Current Drug Metabolism*, vol. 14, no. 10, pp. 1083–1097, 2013.

[65] M. K. Chahar, N. Sharma, M. P. Dobhal, and Y. C. Joshi, "Flavonoids: a versatile source of anticancer drugs," *Pharmacognosy Reviews*, vol. 5, no. 9, pp. 1–12, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 964795, 8 pages
http://dx.doi.org/10.1155/2015/964795

Research Article

Identification of Novel Thyroid Cancer-Related Genes and Chemicals Using Shortest Path Algorithm

Yang Jiang,[1] Peiwei Zhang,[2] Li-Peng Li,[1] Yi-Chun He,[1] Ru-jian Gao,[1] and Yu-Fei Gao[1]

[1] *Department of Surgery, China-Japan Union Hospital of Jilin University, Changchun 130033, China*
[2] *The Key Laboratory of Stem Cell Biology, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*

Correspondence should be addressed to Yu-Fei Gao; gaoyufei1975@sina.cn

Received 20 September 2014; Accepted 5 December 2014

Academic Editor: Tao Huang

Thyroid cancer is a typical endocrine malignancy. In the past three decades, the continued growth of its incidence has made it urgent to design effective treatments to treat this disease. To this end, it is necessary to uncover the mechanism underlying this disease. Identification of thyroid cancer-related genes and chemicals is helpful to understand the mechanism of thyroid cancer. In this study, we generalized some previous methods to discover both disease genes and chemicals. The method was based on shortest path algorithm and applied to discover novel thyroid cancer-related genes and chemicals. The analysis of the final obtained genes and chemicals suggests that some of them are crucial to the formation and development of thyroid cancer. It is indicated that the proposed method is effective for the discovery of novel disease genes and chemicals.

1. Introduction

Thyroid cancer (TC) is a typical endocrine malignancy. During the past three decades, its incidence has been nearly tripled in the whole world, such as the United States and other developed countries [1]. Thus, it has been a formidable and urgent task to uncover the mechanism behind it, thereby efficiently improving the medical treatment. Research has been focused on the findings of possible driving genes of this disease, especially those genes with high frequent mutations, over-expressions, or fusions for a long time. Until recent years, this research process just started to accelerate.

With the advent of advanced technology including the next-generation sequencing technologies, findings of genetic and epigenetic alterations are speeding up [2]. In other words, the gradual accumulation of somatic mutations and chromosomal rearrangements that are related to many crucial tumor initiation and development genes has been found [3]. For example, high prevalence of mutations and gene fusions in effectors of the PI3K-AKT and MAPK pathway occurred in most patients with TC, suggesting its important contributions to tumor initiation and development. Meanwhile, dysregulation of hundreds of gene expressions, such as DPP4, MET, LGALS3, and TIMP1, have been common events in this disease [4]. This achievement towards the uncovering of mechanism behind TC is inspiring. However, despite the unprecedented rate of discovery of novel mutations and gene fusions in TC, evidence towards the tumor genesis of TC is still not convincing because of the still large search space.

In addition to the influence of our genomes, it is evident that cancer is also influenced by environmental chemicals from our daily lives. This is partly because environmental exposures can cause DNA mutations and change epigenetic mechanisms [5]. For example, we might contact fluoride and arsenic in drinking water, and toxic gases from burning of fuel and industrial emissions. Current studies show that outdoor air pollution and second-hand smoke often contain chemicals, such as arsenic and polycyclic aromatic hydrocarbons, which further increase risks of numerous cancers [6]. Exposure to toxic level of arsenic can significantly increase DNA methylation of p16 and p53 promoter regions [7] and change miRNA expression [8]. However, many chemicals' effects towards cancer have not been researched and illustrated. Considering the important influences of chemicals

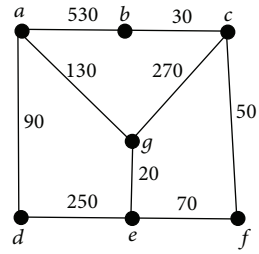
towards cancer, we are also interested in searching for novel chemicals related to TC.

We realized that with the simple results from experiments, it would be difficult to meet up our expectation on the detection of novel genes and chemicals related to TC due to the time- and money-consuming process. Thus, more effective and rapid alternative methods must be used to assist the searching process of genes and chemicals related to TC. Considering the efficiency of computational approach, it might be a potential way, which can be used to complete this arduous searching task in a more effective and time-saving way. Until now, several computational methods have been developed in the field of biological network analysis and other related areas, such as construction and analysis of gene regulation, gene coexpression or other biological networks [9–14], and drug designs [15–21]. Recently, some computation methods were proposed to identify new candidate disease genes based on the knowledge of the known disease genes [22–25]. These methods only considered the disease genes. However, it is easy to improve their methods to identify both genes and chemicals that were related to certain disease. In this study, we generalized their methods by constructing a weighted graph containing the information of protein-protein interactions, chemical-chemical interactions, and chemical-protein interactions and applied this method to study TC. Similar to the methods in [22–25], according to known TC-related genes that were collected from TSGene Database [26], UniPort [27], and NCI (National Cancer Institute) [28] and known TC-related chemicals retrieved from CTD (Comparative Toxicogenomics Database) [29], some new candidate genes and chemicals were discovered by our method. The analysis results of these new candidate genes and chemicals showed that some of them are crucial to the formation and development of TC. We hope that this method could contribute to uncovering the mechanism of TC.

2. Materials and Methods

2.1. Materials. The TC-related genes were collected from three sources: 209 TC-related genes were achieved from UniProt (http://www.uniprot.org/) [27] after we input "human thyroid cancer reviewed" as the keywords; 16 genes were chosen in the catalogue of thyroid cancer from TSGene database (http://bioinfo.mc.vanderbilt.edu/TSGene/search.cgi) [26]; 251 TC-related genes were retrieved from NCI (https://gforge.nci.nih.gov/, released April 2009) [28]. After integrating the above 476 genes, we finally obtained 444 different TC-related genes, which were provided in Online Supporting Information (see S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/964795).

The TC-related chemicals were retrieved from CTD (http://ctdbase.org/) [29], which included the interactions between chemicals and genes and their associations with diseases that were manually curated from 110,142 articles (http://ctdbase.org/about/dataStatus.go, accessed 2014 August). Only the 44 chemicals that were markers of TC, were therapeutic to TC, or had known mechanism in TC were analyzed. The pubchem IDs of these chemicals were also provided in Online Supporting Information S1.



Protein or chemical	Protein or chemical	Combined_score
a	b	470
b	c	970
d	e	750
e	f	930
e	g	980
a	g	870
c	f	950
c	g	730
a	d	910

Figure 1: An example to display the construction of the weighted graph, where a, b, and c represent chemicals and d, e, f, and g represent proteins.

2.2. A Weighted Graph Constructed from Interactions of Chemicals and Proteins. The core idea of our method is to construct a hybrid weighted graph containing the information of proteins, chemicals, and their associations. This idea has been applied to our previous study on assigning chemicals and enzymes to metabolic pathway [30]. To do that, we employed the information of protein-protein interactions, chemical-chemical interactions, and chemical-protein interactions.

The information concerning protein-protein interaction was retrieved from STRING (Search Tool for the Retrieval of Interacting Genes/Proteins, version 9.1, http://string.embl.de/) [31], a large scale database containing direct (physical) and indirect (functional) interactions of proteins, which are derived from genomic context, high-throughput experiments, (conserved) coexpression, or previous knowledge (refer to http://string.embl.de/). Some computational models have been built based on these information [32–35]. Each obtained interaction contains two proteins and one score, which measures the strength of the interaction, that is, the likelihood of the interaction's occurrence. For latter formulation, let us denote the score of the interaction between proteins p_1 and p_2 by $S_{pp}(p_1, p_2)$. In particular, if proteins p_1 and p_2 cannot comprise an interaction according to the current data in STRING, $S_{pp}(p_1, p_2)$ was set to zero.

The information concerning chemical-chemical interaction and chemical-protein interaction was retrieved from STITCH (Search tool for interactions of chemicals, version 4.0, http://stitch.embl.de/) [36], a sister project of STRING which provides the known and predicted interactions of chemicals and proteins. These interactions are confirmed by evidence derived from experiments, databases, and the

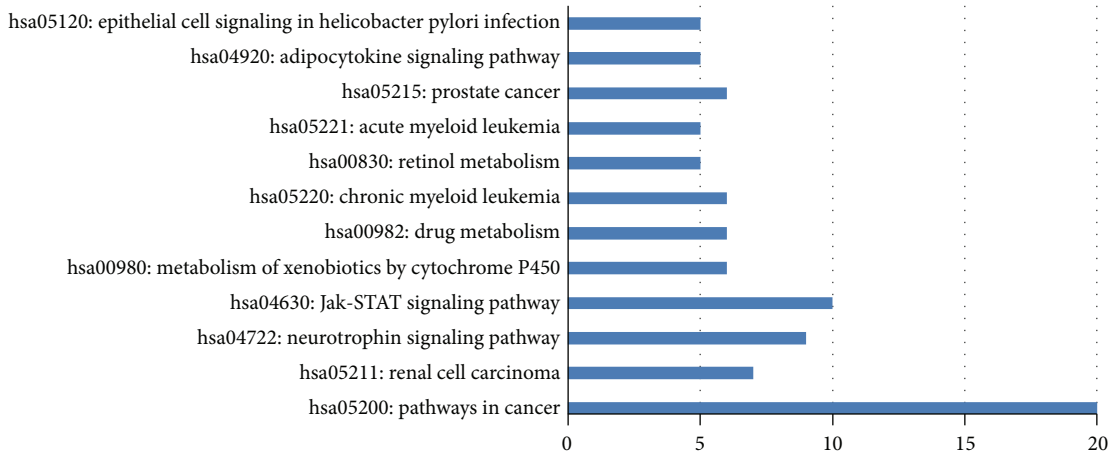


Figure 2: The top twelve KEGG pathways that were enriched by 169 significant candidate genes.

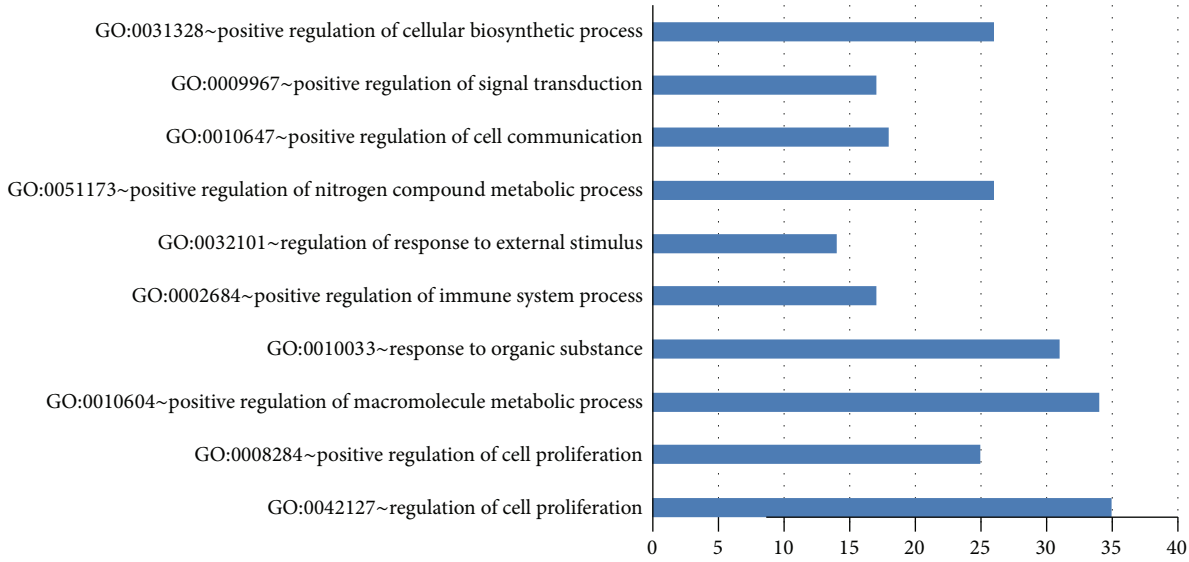


Figure 3: The top ten GO terms that were enriched by 169 significant candidate genes.

literature (refer to http://stitch.embl.de/). Each obtained chemical-chemical interaction contains two chemicals and five scores: "Similarity," "Experiment," "Database," "Textmining," and "Combined_score," which measure the strength of the interaction from different aspects, such as their structures, activities, reactions, cooccurrence in literature, and integration of the above information. To widely indicate the interaction between chemicals, we used the last score, that is, "Combined_score," to measure the strength of the interaction. For two chemicals c_1 and c_2, the "Combined_score" of the interaction between them was denoted by $S_{cc}(c_1, c_2)$. Similarly, $S_{cc}(c_1, c_2)$ was set to zero if c_1 and c_2 do not occur as an interaction in STITCH. Each obtained chemical-protein interaction contains one chemical, one protein, and five scores. With the similar argument, we used the "Combined_score" to indicate the strength of the interaction between one chemical and one protein. Let $S_{cp}(c, p)$ denote the "Combined_score" of the interaction between chemical c and protein p. Also, $S_{cp}(c, p) = 0$ if c and p cannot comprise a chemical-protein interaction. It is necessary to point out that all chemicals in the retrieved chemical-chemical and chemical-protein interactions must have records in KEGG (Kyoto Encyclopedia of Genes and Genomes) [37] because the number of chemicals in STITCH is too large and most of chemicals have few associations with human tissues.

Based on the information concerning protein-protein interactions, chemical-chemical interactions, and chemical-protein interactions, a weighted graph $G = (V, E)$ was constructed as follows: V contained all proteins and chemicals occurring in the above three kinds of information and E consisted of all pairs of nodes such that the corresponding proteins or chemicals can comprise an interaction. It is easy to know that each edge in G represented an interaction. On the other hand, as mentioned in the above paragraph, each interaction was assigned a score to indicate its strength; that is, different interactions may have different strength. To note

this fact in G and use the shortest path algorithm to search for new candidate genes and chemicals, each edge was labeled a weight as follows. Since the range of the interaction scores is between 1 and 999, the weight of an edge e with endpoints n_1 and n_2 was defined by

$$w(e) = \begin{cases} 1000 - S_{pp}(p_1, p_2) \\ \text{If } n_1 \text{ and } n_2 \text{ represented proteins } p_1 \text{ and } p_2 \\ 1000 - S_{cp}(c, p) \\ \text{If } n_1 \text{ and } n_2 \text{ represented chemical } c \text{ and protein } p \\ 1000 - S_{cc}(c_1, c_2) \\ \text{If } n_1 \text{ and } n_2 \text{ represented chemicals } c_1 \text{ and } c_2. \end{cases} \quad (1)$$

To clearly display the procedures of construction of the graph, a small example is shown in Figure 1. In the example, there are three chemicals a, b, and c and four proteins d, e, f, and g. The interactions, including their "Combined_score," between them are listed in the table of Figure 1 and the constructed graph based on these interactions is shown at the top of Figure 1.

2.3. Method for Discovery of New Candidate Genes and Chemicals. The following method for finding new candidate TC-related genes and chemicals was almost same as that in our previous study [25]. The only difference was that the input of the current method contained both genes and chemicals, while the method in [25] only considered genes. Readers can refer to our previous study [25] for the detailed procedures of the method and its principle. For the integrity of this study, the brief description of the method was as follows: (I) search all shortest paths connecting any pair of TC-related genes and chemicals using Dijkstra's algorithm [38]; (II) for each node (gene or chemical) in G, count its betweenness that was defined as the number of paths containing it as an inner node; (III) select the nodes (genes or chemicals) with betweenness larger than zero as candidate genes and chemicals; (IV) produce 1,000 sets by randomly selecting nodes (genes or chemicals) from G; the numbers of genes and chemicals in each set were the same as those in known TC-related gene and chemical set; (V) for each set, search all shortest paths connecting any pair of genes or chemicals in G; (VI) count the betweenness of candidate genes and chemicals on each randomly produced sets; (VII) for each candidate gene and chemical, compare its betweenness on known TC-related gene and chemical set and those on randomly produced sets, thereby calculating its permutation FDR that was defined as "the number of randomly produced sets on which the betweenness was larger than that on the known TC-related gene and chemical set"/1000.

3. Results and Discussions

3.1. Candidate Genes and Chemicals. Of the 444 TC-related genes and 44 TC-related chemicals, we searched the shortest paths in G such that the endpoints of them were TC-related genes or TC-related chemicals. Accordingly, the betweenness of each gene and chemical in G was computed, obtaining 636 candidate genes and 174 candidate chemicals whose betweenness was larger than zero; that is, these genes and chemicals occurred in at least one shortest path as inner nodes. These genes and chemicals are listed in Online Supporting Information S2, in which their betweenness is also provided.

To exclude false discoveries, the permutation test was executed by constructing 1,000 randomly selected gene and chemical sets for calculating the permutation FDR of each candidate gene and chemical, which is also provided in Online Supporting Information S2. Then, we selected 0.05 as a threshold to exclude false discoveries, obtaining 169 candidate genes and 49 candidate chemicals with permutation FDRs smaller than 0.05. The information of these genes and chemicals is available in Online Supporting Information S3. For convenience, we termed these genes and chemicals as significant candidate genes and significant candidate chemicals, respectively. The following discussion was based on these significant candidate genes and significant candidate chemicals.

3.2. Gene Enrichment Analysis. DAVID [39] is a powerful tool that could be used to make integrative and systematic of large gene lists. Thus, it was used in this study to analyze the 169 significant candidate genes. The analysis results included two parts: KEGG pathway enrichments (Online Supporting Information S4) and gene ontology (GO) enrichments (Online Supporting Information S5). GO enrichments include three parts: biological process enrichment (BP enrichment), cellular component enrichment (CC enrichment), and molecular function enrichment (MF enrichment). Since our method was mainly based on protein-protein interactions, BP enrichment analysis was more convincing, while other two results were not very reasonable. Thus, we only gave the discussion based on the BP enrichment.

For the KEGG pathway enrichment analysis results, 169 candidate genes are enriched in 19 KEGG pathways (see Online Supporting Information S4). Among these 19 KEGG pathways, twelve of them were with P value (modified Fisher exact P value) less than 0.05. Figure 2 shows these twelve KEGG pathways and the number of enriched genes among the significant candidate genes ("count"). Hsa05200 (pathways in cancer, "count" = 20) is the most significant pathway, which enriched 20 significant candidate genes, such as FGFR2, FGF6, DVL3, EPAS1, and PPARG. Since all these genes enriched in this pathway were reported related to cancer formation and development, it further revealed the validity of our method. Hsa05211 (renal cell carcinoma, "count" = 7) is the second significant pathway with 7 genes related to renal cell carcinoma. Hsa04722 (neurotrophin signaling pathway, "count" = 7) is the third significant pathway, enriching 7 genes, such as KRAS, PLCG1, and NTF3. Among

them, NTF3 in neurotrophin signaling pathway has been reported with the association to cancer [40]. Other pathways, such as hsa05221 (acute myeloid leukemia, "count" = 5) and hsa05215 (prostate cancer, "count" = 6), also revealed that the significant candidate genes are associated with cancer.

For the BP enrichment analysis, results are shown in Online Supporting Information S5. Ranked by P value, top ten BP GO terms are depicted in Figure 3. The mainly enriched GO terms are associated with cell proliferation. For example, genes in GO:0042127 (regulation of cell proliferation, "count" = 35) and GO:0008284 (positive regulation of cell proliferation, "count" = 25) are all reported related to cell proliferation. Also, GO:0010604 (positive regulation of macromolecule metabolic process, "count" = 34) and GO:0051173 (positive regulation of nitrogen compound metabolic process, "count" = 26) are associated with metabolic process. Since proliferative signaling and activating metastasis are two hallmarks of cancer [41], it is convincing that the result of BP enrichment analysis further supports the validity of our method.

Thus, this enrichment analysis further proved the importance of genes discovered by our method. We hope that it could be used to gain better understandings of the mechanism of TC.

3.3. Analysis of Some Significant Candidate Genes. Among 169 significant candidate genes, we selected some important genes to elucidate their potential values to be TC-related genes. Since they have been reported to be associated with the tumorigenesis or development of other types of cancers, we thought it might lend credence to our method and make our findings more convincing.

The gene CYP2B6 (cytochrome P450, family 2, subfamily B, polypeptide 6) mainly encodes enzymes which are involved in many reactions, specifically in anticancer drug metabolism. A report based on one Japanese population showed that polymorphism of CYP2B6 is significantly associated with prostate cancer risk [42]. Also, decreased expression of CYP2B6 is shown in prostate cancer, and it has been recognized as growth inhibitory [43].

FURIN, also known as PACE, encodes furin protein. High expression of furin has been detected in different cancer types, such as ovarian cancer [44] and head and neck cancer cells [45]. And the inhibition of its expression can help decrease the tumorigenesis of cancers [46]. Also, furin overexpression can promote cell invasion in human hepatoma cell lines, which plays a role in the development of hepatocellular carcinoma [47]. Moreover, the gene may involve in the activity of Notch, and the Notch pathway is important during the medullary thyroid cancer (MTC) [48].

MERTK (c-mer proto-oncogene tyrosine kinase) is a protein-coding gene, which belongs to the MER/AXL/TYRO3 receptor kinase family and encodes cell-surface transmembrane receptors that contain regulated kinase activity [49]. Research has found that MERTK is overexpressed in a variety of cancers, such as prostate cancer, non-small-cell lung cancer, and breast cancer [50]. Also, its overexpression can result in the activation of oncogenic signaling pathways and drive cell transformation in cancer cells [51].

OAS2 ($2'$-$5'$-oligoadenylate synthetase 2, 69/71 kDa) is involved in immune response of viral infection, because it activates RNase L as a result of the elimination of viruses. In a recent study of cervical cancer, researchers found that genes related to antiviral response were increasingly expressed, including OAS2 which is directly involved in viral RNA degradation [52].

PPARG (peroxisome proliferator-activated receptor gamma) is a member of PPAR subfamily of nuclear receptors, which plays a crucial role in the regulation of gene transcription and adipocyte differentiation. Currently, the activation of PPARG has been recognized as one key step in colorectal cancer progression [53], and its deacetylation can determine lipid synthesis and growth in breast tumor [54].

To summarize, even though these 169 significant candidate genes have not been found associated with TC until now, a wealth of evidence has proved their relations to other types of cancer. Therefore, previous researches have validated the reliability of our method and the importance of our findings. We hope our method will be helpful to search novel TC-related genes and be further promoted to the exploration of other biological questions.

3.4. Analysis of Some Candidate Chemicals. Besides the significant candidate genes, we also discovered 49 significant candidate chemicals that are deemed to be related to thyroid cancer development. Most of them (29 out of 49) can be supported by published literatures. Here, we only gave detailed discussions for three of them. All of these 29 chemicals are briefly discussed in Online Supporting Information S6.

Chloride ion (CID000000312) is a common ion in human cells, which plays a crucial role in cell invasion due to its ability to change the osmotic balance between the inner- and extra-cellular space [55]. The reason behind invading cancer cells that can pass though extracellular matrix is partly because it has the ability to reduce its volume. Several major chloride channels on the cell membrane are responsible for this invasive behavior of cancer cells. Research has found that inhibition of the sodium-potassium-chloride cotransporter isoform-1 (NKCC1) can decrease cell invasion by 50% [56].

Hydrogen cyanide (HCN, CID000000768) is the product of various tobaccos, existing in the smoke as a colorless gas. In the study of gastroesophageal cancer based on selected ion flow tube mass spectrometry (SIFT-MS), hydrogen cyanide is significantly different between cancer and healthy groups [57]. Hydrogen cyanide is also recognized to have cardiovascular and respiratory toxicity, which might be a potential factor to cause lung cancer [58].

Aniline (CID000006115) consists of a phenyl group attached to an amino group, and it is the precursor of industrial chemicals. It is reported that the incidence of bladder cancer is clearly related to exposure to aniline [59]. Potential reasons might be due to an increase in iron overload in the spleen and upregulation of TNF-α, IL-1, and IL-6. Also, the expression of cyclin dependent kinases (CDKs) is upregulated by aniline [60].

4. Conclusion

During the fight with thyroid cancer, discovery of its related genes and chemicals and uncovering the mechanism behind it are important to today's research and future's drug design for designing effective treatments. Only with the assistance of experiment methods would be an onerous and low efficient way. In this study, we sufficiently used known resource, such as protein-protein interactions, chemical-chemical interactions, chemical-protein interactions, and known thyroid cancer-related genes and chemicals, to search new candidate thyroid cancer-related genes and chemicals by the shortest path algorithm. The proposed method generalized our previous method that can only discover disease genes. Further analysis of the selected genes and chemicals implies that some of them have direct or indirect relationship with the formation and development of thyroid cancer, thereby suggesting the effectiveness of our method. We hope that our method and the findings could shed new light on the mechanism research of thyroid cancer.

Conflict of Interests

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

Authors' Contribution

Yang Jiang and Peiwei Zhang contributed equally to this work.

Acknowledgments

This paper is supported by National Natural Science Foundation of China (81372696), China Postdoctoral Science Foundation (2013M541314), Jilin Provincial Science & Technology Department (20090175 and 20100733), Natural Science Fund Projects of Jilin province (201215059), Development of Science and Technology Plan Projects of Jilin province (20100733 and 201101074), Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (2009-36), Scientific Research Foundation (Jilin Department of Science & Technology, 200705314, 20090175, and 20100733), Health and Family Planning Commission of Jilin Province (2010Z068), Human resources and Social Security Department of Jilin Province (2012–2014), Postdoctoral Science Foundation of Jilin Province, and Human resources and Social Security Department of Jilin Province (2012).

References

[1] Y. E. Nikiforov and M. N. Nikiforova, "Molecular genetics and diagnosis of thyroid cancer," *Nature Reviews Endocrinology*, vol. 7, no. 10, pp. 569–580, 2011.

[2] M. A. Ginos, G. P. Page, B. S. Michalowicz et al., "Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck," *Cancer Research*, vol. 64, no. 1, pp. 55–63, 2004.

[3] A. Mathur, W. Moses, R. Rahbari et al., "Higher rate of BRAF mutation in papillary thyroid cancer over time: a single-institution study," *Cancer*, vol. 117, no. 19, pp. 4390–4395, 2011.

[4] S. Chevillard, N. Ugolin, P. Vielh et al., "Gene expression profiling of differentiated thyroid neoplasms: diagnostic and clinical implications," *Clinical Cancer Research*, vol. 10, no. 19, pp. 6586–6597, 2004.

[5] A. Baccarelli and V. Bollati, "Epigenetics and environmental chemicals," *Current Opinion in Pediatrics*, vol. 21, no. 2, pp. 243–251, 2009.

[6] A. Prüss-Ustün, C. Vickers, P. Haefliger, and R. Bertollini, "Knowns and unknowns on burden of disease due to chemicals: a systematic review," *Environmental Health: A Global Access Science Source*, vol. 10, no. 1, article 9, 2011.

[7] S. Chanda, U. B. Dasgupta, D. GuhaMazumder et al., "DNA hypermethylation of promoter of gene p53 and p16 in arsenic-exposed people with and without malignancy," *Toxicological Sciences*, vol. 89, no. 2, pp. 431–437, 2006.

[8] C. J. Marsit, K. Eddy, and K. T. Kelsey, "MicroRNA responses to cellular stress," *Cancer Research*, vol. 66, no. 22, pp. 10843–10848, 2006.

[9] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, no. 5643, pp. 249–255, 2003.

[10] L. Chen, B.-Q. Li, and K.-Y. Feng, "Predicting biological functions of protein complexes using graphic and functional features," *Current Bioinformatics*, vol. 8, no. 5, pp. 545–551, 2013.

[11] D. Warde-Farley, S. L. Donaldson, O. Comes et al., "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Research*, vol. 38, no. 2, Article ID gkq537, pp. W214–W220, 2010.

[12] L. Chen, W.-M. Zeng, Y.-D. Cai, and T. Huang, "Prediction of metabolic pathway using graph property, chemical functional group and chemical structural set," *Current Bioinformatics*, vol. 8, no. 2, pp. 200–207, 2013.

[13] D. Marbach, J. C. Costello, R. Küffner et al., "Wisdom of crowds for robust gene network inference," *Nature Methods*, vol. 9, no. 8, pp. 796–804, 2012.

[14] T. Huang, L. Chen, Y.-D. Cai, and K.-C. Chou, "Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property," *PLoS ONE*, vol. 6, no. 9, Article ID e25297, 2011.

[15] L. Chen, W.-M. Zeng, Y.-D. Cai, K.-Y. Feng, and K.-C. Chou, "Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities," *PLoS ONE*, vol. 7, no. 4, Article ID e35254, 2012.

[16] Y. C. Wang, N. Deng, S. Chen, and Y. Wang, "Computational study of drugs by integrating omics data with kernel methods," *Molecular Informatics*, vol. 32, no. 11-12, pp. 930–941, 2013.

[17] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.

[18] L. Chen, J. Lu, T. Huang et al., "Finding candidate drugs for hepatitis C based on chemical-chemical and chemical-protein interactions," *PLoS ONE*, vol. 9, no. 9, Article ID e107767, 2014.

[19] F. Napolitano, Y. Zhao, V. M. Moreira et al., "Drug repositioning: a machine-learning approach through data integration," *Journal of Cheminformatics*, vol. 5, article 30, 2013.

[20] L. Chen, J. Lu, N. Zhang, T. Huang, and Y.-D. Cai, "A hybrid method for prediction and repositioning of drug anatomical therapeutic chemical classes," *Molecular BioSystems*, vol. 10, no. 4, pp. 868–877, 2014.

[21] L. Wu, N. Ai, Y. Liu, Y. Wang, and X. Fan, "Relating anatomical therapeutic indications by the ensemble similarity of drug sets," *Journal of Chemical Information and Modeling*, vol. 53, no. 8, pp. 2154–2160, 2013.

[22] B.-Q. Li, T. Huang, L. Liu, Y.-D. Cai, and K.-C. Chou, "Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network," *PLoS ONE*, vol. 7, no. 4, Article ID e33393, 2012.

[23] M. Jiang, Y. Chen, Y. Zhang et al., "Identification of hepatocellular carcinoma related genes with k-th shortest paths in a protein-protein interaction network," *Molecular BioSystems*, vol. 9, no. 11, pp. 2720–2728, 2013.

[24] J. Zhang, M. Jiang, F. Yuan et al., "Identification of age-related macular degeneration related genes by applying shortest path algorithm in protein-protein interaction network," *BioMed Research International*, vol. 2013, Article ID 523415, 8 pages, 2013.

[25] Y.-F. Gao, Y. Shu, L. Yang et al., "A graphic method for identification of novel glioma related genes," *BioMed Research International*, vol. 2014, Article ID 891945, 8 pages, 2014.

[26] M. Zhao, J. Sun, and Z. Zhao, "TSGene: a web resource for tumor suppressor genes," *Nucleic Acids Research*, vol. 41, no. 1, pp. D970–D976, 2013.

[27] UniProt Consortium, "Update on activities at the Universal Protein Resource (UniProt) in 2013," *Nucleic Acids Research*, vol. 41, pp. D43–D47, 2012.

[28] S. McNeil, A. Budhu, N. Grantees et al., *Imaging*, National Cancer Institute, 2013.

[29] A. P. Davis, C. G. Murphy, R. Johnson et al., "The comparative toxicogenomics database: update 2013," *Nucleic Acids Research*, vol. 41, no. 1, pp. D1104–D1114, 2013.

[30] Y. F. Gao, L. Chen, Y. D. Cai, K. Y. Feng, T. Huang, and Y. Jiang, "Predicting metabolic pathways of small molecules and enzymes based on interaction information of chemicals and proteins," *PLoS ONE*, vol. 7, no. 9, Article ID e45944, 2012.

[31] L. J. Jensen, M. Kuhn, M. Stark et al., "STRING 8—a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Research*, vol. 37, no. 1, pp. D412–D416, 2009.

[32] L. Hu, T. Huang, X.-J. Liu, and Y.-D. Cai, "Predicting protein phenotypes based on protein-protein interaction network," *PLoS ONE*, vol. 6, no. 3, Article ID e17668, 2011.

[33] L.-L. Hu, T. Huang, Y.-D. Cai, and K.-C. Chou, "Prediction of body fluids where proteins are secreted into based on protein interaction network," *PLoS ONE*, vol. 6, no. 7, Article ID e22989, 2011.

[34] P. Gao, Q.-P. Wang, L. Chen, and T. Huang, "Prediction of human genes' regulatory functions based on proteinprotein interaction network," *Protein and Peptide Letters*, vol. 19, no. 9, pp. 910–916, 2012.

[35] L. L. Hu, T. Huang, X. Shi, W.-C. Lu, Y.-D. Cai, and K.-C. Chou, "Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties," *PLoS ONE*, vol. 6, no. 1, Article ID e14556, 2011.

[36] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: interaction networks of chemicals and proteins," *Nucleic Acids Research*, vol. 36, no. 1, pp. D684–D688, 2008.

[37] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[38] T. H. Gormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Eds., *Introduction to Algorithms*, MIT Press, Cambridge, Mass, USA, 1990.

[39] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.

[40] A. A. Bapat, G. Hostetter, D. D. von Hoff, and H. Han, "Perineural invasion and associated pain in pancreatic cancer," *Nature Reviews Cancer*, vol. 11, no. 10, pp. 695–707, 2011.

[41] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[42] T. Kurosaki, M. Suzuki, Y. Enomoto et al., "Polymorphism of cytochrome P450 2B6 and prostate cancer risk: a significant association in a Japanese population," *International Journal of Urology*, vol. 16, no. 4, pp. 364–368, 2009.

[43] J. Kumagai, T. Fujimura, S. Takahashi et al., "Cytochrome P450 2B6 is a growth-inhibitory and prognostic factor for prostate cancer," *The Prostate*, vol. 67, no. 10, pp. 1029–1037, 2007.

[44] R. E. Page, A. J. P. Klein-Szanto, S. Litwin et al., "Increased expression of the pro-protein convertase furin predicts decreased survival in ovarian cancer," *Cellular Oncology*, vol. 29, no. 4, pp. 289–299, 2007.

[45] D. E. Bassi, H. Mahloogi, R. L. de Cicco, and A. Klein-Szanto, "Increased furin activity enhances the malignant phenotype of human head and neck cancer cells," *The American Journal of Pathology*, vol. 162, no. 2, pp. 439–447, 2003.

[46] D. E. Bassi, R. L. de Cicco, H. Mahloogi, S. Zucker, G. Thomas, and A. J. P. Klein-Szanto, "Furin inhibition results in absent or decreased invasiveness and tumorigenicity of human cancer cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 18, pp. 10326–10331, 2001.

[47] R.-N. Chen, Y.-H. Huang, Y.-C. Lin et al., "Thyroid hormone promotes cell invasion through activation of furin expression in human hepatoma cell lines," *Endocrinology*, vol. 149, no. 8, pp. 3817–3831, 2008.

[48] M. Cook, X.-M. Yu, and H. Chen, "Notch in the development of thyroid C-cells and the treatment of medullary thyroid cancer," *American Journal of Translational Research*, vol. 2, no. 1, pp. 119–125, 2010.

[49] A. Verma, S. L. Warner, H. Vankayalapati, D. J. Bearss, and S. Sharma, "Targeting Axl and Mer kinases in cancer," *Molecular Cancer Therapeutics*, vol. 10, no. 10, pp. 1763–1773, 2011.

[50] C. T. Cummings, D. DeRyckere, H. S. Earp, and D. K. Graham, "Molecular pathways: MERTK signaling in cancer," *Clinical Cancer Research*, vol. 19, no. 19, pp. 5275–5280, 2013.

[51] K. Q. Nguyen, W. I. Tsou, D. A. Calarese et al., "Overexpression of MERTK receptor tyrosine kinase in epithelial cancer cells drives efferocytosis in a gain-of-function capacity," *The Journal of Biological Chemistry*, vol. 289, no. 37, pp. 25737–25749, 2014.

[52] K. L. Mine, N. Shulzhenko, A. Yambartsev et al., "Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer," *Nature Communications*, vol. 4, article 1806, 2013.

[53] L. Sabatino, A. Fucci, M. Pancione et al., "UHRF1 coordinates peroxisome proliferator activated receptor gamma (PPARG) epigenetic silencing and mediates colorectal cancer progression," *Oncogene*, vol. 31, no. 49, pp. 5061–5072, 2012.

[54] R. Pestell, L. Tian, C. Wang et al., "Abstract P2-06-02: Pparg deacetylation by SIRT1 determines breast tumor lipid synthesis and growth," *Cancer Research*, vol. 73, p. P2-06-02, 2014.

[55] O. Veiseh, F. M. Kievit, R. G. Ellenbogen, and M. Zhang, "Cancer cell invasion: treatment and monitoring opportunities in nanomedicine," *Advanced Drug Delivery Reviews*, vol. 63, no. 8, pp. 582–596, 2011.

[56] B. R. Haas and H. Sontheimer, "Inhibition of the sodium-potassium-chloride cotransporter isoform-1 reduces glioma invasion," *Cancer Research*, vol. 70, no. 13, pp. 5597–5606, 2010.

[57] S. Kumar, J. Huang, J. R. Cushnir, P. Španěl, D. Smith, and G. B. Hanna, "Selected ion flow tube-MS analysis of headspace vapor from gastric content for the diagnosis of gastro-esophageal cancer," *Analytical Chemistry*, vol. 84, no. 21, pp. 9550–9557, 2012.

[58] P. J. Branton, K. G. McAdam, D. B. Winter, C. Liu, M. G. Duke, and C. J. Proctor, "Reduction of aldehydes and hydrogen cyanide yields in mainstream cigarette smoke using an amine functionalised ion exchange resin," *Chemistry Central Journal*, vol. 5, no. 1, article 15, 2011.

[59] T. Carreón, M. Hein, K. Hanley, S. Viet, and A. Ruder, "0094 Bladder cancer incidence among workers exposed to o-toluidine, aniline and nitrobenzene at a rubber chemical manufacturing plant," *Occupational & Environmental Medicine*, vol. 71, pp. A9–A10, 2014.

[60] S. Kannan, R. Fielder, J. Tristan, E. Longoria, and A. Castillon, "Molecular mechanism of aniline induced bladder cancer," *The FASEB Journal*, vol. 27, pp. 793–791, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 576349, 8 pages
http://dx.doi.org/10.1155/2015/576349

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

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:

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

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 $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$,

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

(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 $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$, calculate the following:

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

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 $i = 1, 2, \ldots, n$, 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 $<, >$ indicates the inter product and $\|\cdot\|$ is the Euclidean norm. Consider

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

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

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

where $r_{g,i}$ is the rank of gene g in the score list of the ith 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:

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

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.

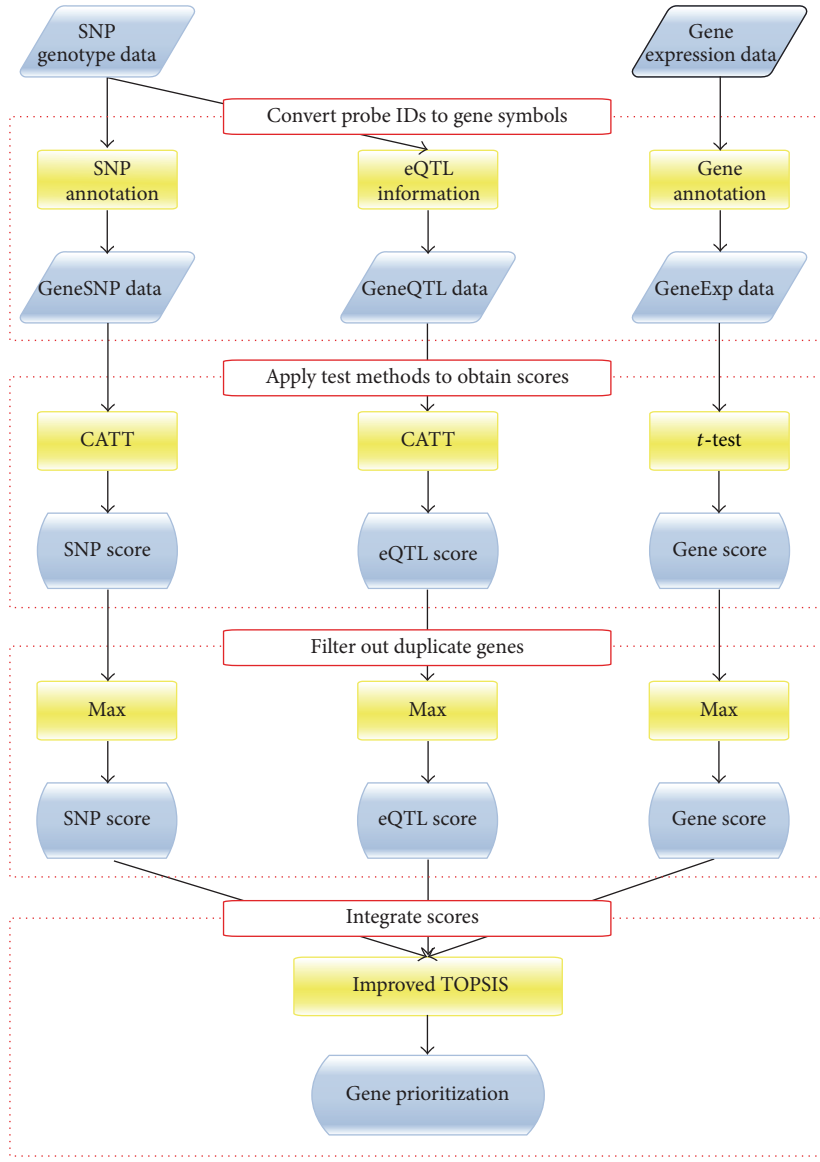
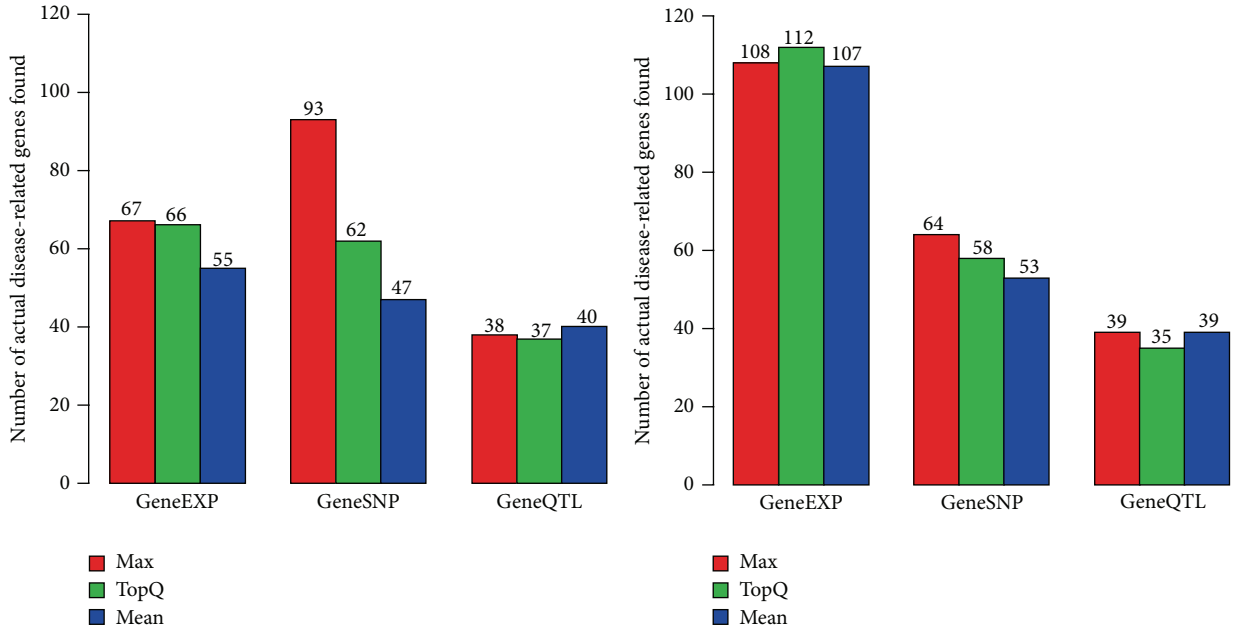


Figure 1: Overall procedure of the proposed strategy for gene prioritization which consists of four steps: (1) convert probe IDs to gene symbols, (2) apply test methods to obtain scores, (3) filter out duplicate genes in each score list, and (4) integrate scores with improved TOPSIS.

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.

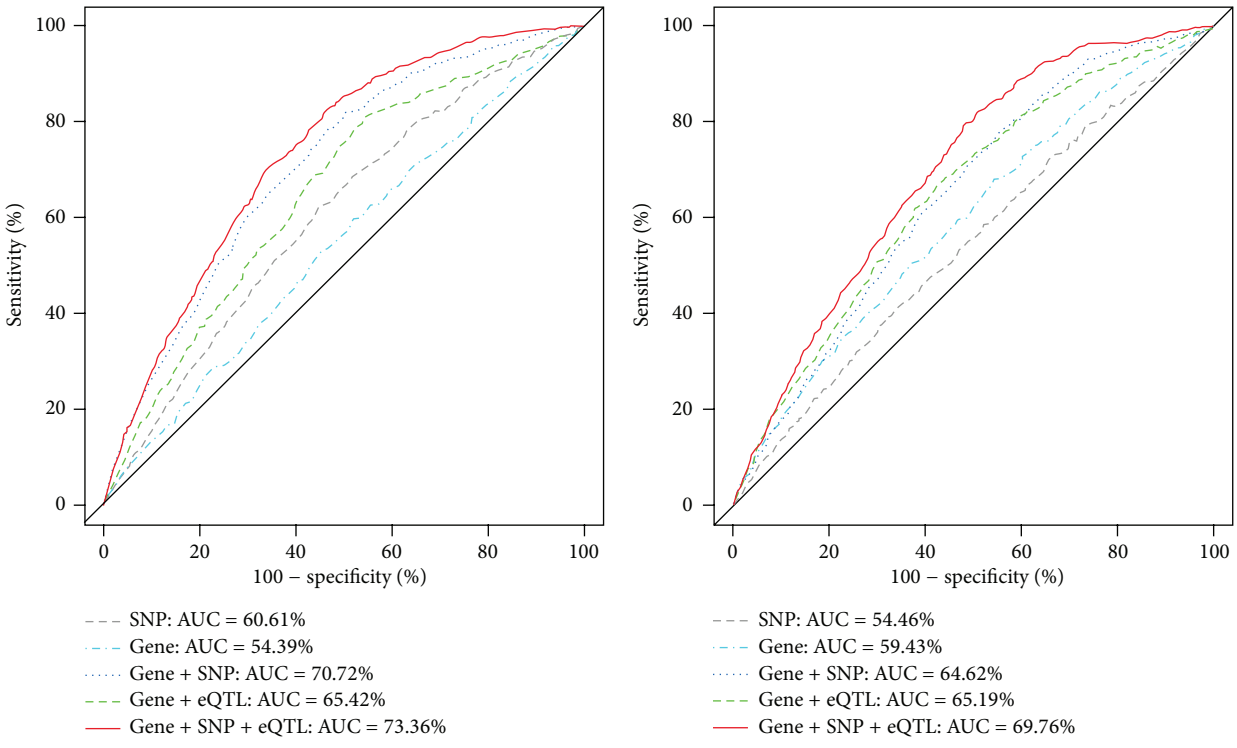
Our experiment results of gene prioritization are summarized in Figures 2, 3, and 4, where the performance was evaluated in terms of the receiver operating characteristic (ROC) curves and the area under the curve (AUC) estimates. In particular, Figure 3 shows the effects of integrating distinct genetic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



(a) Comparison of filtering methods (prostate cancer)

(b) Comparison of filtering methods (lung cancer)

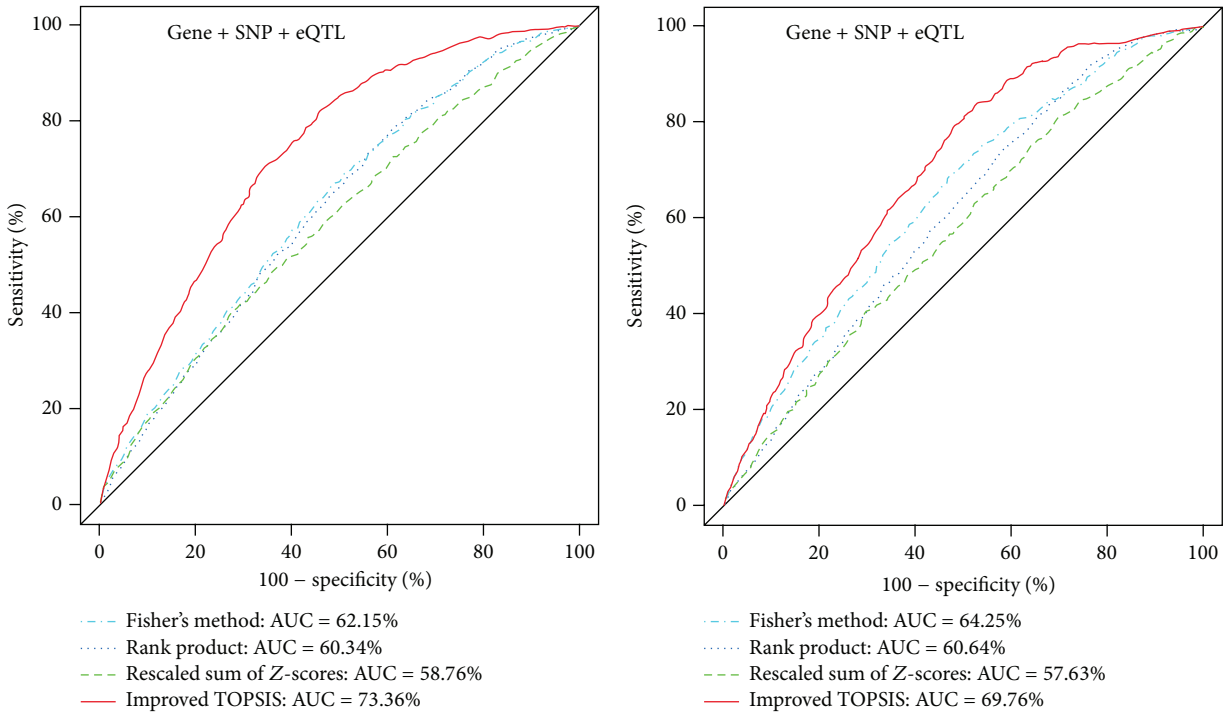
FIGURE 2: Comparison of the Max, TopQ, and Mean methods in filtering out duplicate genes: (a) prostate cancer results, (b) lung cancer results.



(a) Integration by improved TOPSIS (prostate cancer)

(b) Integration by improved TOPSIS (lung cancer)

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



(a) Comparison of integration methods (prostate cancer)

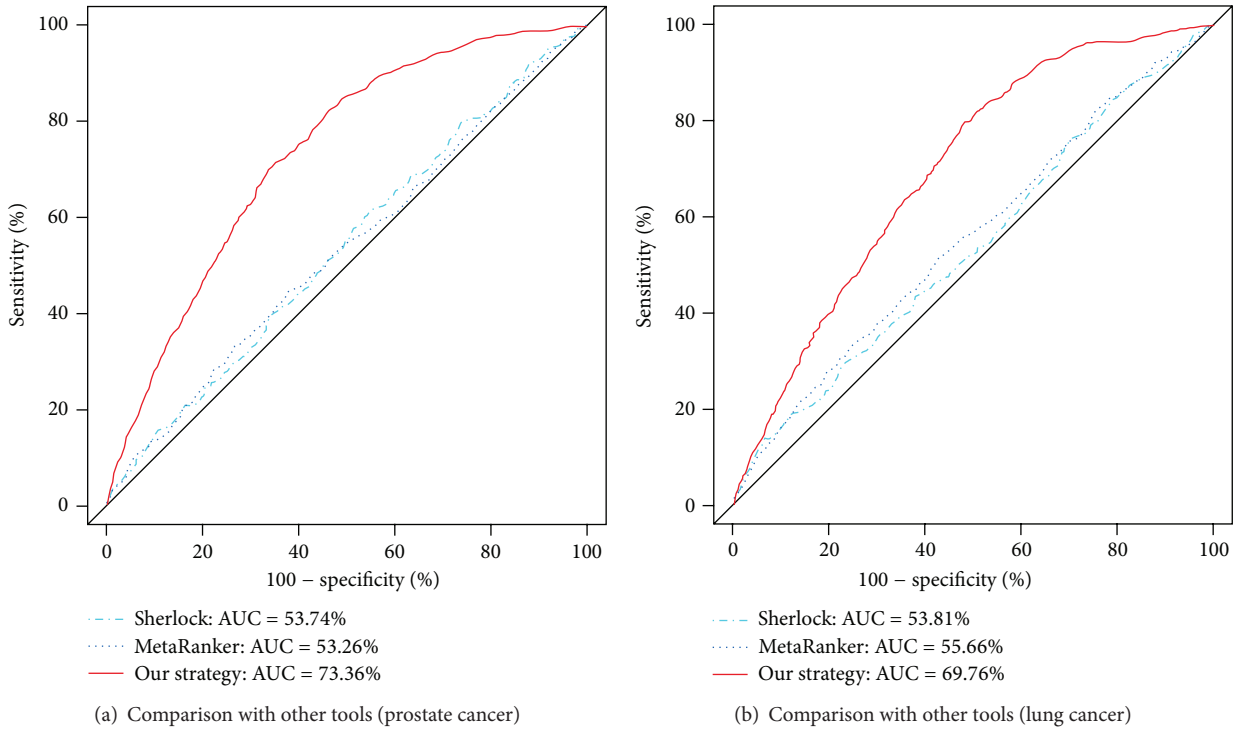
(b) Comparison of integration methods (lung cancer)

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



(a) Comparison with other tools (prostate cancer)

(b) Comparison with other tools (lung cancer)

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. Dworzyński, 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.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 707453, 9 pages
http://dx.doi.org/10.1155/2015/707453

Research Article

Probabilistic Inference of Biological Networks via Data Integration

Mark F. Rogers,[1] Colin Campbell,[1] and Yiming Ying[2]

[1]*Intelligent Systems Laboratory, University of Bristol, Merchant Venturers Building, Bristol BS8 1UB, UK*
[2]*College of Engineering, Mathematics and Physical Sciences, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QF, UK*

Correspondence should be addressed to Mark F. Rogers; mark.rogers@bristol.ac.uk

Received 27 August 2014; Accepted 5 November 2014

Academic Editor: Lei Chen

There is significant interest in inferring the structure of subcellular networks of interaction. Here we consider supervised interactive network inference in which a reference set of known network links and nonlinks is used to train a classifier for predicting new links. Many types of data are relevant to inferring functional links between genes, motivating the use of data integration. We use pairwise kernels to predict novel links, along with multiple kernel learning to integrate distinct sources of data into a decision function. We evaluate various pairwise kernels to establish which are most informative and compare individual kernel accuracies with accuracies for weighted combinations. By associating a probability measure with classifier predictions, we enable cautious classification, which can increase accuracy by restricting predictions to high-confidence instances, and data cleaning that can mitigate the influence of mislabeled training instances. Although one pairwise kernel (the tensor product pairwise kernel) appears to work best, different kernels may contribute complimentary information about interactions: experiments in *S. cerevisiae* (yeast) reveal that a weighted combination of pairwise kernels applied to different types of data yields the highest predictive accuracy. Combined with cautious classification and data cleaning, we can achieve predictive accuracies of up to 99.6%.

1. Introduction

There is a significant interest in determining subcellular network structures, from metabolic and protein-protein interaction networks, through to signalling pathways. Two broad interactive inference approaches are unsupervised and supervised network inference. With unsupervised inference, no prior knowledge of network linkage is assumed. Supervised inference is a more tractable alternative in which there is a training set of links and nonlinks, believed to be reliably known, and the task is to train a classifier using this information. We then make predictions for additional possible links where interactive network structure is less clearly resolved. One advantage of supervised inference is that there are a variety of pathways where the structure is fairly reliably determined and thus this prior structural knowledge could give a viable training set. A further advantage of supervised inference is that different types of data are informative about whether a functional link may exist, allowing practitioners to integrate data from diverse sources [1]. Furthermore we can weight these different data sources according to their relative significance. With unsupervised learning, it is much more difficult integrating different types of data into a predictive model, though various schemes have been suggested.

In this paper we will consider supervised network inference and we evaluate a variety of strategies to improve predictive performance. First we consider multiple kernel learning (MKL) in which different types of data are encoded into different pairwise base kernels. Using a weighted combination of base kernels, we construct a composite kernel that is used in a kernel-based classifier, for example, a Support Vector Machine (SVM) [2]. In Section 3 we show that this integrative approach gives better performance over a uniform weighting of kernels or classifiers constructed using only one type of data. Secondly, we discuss both established and a novel pairwise kernel for use with MKL. In this study we are interested in functional links between pairs of nodes in an interactive network, so the kernels we use encode similarity

between pairs. Our goal is to investigate which pairwise kernel is best and whether a variety of such pairwise kernels should be used in combination with MKL. Next we associate a probability measure with the predicted class assignment. This facilitates cautious classification and motivates a novel data cleaning method. We demonstrate dramatic improvements in accuracy via cautious classification, in which test accuracy is improved at the expense of making predictions for only a subset of possible links or nonlinks. This probability measure also motivates a method for data cleaning: we train a classifier incrementally and predict a new link-label prior to adding it to our training set. If, with a high confidence prediction, the predicted link-label disagrees with the actual label then this may indicate an outlier (a wrong link-label) and the datapoint should not be learnt. We investigate a method of incremental data cleaning for SVMs in which we sequentially add training data to the training set by selecting the next example closest to the current separating hyperplane: these are necessarily low confidence predictions and, by this means, we defer encounter with potential outliers toward the end of the sequential learning process. For the data set considered we show that this strategy leads to a small improvement in test accuracy.

2. Methods

2.1. Pairwise Kernels. *Kernels* [2, 3] encode the similarity of data objects and they can be constructed for a variety of different types of data, from continuously valued to sequence or graph information [2, 4]. For network inference, we will use a label $y_{i_1,i_2} = +1$ for a functional interaction between a pair of nodes (e.g., genes), labelled i_1 and i_2. $y_{i_1,i_2} = -1$ will label a noninteracting pair. Thus, with supervised inference, we have an adjacency matrix with components +1 and −1 and a number of unknown elements which we wish to estimate.

Our data is in the form $\mathbf{x}_i$ (where $i = 1, \ldots, m$). Linkage patterns in the data are classified in terms of pairings of nodes and appropriate kernels quantify a similarity between pairs. Thus, a comparison between a pair $(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$ and a further pair $(\mathbf{x}_{i_3}, \mathbf{x}_{i_4})$ could be performed through a comparison of $\mathbf{x}_{i_1}$ with $\mathbf{x}_{i_3}$ and $\mathbf{x}_{i_2}$ with $\mathbf{x}_{i_4}$ and, secondly, $\mathbf{x}_{i_1}$ with $\mathbf{x}_{i_4}$ and $\mathbf{x}_{i_2}$ with $\mathbf{x}_{i_3}$. If we write a general pairwise kernel as $\widehat{K}_P = \widehat{K}_P((\mathbf{x}_{i_1}, \mathbf{x}_{i_2}), (\mathbf{x}_{i_3}, \mathbf{x}_{i_4}))$ then an appropriate pairwise kernel would be

$$\widehat{K}_{P1} = K\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_3}\right) K\left(\mathbf{x}_{i_2}, \mathbf{x}_{i_4}\right) + K\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_4}\right) K\left(\mathbf{x}_{i_2}, \mathbf{x}_{i_3}\right). \quad (1)$$

Subsequently, we will use the loose convention that the arguments of the pairwise kernel can be data vectors, $\mathbf{x}_i$, or derived kernel matrices, $K(\mathbf{x}_i, \mathbf{x}_j)$. Ben-Hur and Noble [5] proposed kernel $\widehat{K}_{P1}$ and called it the tensor product pairwise kernel (TPPK). This pairwise kernel can be viewed as the weighted adjacency matrix of a Kronecker product graph of two graphs associated with the constituent kernels [6].

The second pairwise kernel we consider is [7]

$$\widehat{K}_{P2} = K\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_3}\right) + K\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_4}\right) + K\left(\mathbf{x}_{i_2}, \mathbf{x}_{i_3}\right) + K\left(\mathbf{x}_{i_2}, \mathbf{x}_{i_4}\right). \quad (2)$$

Assuming $K(\mathbf{x}_i, \mathbf{x}_j)$ is a positive semidefinite (PSD) kernel then the sum or the product of two such PSD kernels is also a PSD kernel, hence establishing $\widehat{K}_{P1}$ and $\widehat{K}_{P2}$ as allowable PSD kernels. Our third pairwise kernel is called the metric learning pairwise kernel (MLPK) [8]:

$$\widehat{K}_{P3} = \left[K(\mathbf{x}_{i_1}, \mathbf{x}_{i_3}) - K(\mathbf{x}_{i_1}, \mathbf{x}_{i_4}) - K(\mathbf{x}_{i_2}, \mathbf{x}_{i_3}) + K\left(\mathbf{x}_{i_2}, \mathbf{x}_{i_4}\right)\right]^2. \quad (3)$$

A kernel is a mapped inner product $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$; hence, $\widehat{K}_{P3}$ follows from

$$\widehat{K}_{P3} = \left[\left(\Phi(\mathbf{x}_{i_1}) - \Phi(\mathbf{x}_{i_2})\right)^T \left(\Phi(\mathbf{x}_{i_3}) - \Phi(\mathbf{x}_{i_4})\right)\right]^2. \quad (4)$$

Thus, for this kernel, the pair $(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$ is mapped to the vector $\Phi(\mathbf{x}_{i_1}) - \Phi(\mathbf{x}_{i_2})$ in feature space and the kernel is the inner product between these mapped vectors (subsequently squared). Extending this idea we can introduce a new kernel that is based on the inner product between the normalised pairs of vectors $\Phi(\mathbf{x}_{i_1}) - \Phi(\mathbf{x}_{i_2})$ and $\Phi(\mathbf{x}_{i_3}) - \Phi(\mathbf{x}_{i_4})$. This kernel is then based on the cosine similarity measure; that is,

$$\begin{aligned}\widehat{K}_{P4} = {} & \left(\Phi\left(\mathbf{x}_{i_1}\right) - \Phi\left(\mathbf{x}_{i_2}\right)\right)^T \left(\Phi\left(\mathbf{x}_{i_1}\right) - \Phi\left(\mathbf{x}_{i_2}\right)\right) \\ & \times \left[\sqrt{\left(\Phi\left(\mathbf{x}_{i_1}\right) - \Phi\left(\mathbf{x}_{i_2}\right)\right)^T \left(\Phi\left(\mathbf{x}_{i_3}\right) - \Phi\left(\mathbf{x}_{i_4}\right)\right)}\right. \\ & \left. \times \sqrt{\left(\Phi\left(\mathbf{x}_{i_3}\right) - \Phi\left(\mathbf{x}_{i_4}\right)\right)^T \left(\Phi\left(\mathbf{x}_{i_3}\right) - \Phi\left(\mathbf{x}_{i_4}\right)\right)}\right]^{-1},\end{aligned} \quad (5)$$

so

$$\begin{aligned}\widehat{K}_{P4} = {} & \left(K\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_3}\right) - K\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_4}\right) - K\left(\mathbf{x}_{i_2}, \mathbf{x}_{i_3}\right) + K\left(\mathbf{x}_{i_2}, \mathbf{x}_{i_4}\right)\right) \\ & \times \left[\sqrt{K\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_1}\right) - 2K\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}\right) + K\left(\mathbf{x}_{i_2}, \mathbf{x}_{i_2}\right)}\right. \\ & \left. \times \sqrt{K\left(\mathbf{x}_{i_3}, \mathbf{x}_{i_3}\right) - 2K\left(\mathbf{x}_{i_3}, \mathbf{x}_{i_4}\right) + K\left(\mathbf{x}_{i_4}, \mathbf{x}_{i_4}\right)}\right]^{-1}.\end{aligned} \quad (6)$$

For $\widehat{K}_{P1}$, we mentioned the relation between this pairwise kernel and a Kronecker product graph. This motivates consideration of other types of product graphs and one based on a Cartesian product graph (CSPK) has been proposed by [6]. This kernel is defined by

$$\begin{aligned}\widehat{K}_{P5} = {} & [K]_{i_1,i_3} I\left(i_2 = i_4\right) + [K]_{i_2,i_4} I\left(i_1 = i_3\right) \\ & + [K]_{i_1,i_4} I\left(i_2 = i_3\right) + [K]_{i_2,i_3} I\left(i_1 = i_4\right),\end{aligned} \quad (7)$$

where the (i, j)th component of a kernel matrix $[K]$ quantifies the similarity between the i'th and j'th nodes and where $I(\gamma)$ is an indicator function (1 if its argument is true and 0 otherwise). We include this kernel for completeness, since it will be included in our usage of MKL later. The information encapsulated in these product graphs can overlap substantially depending on the nature of the base kernels. The tensor

product $G \times G$ and the Cartesian product $G \square G$ of a graph $G(V_G, E_G)$ use the same vertex set, defined as a Cartesian product over the vertices in V_G ($\{(g,h) \mid g,h \in V_G\}$). However, their edge sets are defined as follows [9]:

$$E(G \times G) = \left\{(g,h)\left(g',h'\right) \mid gg' \in E_G,\ hh' \in E_G\right\},$$

$$E(G \square G) = \left\{(g,h)\left(g',h'\right) \mid g = g',\ hh' \in E_G, \text{ or } gg' \in E_G,\ h = h'\right\}. \tag{8}$$

A base kernel with nonzero diagonal elements corresponds to a graph with self-edges (i.e., $gg \in E_G$). In these cases a tensor product kernel will subsume a Cartesian product kernel over the same graph.

It is possible to further combine these types of pairwise kernels with other standard kernels, for example, Gaussian kernels or kernels based on polynomials; for example,

$$K = \left[K\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_3}\right) + K\left(\mathbf{x}_{i_2}, \mathbf{x}_{i_4}\right) + r\right]^d. \tag{9}$$

However, these types of kernels also require the use and determination of a *kernel parameter*, for example, r in (9), via a further cross-validation study, and so we will not consider them further in this study. There are further non-PSD (infinite) symmetric pairwise kernels which have been considered [7]. Though it is possible to project these to the cone of positive semidefinite kernels and use a proxy kernel [10], we investigated these and did not find consistently good performance, so they are not considered further in this study.

To give equal weight to different types of data we can further normalize the base kernels. Thus, viewing the kernel as a mapped inner product [2], we used the mapping $\mathbf{x} \rightarrow \Phi(\mathbf{x})/\|\Phi(\mathbf{x})\|_2$; then,

$$\widehat{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \frac{\Phi\left(\mathbf{x}_i\right) \cdot \Phi\left(\mathbf{x}_j\right)}{\sqrt{\Phi\left(\mathbf{x}_i\right) \cdot \Phi\left(\mathbf{x}_i\right)}\sqrt{\Phi\left(\mathbf{x}_j\right) \cdot \Phi\left(\mathbf{x}_j\right)}} = \frac{K\left(\mathbf{x}_i, \mathbf{x}_j\right)}{\sqrt{K\left(\mathbf{x}_i, \mathbf{x}_i\right) K\left(\mathbf{x}_j, \mathbf{x}_j\right)}}. \tag{10}$$

2.2. Multiple Kernel Learning. Different sources of data can be encoded into different types of *data kernel* [2], which we denote by $K(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$. Examples include diffusion kernels or standard kernels such as linear or Gaussian kernels [2] for encoding the similarity between data objects $\mathbf{x}_{i_1}$ and $\mathbf{x}_{i_2}$. These data kernels are, in turn, embedded in pairwise kernels, as described in the previous section. The resultant *pairwise kernels* will be denoted by $\widehat{K}_\ell(\mathbf{x}_i, \mathbf{x}_j)$ (where $\ell = 1, \ldots, p$) and are the base kernels used to construct a *composite kernel*, denoted by $\overline{K}$, for MKL learning. Two distinct base kernels may be different pairwise kernels representing the same source of data (i.e., the same data kernel) or they could be the same type of pairwise kernel applied to two different sources of data.

With *multiple kernel learning* [3, 11, 12], we can derive a composite kernel, $\overline{K}$, as a linear combination of these base kernels:

$$\overline{K}\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \mathbf{x}_{i_3}, \mathbf{x}_{i_4}\right) = \sum_{\ell=1}^{p} \lambda_\ell \widehat{K}_\ell\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \mathbf{x}_{i_3}, \mathbf{x}_{i_4}\right), \tag{11}$$

where λ_ℓ are the *kernel weights* that are restricted to lie on the simplex:

$$\sum_{\ell=1}^{p} \lambda_\ell = 1, \quad \lambda_\ell \geq 0. \tag{12}$$

The kernel weight λ_ℓ indicates the relative informativeness of data source ℓ. Aside from these weights, we must find the values of the learning parameters $\alpha_{i,j}$ during the training process. These learning parameters are the same learning parameters as for a standard Support Vector Machine [3]. However, in this case, rather than a single sample index, we use two indices, denoting the link between node i and j, since a data vector is attached to a link between two nodes and carries information about a possible interaction between these nodes. Here, we are interested in binary classification (link or nonlink) so $y_{i,j} = \pm 1$. Both $\alpha_{i,j}$ and λ_ℓ are found during the learning process through the following optimisation task:

$$\min_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}} \left\{ \sum_{i_1,i_2=1}^{m} \alpha_{i_1,i_2} - \frac{1}{2} \sum_{i_1,\ldots,i_4=1}^{m} \alpha_{i_1,i_2}\alpha_{i_3,i_4} y_{i_1,i_2} y_{i_3,i_4} \overline{K}\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \mathbf{x}_{i_3}, \mathbf{x}_{i_4}\right) \right\} \tag{13}$$

subject to

$$\sum_{i_1,i_2=1}^{m} \alpha_{i_1,i_2} y_{i_1,i_2} = 0, \quad 0 \leq \alpha_{i_1,i_2} \tag{14}$$

and the constraints in (12). This optimisation problem for MKL [3] can be tackled via quadratically constrained linear programming [13] and other methods [11, 12]. If $\{\alpha^\star_{i_1,i_2}, \lambda^\star_\ell\}$ is the solution to the optimisation problem in (13), then the predicted class label for novel input data, $\mathbf{z}_i$, is given by the sign of

$$\phi\left(\mathbf{z}_{i_1}, \mathbf{z}_{i_2}\right) = \sum_{j_1,j_2=1}^{m} \alpha^\star_{j_1,j_2} y_{j_1,j_2} \overline{K}\left(\mathbf{x}_{j_1}, \mathbf{x}_{j_2}, \mathbf{z}_{i_1}, \mathbf{z}_{i_2}\right) + b^\star, \tag{15}$$

where

$$b^\star = -\frac{1}{2}\left[\max_{\{i \mid y_i = -1\}}\left(\sum_{j=1}^{m} \alpha^\star_j y_j \overline{K}\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \mathbf{x}_{i_3}, \mathbf{x}_{i_4}\right)\right) + \min_{\{i \mid y_i = +1\}}\left(\sum_{j=1}^{m} \alpha^\star_j y_j \overline{K}\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \mathbf{x}_{i_3}, \mathbf{x}_{i_4}\right)\right)\right] \tag{16}$$

which is an adapted version of the decision function and bias, b, of a Support Vector Machine [3], appropriate to the context presented here.

Table 1: Kernel weights for the pairwise kernels used in this study. The weights selected for each kernel were those at the highest C-value that had two or more nonzero weights.

Kernel	Kernel weights for individual models					
	$\widehat{K}_M$	$\widehat{K}_P$	$\widehat{K}_S$	$\widehat{K}_{\mathrm{GI}}$	$\widehat{K}_{\mathrm{YH}}$	$\widehat{K}_{\mathrm{MS}}$
$\overline{K}_{P1}$	0	0.449	0.099	0.033	0	0.419
$\overline{K}_{P2}$	0.362	0.198	0	0.096	0.308	0.035
$\overline{K}_{P3}$	0	0.258	0	0.200	0	0.542
$\overline{K}_{P4}$	0	0.170	0.176	0.211	0.191	0.252
$\overline{K}_{P5}$	0	0.266	0	0	0	0.734

2.3. Introduction of a Probability Measure. In later experiments, we will introduce a confidence measure associated with linkage prediction. Most MKL methods have an intrinsic measure of confidence, namely, the margin measure $\phi(\mathbf{z})$ given in (15). The larger the absolute value of $\phi(\mathbf{z})$ the greater the degree of confidence in the predicted label. We can relate $\phi(\mathbf{z})$ to a probability measure by fitting a posterior probability distribution [14]. For binary classification, we use the sigmoid $p(y = +1 \mid \phi) = [1 + \exp(A\phi + B)]^{-1}$. With binary labels for link l, $y_l \in \{-1, 1\}$, we define $t_l = 0.5(y_l + 1) \in \{0, 1\}$. The parameters A and B are then found by minimizing the negative log likelihood of the training data via the cross entropy error function:

$$\min_{A,B} \left[-\sum_l t_l \log(p_l) + (1 - t_l) \log(1 - p_l) \right], \quad (17)$$

where p_l is the sigmoid probability function evaluated from $\phi(\mathbf{z})$ for the link considered. To minimize this function, we used the Levenberg-Marquardt algorithm [15].

3. Results

In this paper, we set out to investigate the following questions. Firstly, which pairwise kernel is the most accurate. As a second objective, we considered MKL and the gain to be made by using a weighted combination of different types of data over using a uniform combination. Combined with our first objective, a further objective was to understand if one type of pairwise kernel is the best or if higher accuracy is achieved by using a weighted combination of pairwise kernels. Our results are reported in Section 3.1. We then place a probability measure on $\phi(\mathbf{z}_{i_1}, \mathbf{z}_{i_2})$ in (15) and briefly consider prediction restricted to high confidence inference (Section 3.2) and strategies for removing possibly wrongly labelled datapoints in the training data (Section 3.3).

3.1. Multiple Kernel Learning. For our analysis, we used kernels from six heterogeneous data sets that have been used for supervised interactive network inference in a previous study [1]: three based on protein sequence kernels and three based on diffusion kernels. Borrowing notation from these authors, we used three data kernels based on sets of amino acid sequences (spectrum (K_S) [4], motif (K_M) [16], and Pfam (K_P)) [17] and three diffusion data kernels based on interaction networks from the BioGRID database [18] (yeast two-hybrid assay (K_{YH}), genetic interactions (K_{GI}), and affinity capture-MS (K_{MS})) [1].

In their original study, Qiu and Noble [1] used a uniformly weighted combination of kernels: the average value of the three sequence kernels was added to the average of the three diffusion kernels (we omit using their RBF kernels, given the latter contain a kernel parameter). A tensor product pairwise kernel (TPPK or $P1$ in our classification) was applied as follows:

$$\overline{K} = \widehat{K}_{P1} \left(\frac{K_M + K_S + K_P}{3} + \frac{K_{\mathrm{YH}} + K_{\mathrm{GI}} + K_{\mathrm{MS}}}{3} \right). \quad (18)$$

Here, we use MKL to assign weights according to the contribution of each data source for predicting edges in a gene interaction network. Since uniform weighting is a subinstance of using variable kernel weights, MKL will inevitably improve on (or equal) a uniform weighting scheme. The data we are using provides information on individual proteins, rather than protein pairs, and hence we use pairwise kernels, as outlined above. Since we have kernel weights λ_ℓ and sequence or diffusion kernels K_ℓ, for a given pairwise kernel, $\widehat{K}_P$, our composite kernel after MKL training will be

$$\overline{K}_{\mathscr{P}} = \sum_{\ell=1}^{p} \lambda_\ell \widehat{K}_P(K_\ell). \quad (19)$$

We used the simple MKL Matlab package [19]. Training is compute-intensive, even with an efficient implementation, so we learned the kernel weights using relatively small sets of 1,000 to 4,000 examples. We found that the kernel weights for data sets larger than 4,000 examples were barely altered, so we did not use larger data sets for this purpose. The learnt weights for each individual pairwise kernel appear in Table 1. Of the three sequence data kernels, the Pfam kernel (K_P) achieves the highest weight for the TPPK kernel $\overline{K}_{P1}$. By contrast, the motif kernel (K_M) was assigned zero weight in all cases but $\overline{K}_{P2}$. There is a greater difference in the way these pairwise kernels apply information from the diffusion kernels. The TPPK ($\overline{K}_{P1}$) and CSPK ($\overline{K}_{P5}$) kernels rely almost entirely on the affinity capture-MS data, while the $\overline{K}_{P2}$ and $\overline{K}_{P4}$ kernels are able to leverage information from the yeast two-hybrid assay and gene interaction data as well. No pairwise kernel uses more than five of the component data kernels. The $\overline{K}_{P1}$ kernel weights exhibit the highest variation, while the $\overline{K}_{P4}$ kernel has a more even distribution of weights. Once the MKL algorithm had learned the weights,

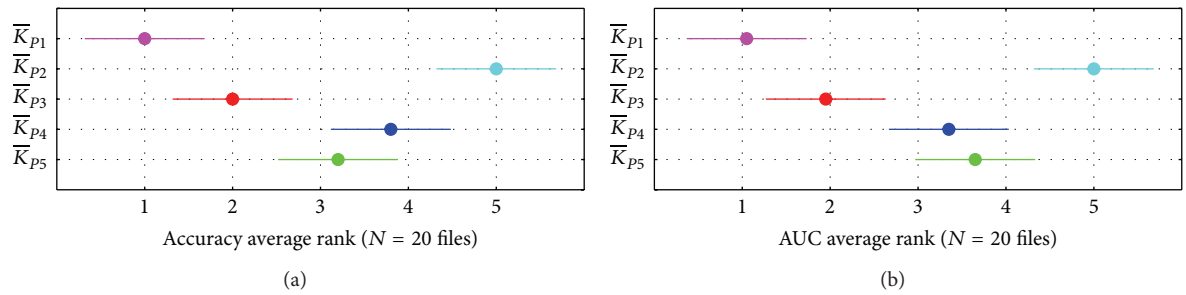


Figure 1: Comparison of average rankings for accuracy (a) and AUC (b) for 20 small data sets using unweighted pairwise kernels. The dot for each kernel identifies its mean rank; horizontal bars depict the Nemenyi test critical region for $\alpha = 0.05$. The tensor product kernel ($\overline{K}_{P1}$) consistently had the highest ranking while the symmetric direct sum kernel ($\overline{K}_{P2}$) had the lowest. The differences between the remaining three kernels become clearer when we consider AUC as well as accuracy: the metric learning ($\overline{K}_{P3}$) kernel has higher rankings than the other two on both measures.

we recomputed the kernels as described in (19) and compared the kernels' performance.

The *S. cerevisiae* data from [1] form a balanced set consisting of 10,980 positive and 10,980 negative pairs of interacting genes (21,960 total pairs). Given this relatively large data set, we wished to see how well each kernel would perform when trained on subsets of different size. Thus, we ran three different experiments on these data. To assess performance on small data sets, we split the original set into 20 subsets of 1,098 examples each, randomly assigning an equal number of positive and negative examples to each subset. We ran 5-fold cross-validation to obtain average accuracy and AUC (area under the ROC curve) values for each kernel on each subset. Following the recommendations in [20] for comparing multiple classifiers on multiple data sets, we ranked the kernels for each data set and used nonparametric tests to assess differences between the kernels. We used the Friedman test to determine the significance of differences between all five kernels and then used the post hoc Nemenyi test to assess pairwise differences [12, 20]. To evaluate the kernels' performance on medium and large data sets, we used the same procedure, splitting the original data set into 10 subsets of 2,196 examples (1,757 training/439 test per fold) or 5 subsets of 4,392 examples (3,514 training/878 test per fold).

We expect this experimental design to yield realistic results for the data used in our study [21], but to extend this work to general-purpose classifiers, we recommend separating test data into separate classes as outlined in [22].

3.1.1. Comparison of Different Pairwise Kernels. For small data sets, the tensor product kernel ($\overline{K}_{P1}$) consistently yields the highest accuracy ranking of any pairwise kernel (mean 1.0) while the symmetric direct sum kernel ($\overline{K}_{P2}$) consistently yields the lowest (Figure 1). The metric learning ($\overline{K}_{P3}$), cosine-like ($\overline{K}_{P4}$), and Cartesian graph product ($\overline{K}_{P5}$) pairwise kernels yield intermediate rankings, though the $\overline{K}_{P3}$ kernel (mean 2.0) was consistently ranked higher than the other two. When we rank the kernels based on AUC score as well as accuracy, we again see that the $\overline{K}_{P3}$ kernel yields higher performance than $\overline{K}_{P4}$ or $\overline{K}_{P5}$, but here the $\overline{K}_{P4}$ ranking is higher than that for $\overline{K}_{P5}$, making it difficult to identify a clear winner between them. The $\overline{K}_{P1}$ kernel's high accuracy and AUC rankings are statistically significant ($\alpha = 0.01$) when compared to all but the $\overline{K}_{P3}$ kernels, but the differences between $\overline{K}_{P1}$ and $\overline{K}_{P3}$ are not statistically significant at $\alpha = 0.05$. Results for medium and large data sets (not shown) are nearly identical, but the smaller data size yields less statistical power.

3.1.2. Performance of Individual Pairwise Kernels with Multiple Types of Input Data. We compared the performance of each individual pairwise kernel with and without MKL weights using the same cross-validation procedure outlined above. To determine whether MKL yields significant improvements for any of the kernels, we use a Wilcoxon signed rank test for $N = 10$ and $N = 20$ files and a paired t-test for $N = 5$ data files (there are no critical values for the Wilcoxon test for $\alpha \leq 0.05$ and $N = 5$). Table 2 shows the relative performance of the weighted and averaged kernels. In many cases we find a statistically significant increase in performance if we use weighted kernels (weighted over the 6 constituent kernels); even if the difference is not significant, it is rare that weighted kernels limit performance. In particular, the weighted version of the $\overline{K}_{P3}$ kernel exhibits significantly higher accuracy than the unweighted version in all of our experiments. On large training sets, we see a significant improvement with the weighted versions of the $\overline{K}_{P2}$, $\overline{K}_{P3}$, and $\overline{K}_{P4}$ kernels: increases in accuracy range from 2.2% to 3.6%. We note that the weighted version of the $\overline{K}_{P5}$ kernel yields slightly lower accuracy on average than the unweighted version, but these differences are not statistically significant.

Secondly, we compared the relative performance of these composite MKL kernels with their corresponding base kernels. We ran the same experiment outlined above on the individual base kernels. In general, we see a significant difference between the MKL-weighted kernels and their individual base kernels. For example, the top-performing combined kernel $\overline{K}_{P1}$ yields accuracy that is at least 4% higher than the nearest corresponding base kernel (Figure 2). We note that the weights used for the constituent kernels roughly track the relative performance of the kernels: for example,

Table 2: Cross-validation results for the pairwise kernels using unweighted (U) and weighted (W) combinations of the six unpaired kernels for data sets of different sizes. Shown is test accuracy averaged over $N = 20$, $N = 10$, or $N = 5$ data sets (1,098, 2,196, or 4,392 examples, respectively, split into 80% training and 20% test sets). In many cases, the MKL weights yield a significant improvement while in other cases there is no significant change. Significant values are denoted as follows: **Wilcoxon signed rank $\alpha = 0.01$ or $^{*}\alpha = 0.05$, and †paired t-test $\alpha < 0.01$. Statistically significant values are marked in bold type.

Kernel	$N = 20$		$N = 10$		$N = 5$	
	U	W	U	W	U	W
$\overline{K}_{P1}$	0.826	**0.836***	0.860	0.867	0.895	0.901
$\overline{K}_{P2}$	0.667	0.662	0.663	**0.681****	0.694	**0.716**†
$\overline{K}_{P3}$	0.764	**0.801****	0.802	**0.837****	0.852	**0.883**†
$\overline{K}_{P4}$	0.731	0.740	0.756	0.764	0.755	**0.791**†
$\overline{K}_{P5}$	0.764	0.759	0.817	0.807	0.862	0.849

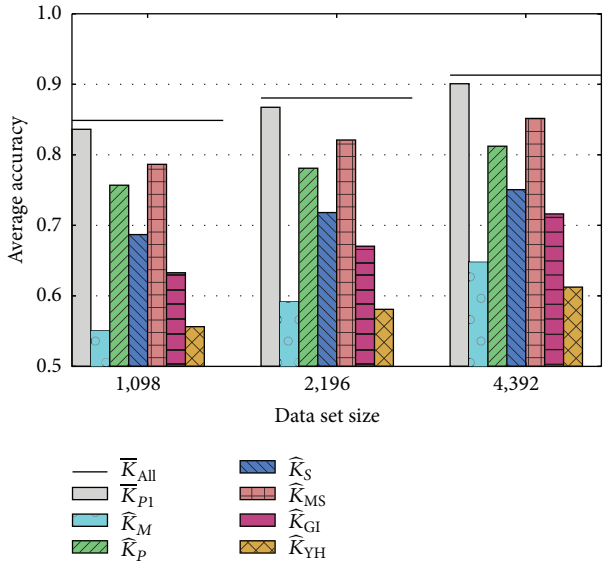


Figure 2: Graphical depiction showing the typical improvement in accuracy we see when using a weighted sum of base kernels via MKL. Here, we compare the average performance of the best-performing composite kernel, $\overline{K}_{P1}$ (solid grey bars), with the corresponding base kernels (hashed bars) on data sets of three different sizes. By leveraging information from multiple kernels, $\overline{K}_{P1}$ provides an accuracy increase of 4% to 5% over the best of the base kernels. When we use MKL over all 30 base kernels combined ($\overline{K}_{\text{All}}$), we achieve a further 1.2% to 1.4% increase (black bars). Differences between $\overline{K}_{P1}$ and its base kernels are significant at $\alpha < 0.001$; differences between $\overline{K}_{\text{All}}$ and $\overline{K}_{P1}$ are significant at $\alpha < 0.01$.

K_P and K_{MS} yield the highest accuracy and also have the largest weights for $\overline{K}_{P1}$ (see Table 1), while the two weakest base kernels, K_M and K_{YH}, have zero weights and do not contribute to the final composite kernel.

3.1.3. Performance Using All Pairwise Kernels and All Types of Input Data. Next we use MKL with all five pairwise kernels and all six different types of input data to produce a comprehensive kernel, $\overline{K}_{\text{All}}$. This gave 30 possible kernels but only 11 of these have nonzero kernel weights (Table 3). Notably, the tensor product kernel ($\overline{K}_{P1}$) and the metric learning kernel ($\overline{K}_{P3}$) contribute 4 and 3 base kernels, respectively. None of the motif base kernels (K_M) are included, nor are any of the Cartesian product base kernels ($\overline{K}_{P5}$). The resulting $\overline{K}_{\text{All}}$ kernel yields accuracy that is 1.2% to 1.4% higher than the best individual pairwise kernel (horizontal lines in Figure 2). For all data set sizes tested, this difference is statistically significant. The kernel weights and the improved performance both indicate that there is complimentary information provided by the different pairwise kernels. By contrast, the closely related Cartesian product kernels and tensor product kernels likely yield redundant information (Section 2.1), resulting in zero weights for Cartesian product base kernels.

3.2. Cautious Classification. We now introduce the probability measure considered in Section 2.3. A confidence measure is of interest in its own right. However, our interest here is in its use to further improve test accuracy for the pairwise-kernel based MKL scheme already introduced. Specifically, we consider *cautious classification* in which we decline to make predictions if the confidence is sufficiently low but make predictions of a link or nonlink in high confidence instances. For the *S. cerevisiae* data set, we show that this strategy can yield significant improvements in test accuracy, though at the cost of a reduced set of predictions.

In Figure 3 we plot the test accuracy (as a fraction) versus the p-value cutoff (a) when using all the above mentioned pairwise and data kernels. The test accuracy increased up to 0.996 as we increased the p-value cutoff, while the number of points predicted dropped to 246 (11.2%). If we used individual pairwise kernels with all the available data (we illustrate with $\overline{K}_{P1}$ in this figure), then the test accuracy was lower (0.86 to 0.97 for $\overline{K}_{P1}$), but, as illustrated, we also noticed a greater sensitivity to outliers (incorrect link-labels) for high values of the p-value cutoff. These numerical simulations are for $m = 2,196$ and so they correspond to the weighted values for $N = 10$ in Table 2 when the cutoff is $p = 0.50$.

3.3. Data Cleaning. To address the impact of outliers on our classifiers, we investigated two data cleaning methods. In each method, our goal was to train an SVM using as many informative examples as possible while eliminating counter-productive examples (outliers). In both cases, we initiated training with a small subset of reliably labelled datapoints, where the *label* of link (positive) or nonlink (negative) is

Table 3: Kernel weights learned for a comprehensive kernel, $\overline{K}_{\text{All}}$, that combines all base pairwise kernels. For each pairwise kernel, we show the final weight assigned to each of its base kernels. The tensor product kernel ($\overline{K}_{P1}$) and the metric learning kernel ($\overline{K}_{P3}$) contribute the most information to this comprehensive kernel. None of the motif base kernels (K_M) contribute, nor do any of the Cartesian product base kernels ($\overline{K}_{P5}$). The kernel weights sum to unity.

Kernel	Kernel weights for combined model					
	K_M	K_P	K_S	K_{GI}	K_{YH}	K_{MS}
$\overline{K}_{P1}$	0	0.193	0	0.103	0.075	0.372
$\overline{K}_{P2}$	0	0.002	0	0.010	0	0
$\overline{K}_{P3}$	0	0.044	0	0.023	0	0.153
$\overline{K}_{P4}$	0	0.006	0	0.019	0	0
$\overline{K}_{P5}$	0	0	0	0	0	0

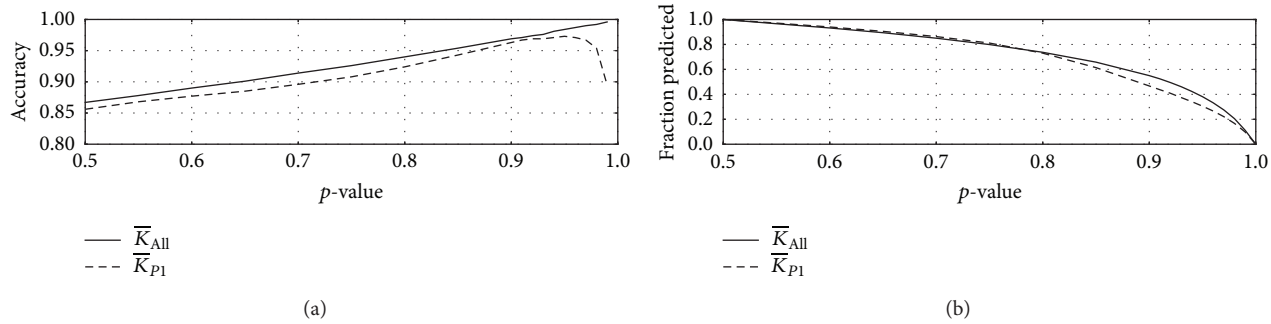


Figure 3: Plot of the test accuracy ((a) y-axis) and fraction of pairs predicted ((b) y-axis) as a function of the p-value cutoff (x-axis) for (i) using all available pairwise and data kernels ($\overline{K}_{\text{All}}$, solid curve) and (ii) the top-performing pairwise kernel ($\overline{K}_{P1}$, dashed curve). By increasing the p-value cutoff, we increase the accuracy in our predictions but decrease the fraction of pairs for which we can make predictions.

known. To obtain reliable representatives from both positive and negative example classes, we estimated the centroids of each class and chose the 10 datapoints in each class that were closest to their centroids (alternatively, biological insight may give a reliable starting set). We then learnt the remaining datapoints sequentially and avoided potential outliers using one of two strategies. Our first approach, introduced by [3], is to predict the labels for all currently unlearnt links in the training data and use the datapoint with the lowest associated confidence for training in the next iteration. This procedure tends to postpone learning potential outliers to the end of the learning process but incurs a high computational cost as it makes predictions for all unlearnt links at each iteration. A second and less computationally costly approach is to select the next training example randomly at each iteration and predict its label using the current classifier. If the prediction is high confidence but the actual label is of opposite sign, we omit the datapoint since it may be an outlier.

For the data set considered [1], there appear to be few anomalous links in the data, so there is at most a small gain in test accuracy when we use these methods. In Figure 4, we give the test error achieved on held-out data, averaged over 10 distinct data sets from the experiments described in Section 3.1. In this case, we are making predictions of link-labels over all currently unlearnt datapoints and learning that datapoint with the lowest associated confidence for the link-label. The learning curve has a shallow minimum of the test error with a fractional test error of 0.1380 at $m = 1563$, against a final test error of 0.1490 at $m = 2000$, having learnt all the data in the training set. Of course, we can also lessen the influence of outliers by using an L_1 or L_2 soft margin with a margin-based classifier [2, 3]. However, when using a soft margin, we need to pursue a validation study, using some held-out data, to establish the most appropriate value for the soft margin parameter. With the proposed data cleaning method, there is no need to use validation data since there is a suitable stopping criterion available. Specifically, we can stop learning new datapoints when the equivalent of the margin band is empty [3], that is, when $|\phi(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})| > 1$ in (15). At this point, we would be learning two types of link-labels. Either we learn a link-label of the expected sign, that is, the predicted link-label and actual label agree, or the predicted link-label and actual label disagree. If the predicted and actual link-labels agree then this potential link is the equivalent of a non-support vector, with $\alpha^*_{i_1,i_2} = 0$, and so it will not contribute to the decision function stated in (15). We therefore do not need to learn this datapoint. Alternatively, the new link will have a label that is substantially out-of-alignment with the current hypothesis (after having learnt a number of link-labels). With $|\phi(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})| > 1$, it is being placed within the data space of the oppositely labelled datapoints. Such a link could be correct, but it does have a strong possibility of being an outlier. We would not stop before the margin band is empty because the newly learnt datapoints will have $\alpha^*_{i_1,i_2} > 0$ and thus will contribute to the decision criterion stated in (15). This stopping criterion gave a termination point that is within 0.1% of the empirically observed minimum error, with cessation of learning after 1,642 samples, with a test error of 0.1323, as

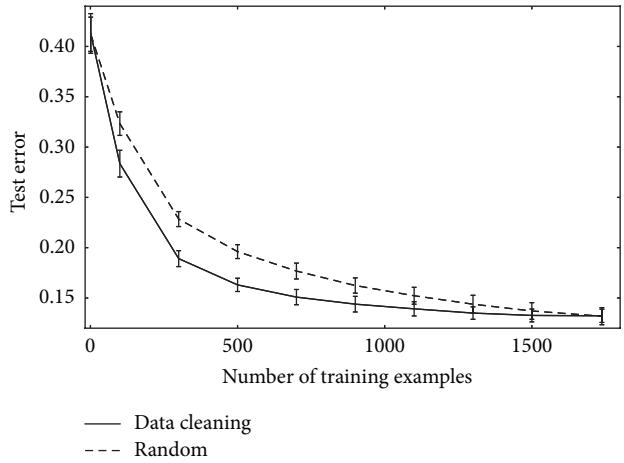


FIGURE 4: Mean test error as a fraction (y-axis) versus the number of patterns learnt (x-axis) for the top-performing pairwise kernel, $\overline{K}_{P1}$. Error bars depict a 95% confidence interval for 5-fold cross-validation test error averaged over 10 distinct data subsets, each with $m = 2,196$. The upper curve gives the performance if we learn all the data sequentially (from a common start set) in random order. The lower curve gives the test accuracy if the next addition to the training set is chosen based on having the lowest confidence predicted link-label.

against the observed minimum test error of 0.1319 at 1,565 samples learnt. Beyond this stopping point, the test error can rise as we may start learning links (or nonlinks) which are anomalously labeled.

An additional advantage of using this sequential learning method is that the prospects of achieving convergence with a linear kernel are enhanced. Specifically, a mislabelled datapoint can appear as a wrongly labelled datapoint within a cluster of datapoints of the opposite sign. This would mean the two classes of data can become nonseparable, requiring the use of a nonlinear kernel (e.g., an RBF kernel), with an associated validation study to find the appropriate value of the kernel parameter.

4. Conclusion

In this paper, we have investigated supervised interactive network inference using multiple kernel learning. Our objective was to consider ways to improve prediction performance and there are five main conclusions drawn from our study. Firstly, we compared five different types of pairwise kernel, which did not require adjustment of a kernel parameter, on six different types of data for supervised network inference. Our conclusion was that the pairwise kernel $P1$ (TPPK) worked best. Next, we considered whether use of a weighted combination of kernels (data sources) performed better than a uniformly weighted combination (Table 2) and, as expected, we found this was the case. Thirdly, for each pairwise kernel, we established performance using MKL over these six different data kernels and then compared this with the performance of MKL, when using all five different types of pairwise kernel and taken over all six different types of data; that is, the algorithm could use a weighted combination of 30 different types of kernel. At a statistically significant level, we found that this 30-base kernel combination outperformed the best of the individual pairwise kernels taken in isolation by between 1.2 and 1.4 percentage points. Thus, TPPK may look like the most effective pairwise kernel, but there must be complementary information among these different types of pairwise kernels and they are best used in combination with kernel-selection being made by the algorithm. To further improve predictive test accuracy, we next introduced a confidence measure associated with the class assignment. We showed that there are significant gains from using cautious classification, where prediction is confined to a high confidence instance. Our fifth study was to investigate the use of this probability measure with data cleaning. The *S. cerevisiae* data set considered appears clean, with only a few link-labels suggested as being possibly mislabelings. Thus, this strategy only gave a gain of 1.7% in our study in Section 3.3. However, label noise may be a more substantial problem in the understanding of pathways in more advanced organisms. This strategy would therefore likely yield better gains in these contexts.

In short, each component strategy has delivered modest through to more substantive improvements in predictive accuracy. Taken together, though, they lead to a substantial improvement in predictive accuracy over previous studies [1] and a highly accurate predictor.

As a consequence of this investigation, we have identified several potentially fruitful avenues for future work. We selected the SimpleMKL method for its speed and relatively sparse kernel weights, but other weighting methods conceivably could provide better performance [12, 23]. Further, recently proposed methods for predicting protein interactions such as coevolutionary divergence [24] and remote homology [25] could be used to extend our model. Finally, we have enumerated several approaches to data cleaning that could become increasingly effective as novel data sets become available.

Conflict of Interests

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

Acknowledgment

The authors acknowledge the support of EPSRC Grant EP/K008250/1.

References

[1] J. Qiu and W. S. Noble, "Predicting co-complexed protein pairs from heterogeneous data," *PLoS Computational Biology*, vol. 4, no. 4, Article ID e1000054, 2008.

[2] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[3] C. Campbell and Y. Ying, *Learning with Support Vector Machines*, Morgan & Claypool, 2011.

[4] C. S. Leslie, E. Eskin, and W. S. Noble, "The spectrum kernel: a string kernel for SVM protein classification," in *Proceedings of the Pacific Symposium on Biocomputing*, vol. 7, pp. 564–575, World Scientific, 2002.

[5] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics*, vol. 21, supplement 1, pp. i38–i46, 2005.

[6] H. Kashima, S. Oyama, Y. Yamanishi, and K. Tsuda, "Cartesian kernel: an efficient alternative to the pairwise kernel," *IEICE Transactions on Information and Systems*, vol. E93-D, no. 10, pp. 2672–2679, 2010.

[7] C. Brunner, A. Fischer, K. Luig, and T. Thies, "Pairwise support vector machines and their application to large scale problems," *Journal of Machine Learning Research*, vol. 13, pp. 2279–2292, 2012.

[8] J.-P. Vert, J. Qiu, and W. S. Noble, "A new pairwise kernel for biological network inference with support vector machines," *BMC Bioinformatics*, vol. 8, no. 10, article S8, 2007.

[9] R. Hammack, I. Wilfried, and S. Klavzar, *Handbook of Product Graphs*, Taylor & Francis, Boca Raton, Fla, USA, 2011.

[10] Y. Ying, C. Campbell, and M. Girolami, "Analysis of svm with indefinite kernels," in *Advances in Neural Information Processing Systems*, pp. 2205–2213, 2009.

[11] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2003/04.

[12] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[13] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, pp. 41–48, Alberta, Canada, July 2004.

[14] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularised likeli-hood methods," in *Advances in Large Margin Classifiers*, pp. 61–74, MIT Press, 1999.

[15] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, NY, USA, 2000.

[16] A. Ben-Hur and D. Brutlag, "Remote homology detection: a motif based approach," *Bioinformatics*, vol. 19, no. 1, pp. i26–i33, 2003.

[17] S. M. Gomez, W. S. Noble, and A. Rzhetsky, "Learning to predict protein-protein interactions from protein sequences," *Bioinformatics*, vol. 19, no. 15, pp. 1875–1881, 2003.

[18] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Research*, vol. 34, supplement 1, pp. D535–D539, 2006.

[19] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[20] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[21] Y. Park and E. M. Marcotte, "Revisiting the negative example sampling problem for predicting protein-protein interactions," *Bioinformatics*, vol. 27, no. 21, Article ID btr514, pp. 3024–3028, 2011.

[22] Y. Park and E. M. Marcotte, "Flaws in evaluation schemes for pair-input computational predictions," *Nature Methods*, vol. 9, no. 12, pp. 1134–1136, 2012.

[23] Y. Ying, C. Campbell, T. Damoulas, and M. Girolami, "Class prediction from disparate biological data sources using an iterative multi-Kernel algorithm," in *Pattern Recognition in Bioinformatics*, vol. 5780 of *Lecture Notes in Computer Science*, pp. 427–438, Springer, Berlin, Germany, 2009.

[24] C. Hsin Liu, K.-C. Li, and S. Yuan, "Human protein-protein interaction prediction by a novel sequence-based co-evolution method: co-evolutionary divergence," *Bioinformatics (Oxford, England)*, vol. 29, no. 1, pp. 92–98, 2013.

[25] B. Liu, D. Zhang, R. Xu et al., "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 501262, 15 pages
http://dx.doi.org/10.1155/2015/501262

Research Article

Identification of Subtype Specific miRNA-mRNA Functional Regulatory Modules in Matched miRNA-mRNA Expression Data: Multiple Myeloma as a Case

Yunpeng Zhang,[1] Wei Liu,[1,2] Yanjun Xu,[1] Chunquan Li,[1] Yingying Wang,[1] Haixiu Yang,[1] Chunlong Zhang,[1] Fei Su,[1] Yixue Li,[1] and Xia Li[1]

[1] *College of Bioinformatics Science and Technology, Harbin Medical University, 194 Xuefu Road, Harbin 150081, China*
[2] *Department of Mathematics, Heilongjiang Institute of Technology, Harbin 150050, China*

Correspondence should be addressed to Yixue Li; yxli@sibs.ac.cn and Xia Li; lixia@ems.hrbmu.edu.cn

Received 17 July 2014; Revised 19 October 2014; Accepted 27 October 2014

Academic Editor: Lei Chen

Identification of miRNA-mRNA modules is an important step to elucidate their combinatorial effect on the pathogenesis and mechanisms underlying complex diseases. Current identification methods primarily are based upon miRNA-target information and matched miRNA and mRNA expression profiles. However, for heterogeneous diseases, the miRNA-mRNA regulatory mechanisms may differ between subtypes, leading to differences in clinical behavior. In order to explore the pathogenesis of each subtype, it is important to identify subtype specific miRNA-mRNA modules. In this study, we integrated the Ping-Pong algorithm and multiobjective genetic algorithm to identify subtype specific miRNA-mRNA functional regulatory modules (MFRMs) through integrative analysis of three biological data sets: GO biological processes, miRNA target information, and matched miRNA and mRNA expression data. We applied our method on a heterogeneous disease, multiple myeloma (MM), to identify MM subtype specific MFRMs. The constructed miRNA-mRNA regulatory networks provide modular outlook at subtype specific miRNA-mRNA interactions. Furthermore, clustering analysis demonstrated that heterogeneous MFRMs were able to separate corresponding MM subtypes. These subtype specific MFRMs may aid in the further elucidation of the pathogenesis of each subtype and may serve to guide MM subtype diagnosis and treatment.

1. Introduction

MicroRNAs (miRNAs) are a class of short, noncoding RNAs of ~22 nucleotides RNA molecules that play important roles in gene regulation during physiological or disease-associated processes [1]. By regulating gene expression, miRNAs are involved in most biological processes, such as cell cycle regulation, development, apoptosis, stress response, and tumourigenesis [2, 3]. Accordingly, miRNA alterations may contribute to many human diseases [4]. In fact, deregulated miRNA expression has been observed in various cancer types, such as multiple myeloma (MM) [5, 6]. miRNAs can act as both tumor suppressors and oncogenes, depending on the context and target genes [7, 8]. The regulatory mechanisms underlying miRNAs and their target mRNAs remain unclear: a single miRNA is capable of regulating >200 mRNAs, and a single mRNA may be regulated by multiple miRNAs [9]. Some studies have shown that miRNAs may not primarily act by repressing a few cancer-related genes but by disturbing a regulatory network in which these cancer-related genes play crucial functional roles [10, 11]. Thus, identification of context-dependent miRNA-mRNA modules is an important step to elucidate their synergistic effect on the pathogenesis of complex diseases.

Several computational methods have been previously developed for the discovery of miRNA-mRNA modules [12–18]. Early efforts primarily focused on computational predicted miRNA-mRNA pairs and detection of miRNA regulatory modules at the sequence level [16]. However, miRNA and mRNA expression were not taken into consideration. MiRNAs that are regulatory in one experimental scenario may not be regulatory in another [12]; expression

information is essential for the identification of biologically meaningful miRNA-mRNA modules. Recently, integrated analysis of both sequence information and expression profiles of miRNAs and mRNAs was proposed to identify functional miRNA-mRNA regulatory modules [12–15, 19, 20]. Joung et al. [13, 14] discovered miRNA-mRNA modules using a combination of putative miRNA-mRNA pairs and expression data; however, correlations between the expression of miRNAs and mRNAs were not considered. Liu et al. [15] identified modules in two steps: (i) discovering the putative networks given the target information of miRNAs and mRNAs and (ii) deriving functional miRNA-mRNA regulatory modules on expression data given the putative networks. Considering that the computational predicted miRNA targets exhibit a high false positive discovery rate and that the targeting relationship between miRNAs and genes is far from complete, the first step that is based on target information exhibits an innate defect concerning the identification of modules. Jayaswal et al. [12] proposed an improved method: first, identification of miRNA and mRNA clusters using both target information and expression data; and second, estimation of the association between the two types of clusters to select potential regulatory miRNA-mRNA modules with statistically significant associations. However, this method was based on expression correlations under all available conditions rather than a subset of conditions, and the procedures for identification of miRNA and mRNA clusters were separated:; thus, it was limited to identification of miRNA-mRNA functional modules under the same specific conditions.

For heterogeneous diseases, the miRNA-mRNA regulatory mechanism may be different in various subtypes, leading to differences in clinical behavior. In order to illustrate the pathogenesis of different subtypes, identification of context-dependent miRNA-mRNA functional regulatory modules (MFRMs) is important. In this study, we propose a novel method (Figure 1) for the genome-wide identification of MFRMs for different genetic subtypes of heterogeneous diseases. We applied the novel method on MM, which is characterized by significant heterogeneity at the molecular level [21] and divided into several subtypes on the basis of chromosomal abnormalities, such as t(4;14), t(14;16), t(11;14), and *RB* deletion [22]. We identified abundant subtype specific MFRMs associated with MM pathogenesis. The miRNA-mRNA regulatory networks were constructed based on MFRMs and provided numerous subtype specific miRNA-mRNA interactions. Clustering analysis showed that the MFRMs involved in multiple MM subtypes could separate the corresponding MM subtypes, indicating that these MFRMs could potentially aid in elucidation of the mechanisms underlying differences in clinical behavior.

2. Materials and Methods

2.1. Preparation of the Data Set. The matched expression profiles of miRNAs and mRNAs of MM were obtained from the studies of Gutiérrez et al. [23] (GSE16558). According to cytogenetic abnormalities, the 60 patients were classified into five subtypes: 17 patients with t(4;14); 11 with t(11;14); four with t(14;16); 15 with *RB* DEL (*RB* deletion as a unique abnormality, *RB* deletion, and P53 deletion); and 13 with NFISH (Normal Fish).

The seven target prediction data sources were obtained from DIANA-microT [24, 25], PicTar5 [26], RNA22 (R3/R5) [27], RNAhybrid [28], TargetScan [29, 30], and miRanda [31, 32]. The MM associated genes (see Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/501262) were collected from three databases: Online Mendelian Inheritance in Man (OMIM), the Cancer Genome Project (CGP), and Genetic Association Database (GAD).

2.2. MFRMs Identification and Analysis. The methodology utilized in this study is illustrated in Figure 1. First, we identified initial comodules on matched mRNA and miRNA expression profiles using PPA [33]. A comodule is an ensemble of certain miRNAs, mRNAs, and samples, in which miRNAs and mRNAs exhibit similar patterns of expression across the same samples. The samples in the comodule imply the specific conditions under which the miRNAs and mRNAs act cooperatively. Second, to derive coherent modules associated with the pathogenesis of MM, we integrated GO BP [34] and miRNA target information to identify MFRMs in each comodule by multiobjective GA. Three optimization objectives were defined: (i) the minimum enriched P value on MM associated GO terms; (ii) the correlation coefficient and target coefficient (see Methods 2.3) of the module; and (iii) variations of expression values of miRNAs and mRNAs in the module. The multiobjective GA iteratively searched Pareto optimal solutions with three objectives and obtained noninferior MFRMs for each comodule. Next, we sorted the MFRMs according to their scores on three objectives. Finally, the top modules in the ranking results were identified and utilized to construct miRNA-mRNA regulatory networks or for clustering analysis.

2.3. Discovery of Comodules by PPA. We utilized the PPA [33] to identify comodules. The PPA is a modular analysis approach operating on two large-scale data sets that share one common dimension. Kutalik et al. [33] demonstrated that PPA could identify coherent patterns across paired data sets more effectively compared to classical approaches like clustering, regression, or SVD. A further advantage is that PPA provides context-dependent modules across paired data sets.

Let $\mathbf{E}_{N_G \times N_C}$ and $\mathbf{R}_{N_D \times N_C}$ represent paired gene expression data matrix and miRNA expression data matrix, respectively. N_G, N_D, and N_C represent the number of genes, miRNAs, and samples, respectively. Then the PPA is summarized in Pseudocode 1, where $|\mathbf{x}|$, $\mu(\mathbf{x})$, and $\sigma(\mathbf{x})$ denote the norm, mean, and standard deviation of the components x_i in the vector $\mathbf{x}$; $\hat{\mathbf{x}} = \mathbf{x}/|\mathbf{x}|$; t_G, t_D, and t_C denote the threshold of genes, miRNAs, and samples, respectively; $\mathbf{E}_G$ and $\mathbf{E}_C$ represent the gene expression matrix normalized across genes and samples, respectively; $\mathbf{R}_D$ and $\mathbf{R}_C$ represent the miRNA expression matrix normalized across miRNAs and samples, respectively.

Starting with the candidate set of genes ($\mathbf{g}^{(0)}$), the mRNA expression profile ($\mathbf{E}_{N_G \times N_C}$) was used to identify samples ($\bar{\mathbf{c}}^{(n)}$) in which these genes were coexpressed. Next, the

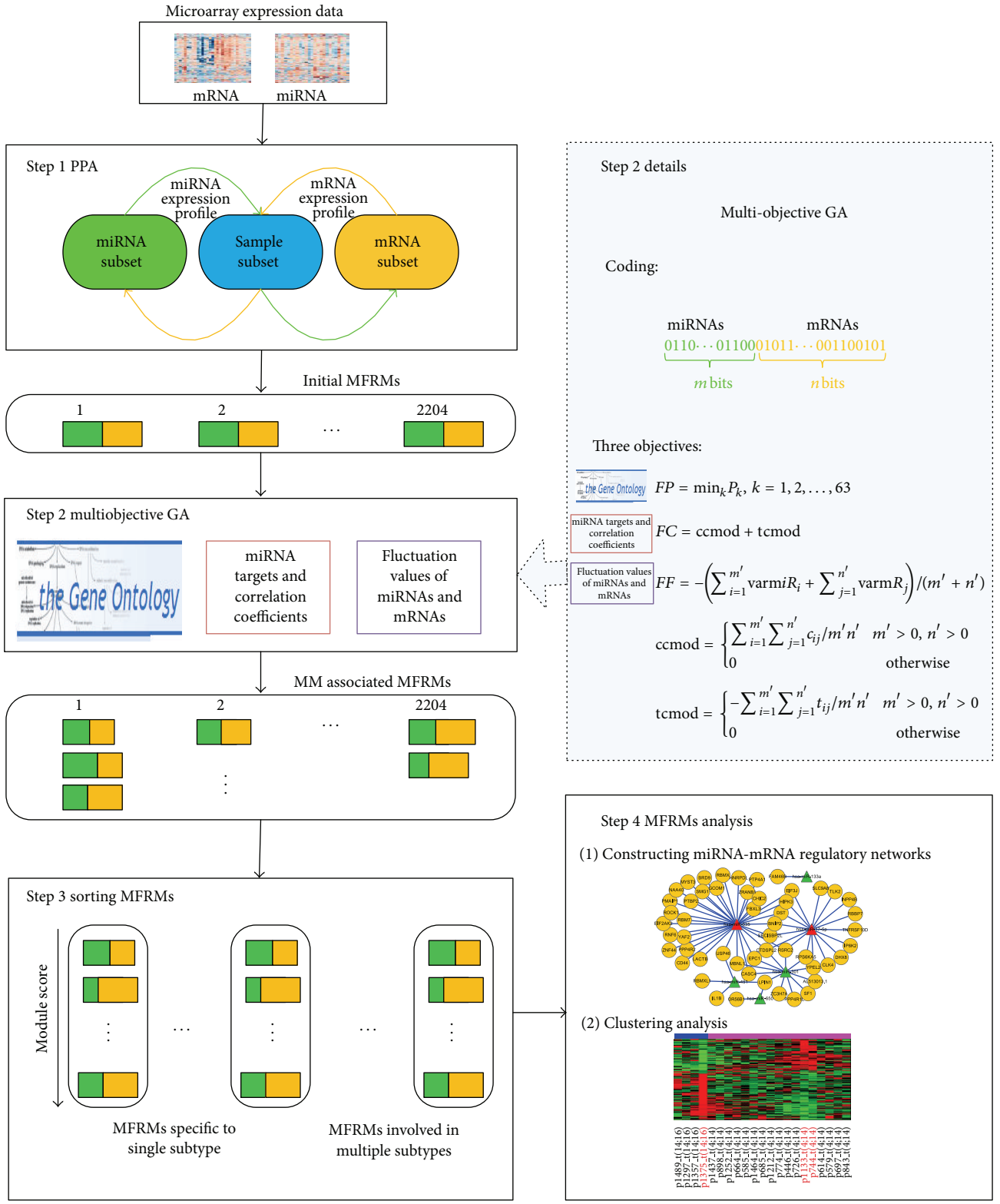


Figure 1: Workflow to identify MFRMs and MFRMs analysis.

$n = 0$; $\mathbf{g}^{(0)} = random(N_G) \in [0,1]^{N_G}$ (initial random *seed*)

While $\left(\left|\widehat{\mathbf{g}}^{(n)} - \widehat{\mathbf{g}}^{(n-1)}\right| + \left|\widehat{\mathbf{d}}^{(n)} - \widehat{\mathbf{d}}^{(n-1)}\right| + \left|\widehat{\mathbf{c}}^{(n)} - \widehat{\mathbf{c}}^{(n-1)}\right| + \left|\widehat{\widetilde{\mathbf{c}}}^{(n)} - \widehat{\widetilde{\mathbf{c}}}^{(n-1)}\right| > \varepsilon\right)$

(1) $\mathbf{c} = \mathbf{E}_G^T \cdot \widehat{\mathbf{g}}^{(n)}$; $c_j^{(n+1)} = \begin{cases} c_j & \text{if } \left|c_j - \mu(\mathbf{c})\right| > t_C\sigma(\mathbf{c}) \\ 0 & \text{otherwise} \end{cases}$ $(j = 1, \ldots, N_C)$

(2) $\mathbf{d} = \mathbf{R}_C \cdot \widehat{\mathbf{c}}^{(n)}$; $d_k^{(n+1)} = \begin{cases} d_k & \text{if } \left|d_k - \mu(\mathbf{d})\right| > t_D\sigma(\mathbf{d}) \\ 0 & \text{otherwise} \end{cases}$ $(k = 1, \ldots, N_D)$

(3) $\widetilde{\mathbf{c}} = \mathbf{R}_D^T \cdot \widehat{\mathbf{d}}^{(n)}$; $\widetilde{c}_l^{(n+1)} = \begin{cases} \widetilde{c}_l & \text{if } \left|\widetilde{c}_l - \mu(\widetilde{\mathbf{c}})\right| > \widetilde{t}_C\sigma(\widetilde{\mathbf{c}}) \\ 0 & \text{otherwise} \end{cases}$ $(l = 1, \ldots, N_C)$

(4) $\mathbf{g} = \mathbf{E}_C \cdot \widehat{\widetilde{\mathbf{c}}}^{(n)}$; $g_m^{(n+1)} = \begin{cases} g_m & \text{if } \left|g_m - \mu(\mathbf{g})\right| > t_G\sigma(\mathbf{g}) \\ 0 & \text{otherwise} \end{cases}$ $(m = 1, \ldots, N_G)$

(5) $n = n + 1$;

$\mathbf{g}^* = \mathbf{g}^{(n)}$; $\mathbf{c}^* = \mathbf{c}^{(n)}$; $\mathbf{d}^* = \mathbf{d}^{(n)}$

Pseudocode 1

miRNA expression profile ($\mathbf{R}_{N_D \times N_C}$) was utilized to select miRNAs ($\widehat{\mathbf{d}}^{(n)}$) that also exhibited a coherent expression in these samples ($\widehat{\mathbf{c}}^{(n)}$). This set of miRNAs ($\widehat{\mathbf{d}}^{(n)}$) was then utilized to refine the set of samples ($\widehat{\widetilde{\mathbf{c}}}^{(n)}$) by eliminating those which had an incoherent miRNA expression and adding others that behave similarly across these miRNAs. Finally, this refined set of samples ($\widehat{\widetilde{\mathbf{c}}}^{(n)}$) was used to probe for mRNAs ($\widehat{\mathbf{g}}^{(n)}$) coexpressed in these samples. This alternating procedure was reiterated until it converged to stable sets of mRNAs, samples, and miRNAs: comodules.

2.4. Identification of MM Associated GO BP. To identify MM associated GO BP, we conducted cumulative hypergeometric distribution test to identify specific biological processes enriched with the MM associated genes. A total of 63 MM associated GO BP were identified ($P < 0.05$, Bonferroni corrected, Table S2).

2.5. Identification of MM Associated MFRMs Based on Multiobjective GA. To identify biologically meaningful coherent modules, we utilized a multiobjective genetic algorithm to extract MFRMs for each comodule. Let m be the number of miRNAs in a comodule and n be the number of mRNAs. Our aim is to extract a subset of miRNAs from the m miRNAs and a subset of mRNAs from the n mRNAs and construct a MFRM in which (i) the extracted subset of mRNAs is significantly enriched in the MM associated GO BP, (ii) miRNA expression exhibits a significant negative correlation with mRNA expression across the samples in the comodule, and, concurrently, the miRNAs and mRNAs exhibit a strong targeting relationship, and (iii) their expression values vary greatly among different subtypes. To this end, we defined three optimization objectives (i.e., the fitness function) as follows:

$$FP = \min_k P_k, \quad k = 1, 2, \ldots, 63,$$

$$FC = \text{ccmod} + \text{tcmod},$$

$$FF = -\frac{\left(\sum_{i=1}^{m'} \text{varmi}R_i + \sum_{j=1}^{n'} \text{varm}R_j\right)}{\left(m' + n'\right)}, \tag{1}$$

where P_k was the P value (Bonferroni corrected) of the kth GO term enrichment on the subset of mRNAs. The first objective function FP represented the minimum P value of 63 MM associated GO term enrichments. The second objective function, FC, reflected the coherence of the module, where ccmod and tcmod were the correlation coefficient and target coefficient of the module, respectively:

$$\text{ccmod} = \begin{cases} \displaystyle\sum_{i=1}^{m'}\sum_{j=1}^{n'} \frac{c_{ij}}{m'n'} & m' > 0,\ n' > 0 \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{tcmod} = \begin{cases} \displaystyle -\sum_{i=1}^{m'}\sum_{j=1}^{n'} \frac{t_{ij}}{m'n'} & m' > 0,\ n' > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where m' and n' were the number of selected miRNAs and mRNAs, respectively. The term c_{ij} represented the Pearson correlation coefficient of the ith miRNA and jth mRNA' expression value across the samples in the comodule (P value < 0.05). The term t_{ij} represented the target coefficient of the ith miRNA and jth mRNA. The target coefficient between miRNA and mRNA was defined as the frequency that the mRNA was predicted as the target of the miRNA in seven target prediction data sources. The correlation coefficient ccmod and target coefficient tcmod of a module were defined as the average correlation coefficient and target coefficient of all miRNA-mRNA pairs in the module, respectively. The third objective function, FF, denoted the variations of miRNAs and mRNAs expression in the module, where varmiR_i and varmR_j were the between class variances of the ith miRNA's expression value and jth mRNA's expression value across five subtypes of MM, respectively. Both variances were normalized to between 0 and 1.

We used the "bit string" type to encode the individuals in the population X. Every individual x in X was encoded as a bit string with length $m + n$.

$$\underbrace{0110\ldots0110}_{m \text{ bits}}\underbrace{001011\ldots001}_{n \text{ bits}}\underbrace{100101}. \quad (3)$$

The first m bits represented m miRNAs in the comodule, and the remaining n bits represented n mRNAs in the comodule. The "0" represented the miRNA or mRNA selected into MFRM and "1" represented the miRNA or mRNA not selected. The number of "1" in the first m bits and remaining n bits was m' and n', respectively. The multiobjective optimization is formulated as

$$\begin{aligned} \min_{x \in X} \quad & F(x) = [FP(x), FC(x), FF(x)]^T \\ \text{s.t.} \quad & FP(x) < 0.05 \\ & m', n' \geq 1, \end{aligned} \quad (4)$$

where FP, FC, and FF are the objective functions defined as above. The solutions with fitness function $FP(x) \geq 0.05$ were not kept for further investigation as the mRNAs lists of these solutions were not significantly enriched on any MM associated biological process.

2.6. Sorting MFRMs. MFRMs can be classified into six categories according to the condition under which they act: t(11;14), *RB* DEL, t(4;14), NFISH and t(14;16) subtype specific MFRMs, and heterogeneous MFRMs. The first five categories of MFRMs are specific to a single subtype, whereas the last category is involved in multiple subtypes (Figure 1). We sorted the MFRMs in each category separately. First, we sorted the MFRMs according to each objective and achieved the ranks R_1, R_2, and R_3 on three objectives, respectively. The final score of a MFRM was then defined as the weighted sum of the three ranks:

$$S = \alpha R_1 + \beta R_2 + \gamma R_3. \quad (5)$$

We set $\alpha = \beta = \gamma = 1/3$. Finally, the MFRMs were sorted by their final scores in descending order.

3. Results

3.1. The Comodules Discovered by PPA. We applied the PPA to the matched miRNA and mRNA expression profiles of MM and produced 2204 comodules which contains mRNAs, miRNAs, and samples. In each comodule, mRNAs and miRNAs exhibit coherent expression across the same samples. These samples imply the specific conditions under which the miRNA-mRNA module acts. For example, if the samples in a comodule all belong to subtype t(4;14), we refer to the miRNA-mRNA module as t(4;14) specific module. The miRNAs and mRNAs in the t(4;14) specific module are coexpressed only in t(4;14) samples but not in samples with other subtypes. Thus, the miRNAs and mRNAs in the t(4;14) specific module may exhibit a function specific to t(4;14). The miRNA-mRNA modules can be classified into six categories according to the condition under which they act, that is, t(11;14), *RB* DEL, t(4;14), NFISH and t(14;16) subtype specific modules, and heterogeneous modules (in other words, the samples in the corresponding comodule belong to different MM subtypes; Figure S1 shows a heterogeneous module). Among the 2204 comodules, we identified 14, 58, 41, 15, and two comodules specific to MM subtype t(11;14), *RB* DEL, t(4;14), and NFISH and t(14;16), respectively. Figure 2 describes the distribution of the number of samples, mRNAs, and miRNAs attributed to 2204 comodules. The majority of comodules contained less than 2000 mRNAs and 60 miRNAs, and a few mRNAs and miRNAs acted as "hubs" by being part of up to 600 different comodules.

To assess the biological relevance of the mRNAs in the modules, we tested the functional homogeneity of the mRNAs in each module. A set of mRNAs is defined as functionally homogeneous if it is significantly enriched in at least one GO biological process category [34, 35]. Among the 2204 modules, 1679 (76.2%) were functionally homogeneous (q-value < 0.05, FDR correction), indicating that the majority of modules discovered by PPA were biologically meaningful. Thus, the PPA was reliable to perform on matched miRNA and mRNA expression profiles and identify biologically meaningful miRNA-mRNA modules.

3.2. MFRMs Associated with MM Identified by Multiobjective GA. The miRNA-mRNA modules in the above section were identified only based on the expression correlation of miRNAs and mRNAs. To identify modules that are more biologically meaningful, there are still two important aspects need to be considered: the miRNA-target relationships and identification of modules that associated with the pathogenesis of given disease. To this end, we applied multiobjective GA on each comodule to extract MFRMs by integrating miRNA target information and MM associated GO BP (See Methods).

For each comodule, the multiobjective GA produced a Pareto optimal solution set of noninferior MFRMs. More significant expression correlations and stronger target relationships between the miRNAs and mRNAs were observed in the extracted MFRMs. For example, the multiobjective GA got four MFRMs on comodule 1680 (Table 1). Each MFRM was enriched on MM associated GO BP ($FP < 0.05$). Both the expression correlation coefficient of miRNAs and mRNAs and the target coefficient of the module were optimized. The second objective FC which reflected the expression correlation and target relationship was improved from −0.1599 in the original comodule to −0.1774, −0.2049 and −0.3456 in three functional modules, respectively. Although FC of the third MFRM was inferior, the variations of miRNAs and mRNAs expression (FF, the third objective) in this MFRM were the best. The larger the FF, the larger the variation of miRNA and mRNA expression among different MM subtypes and the more subtype specific the MFRM. Comodule 1680 contained four samples: p709, p831, p841 and p1204 which all belonged to subtype *RB* DEL. This indicated that the miRNAs and mRNAs in the MFRM were only co-expressed in samples with MM subtype *RB* DEL. Because the mRNAs in the MFRMs were significantly enriched on MM associated GO BP, the MFRMs extracted from comodule 1680 were *RB* DEL

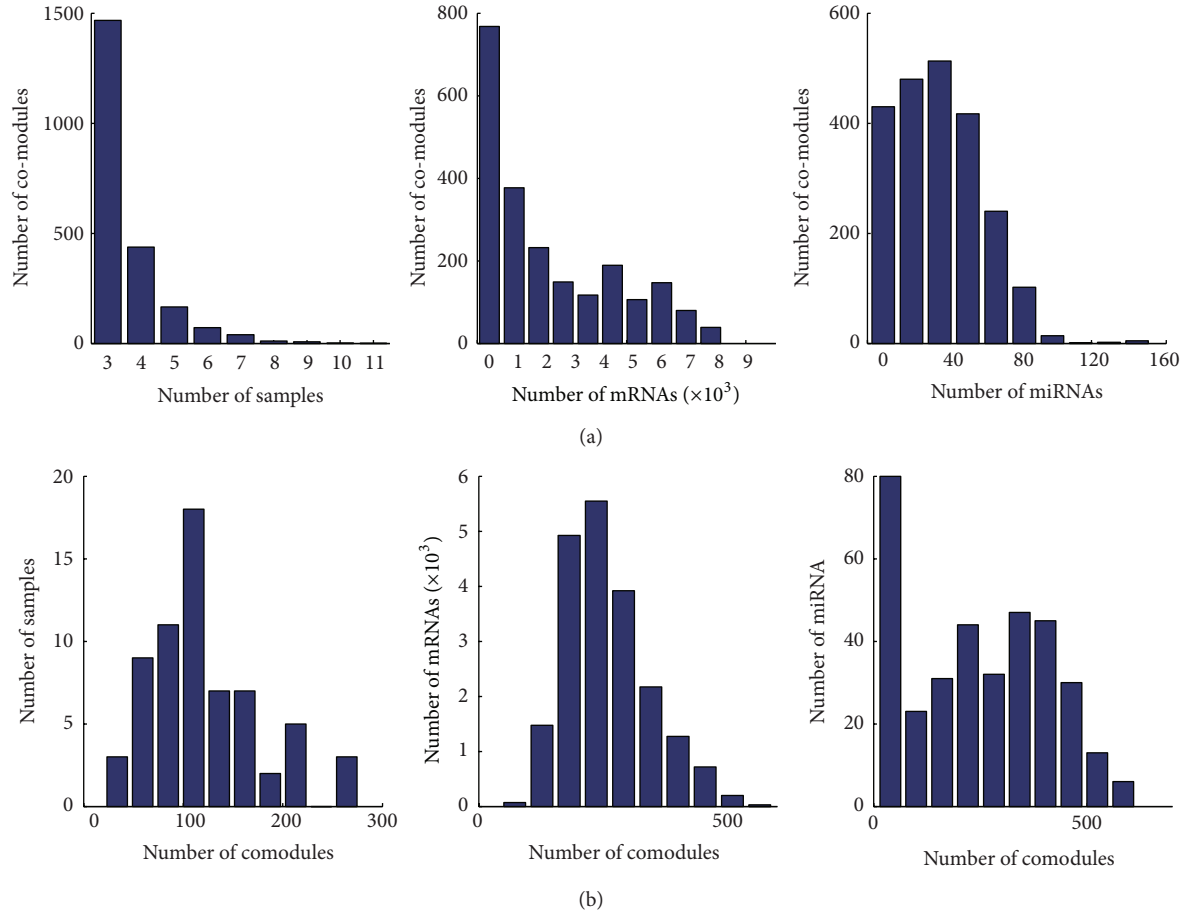


Figure 2: Comodule statistics. (a) The distribution of the number of comodules according to the number of samples, mRNAs, and miRNAs they contained. (b) The distribution of the number of samples, mRNAs, and miRNAs according to the number of comodules in which they were included.

Table 1: Multiobjective GA on comodule 1860.

	Comodule 1680	MFRM 1860-1	MFRM 1860-2	MFRM 1860-3	MFRM 1860-4
Number of miRNAs	20	14	15	16	3
Number of mRNAs	511	224	347	209	283
FP	None[a]	0	0.031	0.024	0.020
FC	−0.1599	−0.1774	−0.2049	−0.1442	−0.3456
FF	−0.0469	−0.0610	−0.0548	−0.0665	−0.0339

[a]The mRNAs were not enriched on any MM associated biological process.

specific MFRMs and may represent a regulatory mechanism leading to the specific pathogenesis of subtype *RB* DEL. Similarly, we obtained subtype specific MFRMs for other MM subtypes, such as t(4;14), t(11;14), t(14;16), and NFISH. Multiobjective GA may produce more than one MFRM for each comodule. Figure S2 shows the distribution of the number of MFRMs extracted from each comodule. Most comodules produced no more than five MFRMs. The MFRMs were sorted according to the three objectives (see Methods) and those with the highest rank had priority for further investigation.

3.3. The miRNA-mRNA Regulatory Networks Provided a Modular Outlook at Subtype Specific miRNA-mRNA Interactions: Two Case Studies. We obtained abundant subtype specific MFRMs for each MM subtype. We focused on the MFRMs that ranked the highest, and then constructed miRNA-mRNA regulatory networks based on the expression correlation and target relationship between miRNAs and mRNAs in the MFRM. A miRNA-mRNA pair was connected with an edge if it concurrently satisfied two condition: (i) the miRNA exhibited a significant negative correlation with the mRNA across the samples in the comodule; and (ii)

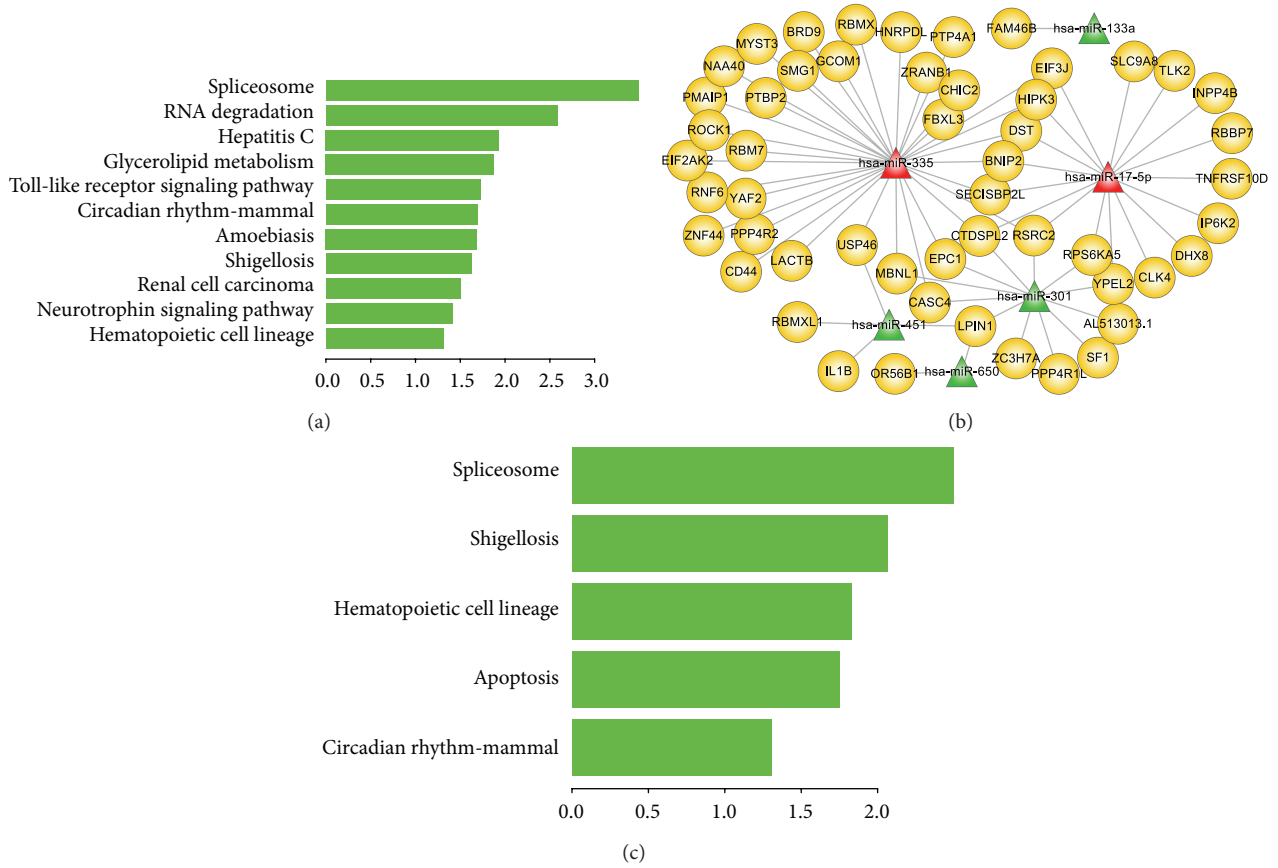


FIGURE 3: MiRNA-mRNA functional regulatory network of *RB* DEL specific MFRM 1680-1 and functions of miRNAs and mRNAs. (a) Significant biological pathways that genes of MFRM 1680-1 identified by multiobjective GA participated in. The *X*-axis represents − log 10 transformation of *P* value. (b) miRNA-mRNA functional regulatory network of *RB* DEL specific MFRM 1680-1. Six miRNAs and 52 genes were involved in the network. miRNA and mRNA were connected with an edge if and only if the mRNA was targeted by the miRNA and there was a significant negative correlation between miRNA and mRNA expression. The miRNAs in red color were previously reported to play key roles in the pathogenesis of MM. (c) Significant biological pathways that genes of miRNA-mRNA functional regulatory network of *RB* DEL specific MFRM 1680-1 participated in. The *X*-axis represents − log 10 transformation of *P* value.

the mRNA was predicted as a target of the miRNA by at least one miRNA target prediction algorithm. Two cases are presented below: a *RB* DEL specific MFRM and a t(4;14) specific MFRM.

3.3.1. RB DEL Specific MFRMs. *RB* DEL was a MM subtype that exhibited high morbidity rate, increased proliferative activity, and shorter overall survival [36]. We focused on the MFRM 1680-1 which ranked first among the *RB* DEL specific MFRMs. Firstly, we performed the functional enrichment analysis of mRNAs, indicating the functional roles of mRNAs belonged to this *RB* DEL specific MFRM. As shown in Figure 3(a), mRNAs in the MFRM significantly participated in several biological pathways that directly related with tumor. For example, Spliceosome is the most significant pathway; the study of Quidville et al. suggested that the deregulation of spliceosome induces mTOR Blockade and they provided the component of spliceosome as new therapeutic target of tumor [37]. Then, the miRNA-mRNA regulatory network was constructed based on the reverse expression and miRNA-target relationships (Figure 3(b)). There were 6 miRNAs and 52 genes in this MFRM (see Supplementary Materials for details). The functional enrichment analysis of these genes indicating that miRNAs and mRNAs in this MFRM significantly involved in the spliceosome and apoptosis biological pathways (Figure 3(c)), which are directly related to the occurrence and progression of tumor [37, 38]. Among these 6 miRNAs in MFRM, up to four miRNAs including miR-335, miR-17-5p, miR-451, and miR-301 were involved in a broad range of cancers [39, 40], such as acute lymphoblastic leukemia (ALL) [41], acute myeloid leukemia (AML) [42], and chronic myeloid leukemia (CML) [43]. In particular, miR-335 and miR-17-5p were connected with 33 and 17 mRNAs, respectively, thus exhibiting the important roles played in the network. Ronchetti et al. [44] reported that miR-335 was recurrently overexpressed in a fraction of primary tumors, possibly influencing plasma cell homing and/or interactions with the bone marrow microenvironment. miR-17-5p was a key regulator of the G1/S-phase cell cycle transition [45]. The study of Zhou et al. indicated that miR-17-5p exhibits a high expression level in myeloma cells and it may participate in the induction of p21Waf1/Cip1 expression,

which relevant to the cell-cycle arrest process [46]. *MYC* has been reported to play a causal role in the progression of monoclonal gammopathy to MM [47]. *EPC1*, which interacts with miR-335 in the miRNA-mRNA regulatory network, has been shown to participate in growth regulation and has been suggested to be involved in a *MYC*-centered regulatory network [48]. CD44 was also directly connected with miR-335 and relevant to tumor. Purushothaman and Toole indicated CD44 serves as the binding partner of serglycin participate in the progression of MM [49]. Bjorklund et al. suggested that CD44 may contribute to the lenalidomide resistance in MM [50]. Moreover, miR-451 has previously been identified as one of the signatures capable of accurate discrimination of ALL from AML [41]. Because ALL, AML, CLL, MM, and lymphoma are all hematological malignancies, it is likely that these miRNAs also played special functional roles in MM. Of the 52 genes, many genes have been reportedly involved in various cancers, such as *HIPK3* [51], *RSRC2* [52], *BPAG1* [53], and *EPC1* [48]. Eight genes were annotated on apoptosis process, including *BNIP2*, *CD44*, *HIPK3*, *IP6K2*, *IL1B*, *PMAIP1*, *ROCK1*, and *TNFRSF10D*. They interacted with miR-335, miR-17-5p, and miR-451, further indicating the central role of these miRNAs in the regulatory network. This suggests that the miRNAs and mRNAs in the network worked together and contributed to the pathogenesis MM subtype Del *RB*.

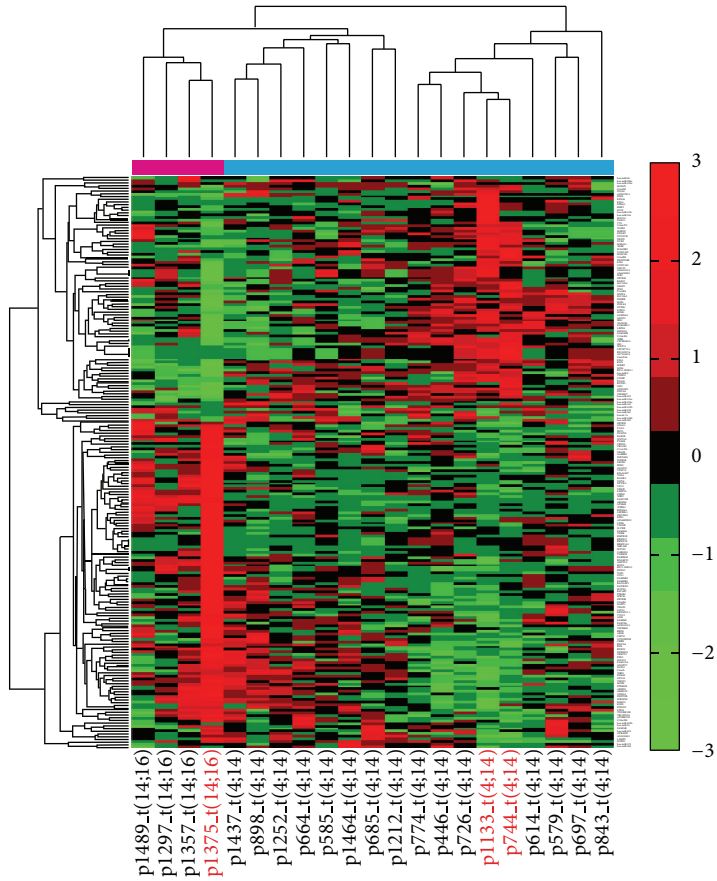
3.3.2. The t(4;14) Specific MFRMs. MM subtype t(4;14), translocation of a region of chromosome 4 to chromosome 14, was highly associated with poor prognosis [54–57]. We firstly carried out functional enrichment analysis on mRNAs in MFRM 1121-2 which is the top ranked t(4;14) specific MFRMs. As a result, genes in the module significantly involved in many biological pathways such as "NOD-like receptor signaling pathway", "MAPK signaling pathway", "apoptosis," and "tight junction" that have been known as hallmark processes of tumor (Figure 4(a)). Then, we constructed the miRNA-mRNA regulatory network of this t(4;14) specific MFRM, which including 36 miRNAs, 382 mRNAs and 983 edges (Figure 4(b)). Pathway enrichment analysis were also performed on these 382 mRMAs, the results suggest that miRNAs and genes in the MFRM were significantly involved in the biological pathways that directly related with tumor (Figure 4(c)). There were eight sub-networks identified, and most miRNAs and mRNAs were incorporated into the largest subnetwork. In the largest subnetwork, several miRNAs (let-7a, miR-125a, miR-193b, miR-25, and miR-181c) that acted as hubs in the network were previously reported to be associated with MM pathogenesis [6, 58]. Let-7a and mir-125a played important role in the t(4;14) regulatory network, in concordance with a previous study by Lionetti et al. [58]. They found that patients with t(4;14) exhibited specific overexpression of the miRNA cluster with let-7e, miR-125a, and miR-99b. Bakkus et al. also reported that Let-7a has a higher expression level in both the MM patients and cell lines [59]. Changes expression of miR-125a and let-7f which is in the same family of let-7a contributes to the myelomagenesis and are also relevant to overall prognosis [60]. Furthermore, the expression of miR-125b which is the same family member of miR-125a is associated with the chemotherapeutic-induced cell death in MM [61]. MiR-193b was a member of the miR-193b-365 cluster, which was previously identified as part of the unique miRNA signature in MM [62]. Mir-25 was a member of the oncogenic cluster miR-106b-25. Pichiorri et al. [6] determined that the oncogenic cluster miR-106b-25, miR-181a and miR-181b, which belonged to the same gene family with mir-181c, was a miRNA signature in the malignant transformation from MGUS to MM. Upregulation of miR-25 and miR-181-a/b and inactivation of miR-34, a central player in a smaller subnetwork, could negatively regulate the expression of the tumor suppressor gene *p53* [6, 63, 64], and contribute to MM progression. Alteration in miRNA expression (such as miR-34, miR-25, miR-181a/b, and miR-30d) during the progression from MGUS to newly diagnosed MM could be partially responsible for *p53* inactivation [5]. miR-25 is connected with 42 mRNAs. Many of these mRNAs were cotargeted by other miRNAs in the network, such as *RALA*, *BAK1*, *BMF*, and *JARID2*. *RALA* was targeted by 11 miRNAs. The product of *RALA* belonged to the oncogene *RAS* family of proteins and was involved in the MAPK/ERK signal transduction pathway which is the hallmark process of tumor. The study of Lim et al. indicated that activation of RALA play important role in the Ras-induced tumorigenesis [65]. *BAK1* and *BMF* were targeted by four and six miRNAs, respectively. *JARID2* is an ortholog of the mouse *jumonji* gene that negatively regulates cell proliferation: it was targeted by nine miRNAs in the network, suggesting that it may also play an important role in human MM. Aside from these genes, 56.3% of genes in the network were targeted by multiple miRNAs, exhibiting the synergistic regulatory mechanism of miRNAs; miRNAs, along with genes, comprised the complex network specific to t(4;14).

3.4. Heterogeneous MFRMs Were Able to Separate Corresponding Subtypes. Aside from MFRMs involved in a single subtype, we obtained a heterogeneous MFRM collection covering multiple MM subtypes. We found that some heterogeneous MFRMs exhibited differences in corresponding subtypes. For example, in heterogeneous comodule 1649 (Figure S1), expression of miRNAs and mRNAs between subtypes t(14;16) and t(4;14) was negatively correlated. The heterogeneous MFRMs extracted from this comodule potentially contain a mechanism that leads to a difference in subtype. We performed hierarchical clustering on both miRNAs and mRNAs in MFRM 1649-2 (ranked first among the MFRMs extracted from comodule 1649) for all t(4;14) and t(14;16) samples (Figure 5(a)). The clustering results confirmed that the 21 miRNAs and 196 mRNAs in this MFRM could separate t(4;14) from t(14;16) patients. Two miRNAs (hsa-let-7e; hsa-miR-125a) in this MFRM have been previously reported as miRNA signatures for their specific overexpression in t(4;14). Another miRNA miR-25, which was discussed above in t(4;14) specific MFRM 1121-2, was also incorporated in this MFRM. The 196 genes were significantly enriched in regulation of cell proliferation (P value = 6.9×10^{-3}), regulation of ossification (P value = 8.2×10^{-4}), blood vessel morphogenesis (P value = 3.7×10^{-3}), blood vessel development (P value = 8.3×10^{-3}), angiogenesis (P value = 1.3×10^{-2}), and

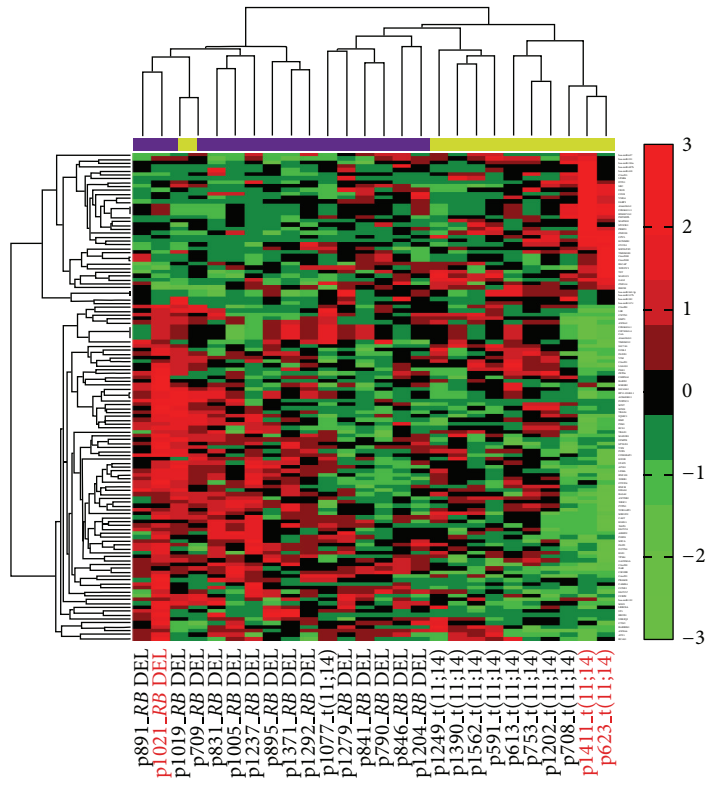
Figure 4: miRNA-mRNA functional regulatory network of t(4;14) specific MFRM 1121-2 and functions of miRNAs and mRNAs. (a) Significant biological pathways that genes of MFRM 1121-2 identified by multiobjective GA participated in. The X-axis represents $-\log 10$ transformation of P value. (b) miRNA-mRNA functional regulatory network of t(4;14) specific MFRM 1121-2. There were 36 miRNAs, 382 mRNAs and 983 edges involved in the network. The miRNAs in red color were previously reported to play key roles in the pathogenesis of MM. (c) Significant biological pathways that genes of miRNA-mRNA functional regulatory network of t(4;14) specific MFRM 1121-2 participated in. The X-axis represents $-\log 10$ transformation of P value.

others, suggesting that the functional module contributes to the difference of t(4;14) and t(14;16). Next, we investigated the heterogeneous MFRM 1953-3, identified in patients with *RB* DEL and patients with t(11;14). Unsupervised hierarchical clustering showed that the 10 miRNAs and 115 genes could separate *RB* DEL from t(11;14), aside from one sample (Figure 5(b)). Interestingly, the heterogeneous MFRM 1962-4 acted in three subtypes: *RB* DEL, t(11;14), and t(14;16). The expression of miRNAs and genes was positively correlated between samples in *RB* DEL and t(11;14), but negatively correlated between samples in t(14;16) and *RB* DEL, t(11;14), suggesting that this module could lead to functional differences between t(14;16) and the other two subtypes. Clustering analysis using integrated miRNA and mRNA expression profiles showed that seven miRNAs and 138 mRNAs could separate t(14;16) from *RB* DEL and t(11;14) patients (Figure 5(c)).

3.5. The MFRMs Revealed Active miRNAs and mRNAs in Each MM Subtype. Overall, a few miRNAs and mRNAs act as "hubs" by being part of the majority of MFRMs. Further investigation of the miRNAs and mRNAs that appeared most frequently in subtype specific MFRMs will be helpful to elucidate the pathogenesis underlying each subtype. We referred to these miRNAs/mRNAs as subtype dependent active miRNAs/mRNAs. For subtype *RB* DEL, *CCDC50* was an active gene included in the majority of *RB* DEL specific MFRMs. It has been reported that tyrosine phosphorylation of *CCDC50* is important for inhibition of the *NFkB*-mediated

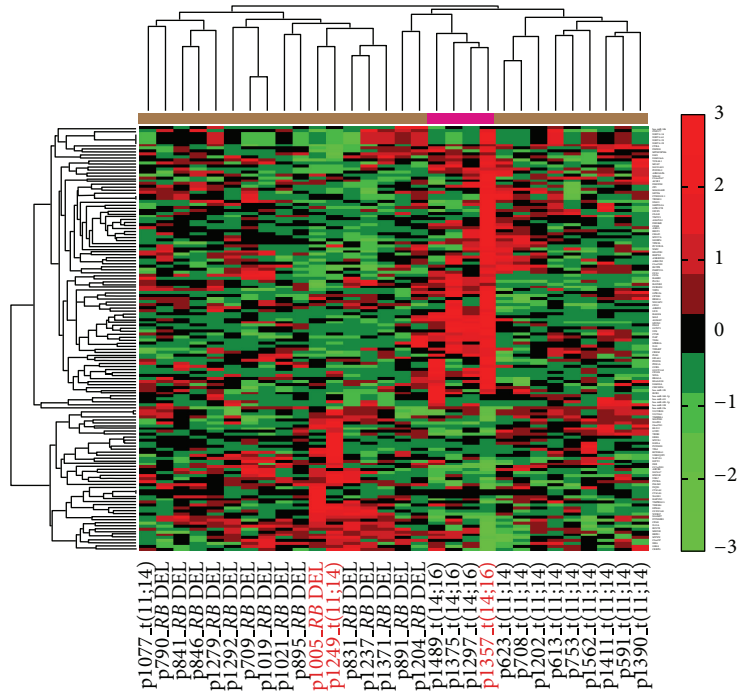


(a)



(b)

Figure 5: Continued.



(c)

Figure 5: Hierarchical clustering diagrams. (a) The clustering diagram on t(14;16) and t(4;14) samples using the miRNAs and mRNAs in MFRM 1649-2. (b) The clustering diagram on *RB* DEL and t(11;14) samples using the miRNAs and mRNAs in MFRM 1653-3. (c) The clustering diagram on *RB* DEL, t(11;14) and t(14;16) samples using the miRNAs and mRNAs in MFRM 1962-4. Clustering could separate t(14;16) samples from other two subtypes. The samples in red color were from respective MFRMs. Although the MFRMs were identified in small subsets of samples only, they captured the principle characteristics of different MM subtypes.

apoptotic pathway [66] and *CCDC50* is required for survival in mantle cell lymphoma (MCL) and CLL cells [67]. *KAT5* was another active gene: Zhao et al. [68] demonstrated that *KAT5* negatively modulated *c-Myb* transcriptional activity by recruiting histone deacetylases in human hematopoietic cells. Other active genes, like *NFKBIB*, *PIK3CA*, *RELA*, *LYN*, and *MAP2K7*, were involved in B cell and T cell receptor signaling pathways. These genes frequently appeared in *RB* DEL specific MFRMs, demonstrating that they played critical roles in *RB* DEL. The top 15 miRNAs and 50 mRNAs frequently included in each type of subtype specific MFRMs are listed in Table S3.

4. Discussion

Identification of subtype specific miRNA-mRNA modules is important for the study of heterogeneous diseases. Several points need to be considered regarding miRNA-mRNA modules: (i) the mRNAs are targeted by miRNAs in the same module; (ii) there may be a significant expression correlation of miRNAs and mRNAs; (iii) the functions that the miRNA-mRNA modules perform; (iv) the conditions under which the modules work. Most methods [12–16, 19] considered (i) and (ii) but ignored (iii) and (iv). Besides, the methods currently employed assign a miRNA/mRNA to only one module. However, a miRNA/mRNA may participate in different biological processes working with different genes and miRNAs. In this study, we used the PPA algorithm to identify miRNA-mRNA modules. The advantage of PPA is twofold. First, it can assign miRNAs and genes to multiple modules, which is well motivated from the biological point of view as the same gene can function in multiple processes under different conditions. Second, the PPA could identify context-dependent modules in which the miRNAs and genes are coexpressed in a subset of samples. These modules are widely ignored by many other clustering algorithms which calculate correlations over all samples. Another improvement we propose is the ability to identify condition-related MFRMs associated with predefined biological processes (e.g., MM associated GO BP). This process utilizes an integrated analysis of three pieces of biological data: GO BP, miRNA target information, and expression data based on multiobjective GA. The first objective *FP* utilizes the predefined MM associated GO BP to optimize the MFRMs. It ensures that the MFRMs are biologically meaningful and associated with the pathogenesis of MM. The second objective *FC* integrates both expression profiles and miRNA target information. It guarantees that the miRNA expression is significantly negatively correlated with the mRNA expression in the MFRM, and concurrently the miRNAs and mRNAs exhibit a strong targeting relationship. The last objective *FF* is based on the intuition that the expression values of miRNAs and genes which lead to pathogenesis and heterogeneity of MM may vary greatly among the different subtypes. Our method

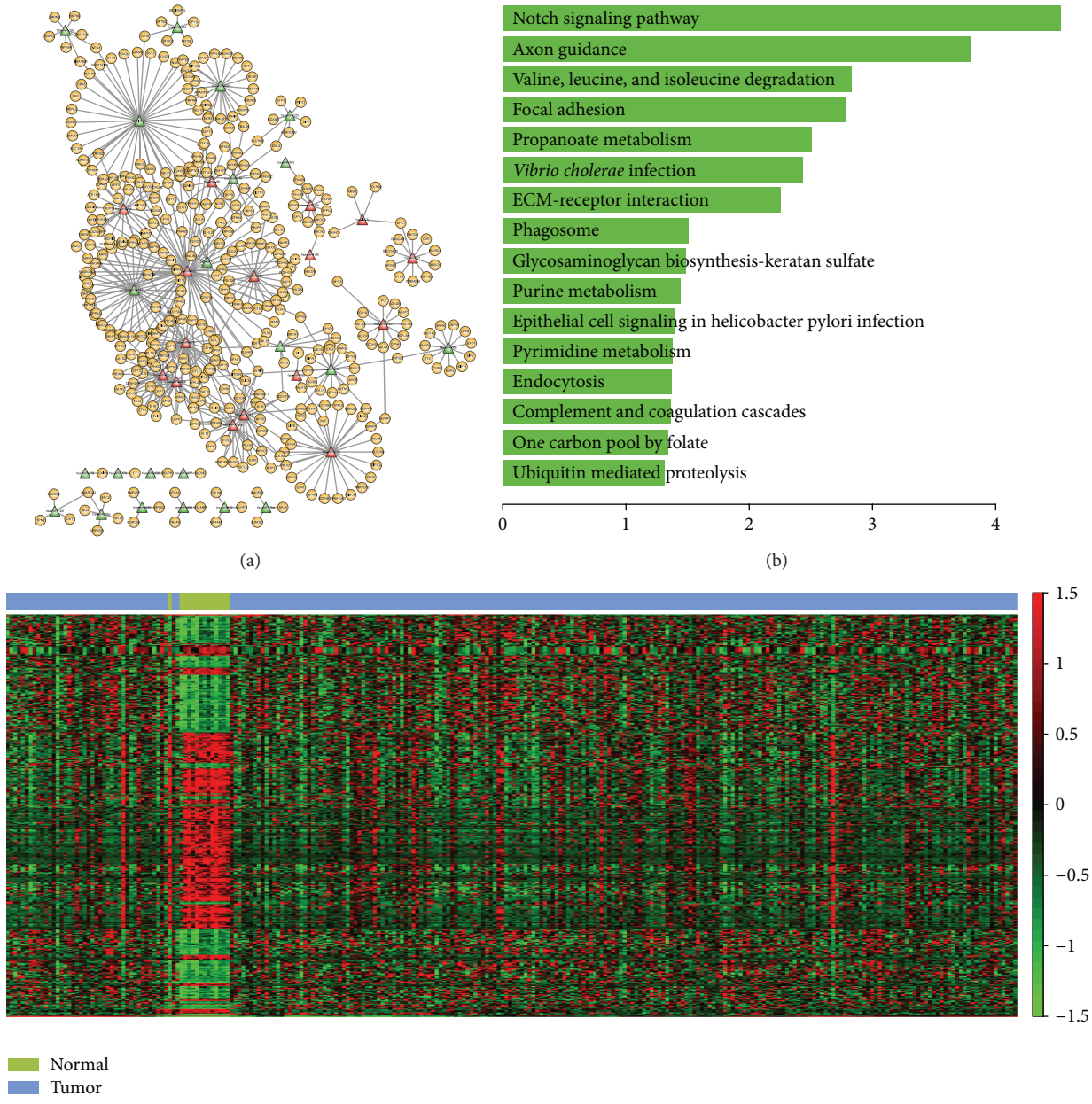


Figure 6: miRNA-mRNA functional regulatory network of breast cancer and functions of miRNAs and mRNAs. (a) miRNA-mRNA functional regulatory network. The miRNAs in red color were reported to relevant with breast cancer in miR2Disease database. (b) Significant biological pathways that genes of miRNA-mRNA functional regulatory network of breast cancer participated in. The X-axis represents $-\log 10$ transformation of P value. (c) The hierarchical clustering diagram on breast cancer dataset using the 38 miRNAs and 418 genes in MFRM.

captures a resource of subtype specific MFRMs that constitute various specific functional mechanisms in each MM subtype that may lead to differences in clinical behavior.

In order to examine the robustness and extensive application of our method, we performed it on breast cancer data set, which is RNA-seq data of TCGA including 14 normal samples and 248 cancer samples (http://cancergenome.nih.gov/). In total, PPA algorithm obtained 66 modules, 4 of these modules (modules 1, 12, 27, and 42) were normal samples specific indicating the dysregulation of these modules were associated with breast cancer. We then constructed the miRNA-mRNA regulatory network and carried out functional analysis of module 1. The module 1 regulatory network contained 38 miRNAs, 418 genes, and 537 edges (Figure 6(a)). These miRNAs and genes were involved in several biological pathways that directly associated with tumor such as "focal adhesion", "notch signaling pathway", "purine metabolism," and "ECM-receptor interaction" (Figure 6(b)). Of these 38 miRNAs in the network, up to 16 miRNAs were recoded to be relevant with breast cancer in miR2Disease database [69].

For example, the study of Guttilla and White indicate that the coordinately regulation of FOXO1 by miR-96 and miR-182 which involved in the miRNA-mRNA regulatory network was associated with the oncogenic state in breast cancer cells [70]. Furthermore, regulation of Rac1 signaling by ARF1 which directly interact with miR-96 in the regulatory network is associated with invasive breast cancer cells [71]. We used two-dimensional hierarchical clustering analysis to visualize the expression pattern of miRNAs and genes in the MFRM. As shown in Figure 6(c), these miRNAs and genes exhibit different expression pattern in normal and tumor samples. In summary, these results suggest that our method can robustly capture important MFRMs relevant to diseases when applied to RNA-Seq data.

We identified a large number of subtype specific MFRMs, such as MFRM 1680-1, 1121-2. The regulatory networks built on these two MFRMs were specific to *RB* DEL and t(4;14), respectively. The links in the regulatory networks predicted new potential subtype dependent miRNA-mRNA interactions. The genes in the two regulatory networks were significantly enriched in MM associated biological processes, such as apoptosis, and regulation of cell death. Although the miRNAs, genes and the regulatory mechanisms were different, they all contributed to the pathogenesis of their respective subtypes. Further investigation of other subtype specific MFRMs may uncover a different pathogenesis in each subtype.

For heterogeneous MFRMs involved in multiple subtypes, the miRNAs and mRNAs acted in different ways between the subtypes. For example, in MFRM 1649-2, approximately one third of mRNAs were upregulated in t(4;14) but downregulated in t(14;16). Clustering on three heterogeneous MFRMs showed that these MFRMs could separate different subtypes, although this only involved a small subset of the corresponding subtypes; the reason for this may be that all of the samples of the corresponding subtypes were not covered due, in part, to individual differences. Clustering results indicated that heterogeneous MFRMs captured natural differences and led to different subtypes. These MFRMs could potentially be helpful for identifying functional biomarkers of MM subtypes.

Conflict of Interests

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

Acknowledgments

This work was supported in part by the National High Technology Research and Development Program of China [863 Program, Grant no. 2014AA021102], the National Program on Key Basic Research Project [973 Program, Grant no. 2014CB910504], the National Natural Science Foundation of China [Grant nos. 91129710, 31200996, and 61170154], the National Science Foundation for Young Scientists of Heilongjiang Institute of Technology [Grant no. 2013QJ14], and the Scientific Research Fund of Heilongjiang Education Department [Grant nos. 12541684 and YJSCX2012-252HLJ].

References

[1] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[2] E. A. Miska, "How microRNAs control cell division, differentiation and death," *Current Opinion in Genetics & Development*, vol. 15, no. 5, pp. 563–568, 2005.

[3] J. Hayes, P. P. Peruzzi, and S. Lawler, "MicroRNAs in cancer: biomarkers, functions and therapy," *Trends in Molecular Medicine*, vol. 20, no. 8, pp. 460–469, 2014.

[4] G. Di Leva and C. M. Croce, "Roles of small RNAs in tumor formation," *Trends in Molecular Medicine*, vol. 16, no. 6, pp. 257–267, 2010.

[5] F. Pichiorri, L. De Luca, and R. I. Aqeilan, "MicroRNAs: new players in multiple myeloma," *Frontiers in Genetics*, vol. 2, p. 22, 2011.

[6] F. Pichiorri, S.-S. Suh, M. Ladetto et al., "MicroRNAs regulate critical genes associated with multiple myeloma pathogenesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 35, pp. 12885–12890, 2008.

[7] C. M. Croce and G. A. Calin, "miRNAs, cancer, and stem cell division," *Cell*, vol. 122, no. 1, pp. 6–7, 2005.

[8] A. Esquela-Kerscher and F. J. Slack, "Oncomirs—microRNAs with a role in cancer," *Nature Reviews Cancer*, vol. 6, no. 4, pp. 259–269, 2006.

[9] H. Jin, W. Tuo, H. Lian, Q. Liu, X.-Q. Zhu, and H. Gao, "Strategies to identify microRNA targets: new advances," *New Biotechnology*, vol. 27, no. 6, pp. 734–738, 2010.

[10] P. M. Voorhoeve, "MicroRNAs: oncogenes, tumor suppressors or master regulators of cancer heterogeneity?" *Biochimica et Biophysica Acta*, vol. 1805, no. 1, pp. 72–86, 2010.

[11] L. Hiddingh, R. S. Raktoe, J. Jeuken et al., "Identification of temozolomide resistance factors in glioblastoma via integrative miRNA/mRNA regulatory network analysis," *Scientific Reports*, vol. 4, p. 5260, 2014.

[12] V. Jayaswal, M. Lutherborrow, D. D. F. Ma, and Y. H. Yang, "Identification of microRNA-mRNA modules using microarray data," *BMC Genomics*, vol. 12, article 138, 2011.

[13] J.-G. Joung and Z. Fei, "Identification of microRNA regulatory modules in *Arabidopsis* via a probabilistic graphical model," *Bioinformatics*, vol. 25, no. 3, pp. 387–393, 2009.

[14] J.-G. Joung, K.-B. Hwang, J.-W. Nam, S.-J. Kim, and B.-T. Zhang, "Discovery of microRNA-mRNA modules via population-based probabilistic learning," *Bioinformatics*, vol. 23, no. 9, pp. 1141–1147, 2007.

[15] B. Liu, J. Li, and A. Tsykin, "Discovery of functional miRNA-mRNA regulatory modules with computational methods," *Journal of Biomedical Informatics*, vol. 42, no. 4, pp. 685–691, 2009.

[16] S. Yoon and G. de Micheli, "Prediction of regulatory modules comprising microRNAs and target genes," *Bioinformatics*, vol. 21, supplement 2, pp. ii93–ii100, 2005.

[17] C. Zhang, C. Li, J. Li et al., "Identification of miRNA-mediated core gene module for glioma patient prediction by integrating high-throughput miRNA, mRNA expression and pathway structure," *PLoS ONE*, vol. 9, no. 5, Article ID e96908, 2014.

[18] Y. Xiao, Y. Ping, H. Fan et al., "Identifying dysfunctional miRNA-mRNA regulatory modules by inverse activation, cofunction, and high interconnection of target genes: a case study of glioblastoma," *Neuro-Oncology*, vol. 15, no. 7, pp. 818–828, 2013.

[19] X. Peng, Y. Li, K.-A. Walters et al., "Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers," *BMC Genomics*, vol. 10, article 373, 2009.

[20] K. Bryan, M. Terrile, I. M. Bray et al., "Discovery and visualization of miRNA-mRNA functional modules within integrated data using bicluster analysis," *Nucleic Acids Research*, vol. 42, no. 3, article e17, 2014.

[21] H. Avet-Loiseau, F. Magrangeas, P. Moreau et al., "Molecular heterogeneity of multiple myeloma: pathogenesis, prognosis, and therapeutic implications," *Journal of Clinical Oncology*, vol. 29, no. 14, pp. 1893–1897, 2011.

[22] F. Zhan, Y. Huang, S. Colla et al., "The molecular classification of multiple myeloma," *Blood*, vol. 108, no. 6, pp. 2020–2028, 2006.

[23] N. C. Gutiérrez, M. E. Sarasquete, I. Misiewicz-Krzeminska et al., "Deregulation of microRNA expression in the different genetic subtypes of multiple myeloma and correlation with gene expression profiling," *Leukemia*, vol. 24, no. 3, pp. 629–637, 2010.

[24] M. Maragkakis, P. Alexiou, G. L. Papadopoulos et al., "Accurate microRNA target prediction correlates with protein repression levels," *BMC Bioinformatics*, vol. 10, article 295, 2009.

[25] M. Maragkakis, M. Reczko, V. A. Simossis et al., "DIANA-microT web server: elucidating microRNA functions through target prediction," *Nucleic Acids Research*, vol. 37, no. 2, pp. W273–W276, 2009.

[26] A. Krek, D. Grün, M. N. Poy et al., "Combinatorial microRNA target predictions," *Nature Genetics*, vol. 37, no. 5, pp. 495–500, 2005.

[27] K. C. Miranda, T. Huynh, Y. Tay et al., "A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes," *Cell*, vol. 126, no. 6, pp. 1203–1217, 2006.

[28] M. Rehmsmeier, P. Steffen, M. Höchsmann, and R. Giegerich, "Fast and effective prediction of microRNA/target duplexes," *RNA*, vol. 10, no. 10, pp. 1507–1517, 2004.

[29] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[30] B. P. Lewis, I.-H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge, "Prediction of mammalian microRNA targets," *Cell*, vol. 115, no. 7, pp. 787–798, 2003.

[31] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks, "MicroRNA targets in *Drosophila*," *Genome biology*, vol. 5, no. 1, article R1, 2003.

[32] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, "The role of site accessibility in microRNA target recognition," *Nature Genetics*, vol. 39, no. 10, pp. 1278–1284, 2007.

[33] Z. Kutalik, J. S. Beckmann, and S. Bergmann, "A modular approach for integrative analysis of large-scale gene-expression and drug-response data," *Nature Biotechnology*, vol. 26, no. 5, pp. 531–539, 2008.

[34] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[35] S. Zhang, C.-C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou, "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data," *Nucleic Acids Research*, vol. 40, no. 19, pp. 9379–9391, 2012.

[36] N. Zojer, R. Königsberg, J. Ackermann et al., "Deletion of 13q14 remains an independent adverse prognostic variable in multiple myeloma despite its frequent detection by interphase fluorescence in situ hybridization," *Blood*, vol. 95, no. 6, pp. 1925–1930, 2000.

[37] V. Quidville, S. Alsafadi, A. Goubar et al., "Targeting the deregulated spliceosome core machinery in cancer cells triggers mTOR blockade and autophagy," *Cancer Research*, vol. 73, no. 7, pp. 2247–2258, 2013.

[38] T. Li, N. Kon, L. Jiang et al., "Tumor suppression in the absence of p53-mediated cell-cycle arrest, apoptosis, and senescence," *Cell*, vol. 149, no. 6, pp. 1269–1283, 2012.

[39] D. Serpico, L. Molino, and S. Di Cosimo, "microRNAs in breast cancer development and treatment," *Cancer Treatment Reviews*, vol. 40, no. 5, pp. 595–604, 2014.

[40] J. Wang, K.-Y. Zhang, S.-M. Liu, and S. Sen, "Tumor-Associated circulating micrornas as biomarkers of cancer," *Molecules*, vol. 19, no. 2, pp. 1912–1938, 2014.

[41] S. Mi, J. Lu, M. Sun et al., "MicroRNA expression signatures accurately discriminate acute lymphoblastic leukemia from acute myeloid leukemia," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 50, pp. 19971–19976, 2007.

[42] H. Zhang, X.-Q. Luo, P. Zhang et al., "MicroRNA patterns associated with clinical prognostic parameters and CNS relapse prediction in pediatric acute leukemia," *PLoS ONE*, vol. 4, no. 11, Article ID e7826, 2009.

[43] L. Venturini, K. Battmer, M. Castoldi et al., "Expression of the miR-17-92 polycistron in chronic myeloid leukemia (CML) CD34+ cells," *Blood*, vol. 109, no. 10, pp. 4399–4405, 2007.

[44] D. Ronchetti, M. Lionetti, L. Mosca et al., "An integrative genomic approach reveals coordinated expression of intronic miR-335, miR-342, and miR-561 with deregulated host genes in multiple myeloma," *BMC Medical Genomics*, vol. 1, article 37, 2008.

[45] N. Cloonan, M. K. Brown, A. L. Steptoe et al., "The miR-17-5p microRNA is a key regulator of the G1/S phase cell cycle transition," *Genome Biology*, vol. 9, no. 8, article R127, 2008.

[46] Y. Zhou, L. Chen, and B. Barlogie, "High-risk myeloma is associated with global elevation of miRNAs and overexpression of *EIF2C2*/*AGO2*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 17, pp. 7904–7909, 2010.

[47] M. Chesi, D. F. Robbiani, M. Sebag et al., "AID-dependent activation of a MYC transgene induces multiple myeloma in a conditional mouse model of post-germinal center malignancies," *Cancer Cell*, vol. 13, no. 2, pp. 167–180, 2008.

[48] J. Kim, A. J. Woo, J. Chu et al., "A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs," *Cell*, vol. 143, no. 2, pp. 313–324, 2010.

[49] A. Purushothaman and B. P. Toole, "Serglycin proteoglycan is required for multiple myeloma cell adhesion, in vivo growth, and vascularization," *Journal of Biological Chemistry*, vol. 289, no. 9, pp. 5499–5509, 2014.

[50] C. C. Bjorklund, V. Baladandayuthapani, H. Y. Lin et al., "Evidence of a role for CD44 and cell adhesion in mediating resistance to lenalidomide in multiple myeloma: therapeutic implications," *Leukemia*, vol. 28, no. 2, pp. 373–383, 2014.

[51] J. F. Curtin and T. G. Cotter, "JNK regulates HIPK3 expression and promotes resistance to Fas-mediated apoptosis in DU 145 prostate carcinoma cells," *Journal of Biological Chemistry*, vol. 279, no. 17, pp. 17090–17100, 2004.

[52] H. Kurehara, H. Ishiguro, M. Kimura et al., "A novel gene, RSRC2, inhibits cell proliferation and affects survival in esophageal cancer patients," *International Journal of Oncology*, vol. 30, no. 2, pp. 421–428, 2007.

[53] T. Shimbo, A. Tanemura, T. Yamazaki, K. Tamai, I. Katayama, and Y. Kaneda, "Serum anti-BPAG1 auto-antibody is a novel marker for human melanoma," *PLoS ONE*, vol. 5, no. 5, Article ID e10566, 2010.

[54] H. Avet-Loiseau, T. Facon, B. Grosbois et al., "Oncogenesis of multiple myeloma: 14q32 and 13q chromosomal abnormalities are not randomly distributed, but correlate with natural history, immunological features, and clinical presentation," *Blood*, vol. 99, no. 6, pp. 2185–2191, 2002.

[55] A. M. Dring, F. E. Davies, J. A. L. Fenton et al., "A global expression-based analysis of the consequences of the t(4;14) translocation in myeloma," *Clinical Cancer Research*, vol. 10, no. 17, pp. 5692–5701, 2004.

[56] R. Fonseca, C. S. Debes-Marun, E. B. Picken et al., "The recurrent IgH translocations are highly associated with nonhyperdiploid variant multiple myeloma," *Blood*, vol. 102, no. 7, pp. 2562–2567, 2003.

[57] J. J. Keats, T. Reiman, C. A. Maxwell et al., "In multiple myeloma, t(4;14)(p16;q32) is an adverse prognostic factor irrespective of FGFR3 expression," *Blood*, vol. 101, no. 4, pp. 1520–1529, 2003.

[58] M. Lionetti, M. Biasiolo, L. Agnelli et al., "Identification of microRNA expression patterns and definition of a microRNA/mRNA regulatory network in distinct molecular groups of multiple myeloma," *Blood*, vol. 114, no. 25, pp. e20–e26, 2009.

[59] M. Bakkus, S. Dujardin, I. van Riet, and M. de Waele, "MicroRNA expression analysis in multiple myeloma plasma cells and cell lines by a quantitative real-time PCR approach," *Blood*, vol. 110, abstract 2472, pp. 1330–1333, 2007.

[60] S. Adamia, H. Avet-Loiseau, S. B. Amin et al., "Clinical and biological significance of microRNA profiling in patients with myeloma," *Journal of Clinical Oncology*, vol. 27, supplement 15, abstract 8539, 2009.

[61] M. Y. Murray, S. A. Rushworth, L. Zaitseva, K. M. Bowles, and D. J. MacEwan, "Attenuation of dexamethasone-induced cell death in multiple myeloma is mediated by miR-125b expression," *Cell Cycle*, vol. 12, no. 13, pp. 2144–2153, 2013.

[62] K. Unno, Y. Zhou, T. Zimmerman, L. C. Platanias, and A. Wickrema, "Identification of a novel microRNA cluster *miR-193b-365* in multiple myeloma," *Leukemia & Lymphoma*, vol. 50, no. 11, pp. 1865–1871, 2009.

[63] C. S. Chim, K. Y. Wong, Y. Qi et al., "Epigenetic inactivation of the miR-34a in hematological malignancies," *Carcinogenesis*, vol. 31, no. 4, pp. 745–750, 2010.

[64] M. Kumar, Z. Lu, A. A. L. Takwi et al., "Negative regulation of the tumor suppressor p53 gene by microRNAs," *Oncogene*, vol. 30, no. 7, pp. 843–853, 2011.

[65] K. H. Lim, A. T. Baines, J. J. Fiordalisi et al., "Activation of RalA is critical for Ras-induced tumorigenesis of human cells," *Cancer Cell*, vol. 7, no. 6, pp. 533–545, 2005.

[66] H. Kameda, M. Watanabe, M. Bohgaki, T. Tsukiyama, and S. Hatakeyama, "Inhibition of NF-κB signaling via tyrosine phosphorylation of Ymer," *Biochemical and Biophysical Research Communications*, vol. 378, no. 4, pp. 744–749, 2009.

[67] A. Farfsing, F. Engel, M. Seiffert et al., "Gene knockdown studies revealed CCDC50 as a candidate gene in mantle cell lymphoma and chronic lymphocytic leukemia," *Leukemia*, vol. 23, no. 11, pp. 2018–2026, 2009.

[68] H. Zhao, S. Jin, and A. M. Gewirtz, "The histone acetyltransferase TIP60 interacts with c-Myb and inactivates its transcriptional activity in human leukemia," *The Journal of Biological Chemistry*, vol. 287, no. 2, pp. 925–934, 2012.

[69] Q. Jiang, Y. Wang, Y. Hao et al., "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, no. 1, pp. D98–D104, 2009.

[70] I. K. Guttilla and B. A. White, "Coordinate regulation of FOXO1 by miR-27a, miR-96, and miR-182 in breast cancer cells," *Journal of Biological Chemistry*, vol. 284, no. 35, pp. 23204–23216, 2009.

[71] S. Lewis-Saravalli, S. Campbell, and A. Claing, "ARF1 controls Rac1 signaling to regulate migration of MDA-MB-231 invasive breast cancer cells," *Cellular Signalling*, vol. 25, no. 9, pp. 1813–1819, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 394260, 13 pages
http://dx.doi.org/10.1155/2015/394260

Research Article

The Construction of Common and Specific Significance Subnetworks of Alzheimer's Disease from Multiple Brain Regions

Wei Kong,[1] Xiaoyang Mou,[2] Na Zhang,[1] Weiming Zeng,[1] Shasha Li,[3] and Yang Yang[4]

[1]*Information Engineering College, Shanghai Maritime University, Shanghai 201306, China*
[2]*DNJ Pharma and Rowan University, Glassboro, NJ 08028, USA*
[3]*Psychology Department, The Second People's Hospital of Guizhou Province, Guiyang 550004, China*
[4]*Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*

Correspondence should be addressed to Wei Kong; weikong@shmtu.edu.cn

Received 28 August 2014; Accepted 7 October 2014

Academic Editor: Tao Huang

Alzheimer's disease (AD) is a progressively and fatally neurodegenerative disorder and leads to irreversibly cognitive and memorial damage in different brain regions. The identification and analysis of the dysregulated pathways and subnetworks among affected brain regions will provide deep insights for the pathogenetic mechanism of AD. In this paper, commonly and specifically significant subnetworks were identified from six AD brain regions. Protein-protein interaction (PPI) data were integrated to add molecular biological information to construct the functional modules of six AD brain regions by Heinz algorithm. Then, the simulated annealing algorithm based on edge weight is applied to predicting and optimizing the maximal scoring networks for common and specific genes, respectively, which can remove the weak interactions and add the prediction of strong interactions to increase the accuracy of the networks. The identified common subnetworks showed that inflammation of the brain nerves is one of the critical factors of AD and calcium imbalance may be a link among several causative factors in AD pathogenesis. In addition, the extracted specific subnetworks for each brain region revealed many biologically functional mechanisms to understand AD pathogenesis.

1. Introduction

Alzheimer's disease (AD) is a complex progressive and irreversible neurodegenerative disease. The characteristic pathology change in AD is the deposition of beta-amyloid (Aβ) and poly-Tau protein in the cell. The pathomorphism features of AD are the senile plaques (SP) and neurofibrillary tangles (NFT), cerebrovascular amyloid, dystrophic neuritis, and loss of synaptic connections [1, 2]. AD is a complex neurodegenerative disorder with largely unknown genetic mechanisms. Lots of transcriptome studies show that AD can lead to the dysfunction of multiple brain regions and progressively destroys remembering, thinking, and reasoning skills [3]. Identifying altered gene expression and molecular mechanism in brain regions differentially affected by AD would represent a significant advance in the genetics of AD.

Most of the current genome-wide studies for AD pathogenesis focus on the hippocampus (HIP) since it is the first and most degraded region in AD brain. However, the changes of gene expression profiles, pathways, and regulatory networks are related to many brain regions which have close relationship to human learning and memory. For example, entorhinal cortex (EC) works as a hub in a widespread network for memory and navigation. The EC-hippocampus system plays an important role in declarative and spatial memories. Posterior cingulate cortex (PC) is a polymodal association area that contributes importantly to normal recognition memory and plays a critical role in visual perception. The functions of middle temporal gyrus (MTG) are associated with brain processes like recognizing familiar faces, ascertaining distance, and understanding meaning of words while reading. Superior frontal gyrus (SFG) is involved

in self-awareness and in coordination with the action of the sensory system. The visual cortex (VCX) of the brain, which is responsible for processing visual information, has shown many changes in aging and AD.

In the past decades, many efforts have been made to explore the differentially expressed genes of different AD-affected brain regions. For instance, Loring et al. found that 118 significant genes were differentially expressed in the amygdala and cingulate cortex [4]. Dunckley et al. found 225 differentially expressed genes up- or downregulated in the early stages of NFT formation by comparing gene expression profiles of NFT-bearing with non-NFT-bearing entorhinal cortex neurons [5]. Liang et al. provided gene expression profiles of six brain regions from the healthy and AD-affected individuals. And then they identified differential expression changes of genes in AD pathogenesis, particularly with regard to tangle and plaque formation [6].

Recently, fast development of statistically computational tools enables large-scale discovery of coregulated gene groups and reveal of functional subnetworks for AD. Ray et al. identified 6 coexpressed gene modules, each of which represented some biological processes perturbed in the HIP of AD brain [7]. By using a weighted gene coexpression network analysis method, Miller et al. identified 12 distinct modules related to synaptic and metabolic processes and immune response in the HIP of AD patients [8]. Ray and Zhang developed a novel differential topological method to identify the coexpression network of four regions, HIP, EC, PC, and MTG from AD-affected brain, and built the topological overlap between them [9]. Liu et al. discovered the relationships among AD related pathways and dysfunctions in the six brain regions and identified the similarities and differences of these dysfunctional pathways by integrating protein-protein interaction (PPI) data [10, 11]. Liang et al. identified hub genes and the significantly perturbed subnetwork closely related to plaques and tangles from six AD-affected brain regions by using a heaviest induced subgraph algorithm with a modular scoring function [12]. Chen et al. identified gene signatures associated with six different brain regions. Functional analyses revealed that the biological processes involved with metabolism, protein ubiquitination, vasculature, and synaptic signaling pathways were dysregulated and perturbed in AD [13].

In short, the common features of the dysfunctional pathways and subnetworks extracted from various AD brain regions can provide cooperativities for potential pathogenesis. On the other hand, the distinct features and the differences among different brain regions may provide more enlightenment for finding the pathogenesis of AD. In this study, both common coexpression and specific subnetworks of six AD-affected brain regions were extracted and analyzed to find the underlying AD pathogenesis. Considering that the gene expression networks have been constructed with poor precision and accuracy due to the inherent shortcomings of small sample size and strong noise, protein-protein interaction (PPI) data was added to provide abundant translation information for extracting significant genes and functional subnetworks in this research. The heuristic algorithms and node-scoring functions were applied to integrating gene expression profiles with PPI network information to find out not only common coexpression networks, but also specific dysregulated pathways for six AD-affected brain regions, including HIP, EC, PC, MTG, SFG, and VCX. Then, simulated annealing algorithm was applied to building and optimizing the functional modules. The molecular biological analysis revealed that the inflammation reaction and calcium ions metabolism constructed by the common genes of six brain regions play important roles in AD pathogenesis. Moreover, the identified specific subnetworks for each brain region revealed many biological pathways perturbed in AD which will lead to greater insight into AD pathogenesis.

2. Methods

2.1. Score Function Principle. Score function can be used as a method for measuring the significance of genes and subnetworks. It includes edge score and node score, of which edge score represents the strength of the correlation between two node-genes and the node score represents the differential significance of each individual gene [12, 14]. The node score function can be given as follows:

$$\text{Score\_}n = \log\left(\frac{ax^{a-1}}{a\tau^{a-1}}\right) = (a-1)\left(\log(x) - \log(\tau(\text{FDR}))\right), \tag{1}$$

where a represents the maximum-likelihood estimation of the shape parameter for the beta-uniform mixture (BUM) model, which indicates that the signal component is equal to the $\beta(a, 1)$ density, x denotes the raw P values, and τ represents the significance threshold, which controls the false discovery rate (FDR) for the positively scoring P values and fine-tunes the discrimination of signal and noise. The raw P values, which are considered as a mixture of signal and noise, can be calculated from the raw gene expression data. By this method, the noise of raw P values can be easily separated since the signal component is assumed to be beta $(a, 1)$ distributed, and the noise is uniform $(0, 1)$ distributed [15].

The values of edge score represent the strength significance of the interaction between genes. Positive score represents activation and negative score represents inhibition. The edge significance score is given as follows:

$$\text{Score\_}e = \text{cov}(X, Y) = \text{corr}(X, Y)\,\text{std}(X)\,\text{std}(Y), \tag{2}$$

where X and Y denote two different genes X and Y, respectively, and $\text{corr}(X, Y)$ denotes the Pearson correlation coefficient of the gene expression profiles of X and Y. The differential expressions of the genes are measured as the overall expression variation ($\text{std}(X)$ and $\text{std}(Y)$). In order to avoid the influence from the number of edges, the edge score function is defined as follows:

$$\text{Score\_}e(G) = \frac{\sum_{e\in E}\text{Score}(e) - \text{avg}_k}{\text{std}_k}, \tag{3}$$

where avg_k denotes the mean of the edge score of the network and std_k represents the standard deviation of edge scores.

2.2. The Algorithm of Identifying Differential Significance Subnetworks. The heaviest induced subgraph algorithm (Heinz) based on the node scoring was applied to our study to find out differentially significant genes and optimal subnetworks from PPI data for different brain regions. The theoretical model of Heinz algorithm belongs to a Steiner-tree problem. The main task of the model is to find an optimal network from a very complex network. In this paper, relevant subnetworks with maximal score are captured from the PPI network with negative and positive scores.

The steps of identifying a significant subnetwork by Heinz algorithm are as follows: firstly, calculate the scores of all the nodes by the score function. Next, define the edge scores based on the node scores connected to the edge. Based on these edge scores, a minimum spanning tree (MST) was calculated. Then, identify all the paths between positive nodes and at the same time the negative nodes involved in these paths were caught. Finally, calculate MST again based on the negative nodes from the obtained maximal significance subnetwork; then, the maximal subnetwork can be finally identified according to the scores of the final positive and negative nodes.

In order to increase the accuracy of the significance subnetwork, in our study, simulated annealing algorithm based on edge scores was applied to removing the weak interactions and enhancing the strong interactions of the calculated significance subnetwork. Guo et al. applied this method to analyzing human prostate cancer and yeast cell cycle. Their results demonstrated that the edge-based method was able to efficiently capture relevant protein interaction behaviors under the investigated conditions [14]. Simulated annealing algorithm is a widely used intelligent optimization algorithm in a number of fields [16]. The modular analysis of biological networks in the bioinformatics research can be considered as a large-scale combinatorial optimization problem essentially. Meanwhile the simulated annealing algorithm is an effective approximation algorithm for solving these kinds of large-scale combinatorial optimization problems with the advantage of avoiding falling into the local optimization.

3. Results and Discussion

3.1. Data and Preprocessing. The gene expression datasets of healthy elders and AD patients we used in this study were downloaded from NCBI GEO Datasets-record of GSE5281. The neurons were collected by laser-capture microdissection from six different brain regions, including HIP, EC, MTG, PC, SFG, and primary visual cortex (VCX). The human GeneChips Affymetrix U133 Plus 2.0 array was used to provide the gene expression data. Each gene chip involved 54675 genes probes for each sample. The datasets consisted of 13 control (normal aging) and 10 AD-affected samples for HIP, the same sample number for EC, 12 control and 16 AD-affected samples for MTG, 13 control and 9 AD-affected samples for PC, 11 control and 23 AD-affected samples for SFG, and 12 control and 19 AD-affected samples for VCX. Moreover, the PPI datasets we utilized in this research are obtained from the Human Protein Reference Database (HPRD) [17], which consisted of 36504 interactions among 9386 genes.

Before searching for differential significance subnetworks with maximal scores, we matched the preprocessed gene expression data with PPI dataset to get the raw interactions of genes (nodes) with the related edges, and the raw P values of all the nodes were calculated as well. Secondly, we processed the gene expression data by gene annotation and variance analysis. For PPI dataset, self-loops and proteins without expression values were removed for simplifying the raw protein interaction networks. Next, the Affymetrix probe set IDs and HPRD gene symbols were mapped to Entrez Gene IDs to extract maximal network. After the preprocessing, there are around 6100–6400 genes left for each brain region.

3.2. Results and Discussion. According to our experiments, adjusting the FDR into different values will obtain different number of genes with positive scores; in order to insure plenty of gene nodes with positive scores, for each brain region, we selected different suitable FDR for the PPI network by which each raw network can contain about 15% positive score nodes. The FDR for HIP, EC, MTG, PC, SFG, and VCX region were set to be 0.008, 0.004, 0.0007, 0.01, 0.06, and 0.09, respectively. The node scores and edge scores for each brain region dataset were calculated. Starting from the positive score nodes, Heinz algorithm was applied to searching for the maximal scoring subnetworks in each brain region. After that the simulated annealing algorithm was applied to optimizing the networks with the threshold value of 0.8. By using the simulated annealing algorithm, the interactions with the edge strength exceeding the threshold were added to the extracted subnetworks, while the weak strengths whose value was less than the threshold were removed from the subnetworks. Six differential significance subnetworks were finally identified for the six brain regions, respectively. Based on that, we carried out the functional enrichment analysis for the identified subnetworks by Gene Ontology (GO) and DAVID [18]. Many known risk genes and pathways were extracted in our results such as APP and GAPDH. Additionally, NF-κB signaling pathway, pathways associated with mitochondria, nerve tissue, calcium ion metabolism, and process of acetylation were also identified and shown to be closely associated with the pathogenetic mechanism of AD.

By observing the genes and interactions of these six identified significance subnetworks, it was found that they were overlapped with each other. The overlapped genes and interactions may suggest that the similarities may play important roles in the dysregulated networks in AD. Figure 1 provided the Venn diagram of the overlap of the significance subnetworks among five brain regions.

From Figure 1 we can see that many significantly expressed genes overlapped between different brain regions. Additionally, many other genes were specifically differentially expressed in each brain region as well. Therefore, the consideration for both common and specific genes and subnetworks were necessary to discover the pathogenetic mechanism of AD.

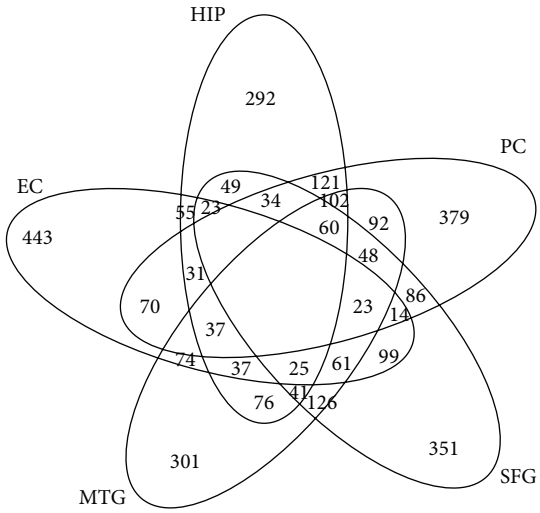


Figure 1: The Venn diagram of five brain regions with significant genes.

3.2.1. Molecular Biological Analysis of Common Functional Subnetworks. For the genes overlapped in overall the six brain regions, we selected a gene as a common gene when the number of interactions with other genes exceeds 90% quantile in different subnetworks. With this criterion, 206 common genes were extracted. It showed that most of the common genes play important roles in different brain regions. The molecular biological analysis revealed that many common genes were functionally related to metabolism, synaptic vesicle-mediated transport, transcriptional regulation, protein kinase phosphorylation, apoptosis, intracellular signaling, and cell cycle. Particularly, two functional subnetworks consisted of the common genes, which closely related to inflammation and calcium imbalance, were found distinctly dysregulated in all of the AD-affected brain regions. Figure 2 showed the inflammatory response subnetwork constructed by the related common genes. Diamonds in Figure 2 denoted the known risk genes of AD, circles presented our extracted common genes related to inflammation, and their different colors indicated different numbers of brain regions in which the gene was upregulated or downregulated. Table 1 provided the KEGG pathway analysis of genes in Figure 2.

From Figure 2 and Table 1 we can see that many regions of the AD brain suffered from inflammation. The degeneration of tissue and the deposition of beta-amyloid (Aβ) and poly-Tau protein are known as the classical stimulants of inflammation [19]. Mitogen-activated protein kinases (MAPKs) are serine-threonine kinases that mediate various types of cellular activities including cell proliferation, differentiation, survival, death, and transformation [20, 21]. The dysregulation of MAPK signaling pathways was found to have involved in many human diseases including AD, Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), and many kinds of cancers [22]. The activation of ERK, JNK, and p38 signaling pathways may lead to neuronal apoptosis in AD [23]. In Figure 2, our results indicated that the crowd MAPK1 and MAKP3 expressions were lower than the normal samples obviously in most of the six AD brain regions.

Our results in Figure 2 also showed that caspase-3, Bcl2, caspase-6, and caspase-8 were overexpressed in most of the six AD brain regions. Caspases are a family of cysteine proteases that plays an important role in apoptosis (programmed cell death), necrosis, and inflammation [24, 25]. The predominant caspase involved in the cleavage of amyloid-beta precursor protein (APP) is suggested to be associated with neuronal death in AD [26]. The Bax gene was the first

Table 1: KEGG pathway analysis results.

Pathway	Number of genes
Pathways in cancer	11
Renal cell carcinoma	6
Neurotrophin signaling pathway	7
Chronic myeloid leukemia	6
Fc epsilon RI signaling pathway	6
Focal adhesion	8
ErbB signaling pathway	6
Prostate cancer	6
Jak-STAT signaling pathway	7
T cell receptor signaling pathway	6
Chemokine signaling pathway	7
Insulin signaling pathway	6
Fc gamma R-mediated phagocytosis	4
GnRH signaling pathway	4
Melanogenesis	4
Toll-like receptor signaling pathway	4
Endometrial cancer	3
Nonsmall cell lung cancer	3
Wnt signaling pathway	4
NOD-like receptor signaling pathway	3
Epithelial cell signaling in *Helicobacter pylori* infection	3
Long-term depression	3
Pancreatic cancer	3
B cell receptor signaling pathway	5
Adherens junction	5
Prion diseases	4
Colorectal cancer	5
Apoptosis	5
Acute myeloid leukemia	4
Glioma	4
Natural killer cell mediated cytotoxicity	5
Long-term potentiation	4
Alzheimer's disease	5
MAPK signaling pathway	6
TGF-beta signaling pathway	3
Gap junction	3

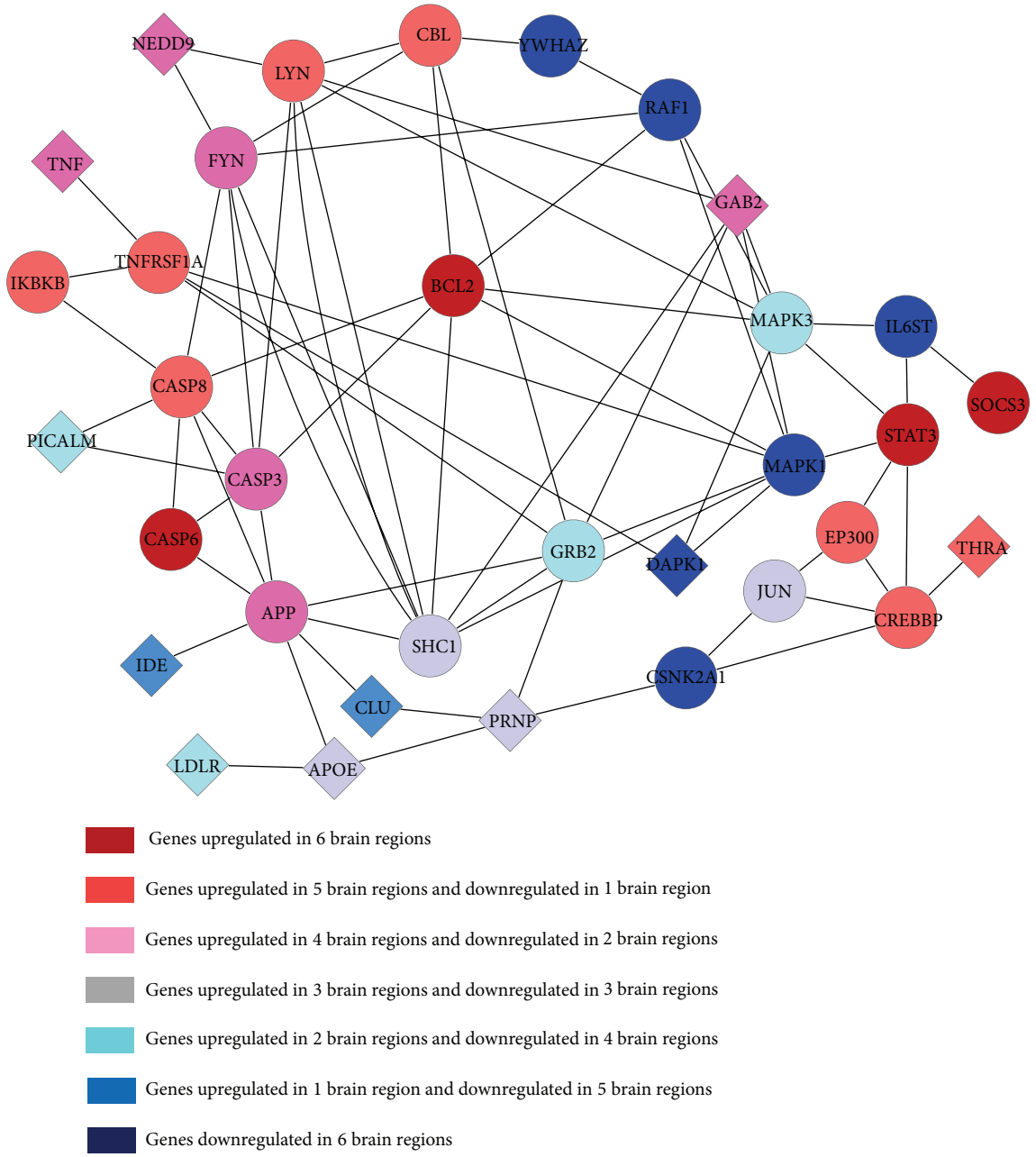


Figure 2: Functional subnetwork of the inflammatory response constructed by common genes.

identified proapoptotic member of the Bcl2 protein family; Bax and caspase-3 are both death effectors in neurodegenerative pathways [27]. It implicates that nervous cells apoptosis will occur. The extracted common genes in our results are in keeping with recently pathological analysis. Furthermore, our results also showed that APP was overexpressed in AD samples in all of the six brain regions. The overexpression of APP would activate caspase-3 in human postmitotic neurons and cause the degeneration of postmitotic neurons in AD [28].

The SOCS (suppressors of cytokine signaling) family of proteins has a dual identity; they are the inhibitors of JAK (Janus kinase) signal transducer and the activator of the STAT signaling pathways as well [29]. The subtype SOCS3 has an important function in inhibiting some inflammatory genes, cytokine signaling, and STAT3 activation. The STAT3 has dual function in different types of cells; on the one hand, it promotes proliferation and prevents apoptosis; on the other hand, it can induce growth arrest and apoptosis [30, 31]. Our data exhibited that SOCS3 and STAT3 were both overexpressed in six AD brain regions.

Furthermore, TNFRSF1A protein works as a regulator of inflammation, and as a receptor of TNF-α (tumor necrosis factor-alpha) it can activate NF-κB (nuclear factor

Table 2: KEGG pathway analysis results.

Pathway	Number of genes
ErbB signaling pathway	7
Gap junction	7
Calcium signaling pathway	8
Focal adhesion	8
Glioma	6
GnRH signaling pathway	6
Nonsmall cell lung cancer	5
Natural killer cell mediated cytotoxicity	5
Tight junction	5
Vibrio cholerae infection	4
Chemokine signaling pathway	5
Pathways in cancer	6
Melanogenesis	4
Leukocyte transendothelial migration	4
Neurotrophin signaling pathway	4
MAPK signaling pathway	5
Insulin signaling pathway	4
Wnt signaling pathway	4
Endometrial cancer	3
Epithelial cell signaling in *Helicobacter pylori* infection	3
Phosphatidylinositol signaling system	4
VEGF signaling pathway	4
Fc epsilon RI signaling pathway	4
Long-term potentiation	3
Adherens junction	3
Colorectal cancer	3
Prostate cancer	3
Fc gamma R-mediated phagocytosis	3
Vascular smooth muscle contraction	3
Dorsoventral axis formation	2

kappa-light-chain-enhancer of activated B cells) and mediate apoptosis [32]. Our data show that TNFRSF1A and TNF-α were both overexpressed. The results confirmed that the six AD brain regions closely associated with inflammation and apoptosis.

In addition to inflammatory response, another important subnetwork we found from the common genes was calcium ion metabolism subnetwork. Figure 3 showed the subnetwork of the calcium ion mechanism with the extracted common genes, and the KEGG pathway analysis of this subnetwork was provided in Table 2.

In AD-affected brain regions, the calcium ion metabolism related signaling were found dysregulated. It was reported that calcium can modulate many neural processes, including synaptic plasticity and apoptosis. With the increase of the intracellular calcium, the accumulation of amyloid-β, the hyperphosphorylation of Tau, and neuronal death will occur in the affected brain regions [33]. Particularly, in this subnetwork, our results showed that APP was greatly upregulated in HIP, PC, MTG, and VCX. In addition, Annexin is a Ca (2+)-effector protein which plays an important role in the metabolism of intracellular Ca (2+) [34]. Annexin A2 is a calcium-dependent phospholipid-binding protein whose function is to help organize exocytosis of intracellular proteins to the extracellular domain. Figure 3 showed that ATP2A2 and ATP2B2 were lower expressed than normal samples in overall the six AD brain regions. ATP2A2 and ATP2B2 are enzymes that can remove bivalent calcium ions from eukaryotic cells against very large concentration gradients and play a critical role in intracellular calcium homeostasis [35, 36].

3.2.2. Specific Subnetworks in Each Brain Region. Figure 1 revealed that many significant genes were specially expressed in different brain regions. It suggested that some specific dysregulated pathways and subnetworks among them will provide deep insights into the pathogenetic mechanism of AD. By getting rid of the common genes overlapped in different brain regions, the maximal scoring function and the simulated annealing method were used again to construct the specific functional subnetwork by the specifically differential significant genes for each brain region including HIP, EC, PC, MTG, and SFG. Therefore, the sizes of the constructed specific subnetworks are much smaller. Since there were not enough significant genes that can be discovered to construct any functional subnetwork, the result of primary visual cortex (VCX) was absence. Figures 4–8 showed the specific functional subnetworks in HIP, EC, PC, MTG, and SFG, respectively. In Figures 4–8, red circles represented the genes upregulated in this brain region for AD samples, blue circles denoted the genes downregulated, and grey ones denoted that this gene had no great changes compared with normal samples in this brain area.

For HIP, among the specific significant genes, 33 genes can be constructed to the maximal scoring subnetwork (Figure 4). It is known that for AD patients the hippocampus is one of the first regions of the brain to suffer damage including memory loss and disorientation. In this region, some genes were specially differentially overexpressed, such as CAPN1, FXR1, GRIN2B, ITPR1, KDR, KIAA1377, NBN, PRKG1, RUNX1T1, SGSM2, SREBF2, TAF15, and U2AF2. CAPN1 (calcium-activated neutral protease) is a kind of nonlysosomal intracellular cysteine protease. The overexpression of CAPN1 has a relationship with intractable epilepsy as well as the clinicopathological characteristics in AD patients [37]. It was interesting to note that kinase 1 was low expressed in AD, but it was known to be high expressed in cancer.

In EC area, 40 specially expressed genes were constructed to the maximal scoring subnetwork (Figure 5). The changes in the Tau protein and the cleaved fragments of APP have been found in EC region in the early stages of AD [38]. In this brain area, our data showed that BRAF, C21orf91, CBL, CSF2RB, LYN, MDM2, SLA, and KANK1 were all overexpressed. BRAF, as a member of the RAF kinase family of growth signal transduction protein kinases, can affect cell

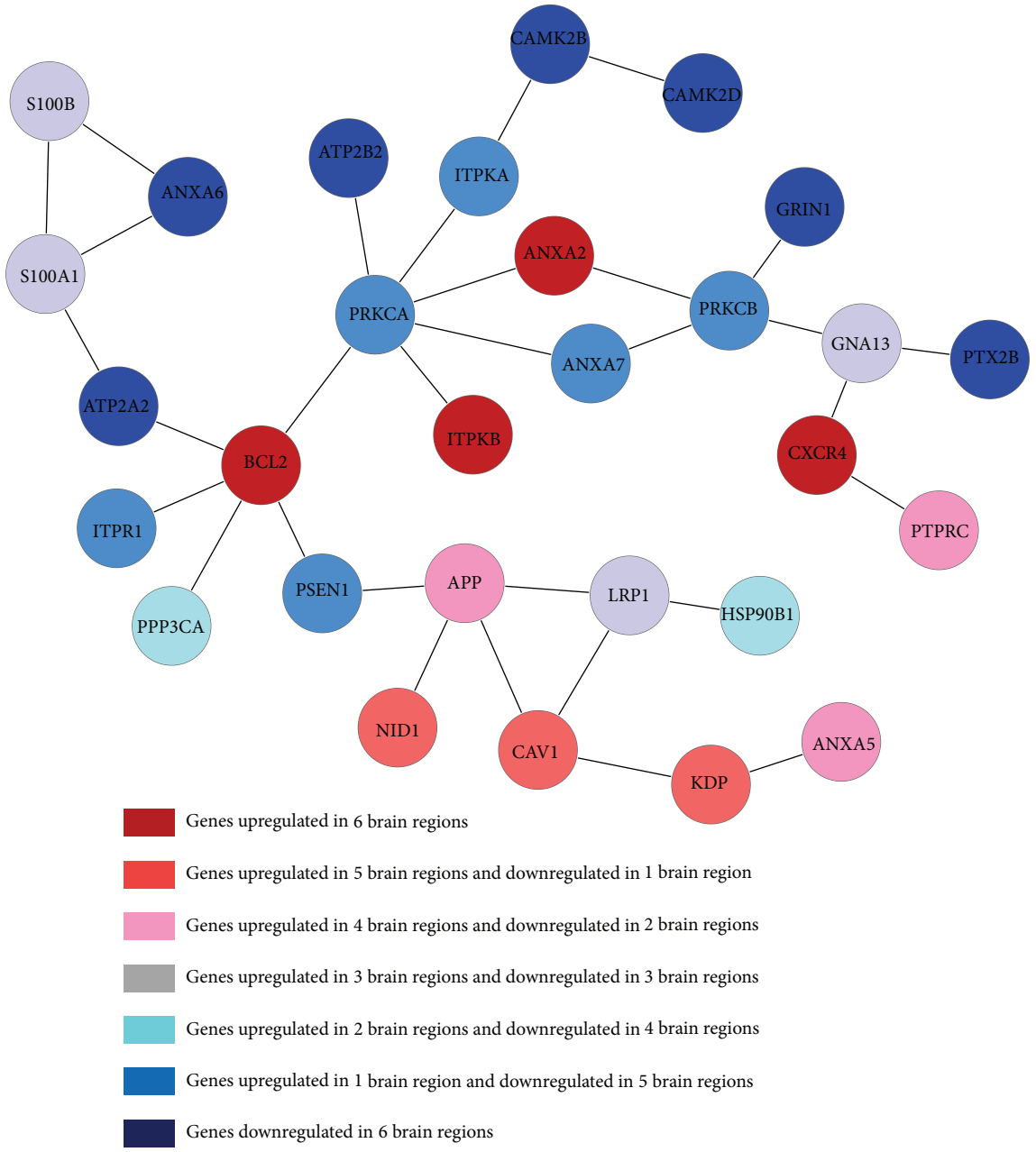


Figure 3: Functional subnetwork of the calcium ion mechanism constructed by common genes.

division, differentiation, and secretion by regulating the MAP kinase/ERKs signaling pathway [39]. Members of the Casitas B-lineage lymphoma (Cbl) protein family are evolutionarily conserved multidomain regulators of signal transduction. Colony stimulating factor 2 receptor β (CSF2RB) is a risk factor in both schizophrenia and major depression since the overexpression of it has relationship with the disturbance of nerve signal conduction [40]. AKT1 is a survival factor and the activated AKT1 plays important roles in inhibiting apoptosis and promoting cell survival [41, 42]. Our results showed that AKT1 was low expressed in EC of AD patients which indicated that apoptosis was happening.

Figure 6 shows the maximal scoring subnetwork of PC area constructed by 16 specifically significant genes. PC is a polymodal association area that contributes importantly to normal recognition memory and plays a critical role in visual perception [43, 44]. From Figure 6 we can see that genes CASP3, CASP6, CDKN1A, FLNA, ITGB1, MAPT, PCBP2, PRKACA, and SET were overexpressed in PC area of AD brain. The genes CASP3 and CASP6 are related to apoptosis. CDKN1A (p21) has a function of regulating cell cycle by inhibiting the activity of cyclin-CDK2, cyclin-CDK1, and cyclin-CDK4/6 complexes and it activates CDK2; thus, it leads to apoptosis [45, 46]. The Tau proteins are the product

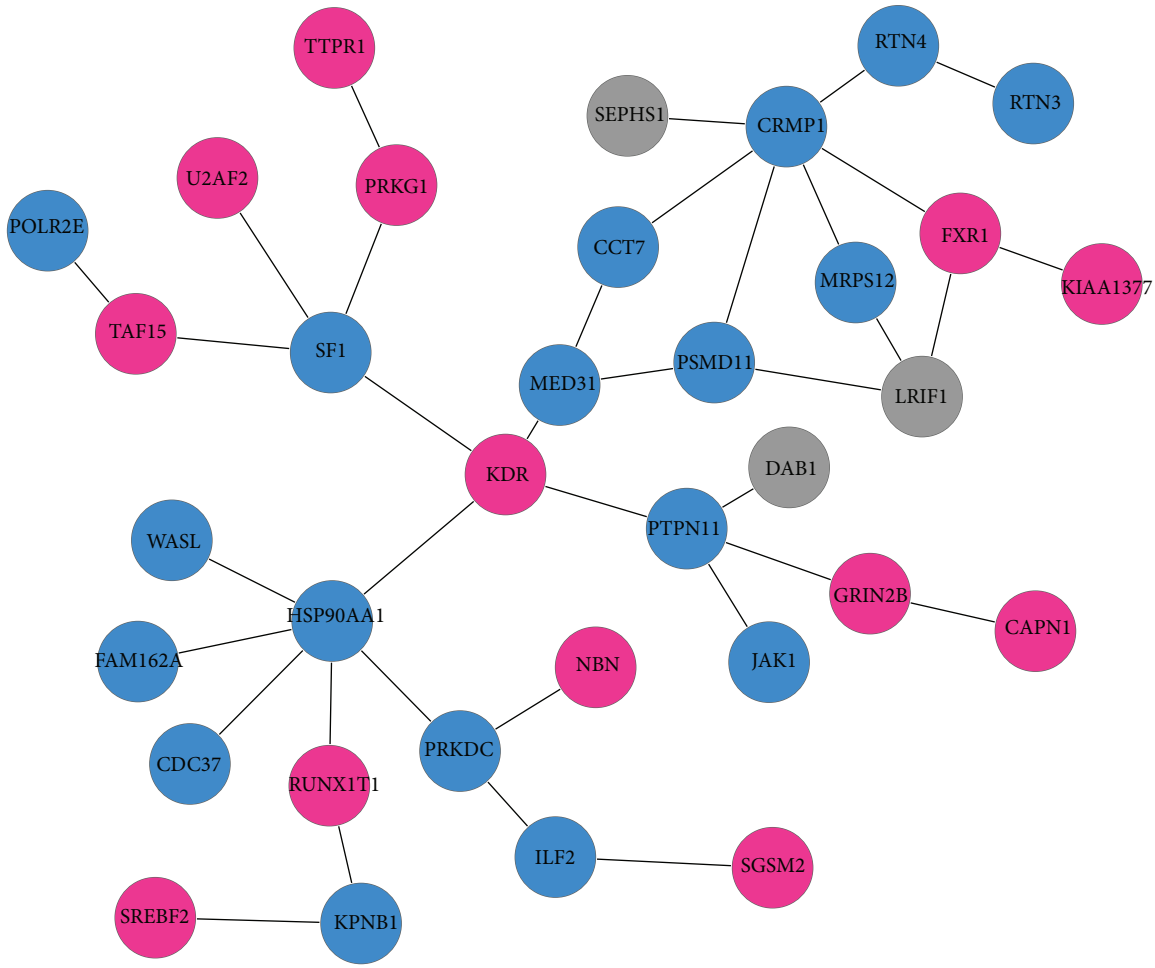


Figure 4: Specific subnetwork of HIP.

of alternative splicing from a single gene that is designated MAPT (microtubule-associated protein tau) in humans, and the overexpression of Tau will impact a neuroinflammation gene expression network perturbed in AD [2, 47, 48]. The major human AP endonuclease APE1 are reported to play an important role in the base excision repair (BER) pathway [49]; however, they were found to be low expressed in PC area of AD brains.

In MTG area, 24 specifically significant genes were extracted to achieve a maximal scoring subnetwork (Figure 7). The functions of MTG are associated with brain processes like recognizing familiar faces, ascertaining distance, and understanding meaning of words while reading. Our data exhibited that ANTXR1, BRCA1, CCND1, GATA2, KMT2D, NEDD9, PITPNM3, PLCG2, PRTFDC1, RB1CC1, SMAD1, and STAT5A were overexpressed. The ANTXR1 is a member of the aldo/keto reductase superfamily, which consists of more than 40 known enzymes and proteins. Aldose reductase contributes to diabetes-mediated mitochondrial dysfunction and damage through the activation of p53. The degree of mitochondrial dysfunction and damage determines whether hyperactivity (mild damage) or apoptosis (severe damage) will ensue [50]. BRCA1 is part of a complex that repairs double-strand breaks in DNA [51]; the overexpression of BRCA1 may suggest that DNA damage is serious; but BRCA1 mutation carriers are at an increased risk of prostate and breast cancer [52]. CCNDBP1 belongs to cyclin D family. The expression of cyclin D suggests that act to link growth factor signals with cell cycle transitions during G1 [53].

For the SFG area, there were 15 specifically significant genes that can be used to construct a maximal scoring subnetwork (Figure 8). SFG is involved in self-awareness and in coordination with the action of the sensory system [54]. In this region, genes CAV1, CDH5, EDNRB, GJB2, JUP, GJB6, and PPAP2B were found overexpressed in AD brains. The CDH5 gene is a classical cadherin from the cadherin superfamily, which provides a molecular system reflecting both early embryonic and mature nervous system architecture. In AD crowd, overexpression of cadherins may be related to restoration of neural epithelium [55, 56]. The PTPN6, which is a member of the PTP (protein tyrosine phosphatase) family, was downregulated in SFG area of AD brain. PTP family is reported to have the ability to

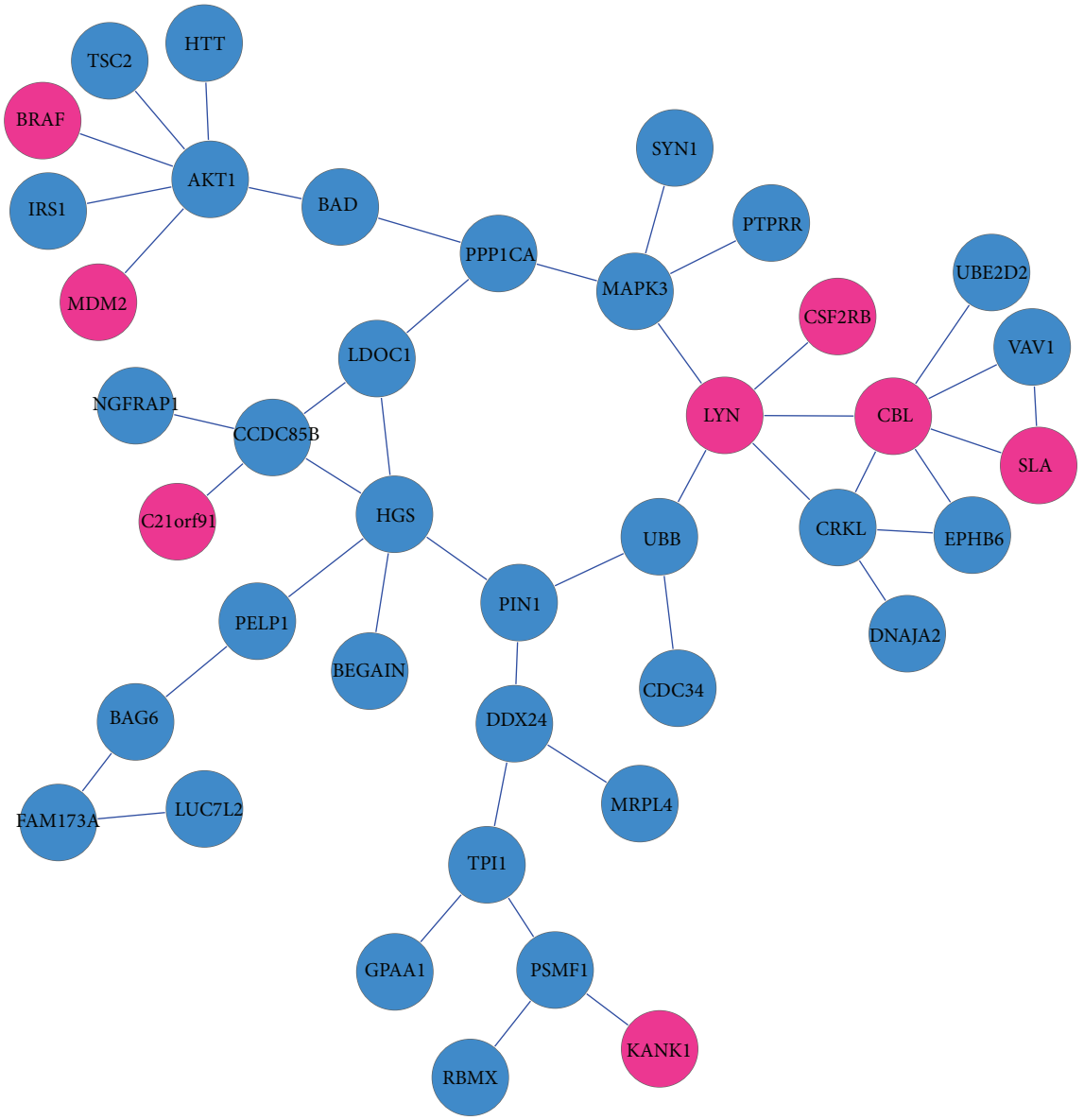


Figure 5: Specific subnetwork of EC.

regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation [57, 58].

4. Conclusions

AD progression is known to occur in many brain regions which have close relationship of human learning and memory with particular features. Discovering the common and specific changes and dysregulated pathways of different brain regions will provide deep insights for finding of the pathogenesis of AD. In this study, we applied a method of scoring function and simulated annealing algorithm to constructing and optimizing the differential significance subnetworks for six brain regions including hippocampus (HIP), entorhinal cortex (EC), middle temporal gyrus (MTG), posterior cingulate cortex (PC), superior frontal gyrus (SFG), and primary visual cortex (VCX). The common genes we identified from overall the six brain regions revealed two significant functional subnetworks which showed that the dysregulation of inflammation and calcium metabolism play important roles in the onset and deterioration of AD. For example, the dysregulated MAPK signaling pathways and JNK or p38 signaling pathways, as parts of inflammation subnetwork, were demonstrated to be associated with many cellular activities and neuronal apoptosis in AD. Many common genes such as caspases, SOCS3, STAT3, TNFRSF1A, and TNF-α were

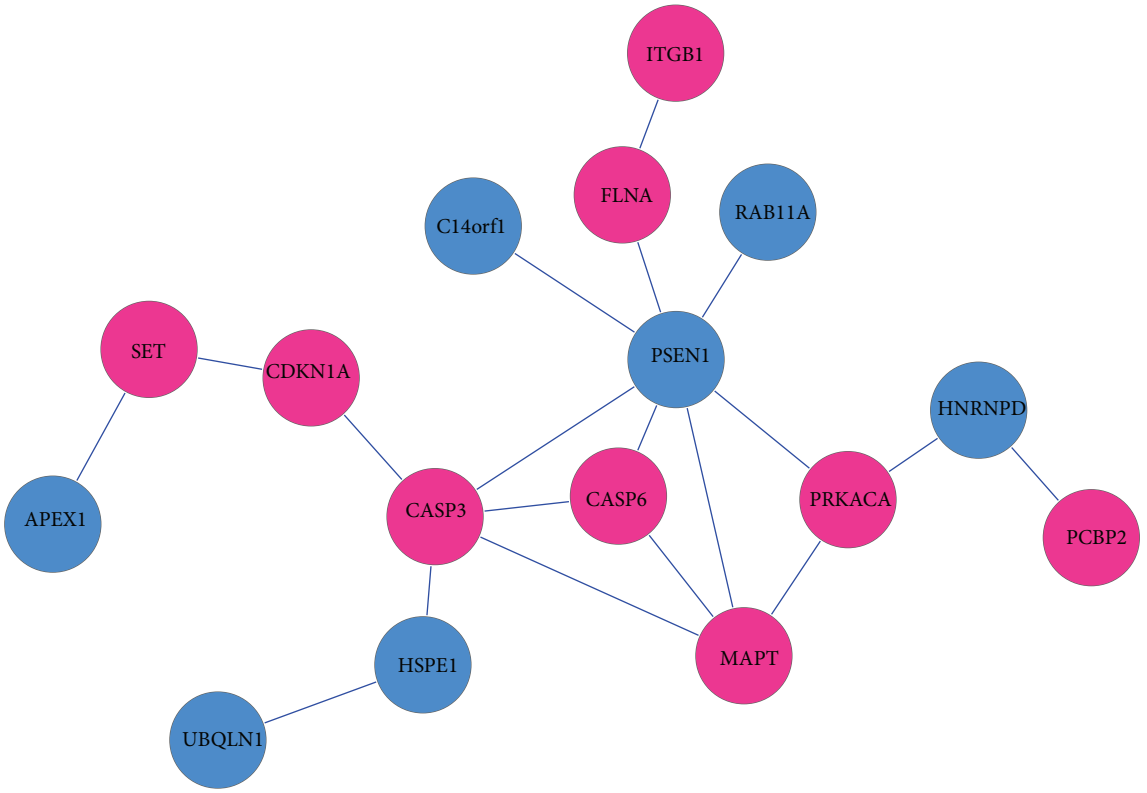


Figure 6: Specific subnetwork of PC.

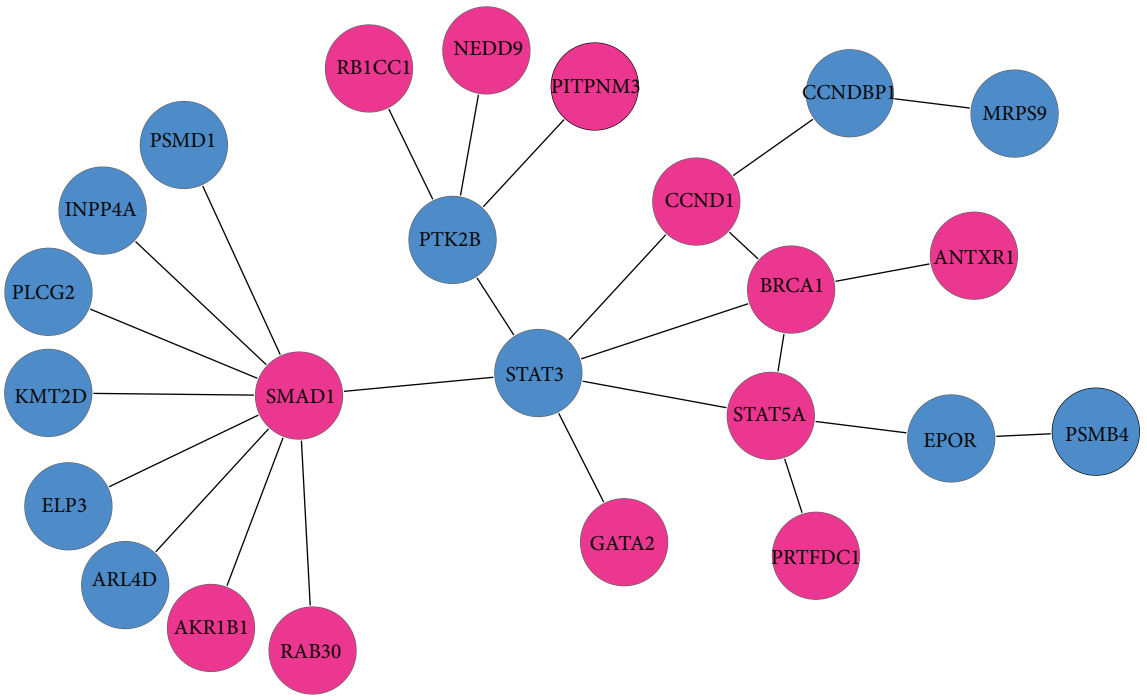


Figure 7: Specific subnetwork of MTG.

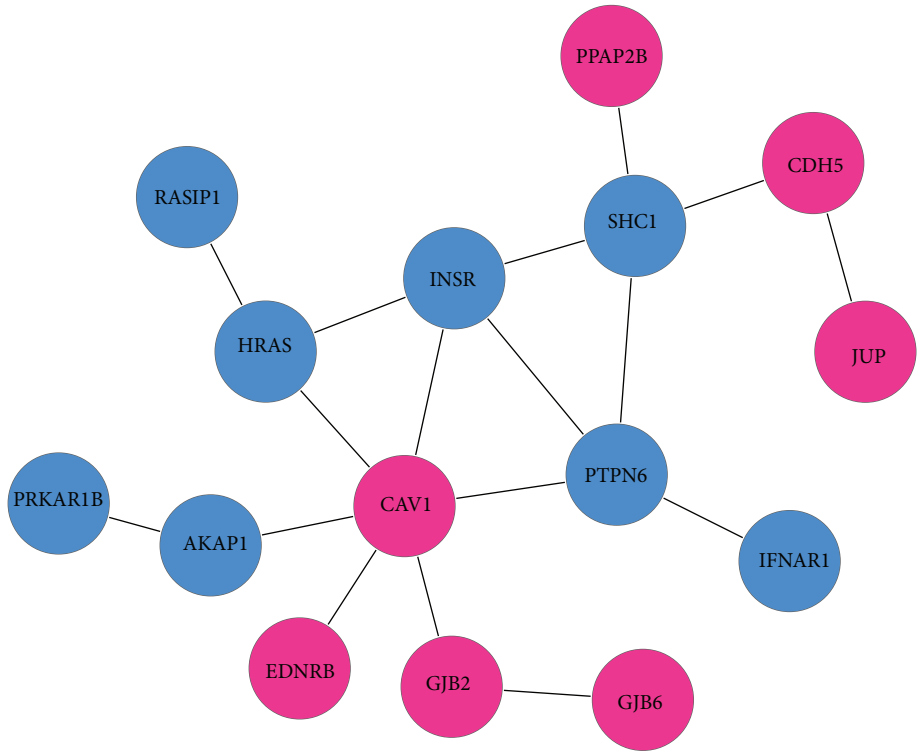


Figure 8: Specific subnetwork of SFG.

confirmed to have close association with inflammation and apoptosis in all the six AD brain regions. In calcium ion mechanism subnetwork, many genes such as ATP2A2 and ATP2B2 played a critical role in intracellular calcium homeostasis. The dysregulation of intracellular calcium would lead to the accumulation of amyloid-β, hyperphosphorylation of Tau, and neuronal death which are parts of the known pathogenesis of AD.

Although we highlighted the contributions of the inflammation and calcium imbalance subnetworks as common features for six brain regions, this paper illustrated the specific dysregulated subnetworks in each AD-affected brain region for HIP, EC, PC, MTG, and SFG. In the subnetwork of HIP, many differentially expressed genes were identified as the clinicopathological characteristics of AD. EC, as another area early affected by AD, was characterized by changes in the Tau protein and APP. Many significant genes in EC played central roles in regulating the MAP kinase/ERKs signaling pathway and affected cell division, differentiation, and secretion. The specific expressed genes in PC area showed close relationship of apoptosis and cell cycle progression. Particularly, the overexpression of Tau will impact a neuroinflammation gene expression network perturbed in AD. The specifically expressed genes in the subnetwork in MTG area of AD showed that they effected the mitochondrial dysfunction and DNA damage. The significant genes related to cadherins in SFG area suggested that this area may contribute to the formation and maintenance of segmental and functional nervous system structures.

In summary, our molecular biological analysis demonstrated that the identified common and specific maximal scoring subnetworks help in comparing biological phenomena across AD-affected brain regions and obtaining a global overview of the disease, which can enable us to further understand the pathogenetic mechanism of AD.

Conflict of Interests

The authors declare no conflict of interests.

Acknowledgment

This work was supported by the National Natural Science Foundations of China (no. 61271446, 31170952, and 61003093).

References

[1] M. Meyer-Luehmann, T. L. Spires-Jones, C. Prada et al., "Rapid appearance and local toxicity of amyloid-β plaques in a mouse model of Alzheimer's disease," *Nature*, vol. 451, no. 7179, pp. 720–725, 2008.

[2] P. D. Wes, A. Easton, J. Corradi et al., "Tau overexpression impacts a neuroinflammation gene expression network perturbed in Alzheimer's disease," *PLoS ONE*, vol. 9, no. 8, Article ID e106050, 2014.

[3] C. Hock, K. Heese, C. Hulette, C. Rosenberg, and U. Otten, "Region-specific neurotrophin imbalances in Alzheimer disease: decreased levels of brain-derived neurotrophic factor and increased levels of nerve growth factor in hippocampus and cortical areas," *Archives of Neurology*, vol. 57, no. 6, pp. 846–851, 2000.

[4] J. F. Loring, X. Wen, J. M. Lee, J. Seilhamer, and R. Somogyi, "A gene expression profile of Alzheimer's disease," *DNA and Cell Biology*, vol. 20, no. 11, pp. 683–695, 2001.

[5] T. Dunckley, T. G. Beach, K. E. Ramsey et al., "Gene expression correlates of neurofibrillary tangles in Alzheimer's disease," *Neurobiology of Aging*, vol. 27, no. 10, pp. 1359–1371, 2006.

[6] W. S. Liang, T. Dunckley, T. G. Beach et al., "Altered neuronal gene expression in brain regions differentially affected by Alzheimer's disease: a reference data set," *Physiological Genomics*, vol. 33, no. 2, pp. 240–256, 2008.

[7] M. Ray, J. Ruan, and W. Zhang, "Variations in the transcriptome of Alzheimer's disease reveal molecular networks involved in cardiovascular diseases," *Genome Biology*, vol. 9, no. 10, article R148, 2008.

[8] J. A. Miller, M. C. Oldham, and D. H. Geschwind, "A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging," *The Journal of Neuroscience*, vol. 28, no. 6, pp. 1410–1420, 2008.

[9] M. Ray and W. Zhang, "Analysis of Alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks," *BMC Systems Biology*, vol. 4, article 136, 2010.

[10] Z.-P. Liu, Y. Wang, X.-S. Zhang, and L. Chen, "Identifying dysfunctional crosstalk of pathways in various regions of Alzheimer's disease brains," *BMC Systems Biology*, vol. 4, no. 2, article 11, 2010.

[11] Z.-P. Liu, Y. Wang, X.-S. Zhang, W. Xia, and L. Chen, "Detecting and analyzing differentially activated pathways in brain regions of Alzheimer's disease patients," *Molecular BioSystems*, vol. 7, no. 5, pp. 1441–1452, 2011.

[12] D. Liang, G. Han, X. Feng, J. Sun, Y. Duan, and H. Lei, "Concerted perturbation observed in a hub network in Alzheimer's disease," *PLoS ONE*, vol. 7, no. 7, Article ID e40498, 2012.

[13] F. Chen, Q. Guan, Z.-Y. Nie, and L.-J. Jin, "Gene expression profile and functional analysis of Alzheimer's disease," *American Journal of Alzheimer's Disease and other Dementias*, vol. 28, no. 7, pp. 693–701, 2013.

[14] Z. Guo, L. Wang, Y. Li et al., "Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network," *Bioinformatics*, vol. 23, no. 16, pp. 2121–2128, 2007.

[15] M. T. Dittrich, G. W. Klau, A. Rosenwald, T. Dandekar, and T. Müller, "Identifying functional modules in protein-protein interaction networks: an integrated exact approach," *Bioinformatics*, vol. 24, no. 13, pp. i223–i231, 2008.

[16] D. W. Ding, P. Yang, and X. H. Wu, "Application of simulated annealing algorithm to biological network research," *Computers and Applied Chemistry*, vol. 28, no. 10, pp. W1302–W1304, 2011.

[17] W. E. Müller, A. Eckert, C. Kurz, G. P. Eckert, and K. Leuner, "Mitochondrial dysfunction: common final pathway in brain aging and alzheimer's disease-therapeutic aspects," *Molecular Neurobiology*, vol. 41, no. 2-3, pp. 159–171, 2010.

[18] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[19] H. Akiyama, S. Barger, S. Barnum et al., "Inflammation and Alzheimer's disease," *Neurobiology of Aging*, vol. 21, no. 3, pp. 383–421, 2000.

[20] J. A. McCubrey, M. M. Lahair, and R. A. Franklin, "Reactive oxygen species-induced activation of the MAP kinase signaling pathways," *Antioxidants & Redox Signaling*, vol. 8, no. 9-10, pp. 1775–1789, 2006.

[21] S. Torii, T. Yamamoto, Y. Tsuchiya, and E. Nishida, "ERK MAP kinase in G1 cell cycle progression and cancer," *Cancer Science*, vol. 97, no. 8, pp. 697–702, 2006.

[22] A. S. Dhillon, S. Hagan, O. Rath, and W. Kolch, "MAP kinase signalling pathways in cancer," *Oncogene*, vol. 26, no. 22, pp. 3279–3290, 2007.

[23] E. K. Kim and E.-J. Choi, "Pathological roles of MAPK signaling pathways in human diseases," *Biochimica et Biophysica Acta—Molecular Basis of Disease*, vol. 1802, no. 4, pp. 396–405, 2010.

[24] E. S. Alnemri, D. J. Livingston, D. W. Nicholson et al., "Human ICE/CED-3 protease nomenclature," *Cell*, vol. 87, no. 2, p. 171, 1996.

[25] D. R. McIlwain, T. Berger, and T. W. Mak, "Caspase functions in cell death and disease," *Cold Spring Harbor Perspectives in Biology*, vol. 5, no. 4, Article ID a008656, 2013.

[26] N. Bulat and C. Widmann, "Caspase substrates and neurodegenerative diseases," *Brain Research Bulletin*, vol. 80, no. 4-5, pp. 251–267, 2009.

[27] M. J. Chong, M. R. Murray, E. C. Gosink et al., "Atm and Bax cooperate in ionizing radiation-induced apoptosis in the central nervous system," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 2, pp. 889–894, 2000.

[28] T. Uetsuki, K. Takemoto, I. Nishimura et al., "Activation of neuronal caspase-3 by intracellular accumulation of wild- type Alzheimer amyloid precursor protein," *Journal of Neuroscience*, vol. 19, no. 16, pp. 6955–6964, 1999.

[29] D. L. Krebs and D. J. Hilton, "SOCS: physiological suppressors of cytokine signaling," *Journal of Cell Science*, vol. 113, part 16, pp. 2813–2819, 2000.

[30] Y. Lu, S. Fukuyama, R. Yoshida et al., "Loss of SOCS3 gene expression converts STAT3 function from anti-apoptotic to pro-apoptotic," *The Journal of Biological Chemistry*, vol. 281, no. 48, pp. 36683–36690, 2006.

[31] H. Qin, W.-I. Yeh, P. de Sarno et al., "Signal transducer and activator of transcription-3/suppressor of cytokine signaling-3 (STAT3/SOCS3) axis in myeloid cells regulates neuroinflammation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 13, pp. 5004–5009, 2012.

[32] H. Li and X. Lin, "Positive and negative signaling components involved in TNFα-induced NF-κB activation," *Cytokine*, vol. 41, no. 1, pp. 1–8, 2008.

[33] F. M. LaFerla, "Calcium dyshomeostasis and intracellular signalling in Alzheimer's disease," *Nature Reviews Neuroscience*, vol. 3, no. 11, pp. 862–872, 2002.

[34] V. Gerke, C. E. Creutz, and S. E. Moss, "Annexins: linking Ca^2+ signalling to membrane dynamics," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 6, pp. 449–461, 2005.

[35] L. Hadri, R. G. Kratlian, L. Benard et al., "Therapeutic efficacy of AAV1.SERCA2a in monocrotaline-induced pulmonary arterial hypertension," *Circulation*, vol. 128, no. 5, pp. 512–523, 2013.

[36] E. M. Lynes, A. Raturi, M. Shenkman et al., "Palmitoylation is the switch that assigns calnexin to quality control or ER Ca^{2+} signaling," *Journal of Cell Science*, vol. 126, no. 17, pp. 3893–3903, 2013.

[37] Z.-H. Feng, J. Hao, L. Ye et al., "Overexpression of μ-calpain in the anterior temporal neocortex of patients with intractable epilepsy correlates with clinicopathological characteristics," *Seizure*, vol. 20, no. 5, pp. 395–401, 2011.

[38] U. A. Khan, L. Liu, F. A. Provenzano et al., "Molecular drivers and cortical spread of lateral entorhinal cortex dysfunction in preclinical Alzheimer's disease," *Nature Neuroscience*, vol. 17, no. 2, pp. 304–311, 2014.

[39] G. Daum, I. Eisenmann-Tappe, H. W. Fries, J. Troppmair, and U. R. Rapp, "The ins and outs of Raf kinases," *Trends in Biochemical Sciences*, vol. 19, no. 11, pp. 474–480, 1994.

[40] P. Chen, K. Huang, G. Zhou et al., "Common SNPs in CSF2RB are associated with major depression and schizophrenia in the Chinese Han population," *World Journal of Biological Psychiatry*, vol. 12, no. 3, pp. 233–238, 2011.

[41] T. F. Franke, S.-I. Yang, T. O. Chan et al., "The protein kinase encoded by the Akt proto-oncogene is a target of the PDGF-activated phosphatidylinositol 3-kinase," *Cell*, vol. 81, no. 5, pp. 727–736, 1995.

[42] H. Dudek, S. R. Datta, T. F. Franke et al., "Regulation of neuronal survival by the serine-threonine protein kinase Akt," *Science*, vol. 275, no. 5300, pp. 661–665, 1997.

[43] W. A. Suzuki, "The anatomy, physiology and functions of the perirhinal cortex," *Current Opinion in Neurobiology*, vol. 6, no. 2, pp. 179–186, 1996.

[44] R. R. Hampton, "Monkey perirhinal cortex is critical for visual memory, but not for visual perception: reexamination of the behavioural evidence from monkeys," *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, vol. 58, no. 3-4, pp. 283–299, 2005.

[45] A. L. Gartel and S. K. Radhakrishnan, "Lost in transcription: p21 repression, mechanisms, and consequences," *Cancer Research*, vol. 65, no. 10, pp. 3980–3985, 2005.

[46] Y. H. Jin, K. J. Yoo, Y. H. Lee, and S. K. Lee, "Caspase 3-mediated cleavage of p21(WAF1/CIP1) associated with the cyclin A-cyclin-dependent kinase 2 complex is a prerequisite for apoptosis in SK-HEP-1 cells," *The Journal of Biological Chemistry*, vol. 275, no. 39, pp. 30256–30263, 2000.

[47] M. Goedert, C. M. Wischik, R. A. Crowther, J. E. Walker, and A. Klug, "Cloning and sequencing of the cDNA encoding a core protein of the paired helical filament of Alzheimer disease: Identification as the microtubule-associated protein tau," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, no. 11, pp. 4051–4055, 1988.

[48] M. Goedert, M. G. Spillantini, R. Jakes, D. Rutherford, and R. A. Crowther, "Multiple isoforms of human microtubule-associated protein tau: sequences and localization in neurofibrillary tangles of Alzheimer's disease," *Neuron*, vol. 3, no. 4, pp. 519–526, 1989.

[49] A. E. Vidal, S. Boiteux, I. D. Hickson, and J. P. Radicella, "XRCC1 coordinates the initial and late stages of DNA abasic site repair through protein-protein interactions," *The EMBO Journal*, vol. 20, no. 22, pp. 6530–6539, 2001.

[50] W. H. Tang, J. Stitham, Y. Jin et al., "Aldose reductase-mediated phosphorylation of p53 leads to mitochondrial dysfunction and damage in diabetic platelets," *Circulation*, vol. 129, no. 15, pp. 1598–1609, 2014.

[51] I. H. Ismail, R. Davidson, J. P. Gagné et al., "Germline mutations in BAP1 impair its function in DNA double-strand break repair," *Cancer Research*, vol. 74, no. 16, pp. 4282–4294, 2014.

[52] A. Liede, B. Y. Karlan, and S. A. Narod, "Cancer risks for male carriers of germline mutations in BRCA1 or BRCA2: a review of the literature," *Journal of Clinical Oncology*, vol. 22, no. 4, pp. 735–742, 2004.

[53] C. J. Sherr, H. Matsushime, and M. F. Roussel, "Regulation of CYL/cyclin D genes by colony-stimulating factor 1," *Ciba Foundation Symposium*, vol. 170, pp. 209–219, 1992.

[54] I. I. Goldberg, M. Harel, and R. Malach, "When the brain loses its self: prefrontal inactivation during sensorimotor processing," *Neuron*, vol. 50, no. 2, pp. 329–339, 2006.

[55] A. Graziani, M. Poteser, W.-M. Heupel et al., "Cell-cell contact formation governs Ca^{2+} signaling by TRPC4 in the vascular endothelium: evidence for a regulatory TRPC4-β-catenin interaction," *The Journal of Biological Chemistry*, vol. 285, no. 6, pp. 4213–4223, 2010.

[56] C. Redies and M. Takeichi, "Cadherins in the developing central nervous system: an adhesive code for segmental and functional subdivisions," *Developmental Biology*, vol. 180, no. 2, pp. 413–423, 1996.

[57] R. Cao, Q. Ding, P. Li et al., "SHP1-mediated cell cycle redistribution inhibits radiosensitivity of non-small cell lung cancer," *Radiation Oncology*, vol. 8, no. 1, article 178, 2013.

[58] L. Sooman, S. Ekman, G. Tsakonas et al., "PTPN6 expression is epigenetically regulated and influences survival and response to chemotherapy in high-grade gliomas," *Tumor Biology*, vol. 35, no. 5, pp. 4479–4488, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 598129, 9 pages
http://dx.doi.org/10.1155/2014/598129

Research Article

Large-Scale Protein-Protein Interactions Detection by Integrating Big Biosensing Data with Computational Model

Zhu-Hong You,[1] Shuai Li,[2] Xin Gao,[3] Xin Luo,[2] and Zhen Ji[1]

[1] *College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China*
[2] *Department of Computing, Hong Kong Polytechnic University, Hong Kong*
[3] *Department of Medical Imaging, Suzhou Institute of Biomedical Engineering and Technology, Suzhou, Jiangsu 215163, China*

Correspondence should be addressed to Shuai Li; shuaili@polyu.edu.hk and Xin Gao; gaox@sibet.ac.cn

Received 23 June 2014; Accepted 24 July 2014; Published 18 August 2014

Academic Editor: Jiangning Song

Protein-protein interactions are the basis of biological functions, and studying these interactions on a molecular level is of crucial importance for understanding the functionality of a living cell. During the past decade, biosensors have emerged as an important tool for the high-throughput identification of proteins and their interactions. However, the high-throughput experimental methods for identifying PPIs are both time-consuming and expensive. On the other hand, high-throughput PPI data are often associated with high false-positive and high false-negative rates. Targeting at these problems, we propose a method for PPI detection by integrating biosensor-based PPI data with a novel computational model. This method was developed based on the algorithm of extreme learning machine combined with a novel representation of protein sequence descriptor. When performed on the large-scale human protein interaction dataset, the proposed method achieved 84.8% prediction accuracy with 84.08% sensitivity at the specificity of 85.53%. We conducted more extensive experiments to compare the proposed method with the state-of-the-art techniques, support vector machine. The achieved results demonstrate that our approach is very promising for detecting new PPIs, and it can be a helpful supplement for biosensor-based PPI data detection.

1. Introduction

Proteins play crucial roles in cellular biology, including signaling cascades, metabolic cycles, and DNA transcription. In most cases, proteins rarely perform their functions alone; instead, they cooperate with other proteins by forming protein-protein interactions (PPIs) networks. PPIs are responsible for the majority of cellular functions. Over the past decades, many innovative techniques and systems for identifying protein interactions have been developed [1]; for example, in the high-throughput experimental technologies such as yeast two-hybrid (Y2H) screens [2], tandem affinity purification (TAP) [3], mass spectrometric protein complex identification (MS-PCI) [4], and other large-scale biological techniques for PPIs detection, a large amount of PPIs data for different species has been accumulated [5–11]. However, the experimental methods are costly and time consuming; therefore, current PPI pairs obtained from biological experiments only cover a small fraction of the complete PPI networks [12–14]. In addition, large-scale experimental methods usually suffer from high rates of both false positives and false negatives [12, 15–20]. Hence, it is of great practical significance to build low cost protein detection systems and establish the reliable computational methods to facilitate the detection of PPIs [21–25].

A number of computational methods have been proposed for the prediction of PPIs based on different data types, including phylogenetic profiles, gene neighborhood, gene fusion, sequence conservation between interacting proteins, and literature mining knowledge [12, 26–33]. There are also methods that combine interaction information from several different data sources [27]. However, the aforementioned methods cannot be carried out if such biological information about the proteins is not available. Recently, a number of methods which derive information directly from protein sequence are of particular interest [26, 28–30]. Researchers are committed to develop the sequences-based method for discovering new PPIs, and the experimental results showed

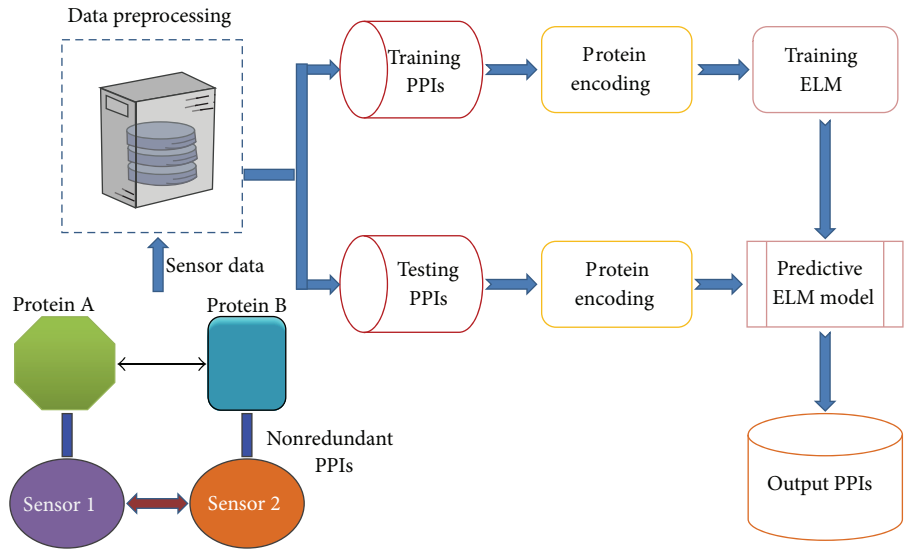


Figure 1: The schematic diagram for mapping large-scale protein-protein interactions by integrating biosensor data with ELM model.

that the information of amino acid sequences of proteins alone is sufficient to predict PPIs. Among them, one of the excellent works is a support vector machine based method developed by Shen et al. [29]. In that study, the twenty amino acids were firstly clustered into 7 classes according to their volumes and dipoles of the side chains. Then the conjoint triad approach extracts the features of protein pairs based on the classification of amino acids. When applied to predict *human* PPIs, this method yields a high prediction accuracy of about 84%.

Because the conjoint triad approach did not take neighboring effect into account and the interactions usually occur in the discontinuous amino acids segments in the sequence, on the other work Guo et al. developed a method based on SVM and autocovariance to extract the interactions information in the discontinuous amino acids segments in the sequence [26]. Their method yielded a prediction accuracy of 86.55%, when applied to predicting *Saccharomyces cerevisiae* PPIs. Lately, Pan et al. proposed a novel hierarchical LDA-RF model to predict *human* PPIs from protein primary sequences directly. In this study, the local sequential features represented by conjoint triads are firstly extracted from sequences. Then the generative LDA model is used to project the original feature space into the latent semantic space to obtain low dimensional latent topic features. Finally, the random forest model is used to predict the interactions between two proteins. The experimental results show that it is a very promising scheme for PPIs prediction [28].

The general trend in the current study for predicting PPIs has focused on high accuracy but has not considered the running time taken to train the classification model, which should be an important factor of developing a sequence-based method for predicting PPIs because the total number of possible PPIs is very large. For example, if we assume that the *human* genome consists of 22,500 protein-coding genes, then the total number of possible PPIs is estimated to be around 253,113,750 ($N = 22,500 \times (22,500 - 1)/2$), which indicates that some classification models with high classification accuracy may not be satisfactory when considering the tradeoff between the classification accuracy and the time for training the models. Here, in addition to exploring the local and global descriptors to mine interaction information from the multiscale amino acids segments at the same time, we also investigate the use of a novel paradigm of learning machine called extreme learning machine (ELM) [34], in order to obtain a balance between high classification accuracy and short training time.

In the present work, we report a novel sequence-based method for the prediction of interacting protein pairs using ELM combined with local and global descriptors. More specifically, we first represent each protein sequence as a vector by utilizing the novel representation of local and global protein sequence descriptors which provides us with a chance to mine interaction information from the multiscale amino acids segments at the same time. Then we characterize a protein pair in different feature vectors by coding the vectors of two proteins in this protein pair. Finally, an ELM model is constructed using these feature vectors of the protein pair as input. To evaluate the performance, the proposed method was applied to *human* PPI dataset. The experiment results show that our method achieved 84.8% prediction accuracy with 84.08% sensitivity at the specificity of 85.53%.

2. Materials and Methodology

In this section, we outline the main idea behind the proposed method. The flowchart intuitively showing how to map large-scale PPIs by integrating biosensor-based PPI data with computational model is given in Figure 1. Firstly, we discuss the PPI dataset which is used in the study to evaluate the performance of the proposed method. Next we introduce the novel sequence-based protein representation method.

Finally, we briefly descript the computational model, ELM, used in this study.

2.1. Golden Standard Datasets. We evaluated the proposed method with the *human* PPI dataset, which was downloaded from the Human Protein References Database (HPRD). After self-interactions and duplicate interactions were removed, the remaining 36,630 PPI pairs between 9,630 different human proteins comprise the final positive dataset.

The chosen golden negative dataset has a variable impact on the prediction performance, and it can be artificially inflated by a bias towards dominant samples in the positive data. For golden negative set, we followed the previous work [28] assuming that the proteins in separate subcellular compartments do not interact with each other. In this study, the golden negative dataset is generated from Swiss-Prot database version 57.3 according to four criteria: (1) protein sequences annotated with uncertain subcellular location terms were removed. (2) Protein sequences annotated by multiple locations were removed because of lack of the uniqueness. (3) Protein sequences annotated with "fragment" were removed. (4) Protein sequences with less than 50 amino acid residues were also removed because they might be fragments. After strictly following the above steps, we finally obtained 1,773 human proteins from six subcellular localizations. Then the noninteracting protein pairs were constructed by randomly pairing the proteins from separate subcellular compartments.

We also downloaded the golden negative dataset of human with experimental evidence used in the study of Smialowski et al. [35]. By combining the above two negative datasets, the whole final golden negative dataset consists of 36,480 noninteracting protein pairs. The whole dataset consists of 73,110 protein pairs, where nearly half are from the positive dataset and half are from the negative dataset. Four-fifths of the protein pairs from the positive and negative dataset were, respectively, randomly selected as the training dataset and the remaining one-fifths were used as the testing dataset.

2.2. Representing Proteins with Descriptors from Primary Protein Sequences. To successfully use the machine learning methods to identify PPIs from primary protein amino acids sequences, one of the most important computational challenges is how to effectively represent a protein sequence by a fixed length feature vector in which the important information content of proteins is fully encoded [36, 37]. In this study, two kinds of sequence representation approach are used to transform the protein sequences into feature vectors, including amino acid composition and a novel local descriptor. For amino acid composition, it is evident that 20 amino acid composition descriptors reflecting the fraction of each kind of amino acid in a protein sequence are directly calculated. Then, a local multiscale decomposition technique is used to divide protein sequence into multiple sequence segments of varying length to describe local regions. Here, the continuous sequence segments are composed of residues which are local in the polypeptide sequence [38].

In order to extract local information, we first divided the entire protein sequence into seven equal length fractions. Then a novel binary coding scheme was adopted to construct a set of continuous regions on the basis of the above partition. For example, consider a protein sequence "CCYGGGYY-CYYYCGGCCYYCG" containing 21 residues. To represent the sequence by a feature vector, let us first divide each protein sequence into multiple regions. For simplicity, the protein sequence is divided into four equal length segments (denoted as S_1, S_2, S_3, and S_4). Then it is encoded as a sequence of 1's and 0's of 4-bit binary form. In binary format, these combinations are written as *0000*, *0001*, *0010*, *0011*, *0100*, *0101*, *0110*, *0111*, *1000*, *1001*, *1010*, *1011*, *1100*, *1101*, *1110*, and *1111*. The number of states of a group of bits can be found by the expression 2^n, where n is the number of bits. It should be noticed that here 0 or 1 denotes one of the four equal length regions, and S_1–S_4 are excluded or included in constructing the continuous regions, respectively. For example, 1100 denotes a continuous region constructed by S_1 and S_2 (the first 50% of the sequence). Similarly, 0011 represents a continuous region constructed by S_3 and S_4 (the final 50% of the sequence).

It should be noticed that the proposed representation can be simply and conveniently edited at multiple scales, which offers a promising new approach for addressing these difficulties in a simple, unified, and theoretically sound way when presenting a protein sequence. For a given number of bits, each protein sequence may take on only a finite number of continuous or discontinuous regions. This limits the resolution of the sequence. If more bits are used for each protein sequence, then a higher degree of resolution is obtained. In this study, the protein sequence is encoded by 7-bit binary form; each protein sequence may take on 126 (2^7–2) different regions. Higher bit encoding requires more storage for data and requires more computing resource to process. In this study, only the continuous regions are used and the discontinuous regions are discarded.

For each continuous region, three types of descriptors, composition (C), transition (T), and distribution (D), are used to represent its characteristics. C denotes the amino acids number of a particular property (e.g., hydrophobicity) divided by the total amino acids number in a local region. T is the percentage frequency with which amino acids for a particular property are followed by protein amino acids of another property. D characterizes the chain length within which the first 25 percent, 50 percent, 75 percent, and 100 percent of the protein amino acids of a particular property are located, respectively [39].

The three descriptors can be calculated in the following ways. Firstly, in order to reduce the complexity inherent in the representation of the 20 standard protein amino acids, we firstly clustered them into seven clusters based on the volumes and dipoles of the side chains. Amino acids within the same groups likely involve synonymous mutations because of their similar characteristics [29]. The amino acids belonging to each group are shown in Table 1.

Then, every amino acid in each protein sequence is replaced by the index depending on its grouping. For example, protein sequence "CCYGGGYYCYYYCGGCCYYCG" is replaced by 77311133733371177337 1 based on this classification of amino acids (see Figure 2). There are six "1," eight "3," and seven "7" in this protein sequence. The composition for these

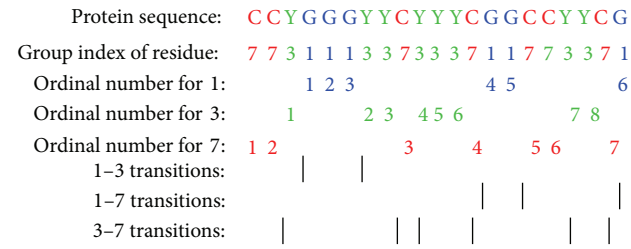


Figure 2: Sequence of a hypothetic protein indicating the construction of composition, transition, and distribution descriptors of a protein region.

Table 1: Division of amino acids into seven groups based on the dipoles and volumes of the side chains.

Group	Class	Dipole scale	Volume scale
1	Ala, Gly, Val	Dipole < 1.0	Volume < 50
2	Ile, Leu, Phe, Pro	Dipole < 1.0	Volume > 50
3	Tyr, Met, Thr, Ser	1.0 < dipole < 2.0	Volume > 50
4	His, Asn, Gln, Trp	2.0 < dipole < 3.0	Volume > 50
5	Arg, Lys	Dipole > 3.0	Volume > 50
6	Asp, Glu	Dipole > 3.0 (opposite orientation)	Volume > 50
7	Cys	1.0 < dipole < 2.0 (form disulphide bonds)	Volume > 50

three symbols is $6/(6+7+8) \times 100\% = 28.57\%$, $8/(6+7+8) \times 100\% = 38.10\%$, and $7/(6+7+8) \times 100\% = 33.33\%$, respectively. There are 2 transitions from "1" to "3" or from "3" to "1" in this sequence, and the percentage frequency of these transitions is $(2/20) \times 100\% = 10\%$. The transitions from "1" to "7" or from "7" to "1" in this sequence can similarly be calculated as $(3/20) \times 100\% = 15\%$. The transitions from "3" to "7" or from "7" to "3" in this sequence can also similarly be calculated as $(6/20) \times 100\% = 30\%$.

For distribution D, there are 6 residues encoded as "1" in the example of Figure 3, the positions for the first residue "1," the 2nd residue "1" ($25\% \times 6 = 2$), the 4th "1" residue ($50\% \times 6 = 3$), the 6th "1" ($75\% \times 6 = 5$), and the 8th residue "1" ($100\% \times 6 = 6$) in the encoded sequence are 4, 5, 6, 15, and 21, respectively, so the D descriptors for "1" are $(4/21) \times 100\% = 19.05\%$, $(5/21) \times 100\% = 23.81\%$, $(6/21) \times 100\% = 28.57\%$, $(15/21) \times 100\% = 71.43\%$, and $(21/21) \times 100\% = 100\%$, respectively. Similarly, the D descriptor for "3" and "7" is 14.29%, 33.33%, 47.62%, 57.14%, and 90.48% and 4.76%, 9.52%, 61.9%, 76.19%, and 95.24%, respectively.

For each continuous local region, the three descriptors (C, T, and D) were calculated and concatenated, and a total of 63 descriptors are generated: 7 for C, 21 ($(7 \times 6)/2$) for T, and 35 (7×5) for D. Then, the local descriptor from 27 regions (7-bit) was concatenated and a total 1701 dimensional vector has been built to represent each protein sequence. Finally,

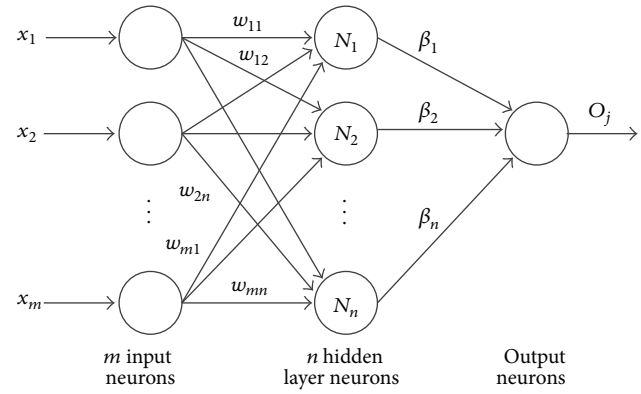


Figure 3: The structure of extreme learning machine.

the PPI pair is characterized by concatenating the local and global descriptors of two individual proteins. Thus, a 3442-dimensional vector has been constructed to represent each protein pair and was used as a feature vector for input into SVM classifier.

2.3. Extreme Learning Machine. By virtue of their approximation capabilities for nonlinear mappings, the feed-forward neural networks (FNN) have become ideal classifiers in many applications. Huang et al. proved that the single-hidden-layer FNN could exactly learn M distinct observations for almost any nonlinear activation function with almost M hidden nods [34, 40, 41]. However, the hidden layer biases and input weights of FNN have usually to be tuned using some parameter adjusting approach, which are generally time-consuming due to inappropriate learning steps with significantly large latency to converge to local maxima. Therefore, the slow learning speed of FNN has been a major bottleneck in different applications.

Extreme learning machine (ELM) was originally developed for the single hidden layer feed-forward neural network (SLFNN) and then extended to the generalized SLFNN where the hidden layer need not be neuron alike [34, 40]. As shown in Figure 3, its architecture is similar to that of a SLFNN. Recently the ELM algorithm has been increasingly popular in classification tasks due to its high generalization ability and fast learning speed. Different from the popular thinking that network parameters need to be adjusted, the input weights and first hidden layer biases need not be adjusted but they are randomly assigned in ELM. It has been proved that the ELM algorithm performs learning at an extremely fast speed and achieves a good generalization performance with activation functions which are infinitely differentiable in hidden layers [40, 42, 43].

The ELM algorithm transforms the learning problem into a simple linear system; that is, the output weights of ELM can be analytically determined through a generalized inverse operation of the hidden layer weight matrices. Compared with traditional learning frameworks such a learning scheme can operate at extremely much fast speed. Improved generalization performance of ELM with the smallest training error shows its superior classification capability for real-time

applications at an exceptionally fast pace without any learning bottleneck [44].

The basic idea behind ELM algorithm is briefly descripted as follows: suppose learning N arbitrary distinct samples $(x_i, t_i) \in R^n \times R^m$, where $x_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^T \subseteq R^n$, $t_i = [t_{i1}, t_{i2}, \ldots, t_{im}]^T \subseteq R^m$, a standard ELM with L hidden neurons and activation function $g(x)$ are mathematically modeled by

$$\sum_{i=1}^{L} \beta_i g\left(x_j\right) = \sum_{i=1}^{L} \beta_i g\left(w_i \cdot x_j + b_i\right) = o_j, \quad j = 1, \ldots, N, \tag{1}$$

where $w_i = [w_{i1}, w_{i2}, \ldots, w_{in}]^T$ represents the weight vector connecting the ith hidden node and the input nodes, $\beta_i = [\beta_{i1}, \beta_{i2}, \ldots, \beta_{im}]^T$ represents the weight vector connecting the ith hidden neuron and the output neurons, and b_i is the bias of the ith hidden neuron. $w_i \cdot x_j$ denotes the inner product of w_i and x_j. A wide variety of functions could be selected as the activation function, including sigmoid function, radial basis function, sine function, hardlim function, and triangular basis function. The architecture of ELM is shown in Figure 3. Equation (1) can be written compactly as

$$H\beta = T, \tag{2}$$

where

$$H\left(w_1, \ldots, w_L, b_1, \ldots, b_L, \ldots, x_1, \ldots, x_N\right) = \begin{bmatrix} g\left(w_1 \cdot x_1 + b_1\right) & \cdots & g\left(w_L \cdot x_1 + b_L\right) \\ \vdots & \cdots & \vdots \\ g\left(w_1 \cdot x_N + b_1\right) & \cdots & g\left(w_L \cdot x_N + b_L\right) \end{bmatrix}_{N \times L} \tag{3}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L \times m}, \qquad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m}.$$

H is termed as the hidden layer output matrix of the SLFNN; the ith column of H is the ith hidden neuron's output vector with respect to inputs $x_1, x_2, \ldots, x_N$. Hence for fixed arbitrary input weights w_i and the hidden layer bias b_i, training a SLFNN equals finding a least-squares solution $\widehat{\beta}$ of the linear system $H\beta = T$; that is,

$$\left\|H\left(w_1, \ldots, w_L, b_1, \ldots, b_L, x_1, \ldots, x_N\right)\widehat{\beta} - T\right\| = \min_{\beta} \left\|H\left(w_1, \ldots, w_L, b_1, \ldots, b_L, x_1, \ldots, x_N\right)\beta - T\right\|. \tag{4}$$

Equation (12) becomes a linear system and the solution is estimated as

$$\widehat{\beta} = H^{\dagger}T, \tag{5}$$

where $H^{\dagger}$ is the Moore-Penrose generalized inverse of the hidden layer output matrix H.

In summary, given a training dataset $\aleph = \{(x_i, t_i) \mid x_i \in R^n, t_i \in R^m, i = 1, \ldots, N\}$, activation function $g(x)$, and hidden neuron number L, the ELM-based learning procedure can be summarized as follows.

Step 1. Assign arbitrary input weight w_i and bias b_i, $i = 1, \ldots, L$.

Step 2. Calculate the hidden layer output matrix H.

Step 3. According to (13), calculate the output weight β.

3. Results and Discussion

In this section, we describe our simulation methodology and present the experimental results that evaluate the effectiveness of our schemes. The proposed sequence-based PPI predictor was implemented using MATLAB platform. For ELM algorithm, the implementation by Zhu and Huang available from http://www.ntu.edu.sg/home/egbhuang was used. Regarding SVM, LIBSVM implementation available from http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html was utilized, which was originally developed by Chang et al. [33]. Tree kinds of kernel functions were chosen and the optimized parameters were obtained with a grid search approach. All the simulations were carried out on a computer with 3.1 GHz 2-core CPU, 8 GB memory, and Windows operating system.

3.1. Cross Validation and Performance Evaluation. In the study, fivefold cross-validation technique has been employed to evaluate the performance of the proposed model. In fivefold cross-validation technique, the whole dataset is randomly divided into five subsets, where each subset consists of nearly equal number of interacting and noninteracting protein pairs. Four subsets are used for training and the remaining set for testing. This process is repeated five times so that each subset is used once for testing. The performance of method is average performance of method on five sets.

Seven metrics have been used in the study to measure the predictive ability of the proposed method. The parameters are as follows: (1) the overall prediction accuracy (ACC) is the percentage of correctly identified interacting and noninteracting protein pairs; (2) the sensitivity (SN) is the percentage of correctly identified interacting protein pairs; (3) the specificity (SP) is the percentage of correctly identified noninteracting protein pairs; (4) the positive predictive value (PPV) is the positive prediction value; (5) the negative predictive value (NPV) is the negative prediction value; (6) the F-score is a weighted average of the PPV and sensitivity, where an F-score reaches its best value at 1 and worst score at 0; (7) Matthew's correlation coefficient (MCC) is a more stringent measure of prediction accuracy accounts for both under- and overpredictions. These parameters are defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}, \tag{6}$$

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{7}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{8}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{9}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}, \tag{10}$$

$$F1 = 2 \times \frac{\mathrm{SN} \times \mathrm{PPV}}{\mathrm{SN} + \mathrm{PPV}}, \tag{11}$$

$$\begin{aligned}\mathrm{MCC} = {} & (\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}) \\ & \times ((\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \\ & \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN}))^{-1/2},\end{aligned} \tag{12}$$

where true positive (TP) is the number of true PPIs that are predicted correctly; false negative (FN) is the number of true PPIs that are predicted to be noninteracting pairs; false positive (FP) is the number of true noninteracting pairs that are predicted to be PPIs, and true negative (TN) is the number of true noninteracting pairs that are predicted correctly.

The above mentioned parameters rely on the selected threshold. The area under the ROC curve (AUC), which is threshold-independent for evaluating the performances, can be easily calculated according to the following formula [45]:

$$\mathrm{AUC} = \frac{S_0 - n_0 (n_0 + 1)/2}{n_0 \times n_1}, \tag{13}$$

where n_0 and n_1 denote the number of positive and negative samples, respectively, and S_0 is the sum of the ranks of all positive samples in the list of all samples ranked in increasing order by estimated probabilities belonging to positive. AUC values can give us a good insight into performance comparison of different prediction methods. Although the AUC is threshold-independent, an appropriate threshold must be selected for the final decision. For the classifier which outputs a continuous numeric value to represent the confidence or probability of a sample belonging to the predicted class, adjusting the classification threshold will lead to different confusion matrices which decide different ROC points [29].

3.2. Determination of ELM Parameter. The number of hidden nodes is a critical factor for the generalization of ELM. To determine the parameter, four-fifths of the whole dataset are randomly chosen to train the ELM classifiers with different number of hidden nodes, while the rest one-fifths of the dataset are used as the validation set to compute the accuracy.

Here the sigmoid function was used as the activation function of the ELM classifier. The results are plotted in Figure 4, which shows that the accuracy value reaches about 0.9 and increases slowly when the number of hidden neurons was set to 9 percent of the amount of samples. Based on Figure 4, we finally set 9 percent of the sample number as the number of hidden neurons for the ELM classifier. The second experiment was to examine how the running time scales with

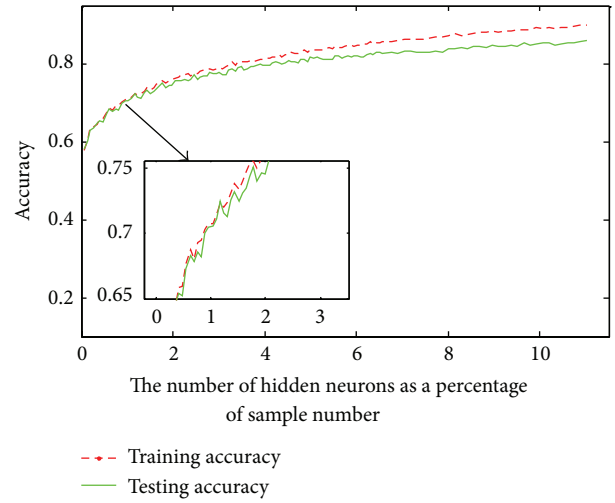


Figure 4: The relationship between the prediction accuracy and the number of hidden neurons. The x-axis denotes the number of hidden neurons as a percentage of sample number and the y-axis is the corresponding accuracy values.

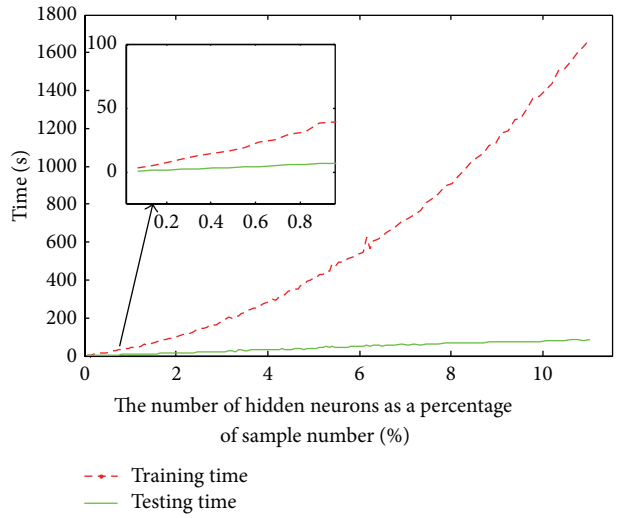


Figure 5: The relationship between the consuming time and the number of hidden neurons. The x-axis denotes the number of hidden neurons as a percentage of sample number and the y-axis is the running time.

the number of hidden neurons. We increase the number of hidden neurons from 1 to 11 percent of the amount of samples and measure the average time overhead. Figure 5 shows that the running time of proposed ELM model scales nearly linear as the hidden neuron size increases.

3.3. Prediction Performance of Proposed Model. We evaluated the performance of the proposed model using the PPIs dataset as described in the aforementioned section. To guarantee that the experimental results are valid and can be generalized for making predictions regarding new data, we adopted the fivefold cross-validation in this study. The advantages of cross-validation are that the impact of data

Table 2: Comparison of the prediction performance by the proposed method and state-of-the-art SVM classifier on the human dataset.

Method	Kernel	Mean/std	Time (s)	ACC	SN	SP	PPV	NPV	F1	MCC	AUC
							Testing				
ELM	Sigmoid	Mean	72.7901	0.8480	0.8408	0.8553	0.8547	0.8415	0.8477	0.7422	0.9232
		Variance	1.9062	0.0022	0.0019	0.0028	0.0040	0.0038	0.0029	0.0030	0.0028
	Hardlim	Mean	77.4139	0.8206	0.8171	0.8242	0.8227	0.8185	0.8199	0.7056	0.9020
		Variance	3.7710	0.0050	0.0040	0.0063	0.0088	0.0026	0.0063	0.0064	0.0031
	Gaussian	Mean	76.9615	0.7257	0.7328	0.7186	0.7232	0.7283	0.7279	0.6018	0.7624
		Variance	4.1012	0.0036	0.0048	0.0054	0.0085	0.0077	0.0044	0.0033	0.0017
							Training				
ELM	Sigmoid	Mean	1282.12	0.8887	0.8831	0.8944	0.8933	0.8843	0.8882	0.8022	0.9561
		Variance	17.25	0.0006	0.0010	0.0018	0.0014	0.0001	0.0008	0.0010	0.0012
	Hardlim	Mean	1330.33	0.8668	0.8655	0.8682	0.8683	0.8654	0.8669	0.7691	0.9397
		Variance	46.28	0.0027	0.0021	0.0033	0.0027	0.0027	0.0024	0.0039	0.0031
	Gaussian	Mean	1435.45	0.7824	0.7896	0.7753	0.7790	0.7860	0.7843	0.6595	0.8626
		Variance	94.85	0.0033	0.0022	0.0053	0.0040	0.0026	0.0029	0.0037	0.0038
							Testing				
SVM	Sigmoid	Mean	2794.29	0.8177	0.8119	0.8232	0.8215	0.8144	0.8165	0.7018	0.8878
		Variance	16.71	0.0127	0.0266	0.0128	0.0067	0.0200	0.0155	0.0160	0.0143
	Gaussian	Mean	5237.89	0.6947	0.4714	0.9191	0.8535	0.6348	0.6064	0.5320	0.8997
		Variance	67.82	0.0228	0.0412	0.0112	0.0178	0.0265	0.0340	0.0276	0.0364
	Polynomial	Mean	3612.98	0.8019	0.8219	0.7819	0.7903	0.8144	0.8057	0.6820	0.8838
		Variance	20.16	0.0101	0.0126	0.0117	0.0165	0.0114	0.0125	0.0122	0.0138

dependency is minimized and the reliability of the results can be improved.

The prediction performance of ELM predictor with novel representation of protein sequence across five runs is shown in Table 2. It can be observed from Table 2 that high prediction accuracy of 84.8% is achieved for the ELM model with sigmoid function. To better investigate the prediction ability of our model, we also calculated the values of sensitivity, specificity, PPV, NPV, F-score, MCC, and AUC. From Table 2, we can see that our model gives good prediction performance with an average sensitivity value of 84.08%, specificity value of 85.53%, PPV value of 85.47%, NPV value of 84.15%, F-score value of 84.77%, MCC value of 74.22%, and AUC value of 0.9232. Further, it can also be seen in Table 2 that the standard deviation of accuracy, sensitivity, specificity, PPV, NPV, F-score, MCC, and AUC is as low as 0.0022, 0.0019, 0.0028, 0.0040, 0.0038, 0.0029, 0.0030, and 0.0028, respectively.

To demonstrate the performance of the proposed model, we further compared our method with the state-of-the-art predictor SVM. From Table 2, we can see the performance of ELM and SVM model. As observed from Table 2, the testing time of SVM algorithm (2794.29 s) is roughly 38 times the testing time of ELM algorithm (72.7901 s) for sigmoid activation function. In addition, the prediction performance of ELM is also promising. The AUC of the SVM algorithm is 0.8878, which is lower than the ELM. The overall accuracy, sensitivity, specificity, PPV, NPV, $F1$ score, and MCC of SVM algorithm are, respectively, 81.77%, 81.19%, 82.32%, 82.15%, 81.44%, 81.65%, and 70.18% as illustrated in Table 2. Hence, it can be seen that almost all evaluation measures of ELM

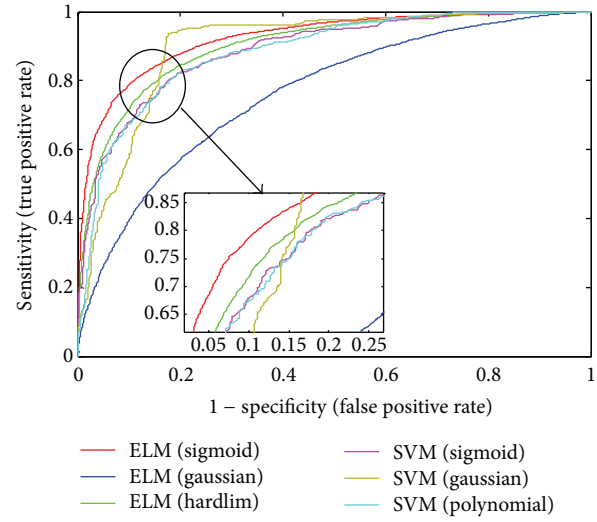


Figure 6: The ROC (receiver operator characteristic) curve illustrating the performance of different activation functions. The curve presents the true positive rate (sensitivity) against the false positive rate (1 − specificity).

algorithm are a little better than those of SVM algorithm, while its learning speed is much more faster than SVM.

We also conduct an experiment to characterize the sensitivity (i.e., the size of true positives that can be detected by our method) and specificity (i.e., 1 − false positive rate) of proposed approach for different activation functions

(Figure 6). The results in Figure 6 are reported using receiver operator characteristic (ROC) curves, which plot the achievable sensitivity at a given specificity (1 − false positive rate). Good performance is reflected in curves with a stronger bend towards the upper-left corner of the ROC graph (i.e., high sensitivity is achieved with a low false positive rate). We found that the proposed method achieved over 83 percent detection rate with less than 10 percent false positive rate. The results demonstrate that the proposed ELM can successfully classify positive and negative samples in all five activation functions that we investigated. Our algorithm can perfectly classify interacting and noninteracting protein pairs with only a few exceptions.

To sum up, considering the high efficiency as well as the good performance we can readily conclude that the proposed approach generally outperforms the state-of-the-art model with higher discrimination power for predicting PPIs based on the information of protein sequences. Therefore, we can see clearly that our model is a much more appropriate method for predicting new protein interactions compared with the other methods. Consequently, it makes us be more convinced that the proposed method can be very helpful in assisting the biologist to assist in the design and validation of experimental studies and for the prediction of interaction partners.

4. Conclusions

In this paper, we have developed an efficient and fast learning technique, which utilizes global and local information of protein amino acid sequence, for accurate identification PPIs at considerably high speed both in training and testing phase. The first contribution of this work is a novel protein amino acids sequence representation using amino acid composition and a descriptor to represent global and local information of a protein sequence, respectively. Then, the application of extreme learning machine ensures reliable recognition with minimum error and learning speed approximately thousands of times faster than the state-of-the-art classification method SVM. Experimental results demonstrated that the proposed method performed significantly well in distinguishing interacting and noninteracting protein pairs. It was observed that the proposed method achieved the mean classification accuracy of 84.8% using 5-fold cross-validation. Meanwhile, comparative study was conducted on the proposed method and the state-of-the-art SVM. The experimental results showed that our method significantly outperformed SVM in terms of classification accuracy with shorter running time.

Conflict of Interests

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

Acknowledgments

This work is supported by the National Science Foundation of China, under Grants 61102119, 61373086, 61133010, U1201256, and 61171125. The authors would like to thank all the guest editors and anonymous reviewers for their constructive advices.

References

[1] M. Vestergaard, K. Kerman, and E. Tamiya, "An overview of label-free electrochemical protein sensors," *Sensors*, vol. 7, no. 12, pp. 3442–3458, 2007.

[2] P. Uetz, L. Glot, G. Cagney et al., "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.

[3] S. R. Collins, P. Kemmeren, X. Zhao et al., "Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*," *Molecular and Cellular Proteomics*, vol. 6, no. 3, pp. 439–450, 2007.

[4] Y. Ho, A. Gruhler, A. Heilbut et al., "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[5] N. Simonis, J. Rual, A. Carvunis et al., "Empirically controlled mapping of the caenorhabditis elegans protein-protein interactome network," *Nature Methods*, vol. 6, no. 1, pp. 47–54, 2009.

[6] K. Venkatesan, J. Rual, A. Vazquez et al., "An empirical framework for binary interactome mapping," *Nature Methods*, vol. 6, no. 1, pp. 83–90, 2009.

[7] H. Yu, P. Braun, M. A. Yildirim et al., "High-quality binary protein interaction map of the yeast interactome network," *Science*, vol. 322, no. 5898, pp. 104–110, 2008.

[8] L. Giot, J. S. Bader, C. Brouwer et al., "A Protein Interaction Map of Drosophila melanogaster," *Science*, vol. 302, no. 5651, pp. 1727–1736, 2003.

[9] V. Schachter, "Construction and prediction of protein: protein interaction maps," in *Ernst Schering Research Foundation Workshop. Bioinformatics and Genome Analysis*, H. W. Mewes, H. Seidel, and B. Weiss, Eds., vol. 38, pp. 191–220, 2002.

[10] L. Giot, J. S. Bader, C. Brouwer et al., "A protein interaction map of Drosophila melanogaster," *Science*, vol. 302, no. 5651, pp. 1727–1736, 2003.

[11] T. Huang, S. Wan, Z. Xu et al., "Analysis and prediction of translation rate based on sequence and functional features of the mRNA," *PLoS ONE*, vol. 6, no. 1, Article ID e16036, 2011.

[12] Z. You, Y. Lei, J. Gui, D. Huang, and X. Zhou, "Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data," *Bioinformatics*, vol. 26, no. 21, pp. 2744–2751, 2010.

[13] A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein, "Bridging structural biology and genomics: assessing protein interaction data with known complexes," *Drug Discovery Today*, vol. 9, no. 2, pp. S32–S40, 2004.

[14] G. Liu, J. Li, and L. Wong, "Assessing and predicting protein interactions using both local and global network topological metrics," *Genome Informatics*, vol. 21, pp. 138–149, 2008.

[15] H. N. Chua and L. Wong, "Increasing the reliability of protein interactomes," *Drug Discovery Today*, vol. 13, no. 15-16, pp. 652–658, 2008.

[16] T. Huang, J. Zhang, Z. Xu et al., "Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches," *Biochimie*, vol. 94, no. 4, pp. 1017–1025, 2012.

[17] Z.-H. You, Y.-K. Lei, L. Zhu, J. Xia, and B. Wang, "Prediction of protein-protein interactions from amino acid sequences with

ensemble extreme learning machines and principal component analysis," *BMC bioinformatics*, vol. 14, supplement 8, article S10, 2013.

[18] T. Huang, C. Wang, G. Zhang, L. Xie, and Y. Li, "SySAP: a system-level predictor of deleterious single amino acid polymorphisms," *Protein and Cell*, vol. 3, no. 1, pp. 38–43, 2012.

[19] Z. You, Z. Yin, K. Han, D. Huang, and X. Zhou, "A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network," *BMC Bioinformatics*, vol. 11, no. 1, article 343, 2010.

[20] T. Huang, J. Wang, Y. Cai, H. Yu, and K. Chou, "Hepatitis c virus network based classification of hepatocellular cirrhosis and carcinoma," *PLoS ONE*, vol. 7, no. 4, Article ID e34460, 2012.

[21] J. Song, Z. Yuan, H. Tan, T. Huber, and K. Burrage, "Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure," *Bioinformatics*, vol. 23, no. 23, pp. 3147–3154, 2007.

[22] T. Huang, X. Shi, P. Wang et al., "Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks," *PLoS ONE*, vol. 5, no. 6, Article ID e10972, 2010.

[23] J. Song, H. Tan, H. Shen et al., "Cascleave: towards more accurate prediction of caspase substrate cleavage sites," *Bioinformatics*, vol. 26, no. 6, Article ID btq043, pp. 752–760, 2010.

[24] T. Huang, K. Tu, Y. Shyr, C.-C. Wei, L. Xie, and Y.-X. Li, "The prediction of interferon treatment effects based on time series microarray gene expression profiles," *Journal of Translational Medicine*, vol. 6, article 44, 2008.

[25] L. Zhu, Z. You, D. Huang, and B. Wang, "*t*-LSE: a novel robust geometric approach for modeling protein-protein interaction networks," *PLoS ONE*, vol. 8, no. 4, Article ID e58368, 2013.

[26] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences," *Nucleic Acids Research*, vol. 36, no. 9, pp. 3025–3030, 2008.

[27] Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph, "A mixture of feature experts approach for protein-protein interaction prediction," *BMC Bioinformatics*, vol. 8, no. 10, article S6, 2007.

[28] X.-Y. Pan, Y.-N. Zhang, and H.-B. Shen, "Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features," *Journal of Proteome Research*, vol. 9, no. 10, pp. 4992–5001, 2010.

[29] J. Shen, J. Zhang, X. Luo et al., "Predicting protein-protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 11, pp. 4337–4341, 2007.

[30] S. Pitre, M. Hooshyar, A. Schoenrock et al., "Short co-occurring polypeptide regions can predict global protein interaction maps," *Scientific Reports*, vol. 2, article 239, 2012.

[31] Y.-K. Lei, Z.-H. You, Z. Ji, L. Zhu, and D.-S. Huang, "Assessing and predicting protein interactions by combining manifold embedding with multiple information integration," *BMC Bioinformatics*, vol. 13, supplement 7, article S3, 2012.

[32] J. Song, H. Tan, M. Wang, G. I. Webb, and T. Akutsu, "Tangle: two-Level support vector regression approach for protein backbone torsion angle prediction from primary sequences," *PLoS ONE*, vol. 7, no. 2, Article ID e30361, 2012.

[33] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[34] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.

[35] P. Smialowski, P. Pagel, P. Wong et al., "The Negatome database: a reference set of non-interacting protein pairs," *Nucleic Acids Research*, vol. 38, no. 1, pp. D540–D544, 2009.

[36] J. Song, H. Tan, A. J. Perry et al., "PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites," *PLoS ONE*, vol. 7, no. 11, Article ID e50300, 2012.

[37] T. Huang, Z. Xu, L. Chen, Y. Cai, and X. Kong, "Computational analysis of HIV-1 resistance based on gene expression profiles and the virus-host interaction network," *PLoS ONE*, vol. 6, no. 3, Article ID e17291, 2011.

[38] T. Huang, M. Jiang, X. Kong, and Y. Cai, "Dysfunctions associated with methylation, microrna expression and gene expression in lung cancer," *PLoS ONE*, vol. 7, no. 8, Article ID e43441, 2012.

[39] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. Kim, "Prediction of protein folding class using global description of amino acid sequence," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 19, pp. 8700–8704, 1995.

[40] G. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

[41] G. Huang, Q. Zhu, and C. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, vol. 1–4, pp. 985–990, July 2004.

[42] G. B. Huang, X. Ding, and H. Zhou, "Optimization method based extreme learning machine for classification," *Neurocomputing*, vol. 74, no. 1–3, pp. 155–163, 2010.

[43] G. Huang, M. Li, L. Chen, and C. Siew, "Incremental extreme learning machine with fully complex hidden nodes," *Neurocomputing*, vol. 71, no. 4–6, pp. 576–583, 2008.

[44] R. Minhas, A. A. Mohammed, and Q. M. Jonathan Wu, "A fast recognition framework based on extreme learning machine using hybrid object information," *Neurocomputing*, vol. 73, no. 10-12, pp. 1831–1839, 2010.

[45] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 450386, 10 pages
http://dx.doi.org/10.1155/2014/450386

Research Article

Gene Ontology and KEGG Enrichment Analyses of Genes Related to Age-Related Macular Degeneration

Jian Zhang,[1,2] ZhiHao Xing,[3] Mingming Ma,[1,2] Ning Wang,[1,2] Yu-Dong Cai,[4] Lei Chen,[5] and Xun Xu[1,2]

[1] *Department of Ophthalmology, Shanghai First People's Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai 200080, China*
[2] *Shanghai Key Laboratory of Ocular Fundus Diseases, Shanghai First People's Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai 200080, China*
[3] *The Key Laboratory of Stem Cell Biology, Institute of Health Sciences, Shanghai Jiaotong University School of Medicine and Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200025, China*
[4] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[5] *College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*

Correspondence should be addressed to Lei Chen; chen_lei1@163.com and Xun Xu; drxuxun@tom.com

Received 13 June 2014; Accepted 21 July 2014; Published 6 August 2014

Academic Editor: Tao Huang

Identifying disease genes is one of the most important topics in biomedicine and may facilitate studies on the mechanisms underlying disease. Age-related macular degeneration (AMD) is a serious eye disease; it typically affects older adults and results in a loss of vision due to retina damage. In this study, we attempt to develop an effective method for distinguishing AMD-related genes. Gene ontology and KEGG enrichment analyses of known AMD-related genes were performed, and a classification system was established. In detail, each gene was encoded into a vector by extracting enrichment scores of the gene set, including it and its direct neighbors in STRING, and gene ontology terms or KEGG pathways. Then certain feature-selection methods, including minimum redundancy maximum relevance and incremental feature selection, were adopted to extract key features for the classification system. As a result, 720 GO terms and 11 KEGG pathways were deemed the most important factors for predicting AMD-related genes.

1. Introduction

Age-related macular degeneration (AMD or ARMD) is a chronic, progressive eye disorder that primarily occurs in elders (>50 years) and has become a major cause of blindness and visual impairment in developed countries as well as the third major cause globally [1, 2]. In an Asian population aged 40–79 years, the morbidities of early and late AMD were 6.8% and 0.56%, respectively [3]. Further, AMD is likely to increase with a longer life expectancy. Due to retina damage, AMD typically results in vision loss, which can render daily activities difficult, such as reading, watching TV, and recognizing faces [4]. There are two typical types of AMD: dry AMD and wet AMD. Dry AMD is the major type of AMD and accounts for approximately 80% of cases; no efficient surgical or medical treatments are available. It typically causes mild vision loss, which develops slowly. However, it can cause vision loss through retinal pigment epithelial layer atrophy, which results in photoreceptor loss (rods and cones) in the central portion of the eye. Wet AMD is caused by choroidal neovascularization (CNV), wherein new blood vessels grow in choriocapillaries through the Bruch's membrane. Leaking and bleeding of these vessels can damage the rods and cones, which lead to rapidly deteriorating vision. Thus, wet AMD accounts for 90% of AMD cases with severe visual impairment.

The AMD etiology is complex. AMD results from both genetic and environmental factors; however, the underlying

mechanisms are unclear. Moreover, previous studies have demonstrated strong correlations between AMD and multiple environmental factors. In addition to age, many risk factors are correlated with AMD, such as cigarette smoking [5], oxidative stress [6–8], hypertension, previous cataract surgery, higher body mass index, a history of cardiovascular disease, and higher plasma fibrinogen [9].

AMD is characterized by complex traits. Moreover, mutant protein expression may begin early in AMD patients, and symptoms associated with AMD do not manifest until a long time thereafter. Often only clinical information for a single generation is available for studies; thus, it is difficult to detect AMD phenotypic heterogeneity and determine the underlying mechanisms. Initially, through early linkage studies on small families, several genetic loci at chromosomes 9p24, 10q26, and 15q21 [10] and 1q31, 10q26, and 17q25 [11] were identified and verified. A GWAS study greatly increased our understanding of AMD risk loci. Subsequently, more AMD-related genes have been identified, such as *C2* [12], *CFH* [13], *CFI* [14], *LIPC* [15], *CETP*, *TIMP3* [16], and *TNFRSF10A* [17]. Recently a large-scale GWAS analysis of more than 17,000 AMD cases indicated 19 other AMD loci, in which 7 loci were novel and near the genes *IER3-DDR1*, *COL8A1-FILIP1L*, *SLC16A8*, *TGFBR1*, *ADAMTS9*, *RAD51B*, and *B3GALTL* [18]. Several studies have evaluated the impact of susceptibility genes on AMD onset and progression. For instance, *CFH* gene mutations yield a high risk of AMD. Compared with the normal homozygous genotype, individuals with heterozygotic and homozygotic *CFH* exhibited a 4.6-fold or 7.4-fold increased AMD risk, respectively [19].

AMD is a disease with complex inheritance patterns, and it may be difficult to discover individual susceptibility genes due to multiple genetic and environmental effects and interactions. Identifying several genetic loci revealed that several important biological pathways are involved in AMD pathogenesis, such as the cholesterol, lipid metabolism pathway, complement pathway, extracellular matrix pathway, oxidative stress pathway, and angiogenesis signaling pathway in [20–22], which provides a foundation for systematically analyzing the biological processes underlying AMD. Gene ontology (GO) is a major bioinformatics tool that standardizes representation and the product attributes of genes across species [23]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [24, 25] pathway database is a collection of manually drawn diagrams and comprehensive inferences for pathway mapping. Based on the gene ontology and KEGG pathway materials, we analyzed the GO and KEGG enrichments for known AMD-related genes, which were retrieved from the Retina International website (http://www.retina-international.org/files/sci-news/remacdy.htm) or the published literature. To extract the distinctive features of these genes, certain genes, which were not reported as AMD-related genes, were randomly selected from Ensemble. Each investigated gene was encoded into numeric vectors consisting of enrichment scores of the gene set, including it and its direct neighbors in STRING, and the GO terms or KEGG pathways. Based on certain feature-selection methods and SMO as the prediction engine, certain important GO terms and KEGG pathways were discovered that were deemed important for identifying AMD-related genes. Analyses suggest that certain such genes relate directly or indirectly to AMD formation or development.

2. Materials and Methods

2.1. Dataset. The known AMD-related genes were retrieved from the Retina International website (http://www.retina-international.org/files/sci-news/remacdy.htm, recent update from March 24, 2010) and the literature. Specifically, 16 genes are from Retina International; three genes for the complement system proteins factor H (*CFH*), factor 3 (*C3*), and factor B (*CFB*), which are strongly related with a person's risk for developing AMD, are employed; *HTRA1* is from [26, 27]; *ABCR* is from [28]; 2 genes are from [29, 30]; and 23 genes are from [18]. Finally, 39 known AMD-related genes were collected; these genes are referred to as "positive genes" and compose the gene set S_p. To analyze the differences between the positive genes and other genes, we randomly selected 1,950 genes (50 times the number of positive genes) from Ensemble that were not in S_p; these 1,950 genes are referred to as "negative genes" and compose the set S_n. The Ensemble IDs for the positive and negative genes are in Supplementary Material I available online at http://dx.doi.org/10.1155/2014/450386.

The negative genes outnumbered the positive genes; thus, we confronted an imbalanced dataset. Encouraged by certain studies that have managed this type of data [31, 32], the following strategy was adopted. The negative genes were equally and randomly split into 10 portions $S_n^1, S_n^2, \ldots, S_n^{10}$ (i.e., $S_n = S_n^1 \cup S_n^2 \cup \cdots \cup S_n^{10}$ and $S_n^i \cap S_n^j = \phi$ for $i \neq j$). For each S_n^i, we combined the genes in S_p and S_n^i to comprise the ith datasets D_i (i.e., $D_i = S_n^i \cup S_p$).

2.2. Feature Construction. To analyze the differences between the positive and negative genes, each gene must be represented by certain features that can then be processed by certain computer programs. Here, we adopted gene ontology (GO) and KEGG enrichment to compute numerical values that represent each gene.

GO enrichment indicates the relationship between genes and GO terms. For each gene g and each GO term GO_j, a score is generated, which is typically referred to as the gene ontology enrichment score and defined as the $-\log_{10}$ of the hypergeometric test P value [33–35] for a gene set G consisting of g's direct neighbors in STRING and the GO term GO_j that can be computed as follows:

$$\mathrm{ES}_{\mathrm{GO}}\left(g, \mathrm{GO}_j\right) = -\log_{10}\left(\sum_{k=m}^{n} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}\right), \quad (1)$$

where N denotes the overall number of proteins in humans, M denotes the number of proteins annotated in the gene ontology term GO_j, n denotes the number of proteins in G, and m denotes the number of proteins in G that are annotated in the gene ontology term GO_j. If the score is large for one gene and one GO term, the gene and GO term likely

have a strong relationship; there were 12,877 gene ontology enrichment scores.

Similarly, for each gene g and each KEGG pathway P_j, the KEGG enrichment score is defined as the $-\log_{10}$ of the hypergeometric test P value [35, 36] for a gene set G that consists of g's direct neighbors in STRING and the KEGG pathway P_j, which can be calculated as follows:

$$\mathrm{ES}_{\mathrm{KEGG}}\left(g, P_j\right) = -\log_{10}\left(\sum_{k=m}^{n} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}\right), \quad (2)$$

where N denotes the overall number of proteins in humans, M denotes the number of proteins annotated in the KEGG pathway P_j, n denotes the number of proteins in G, and m denotes the number of proteins in G that are annotated in the KEGG pathway P_j. Additionally, a higher KEGG enrichment score between g and P_j indicates a stronger relationship; 239 features were KEGG enrichment scores.

Accordingly, each gene g can be represented by 12,877 gene ontology enrichment scores and 239 KEGG enrichment scores, which can be formulated as follows:

$$\begin{aligned} v(g) = &\left(\mathrm{ES}_{\mathrm{GO}}\left(g, \mathrm{GO}_1\right), \ldots, \mathrm{ES}_{\mathrm{GO}}\left(g, \mathrm{GO}_{12877}\right),\right. \\ &\left.\mathrm{ES}_{\mathrm{KEGG}}\left(g, P_1\right), \ldots, \mathrm{ES}_{\mathrm{KEGG}}\left(g, P_{239}\right)\right)^{\mathrm{T}}. \end{aligned} \quad (3)$$

2.3. Prediction Method and Accuracy Measurement. Weka [37] is a collection of many state-of-the-art machine-learning algorithms and has been used to solve various biological problems [38–42]. One classifier, which is referred to as SMO, was adopted herein as the classification method; it implements John Platt's sequential minimal optimization algorithm to solve the optimization problem that should be settled during training of a support vector classifier. The kernel function can be polynomial or Gaussian [43, 44].

The predicted results for a two-class classification problem can be represented by a confusion matrix consisting of four entries: a true positive (TP), a true negative (TN), false positives (FP), and a false negative (FN) [45, 46]. Accordingly, the prediction accuracy (ACC), specificity (SP), and sensitivity (SN) can be computed as follows:

$$\begin{aligned} \mathrm{ACC} &= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \\ \mathrm{SP} &= \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \\ \mathrm{SN} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \end{aligned} \quad (4)$$

However, in each dataset D_i, the number of negative genes was 5 times as many as the number of positive genes, which is still imbalanced. Thus, an additional measurement, Matthews's correlation coefficient (MCC) [47], was employed to solve the problem; the coefficient can be computed as follows:

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TN} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TP} + \mathrm{FP})}}. \quad (5)$$

2.4. 10-Fold Cross Validation. Ten-fold cross validation is often used to examine the performance of various classification models [48]. In 10-fold cross validation, the dataset is equally and randomly divided into ten portions. Each portion is used as testing data, and the samples in the remaining nine portions compose the training dataset. Each sample is tested once because each portion is tested once. Compared with the Jackknife test [49, 50], a 10-fold cross-validation test is more efficient and provides similar results for a given dataset. Thus, it was adopted herein to examine the classification model.

2.5. Feature Selection. As described in Section 2.2, each gene is represented by 12,877 + 239 = 13,116 enrichment scores. To analyze these features and extract key features that contribute the most to the positive and negative gene classification, certain feature-selection methods were employed. This procedure included two stages: (1) using Cramer's coefficient [51, 52] to exclude nonsignificant features and (2) using the minimum redundancy maximum relevance (mRMR) method as well as incremental feature selection (IFS) [53] for additional selection.

Cramer's coefficient [51, 52] is a statistical measure of two variables that was derived from the Pearson Chi-square test [54]; it ranges from 0 to 1. A high Cramer's coefficient for two variables indicates a strong association. Here, for each feature and samples' class labels, Cramer's coefficient was calculated, and features with a Cramer's coefficient lower than 0.1 were excluded.

The remaining features were further refined using the minimum redundancy maximum relevance (mRMR) method and incremental feature selection (IFS), which are feature selection methods that have been widely used in recent years [34, 55–58]. By evaluating a classification model, key features can be extracted from a complicated biological system. The mRMR method has two criteria: max-relevance and min-redundancy. Accordingly, two feature lists can be generated using this method: (1) the MaxRel feature list and (2) the mRMR feature list. Specifically, the former list sorts features according to their contributions to the classification (i.e., only considering the criterion of max-relevance), while the latter list sorts features by considering both the max-relevance and min-redundancy criteria. The MaxRel and mRMR features lists were formulated as follows:

$$\begin{aligned} &\text{MaxRel features list} : F_M = \left[f_1^M, f_2^M, \ldots, f_N^M\right], \\ &\text{mRMR features list} : F_m = \left[f_1^m, f_2^m, \ldots, f_N^m\right], \end{aligned} \quad (6)$$

Table 1: The number of remaining features for each dataset after the first stage of feature selection.

Dataset	Number of remaining features
D_1	4,288
D_2	3,919
D_3	4,549
D_4	4,663
D_5	4,371
D_6	5,012
D_7	4,877
D_8	3,787
D_9	4,701
D_{10}	4,473

where N denotes the total number of features. A detailed description of the mRMR method can be found in Peng et al.'s paper [53].

Only the mRMR features list was used to extract key features. The extraction procedure is described as follows.

(1) For the mRMR features list F_m, construct N feature set, say $F_m^1, F_m^2, \ldots, F_m^N$, such that $F_m^i = [f_1^m, f_2^m, \ldots, f_i^m]$ $(1 \le i \le N)$ (i.e., F_m^i contained the first i features in F_m).

(2) The classifier SMO was evaluated through 10-fold cross validation using features in F_m^i. As described in Section 2.3, ACC, SP, SN and MCC can be obtained.

(3) The feature set with the maximum MCC is deemed the optimal feature set. For ease in observation, an IFS-curve can be plotted with MCC values as the y-axis and the superscript i of F_m^i as the x-axis.

3. Results and Discussion

3.1. Results of the First Stage of Feature Selection. For each of the 10 datasets $D_1, D_2, \ldots, D_{10}$, Cramer's coefficients of the features and samples' class labels were calculated. Accordingly, features with Cramer's coefficients less than 0.1 were excluded, while the remaining features were processed further. The number of remaining features in each dataset is listed in Table 1.

3.2. Results of the Second Stage of Feature Selection. For each dataset D_i, the mRMR, IFS, and SMO methods were used to process the remaining features. The mRMR program was retrieved from http://research.janelia.org/peng/proj/mRMR/ and was executed with its default parameters. As a result, we generated two feature lists: the MaxRel and mRMR features lists. To reduce the computation time, only the first 500 features in each of the two feature lists were obtained, and they are available in Supplementary Material II.

The IFS and SMO methods were used in accordance with the mRMR features list for each dataset D_i evaluated using 10-fold cross validation. The SNs, SPs, ACCs, and MCCs obtained for each dataset D_i are available in Supplementary Material III. For clarity, we plotted an IFS-curve for each dataset D_i, which is referred to as IFS-curve-D_i. The five IFS-curves for D_1, D_2, D_3, D_4, and D_5 are shown in Figure 1(a), while the other five IFS-curves for D_6, D_7, D_8, D_9, and D_{10} are shown in Figure 1(b); the ten IFS-curves that are plotted in separate coordinates are available in Supplementary Material IV. Generating the maximum MCC for each dataset from Supplementary Material III and IV (listed in column 3 of Table 2) was a straightforward process. Clearly, most MCCs are in the range 0.7 to 0.8, and the mean value was 0.76139. As mentioned in Section 2.5, the features used to obtain the maximum MCC compose the optimal feature set. The number of features in the optimal feature set for each dataset is listed in column 2 of Table 2. The results for dataset D_1 are described as follows. The maximum MCC for the dataset D_1 is 0.712699 (listed in row 2 and column 3 of Table 2) using the first 344 (listed in row 2 and column 2 of Table 2) features in the mRMR features list of dataset D_1 (see Supplementary Material II).

Table 2: The number of features in the optimal feature set for each dataset and the MCC value obtained using these features.

Dataset	Number of features in the optimal feature set	Maximum MCC value
D_1	344	0.712699
D_2	226	0.723116
D_3	104	0.873086
D_4	57	0.77142
D_5	146	0.744851
D_6	26	0.699118
D_7	136	0.788893
D_8	462	0.789865
D_9	55	0.704687
D_{10}	70	0.806162
Mean		0.76139

3.3. Analysis of the Optimal Feature Set. As mentioned in Section 3.2, we generated an optimal feature set for each dataset, thereby obtaining 10 optimal feature sets. We combined these optimal feature sets to compose the final optimal feature set, which includes 720 GO terms and 11 KEGG pathways that are available in Supplementary Material V. To discern the distribution of these 731 optimal features, we counted the number of optimal feature sets containing each of 731 features. Figure 2 shows the number of features against the number of optimal feature sets, from which we can see that 400 features were exactly contained in one optimal feature set, 131 features were exactly contained in two optimal feature sets, while others were contained in at least three optimal feature sets. Accordingly, 45.28% (331/731) features were contained in at least two optimal feature sets, indicating that different datasets may induce some common features. It also suggested that some important features for distinguishing AMD-related genes were contained in the final optimal feature set. In

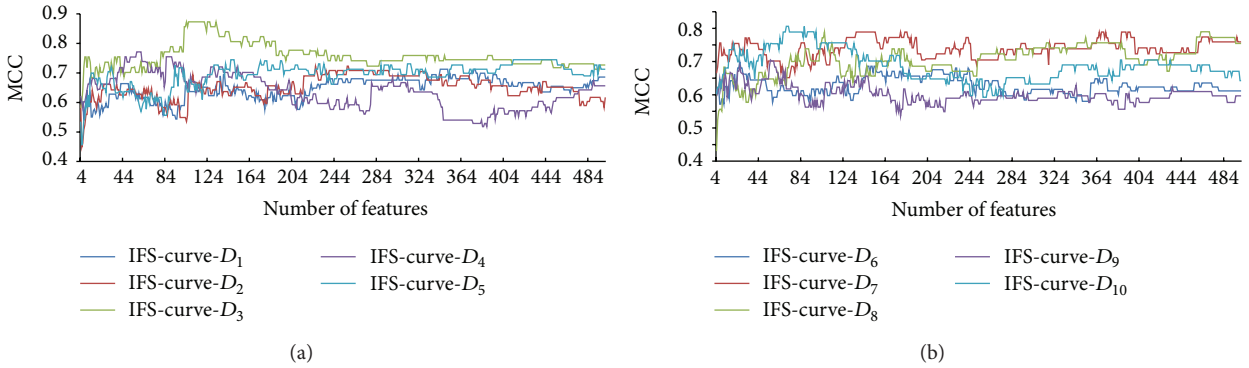


Figure 1: IFS-curve for each dataset. Specifically, (a) shows the IFS-curves for the datasets D_1, D_2, D_3, D_4, and D_5, while (b) shows the IFS-curves for the datasets D_6, D_7, D_8, D_9, and D_{10}. The y-axis represents Matthews's correlation coefficient (MCC), and the x-axis represents the number of features involved in the classification model.

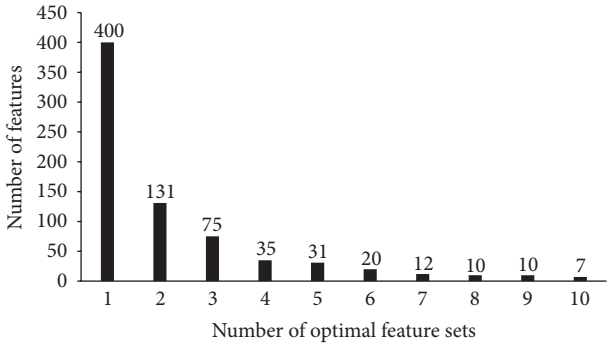


Figure 2: The number of features against the number of optimal feature sets.

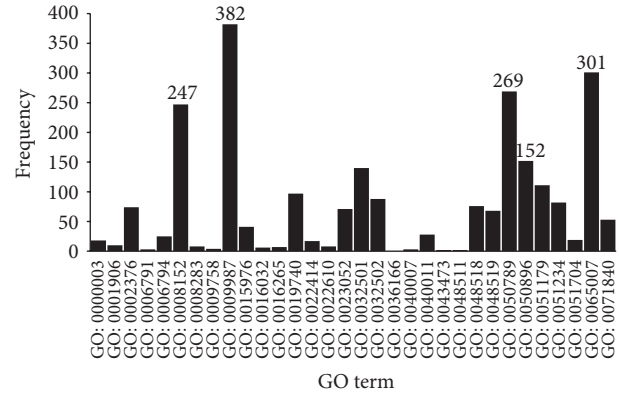


Figure 3: Frequency of children terms of biological process GO terms in the final optimal feature set.

the following sections, features in the final optimal feature set were discussed.

3.3.1. GO Number and Percentage. It is known that GO terms can be divided into the following three types: (1) biological process (BP) GO term, (2) cellular component (CC) GO term, and (3) molecular function (MF) GO term. To efficiently discern the biological meanings and characterize the functional essentiality of the GO terms in the final optimal feature set, we considered the children terms of the aforementioned three types. For clarity, let S_o be the 720 GO terms in the final optimal feature set and S be the children terms of any children term of BP GO term, CC GO term, or MF GO term. To display the distribution of the GO terms in S_o, we calculated the frequency and percentage for each children term of BP GO term, CC GO term, or MF GO term which were defined as $|S_o \cap S|$ and $|S_o \cap S|/|S|$, respectively. Figures 3–8 display the frequency and percentage of children terms of BP GO term, CC GO term, or MF GO term in the final optimal feature set.

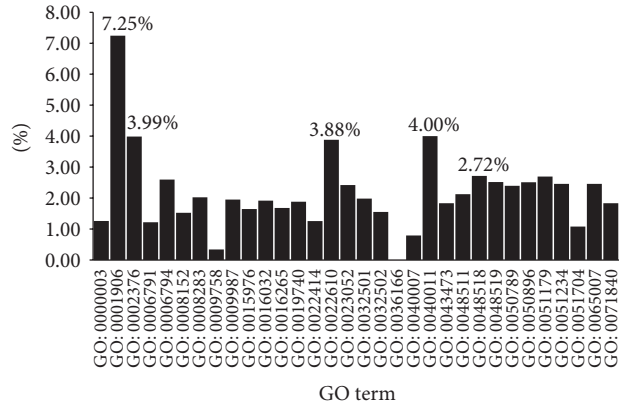


Figure 4: Percentage of children terms of biological process GO terms in the final optimal feature set.

(1) *BP GO Terms.* In Figure 3, based on the BP term frequencies, the top five biological process terms are (I) GO: 0009987: cellular process (382); (II) GO: 0065007: biological regulation (301); (III) GO: 0050789: regulation of biological process (269); (IV) GO: 0008152: metabolic process (247); and (V) GO: 0050896: response to stimulus (152).

The top four BP terms may indicate that these biological processes are necessary to maintain normal cellular functions

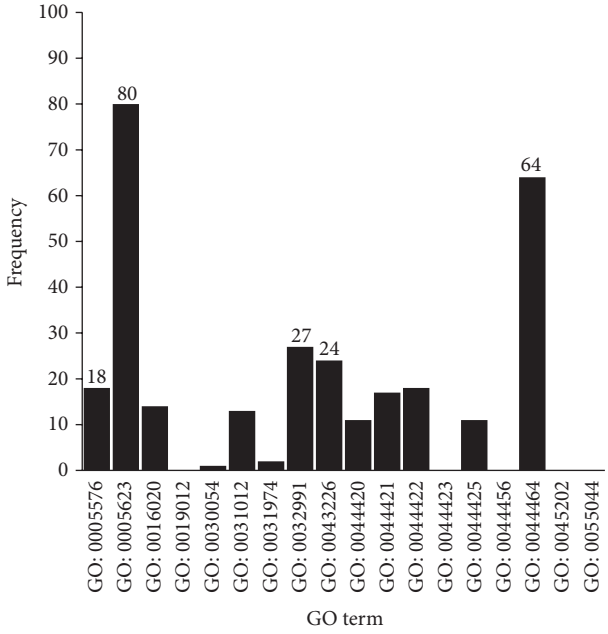


Figure 5: Frequency of children terms of cellular component GO terms in the final optimal feature set.

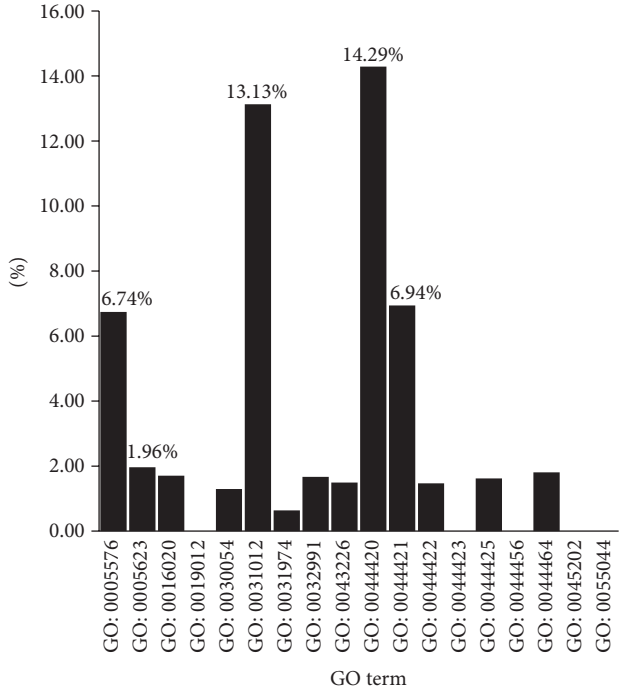


Figure 6: Percentage of children terms of cellular component GO terms in the final optimal feature set.

and may lead to AMD due to aberrant behavior in relevant cells.

"Response to stimulus" refers to any process that results from a stimulus, which leads to a change in a state or activity, such as movement and secretion.

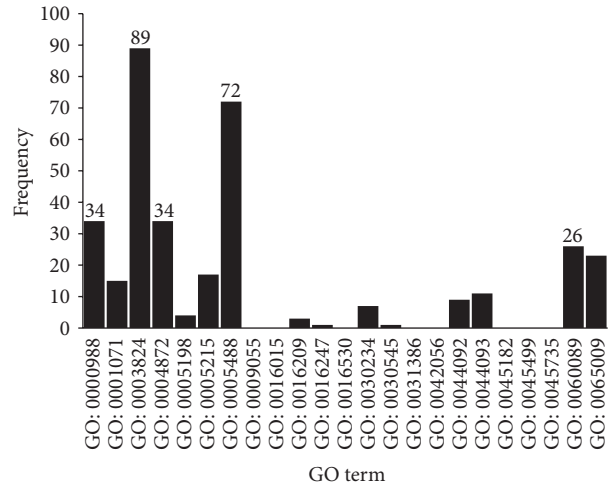


Figure 7: Frequency of children terms of molecular function GO terms in the final optimal feature set.

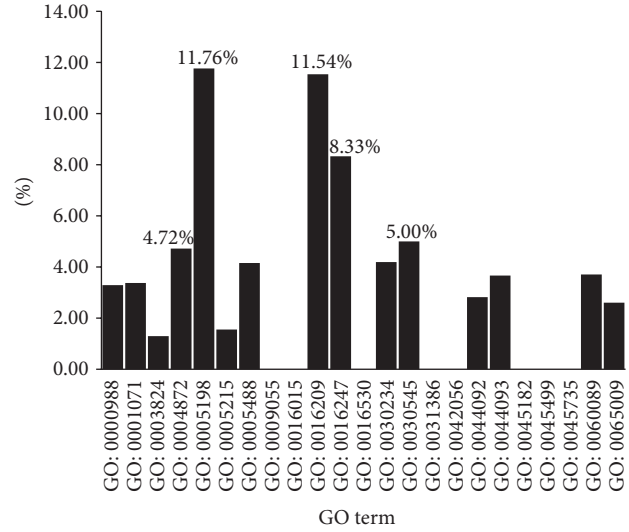


Figure 8: Percentage of children terms of molecular function GO terms in the final optimal feature set.

For the BP term percentages, as shown in Figure 4, the top five biological process terms are (I) GO: 0001906: cell killing (7.25%); (II) GO: 0040011: locomotion (4.00%); (III) GO: 0002376: immune system process (3.99%); (IV) GO: 0022610: biological adhesion (3.88%), and (V) GO: 0048518: positive regulation of a biological process (2.72%).

Biological adhesion between substrate and cells modulates several critical cellular processes, such as cell locomotion and gene expression [59]. Biological adhesion- and locomotion-related gene dysfunction may result in AMD. Previous research has shown that the immune system, particularly the complement system, is relevant to AMD. Genetic studies also indicate that several complement-related genes, including *CFH*, *complement component 2*, *complement component 3*, *CFHR1,* and *CFHR3*, are highly associated with AMD [60]. Further, complement can enhance the generation

of VEGF (vascular endothelial growth factor), which may strongly facilitate AMD development [61]. Histological studies show the presence of macrophages, lymphocytes, mast cells, and fibroblasts in both atrophic lesions and with retinal neovascularization [61].

(2) *CC GO Terms.* In Figure 5, for the cellular component GO term frequency, the top five CC terms are (I) GO: 0005623: cell (80); (II) GO: 0044464: cell part (64); (III) GO: 0032991: macromolecular complex (27); (IV) GO: 0043226: organelle (24); and (V) GO: 0005576: extracellular region (18). Cell, cell part, organelle, and macromolecular complex inclusion may be attributed to large base numbers of these GO terms.

For the percentage of cellular component terms, as shown in Figure 6, the top five CC terms include (I) GO: 0044420: extracellular matrix part (14.29%); (II) GO: 0031012: extracellular matrix (13.13%); (III) GO: 0044421: extracellular region part (6.94%); (IV) GO: 0005576: extracellular region (6.74%); and (V) GO: 0005623: cell (1.96%).

From the distribution of CC terms, except for the cell term (GO: 0005623), the top four CC terms are associated with the extracellular matrix. Moreover, the extracellular region is relevant to cell adhesion and locomotion, which were mentioned in the biological process GO terms.

The results are also consistent with a recent GWAS study, which identified several new loci with enrichment for genes involved in the extracellular matrix and other activities [18]. Structural damage of extracellular matrix in retinal cells may lead to break point of AMD [62]. Matrix metalloproteinases result in extracellular matrix degradation and are highly related to AMD pathogenesis [63]. Therefore, taken together, these facts suggest that the extracellular matrix plays an important role in AMD.

(3) *MF GO Terms.* In Figure 7, based on the frequency of molecular function terms, the top five MF terms are (I) GO: 0003824: catalytic activity (89); (II) GO: 0005488: binding (72); (III) GO: 0000988: protein binding transcription factor activity (34); (IV) GO: 0004872: receptor activity (34); and (V) GO: 0060089: molecular transducer activity (26).

MF terms related to catalytic activity and binding were highlighted partly due to the large base numbers of these terms. However, this finding may suggest that genes assigned to these two terms are essential to maintain normal function. For example, matrix metalloproteinases, which can degrade extracellular matrix proteins, play an important role in AMD [63]. In addition, highlighting receptor activity and molecular transducer activity indicates that abnormal cellular signal pathway behaviors are involved in AMD patients. For example, the Aryl hydrocarbon receptor, which is responsible for clearing cellular debris and for toxin metabolism, is essential to maintaining normal function in RPE cells, and deficiency of this receptor causes AMD in mice [64].

For the percentage of molecular function terms, as shown in Figure 8, the top five MF terms are (I) GO: 0005198: structural molecule activity (11.76%); (II) GO: 0016209: antioxidant activity (11.54%); (III) GO: 0016247: channel regulator activity (8.33%); (IV) GO: 0030545: receptor regulator activity (5.00%); and (V) GO: 0004872: receptor activity (4.72%). To our surprise, receptor activity was highlighted in both the frequency and percentage of molecular function terms, which is further evidence of the important role that receptor activity plays in AMD. Antioxidant activity is also highlighted, and oxidative stress [6] is a risk factor correlated with AMD. Channel regulator activity and structural molecule activity may also be involved in AMD.

3.3.2. The KEGG Pathways in the Final Optimal Set. Based on the final optimal set, we obtained 11 KEGG pathways, which are (I) hsa00290 (valine, leucine, and isoleucine biosynthesis); (II) has00450 (selenocompound metabolism); (III) hsa00512 (mucin-type O-glycan biosynthesis); (IV) hsa03013 (RNA transport); (V) hsa04145 (phagosome); (VI) hsa04610 (complement and coagulation cascades); (VII) hsa04962 (vasopressin-regulated water reabsorption); (VIII) hsa05133 (pertussis); (IX) hsa05146 (viral myocarditis); and (X) hsa05150 (*Staphylococcus aureus* infection); and (XI) hsa05416 (viral myocarditis).

Valine, leucine, and isoleucine biosynthesis (hsa00290) and selenocompound metabolism (hsa00450) are related to amino acid metabolism. Mucin-type O-glycan biosynthesis is associated with modifications of serine or threonine residues of certain proteins. RNA transport from nucleus to cytoplasm is also essential for gene expression. These terms may not be the key factors in AMD, but they may give us suggestions about the AMD development. Phagosome (hsa04145) is also associated with AMD. There are various forms of cell death and phagocytosis in the retina [65]. But failure of retinal pigment epithelial cells and macrophages to phagocytize dying retinal pigment epithelial cells may result in drusen formation and development of AMD [66]. The underlying mechanism of AMD is still unclear, but many studies have highlighted the essential role of the immune system in the development and progression of AMD [67]. Previous studies have revealed a strong association between complement pathway and AMD [20]. Several complement genes including complement 2 (*C2*) and complement 3 (*C3*) have been strongly associated with AMD [12, 68]. Except vasopressin-regulated water reabsorption, viral myocarditis (hsa05146) and *Staphylococcus aureus* infection (hsa05150) are all correlated with immunity, which further emphasizes the effect of immunity in AMD.

4. Conclusions

In this study, we performed GO and KEGG enrichment analyses of AMD-related genes. The results suggest that 720 GO terms and 11 KEGG pathways are important factors that contribute to identifying AMD-related genes.

Conflict of Interests

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

Acknowledgments

This paper was supported by the National Basic Research Program of China (2011CB510101), Doctoral Innovation Fund of Shanghai Jiaotong University School of Medicine (BXJ201337 and BXJ201234), and National Natural Science Foundation of China (81100679, 81273424, 31371335, and 61202021).

References

[1] D. Pascolini, S. P. Mariotti, G. P. Pokharel et al., "2002 Global update of available data on visual impairment: a compilation of population-based prevalence studies," *Ophthalmic Epidemiology*, vol. 11, no. 2, pp. 67–115, 2004.

[2] J. Mitchell and C. Bradley, "Quality of life in age-related macular degeneration: a review of the literature," *Health and Quality of Life Outcomes*, vol. 4, article 97, 2006.

[3] R. Kawasaki, M. Yasuda, S. J. Song et al., "The prevalence of age-related macular degeneration in Asians: a systematic review and meta-analysis," *Ophthalmology*, vol. 117, no. 5, pp. 921–927, 2010.

[4] B. Meyer-Ruesenberg and G. Richard, "New insights into the underestimated impairment of quality of life in age-related macular degeneration—a review of the literature," *Klinische Monatsblatter fur Augenheilkunde*, vol. 227, no. 8, pp. 646–652, 2010.

[5] J. Thornton, R. Edwards, P. Mitchell, R. A. Harrison, I. Buchan, and S. P. Kelly, "Smoking and age-related macular degeneration: a review of association," *Eye*, vol. 19, no. 9, pp. 935–944, 2005.

[6] J. G. Hollyfield, V. L. Bonilha, M. E. Rayborn et al., "Oxidative damage-induced inflammation initiates age-related macular degeneration," *Nature Medicine*, vol. 14, no. 2, pp. 194–198, 2008.

[7] S. Beatty, H. Koh, M. Phil, D. Henson, and M. Boulton, "The role of oxidative stress in the pathogenesis of age-related macular degeneration," *Survey of Ophthalmology*, vol. 45, no. 2, pp. 115–134, 2000.

[8] J. K. Shen, A. Dong, S. F. Hackett, W. R. Bell, W. R. Green, and P. A. Campochiaro, "Oxidative damage in age-related macular degeneration," *Histology and Histopathology*, vol. 22, no. 12, pp. 1301–1308, 2007.

[9] U. Chakravarthy, T. Y. Wong, A. Fletcher, E. Piault, and C. Evans, "Clinical risk factors for age-related macular degeneration: a systematic review and meta-analysis," *BMC Ophthalmology*, vol. 10, article 31, 2010.

[10] S. K. Iyengar, D. Song, B. E. K. Klein et al., "Dissection of genomewide-scan data in extended families reveals a major locus and oligogenic susceptibility for age-related macular degeneration," *The American Journal of Human Genetics*, vol. 74, no. 1, pp. 20–39, 2004.

[11] D. E. Weeks, Y. P. Conley, H. Tsai et al., "Age-related maculopathy: A genomewide scan with continued evidence of susceptibility loci within the 1q31, 10q26, and 17q25 regions," *The American Journal of Human Genetics*, vol. 75, no. 2, pp. 174–189, 2004.

[12] B. Gold, J. E. Merriam, J. Zernant et al., "Variation in factor B (*BF*) and complement component 2 (*C2*) genes is associated with age-related macular degeneration," *Nature Genetics*, vol. 38, no. 4, pp. 458–462, 2006.

[13] A. O. Edwards, R. Ritter III, K. J. Abel, A. Manning, C. Panhuysen, and L. A. Farrer, "Complement factor H polymorphism and age-related macular degeneration," *Science*, vol. 308, no. 5720, pp. 421–424, 2005.

[14] J. A. Fagerness, J. B. Maller, B. M. Neale, R. C. Reynolds, M. J. Daly, and J. M. Seddon, "Variation near complement factor I is associated with risk of advanced AMD," *European Journal of Human Genetics*, vol. 17, no. 1, pp. 100–104, 2009.

[15] B. M. Neale, J. Fagerness, R. Reynolds et al., "Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (LIPC)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 16, pp. 7395–7400, 2010.

[16] W. Chen, D. Stambolian, A. O. Edwards et al., "Genetic variants near *TIMP3* and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 16, pp. 7401–7406, 2010.

[17] S. Arakawa, A. Takahashi, K. Ashikawa et al., "Genome-wide association study identifies two susceptibility loci for exudative age-related macular degeneration in the Japanese population," *Nature Genetics*, vol. 43, no. 10, pp. 1001–1004, 2011.

[18] L. G. Fritsche, W. Chen, M. Schu et al., "Seven new loci associated with age-related macular degeneration," *Nature Genetics*, vol. 45, pp. 433–439, 2013.

[19] R. J. Klein, C. Zeiss, E. Y. Chew et al., "Complement factor H polymorphism in age-related macular degeneration," *Science*, vol. 308, no. 5720, pp. 385–389, 2005.

[20] A. Swaroop, E. Y. Chew, C. B. Rickman, and G. R. Abecasis, "Unraveling a multifactorial late-onset disease: from genetic susceptibility to disease mechanisms for age-related macular degeneration," *Annual Review of Genomics and Human Genetics*, vol. 10, pp. 19–43, 2009.

[21] M. B. Gorin, "Genetic insights into age-related macular degeneration: controversies addressing risk, causality, and therapeutics," *Molecular Aspects of Medicine*, vol. 33, no. 4, pp. 467–486, 2012.

[22] R. R. Priya, E. Y. Chew, and A. Swaroop, "Genetic studies of age-related macular degeneration: lessons, challenges, and opportunities for disease management," *Ophthalmology*, vol. 119, no. 12, pp. 2526–2536, 2012.

[23] Consortium GO, "The gene ontology (GO) project in 2006," *Nucleic Acids Research*, vol. 34, pp. D322–D326, 2006.

[24] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[25] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.

[26] Z. Yang, N. J. Camp, H. Sun et al., "A variant of the *HTRA1* gene increases susceptibility to age-related macular degeneration," *Science*, vol. 314, no. 5801, pp. 992–993, 2006.

[27] A. DeWan, M. Liu, S. Hartman et al., "*HTRA1* promoter polymorphism in wet age-related macular degeneration," *Science*, vol. 314, no. 5801, pp. 989–992, 2006.

[28] T. P. Dryja, C. E. Briggs, E. L. Berson, P. J. Rosenfeld, and M. Abitbol, "ABCR gene and age-related macular degeneration," *Science*, vol. 279, article 1107, 1998.

[29] A. E. Hughes, N. Orr, H. Esfandiary, M. Diaz-Torres, T. Goodship, and U. Chakravarthy, "A common *CFH* haplotype, with deletion of *CFHR1* and *CFHR3*, is associated with lower risk of age-related macular degeneration," *Nature Genetics*, vol. 38, no. 10, pp. 1173–1177, 2006.

[30] L. G. Fritsche, N. Lauer, A. Hartmann et al., "An imbalance of human complement regulatory proteins CFHR1, CFHR3 and

factor H influences risk for age-related macular degeneration (AMD)," *Human Molecular Genetics*, vol. 19, no. 23, pp. 4694–4704, 2010.

[31] Z. He, T. Huang, X. Shi et al., "Computational analysis of protein tyrosine nitration," in *Proceedings of the 4th International Conference on Computational Systems Biology (ISB '10)*, pp. 35–42, 2010.

[32] L. Chen, Z. Qian, K. Fen, and Y. Cai, "Prediction of interactiveness between small molecules and enzymes by combining gene ontology and compound similarity," *Journal of Computational Chemistry*, vol. 31, no. 8, pp. 1766–1776, 2010.

[33] P. Carmona-Saez, M. Chagoyen, F. Tirado, J. M. Carazo, and A. Pascual-Montano, "GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists," *Genome Biology*, vol. 8, no. 1, article R3, 2007.

[34] T. Huang, L. Chen, Y. Cai, and K. Chou, "Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property," *PLoS ONE*, vol. 6, no. 9, Article ID e25297, 2011.

[35] T. Huang, J. Zhang, Z. Xu et al., "Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches," *Biochimie*, vol. 94, no. 4, pp. 1017–1025, 2012.

[36] L. Chen, B.-Q. Li, and K.-Y. Feng, "Predicting biological functions of protein complexes using graphic and functional features," *Current Bioinformatics*, vol. 8, pp. 545–551, 2013.

[37] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2005.

[38] L. Chen, L. Lu, K. Feng et al., "Multiple classifier integration for the prediction of protein structural classes," *Journal of Computational Chemistry*, vol. 30, no. 14, pp. 2248–2254, 2009.

[39] B. Li, K. Feng, L. Chen, T. Huang, and Y. Cai, "Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS," *PLoS ONE*, vol. 7, no. 8, Article ID e43927, 2012.

[40] M. Shugay, I. O. de Mendibil, J. L. Vizmanos, and F. J. Novo, "Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions," *Bioinformatics*, vol. 29, pp. 2539–2546, 2013.

[41] A. Holzinger and M. Zupan, "KNODWAT: a scientific framework application for testing knowledge discovery methods for the biomedical domain," *BMC Bioinformatics*, vol. 14, no. 1, article 91, 2013.

[42] C. Yan, D. Dobbs, and V. Honavar, "A two-stage classifier for identification of protein-protein interface residues," *Bioinformatics*, vol. 20, no. 1, pp. i371–i378, 2004.

[43] J. Platt, Ed., *Fast Training of Support Vector Machines Using Sequential Minimal Optimization*, MIT Press, Cambridge, Mass, USA, 1998.

[44] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," *Neural Computation*, vol. 13, no. 3, pp. 637–649, 2001.

[45] L. Chen, K. Feng, Y. Cai, K. Chou, and H. Li, "Predicting the network of substrate-enzyme-product triads by combining compound similarity and functional domain composition," *BMC Bioinformatics*, vol. 11, article 293, 2010.

[46] . Baldi P, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, pp. 412–424, 2000.

[47] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[48] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI '95)*, pp. 1137–1143, San Mateo, Calif, USA.

[49] L. Chen, W. Zeng, Y. Cai, K. Feng, and K. Chou, "Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities," *PLoS ONE*, vol. 7, no. 4, Article ID e35254, 2012.

[50] L. Chen, J. Lu, N. Zhang, T. Huang, and Y.-D. Cai, "A hybrid method for prediction and repositioning of drug Anatomical Therapeutic Chemical classes," *Molecular BioSystems*, vol. 10, pp. 868–877, 2014.

[51] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[52] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, vol. 2 of *Inference and Relationship*, Macmillan, New York, NY, USA, 1973.

[53] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundanc," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[54] K. M. Harrison, T. Kajese, H. I. Hall, and R. Song, "Risk factor redistribution of the national HIV/AIDS surveillance data: an alternative approach," *Public Health Reports*, vol. 123, no. 5, pp. 618–627, 2008.

[55] L. Chen, W. Zeng, Y. Cai, and T. Huang, "Prediction of metabolic pathway using graph property, chemical functional group and chemical structural set," *Current Bioinformatics*, vol. 8, no. 2, pp. 200–207, 2013.

[56] Y. Zhang, C. Ding, and T. Li, "Gene selection algorithm by combining reliefF and mRMR," *BMC Genomics*, vol. 9, no. 2, article S27, 2008.

[57] X. Zhou, Z. Dai, and X. Zou, "Classification of G-protein coupled receptors based on support vector machine with maximum relevance minimum redundancy and genetic algorithm," *BMC Bioinformatics*, vol. 11, article 325, 2010.

[58] L. Lu, X. Shi, S. Li et al., "Protein sumoylation sites prediction based on two-stage feature selection," *Molecular Diversity*, vol. 14, no. 1, pp. 81–86, 2010.

[59] R. L. Juliano and S. Haskill, "Signal transduction from the extracellular matrix," *Journal of Cell Biology*, vol. 120, no. 3, pp. 577–585, 1993.

[60] D. H. Anderson, M. J. Radeke, N. B. Gallo et al., "The pivotal role of the complement system in aging and age-related macular degeneration: Hypothesis re-visited," *Progress in Retinal and Eye Research*, vol. 29, no. 2, pp. 95–112, 2010.

[61] S. M. Whitcup, A. Sodhi, J. P. Atkinson et al., "The role of the immune response in age-related macular degeneration," *International Journal of Inflammation*, vol. 2013, Article ID 348092, 10 pages, 2013.

[62] M. R. Al-Ubaidi, M. I. Naash, and S. M. Conley, "A perspective on the role of the extracellular matrix in progressive retinal degenerative disorders," *Investigative Ophthalmology & Visual Science*, vol. 54, pp. 8119–8124, 2013.

[63] R. Liutkeviciene, V. Lesauskaite, G. Sinkunaite-Marsalkiene et al., "The role of matrix metalloproteinases polymorphisms in age-related macular degeneration," *Ophthalmic Genetics*, 2013.

[64] P. Hu, R. Herrmann, A. Bednar, P. Saloupis, and M. A. Dwyer, "Aryl hydrocarbon receptor deficiency causes dysregulated cellular matrix metabolism and age-related macular degeneration-like pathology," *Proceedings of the National Academy of Sciences*, vol. 110, pp. E4069–E4078, 2013.

[65] W. R. Green and C. Enger, "Age-related macular degeneration histopathologic studies: the 1992 Lorenz E. Zimmerman Lecture. 1992.," *Retina (Philadelphia, Pa.)*, vol. 25, no. 5, pp. 1519–1535, 2005.

[66] J. V. Forrester, "Macrophages eyed in macular degeneration," *Nature Medicine*, vol. 9, no. 11, pp. 1350–1351, 2003.

[67] J. Ambati, J. P. Atkinson, and B. D. Gelfand, "Immunology of age-related macular degeneration," *Nature Reviews Immunology*, vol. 13, no. 6, pp. 438–451, 2013.

[68] J. R. W. Yates, T. Sepp, B. K. Matharu et al., "Complement C3 variant and the risk of age-related macular degeneration," *The New England Journal of Medicine*, vol. 357, no. 6, pp. 553–561, 2007.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 424509, 10 pages
http://dx.doi.org/10.1155/2014/424509

Research Article

Identifying the Gene Signatures from Gene-Pathway Bipartite Network Guarantees the Robust Model Performance on Predicting the Cancer Prognosis

Li He,[1] Yuelong Wang,[2] Yongning Yang,[2] Liqiu Huang,[2] and Zhining Wen[2]

[1] *Biogas Institute of Ministry of Agriculture, Chengdu 610041, China*
[2] *College of Chemistry, Sichuan University, Chengdu 610064, China*

Correspondence should be addressed to Zhining Wen; w_zhining@163.com

Received 16 April 2014; Revised 21 June 2014; Accepted 24 June 2014; Published 14 July 2014

Academic Editor: Mingyue Zheng

For the purpose of improving the prediction of cancer prognosis in the clinical researches, various algorithms have been developed to construct the predictive models with the gene signatures detected by DNA microarrays. Due to the heterogeneity of the clinical samples, the list of differentially expressed genes (DEGs) generated by the statistical methods or the machine learning algorithms often involves a number of false positive genes, which are not associated with the phenotypic differences between the compared clinical conditions, and subsequently impacts the reliability of the predictive models. In this study, we proposed a strategy, which combined the statistical algorithm with the gene-pathway bipartite networks, to generate the reliable lists of cancer-related DEGs and constructed the models by using support vector machine for predicting the prognosis of three types of cancers, namely, breast cancer, acute myeloma leukemia, and glioblastoma. Our results demonstrated that, combined with the gene-pathway bipartite networks, our proposed strategy can efficiently generate the reliable cancer-related DEG lists for constructing the predictive models. In addition, the model performance in the swap analysis was similar to that in the original analysis, indicating the robustness of the models in predicting the cancer outcomes.

1. Introduction

In the past decade, DNA microarray technology has been widely used in clinical researches to predict the cancer outcomes because of its capability of monitoring tens of thousands of genes simultaneously [1–9]. Accurately identifying the genes, for which the changes of their expression levels are significantly correlated with the phenotypic differences between the clinical conditions, plays an important role in the procedure of clinical model construction. The statistical methods and the machine learning algorithms that are routinely used for gene selection mainly identify the differentially expressed genes (DEGs) according to the changes of the gene expression levels between the compared biological samples. However, because of the heterogeneity of the clinical samples, the changes of the gene expression levels may be not only caused by the changes in the status of the cancer cells but also by those of the cells unrelated to the cancers. In addition, the intensities detected by the microarrays for a gene will vary to some extent among the technical replicates due to the complex procedures of the microarray experiment, such as labeling, hybridization, and scanning. Consequently, the DEG list generated by the statistical methods or the machine learning algorithms often involves a number of false positive genes, which are not associated with the phenotypic differences between the compared clinical samples, and subsequently impacts the reliability of the predictive models.

The network-based methodologies can efficiently integrate the biological information with the computational techniques and link the disease-related genes to relevant proteins and disease types. In recent years, the network-based methodologies have been successfully introduced into the systems biology researches for drug discovering [2], identifying disease-related genes [10–15], and revealing the molecular mechanisms of tumorigenesis [16–19]. In clinical researches, the prediction models constructed with the cancer-related

gene markers, which were selected only by the statistical methods or the machine learning algorithms, cannot ensure the accuracy and the reproducibility in predicting the clinical outcomes of cancer patients. Therefore, for the sake of improving the model performance and interpreting the biological relevance of the gene markers and the specific cancer, a number of network-based algorithms were developed to prioritize the prognostic genes [20–38].

In our study, a new strategy, which combined the statistical algorithm with a gene-pathway bipartite network, was proposed to prioritize the reliable gene signatures and the supported vector machine (SVM) was used to construct the models for predicting the clinical outcomes of cancer patients. The DEG list was firstly generated by the statistical methods, for example, Student's *t*-test. Then, the bipartite network that connected the genes and the cancer-related pathways was constructed to score each of the DEGs according to its connectivity in the network. Finally, the DEGs were ranked by the scores in descending order and those, for which the scores were greater than a given cutoff, were selected as features for predicting the cancer prognosis.

To evaluate the performance of the predictive models with the gene signatures generated by our strategy, three data sets including the gene expression data of the clinical samples collected from the patients of breast cancer, acute myeloma leukemia, and glioblastoma were downloaded from the gene expression omnibus (GEO) database. Gene signatures separately identified from these data sets by our strategy were used as features to predict the reoperative treatment response of breast cancer, the overall survival milestone outcome of acute myeloma leukemia, and the molecular subclasses of high-grade glioblastoma. The results of predicting the reoperative treatment response of breast cancer and the overall survival milestone outcome of acute myeloma leukemia showed that our models performed better than those reported by the data contributors. In addition, the accuracy of predicting the molecular subclasses of high-grade glioblastoma was as high as 87.5%. In the swap analysis, we repeated the model construction and validation process by training the models with the original independent test set and validating them using the original training set with the same gene signatures prioritized in the original analysis. The prediction results were similar to those achieved in original analysis, indicating the robust model performance on predicting the cancer prognosis when using the gene signatures identified by our proposed strategy.

2. Materials and Methods

2.1. Data Sets. All microarray gene expression data (series MATRIX files) generated from the clinical samples of breast cancer, acute myeloma leukemia, and glioblastoma and the corresponding clinical information were downloaded from the National Center for Biotechnology Information's Gene Expression Omnibus (GEO) database (series accession numbers: GSE16716, GSE12417, and GSE13041).

In the human breast cancer data set [4, 9], the gene expression data of 230 clinical samples were generated by using Affymetrix Human Genome U133A (HG-U133A) microarrays, which included 22,283 probesets. In light of the data analysis protocol in MicroArray Quality Control (MAQC)-II Project [9], the gene expression data generated from 130 out of 230 clinical samples of breast cancer patients were used as training set and the rest of the 100 cases were used as independent test set. The response to preoperative chemotherapy, which was divided into two subcategories of no residual invasive cancer in the breast or lymph nodes (pCR) and residual invasive cancer (RD), was used as the clinical endpoint for prediction [4].

The acute myeloma leukemia data set included the gene expression profiling of the clinical samples of 242 patients with cytogenetically normal acute myeloid leukemia (CN-AML) [39]. In the training set, the gene expression data of 163 clinical samples were generated by using Affymetrix Human Genome U133A&B (HG-U133A&B) microarrays, which included a total of 44,760 probesets. The gene expression data of 79 clinical samples in the independent test set were generated by using HG-U133Plus2 microarrays. During the calculation procedure, we only used 44,693 common probesets between HG-U133A&B chips and HG-U133Plus2 chips for the DEG selection and predictive model construction. The clinical endpoint of overall survival times was dichotomized with a "milestone" cutoff because the continuous endpoint values cannot be predicted by the binary classification models. By considering the balance between the number of positive samples and that of negative samples, the patients with the survival time less than one year were categorized into the "high-risk" group and the rest with the survival time equal to or longer than one year were assigned to the "low-risk" group. In addition, a patient was excluded from the data set if the survival time was less than one-year milestone cutoff and censored when he/she was still alive. Eventually, there were 152 patients in the training set and 77 patients in the independent test set.

The gene expression data in the glioblastoma data set [7] were generated by using HG-U133A microarrays. In glioblastoma research, a subcategory of glioblastoma termed ProNeural (PN) was highly related to better survival prognosis when compared to other subcategories [6]. In our study, we collected 50 patients belonging to the PN subcategory with a mean survival of 924 days and 50 patients belonging to non-PN subcategory with a mean survival of 150 days. Among these patients, 60 of them, which included 30 patients in PN subcategory and 30 in non-PN subcategory, were randomly assigned to the training set and the rest were used as the independent test set. A predictive model was constructed to discriminate the PN and non-PN categories based on the microarray gene expression data.

2.2. Probesets Mapping. For Affymetrix microarray platforms, a gene may be detected by multiple probesets. Before identifying DEGs, we mapped the multiple probesets to a unique HUGO gene symbol by using the probeset with the highest fold change value between two groups of samples. Accordingly, 22,283 probesets involved in the data sets of breast cancer and glioblastoma were mapped to 11,285 unique

genes and 44,693 common probesets involved in the acute myeloma leukemia data set were mapped to 14,892 unique genes, respectively.

2.3. Identification of Differentially Expressed Genes. Student's t-test, which can assess how significant a gene is differentially expressed in two compared phenotypes, was used in our study for the DEG selection. The P value for each of the genes was calculated by t-test and directly used for gene filtering without multiple-testing correction. Only the genes with $P < 0.05$ were kept. To ensure the reproducibility of the DEG lists generated by the t-test, a fold change ranking is usually applied to refining the genes with $P < 0.05$. These genes were ranked by their fold changes (the expression intensity of a gene in sample A/its expression intensity in sample B). Only the genes with fold change >1.5 (FC > 1.5) or fold change <0.667 (FC < 0.667) were kept for the subsequent analysis. Note that, in some microarray studies of clinical samples, only a few genes can meet the fold change cutoff because of the minor phenotypic differences between the two groups of clinical samples.

2.4. Construction of Gene-Pathway Bipartite Network. For the purpose of screening out the genes unassociated with the phenotypic differences, we constructed a gene-pathway bipartite network, which can be used to score the genes according to their connections [40] to the cancer-related signaling pathways. All the cancer-related pathways were collected from Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database and listed in Table 1. The first six pathways reflected the overview of cancers and the rest were correlated with the specific types of cancers.

The bipartite network was a particular class of complex networks, in which the nodes were divided into two groups and the connections only existed between two nodes in different groups [41]. So, the nodes in the gene-pathway bipartite network were genes or pathways and were divided into two groups of gene set and pathway set, respectively. The connections between genes and pathways indicated (1) which genes were involved in a specific pathway and (2) which pathways included a specific gene. We scored each of the genes with a weighting method proposed by Zhou et al. [42]. Let us consider a gene-pathway bipartite network $N(\mathbf{G}, \mathbf{P}, \mathbf{E})$, where G and P represent the gene set and pathway set, respectively. $\mathbf{E}$ is the set of connections between genes and pathways. The genes and pathways in $\mathbf{G}$ and $\mathbf{P}$ were denoted by g_1, g_2, …, g_n and p_1, p_2, …, p_m, respectively. The initial score s_0 assigned to each of the genes in $\mathbf{G}$ was set to 1. In the first step, we calculated the weights W ($W = \{w_1, w_2, \ldots, w_m\}$) for the pathways via

$$w_l = \sum_{j=1}^{n} \frac{a_{jl} s_0}{k(g_j)} \quad (l = 1, 2, \ldots, \text{the number of pathways in } P), \tag{1}$$

where $k(g_j)$ was the degree of the jth gene and a_{jl} was the $n \times m$ adjacent matrix:

$$a_{jl} = \begin{cases} 1, & g_j p_l \in E, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

In the second step, we calculated the final scores S ($S = \{s_1, s_2, \ldots, s_n\}$) for all the genes via

$$s_i = \sum_{l=1}^{m} \frac{a_{il} w_l}{k(p_l)} \quad (i = 1, 2, \ldots, \text{the number of genes in } G), \tag{3}$$

where $k(p_l)$ and w_l were the degree and the weight of the lth pathway, respectively, and a_{il} was the $n \times m$ adjacent matrix:

$$a_{il} = \begin{cases} 1, & g_i p_l \in E, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

By combining (1) and (3), we can directly calculate the scores for the genes via

$$\begin{aligned} s_i &= \sum_{l=1}^{m} \frac{a_{il}}{k(p_l)} \sum_{j=1}^{n} \frac{a_{jl} s_0}{k(g_j)} \\ &= \sum_{j=1}^{n} c_{ij} s_0 \quad (i = 1, 2, \ldots, \text{the number of genes in } G), \end{aligned} \tag{5}$$

where

$$c_{ij} = \frac{1}{k(g_j)} \sum_{l=1}^{m} \frac{a_{il} a_{jl}}{k(p_l)}. \tag{6}$$

The matrix $\mathbf{C} = \{c_{ij}\}_{n \times n}$ represented the weighted $\mathbf{G}$ projection. In our study, the DEGs were firstly selected by the statistical methods and subsequently ranked by the scores $\mathbf{S}$ ($\mathbf{S} = \{s_1, s_2, \ldots, s_n\}$). Only the DEGs with $s \geq 1$ were kept as features for the construction of the predictive models.

2.5. Model Construction for Clinical Endpoints Prediction. The binary classification models for predicting the clinical endpoints were constructed by using support vector machine (SVM), which is a popular learning machine based on statistical learning theory [43, 44]. In our study, radial basic function was used as the kernel function in SVM. The regularization parameter c and the kernel width parameter σ were optimized by a grid search approach. For each of the clinical endpoints, the SVM model was built by using the training set and leave-one-out cross-validation and validated by the independent test set. Four performance metrics, namely, specificity, sensitivity, accuracy, and Matthew's correlation

Table 1: The 20 cancer-related signaling pathways collected from KEGG database for the construction of gene-pathway bipartite network.

Pathway entry	KEGG pathway name	Number of genes
hsa05200	Pathways in cancer	327
hsa05202	Transcriptional misregulation in cancer	179
hsa05203	Viral carcinogenesis	206
hsa05204	Chemical carcinogenesis	80
hsa05205	Proteoglycans in cancer	225
hsa05206	MicroRNAs in cancer	296
hsa05210	Colorectal cancer	62
hsa05211	Renal cell carcinoma	66
hsa05212	Pancreatic cancer	66
hsa05213	Endometrial cancer	52
hsa05214	Glioma	65
hsa05215	Prostate cancer	89
hsa05216	Thyroid cancer	29
hsa05217	Basal cell carcinoma	55
hsa05218	Melanoma	71
hsa05219	Bladder cancer	38
hsa05220	Chronic myeloid leukemia	73
hsa05221	Acute myeloid leukemia	57
hsa05222	Small cell lung cancer	86
hsa05223	Non-small-cell lung cancer	56

coefficient (MCC), were considered for model evaluation and defined as follows:

$$\begin{aligned}
\text{Specificity} &= \frac{\text{TN}}{\text{FP} + \text{TN}},\\
\text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}},\\
\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}},\\
\text{MCC} &= (\text{TP} \times \text{TN} - \text{FP} \times \text{FN})\\
&\quad \times ((\text{TP} + \text{FP}) \times (\text{TP} + \text{FN})\\
&\quad \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN}))^{-1/2},
\end{aligned} \tag{7}$$

where TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively. In addition, the areas under the ROC curves (AUCs) were also provided for evaluating the performance of the models on the prediction of the survival milestone outcomes of AML patients and the molecular subclasses of high-grade glioblastoma. The software libsvm 3.17 [45] used in our study for SVM modeling can be freely downloaded from the website http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

3. Results

3.1. Model Performance on Predicting the Reoperative Treatment Response of Breast Cancer. According to the data analysis protocol in MAQC-II project, 130 clinical samples of breast cancer patients were assigned to training set and the rest of the 100 clinical samples were used as independent test set. By comparing the gene expression profiles of the samples in pCR subcategory with those in RD subcategory in training set, 1010 genes with P value < 0.05 and $|\log_2 \text{FC}| > 0.585$ were selected as DEGs and used to construct the gene-pathway bipartite network. Based on the connections between the DEGs and the cancer-related KEGG pathways, 1010 DEGs were scored by a weighted method and then 29 DEGs with the scores ≥1 were kept as features for model construction. The gene-pathway bipartite network, which connected the 29 DEGs with the 20 cancer-related KEGG pathways, was shown in Figure 1. It can be seen from Figure 1 that the gene *CCND1*, which was ranked 1st in the 1010 DEGs, had the most connections to the cancer-related pathways, indicating it was an important feature for the prediction of the clinical endpoint of breast cancer.

A SVM model was constructed by using the training set and leave-one-out cross-validation. The best parameters of c and σ were 2 and 0.03125, respectively. The prediction results of training set and independent test set were listed in Table 2. In swap analysis, we repeated the model construction and validation process by training the models with the original independent test set and validating them using the original training set with the same 29 DEGs identified in the original analysis. Meanwhile, the prediction results achieved by MAQC-II candidate models were also listed in Table 2. Compared with the MAQC-II candidate models, our model was more robust and superior in predicting the breast cancer outcomes.

3.2. Model Performance on Predicting the Overall Survival Milestone Outcome of Acute Myeloma Leukemia. By comparing the gene expression profiles of the clinical samples

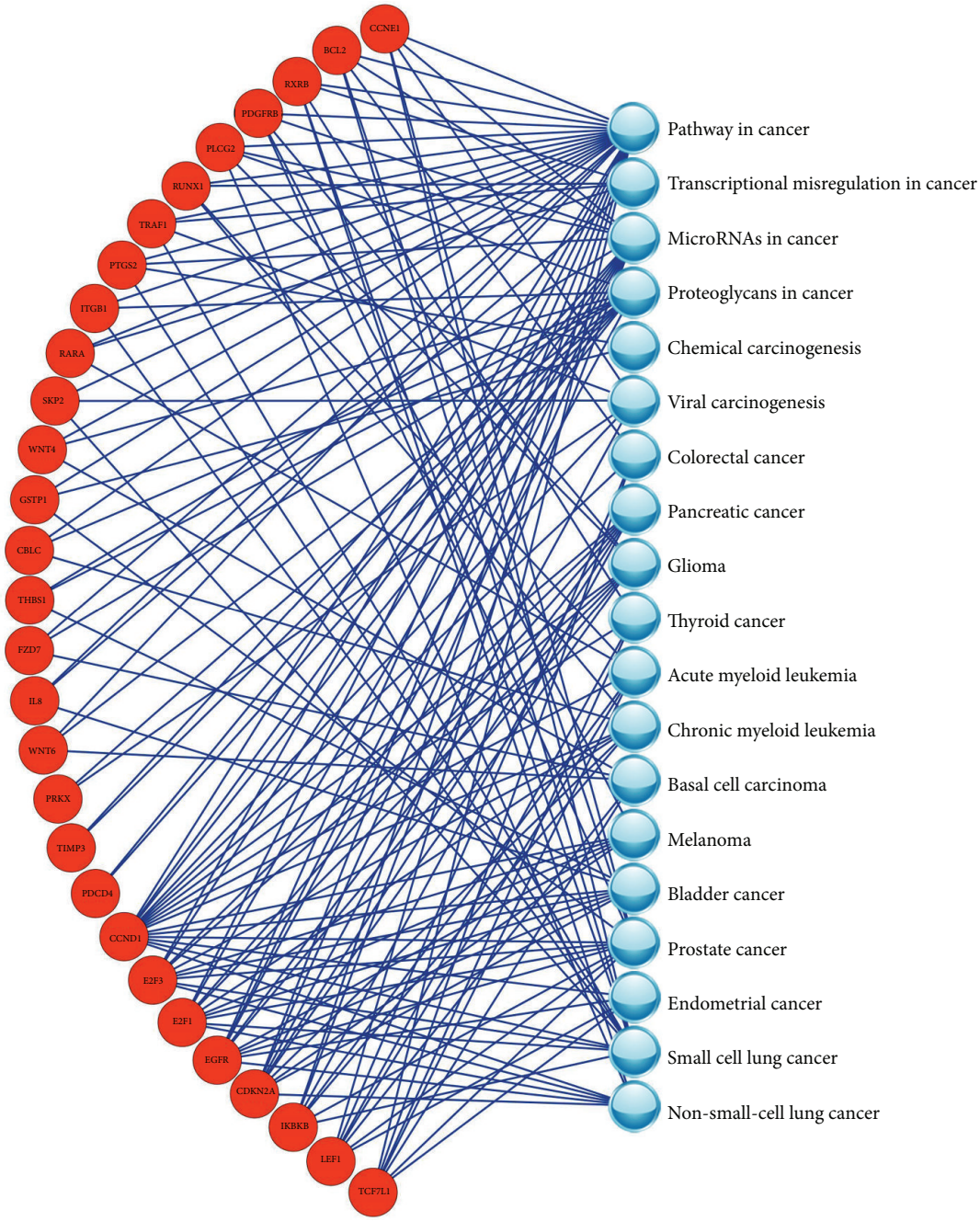


Figure 1: The gene-pathway bipartite network constructed with 29 gene signatures that were used for predicting the reoperative treatment response of breast cancer.

Table 2: The results of predicting the reoperative treatment response of breast cancer in original and swap analyses.

		Our model				MAQC-II candidate model			
		SP	SE	ACC	MCC	SP	SE	ACC	MCC
Original analysis	Training	0.928	0.455	0.808	0.444	0.847	0.569	0.775	0.433
	Validation	0.882	0.467	0.820	0.332	0.729	0.667	0.720	0.301
Swap analysis	Training	0.988	0.200	0.870	0.343	0.899	0.522	0.837	0.454
	Validation	1.000	0.152	0.785	0.343	0.959	0.212	0.769	0.267

In the prediction, pCR was defined as positive sample.
SP, SE, ACC, and MCC represented specificity, sensitivity, accuracy, and Matthew's correlation coefficient, respectively.

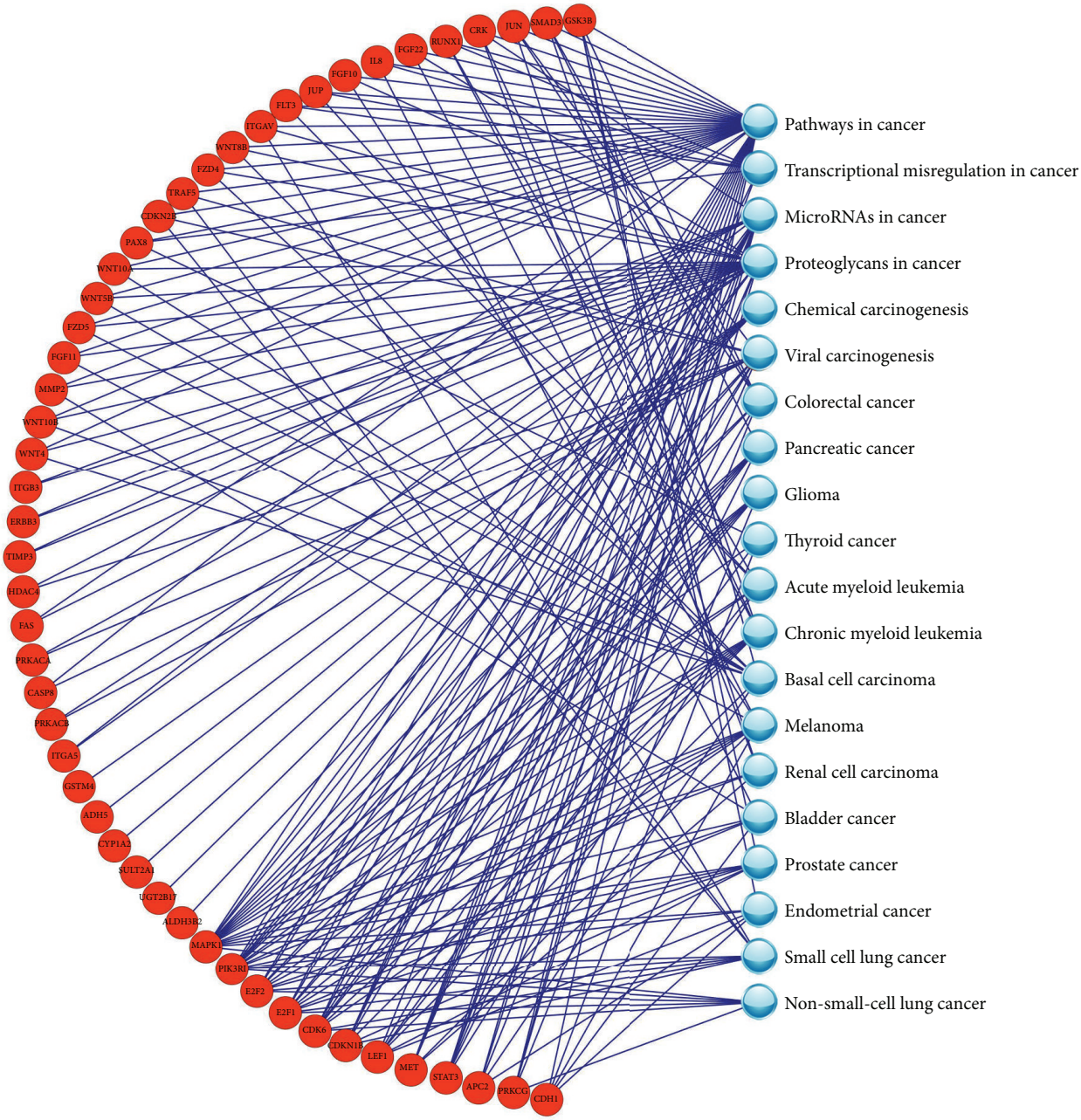


FIGURE 2: The gene-pathway bipartite network constructed with 50 gene signatures that were used for predicting the overall survival milestone outcome of acute myeloma leukemia.

between the high-risk patients and the low-risk patients in the training set, 3234 genes with P value < 0.05 were selected as DEGs. These DEGs were used to construct the gene-pathway bipartite network and ranked by their scores calculated by the weighted method. At last, 50 DEGs with the scores ≥1 were used as features for the subsequent model construction. The gene-pathway bipartite network of 50 DEGs connected to the 20 cancer-related KEGG pathways was shown in Figure 2. The gene *MAPK1* was ranked 1st in the DEG list and had the most connections to the cancer-related pathways.

In both original and swap analyses, the best parameters of c and σ optimized for SVM models were 512 and 0.00195, respectively. The prediction results achieved by our models were listed in Table 3. For the convenience of comparison, we built the SVM models in the original and swap analyses with the expression signatures of 86 probesets proposed by the data contributors [39] and summarized the prediction results (Table 3). In general, our model performed similarly to the 86-probeset model in the original analysis, while the MCC achieved by our model with the validation set in the swap

Table 3: The results of predicting the overall survival milestone outcome of acute myeloma leukemia in original and swap analyses.

		Our model					86-probe-set model				
		SP	SE	ACC	MCC	AUC	SP	SE	ACC	MCC	AUC
Original analysis	Training	0.697	0.837	0.776	0.542	0.776	0.758	0.733	0.743	0.486	0.746
	Validation	0.574	0.600	0.584	0.170	0.587	0.362	0.800	0.532	0.172	0.581
Swap analysis	Training	0.830	0.700	0.779	0.533	0.765	1.000	0.433	0.779	0.564	0.717
	Validation	0.545	0.756	0.664	0.308	0.655	0.712	0.523	0.605	0.236	0.618

In the prediction, high-risk patient was defined as positive sample.
AUC represented the area under the ROC curve.
See notes under Table 2 for more information.

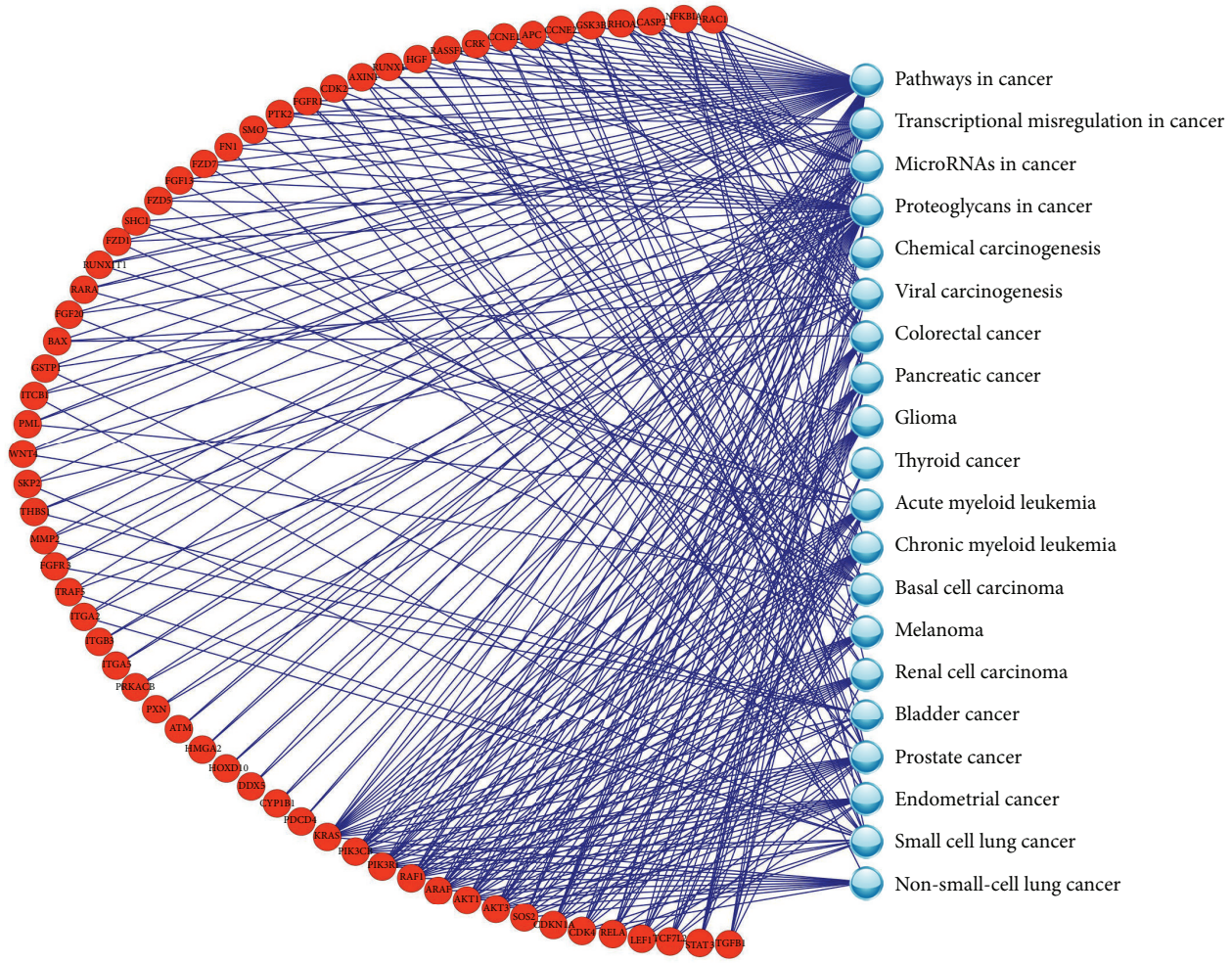


Figure 3: The gene-pathway bipartite network constructed with 62 gene signatures that were used for predicting the molecular subclasses of high-grade glioblastoma.

analysis was 0.308, which was higher than that achieved by the 86-probeset model.

3.3. Model Performance on Predicting the Molecular Subclasses of High-Grade Glioblastoma. For the high-grade glioblastoma data set, 2712 genes with P value < 0.05 were selected as DEGs and 62 of them with scores ≥1 were used to construct the SVM models. The gene-pathway bipartite network of 62 DEGs connected to the 20 cancer-related KEGG pathways was shown in Figure 3. The gene *KRAS* was ranked 1st in the DEG list and had the most connections to the cancer-related pathways. The best parameters of c and σ optimized for SVM models in original analysis were 8 and 0.00781, respectively, and were 0.5 and 0.125 in swap analysis, respectively. The prediction results in original and swap analyses were listed in Table 4. In the original analysis, the prediction accuracy

Table 4: The results of predicting the molecular subclasses of high-grade glioblastoma in original and swap analyses.

		Our model				
		SP	SE	ACC	MCC	AUC
Original analysis	Training	0.900	0.900	0.900	0.800	0.900
	Validation	0.750	1.000	0.875	0.775	0.875
Swap analysis	Training	0.950	0.900	0.925	0.851	0.925
	Validation	0.800	0.567	0.683	0.377	0.684

In the prediction, the gene expression profile termed ProNeural (PN) was defined as positive sample.
AUC represented the area under the ROC curve.
See notes under Table 2 for more information.

for the validation set was as high as 87.5% and similar to that achieved in the training procedure, indicating the superior performance of the SVM model in predicting the molecular subclasses of high-grade glioblastoma. In the swap analysis, the MCC for validation set was dropped to 0.377. This was mainly because the number of samples for the model construction was limited.

4. Discussion

In clinical researches, the microarray-based gene expression profiling is often used to construct the models for predicting cancer prognosis. Identifying the DEGs accurately plays a key role in the procedure of clinical model construction. Current statistical methods and machine learning algorithms used for DEG selection only focus on the changes of the gene expression levels between the two groups of clinical samples instead of the causes behind these changes and subsequently result in a number of false positive genes unrelated to the phenotypic differences involved in the DEG list and the predictive models becoming unreliable. In our current study, we described a weighted method, which scored each of the genes according to their connections to the cancer-related pathways in the gene-pathway bipartite network, for the purpose of refining the DEG list generated by the statistical methods. By considering the two facts of (1) how many genes connected to a specific pathway and (2) how many pathways involved a specific gene, all the DEGs in the bipartite network were scored by the weighted method. The DEGs with scores ≥ 1 were considered as the specific cancer-related genes and used to construct the predictive models.

In order to validate the performance of the predictive models, the gene expression data of the clinical samples were collected in our study to predict three clinical endpoints, namely, the reoperative treatment response of breast cancer, the overall survival milestone outcome of acute myeloma leukemia, and the molecular subclasses of high-grade glioblastoma. For the prediction of reoperative treatment response of breast cancer, 29 DEGs were selected from the bipartite network as features to construct SVM models. In both original and swap analyses, our model performed (MCC = 0.332 and 0.343, resp.) better than the MAQC-II candidate model (MCC = 0.301 and 0.267, resp.). Moreover, in the swap analysis, the MCC achieved by our model in training procedure (MCC = 0.343) was equal to that achieved in validation procedure, indicating the robust model performance. When predicting the overall survival milestone outcome of acute myeloma leukemia, the performance of our model with 50 DEGs was similar to that of the 86-probeset model in original analysis. In the swap analysis, the MCC achieved by our model (MCC = 0.308) was higher than that (MCC = 0.236) achieved by the 86-probeset model. As to the prediction of the molecular subclasses of high-grade glioblastoma, 62 DEGs were used for SVM model construction. The accuracy achieved in the original analysis was as high as 87.5%. Meanwhile, the model performance was robust in the original analysis (MCC = 0.800 and 0.775 in training and validation procedures, resp.). Note that, in the swap analysis, the MCC for validation set was only 0.377. This was mainly because the number of samples used for model construction was limited. In the swap analysis, only 40 samples were used to construct the predictive model, which was insufficient to ensure the reliability of the predictive model.

5. Conclusions

In this study, we suggested a strategy to identify the gene signatures, which not only were differentially expressed between two groups of clinical samples but also highly correlated with a specific cancer, from a gene-pathway bipartite network. The predictive models constructed with these gene signatures performed better than those models reported in previous studies. Moreover, in both original and swap analyses, our models achieved similar prediction results, indicating the robust model performance on predicting the cancer prognosis.

Conflict of Interests

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

Acknowledgment

This work was supported by the National Nature Science Foundation of China (nos. 21205085 and 31370060).

References

[1] L. J. van't Veer, H. Dai, M. J. van de Vijver et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[2] E. L. Leung, Z.-W. Cao, Z.-H. Jiang, H. Zhou, and L. Liu, "Network-based drug discovery by integrating systems biology and computational technologies," *Briefings in Bioinformatics*, vol. 14, no. 4, pp. 491–505, 2013.

[3] F. Zhan, J. Hardin, B. Kordsmeier et al., "Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells," *Blood*, vol. 99, no. 5, pp. 1745–1757, 2002.

[4] K. R. Hess, K. Anderson, W. F. Symmans et al., "Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer," *Journal of Clinical Oncology*, vol. 24, no. 26, pp. 4236–4244, 2006.

[5] J. D. Shaughnessy Jr., F. Zhan, B. E. Burington et al., "Avalidated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1," *Blood*, vol. 109, no. 6, pp. 2276–2284, 2007.

[6] H. S. Phillips, S. Kharbanda, R. Chen et al., "Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis," *Cancer Cell*, vol. 9, no. 3, pp. 157–173, 2006.

[7] Y. Lee, A. Scheck, T. Cloughesy et al., "Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age," *BMC Medical Genomics*, vol. 1, no. 1, article 52, 2008.

[8] A. Oberthuer, F. Berthold, P. Warnat et al., "Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification," *Journal of Clinical Oncology*, vol. 24, no. 31, pp. 5070–5078, 2006.

[9] L. Shi, G. Campbell, W. D. Jones et al., "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models," *Nature Biotechnology*, vol. 28, no. 8, pp. 827–838, 2010.

[10] M. A. van Driel, K. Cuelenaere, P. P. C. W. Kemmeren, J. A. M. Leunissen, and H. G. Brunner, "A new web-based data mining tool for the identification of candidate genes for human genetic disorders," *European Journal of Human Genetics*, vol. 11, no. 1, pp. 57–63, 2003.

[11] S. Aerts, D. Lambrechts, S. Maity et al., "Gene prioritization through genomic data fusion," *Nature Biotechnology*, vol. 24, no. 5, pp. 537–544, 2006.

[12] K. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–8690, 2007.

[13] E. Wang, A. Lenferink, and M. O'Connor-McCourt, "Genetic studies of diseases - Cancer systems biology: Exploring cancer-associated genes on cellular networks," *Cellular and Molecular Life Sciences*, vol. 64, no. 14, pp. 1752–1762, 2007.

[14] Y. Moreau and L. C. Tranchevent, "Computational tools for prioritizing candidate genes: boosting disease gene discovery," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 523–536, 2012.

[15] J. I. F. Bass, A. Diallo, and J. Nelson, "Corrigendum: using networks to measure similarity between genes: association index selection," *Nature Methods*, vol. 11, no. 3, pp. 349–349, 2014.

[16] P. Hernández, J. Huerta-Cepas, D. Montaner et al., "Evidence for systems-level molecular mechanisms of tumorigenesis," *BMC Genomics*, vol. 8, article 185, 2007.

[17] F. M. Lopes and J. Cesar, "Gene expression complex networks: synthesis, identification, and analysis," *Journal of Computational Biology*, vol. 18, no. 10, pp. 1353–1367, 2011.

[18] O. Rozenblatt-Rosen, R. C. Deo, M. Padi et al., "Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins," *Nature*, vol. 487, no. 7408, pp. 491–495, 2012.

[19] T. Wang, J. Gu, J. Yuan, R. Tao, Y. Li, and S. Li, "Inferring pathway crosstalk networks using gene set co-expression signatures," *Molecular BioSystems*, vol. 9, no. 7, pp. 1822–1828, 2013.

[20] W. Jiang, X. Li, S. Q. Rao et al., "Constructing disease-specific gene networks using pair-wise relevance metric: application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements," *BMC Systems Biology*, vol. 2, article 72, 2008.

[21] L. Li, K. Zhang, J. Lee, S. Cordes, D. P. Davis, and Z. Tang, "Discovering cancer genes by integrating network and functional properties," *BMC Medical Genomics*, vol. 2, no. 1, article 61, 2009.

[22] T. Milenković, V. Memišević, A. K. Ganesan, and N. Pršulj, "Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data," *Journal of the Royal Society Interface*, vol. 7, no. 44, pp. 423–437, 2010.

[23] G. Östlund, M. Lindskog, and E. L. L. Sonnhammer, "Network-based identification of novel cancer genes," *Molecular and Cellular Proteomics*, vol. 9, no. 4, pp. 648–655, 2010.

[24] L. Agnelli, M. Forcato, F. Ferrari et al., "The reconstruction of transcriptional networks reveals critical genes with implications for clinical outcome of multiple myeloma," *Clinical Cancer Research*, vol. 17, no. 23, pp. 7402–7412, 2011.

[25] J. Ahn, Y. Yoon, C. Park, E. Shin, and S. Park, "Integrative gene network construction for predicting a set of complementary prostate cancer genes," *Bioinformatics*, vol. 27, no. 13, pp. 1846–1853, 2011.

[26] L. Chen, J. Xuan, R. B. Riggins, R. Clarke, and Y. Wang, "Identifying cancer biomarkers by network-constrained support vector machines," *BMC Systems Biology*, vol. 5, article 161, 2011.

[27] J. X. Wang, G. Chen, M. Li, and Y. Pan, "Integration of breast cancer gene signatures based on graph centrality," *BMC Systems Biology*, vol. 5, no. 3, article S10, 2011.

[28] M. D'Antonio, V. Pendino, S. Shruti, and F. D. Ciccarelli, "Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes," *Nucleic Acids Research*, vol. 40, no. 1, pp. D978–D983, 2012.

[29] J. Roy, C. Winter, Z. Isik, and M. Schroeder, "Network information improves cancer outcome prediction," *Briefings in Bioinformatics*, 2012.

[30] C. Staiger, S. Cadot, R. Kooter et al., "A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer," *PLoS ONE*, vol. 7, no. 4, Article ID e34796, 2012.

[31] C. Winter, G. Kristiansen, S. Kersting et al., "Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes," *PLoS Computational Biology*, vol. 8, no. 5, Article ID e1002511, 2012.

[32] C. Wu, D. D'Argenio, S. Asgharzadeh, and T. Triche, "TARGET-gene: a tool for identification of potential therapeutic targets in cancer," *PLoS ONE*, vol. 7, no. 8, Article ID e43305, 2012.

[33] G. Wu and L. Stein, "A network module-based method for identifying cancer prognostic signatures," *Genome Biology*, vol. 13, no. 12, article R112, 2012.

[34] Y. Chen, J. Hao, W. Jiang et al., "Identifying potential cancer driver genes by genomic data integration," *Scientific Reports*, vol. 3, p. 3538, 2013.

[35] Y. Nie and J. Yu, "Mining breast cancer genes with a network based noise-tolerant approach," *BMC Systems Biology*, vol. 7, article 49, 2013.

[36] H. Fröhlich, "Including network knowledge into Cox regression models for biomarker signature discovery," *Biometrical Journal*, vol. 56, no. 2, pp. 287–306, 2014.

[37] L. Jiang, L. Huang, Q. Kuang et al., "Improving the prediction of chemotherapeutic sensitivity of tumors in breast cancer via optimizing the selection of candidate genes," *Computational Biology and Chemistry*, vol. 49, pp. 71–78, 2014.

[38] A. Chaiboonchoe, S. Samarasinghe, D. Kulasiri, and K. Salehi-Ashtiani, "Integrated analysis of gene network in childhood leukemia from microarray and pathway databases," *BioMed Research International*, vol. 2014, Article ID 278748, 7 pages, 2014.

[39] K. H. Metzeler, M. Hummel, C. D. Bloomfield et al., "An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia," *Blood*, vol. 112, no. 10, pp. 4193–4201, 2008.

[40] J. Li, N. Zhang, Z. Liu, and G. Zhao, "Based on bipartite graph label gene extraction algorithm of network structure," *International Journal of Biology*, vol. 3, no. 4, p. 64, 2011.

[41] P. Holme, F. Liljeros, C. R. Edling, and B. J. Kim, "Network bipartivity," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 68, no. 5, part 2, Article ID 056107, 2003.

[42] T. Zhou, J. Ren, M. Medo, and Y. Zhang, "Bipartite network projection and personal recommendation," *Physical Review E*, vol. 76, no. 4, Article ID 046115, 2007.

[43] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[44] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, NY, USA, 1998.

[45] C. Chang and C. Lin, "LIBSVM: a Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 172620, 10 pages
http://dx.doi.org/10.1155/2014/172620

Research Article

Risk Factors for Mortality in Patients with Septic Acute Kidney Injury in Intensive Care Units in Beijing, China: A Multicenter Prospective Observational Study

Xin Wang,[1,2] Li Jiang,[1] Ying Wen,[1] Mei-Ping Wang,[1] Wei Li,[3] Zhi-Qiang Li,[4] and Xiu-Ming Xi[1]

[1] *Department of Critical Care Medicine, Fu Xing Hospital, Capital Medical University, Beijing 100038, China*
[2] *Department of Surgical Intensive Care Units, Hepatobiliary Surgery and Liver Transplant Center, Beijing YouAn Hospital, Capital Medical University, Beijing 100069, China*
[3] *Center for Infectious Diseases, Beijing YouAn Hospital, Capital Medical University, Beijing 100069, China*
[4] *Department of Critical Care Medicine, Hospital affiliated to Hebei United University, Tangshan 06300, China*

Correspondence should be addressed to Xiu-Ming Xi; xxm2947@sina.com

Received 12 May 2014; Accepted 14 June 2014; Published 7 July 2014

Academic Editor: Mingyue Zheng

Objective. To discover risk factors for mortality of patients with septic AKI in ICU via a multicenter study. *Background*. Septic AKI is a serious threat to patients in ICU, but there are a few clinical studies focusing on this. *Methods*. This was a prospective, observational, and multicenter study conducted in 30 ICUs of 28 major hospitals in Beijing. 3,107 patients were admitted consecutively, among which 361 patients were with septic AKI. Patient clinical data were recorded daily for 10 days after admission. Kidney Disease: Improving Global Outcomes (KDIGO) criteria were used to define and stage AKI. Of the involved patients, 201 survived and 160 died. *Results*. The rate of septic AKI was 11.6%. Twenty-one risk factors were found, and six independent risk factors were identified: age, APACHE II score, duration of mechanical ventilation, duration of MAP <65 mmHg, time until RRT started, and progressive KIDGO stage. Admission KDIGO stages were not associated with mortality, while worst KDIGO stages were. Only progressive KIDGO stage was an independent risk factor. *Conclusions*. Six independent risk factors for mortality for septic AKI were identified. Progressive KIDGO stage is better than admission or the worst KIDGO for prediction of mortality. This trial is registered with ChiCTR-ONC-11001875.

1. Introduction

Globally, the incidence of acute kidney injury (AKI) has increased steadily in recent years [1–4]. AKI is commonly seen in critically ill patients in ICU [5, 6] and contributes to the failure of other organs and systems in such patients [7]. The duration of AKI can be used to predict disease severity and outcome [8] although even transient AKI is linked to increased mortality [9]. The risk of death in AKI patients shows an incremental increase corresponding to disease stage [10]. Known risk factors of AKI include sepsis, critical illness, circulatory shock, burns, trauma, cardiac surgery, chronic diseases (heart, lung, and liver), major noncardiac surgery, and nephrotoxic drugs [11].

The cause of AKI in critically ill patients is usually multifactorial; however, sepsis is one of the leading causes of AKI, contributing to more than half of all reported cases [12–14]. The mechanism of sepsis-induced AKI is a complex combination of factors such as vascular and glomerular thrombotic processes, inflammation, and shock and is distinct from nonseptic AKI [15–18]. Thus, the clinical presentation, outcome, and responses to therapy may differ between septic and nonseptic AKI. Septic AKI is coupled with a significantly increased risk for hospital death, even after adjustment for relevant covariates [19]. However, only a limited number of clinical studies focusing on septic AKI in ICUs have been reported [19–23]. Thorough investigation is urgently required to reveal the epidemiology, pathophysiology, clinical features,

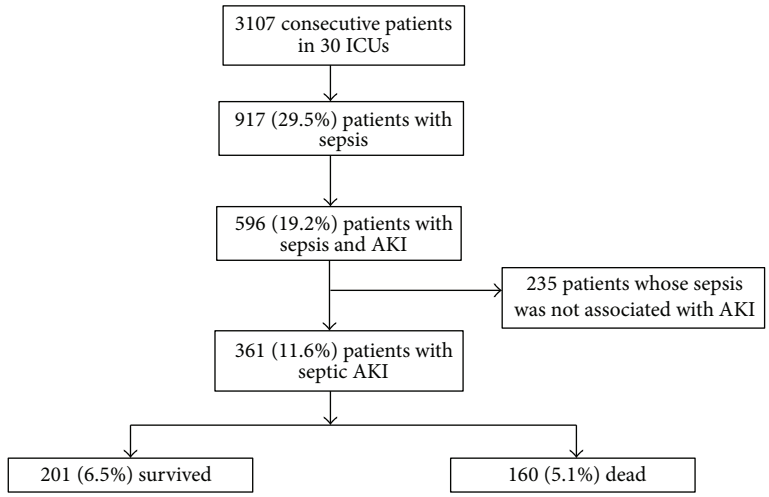


Figure 1: Study protocol flowchart.

and, more importantly, effective therapeutic measures for this disorder to reduce its high mortality.

This study aimed to identify risk factors for mortality in ICU patients with septic AKI and to evaluate the use of the KDIGO staging system for the prediction of prognosis in this group of patients, via a multicenter clinical study.

2. Material and Methods

This observational multicenter study was a retrospective analysis of prospectively collected data from patients in 30 ICUs of 28 major hospitals in Beijing between March 1 and August 31, 2012, as a part of the BAKIT (Beijing AKI Trail) study. Study subjects included all adult patients (age ≥ 18 years) admitted consecutively to the ICU and who received care in the ICU for at least 24 hours. Only the initial ICU admission was considered in this study. The following patients were excluded from the study: patients with preexisting end-stage chronic kidney disease, those already on RRT before admission to ICU, and those who had received kidney transplantation in the previous 3 months.

This study was approved by the Institutional Review Board of the Ethics Committee of the lead study center (Fu Xing Hospital, Capital Medical University, China), which waived the requirement for informed consent for this observational survey. Patient records/information were anonymized and deidentified prior to analysis.

2.1. Case Identification. Nine hundred and seventeen patients diagnosed with sepsis were identified [24]. AKI severity was classified according to the KDIGO guidelines (Kidney Disease: Improving Global Outcomes) [11], as follows: AKI is defined by an increase in serum creatinine (SCr) by ≥26.5 μmol/L within 48 hours or an increase of SCr of ≥1.5 times over baseline (which is known or presumed within the prior 7 days) or urine volume <0.5 mL/kg/h for 6 hours. AKI is staged for severity (3 stages) based on the changes in SCr and urine volume. Patients were staged according to SCr or urine output or both, with the criteria leading to the highest stage being used. Baseline SCr was the last value within the preceding year. For patients without these values or without renal failure, baseline SCr was estimated by the Modification of Diet in Renal Disease (MDRD) equation [25], assuming a glomerular filtration rate of 75 mL/min/1.73 m^2 [6]. For patients with chronic renal failure but not on dialysis, the initial SCr value on admission was used as the baseline value [6].

2.2. Data Collection. A uniform case report form (CRF) was used to collect data. Standard demographic, clinical, and laboratory data collected in the ICU included age, sex, dates and source of admission, BMI, blood pressure, duration of ICU stay, comorbidities, nonrenal organ failures, daily fluid input and output, and serum creatinine. The use of interventions, such as RRT, mechanical ventilation, loop diuretic therapy, and vasoactive agents, was also recorded.

Severity of illness was assessed by the Acute Physiology and Chronic Health Evaluation (APACHE) II score and Sequential Organ Failure Assessment (SOFA) scores, which were calculated based on the worst variables recorded during the first 24 hours after ICU admission to evaluate patient status [26, 27]. AKI severity was evaluated by KDIGO staging. Preexisting comorbidities were diagnosed based on International Classification of Diseases (ICD-10). For all patients included in the study, a thorough follow-up was conducted for the first 10 days after ICU admission. Patient status, laboratory data, interventions, and KDIGO stages were recorded daily. End points of this study included death or being transferred out of the ICU.

2.3. Definitions. Septic AKI was defined as sepsis-associated AKI [20, 28, 29], which meant that sepsis was associated with development and progression of AKI, so the patients (n = 235) with sepsis whose sepsis was not associated with AKI were excluded (Figure 1). We defined sepsis according to the American College of Chest Physicians and the Society of Critical Care Medicine(ACCP/SCCM) consensus [24, 30]. Based on this consensus, SIRS is defined as temperature >38°C

or <36°C, heart rate >90/min, respiratory rate >20/min or PaCO2 <32 mmHg, and white blood cell count >12,000/mm^3 or <4,000/mm^3 or with >10% bands. Sepsis was defined as a condition in which the patient met the criteria for SIRS and presented with either a documented or suspected infection. Admission KDIGO refers to the KDIGO stage on the first day of admission, while worst KDIGO refers to the worst KDIGO stage reached by a patient during their ICU stay. ICU-acquired AKI was defined as the development of AKI at 24 hours or more after admission, with the absence of AKI prior to admission. Progressive AKI was defined as patients reaching a higher KDIGO stage compared with the admission KDIGO stage at any time during their ICU stay. Vasoactive agents used in this study included epinephrine, norepinephrine, dopamine, and dobutamine. Large-dose vasopressor was defined as norepinephrine or epinephrine administered at a dose of >0.1 μg/kg/min, or dopamine or dobutamine administered at a dose of >15 μg/kg/min, or any two or more drugs in combination. Hospital acquired infection was defined as the development of an infection within 48 hours after hospital admission, which was not presented or incubating at the time of admission to the hospital.

2.4. Statistical Analysis. SPSS software (version 15.0) was used for data analysis. All variables were tested for normal distribution using the Kolmogorov-Smirnov test. Normally or near normally distributed variables are presented as means and SD, nonnormally distributed continuous data are presented as medians and interquartile ranges (IQR). Student's t-test was used for analysis of continuous variables. Mann-Whitney U test was used for nonnormally distributed variables. Categorical variables were compared with the chi-square test or Fisher's exact test. A 2-tailed $P < 0.05$ was considered statistically significant. Logistic regression was used to analyze risk factors for mortality. All variables with a P value <0.001 were included in the multivariate model. Backward selection based on the likelihood ratio test was used to select the final multivariate model for risk factors of mortality.

3. Results

3.1. Patient Characteristics. During the 6-month study period, a total of 3,107 patients were admitted to the 30 ICUs involved in this study, of which 29.5% (917/3,107) were diagnosed with sepsis. Of these patients, 39.4% (361/917) of patients were diagnosed with septic AKI; among which 55% (201/361) of patients survived and 44.4% (160/361) died. The rate of septic AKI among all subjects was 11.6% (361/3,107) (Figure 1). The average age was 70.54 ± 16.04 years, and 64.0% were male. Average BMI was 23.16 ± 3.82 and 37.7% were identified as hospital acquired infections. The average first 24 h APPACHE II score in the ICU was 23.59 ± 7.87, and the first 24 h SOFA score was 10.49 ± 5.40. The age, sex, BMI, hospital acquired infection, ways of admission, duration in ICU, nonrenal organ failure, comorbid diseases, and first 24 h APACHE II and SOFA scores in the ICU were compared between survivors and nonsurvivors. The age ($P < 0.001$), hospital acquired infection ($P = 0.001$), surgical admission ($P = 0.004$) and emergency admission ($P = 0.003$), systolic heart failure ($P = 0.007$), malignancy ($P = 0.031$), heart function level IV ($P = 0.021$), first 24 h APACHE II score ($P < 0.001$), and SOFA score ($P < 0.001$) in the ICU were associated with mortality (Table 1).

3.2. Disease Progression in the ICU. Disease progression was observed consecutively in the first 10 days after admission to the ICU in this study, and key interventions and parameters were recorded and analyzed. In total, 78.9% of septic AKI patients were on mechanical ventilation, 35.5% of patients needed RRT. Data on mechanical ventilation, fluid balance, hemodynamic data, and duration of vasoactive agent administration, loop diuretic therapy, and RRT were compared between survivors and nonsurvivors. Mechanical ventilation ($P < 0.001$) and its duration ($P < 0.001$), daily fluid balance ($P = 0.001$), duration of MAP <65 mmHg ($P < 0.001$), days on vasopressors ($P < 0.001$) and high-dose vasopressors ($P < 0.001$), RRT ($P = 0.007$), and time interval between ICU admission and RRT initiation ($P < 0.001$) were associated with patient outcomes (Table 2).

3.3. KDIGO Stages and Patient Outcome. During the first 10 days of ICU care, renal function of the patients was evaluated once a day according to KDIGO stage in this study. A flowchart of the progression of AKI in the ICU measured by KDIGO stages is shown in Figure 2. On admission, 27.7% (100/361) of all patients were at KDIGO stage 0, 29.9% (108/361) were at stage 1, 17.2% (62/361) were at stage 2, and 25.2% (91/361) were at stage 3. For the worst KDIGO stages, none of the patients were at stage 0, 20.8% (75/361) of patients were at stage 1, 25.5% (92/361) were at stage 2, and 53.7% (194/361) were at stage 3. Admission KDIGO stages were not linked to patient outcome, while the worst KDIGO stages were. According to our data, patients categorized into KDIGO stages 1, 2, and 3 by the worst KDIGO stages were strongly associated with patient outcome ($P < 0.001$, $P = 0.038$, and $P < 0.001$, resp.). ICU-acquired AKI was not linked to disease outcome ($P = 0.110$). Progressive AKI was associated with mortality ($P < 0.001$) (Table 3). Mortality rates for patients at different admission and the worst KDIGO stages are shown in Figure 3.

3.4. Risk Factors for Mortality. To identify possible risk factors for mortality in ICU patients with septic AKI, univariate analysis was performed for all the tested factors with a P value <0.05. Multivariate regression analysis was performed for all parameters with a P value <0.001 in the univariate analysis. Six independent risk factors were identified: age (OR = 1.025, 95% CI (1.007–1.042), $P = 0.005$), APACHE II score (first 24 h in ICU) (OR = 1.072, 95% CI (1.037–1.109), $P < 0.001$), duration of mechanical ventilation (OR = 1.080, 95% CI (1.008–1.158), $P = 0.03$), duration of MAP <65 mmHg (OR = 1.149, 95% CI (1.032–1.279), $P = 0.011$), time interval between ICU admission and RRT initiation (OR = 1.238, 95% CI (1.115–1.374), $P < 0.001$), and progressive KIDGO stage (OR = 3.374, 95% CI (1.918–5.933), $P < 0.001$) (Table 4).

Table 1: Patient characteristics.

	Survivors (n = 201)	Nonsurvivors (n = 160)	P value
Age (years; median [IQR])	72 (56–81)	78 (67–83)	<0.001
Gender (male) n (%)	131/201 (65.17%)	100/160 (62.50%)	0.264
BMI (mean ± SD)	23.50 ± 3.56	22.73 ± 4.09	0.058
Hospital acquired infection n (%)	61/201 (30.35%)	75/160 (46.88%)	0.001
Admission, n (%)			
Surgical admission	38/201 (18.91%)	14/160 (8.75%)	0.004
Emergency	83/201 (41.29%)	56/160 (35%)	0.003
Duration in ICU			
Days in ICU (days; median [IQR])	9 (5–16)	9 (5–19)	0.737
Nonrenal organ failure, n (%)			
Respiratory failure	124/201 (61.69%)	112/160 (70%)	0.099
Systolic heart failure	11/201 (5.47%)	22/160 (13.75%)	0.007
Hypovolemia shock	17/201 (8.46%)	19/160 (11.88%)	0.282
Septic shock	80/201 (39.8%)	68/160 (42.50%)	0.604
DIC	13/201 (6.47%)	8/160 (5.00%)	0.554
Hepatic failure	12/201 (5.97%)	8/160 (5.00%)	0.500
MODS (nonrenal)	62/201 (30.85%)	65/160 (40.63%)	0.051
Comorbid disease, n (%)			
Malignancy	34/201 (16.92%)	43/160 (26.88%)	0.031
Hypertension/CHD	108/201 (53.73%)	94/160 (58.75%)	0.340
Diabetes mellitus	48/201 (23.88%)	33/160 (20.63%)	0.461
CKD without renal failure	11/201 (5.47%)	4/160 (2.50%)	0.160
CKD with renal failure	19/201 (9.45%)	21/160 (13.13%)	0.269
Immunosuppression	9/201 (4.48%)	7/160 (4.38%)	0.586
Organ transplant	5/201 (2.49%)	2/160 (1.25%)	0.47
Heart function level IV	14/201 (6.97%)	23/160 (14.36%)	0.021
APACHEII score, first 24 h in ICU (mean ± SD)	21.41 ± 7.74	26.32 ± 7.15	<0.001
SOFA score, first 24 h in ICU (mean ± SD)	9.42 ± 5.33	11.83 ± 5.21	<0.001

BMI: body mass index, RRT: renal replacement therapy, CKD: chronic kidney disease, ICU: intensive care unit, DIC: disseminated intravascular coagulation, MODS: multiple organ dysfunction syndrome, CHD: chronic heart disease, SOFA: sequential organ failure assessment, APACHE: acute physiology and chronic health evaluation.

Table 2: Data on disease progression in the first 10 days after admission to ICU.

	Survivors (n = 201)	Nonsurvivors (n = 160)	P value
Mechanical ventilation			
Patients on mechanical ventilation n (%)	145/201 (72.14%)	140/160 (87.5%)	<0.001
Duration on mechanical ventilation (days; median [IQR])	3 (0–7)	6 (3–10)	<0.001
Fluid management			
Duration for positive fluid balance (days; median [IQR])	5 (3–7)	5 (3–7)	0.583
Daily fluid balance (mL/24 h)	654 ± 794	982 ± 1024	0.001
Hemodynamic data			
Duration for MAP < 65 mmHg (days; mean ± SD)	1.50 ± 1.98	2.42 ± 2.60	<0.001
Vasoactive agents			
Days on vasopressors (median [IQR])	1 (0–4)	3 (2–6)	<0.001
Days on large-dose vasopressor (median [IQR])	0 (0–3)	3 (0–5)	<0.001
Loop diuretic therapy (days) (median [IQR])	2 (0–6)	2 (1–5)	0.693
RRT			
Need for RRT n (%)	59/201 (29.35%)	69/160 (43.13%)	0.007
Duration of RRT (days; median [IQR])	0 (0–3)	0 (0–2)	0.065
Time interval between admission and RRT initiation (days; median [IQR])	0 (0-1)	0 (0–4)	<0.001

RRT: renal replacement therapy; MAP: mean arterial pressure.

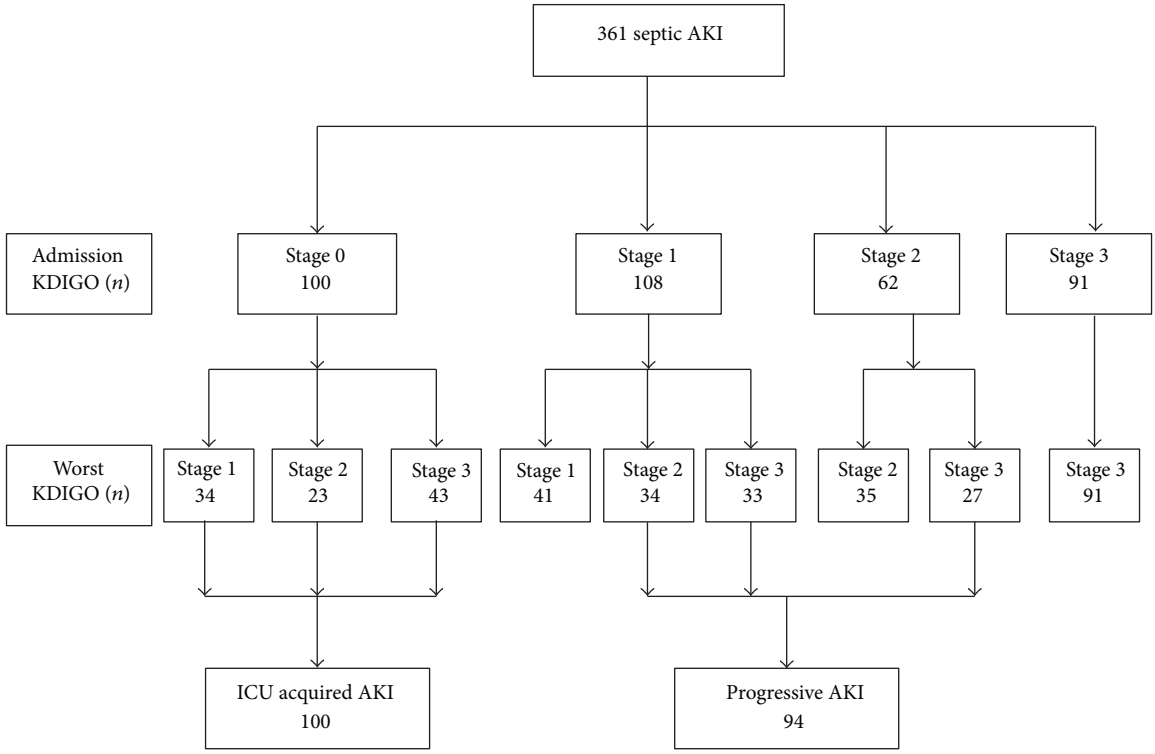


Figure 2: Progression of AKI in ICUs measured by KDIGO stages.

Table 3: AKI classified by KDIGO criteria.

	Survivors (n = 201)	Nonsurvivors (n = 160)	P value
KDIGO stage on admission n (%)			
Stage 1	57/201 (28.36%)	51/160 (31.88%)	0.271
Stage 2	38/201 (18.91%)	24/160 (15%)	0.202
Stage 3	56/201 (27.86%)	35/160 (21.88%)	0.119
Worst KDIGO stage in ICU n (%)			
Stage 1	55/201 (27.36%)	20/160 (12.5%)	<0.001
Stage 2	59/201 (29.35%)	33/160 (20.63%)	0.038
Stage 3	87/201 (43.28%)	107/160 (66.88%)	<0.001
Progress KDIGO stage class n (%)			
ICU acquired AKI	50/201 (24.88%)	50/160 (31.25%)	0.110
Progressive AKI	32/201 (15.92%)	62/160 (38.75%)	<0.001

KDIGO: Kidney Disease: Improving Global Outcomes.

4. Discussion

In this study, we investigated possible risk factors for mortality in critically ill patients with septic AKI via a large, multicenter, and observational study involving 30 ICUs. A total of 21 risk factors and six independent risk factors were identified in a thorough statistical analysis of comparisons between survivors and nonsurvivors among critically ill patients with septic AKI.

Our data showed low mortality among septic AKI patients admitted from the surgical or emergency departments (Table 1). Many surgical patients in the ICU were admitted for routine postoperative care after major operations and were associated with a very low mortality rate. Many patients admitted from emergency departments were in acute conditions, and, after timely interventions in ICU, most of them recovered well.

Mechanical ventilation is a common and important intervention in the ICU. In our study, the use of mechanical ventilation was correlated with increased mortality ($P < 0.001$). This is possibly due to the common complications of mechanical ventilation, such as worsening inflammatory responses, altered systemic hemodynamics, and elevated intrathoracic and intra-abdominal pressure, all of which are

Table 4: Regression analysis of risk factors for mortality in ICU.

	Univariate analysis		Multivariate analysis	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Age (years)	1.026 (1.012–1.041)	<0.001	1.025 (1.007–1.042)	0.005
Hospital acquired infection	2.025 (1.314–3.120)	0.001		
Nonrenal organ failure				
Systolic heart failure	2.754 (1.293–5.866)	0.009		
Comorbid disease				
Malignancy	1.748 (1.050–2.911)	0.032		
Heart function IV	2.242 (1.114–4.516)	0.024		
APACHE II score	1.092 (1.059–1.129)	<0.001	1.072 (1.037–1.109)	<0.001
SOFA score	1.090 (1.046–1.135)	<0.001	0.952 (0.889–1.020)	0.160
Mechanical ventilation				
Patients on mechanical ventilation	2.703 (1.543–4.737)	0.001		
Duration on mechanical ventilation	1.136 (1.071–1.206)	<0.001	1.080 (1.008–1.158)	0.03
Fluid management				
Daily fluid balance (mL/24 h)	1.000 (1.000–1.001)	0.001		
Hemodynamic data				
Duration for MAP < 65 mmHg	1.195 (1.083–1.319)	<0.001	1.149 (1.032–1.279)	0.011
Vasoactive agents				
Vasopressors	1.211 (1.126–1.302)	<0.001	1.082 (0.985–1.188)	0.102
RRT				
Need for RRT *n* (%)	1.825 (1.180–2.822)	0.007		
Time until RRT started (days)	1.261 (1.146–1.388)	<0.001	1.238 (1.115–1.374)	<0.001
Worst KDIGO				
Stage 1 *n* (%)	0.379 (0.216–0.665)	0.001		
Stage 2 *n* (%)	0.625 (0.384–1.019)	0.060		
Stage 3 *n* (%)	2.645 (1.718–4.073)	<0.001	1.466 (0.822–2.613)	0.195
Progress KDIGO class				
Progressive AKI	3.341 (2.039–5.475)	<0.001	3.374 (1.918–5.933)	<0.001

OR: odd ratio; APACHE II score, first 24 h in ICU; SOFA score, first 24 h in ICU.

involved in the development of AKI [31, 32]. In further analysis, we found that the duration of mechanical ventilation was an independent risk factor for mortality in patients with septic AKI, which may be due to the occurrence of ventilator-associated pneumonia (VAP), one of the leading causes of death in mechanically ventilated patients [33].

Compared with septic AKI survivors, nonsurvivors had greater hemodynamic instability: suffering from longer duration of hypotension (MAP < 65 mmHg), receiving more fluid and vasopressor even large-dose vasopressor (Table 2); in addition, multivariate analysis indicated the duration of MAP <65 mmHg as an independent risk factors for mortality in septic AKI patients (Table 4).

Previous studies have shown that, compared with non-septic AKI patients, septic ones came with worse hemodynamic instability and required more vasoactive agent use [20, 23, 34]. Lopes and colleagues discovered that extensive use of vasopressors was found in patients with severe AKI and associated with poor prognosis [35]. This is consistent with our data in Table 2.

In critically ill patients, it has been reported that positive fluid balance impaired cardiac function, led to lung injury, and may contribute to the development of AKI, which, in turn, increase mortality [36]. In patients with sepsis, prior report has shown that cumulative positive fluid balance was associated with increased mortality (odds ratio = 1.2) after adjustment for disease severity [37]. In our study, we came to the same conclusion in septic AKI patients (Table 2).

So, it seems that septic AKI patients with a long duration of low MAP required more vasoactive drug use and positive fluid balance, which implied high risk of shock and poor outcome.

RRT is one of the main approaches to the management of AKI. Recently, a multicenter study shown that in the nonsurvival with septic AKI, proportion of receiving RRT was significantly higher than that in the survival [29]. Our findings are consistent with this study: compared with the septic AKI patients who survived, proportion of receiving RRT was significantly higher in who died (43.13% versus 29.35%, $P = 0.007$). Furthermore, we found that there was no significant difference in duration of RRT between survivors and nonsurvivors with septic AKI ($P = 0.065$).

In addition, it is interesting that the time interval between ICU admission and RRT initiation was significantly longer in the patient who died. Moreover, by multivariate analysis, this delay in initiation of RRT was an independent risk factor for

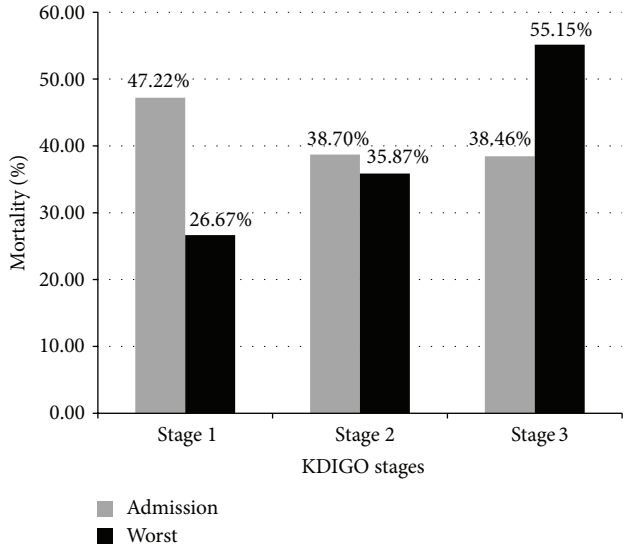


Figure 3: Mortality of septic AKI patients with different admission or the worst KDIGO stages. Gray bars represent mortality rates of patients grouped by admission KDIGO stages; black bars represent mortality rates of the worst KDIGO stages.

mortality (Tables 2 and 4). A large multicenter study about septic AKI came to the same conclusion that the time between ICU admission and start of RRT was significantly longer in the patients with septic AKI and this delay in initiation of RRT was independently associated with hospital mortality [20]. The right time to start RRT is still a topic of debate [38]. Experts recommend beginning RRT earlier, particularly in sepsis where AKI is known to be rapidly progressive [38]. A meta-analysis about timing of RRT clearly favored to begin RRT at early time [39]. In our study, the delay in initiation of RRT associated with mortality might be partly explained that progression of AKI in ICU was also an independent risk factor for mortality (Table 4). Patients with septic AKI who are with progression of AKI in ICU might receive RRT later after ICU admission than patients without progression of AKI. In brief, this observation showed that starting RRT timely is a key factor to reduce the high mortality of patients with septic AKI.

Many previous studies have evaluated AKI in critically ill patients by using the RIFLE classification [40] or AKIN criteria [41] and reported it to be associated with risk for mortality [6, 41–44]. KDIGO criteria is a new scaling system for AKI severity [11] and has been proven to be of prognostic significance [45, 46]. Some studies have indicated that KDIGO classification is better than RIFLE in terms of outcome prediction in certain circumstances [46]. Here we aimed to use KDIGO classification to evaluate critically ill patients with septic AKI. We found that the worst KIDGO stage in the ICU was linked to patient outcome, while no link was identified for the admission classification (Table 3). Furthermore, crude hospital mortality rates showed an incremental increase corresponding to the worst KDIGO stages, but not to the admission classification (Figure 3). This is consistent with previous publications indicating that mortality is not associated with admission RIFLE (risk, 44.7%; injury, 53.2%; failure, 51.0%; $P = 0.58$). However, worst RIFLE is associated within increased 28-day mortality ($P < 0.01$) [21].

Patients with poor admission KDIGO stages can be treated effectively by stage-based management such as hemodynamic monitoring, ensuring volume status and perfusion pressure, monitoring serum creatinine and urine output [11] and early goal-directed therapy (EGDT) [47]. However, later development of AKI or progression to a higher stage of AKI after ICU admission implies poor prognosis [21, 48].

It is interesting that in the worst KDIGO stages (Table 3), we found that only KDIGO stage 3 was associated with a high mortality, while survivors had a greater incidence of KDIGO stages 1 and 2. A multicenter study about septic AKI in Finnish came to the same conclusion; they found that after adjusting for covariates, the worst KDIGO stage 3 was associated with increased risk for 90-day mortality, but stages 1 and 2 were not [29]. It can be explained that although receiving active treatment in ICU, if the severity of septic AKI still progressed to KDIGO stage 3, the mortality would increase significantly. If the worst KDIGO stage of septic AKI only reached stages 1 or 2 in ICU, it would imply a good outcome.

Although the worst KDIGO stages were associated with mortality, they were not independent risk factors, while progressive KDIGO stage was found to be an independent risk factor associated with poor prognosis (Table 4). This is consistent with the results of other septic AKI studies, where progression of AKI has important prognostic implications [21, 49]. This result indicated the necessity of monitoring changes in KDIGO stages when AKI occurred in patients with sepsis in the ICU. On the other hand, ICU-acquired AKI was not a risk factor for mortality in our study ($P = 0.110$) (Table 3) but was an independent risk factor for 28-day mortality in a RIFLE-based study [21]. There were two differences between this study and ours. Firstly, its subject was patients with severe sepsis and septic shock, while the subject of our study was patients with septic AKI. Secondly, this study was a single center study targeting patients from medical ICU, thus limiting the applicability to more heterogeneous populations. In contrast, our study was a multicenter study involving various types of ICUs.

Our study has several limitations. First, baseline creatinine concentration was not measured for all patients; therefore, in such cases this value was estimated using the MDRD equation. Second, use of antibiotics is critical for management of sepsis but was not observed and involved in this study, because this was a substudy of the BAKIT (Beijing AKI Trail) study.

Through a consecutively thorough follow-up for 10 days after ICU admission, we found that the independent risk factors for mortality, except age and APACHE II score, the other four factors were all dynamic observational ones, such as duration (MAP, mechanical ventilation), time interval between admission, and RRT initiation and progression (KDIGO stage), while static factors such as need for RRT and the worst KDIGO stages were not independent risk factors. It suggests that we need to consecutively monitor the conditions of septic AKI patients. Currently, most of

observational studies for septic AKI collected clinical data for only one day [19, 28]; therefore the value of these data for predicting the prognostic for septic AKI is limited.

5. Conclusion

In summary, via a multicenter observational study, we evaluated the use of KDIGO stages on predicting patient outcome, found twenty-one risk factors such as age, hospital acquired infection, systolic heart failure, and mechanical ventilation, and identified six independent risk factors for mortality in ICU patients with septic AKI, which may help make early and accurate diagnosis and adopting preventive and therapeutic interventions that could reduce mortality rates in patients with septic AKI.

Conflict of Interests

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

Acknowledgments

This work was supported by a Grant from the Beijing Municipal Science & Technology Commission (BSTC), a government fund used to improve healthcare quality (no. D101100050010058). It offered financial support for data collection. The authors thank all members of the Beijing Acute Kidney Injury Trial (BAKIT) work group in participating for database management. The Beijing Acute Kidney Injury Trial (BAKIT) workgroup: Bin Du, Medical Intensive Care Unit, Peking Union Medical College Hospital, Beijing, China; Yuan Xu, Department of Critical Care Medicine, Beijing Tongren Hospital, Capital Medical University, Beijing, China; Jianxin Zhou, Department of Critical Care Medicine, Beijing Tiantan Hospital affiliated to Capital Medical University, Beijing, China; Ang Li, Department of Critical Care Medicine, Beijing Friendship Hospital, Capital Medical University, Beijing, China; Jingyuan Liu, Department of Critical Care Medicine, Beijing Ditan Hospital, Capital Medical University, Beijing, China; Wenxiong Li, Surgical Intensive Care Unit, Beijing Chaoyang Hospital, Capital Medical University, Beijing, China; Wenjin Chen, Neurological intensive care unit, Xuanwu Hospital, Capital Medical University, Beijing, China; JianguoJia, Surgical Intensive Care Unit, Xuanwu Hospital, Capital Medical University, Beijing, China; Penglin Ma, Department of Critical Care Medicine, The 309th Hospital of Chinese People's Liberation Army, Beijing, China; Xi Zhu, Department of Critical Care Medicine, Peking University Third Hospital, Beijing, China; Wei Chen, Department of Critical Care Medicine, Beijing Shijitan Hospital, Capital Medical University, Beijing, China; Dongxin Wang, Department of Critical Care Medicine, Peking University First Hospital, Beijing, China; Youzhong An, Department of Critical Care Medicine, Peking University People's Hospital, Beijing, China; Qingyuan Zhan, Department of Critical Care Medicine, China-Japan Friendship Hospital, Beijing, China; Gang Li, Department of Critical Care Medicine, China-Japan Friendship Hospital, Beijing, China; Haitao Zhang, Surgical Intensive Care Unit, Fuwai Hospital, China Academy of Medical Science and Peking Union Medical College, Beijing, China; Bo Ning, Department of Critical Care Medicine, Air Force General Hospital of Chinese People's Liberation Army, Beijing, China; Zhongjie He, Department of Critical Care Medicine, The First Affiliated Hospital of General Hospital of People's Liberation Army, Beijing, China; Zhicheng Zhang, Department of Critical Care Medicine, Navy General Hospital, Beijing, China; Yaxiong Sun, Department of Critical Care Medicine, The Luhe Teaching Hospital of the Capital Medical University, Beijing, China; ShijieJia, Surgical Intensive Care Unit, Beijing Anzhen Hospital, Capital Medical University, Beijing, China; Yalin Liu, Surgical Intensive Care Unit, Beijing Hospital, Capital Medical University, Beijing, China; Rui Cheng, Department of Critical Care Medicine, General Hospital of Armed Police Forces, Beijing, China; Qing Song, Department of Critical Care Medicine, The General Hospital of People's Liberation Army, Beijing, China; Jinning Liu, Surgical Intensive Care Unitin Department of Hepatobiliary Surgery and Liver Transplant Center, Beijing YouAn Hospital, Capital Medical University, Beijing, China; Yangong Chao, Department of Critical Care Medicine, Hua Xin Hospital, First Hospital of Tsinghua University, Beijing, China; Huizhen Li, Department of Critical Care Medicine, Beijing Shunyi Hospital of China Medical University, Beijing, China; Li Feng, Department of Critical Care Medicine, Beijing Geriatric Hospital, Beijing, China; Ruochun Shi, Department of Critical Care Medicine, Beijing No. 6 Hospital, Beijing, China; Department of Critical Care Medicine, Fuxing Hospital, Capital Medical University, Beijing, China: Ying Wen, Meiping Wang, Bo Zhu, Qi Jiang, Yujie Deng, Yan Sun, Peng Wang, Yanyan Yin, Xin Zhang, Li Zhang, Zhen Zhao, Ying Wang, RanLou, and Jing Wang.

References

[1] T. Z. Ali, I. Khan, W. Simpson et al., "Incidence and outcomes in acute kidney injury: a comprehensive population-based study," *Journal of the American Society of Nephrology*, vol. 18, no. 4, pp. 1292–1298, 2007.

[2] S. S. Waikar, G. C. Curhan, R. Wald, E. P. McCarthy, and G. M. Chertow, "Declining mortality in patients with acute renal failure, 1988 to 2002," *Journal of the American Society of Nephrology*, vol. 17, no. 4, pp. 1143–1150, 2006.

[3] J. L. Xue, F. Daniels, R. A. Star et al., "Incidence and mortality of acute renal failure in Medicare beneficiaries, 1992 to 2001," *Journal of the American Society of Nephrology*, vol. 17, no. 4, pp. 1135–1142, 2006.

[4] N. H. Lameire, A. Bagga, D. Cruz et al., "Acute kidney injury: an increasing global concern," *The Lancet*, vol. 382, no. 9887, pp. 170–179, 2013.

[5] S. Uchino, J. A. Kellum, R. Bellomo et al., "Acute renal failure in critically ill patients: a multinational, multicenter study," *The Journal of the American Medical Association*, vol. 294, no. 7, pp. 813–818, 2005.

[6] E. A. J. Hoste, G. Clermont, A. Kersten et al., "RIFLE criteria for acute kidney injury are associated with hospital mortality in critically ill patients: a cohort analysis," *Critical Care*, vol. 10, no. 3, article R73, 2006.

[7] M. E. Grams and H. Rabb, "The distant organ effects of acute kidney injury," *Kidney International*, vol. 81, no. 10, pp. 942–948, 2012.

[8] J. R. Brown, R. S. Kramer, S. G. Coca, and C. R. Parikh, "Duration of acute kidney injury impacts long-term survival after cardiac surgery," *The Annals of Thoracic Surgery*, vol. 90, no. 4, pp. 1142–1148, 2010.

[9] M. Nejat, J. W. Pickering, P. Devarajan et al., "Some biomarkers of acute kidney injury are increased in pre-renal acute injury," *Kidney International*, vol. 81, no. 12, pp. 1254–1262, 2012.

[10] M. Joannidis, B. Metnitz, P. Bauer et al., "Acute kidney injury in critically ill patients classified by AKIN versus RIFLE using the SAPS 3 database," *Intensive Care Medicine*, vol. 35, no. 10, pp. 1692–1702, 2009.

[11] (KDIGO) KDIGO, "Clinical practice guideline for acute kidney injury," *Kidney International Supplements*, vol. 2, pp. 124–138, 2012.

[12] D. C. Angus, W. T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, and M. R. Pinsky, "Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care," *Critical Care Medicine*, vol. 29, no. 7, pp. 1303–1310, 2001.

[13] S. M. Bagshaw, K. B. Laupland, C. J. Doig et al., "Prognosis for long-term survival and renal recovery in critically ill patients with severe acute renal failure: a population-based study," *Critical Care*, vol. 9, no. 6, pp. R700–R709, 2005.

[14] W. Silvester, R. Bellomo, and L. Cole, "Epidemiology, management, and outcome of severe acute renal failure of critical illness in Australia," *Critical Care Medicine*, vol. 29, no. 10, pp. 1910–1915, 2001.

[15] R. Jacobs, P. M. Honore, O. Joannes-Boyau et al., "Septic acute kidney injury: the culprit is inflammatory apoptosis rather than ischemic necrosis," *Blood Purification*, vol. 32, no. 4, pp. 262–265, 2011.

[16] C. Langenberg, L. Wan, M. Egi, C. N. May, and R. Bellomo, "Renal blood flow in experimental septic acute renal failure," *Kidney International*, vol. 69, no. 11, pp. 1996–2002, 2006.

[17] C. Langenberg, L. Wan, S. M. Bagshaw, M. Egi, C. N. May, and R. Bellomo, "Urinary biochemistry in experimental septic acute renal failure," *Nephrology Dialysis Transplantation*, vol. 21, no. 12, pp. 3389–3397, 2006.

[18] M. Brenner, G. L. Schaer, D. L. Mallory, A. F. Suffredini, and J. E. Parillo, "Detection of renal blood flow abnormalities in septic and critically ill patients using a newly designed indwelling thermodilution renal vein catheter," *Chest*, vol. 98, no. 1, pp. 170–179, 1990.

[19] M. Oppert, C. Engel, F.-M. Brunkhorst et al., "Acute renal failure in patients with severe sepsis and septic shock—a significant independent risk factor for mortality: results from the German Prevalence Study," *Nephrology Dialysis Transplantation*, vol. 23, no. 3, pp. 904–909, 2008.

[20] S. M. Bagshaw, S. Uchino, R. Bellomo et al., "Septic acute kidney injury in critically ill patients: clinical characteristics and outcomes," *Clinical Journal of the American Society of Nephrology*, vol. 2, no. 3, pp. 431–439, 2007.

[21] W. Y. Kim, J. W. Huh, C.-M. Lim, Y. Koh, and S.-B. Hong, "Analysis of progression in risk, injury, failure, loss, and end-stage renal disease classification on outcome in patients with severe sepsis and septic shock," *Journal of Critical Care*, vol. 27, no. 1, pp. 104.e1–104.e7, 2012.

[22] M. Plataki, K. Kashani, J. Cabello-Garza et al., "Predictors of acute kidney injury in septic shock patients: an observational cohort study," *Clinical Journal of the American Society of Nephrology*, vol. 6, no. 7, pp. 1744–1751, 2011.

[23] E. A. J. Hoste, N. H. Lameire, R. C. Vanholder, D. D. Benoit, J. M. A. Decruyenaere, and F. A. Colardyn, "Acute renal failure in patients with sepsis in a surgical ICU: predictive factors, incidence, comorbidity, and outcome," *Journal of the American Society of Nephrology*, vol. 14, no. 4, pp. 1022–1030, 2003.

[24] R. C. Bone, R. A. Balk, F. B. Cerra et al., "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine," *Chest*, vol. 101, no. 6, pp. 1644–1655, 1992.

[25] A. S. Levey, J. Coresh, E. Balk et al., "National Kidney Foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification," *Annals of Internal Medicine*, vol. 139, no. 2, pp. 137–147, 2003.

[26] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "APACHE II: a severity of disease classification system," *Critical Care Medicine*, vol. 13, no. 10, pp. 818–829, 1985.

[27] J.-L. Vincent, R. Moreno, J. Takala et al., "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure," *Intensive Care Medicine*, vol. 22, no. 7, pp. 707–710, 1996.

[28] S. M. Bagshaw, C. George, and R. Bellomo, "Early acute kidney injury and sepsis: a multicentre evaluation," *Critical Care*, vol. 12, no. 2, article R47, 2008.

[29] M. Poukkanen, S. T. Vaara, V. Pettilä et al., "Acute kidney injury in patients with severe sepsis in finnish intensive care units," *Acta Anaesthesiologica Scandinavica*, vol. 57, no. 7, pp. 863–872, 2013.

[30] "American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference: definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis," *Critical Care Medicine*, vol. 20, no. 6, pp. 864–874, 1992.

[31] J. L. Koyner and P. T. Murray, "Mechanical ventilation and lung-kidney interactions," *Clinical Journal of the American Society of Nephrology*, vol. 3, no. 2, pp. 562–570, 2008.

[32] A. S. Awad, M. Rouse, L. Huang et al., "Compartmentalization of neutrophils in the kidney and lung following acute ischemic kidney injury," *Kidney International*, vol. 75, no. 7, pp. 689–698, 2009.

[33] A. H. Choudhuri, "Ventilator-Associated Pneumonia: when to hold the breath?" *International Journal of Critical Infrastructure Protection*, vol. 3, no. 3, pp. 169–174, 2013.

[34] S. M. Bagshaw, S. Lapinsky, S. Dial et al., "Acute kidney injury in septic shock: clinical outcomes and impact of duration of hypotension prior to initiation of antimicrobial therapy," *Intensive Care Medicine*, vol. 35, no. 5, pp. 871–881, 2009.

[35] J. A. Lopes, S. Jorge, C. Resina et al., "Acute kidney injury in patients with sepsis: a contemporary analysis," *International Journal of Infectious Diseases*, vol. 13, no. 2, pp. 176–181, 2009.

[36] M. L. Esson and R. W. Schrier, "Diagnosis and treatment of acute tubular necrosis," *Annals of Internal Medicine*, vol. 137, no. 9, pp. 744–752, 2002.

[37] D. Payen, A. C. de Pont, Y. Sakr, C. Spies, K. Reinhart, and J. L. Vincent, "A positive fluid balance is associated with a worse outcome in patients with acute renal failure," *Critical Care*, vol. 12, no. 3, article R74, 2008.

[38] P. M. Honore, R. Jacobs, O. Joannes-Boyau et al., "Septic AKI in ICU patients. Diagnosis, pathophysiology, and treatment type, dosing, and timing: a comprehensive review of recent and future developments," *Annals of Intensive Care*, vol. 1, p. 32, 2011.

[39] V. F. Seabra, E. M. Balk, O. Liangos, M. A. Sosa, M. Cendoroglo, and B. L. Jaber, "Timing of renal replacement therapy initiation in acute renal failure: a meta-analysis," *American Journal of Kidney Diseases*, vol. 52, no. 2, pp. 272–284, 2008.

[40] R. Bellomo, C. Ronco, J. A. Kellum, R. L. Mehta, and P. Palevsky, "Acute renal failure—definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group," *Critical Care*, vol. 8, no. 4, pp. R204–R212, 2004.

[41] R. L. Mehta, J. A. Kellum, S. V. Shah et al., "Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury," *Critical Care*, vol. 11, no. 2, article R31, 2007.

[42] N. Y. Abosaif, Y. A. Tolba, M. Heap, J. Russell, and A. M. El Nahas, "The outcome of acute renal failure in the intensive care unit according to RIFLE: model application, sensitivity, and predictability," *American Journal of Kidney Diseases*, vol. 46, no. 6, pp. 1038–1048, 2005.

[43] Z. Ricci, D. Cruz, and C. Ronco, "The RIFLE criteria and mortality in acute kidney injury: a systematic review," *Kidney International*, vol. 73, no. 5, pp. 538–546, 2008.

[44] A. Kuitunen, A. Vento, R. Suojaranta-Ylinen, and V. Pettilä, "Acute renal failure after cardiac surgery: evaluation of the RIFLE classification," *Annals of Thoracic Surgery*, vol. 81, no. 2, pp. 542–546, 2006.

[45] F. B. Rodrigues, R. G. Bruetto, U. S. Torres, A. P. Otaviano, D. M. T. Zanetta, and E. A. Burdmann, "Incidence and mortality of acute kidney injury after myocardial infarction: a comparison between KDIGO and RIFLE criteria," *PLoS ONE*, vol. 8, no. 7, Article ID e69998, 2013.

[46] A. K. Roy, C. Mc Gorrian, C. Treacy et al., "A comparison of traditional and novel definitions (RIFLE, AKIN, and KDIGO) of acute kidney injury for the prediction of outcomes in acute decompensated heart failure," *Cardiorenal Medicine*, vol. 3, no. 1, pp. 26–37, 2013.

[47] E. Rivers, B. Nguyen, S. Havstad et al., "Early goal-directed therapy in the treatment of severe sepsis and septic shock," *The New England Journal of Medicine*, vol. 345, no. 19, pp. 1368–1377, 2001.

[48] C. Guerin, R. Girard, J. M. Selli, J. P. Perdrix, and L. Ayzac, "Initial versus delayed acute renal failure in the intensive care unit: a multicenter prospective epidemiological study," *American Journal of Respiratory and Critical Care Medicine*, vol. 161, no. 3, pp. 872–879, 2000.

[49] J. A. Lopes, S. Jorge, C. Santos et al., "Acute kidney injury in patients with sepsis: a contemporary analysis," *International Journal of Infectious Diseases*, vol. 13, pp. 176–181, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 469103, 9 pages
http://dx.doi.org/10.1155/2014/469103

Research Article

Combined Analysis with Copy Number Variation Identifies Risk Loci in Lung Cancer

Xinlei Li,[1] Xianfeng Chen,[1] Guohong Hu,[1] Yang Liu,[1] Zhenguo Zhang,[1] Ping Wang,[1] You Zhou,[1] Xianfu Yi,[1] Jie Zhang,[1] Yufei Zhu,[1] Zejun Wei,[1] Fei Yuan,[1] Guoping Zhao,[2] Jun Zhu,[3] Landian Hu,[1,4] and Xiangyin Kong[1,4]

[1] *Institute of Health Sciences, Shanghai Jiao Tong University School of Medicine (SJTUSM) and Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences (CAS), Shanghai 200025, China*
[2] *Chinese National Human Genome Center at Shanghai, Shanghai 201203, China*
[3] *Institute of Bioinformatics, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310029, China*
[4] *State Key Laboratory of Medical Genomics, Ruijin Hospital, Shanghai Jiao Tong University, 197 Ruijin Road II, Shanghai 200025, China*

Correspondence should be addressed to Xiangyin Kong; xykong@sibs.ac.cn

Received 21 May 2014; Revised 11 June 2014; Accepted 11 June 2014; Published 1 July 2014

Academic Editor: Tao Huang

Background. Lung cancer is the most important cause of cancer mortality worldwide, but the underlying mechanisms of this disease are not fully understood. Copy number variations (CNVs) are promising genetic variations to study because of their potential effects on cancer. *Methodology/Principal Findings*. Here we conducted a pilot study in which we systematically analyzed the association of CNVs in two lung cancer datasets: the Environment And Genetics in Lung cancer Etiology (EAGLE) and the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial datasets. We used a preestablished association method to test the datasets separately and conducted a combined analysis to test the association accordance between the two datasets. Finally, we identified 167 risk SNP loci and 22 CNVs associated with lung cancer and linked them with recombination hotspots. Functional annotation and biological relevance analyses implied that some of our predicted risk loci were supported by other studies and might be potential candidate loci for lung cancer studies. *Conclusions/Significance*. Our results further emphasized the importance of copy number variations in cancer and might be a valuable complement to current genome-wide association studies on cancer.

1. Introduction

Lung cancer is the most common cause of cancer-related death in the world [1]. Many types of genetic variation such as single-nucleotide polymorphisms (SNPs) and copy number variations (CNVs) [2–7] have been discovered to be associated with lung cancer. Copy number variations are prevalent in the genome, covering approximately 12% of the human genome [8], which may make it more likely to contribute to disease incidence [9–15]. Thus, a systematic survey of CNVs in lung cancer is essential.

To test the relationship between copy number variations and lung cancer susceptibility, we conducted a pilot combined genome-wide analysis of two case-control datasets in lung cancer, consisting of 1,945 cases and 1,992 controls from the Environment And Genetics in Lung cancer Etiology (EAGLE) project [16] and 803 cases and 848 controls from the Prostate, Lung, Colon and Ovary (PLCO) Study Cancer Screening Trial project [17]. All cases and controls were matched well for age, gender, district, and other characteristics according to the original study design. Genomic DNA samples from somatic cells (blood cells) were probed using an Illumina HumanHap 550K genotyping chip. Before testing, we normalized all data (including a training dataset of 66 individuals from HapMap) using quantile normalization. To infer the copy number state of each SNP site, we trained a well-established hidden Markov model (HMM) using the training dataset. All of the above procedures were performed as previously described [18]. Our combined association analysis was done at two levels. First, in each

Table 1: Descriptions of raw CNVs in EAGLE and PLCO.

	Average CNVs per individual	SNPs per CNV			CNV size		
		Min	Max	Average	Min (bp)	Max (Mb)	Average (kb)
EAGLE							
Case	391.1	3	1325	8.8	23	22.9	36.7
Control	441	3	693	8	23	8.6	32.7
PLCO							
Case	193.2	3	822	8.6	37	22.9	35.7
Control	106.7	3	537	7.5	45	2.2	30.6

Note that the raw CNVs in this table were roughly generated by a simple arbitrary method and might not be reliable.

individual dataset we conducted an SNP-based testing to probe the association of lung cancer with a specific SNP site. Next, window-based testing was performed to probe the association pattern with lung cancer. The details of the SNP-based and window-based testing are in Section 4. Second, to test the association accordance between the two datasets, we conducted a combined analysis in which a new statistic of relative factor (Rf) was calculated for each SNP site. In the calculation of the Rf, the cases and controls from different datasets were hypothesized to have an independent genomic distribution of copy number states. We assumed that the difference between the two copy number state distributions could be tested from their accordance and the combination of related P values could be used to depict such difference between these two datasets (see details in Section 4).

2. Results

2.1. Raw CNVs Prediction for EAGLE and PLCO. We used the full set of study participants as described in Section 4. After quality control of the samples and array data, the probe signals on the chip were transformed to copy number state by our preestablished HMM approach [18]. We first adopted a simple CNV calculation method to roughly generate raw CNVs for both the EAGLE and PLCO datasets (see Section 4). We compared the raw CNVs between EAGLE and PLCO (Table 1) and found that although the average span and size of raw CNVs were comparable between EAGLE and PLCO, the number of raw CNVs predicted by this approach was larger in EAGLE than in PLCO. This might be caused by the smaller sample size of PLCO used in this study than EAGLE when studying potential rare risk loci. Meanwhile, as this simple raw prediction method was based only on individual level copy number state data, false positive noises in each individual might also increase the total number of predicted CNVs when sample size increased. Given these considerations, we assumed that the simple raw CNV calculation method we used might not be appropriate to predict reliable CNVs between different datasets. As a result, we developed another combined strategy following our preestablished CNV association approach to predict CNVs between EAGLE and PLCO. This combined strategy was based on the copy number state distributions in both EAGLE and PLCO individuals, which might help to filter out false positive noises in single individual level prediction. Moreover, since the whole copy number state distributions of both EAGLE and PLCO cohorts were considered and integrated, CNVs predicted in this way were not dependent on single cohorts, which we expected to avoid the impact of different sample sizes on CNVs prediction (see Section 4 for details).

2.2. Genome-Wide Combined Analysis for EAGLE and PLCO. In our previous study [18], we developed a two-step genome-wide CNV association approach based on SNP-based testing and window-based testing to find significant SNP sites with abnormal copy number variations. Here we used the same approach as the initial stage of our combined analysis. In the SNP-based testing, we obtained 509 candidate SNP sites in EAGLE and 573 candidate SNP sites in PLCO (the corresponding FDR $\leq$ 0.15) (Figure S1, available online at http://dx.doi.org/10.1155/2014/469103). We noticed that more SNP candidates were found in PLCO than in EAGLE; the reason might be that larger sample size could help reduce false positive noises in SNP-based testing. All SNP candidates were subjected to window-based testing after SNP-based testing. As expected, we found that statistical power in PLCO was lower than in EAGLE (Figure S2). The reason for this phenomenon might be that PLCO contained a smaller sample size than EAGLE, as a loss of statistical power in small sample sizes had been recognized in other studies [19, 20]. Therefore, in our subsequent analyses EAGLE was used as the discovery dataset and PLCO was used to verify the association accordance between them. In the EAGLE dataset, we identified 355 SNP sites with significant window-based P values (FDR = 0.0702482) (Table S1). In the PLCO dataset, 243 SNP sites passed window-based testing (FDR = 0.102871). Genome-wide association testing in a single dataset might be influenced by population structure and other factors, leading to false positives in many studies [21, 22]. Indeed we observed such population stratification in the EAGLE dataset, although there was no obvious stratification between case and control cohorts (Figure S3). Therefore, a strict combined analysis integrating PLCO with EAGLE was conducted to evaluate the association results in EAGLE. After the combined analysis, 167 SNP sites were obtained as risk loci in EAGLE (Table S2) and good association consistency between EAGLE and PLCO datasets was found in the hypothesis regarding amplification (Figures 1(a) and 1(b)). For the other two hypotheses regarding deletion and abnormal (deletion or

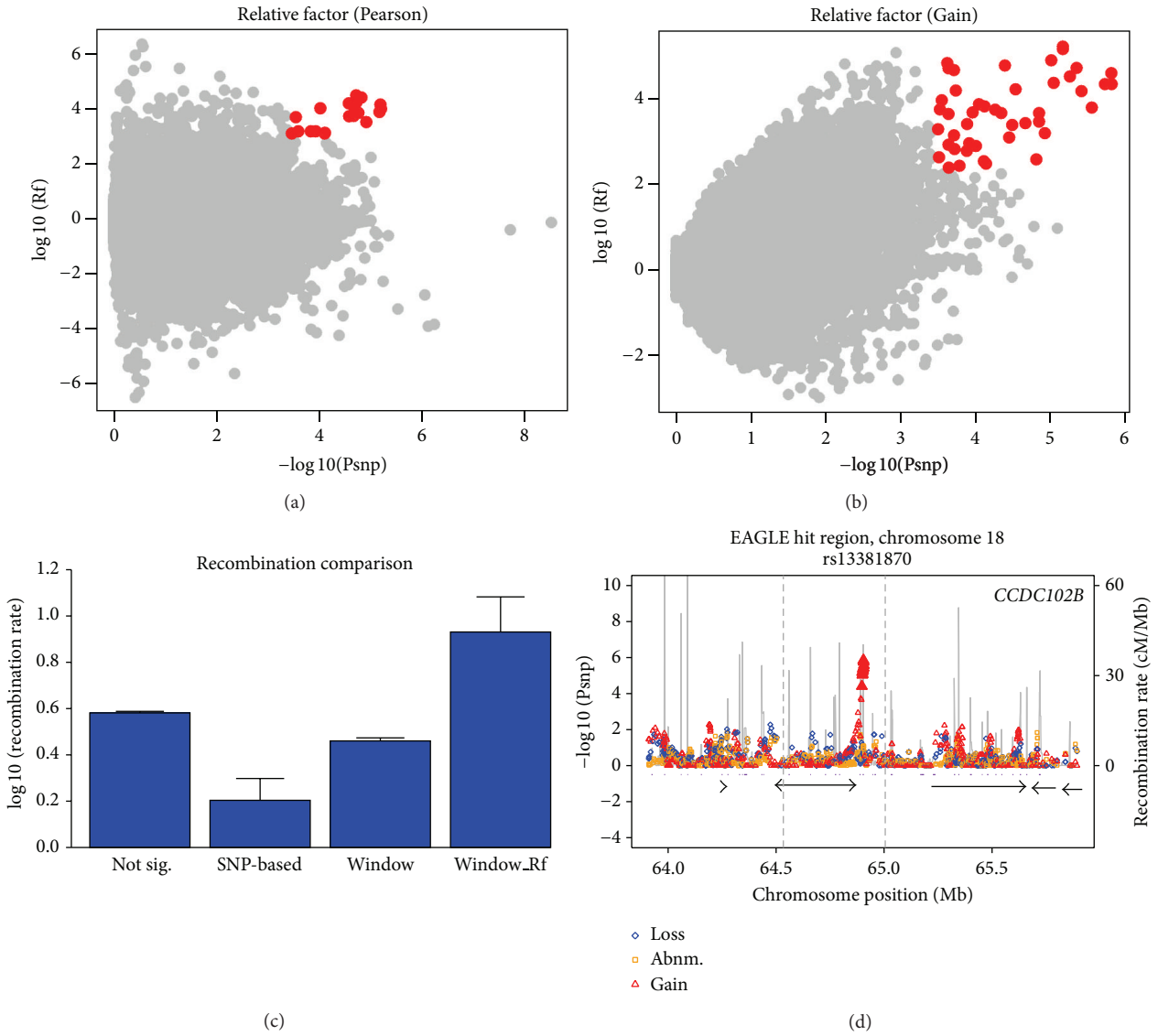


Figure 1: CNV associations regarding amplification are related to recombination hotspot regions in EAGLE and PLCO. The −log 10 (P value) in the (a) Pearson testing and (b) Gain testing are plotted against the corresponding log 10 (relative factors). The SNP sites above a significant level (combining P value and relative factor to ensure that the final false positive is less than 1) are in red. (c) The log 10 (maximum recombination rate) around the SNP sites (in 10 kb) are summed in four categories: not sig. (nonsignificant SNP sites), SNP-based (significant SNP sites that passed SNP-based testing), window (significant SNP sites that passed window-based testing), and window_Rf (risk loci that passed both window-based and combined testing). The log 10 (maximum recombination rate) is prevalent in the category of window_Rf (P value < 0.00001). (d) Most of these risk loci were located around recombination hotspots (plotted in gray lines and with peaks indicating the recombination rates). One of these associated sites, rs13381870, was arbitrarily chosen and is shown here as an example. The −log 10 (P value) of SNP sites in three hypotheses models (*loss*, *abnm.*, and *gain*) in SNP-based testing are plotted in blue, orange, and red, respectively. Grey vertical lines show the high recombination rate sites. Hotspots from HapMap were shown as purple bars between the plot and genes. The names of genes around rs13381870 are shown in the figure.

amplification), the association consistency was not as good as that of amplification (data not shown), which might be caused by the population differences between EAGLE and PLCO or the limitations of our approaches.

2.3. Functional Annotation Clustering Analysis of Genes around Risk Loci. To study the biological meaning of those significant risk sites in EAGLE, we did a functional annotation clustering analysis of genes surrounding those sites. Since there was no more evidence to show which gene will be affected by the candidate risk loci, we roughly glanced at genes located within ±100 kb around those risk loci and retrieved a list of 243 neighboring genes (Table S3) of all the risk loci in Table S2.

Next, we used DAVID (http://david.abcc.ncifcrf.gov/) for a functional clustering analysis of the neighboring gene list to see whether there were some biologically meaningful clusters. From the results we found that there were some

gene sets formed in annotation cluster 4 (see Table S4 for detailed and statistical information) as defined by SMART and INTERPRO classification. We found that in spite of the many gene sets in annotation cluster 4, many of them did not have a significant statistical *P* value. We assumed two reasons for this phenomenon. First, our definition of affected genes was arbitrary which might lead to the inclusion of unrelated background genes or exclusion of truly affected genes, and both situations would lower the clustering power of DAVID. Second, the underlying mechanisms of a complex disease like lung cancer cannot be easily modeled by such a simple study approach and as a result the less significant clustering sets might only reveal a small piece of the whole network.

In DAVID analysis, we compared our neighboring gene list with the published literature. We selected the PubMed ID for DAVID to see which subset of genes was related to previous studies. Two published studies were reported to be significantly related to neighboring gene list (Table S5). PMID 11085536 [23] was a study that included twelve of the genes in neighboring gene list. The authors of that study tested manually and found loss of expression or reduced mRNA levels for *SEMA3B* in both small cell and non-small cell lung cancers, as well as reduced mRNA levels of *CACNA2D2* in non-small cell lung cancer and two or more sequence-altering mutations for *SEMA3B* and *NPRL2*, indicating that those genes might be candidate tumor suppress genes (TSGs). The study in PMID 19140316 [24] found four genes also in our neighboring gene list. They used real-time PCR to analyze the downregulation of four genes, *HYAL1*, *HYAL2*, *RASSF1A*, and *NPRL2*, in lung cancer and found that they were downregulated in non-small cell lung cancer, the first stage of squamous cell lung cancer, and were significantly associated with lung adenocarcinoma progression. They expected the downregulation of those genes to be important for diagnosis and therapeutic strategies development of lung cancer. The fact that our neighboring gene list also contained those previously reported genes revealed that genes around the 167 risk loci were worthy of future functional studies.

2.4. CNVs around Risk Loci and Their Biological Relevance. We carefully investigated the 167 risk loci in EAGLE that passed SNP-based, window-based, and combined analysis with PLCO (Table S2) and found that those risk loci could be classified into two groups depending on their consecutiveness: singular risk loci, which were short of flanking risk loci, and consecutive risk loci, which consisted of many consecutive flanking risk loci, forming a CNV risk region. Based on this classification, we manually checked all 167 risk loci and generated CNVs from the consecutive risk loci blocks (Table 2; see Section 4).

A total of 22 CNVs were summarized from 167 risk loci in EAGLE, including three amplification CNVs, 18 deletions, and 1 abnormal (amplification/deletion) variation. As we mentioned previously, our combined analysis had a good association consistency regarding amplification (Figures 1(a) and 1(b)). Therefore we first focused on the three amplification CNVs, which were located on 8q23.3, 13q21.1, and 18q22.1. We searched PubMed for any published functional or genome-wide studies revealing an association between the region of the three CNVs and lung cancer (Table 2). For 8q23.3, our results indicated 59.4 kb amplification region. Boelens et al. had reported 8q23.3 as a common CNV-related region of lung cancer [25]. We obtained 30.4 kb amplification on 13q21.1, a candidate region containing alterations in esophageal squamous cell carcinoma [26] and also reported by Boelens et al. [25]. We did not get direct evidence from the literature to support the third, a 9.3 kb CNV located on 18q22.1.

We also searched the literature for evidence of the 18 deletions and 1 abnormal variation recovered in our analysis. Only the 12.9 kb deletion on 5q35.2 did not appear in previous studies. Direct or indirect support was found for all other cases (Table 2), although the association consistency between EAGLE and PLCO was better at predicting amplification in EAGLE.

In further support of our CNV findings, we expected that other types of mutations, such as single-nucleotide changes, could validate the physiological significance of our predicted CNVs. We examined the mutation status of neighboring genes around these risk loci (Table S2) in the Cancer Genome Project Data of Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/genetics/CGP/cosmic/). *CSMD3* in 8q23.3 is a large gene encoding a protein with CUB and sushi multiple domains and is associated with somatic mutations in lung cancer (7 mutated in 11 unique samples). Mutations in *CSMD3* were also found to be associated with familial colorectal cancer [27]. *CCDC102B* in 18q22.1 is associated with somatic mutations in lung cancer (1 mutated sample). Moreover, *CSMD3* and *CCDC102B* have been reported to be affected by genomic rearrangement events in autistic patients [28] and in patients with diaphragmatic defects [29], respectively. Given the fact that these two genes in our neighboring gene list were reported with mutations in lung cancer, we expected other genes to be investigated in future studies.

2.5. The Risk Loci Are Located on Genomic Recombination Hotspots. Interestingly, we found that our predicted risk loci were associated with high rates of recombination (Figures 1(c) and 1(d)) compared to other SNP sites. Then we plotted the summarized 22 CNVs against HapMap hotspots on the genome (Figure S4). 13 out of 18 deletion CNVs overlapped with hotspots; 2 out of 3 amplification CNVs overlapped with hotspots; the single abnormal CNV also overlapped with hotspots. Those results revealed that there might be some connections between genomic recombination hotspots and disease risk loci; further studies are necessary to support and confirm such relationships.

3. Discussion

In summary, we developed a combined analysis following our preestablished SNP-based and window-based CNV association methods to conduct a pilot study in two datasets, EAGLE and PLCO. The workflow of this pilot study can be found in Figure 2.

Table 2: Predicted CNVs in EAGLE.

Dataset	Chr.	Band	Start_pos.	End_pos.	Size (kb)	Type	Literature
EAGLE	1	1p36.22	12120766	12129342	8.6	Deletion	19513508; 16142337; 20676096
EAGLE	5	5q35.2	175888783	175901659	12.9	Deletion	—
EAGLE	8	8q23.3	113681735	113741162	59.4	Amplification	19324446
EAGLE	8	8q24.3	144694717	144728743	34.0	Deletion	18990762; 22142333
EAGLE	8	8q24.3	145079175	145118650	39.5	Deletion	18990762; 22142333
EAGLE	9	9q32	114406899	114414974	8.1	Deletion	18798555; 15580306; 7512370
EAGLE	9	9q34.3	138620438	138641922	21.5	Deletion	16740712
EAGLE	10	10q22.3	80766077	80778488	12.4	Deletion	18758299; 20651054
EAGLE	11	11q13.1	65012165	65051406	39.2	Deletion	11274644
EAGLE	13	13q21.1	56772821	56803216	30.4	Amplification	20200074; 19324446
EAGLE	16	16p13.3	1951065	1994156	43.1	Deletion	17086460
EAGLE	17	17q21.1	35509120	35510616	1.5	Deletion	16733218; 11378338
EAGLE	17	17q25.3	73635123	73655682	20.6	Deletion	17086460
EAGLE	17	17q25.3	77848326	78009203	160.9	Deletion	17086460
EAGLE	18	18p11.32	2580764	2629683	48.9	Deletion	19190329
EAGLE	18	18q22.1	64897188	64906488	9.3	Amplification	—
EAGLE	19	19p13.3	1046061	1126396	80.3	Deletion	21521776
EAGLE	19	19p13.3	1994271	2001823	7.6	Deletion	21521776
EAGLE	19	19p13.3	2050820	2079054	28.2	Deletion	21521776
EAGLE	20	20q13.33	61642713	61668792	26.1	Abnormal	17304513
EAGLE	21	21q22.3	45769452	45788806	19.4	Deletion	15900585
EAGLE	22	22q13.1	37667446	37704618	37.2	Deletion	10515681; 15262437

This table reports the 22 predicted CNVs summarized from risk loci (Table S2) in EAGLE. The Literature shows the PubMed unique identifier (PMID) for previous papers that provide the risk evidence for these loci. See Section 4 for detailed information.

3.1. Population Structure Impact. At the study design step, we noted the population stratification in EAGLE samples (Figure S2). Although stratification only occurred within the whole population of cases and controls in EAGLE and no obvious stratification was observed between case and control cohorts, we were still concerned about whether the whole population stratification could lead to false positive results. The strategy we used to overcome this problem was to use another dataset, PLCO, as an independent verification dataset that could be integrated into our combined analysis with EAGLE. As expected, such a combined process indeed gave us meaningful results.

3.2. Fitting Our Results with Other GWA Studies. There was an obvious difference between our approach and other GWA studies. The majority of GWA studies performed association testing using the probe signal of an SNP site. However, our approach first transformed the original probe signal into a copy number state value using a hidden Markov model (HMM), followed by an association analysis using the transformed copy number state of an SNP site. Strictly speaking, our approach was actually a copy number state association testing. This copy number state transformation before association testing made our association risk loci unsuitable to compare directly with other GWA studies. Given this problem, we have chosen to conduct our analysis using SNP-based, window-based, and combined testing until we got a set of potentially reliable SNP risk loci. We found that there were some consecutive blocks formed in this set of risk loci (167 SNP risk loci in EAGLE, Table S2). We extracted the blocks from the set of candidate risk loci and found that these blocks were actually CNVs region predicted by our approach. We could then compare our CNVs with CNVs predicted by other studies including GWA studies.

Due to the strict three-step SNP-based, window-based, and combined analysis, when we did the literature search we were able to find direct or indirect support from previous studies for the majority of our predicted CNVs (see Table 2 for detailed validation information), which indicated the meaningfulness of our CNV predictions. We thought that, compared to other GWA studies, our approach had two advantages. First, in addition to the CNVs described by other GWA studies our approach found some new CNVs validated by other functional studies. In the popular SNP-based genome-wide association studies, some complex CNV regions might be difficult to analyze or filter by SNP site evaluation. Our approach transformed SNP signals to copy number states information, which might help our approach to maintain CNV information. Given the numerous available GWA study strategies, our approach might still give some valuable predictions that other strategies might miss. A second advantage is that our CNVs predictions did not always exactly overlap with the supporting studies' predictions. For functional evidence, our predictions might give more precise boundaries or positions of CNVs than rough functional

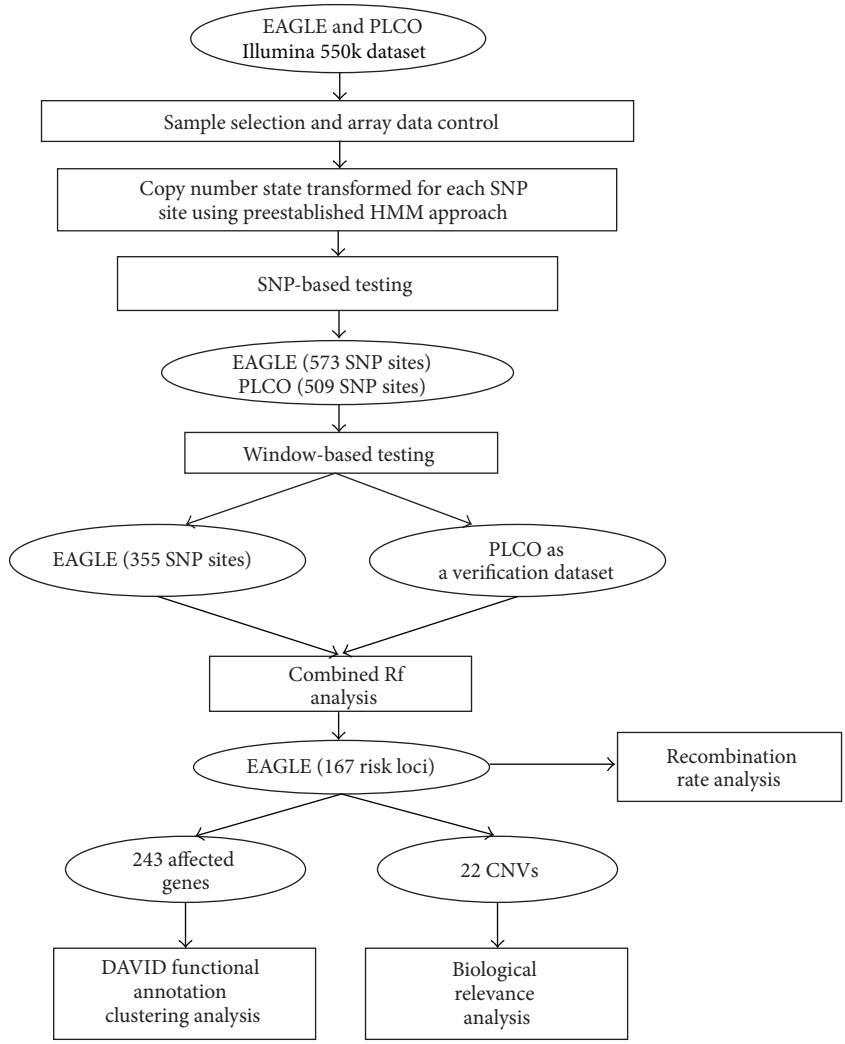


Figure 2: Workflow of the combined study and subsequent analysis in EAGLE.

studies because of the high-resolution array data we used. For other GWA evidence, our predictions might be a useful, complementing tool to locate CNVs more precisely.

When analyzing the risk loci, we mainly focused on CNV blocks extracted using consecutive information. However, there were many singular risk loci left (Table S2). We noted three singular risk loci among them: rs9863274 located on 3q24, rs104554013 located on 5q21.3, and rs952125 located on 21q21.1. The region 3q24 was a well-known CNV-associated locus identified in many studies of lung cancer [25, 30–32] and the amplification of this locus [30] was the most prominent difference between squamous cell carcinomas (SCCs) and adenocarcinomas (ACs) [31]. The loss of copy number in 5q21 had been previously reported to be associated with lung cancer [33] and the CNVs of this locus were implicated in clear-cell renal cell carcinoma in patients who smoked [34]. This locus might be critical in mediating interactions between environmental and genetic factors. Deletions of 21q21.1, which might correspond to a candidate tumor suppressor locus, had also been reported in lung cancer [35]. Given these CNVs supported by other studies and by our predictions, we expected these potential singular risk loci to be a set of candidate loci worthy of further functional validations.

Finally, we expected the approach developed in this study to be a valuable complement to current genome-wide association studies.

4. Materials and Methods

4.1. Data Source and Sample Selection. Our datasets were from the project "A Genome-Wide Scan of Lung Cancer and Smoking" (phs000093, the database of Genotypes and Phenotypes, dbGaP). This project consisted of two parts: (1) Environment And Genetics in Lung cancer Etiology (EAGLE) [16] and (2) the Prostate, Lung, Colon and Ovary (PLCO) Study Cancer Screening Trial [17]. These two datasets were carefully controlled for gender, age, region, and so forth. phs000093 also contained 66 individuals with European ancestry from HapMap which was used as a training dataset to estimate the parameters of the hidden Markov model (HMM).

Individuals with contamination from different genetic backgrounds and duplicated samples were filtered as per the instructions of phs000093. Finally, 1,945 cases and 1,992 controls were obtained for EAGLE and 803 cases and 848 controls were obtained for PLCO.

4.2. Array Data Preprocesses. The blood samples of all individuals were detected using an Illumina HumanHap 550K v3.0 genotyping chip, and these data were quantile normalized to the same baseline for further analysis. After quality control, we processed these data using SNPs annotated in the NCBI build 36 reference genome. As sex chromosomes are different from the autosomes in copy number detection and comparison, only the autosomes were studied in our work.

Finally, 547,458 autosomal SNPs annotated in NCBI build 36 reference genome were used for further analyses.

4.3. Population Stratification Analysis. First, we used PLINK to extract the genotype information of each SNP probe for each individual studied. Next, PLINK was used again to prune out SNPs in the 547,458 autosomal SNPs for linkage disequilibrium between SNPs with $r^2 > 0.2$. We then used EIGENSTRAT 3.0 software suite to do a raw smartpca analysis in EAGLE and PLCO. After the first run of smartpca, we analyzed the output snpweight of each SNP and manually removed large segments of closely flanked SNPs with abs (snpweight) > 3.5. Finally, we reran the smartpca to find top 20 significant eigenvectors in EAGLE and PLCO separately and then plotted the most significant eigenvector against the next four most significant eigenvectors for EAGLE and PLCO.

4.4. Copy Number State Transformation. In our analysis, the SNP probes signal data were first transformed to copy number state with a well-trained hidden Markov model (HMM). The training method and transformation process are described in our previous study [18].

4.5. The Simple Raw CNVs Prediction Method. In this roughly simple raw CNVs calculation step, three or more than three consecutive SNPs with the same abnormal copy number (not equal to 2) in an individual were considered to be a CNV of this sample. The description statistics in Table 1 were then calculated based on the raw CNVs generated. Note that such raw CNVs were not adopted as reliable CNVs and only used to make a comparison between EAGLE and PLCO.

4.6. Statistical Power Comparison between EAGLE and PLCO. Parameters for GWAPower calculation are as follows: CEU population; Illumina 550k platform; predefined P value of $5e-7$; effective size for SNPs: 1.1, 1.15, 1.2, 1.25, 1.3, 1.4, 1.5, and 1.7.2; EAGLE population size, 1945 versus 1992; and PLCO population size, 803 versus 848.

4.7. CNV Association Testing, Recombination Rate, and Relative Factor Calculation. CNV association testing in separate datasets was performed as a two-step statistical testing. *SNP-based testing* was performed to measure the disease association with a specific SNP site and *window-based testing* was performed to measure the CNV pattern differentiation in and around the selected SNP site. The details of these tests can be found in the original paper [18]. Here, the SNP site-based testing also includes *multiple trend testing* for a specific SNP site.

For recombination rate analysis, we downloaded the genome-wide recombination rate data from HapMap phase II (http://hapmap.ncbi.nlm.nih.gov/downloads/index.html.en). For each SNP site analyzed, we computed a sum of log10 (maximum recombination rate) within a 10 kb region of the SNP in order to represent the recombination rate level for this SNP and then compared this level among four groups of SNPs: not sig. (which consisted of nonsignificant SNPs), SNP-based (significant SNPs that passed SNP-based testing), window (significant SNPs that passed both SNP-based and window-based testing), and window_Rf (significant SNPs that passed SNP-based, window-based, and combined testing with PLCO) (see Figure 1(c)). Since HapMap phase II has already analyzed the recombination hotspot regions, we extracted the start and end positions of the hotspot regions and plotted them against the genome in Figure 1(d) and Figure S4.

The relative factor (Rf) was calculated in our analysis to test the association accordance on a specific SNP site between EAGLE and PLCO. The Rf was calculated from four models (M00, M01, M10, and M11) of the comparison between two datasets. In these models, we defined one of the datasets as the "Reference" dataset and the other one as the "Testing" dataset. Hence, the four models describe four comparisons between the two datasets.

M00: the cases' distributions in the Testing dataset were the same as the cases' distributions in the Reference dataset.

M10: the controls' distributions in the Testing dataset were the same as the cases' distributions in the Reference dataset.

M01: the cases' distributions in the Testing dataset were the same as the controls' distributions in the Reference dataset.

M11: the controls' distributions in the Testing dataset were the same as the controls' distributions in the Reference dataset.

Consider

$$\mathrm{Rf} = \frac{P(\mathrm{M00})\,P(\mathrm{M11})}{P(\mathrm{M01})\,P(\mathrm{M10})}. \quad (1)$$

Rf could be calculated using the formula shown above in which P values were calculated in the same manner as in SNP-based testing. We noted that the relative factor is a combination of the distribution accordance of both cases and controls. When the value is higher, the association patterns

in each dataset are more similar. The advantage of using Rf is that it is compatible with the multiple hypotheses model of SNP-based testing, which was effective in our previous study [18].

4.8. Correction of Multiple Tests by Calculating False Discovery Rates (FDR). Confirming the significance of multiple tests for a CNV association study is an important issue in genome-wide association analysis. CNV association P values are not independent of but tend to be related to the neighboring sites because CNVs may span thousands of nucleotides in the human genome. Classical Bonferroni correction was not adopted in our analysis but a permutation-based method was used to calculate false discovery rates (FDR) of a significant level. The FDR of *SNP-based testing* and *window-based testing* were calculated in the same way as in our previous work [18].

In the case of relative factors, the case-control labels for all individuals were permuted 100 times and the calculation of every model was according to the previous work [18]. The FDR could be calculated as follows:

$$\mathrm{FDR}_{\mathrm{Rf}} = \frac{N^{\mathrm{SNP}}_{\mathrm{Rf}^{(m)} \le \mathrm{Rf}_{\mathrm{site}}}}{T_{\mathrm{pm}} \cdot N^{\mathrm{SNP}}_{\mathrm{Rf} \le \mathrm{Rf}_{\mathrm{site}}}}. \quad (2)$$

$\mathrm{Rf}_{\mathrm{site}}$ denotes a designated Rf value in the observed data, Rf and $\mathrm{Rf}^{(m)}$ denote the Rf values in the observed data and the permutated data, respectively, N^{SNP} denotes the number of SNP sites, and T_{pm} denotes the number of permutations.

4.9. Predictions of CNVs Using Risk Loci in EAGLE. We predicted a set of potential reliable CNVs from 167 risk loci in EAGLE (Table S2). We manually investigated the 167 risk loci using the following standards to predict CNVs.

(i) CNVs should span three or more than three consecutive risk loci.

(ii) "Consecutive" means the distance between two neighboring SNPs should not be larger than 30 kb.

(iii) The type of CNVs depends on the related P value in Table S2.

4.10. Software Tools Used in This Study. We used PLINK 1.07 (Shaun Purcell, http://pngu.mgh.harvard.edu/purcell/plink/) [36], EIGENSTRAT [37], GWAPower [38], and DAVID [39] in this study.

Conflict of Interests

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

Authors' Contribution

The study was designed by Xinlei Li, Xianfeng Chen, Landian Hu, and Xiangyin Kong. Data were generated and prepared by Xinlei Li, Xianfeng Chen, Yang Liu, Zhenguo Zhang, Ping Wang, Yufei Zhu, Xianfu Yi, Jie Zhang, You Zhou, Zejun Wei, and Fei Yuan. Data were analyzed by Xinlei Li, Xianfeng Chen, Guohong Hu, Yang Liu, Guoping Zhao, Jun Zhu, Landian Hu, and Xiangyin Kong. The paper was prepared by Xinlei Li, Xianfeng Chen, and Xiangyin Kong. Xinlei Li and Xianfeng Chen contributed equally to this work.

Acknowledgments

The authors acknowledge the project "A Genome-Wide Scan of Lung Cancer and Smoking" (phs000093.v2.p2), CGEMS, for providing EAGLE and PLCO datasets. These data were obtained from and inspected by dbGaP. This work is supported by the National Basic Research Program of China (no. 2011CB510100) and the National Natural Science Foundation of China (no. 81030015).

References

[1] A. Jemal, R. Siegel, J. Xu, and E. Ward, "Cancer statistics, 2010," *CA Cancer Journal for Clinicians*, vol. 60, no. 5, pp. 277–300, 2010.

[2] C. I. Amos, X. Wu, P. Broderick et al., "Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1," *Nature Genetics*, vol. 40, no. 5, pp. 616–622, 2008.

[3] R. J. Hung, J. D. McKay, V. Gaborieau et al., "A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25," *Nature*, vol. 452, pp. 633–637, 2008.

[4] M. T. Landi, N. Chatterjee, K. Yu et al., "A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma," *The American Journal of Human Genetics*, vol. 85, no. 5, pp. 679–691, 2009.

[5] J. D. McKay, R. J. Hung, V. Gaborieau et al., "Lung cancer susceptibility locus at 5p15.33," *Nature Genetics*, vol. 40, no. 12, pp. 1404–1406, 2008.

[6] T. E. Thorgeirsson, F. Geller, P. Sulem et al., "A variant associated with nicotine dependence, lung cancer and peripheral arterial disease," *Nature*, vol. 452, no. 7187, pp. 638–642, 2008.

[7] Y. Wang, P. Broderick, E. Webb et al., "Common 5p15.33 and 6p21.33 variants influence lung cancer risk," *Nature Genetics*, vol. 40, no. 12, pp. 1407–1409, 2008.

[8] R. Redon, S. Ishikawa, K. R. Fitch et al., "Global variation in copy number in the human genome," *Nature*, vol. 444, no. 7118, pp. 444–454, 2006.

[9] P. Stankiewicz and J. R. Lupski, "Genome architecture, rearrangements and genomic disorders," *Trends in Genetics*, vol. 18, no. 2, pp. 74–82, 2002.

[10] C. J. Shaw and J. R. Lupski, "Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease," *Human Molecular Genetics*, vol. 13, no. 1, pp. R57–R64, 2004.

[11] S. J. Diskin, C. Hou, J. T. Glessner et al., "Copy number variation at 1q21.1 associated with neuroblastoma," *Nature*, vol. 459, no. 7249, pp. 987–991, 2009.

[12] H. Vauhkonen, M. Vauhkonen, A. Sajantila, P. Sipponen, and S. Knuutila, "DNA copy number aberrations in intestinal-type gastric cancer revealed by array-based comparative genomic hybridization," *Cancer Genetics and Cytogenetics*, vol. 167, no. 2, pp. 150–154, 2006.

[13] R. P. Kuiper, M. J. L. Ligtenberg, N. Hoogerbrugge, and A. Geurts van Kessel, "Germline copy number variation and cancer risk," *Current Opinion in Genetics and Development*, vol. 20, no. 3, pp. 282–289, 2010.

[14] B. Xu, J. L. Roos, S. Levy, E. J. van Rensburg, J. A. Gogos, and M. Karayiorgou, "Strong association of de novo copy number mutations with sporadic schizophrenia," *Nature Genetics*, vol. 40, no. 7, pp. 880–885, 2008.

[15] E. Gonzalez, H. Kulkarni, H. Bolivar et al., "The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility," *Science*, vol. 307, no. 5714, pp. 1434–1440, 2005.

[16] M. T. Landi, D. Consonni, M. Rotunno et al., "Environment and Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer," *BMC Public Health*, vol. 8, article 203, 2008.

[17] R. B. Hayes, A. Sigurdson, L. Moore et al., "Methods for etiologic and early marker investigations in the PLCO trial," *Mutation Research—Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 592, no. 1-2, pp. 147–154, 2005.

[18] X. Chen, X. Li, P. Wang et al., "Novel association strategy with copy number variation for identifying new risk loci of human diseases," *PLoS ONE*, vol. 5, no. 8, Article ID e12185, 2010.

[19] F. Dudbridge, A. Gusnanto, and B. P. C. Koeleman, "Detecting multiple associations in genome-wide studies," *Human Genomics*, vol. 2, no. 5, pp. 310–317, 2006.

[20] M. García-Closas and J. H. Lubin, "Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches," *The American Journal of Epidemiology*, vol. 149, no. 8, pp. 689–692, 1999.

[21] M. L. Freedman, D. Reich, K. L. Penney et al., "Assessing the impact of population stratification on genetic association studies," *Nature Genetics*, vol. 36, no. 4, pp. 388–393, 2004.

[22] C. Tian, P. K. Gregersen, and M. F. Seldin, "Accounting for ancestry: population substructure and genome-wide association studies," *Human Molecular Genetics*, vol. 17, no. 2, pp. R143–R150, 2008.

[23] M. I. Lerman and J. D. Minna, "The 630-kb lung cancer homozygous deletion region on human chromosome 3p21.3: identification and evaluation of the resident candidate tumor suppressor genes," *Cancer Research*, vol. 60, no. 21, pp. 6116–6133, 2000.

[24] E. A. Anedchenko, A. A. Dmitriev, G. S. Krasnov et al., "Down-regulation of RBSP3/CTDSPL, NPRL2/G21, RASSF1A, ITGA9, HYAL1 and HYAL2 genes in non-small cell lung cancer," *Molekuliarnaia Biologiia*, vol. 42, no. 6, pp. 965–976, 2008.

[25] M. C. Boelens, K. Kok, P. van der Vlies et al., "Genomic aberrations in squamous cell lung carcinoma related to lymph node or distant metastasis," *Lung Cancer*, vol. 66, no. 3, pp. 372–378, 2009.

[26] S. Haruki, I. Imoto, K. Kozaki et al., "Frequent silencing of protocadherin 17, a candidate tumour suppressor for esophageal squamous cell carcinoma," *Carcinogenesis*, vol. 31, no. 6, pp. 1027–1036, 2010.

[27] A. E. Gylfe, J. Sirkia, M. Ahlsten et al., "Somatic mutations and germline sequence variants in patients with familial colorectal cancer," *International Journal of Cancer*, vol. 127, no. 12, pp. 2974–2980, 2010.

[28] C. Floris, S. Rassu, L. Boccone, D. Gasperini, A. Cao, and L. Crisponi, "Two patients with balanced translocations and autistic disorder: CSMD3 as a candidate gene for autism found in their common 8q23 breakpoint area," *European Journal of Human Genetics*, vol. 16, no. 6, pp. 696–704, 2008.

[29] H. Zayed, R. Chao, A. Moshrefi et al., "A maternally inherited chromosome 18q22.1 deletion in a male with late-presenting diaphragmatic hernia and microphthalmia—evaluation of DSEL as a candidate gene for the diaphragmatic defect," *The American Journal of Medical Genetics A*, vol. 152, no. 4, pp. 916–923, 2010.

[30] E. Kettunen, W. El-Rifai, A. Björkqvist et al., "A broad amplification pattern at 3q in squamous cell lung cancer—a fluorescence in situ hybridization study," *Cancer Genetics and Cytogenetics*, vol. 117, no. 1, pp. 66–70, 2000.

[31] J. Pei, B. R. Balsara, W. Li et al., "Genomic imbalances in human lung adenocarcinomas and squamous cell carcinomas," *Genes Chromosomes and Cancer*, vol. 31, no. 3, pp. 282–287, 2001.

[32] J. Qian and P. P. Massion, "Role of chromosome 3q amplification in lung cancer," *Journal of Thoracic Oncology*, vol. 3, no. 3, pp. 212–215, 2008.

[33] C. A. Cooper, V. J. Bubb, N. Smithson et al., "Loss of heterozygosity at 5q21 in non-small cell lung cancer: a frequent event but without evidence of apc mutation," *The Journal of Pathology*, vol. 180, pp. 33–37, 1996.

[34] Y. Korenaga, H. Matsuyama, H. Hirata et al., "Smoking may cause genetic alterations at 5q22.2~q23.1 in clear-cell renal cell carcinoma," *Cancer Genetics and Cytogenetics*, vol. 163, no. 1, pp. 7–11, 2005.

[35] H. Yamada, K. Yanagisawa, S. Tokumaru et al., "Detailed characterization of a homozygously deleted region corresponding to a candidate tumor suppressor locus at 21q11-21 in human lung cancer," *Genes Chromosomes and Cancer*, vol. 47, no. 9, pp. 810–818, 2008.

[36] S. Purcell, B. Neale, K. Todd-Brown et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," *American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[37] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[38] C. C. Spencer, Z. Su, P. Donnelly, and J. Marchini, "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip," *PLoS Genetics*, vol. 5, no. 5, Article ID e1000477, 2009.

[39] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 891945, 8 pages
http://dx.doi.org/10.1155/2014/891945

Research Article

A Graphic Method for Identification of Novel Glioma Related Genes

Yu-Fei Gao,[1] Yang Shu,[2] Lei Yang,[3] Yi-Chun He,[1] Li-Peng Li,[1] GuaHua Huang,[4] Hai-Peng Li,[5] and Yang Jiang[1]

[1] *Department of Surgery, China-Japan Union Hospital of Jilin University, Changchun 130033, China*
[2] *State Key Laboratory of Medical Genomics, Institute of Health Sciences, Shanghai Jiaotong University School of Medicine and Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200025, China*
[3] *Endoscopy Center, China-Japan Union Hospital of Jilin University, Changchun 130033, China*
[4] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[5] *CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*

Correspondence should be addressed to Yang Jiang; jy7555@163.com

Received 18 April 2014; Revised 25 May 2014; Accepted 28 May 2014; Published 23 June 2014

Academic Editor: Tao Huang

Glioma, as the most common and lethal intracranial tumor, is a serious disease that causes many deaths every year. Good comprehension of the mechanism underlying this disease is very helpful to design effective treatments. However, up to now, the knowledge of this disease is still limited. It is an important step to understand the mechanism underlying this disease by uncovering its related genes. In this study, a graphic method was proposed to identify novel glioma related genes based on known glioma related genes. A weighted graph was constructed according to the protein-protein interaction information retrieved from STRING and the well-known shortest path algorithm was employed to discover novel genes. The following analysis suggests that some of them are related to the biological process of glioma, proving that our method was effective in identifying novel glioma related genes. We hope that the proposed method would be applied to study other diseases and provide useful information to medical workers, thereby designing effective treatments of different diseases.

1. Introduction

Glioma is the most common and lethal intracranial tumor. It always revealed itself as malignant glioma which is usually divided into astrocytoma, oligodendroglioma, and oligoastrocytoma. Besides the classification based on histopathological features, glioma could also be graded on a WHO consensus-derived scale of I to IV by means of the degree of malignancy [1]. Clinically, most of the gliomas are high-grade gliomas (HGG). Glioblastoma (GBM), one of the HGG, accounts for more than half of gliomas [1]. Although the knowledge of glioma, especially HGG, has increased dramatically in recent years, many questions are still waiting for further elucidation. On the other hand, the overall 5-year survival rate of GBM remains less than 5% despite the advances in surgery, radiation, and chemotherapy [2].

In the previous reports, glioma always manifested itself with disordered pathways which regulated proliferation, survival, invasion, and angiogenesis. Among these biological processes, RB and p53 pathways are more inclined to be dysfunctional in GBM, and the disrepair could lead to the destruction of cell cycle by regulating the G1-to-S-phase transition [3, 4]. Furthermore, other pathways such as MAPK, PI3K/PTEN/AKT, and NF-κB pathway are also overactivated in glioma and contribute to the uncontrollable cellular proliferation [5–7]. As we know, the tumorigenesis of glioma is a complicated process which involved intricate pathways beyond the above ones. To widely understand the mechanism underlying this disease, identification of its related genes and uncovering the relationship of them and the biological process of glioma are very important. However, it is time-consuming and expensive to identify novel glioma related

genes by conventional experiments. On the other hand, encouraged by the successful application of computational methods to deal with various biological problems such as drug design [8–13] and analysis of complicated biological pathway [14–17], computational methods may address this problem and provide some useful information for investigators.

In this study, we proposed a graphic method and attempted to apply this method to discover novel glioma related genes. The current known glioma related genes collected from various sources were the firsthand information. Based on these genes, some new discovered genes were obtained by the well-known shortest path algorithm. Furthermore, a permutation test was conducted to exclude false positives among them. The analysis of the final remaining genes suggests that some of them had direct or indirect relationship to the biological process of glioma, indicating that this method was effective and may give new insight to study other diseases.

2. Materials and Methods

2.1. Materials. The current known glioma related genes were retrieved from the following sources. (1) All the 11 data sheets listed on the web page of COSMIC (http://cancer.sanger.ac.uk/cancergenome/projects/census/) were downloaded, from which we obtained 18 glioma related genes; (2) search for human diseases in UniProt (http://www.uniprot.org/) with keywords "human glioma oncogene" and "human glioma suppressor gene," thereby obtaining 49 and 32 genes (only reviewed genes were selected), respectively; (3) select "Literature Search" in TSGene (http://bioinfo.mc.vanderbilt.edu/TSGene/search.cgi) and input "glioma" as keyword, obtaining 7 genes. After collecting all of the genes mentioned above, we finally obtained 77 glioma related genes, which were available in Supplementary Material I available online at http://dx.doi.org/10.1155/2014/891945.

2.2. Construction of a Weighted Graph from Protein-Protein Interactions. Protein-protein interaction (PPI) is useful information for investigating various biological problems [18–22]. Many computational methods were proposed based on the fact that proteins that can interact with each other always share similar functions. Since the known glioma related genes must have some common features related to glioma, it is reasonable to discover novel glioma related genes based on protein-protein interaction and known glioma related genes. The data concerning protein-protein interactions was downloaded from STRING (Search Tool for the Retrieval of Interacting Genes/Proteins, http://string.embl.de/) [23], a large database containing known and predicted protein interactions which are derived from genomic context, high-throughput experiments, (Conserved) Coexpression and Previous knowledge. In the obtained file, we extracted all protein-protein interactions of human. Each obtained interaction consists of two proteins and a score with range between 150 and 999, which can quantify the likelihood that an interaction may occur. For later formulation, let $Q(p_1, p_2)$ denote the score of the interaction between two proteins p_1 and p_2. The constructed graph took proteins occurring at least one protein-protein interaction of human as nodes, while two nodes were adjacent if and only if the score of the interaction between the corresponding proteins was greater than zero. The obtained graph consisted of 18,600 nodes and 1,640,707 edges. Furthermore, to correctly reflect the strength of the interaction, each edge with nodes v_1 and v_2 was labeled by a weight, which can be computed by

$$W(v_1, v_2) = 1000 - Q(p_1, p_2), \quad (1)$$

where p_1 and p_2 were two corresponding proteins of nodes v_1 and v_2, respectively.

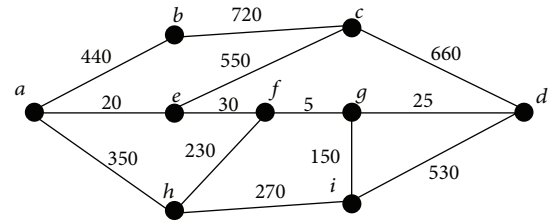


Figure 1: A simple example of the weighted graph.

2.3. Selection of Candidate Genes. It is obvious that the glioma related genes must have some common features which are related to glioma. On the other hand, as mentioned in Section 2.2, two proteins that can interact with each other, that is, they are adjacent in the constructed weighted graph, always share common features. The idea of our method was based on these facts. To clearly elaborate the idea of the method, we constructed a simple weighted graph which is shown in Figure 1, because the original graph was too large to exhibit in the paper. It is easy to observe from Figure 1 that the shortest path connecting a and d contains e, f and g as the inner nodes. Based on the weights of edges on this path, we can obtain that genes a and e can share common functions with high probability, because the confidence score of the interaction between a and e is very high, which is $1000 - 20 = 980$. The similar results also hold for e and f, f and g, g and d. If genes a and d are two known glioma related genes, genes e, f, and g are actual glioma related genes with high likelihood. In view of this, shortest paths between any pair of known glioma related genes obtained by Dijkstra's algorithm, the most famous shortest path algorithm proposed by Dijkstra in 1956 [24], are useful information for further investigation.

After obtaining the shortest paths connecting any pair of known glioma related genes, it can be seen that some nodes/genes occurred in many paths, while most of nodes/genes in the graph were not contained in any path. Thus, for each node/gene in the graph, we counted the number of paths containing the node/gene, termed as betweenness which is defined as the number of shortest paths containing the node/gene as an inner node. The concept of betweenness

has been employed in some studies of natural and man-made networks [25–29]. In fact, the betweenness of some node/gene reflects the direct and indirect relationship of the gene and known glioma related genes. Thus, the likelihood of genes with high betweenness to be related to glioma was higher than those with low betweenness. In view of this, we selected genes with betweenness greater than 0 as the candidate genes, which may be the novel glioma related genes with high probability. It is necessary to point out that the known glioma related genes were not included in the set of candidate genes.

2.4. Filtering Candidate Genes by Permutation Test. As described in Section 2.3, some candidate genes can be obtained by researching the shortest paths connecting any two known glioma related genes. However, some of them may be false positives, because some nodes/genes may easily receive a high betweenness due to their location in the weighted graph even if we randomly select genes in STRING as the known glioma related genes. To exclude these false discoveries, a permutation test should be executed as follows.

(I) 1,000 node/gene sets, denoted by $G_1, G_2, \ldots, G_{1000}$, were randomly selected in the weighted graph such that each of them had the same size of known glioma related gene set.

(II) For each candidate gene discovered in Section 2.3, calculate its betweenness on each set G_i ($1 \leq i \leq 1000$).

(III) Calculate the permutation FDR of each candidate gene p by

$$\mathrm{FDR}(p) = \frac{\sum_{i=1}^{1000} \delta_i}{1000}, \tag{2}$$

where δ_i was set to be 1 if the betweenness of p on G_i was larger than that of p on the known glioma related gene set; otherwise, it was set to be 0.

Obviously, small permutation FDR of one candidate gene implies that it is the true positive with high probability.

3. Results and Discussions

3.1. Candidate Genes. For the 77 genes mentioned in Section 2.1, we searched the shortest path connecting any two of them by Dijkstra algorithm. After calculating the betweenness of each node/gene in the weighted graph, 215 candidate genes with betweenness larger than zero were obtained. These 215 genes and their betweenness were available in Supplementary Material II. To exclude the false positives, the permutation test was conducted as described in Section 2.4. By (2), we can calculate the permutation FDR of each candidate gene. These values were also provided in Supplementary Material II. Since the likelihood of gene with small permutation FDR to be the actual glioma related gene is high, we set the threshold to be 0.05, that is, selecting genes with permutation FDRs lower than 0.05 among 215 candidate genes, thereby obtaining 67 genes, listed in Table 1. These genes were deemed to have strong relationship with glioma and further discussions were based on these genes.

3.2. Analysis of Enriched KEGG Pathways of Candidate Genes. As mentioned in Section 3.1, 67 candidate genes were obtained. To analyze the relationship of them and glioma, we employed DAVID (Database for Annotation, Visualization and Integrated Discovery) [30], a functional annotation tool to understand biological meaning behind large list of genes. 67 candidate genes comprised the input gene list of DAVID, thereby obtaining 9 KEGG pathways that were enriched by these 67 candidate genes. The detailed output of DAVID for KEGG pathway enrichment analysis was available as Supplementary Material III.

The top 5 pathways have the P value less than 0.05 which were discussed below. The most enriched pathway is hsa04360: axon guidance ("count" = 8). Among the 8 genes, 7 ephrins-related genes are enriched whose corresponding proteins include 4 members (EFNA3, EFNB1, EFNB2, and EFNB3) of the ephrins family and 3 members (EPHA1, EPHA4, and EPHA7) of the ephrins receptor subfamily. Eph receptor tyrosine kinases (Ephs) and ephrins (EPH) could navigate cells by controlling cell-cell adhesion and segregation [31]. In other words, with the function of axon guidance Ephs/EPH could regulate the invasion, neoangiogenesis, and metastasis of gliomas [32, 33]. Ding and his colleagues have identified several somatic mutations of Ephs especially EphA7 in lung cancer [34]. Although the close connection between Ephs/EPH and cancer has been reported, its pathogenic mechanism in gliomas is still unknown. The second pathway is hsa04510: focal adhesion ("count" = 7). As we know, infiltration of tumor cells and angiogenesis are critical for the growth of tumor. Zagzag et al. reported that focal adhesion kinase (FAK), highly associated with these biological processes, plays an important role in tumorigenesis of gliomas via enhancing the ability of infiltration and angiogenesis [35]. The FAK-related genes, like CTNNB1, VEGFA, KDR, and FLT4 enriched in this pathway, are always mutated or aberrantly expressed in various types of cancers [36, 37]. The third pathway is hsa04530: tight junction ("count" = 5). In the brain, the expression of the tight junction proteins is important for blood-brain tumor-barrier (BTB) permeability. Hence destruction of the tight junction could facilitate the development of gliomas by increasing BTB permeability [38, 39]. The fourth pathway is hsa04520: adherens junction ("count" = 4). Adherens junction is reported to be disordered in the glioblastoma and to affect the invasive behavior of GBM [40, 41]. The last significantly enriched pathway is hsa05200: pathways in cancer ("count" = 7). The result shows that a common mechanism is shared by the gliomas and other types of cancers. Although these significant enriched pathways have been reported to be related to gliomas more or less, our results might expand the avenues to explore new mechanisms in the tumorigenesis of gliomas.

3.3. Analysis of Enriched GO Terms Candidate Genes. In addition to KEGG pathway enrichment analysis, DAVID also provided the GO terms enrichment analysis of the

Table 1: Candidate genes with permutation FDR lower than 0.05.

Ensemble ID of candidate gene	Gene name	Betweenness	Permutation FDR
ENSP00000227638	PANX1	72	0
ENSP00000235310	MAD2L2	72	0
ENSP00000245323	EFNB2	335	0
ENSP00000258428	REV1	72	0
ENSP00000265727	ADAM22	72	0
ENSP00000281821	EPHA4	339	0
ENSP00000293831	EIF4A1	72	0
ENSP00000302719	KCNAB3	72	0
ENSP00000312697	DMAP1	88	0
ENSP00000329797	CADM1	72	0
ENSP00000335434	WDR20	72	0
ENSP00000341138	EPB41L3	72	0
ENSP00000351697	REV3L	72	0
ENSP00000354778	CNTNAP2	72	0
ENSP00000356150	MDM4	72	0
ENSP00000357177	ARHGEF11	142	0
ENSP00000361366	SFTPD	72	0
ENSP00000369218	RBM17	72	0
ENSP00000370119	SMN2	72	0
ENSP00000245304	RAP2A	72	0.001
ENSP00000275815	EPHA1	72	0.001
ENSP00000328511	KCNA4	72	0.001
ENSP00000377446	SUCLG1	210	0.001
ENSP00000399511	TNIK	72	0.001
ENSP00000229595	ASF1A	72	0.002
ENSP00000252699	ACTN4	72	0.002
ENSP00000263208	HIRA	72	0.002
ENSP00000263923	KDR	104	0.002
ENSP00000304169	PITX2	210	0.002
ENSP00000330633	CNTN2	72	0.002
ENSP00000276072	TAF1	72	0.003
ENSP00000295600	MITF	138	0.003
ENSP00000360157	FOXD3	6	0.003
ENSP00000264010	CTCF	72	0.004
ENSP00000271628	SF3B4	75	0.004
ENSP00000350941	SRC	421	0.004
ENSP00000361125	VEGFA	164	0.004
ENSP00000226091	EFNB3	67	0.006
ENSP00000358309	EPHA7	2	0.006
ENSP00000358918	SUFU	72	0.007
ENSP00000316879	EIF4G1	72	0.011
ENSP00000260653	SIX3	2	0.012
ENSP00000352516	DNMT1	90	0.016
ENSP00000358716	DDX20	72	0.017
ENSP00000357393	EFNA3	50	0.018
ENSP00000341680	DTNBP1	66	0.02
ENSP00000344456	CTNNB1	607	0.02
ENSP00000297904	FIGF	2	0.021
ENSP00000265165	LEF1	134	0.022

Table 1: Continued.

Ensemble ID of candidate gene	Gene name	Betweenness	Permutation FDR
ENSP00000347948	TNFRSF14	68	0.023
ENSP00000288986	NCK1	78	0.027
ENSP00000261937	FLT4	2	0.029
ENSP00000333919	BTLA	68	0.03
ENSP00000332549	GRIN2A	58	0.032
ENSP00000376765	PIAS3	4	0.033
ENSP00000361818	SDC4	1	0.035
ENSP00000386165	CEBPD	38	0.035
ENSP00000348307	SIRPA	34	0.036
ENSP00000344666	NF2	1	0.037
ENSP00000219255	PARD6A	72	0.038
ENSP00000204961	EFNB1	5	0.039
ENSP00000172229	NGFR	72	0.043
ENSP00000344115	CDH5	24	0.043
ENSP00000405041	POU5F1	6	0.045
ENSP00000360532	CDC5L	6	0.046
ENSP00000295897	ALB	72	0.047
ENSP00000340944	PTPN11	112	0.047

67 candidate genes, which were available in Supplementary Material IV.

It can be seen that 227 GO terms were enriched by these 67 genes. Top 10 Go terms sorted by P value are investigated and discussed as below. Among the top 10, 4 GO terms are biological process (BP) which included GO: 0007169: transmembrane receptor protein tyrosine kinase signaling pathway, GO: 0007167: enzyme linked receptor protein signaling pathway, GO: 0042127: regulation of cell proliferation, and GO: 0000904: cell morphogenesis involved in differentiation. From the results, we found that all these processes are connected with receptor-dependent signaling pathways. The cancer genome atlas (TCGA) group has revealed that the receptor tyrosine kinase (RTK) pathway was deregulated in 88% of the patients with glioblastoma [42]. After deregulation of the RTK, its downstream genes could function uncontrollably in the cellular proliferation and morphogenesis which are very pivotal for the growth of gliomas. In the top 10 GO terms, we also find 4 molecular function (MF) GO terms: GO: 0004714: transmembrane receptor protein tyrosine kinase activity, GO: 0005003: ephrin receptor activity, GO: 0046875: ephrin receptor binding, and GO: 0004713: protein tyrosine kinase activity. The MF classification also suggests the importance of RTK signaling pathways especially the Ephs/EPH pathway in the tumorigenesis of gliomas. As the previous reports, RTK pathways could regulate cell proliferation and migration which were indispensable for the development of gliomas [42, 43]. Besides BP and MF GO terms, 2 cellular component (CC) GO terms are also enriched in the top 10 GO terms: GO: 0044459: plasma membrane part and GO: 0005887: integral to plasma membrane. As we know, the transformation of cell membrane is necessary for the migration and invasion process during tumorigenesis of gliomas. Our results pave the way for understanding potential pathogenic mechanism of gliomas.

3.4. Analysis of Some Candidate Genes. Among the 67 genes, several genes are intriguing which may play pivotal role in tumorigenesis of glioma. This section gave the detailed discussion of some candidate genes.

VEGFA, also known as the vascular endothelial growth factor (VEGF), is a member of a large family of growth factors that also includes VEGFB, VEGFC, VEGFD, and placental growth factor (PLGF). VEGF is the only mitogen that specifically acts on endothelial cells and also a tumor angiogenesis factor in human glioma *in vivo* [44]. Knizetova et al. have demonstrated that the autocrine VEGF signaling is mediated via VEGFR2 (KDR), another gene in our list. They found that blockade of VEGFR2 would abrogate the VEFG-mediated enhancement of astrocytoma cell growth and viability [37]. In the *in vivo* level, Millauer et al. found that disrepair of VEGFR2/VEGF system in angiogenesis could prevent tumor growth in nude mice [45]. Another VEGF receptor found in our list is VEGFR3 (also known as FLT4). In contrast to VEGFR1/2, VEGFR3 does not bind VEGFA and mainly functions in lymphangiogenesis as a receptor of VEGFC and VEGFD [46, 47]. Jenny et al. reported that VEFGR3 was expressed in some tumor types such as haemangioblastoma and glioblastoma, despite their lack of lymphatic vessels [48]. Although the roles of VEGF signaling pathway in the tumorigenesis of glioma have been well studied, new findings have been explored in succession recently.

CTNNB1, with more famous name of catenin beta 1, encodes β-catenin protein which plays important roles in cellular morphogenesis, differentiation, and proliferation via regulating the Wnt signaling [49]. Yano et al. induced glioma in rat using N-ethyl-N-nitrosourea display aberrant nuclear accumulation of β-catenin in contrast to normal brains [50]. Moreover, Pu et al. found that the downregulation of β-catenin by siRNA could suppress malignant glioma

cell growth [36]. To elucidate the connection between β-catenin and glioma, Liu and his colleagues performed a systemic research. They found a higher expression level of β-catenin in astrocytic glioma patients with high grade in comparison with the normal controls. Furthermore, they also illustrated that the overexpression of β-catenin may be an important contributing factor to glioma progression by means of facilitating proliferation and inhibiting apoptosis [51]. After Wnt pathway is activated, β-catenin accumulates and enters the nucleus where it can act as a coactivator for TCF/LEF-mediated transcription [52]. As the downstream of beta-catenin, LEF1, another gene in our list, also plays a crucial role in the Wnt signaling pathway. LEF1, with full name of lymphoid enhancer-binding factor 1, tends to be mutated in the tumors. Liu et al. have investigated that MiR-218 could reduce the invasiveness of glioblastoma cells by targeting LEF1 [53].

SRC, whose corresponding protein is a tyrosine-protein kinase, could play a pivotal role in the regulation of embryonic development and cell growth [54]. Besides functioning in the embryonic development, SRC could also regulate the tumorigenesis of various types of cancers like breast cancer, colon cancer, and brain cancer [55, 56]. Src protein always maintains an inactive state until its Y530 residue is dephosphorylated by protein tyrosine phosphatase-α [57]. Src could also be activated by direct binding of its SH2 and SH3 domains to intracellular proteins or activated tyrosine kinase growth factor receptors [58]. Stettner et al. have found elevated SRC activity in GBM compared with normal brain [59]. On the other hand, Lund et al. found that the infiltration of glioma reduced in Src-deficient mice [60]. It is reported that the increased SRC activity in GBM may be due to increased activation of cell surface growth factor receptors and integrins that activate SRC-family kinases (SFKs) rather than the amplification or mutation of SFK genes [42, 61].

4. Conclusion

In biomedicine and genomics, identification of disease genes is an important topic. This contribution proposed a graphic method to identify novel disease genes and the method was applied to glioma, one kind of cancers. The findings indicate that this method is quite effective. It is hopeful that the contribution can provide help for medical workers to discover effective treatments of glioma and give new insight to study various diseases.

Conflict of Interests

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

Authors' Contribution

Yu-Fei Gao and Yang Shu contributed equally to this work.

Acknowledgments

This paper is supported by the National Science Foundation of China (81372696), the Natural Science Fund Projects of Jilin province (201215059), the Development of Science and Technology Plan Projects of Jilin province (20100733, 201101074), and SRF for ROCS, SEM (2009-36), Scientific Research Foundation (Jilin Department of Science & Technology; 200705314, 20090175, and 20100733), Scientific Research Foundation (Jilin Department of Health, 2010Z068), and SRF for ROCS (Jilin Department of Human Resource & Social Security, 2012–2014).

References

[1] D. N. Louis, H. Ohgaki, O. D. Wiestler et al., "The 2007 WHO classification of tumours of the central nervous system," *Acta Neuropathologica*, vol. 114, pp. 97–109, 2007.

[2] T. A. Dolecek, J. M. Propp, N. E. Stroup, and C. Kruchko, "CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2005–2009," *Neuro-Oncology*, vol. 14, supplement 5, pp. v1–v49, 2012.

[3] D. N. Louis, "The p53 gene and protein in human brain tumors," *Journal of Neuropathology and Experimental Neurology*, vol. 53, no. 1, pp. 11–21, 1994.

[4] J. W. Henson, B. L. Schnitker, K. M. Correa et al., "The retinoblastoma gene is involved in malignant progression of astrocytomas," *Annals of Neurology*, vol. 36, no. 5, pp. 714–721, 1994.

[5] A. Guha, M. M. Feldkamp, N. Lau, G. Boss, and A. Pawson, "Proliferation of human malignant astrocytomas is dependent on Ras activation," *Oncogene*, vol. 15, no. 23, pp. 2755–2765, 1997.

[6] C. B. Knobbe, A. Trampe-Kieslich, and G. Reifenberger, "Genetic alteration and expression of the phosphoinositol-3-kinase/Akt pathway genes *PIK3CA* and *PIKE* in human glioblastomas," *Neuropathology and Applied Neurobiology*, vol. 31, pp. 486–490, 2005.

[7] A. L. Rinkenbaugh and A. S. Baldwin, "Monoallelic deletion of NFKBIA in glioblastoma: when less is more," *Cancer Cell*, vol. 19, no. 2, pp. 163–165, 2011.

[8] L. Chen, W. M. Zeng, Y. D. Cai, K. Y. Feng, and K. C. Chou, "Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities," *PLoS ONE*, vol. 7, no. 4, Article ID e35254, 2012.

[9] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.

[10] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. 12, pp. i246–i254, 2010.

[11] L. Chen, J. Lu, X. Luo, and K.-Y. Feng, "Prediction of drug target groups based on chemical-chemical similarities and chemical-chemical/protein connections," *Biochimica et Biophysica Acta: Proteins and Proteomics*, vol. 1844, no. 1, pp. 207–213, 2014.

[12] B. Padhy and Y. Gupta, "Drug repositioning: re-investigating existing drugs for new therapeutic indications," *Journal of Postgraduate Medicine*, vol. 57, no. 2, pp. 153–160, 2011.

[13] L. Chen, J. Lu, N. Zhang, T. Huang, and Y.-D. Cai, "A hybrid method for prediction and repositioning of drug Anatomical

Therapeutic Chemical classes," *Molecular BioSystems*, vol. 10, no. 4, pp. 868–877, 2014.

[14] L. Chen, W.-M. Zeng, Y.-D. Cai, and T. Huang, "Prediction of metabolic pathway using graph property, chemical functional group and chemical structural set," *Current Bioinformatics*, vol. 8, no. 2, pp. 200–207, 2013.

[15] H.-W. Ma and A.-P. Zeng, "The connectivity structure, giant strong component and centrality of metabolic networks," *Bioinformatics*, vol. 19, no. 11, pp. 1423–1430, 2003.

[16] J. M. Dale, L. Popescu, and P. D. Karp, "Machine learning methods for metabolic pathway prediction," *BMC Bioinformatics*, vol. 11, article 15, 2010.

[17] L. Chen, B.-Q. Li, and K.-Y. Feng, "Predicting biological functions of protein complexes using graphic and functional features," *Current Bioinformatics*, vol. 8, no. 5, pp. 545–551, 2013.

[18] Y. F. Gao, L. Chen, Y. D. Cai, K. Y. Feng, T. Huang, and Y. Jiang, "Predicting metabolic pathways of small molecules and enzymes based on interaction information of chemicals and proteins," *PLoS ONE*, vol. 7, no. 9, Article ID e45944, 2012.

[19] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular Systems Biology*, vol. 3, p. 88, 2007.

[20] K. L. Ng, J. S. Ciou, and C. H. Huang, "Prediction of protein functions based on function-function correlation relations," *Computers in Biology and Medicine*, vol. 40, no. 3, pp. 300–305, 2010.

[21] P. Bogdanov and A. K. Singh, "Molecular function prediction using neighborhood features," *IEEE-ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 2, pp. 208–217, 2010.

[22] P. Gao, Q. P. Wang, L. Chen, and T. Huang, "Prediction of human genes' regulatory functions based on proteinprotein interaction network," *Protein and Peptide Letters*, vol. 19, no. 9, pp. 910–916, 2012.

[23] L. J. Jensen, M. Kuhn, M. Stark et al., "STRING 8—a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Research*, vol. 37, no. 1, pp. D412–D416, 2009.

[24] T. H. Gormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Eds., *Introduction to Algorithms*, MIT Press, Cambridge, Mass, USA, 1990.

[25] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 233–240, New York, NY, USA, June 2006.

[26] R. Bunescu, R. Ge, R. J. Kate et al., "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial Intelligence in Medicine*, vol. 33, no. 2, pp. 139–155, 2005.

[27] D. E. Johnson and G. H. I. Wolfgang, "Predicting human safety: screening and computational approaches," *Drug Discovery Today*, vol. 5, no. 10, pp. 445–454, 2000.

[28] B.-Q. Li, B. Niu, L. Chen et al., "Identifying chemicals with potential therapy of HIV based on protein-protein and protein-chemical interaction network," *PLoS ONE*, vol. 8, no. 6, Article ID e65207, 2013.

[29] J. Zhang, M. Jiang, F. Yuan, K. Y. Feng, Y. D. Cai et al., "Identification of age-related macular degeneration related genes by applying shortest path algorithm in protein-protein interaction network," *BioMed Research International*, vol. 2013, Article ID 523415, 8 pages, 2013.

[30] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[31] R. Klein, "Eph/ephrin signaling in morphogenesis, neural development and plasticity," *Current Opinion in Cell Biology*, vol. 16, no. 5, pp. 580–589, 2004.

[32] P. W. Janes, S. Adikari, and M. Lackmann, "Eph/ephrin signalling and function in oncogenesis: Lessons from embryonic development," *Current Cancer Drug Targets*, vol. 8, no. 6, pp. 473–489, 2008.

[33] E. B. Pasquale, "Eph receptors and ephrins in cancer: bidirectional signalling and beyond," *Nature Reviews Cancer*, vol. 10, no. 3, pp. 165–180, 2010.

[34] L. Ding, G. Getz, D. A. Wheeler et al., "Somatic mutations affect key pathways in lung adenocarcinoma," *Nature*, vol. 455, no. 7216, pp. 1069–1075, 2008.

[35] D. Zagzag, D. R. Friedlander, B. Margolis et al., "Molecular events implicated in brain tumor angiogenesis and invasion," *Pediatric Neurosurgery*, vol. 33, no. 1, pp. 49–55, 2000.

[36] P. Pu, Z. Zhang, C. Kang et al., "Downregulation of Wnt2 and β-catenin by siRNA suppresses malignant glioma cell growth," *Cancer Gene Therapy*, vol. 16, no. 4, pp. 351–361, 2009.

[37] P. Knizetova, J. Ehrmann, A. Hlobilkova et al., "Autocrine regulation of glioblastoma cell cycle progression, viability and radioresistance through the VEGF-VEGFR2 (KDR) interplay," *Cell Cycle*, vol. 7, no. 16, pp. 2553–2561, 2008.

[38] Y. T. Gu, Y. X. Xue, X. Y. Wei, H. Zhang, and Y. Li, "Calcium-activated potassium channel activator down-regulated the expression of tight junction protein in brain tumor model in rats," *Neuroscience Letters*, vol. 493, no. 3, pp. 140–144, 2011.

[39] H. Xie, Y. X. Xue, L. B. Liu, and Y. H. Liu, "Endothelial-monocyte-activating polypeptide II increases blood-tumor barrier permeability by down-regulating the expression levels of tight junction associated proteins," *Brain Research*, vol. 1319, pp. 13–20, 2010.

[40] T. Nikuseva-Martic, V. Beros, N. Pecina-Slaus, H. I. Pecina, and F. Bulic-Jakus, "Genetic changes of CDH1, APC, and CTNNB1 found in human brain tumors," *Pathology: Research and Practice*, vol. 203, pp. 779–787, 2007.

[41] C. Perego, C. Vanoni, S. Massari et al., "Invasive behaviour of glioblastoma cell lines is associated with altered organisation of the cadherin-catenin adhesion system," *Journal of Cell Science*, vol. 115, no. 16, pp. 3331–3340, 2002.

[42] The Cancer Genome Atlas Research Network, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061–1068, 2008.

[43] S. A. Rao, A. Arimappamagan, P. Pandey et al., "miR-219-5p inhibits receptor tyrosine kinase pathway by targeting EGFR in glioblastoma," *PLoS ONE*, vol. 8, no. 5, Article ID e63164, 2013.

[44] K. H. Plate, G. Breier, H. A. Weich, and W. Risau, "Vascular endothelial growth factor is a potential tumour angiogenssis factor in human gliomas in vivo," *Nature*, vol. 359, no. 6398, pp. 845–848, 1992.

[45] B. Millauer, L. K. Shawver, K. H. Plate, W. Risau, and A. Ullrich, "Glioblastoma growth inhibited in vivo by a dominant-negative Flk-1 mutant," *Nature*, vol. 367, no. 6463, pp. 576–579, 1994.

[46] S. Koch, S. Tugues, X. Li, L. Gualandi, and L. Claesson-Welsh, "Signal transduction by vascular endothelial growth factor receptors," *The Biochemical Journal*, vol. 437, no. 2, pp. 169–183, 2011.

[47] M. J. Karkkainen and T. V. Petrova, "Vascular endothelial growth factor receptors in the regulation of angiogenesis and lymphangiogenesis," *Oncogene*, vol. 19, no. 49, pp. 5598–5605, 2000.

[48] B. Jenny, J. A. Harrison, D. Baetens et al., "Expression and localization of VEGF-C and VEGFR-3 in glioblastomas and haemangioblastomas," *Journal of Pathology*, vol. 209, no. 1, pp. 34–43, 2006.

[49] J. N. Anastas and R. T. Moon, "WNT signalling pathways as therapeutic targets in cancer," *Nature Reviews Cancer*, vol. 13, no. 1, pp. 11–26, 2013.

[50] H. Yano, A. Hara, J. Shinoda et al., "Immunohistochemical analysis of β-catenin in N-ethyl-N-nitrosourea- induced rat gliomas: implications in regulation of angiogenesis," *Neurological Research*, vol. 22, no. 5, pp. 527–532, 2000.

[51] X. Liu, L. Wang, S. Zhao, X. Ji, Y. Luo, and F. Ling, "β-catenin overexpression in malignant glioma and its role in proliferation and apoptosis in glioblastma cells," *Medical Oncology*, vol. 28, no. 2, pp. 608–614, 2011.

[52] P. Polakis, "Wnt signaling and cancer," *Genes & Development*, vol. 14, no. 15, pp. 1837–1851, 2000.

[53] Y. Liu, W. Yan, W. Zhang et al., "MiR-218 reverses high invasiveness of glioblastoma cells by targeting the oncogenic transcription factor LEF1," *Oncology Reports*, vol. 28, no. 3, pp. 1013–1021, 2012.

[54] S. Nada, T. Yagi, H. Takeda et al., "Constitutive activation of Src family kinases in mouse embryos that lack Csk," *Cell*, vol. 73, no. 6, pp. 1125–1135, 1993.

[55] S. Zhang, W. C. Huang, L. Zhang, C. Zhang, F. J. Lowery et al., "SRC family kinases as novel therapeutic targets to treat breast cancer brain metastases," *Cancer Research*, vol. 73, pp. 5764–5774, 2013.

[56] J. Chen, A. Elfiky, M. Han, C. Chen, and M. W. Saif, "The Role of Src in colon cancer and its therapeutic implications," *Clinical Colorectal Cancer*, vol. 13, pp. 5–13, 2014.

[57] C. Egan, A. Pang, D. Durda, H. Cheng, J. H. Wang, and D. J. Fujita, "Activation of Src in human breast tumor cell lines: elevated levels of phosphotyrosine phosphatase activity that preferentially recognizes the Src carboxy terminal negative regulatory tyrosine 530," *Oncogene*, vol. 18, no. 5, pp. 1227–1237, 1999.

[58] T. J. Yeatman, "A renaissance for SRC," *Nature Reviews Cancer*, vol. 4, no. 6, pp. 470–480, 2004.

[59] M. R. Stettner, W. Wang, L. B. Nabors et al., "Lyn kinase activity is the predominant cellular Src kinase activity in glioblastoma tumor cells," *Cancer Research*, vol. 65, no. 13, pp. 5535–5543, 2005.

[60] C. V. Lund, M. T. Nguyen, G. C. Owens et al., "Reduced glioma infiltration in Src-deficient mice," *Journal of Neuro-Oncology*, vol. 78, no. 1, pp. 19–29, 2006.

[61] M. S. Ahluwalia, J. de Groot, W. M. Liu, and C. L. Gladson, "Targeting SRC in glioblastoma tumors and brain metastases: rationale and preclinical studies," *Cancer Letters*, vol. 298, no. 2, pp. 139–149, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 345605, 12 pages
http://dx.doi.org/10.1155/2014/345605

Research Article

An Integrated Analysis of miRNA, lncRNA, and mRNA Expression Profiles

Li Guo, Yang Zhao, Sheng Yang, Hui Zhang, and Feng Chen

Department of Epidemiology and Biostatistics, School of Public Health, Nanjing Medical University, Nanjing 211166, China

Correspondence should be addressed to Li Guo; gl8008@163.com and Feng Chen; fengchen@njmu.edu.cn

Received 7 March 2014; Revised 24 April 2014; Accepted 25 April 2014; Published 18 June 2014

Academic Editor: Jiangning Song

Increasing amounts of evidence indicate that noncoding RNAs (ncRNAs) have important roles in various biological processes. Here, miRNA, lncRNA, and mRNA expression profiles were analyzed in human HepG2 and L02 cells using high-throughput technologies. An integrative method was developed to identify possible functional relationships between different RNA molecules. The dominant deregulated miRNAs were prone to be downregulated in tumor cells, and the most abnormal mRNAs and lncRNAs were always upregulated. However, the genome-wide analysis of differentially expressed RNA species did not show significant bias between up- and downregulated populations. miRNA-mRNA interaction was performed based on their regulatory relationships, and miRNA-lncRNA and mRNA-lncRNA interactions were thoroughly surveyed and identified based on their locational distributions and sequence correlations. Aberrantly expressed miRNAs were further analyzed based on their multiple isomiRs. IsomiR repertoires and expression patterns were varied across miRNA loci. Several specific miRNA loci showed differences between tumor and normal cells, especially with respect to abnormally expressed miRNA species. These findings suggest that isomiR repertoires and expression patterns might contribute to tumorigenesis through different biological roles. Systematic and integrative analysis of different RNA molecules with potential cross-talk may make great contributions to the unveiling of the complex mechanisms underlying tumorigenesis.

1. Introduction

Large-scale, genome-wide analyses have indicated that much of the human genome is transcribed, yielding a great many nonexonic transcripts [1, 2]. These nonribosomal and non-mitochondrial RNAs, which are metaphorically considered ribosomal dark matter, are quite abundant in cells. The transcription profile of the entire genome at a specific space and time can be obtained using microarray and sequencing technologies [3]. Noncoding RNAs (ncRNAs), including microRNAs (miRNAs), and long noncoding RNAs (lncRNAs) can be obtained to attract considerable attention of researchers in many fields.

miRNAs, a class of small ncRNAs (≈22 nt), are highly important regulatory molecules, and they have seen a great deal of study [4, 5]. Posttranscriptional gene regulation via miRNA is crucial to the regulation of gene expression. These small, single-stranded RNAs negatively regulate gene expression through partial base-pairing with target messenger RNAs (mRNAs). This influences the process of mRNA degradation or repression of translation [4, 6]. They have multiple roles in various biological processes that affect basic cellular functions, including cell proliferation, differentiation, death, and tumorigenesis [7]. Abnormal expression of specific miRNAs has been characterized as a common feature of human diseases, especially for malignancies. In these cases, genes encoding miRNAs may act as oncogenes, oncomiRs, or tumor suppressors [7–9]. Widely concerned lncRNAs are normally longer than 200 nucleotides. Studies have shown them to be involved in a broad range of important cellular processes, including chromatin modification, RNA processing, and gene transcription and that they do so through interaction with DNA and proteins [10–14]. LncRNAs are characterized as complex, diverse ncRNAs. They are usually involved in exons and introns and have 5′ cap and some of the features of mRNAs [15]. The larger ncRNAs have been shown

to regulate gene expression through various mechanisms, such as complementary binding to protein-coding transcripts in the form of cis-antisense lncRNAs [16–20]. They also modulate transcription factors by acting as coregulators [19, 21–25]. Dysregulation of ncRNAs contributes to many biological processes by interfering with gene expression. In recent years, this has become a hot research topic as core regulatory molecules.

Although the many biological roles of ncRNAs have drawn a great deal of concern, systematic and integrative analyses of many kinds of RNA molecules (including functional mRNAs) have been rare. An integrated, genome-wide analysis involving many different RNA molecule levels is necessary if the complex regulatory network and mechanisms underlying tumorigenesis are to be understood. We ever performed analyses of miRNA-mRNA and miRNA-miRNA interactions using miRNA and mRNA expression profiles [26], but it is not enough to further understand the potential relationships between different RNA molecules, especially involving the novel concerned lncRNAs. In the present study, the close relationships between ncRNAs and mRNAs were examined through simultaneously profiling of miRNA, lncRNA, and mRNA in HepG2 and L02 cells using high-throughput technologies. An integrative method of analysis was developed to detect and comprehensively analyze the relationships between RNA molecules, especially between abnormally expressed miRNA, lncRNA, and mRNA molecules in tumor cells. A systematic analysis of miRNAs was also performed at the isomiR level. The results of the present study will enrich the genome-wide analysis of different molecules with potential cross-talk and contribute to further systematic studies of tumorigenesis.

2. Materials and Methods

2.1. Cell Culture and RNA Isolation. HepG2 and L02 cells were obtained from the American Type Tissue Collection. These were maintained in DMEM containing 10% FBS, 100 U/mL benzylpenicillin and 100 U/mL streptomycin at 37°C in a humidified 95% air 5% CO_2 incubator. Total RNAs were isolated using TRIzol reagent (Invitrogen) according to the manufacturer's protocol.

2.2. Small RNA Sequencing and Microarray Experiments. Total RNA from each sample was used to prepare the small RNA sequencing library to perform sequencing on a Genome Analyzer IIx, which was used in accordance with the manufacturer's instructions, and was prepared for microarray hybridization. The raw small RNA sequencing data can be available in the Sequence Read Archive (SRA) database (http://www.ncbi.nlm.nih.gov/sra, accession number SRA 1262121), and the mRNA and lncRNA microarray data can be available in the European Molecular Biology L aboratory-Euro-pean Bioinformatics Institute (EMBL-EBI) database (http://www.ebi.ac.uk/).

2.3. Data Analysis. The total raw miRNA sequencing reads were first filtered using a Solexa CHASTITY quality control filter. The remaining sequencing reads were deleted the 3′ adapter sequences, and tags shorter than 15 nt were discarded. Then, the reads were aligned to the known human miRNA precursors (pre-miRNAs) in the miRBase database (Release 18.0, http://www.mirbase.org/) using Novoalign software (v2.07.11, http://www.novocraft.com/) [27]. Only one mismatch was allowed. Reads with counts under 2 were discarded when miRNA expression was calculated. According to recent reports on multiple isomiRs from a given miRNA locus [28–34], isomiRs (including those isomiRs with 3′ nontemplate additional nucleotides) were also comprehensively surveyed. Sequences that matched the pre-miRNAs in the mature miRNA region ±4 nt (no more than 1 mismatch) were defined as isomiRs. The original sequence counts of miRNAs were normalized to RPM (reads per million), and miRNA expression analysis was performed based on these normalized data at the miRNA and isomiR level.

Images from microarray were analyzed with Agilent Feature Extraction software (version 10.7.3.1). Raw signal intensities of mRNAs and lncRNAs were normalized using the quantile method and the GeneSpring GX v12.0 software package (Agilent Technologies). After quantile normalization of the raw data, lncRNAs and mRNAs for which 2 out of 2 samples had flags in the present or marginal were chosen for further data analysis.

Fold change was calculated to assess expressed miRNA profiles that were differentially expressed between the two samples at miRNA and isomiR levels. Differentially expressed lncRNAs and mRNAs were also identified through fold change filtering. To obtain abnormal ncRNA/mRNA species and filter out rare species with lower expression levels, fold change values were assessed by adding an additional low number (10 units) based on normalized datasets. Hierarchical clustering was performed using Cluster bb3.0 and TreeView 1.60 programs (http://rana.lbl.gov/eisen/) [35, 36]. Experimentally validated target mRNAs of aberrantly expressed miRNAs were collected from the miRTarBase database [37]. For miRNAs with few or no validated target mRNAs, the putative target mRNAs were integrated using the prediction software programs Pictar, TargetScan, and miRanda programs [38]. The threshold values were simultaneously controlled (e.g., in TargetScan, the threshold of total context score was less −0.30). The collected target mRNAs were further screened based on abnormally expressed mRNA profiles. Then, pathway and GO analysis were used to determine the roles of these differentially expressed mRNAs. Using CapitalBio Molecule Annotation System V4.0 (MAS, http://bioinfo.capitalbio.com/mas3/), further functional enrichment analysis was performed. Functional interaction networks were constructed using Cytoscape v2.8.2 Platform [39].

2.4. Schema for Integrative Analysis of ncRNA-mRNA Data. ncRNA-mRNA integrative analysis was performed according to Figure 1. The approach included three steps. First, profiles of aberrantly expressed miRNA, mRNA, and lncRNA in HepG2 cells were comprehensively surveyed using high-throughput datasets. A profile of aberrantly expressed

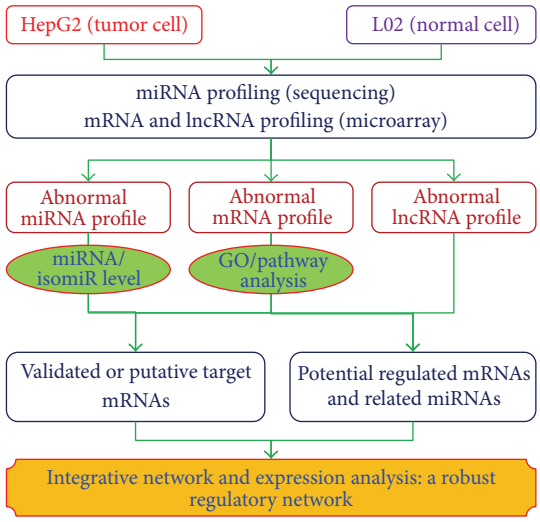


Figure 1: Integrative analysis of ncRNA-mRNA.

isomiRs was obtained at the same time. Pathway and GO analyses were performed for abnormal mRNA and the target mRNAs of abnormal miRNAs. Second, systematic bioinformatic analysis was developed based on possible functional relationships between these molecules. miRNA and mRNA were analyzed in an integrated fashion based on experimentally validated or predicted target mRNAs and on their levels of enrichment. Possible internal relationships among lncRNA-mRNA and lncRNA-miRNA were identified based on their locational distributions and the relationships between their sequences. Finally, integrative regulatory network and expression analyses were performed at different molecular levels based on their possible levels of expression and functional relationships (Figure 1).

3. Results

3.1. Aberrantly Expressed miRNA and isomiR Profiles. As expected, 22 nt was the most common length (see Figure S1A and Figure S1B in Supplementary Material available online at http://dx.doi.org/10.1155/2014/345605). As found by Guo et al., consistent aberrantly expressed miRNAs were obtained based on the most abundant isomiR and all the isomiRs, respectively [34]. However, despite the consistency of the dysregulation pattern, the fold change (log 2) of some miRNAs was found to cover a wide range (miR-200b-3p: 8.56 and 5.45, miR-100-5p: −7.27 and −4.88) (Table 1). Half of the most dominant isomiR sequences had lengths that were different from those of canonical miRNA sequences (data not shown). Generally, the most dominant isomiR sequence may be longer or shorter than registered miRNA sequence through altering 5′ and 3′ ends, especially for the 3′ ends (Figure S1C). These abundantly and abnormally expressed miRNAs were always located on a few specific chromosomes, especially chromosome 9 (Figure S1D). The distribution bias was obvious, even though multicopy pre-miRNAs were also analyzed. Downregulated miRNAs were found to be more common than upregulated species, although the total numbers remained similar across the whole abnormal miRNA profiles.

IsomiR repertoires and expression profiles in the HepG2 and L02 cells were also analyzed. Various isomiR repertoires and expression patterns were detected in different miRNA loci (Figure 2). As found by Guo et al., several dominant isomiRs (always 1–3) were yielded per miRNA locus due to alternative and imprecise cleavage of Drosha and Dicer [34, 40]. Deregulated miRNAs might show abnormal isomiR expression profiles in tumor cells, such as miR-194-5p (upregulated) and miR-24-3p (downregulated) (Figure 2). These findings indicated inconsistent dominant isomiR sequences, even though they were always 5′ isomiRs with the same 5′ ends and seed sequences. This phenomenon was detected primarily in deregulated miRNAs. Generally, those stably expressed species had similar expression profiles between tumor and normal cells. This was true of miR-26a-5p and miR-21-5p (Figure 2). Of the dominant isomiRs, only miR-15a-5p was involved in 3′ addition (Figure 2). Although 3′ addition was quite widespread, especially for adenine and uracil, the presence of type of isomiRs with 3′ additions suggested considerable divergence between miRNAs. For example, miR-103a-3p was not detected any modified isomiRs even though 10 isomiRs were obtained, while three isomiRs with 3′ additions were found in miR-194-5p (Figure 2). At the miRNA locus, these modified isomiRs always possessed lower enrichment levels than dominant isomiR sequences, although they might still show high levels of expression.

Functional enrichment analysis was performed based on targets that had been found to be regulated by at least 2 abnormal miRNAs. The results suggested that these miRNAs play important roles in essential biological processes, including the cell cycle, Wnt, and the MAPK signaling pathway (Table 2). They also contribute to various human diseases, such as chronic myeloid leukemia, prostate cancer, and bladder cancer. According to identified mRNA profiles by using microarray technology, these target mRNAs might be stably expressed, upregulated, or downregulated in tumor cells.

3.2. Aberrantly Expressed mRNA and lncRNA Profiles. Dominant (>10 of normalized data) and significantly differentially expressed (fold change (log 2) >4.0 or <−4.0) mRNAs and lncRNAs (the top deregulated species could be found in Table 3) were collected. Significant divergence was detected between upregulated and downregulated RNA species. 83.98% of deregulated mRNAs and 90.93% deregulated lncRNAs were upregulated in tumor cells. However, the analysis of differentially expressed profiles suggested that 62.06% of mRNAs and 66.49% of lncRNAs were downregulated. Locational distributions of mRNA and lncRNA were analyzed. Consistent distribution patterns were detected, and no bias was found between deregulated mRNA and lncRNA species (Figures S2A, S2B, and S2C). However, inconsistent distributions were detected between the top 100 dominant and deregulated mRNAs and lncRNAs (Figures S2D and S2E). These abnormal species were prone to locate

Table 1: Differentially expressed abundant miRNA species as indicated by the most abundant isomiR and all isomiRs.

miRNA	Chr	Consistent or inconsistent	Fold change (the most)	Fold change (all isomiRs)	Up/down
let-7a-5p	9, 11, 22	Yes	−2.35	−3.24	Down
let-7f-5p	9, X	Yes	−2.54	−2.97	Down
miR-103a-3p	5, 20	Yes	2.33	2.05	Up
miR-146a-5p	5	Yes	7.28	5.52	Up
miR-15a-5p	13	No	−4.85	−3.87	Down
miR-194-5p	1, 11	No	7.26	4.56	Up
miR-200b-3p	1	No	8.56	5.45	Up
miR-23a-3p	19	No	−7.19	−5.02	Down
miR-24-3p	9, 19	No	−4.24	−2.14	Down
miR-27a-3p	19	Yes	−6.61	−6.35	Down
miR-27b-3p	9	Yes	−1.82	−2.61	Down
miR-100-5p	11	Yes	−7.27	−4.88	Down
miR-425-5p	3	No	2.98	2.00	Up

These miRNAs are abundantly expressed in HepG2 and L02 cells. They are the top downregulated and upregulated miRNAs in cancer cells (fold change (log 2) >2.0 or <−2.0). Chr indicates the genomic locations of the miRNA genes (pre-miRNAs), including multicopy pre-miRNAs. let-7a-5p is located on chr9 (let-7a-1), 11 (let-7a-2), and 22 (let-7a-3). The term "consistent" indicates that the sequence of the most abundant isomiR is the same as that of the reference miRNA sequence in the miRBase database. The term "most" indicates the most abundant isomiR from a given locus. The term "all isomiRs" indicates total number of isomiRs from a given locus.

Table 2: Pathway enrichment analysis of experimentally validated mRNA targets of dominant deregulated miRNAs.

Pathway	Number	P value	Target genes
Cell cycle	18	$3.01E-30$	*ATM; **CCNA2**; CCND1; **CCND2**; CCNE1; CDC25A; CDK6; CDKN1A; CDKN1B; CDKN2A; E2F1; **E2F2**; E2F3; EP300; RB1; **RBL2**; TP53; WEE1*
Chronic myeloid leukemia	15	$2.74E-27$	*ACVR1C; AKT1; CCND1; CDK6; CDKN1A; CDKN1B; CDKN2A; E2F1; **E2F2**; E2F3; MYC; NFKB1; **NRAS**; RB1; TP53*
Prostate cancer	15	$3.74E-26$	*AKT1; BCL2; CCND1; CCNE1; CDKN1A; CDKN1B; E2F1; **E2F2**; E2F3; EP300; IGF1R; NFKB1; **NRAS**; RB1; TP53*
Pancreatic cancer	14	$3.03E-25$	*ACVR1C; AKT1; CCND1; **CDC42**; CDK6; CDKN2A; E2F1; **E2F2**; E2F3; NFKB1; RAC1; RB1; TP53; VEGFA*
Bladder cancer	13	$1.47E-26$	*CCND1; CDKN1A; CDKN2A; E2F1; **E2F2**; E2F3; FGFR3; MYC; **NRAS**; RB1; THBS1; TP53; VEGFA*
Melanoma	13	$3.21E-23$	*AKT1; CCND1; CDK6; CDKN1A; CDKN2A; E2F1; **E2F2**; E2F3; IGF1R; **MET**; **NRAS**; RB1; TP53*
Melanoma	13	$3.21E-23$	*AKT1; CCND1; CDK6; CDKN1A; CDKN2A; E2F1; **E2F2**; E2F3; IGF1R; **MET**; **NRAS**; RB1; TP53*
Small-cell lung cancer	13	$4.71E-22$	*AKT1; BCL2; CCND1; CCNE1; CDK6; CDKN1B; E2F1; **E2F2**; E2F3; MYC; NFKB1; RB1; TP53*
Focal adhesion	13	$5.65E-17$	*AKT1; BCL2; CCND1; **CCND2**; **CDC42**; PAK3; IGF1R; **MET**; RAC1; RHOA; ROCK1; THBS1; VEGFA*
Glioma	12	$1.49E-21$	*AKT1; CCND1; CDK6; CDKN1A; CDKN2A; E2F1; **E2F2**; E2F3; IGF1R; **NRAS**; RB1; TP53*
Nonsmall cell lung cancer	10	$3.77E-18$	*AKT1; CCND1; CDK6; CDKN2A; E2F1; **E2F2**; E2F3; **NRAS**; RB1; TP53*
Renal cell carcinoma	10	$7.11E-17$	*AKT1; **CDC42**; PAK3; EP300; ETS1; **HIF1A**; **MET**; **NRAS**; RAC1; VEGFA*
Axon guidance	10	$3.77E-14$	***CDC42**; PAK3; CXCL12; CXCR4; **MET**; NFAT5; **NRAS**; RAC1; RHOA; ROCK1*
p53 signaling pathway	9	$1.83E-14$	*ATM; CCND1; **CCND2**; CCNE1; CDK6; CDKN1A; CDKN2A; THBS1; TP53*
Colorectal cancer	9	$1.02E-13$	*ACVR1C; AKT1; BCL2; CCND1; IGF1R; **MET**; MYC; RAC1; TP53*
Wnt signaling pathway	9	$1.88E-11$	*CCND1; **CCND2**; EP300; MYC; NFAT5; RAC1; RHOA; ROCK1; TP53*
MAPK signaling pathway	9	$2.68E-09$	*ACVR1C; AKT1; **CDC42**; FGFR3; MYC; NFKB1; **NRAS**; RAC1; TP53*
Adherens junction	8	$4.04E-12$	*ACVR1C; **CDC42**; EP300; IGF1R; **MET**; RAC1; RHOA; WASF3*

These target mRNAs are found to be regulated by at least 2 abnormal miRNAs each. Bold type indicates upregulation. Underlining indicates downregulation. Other fonts indicate stable expression or undetectable levels.

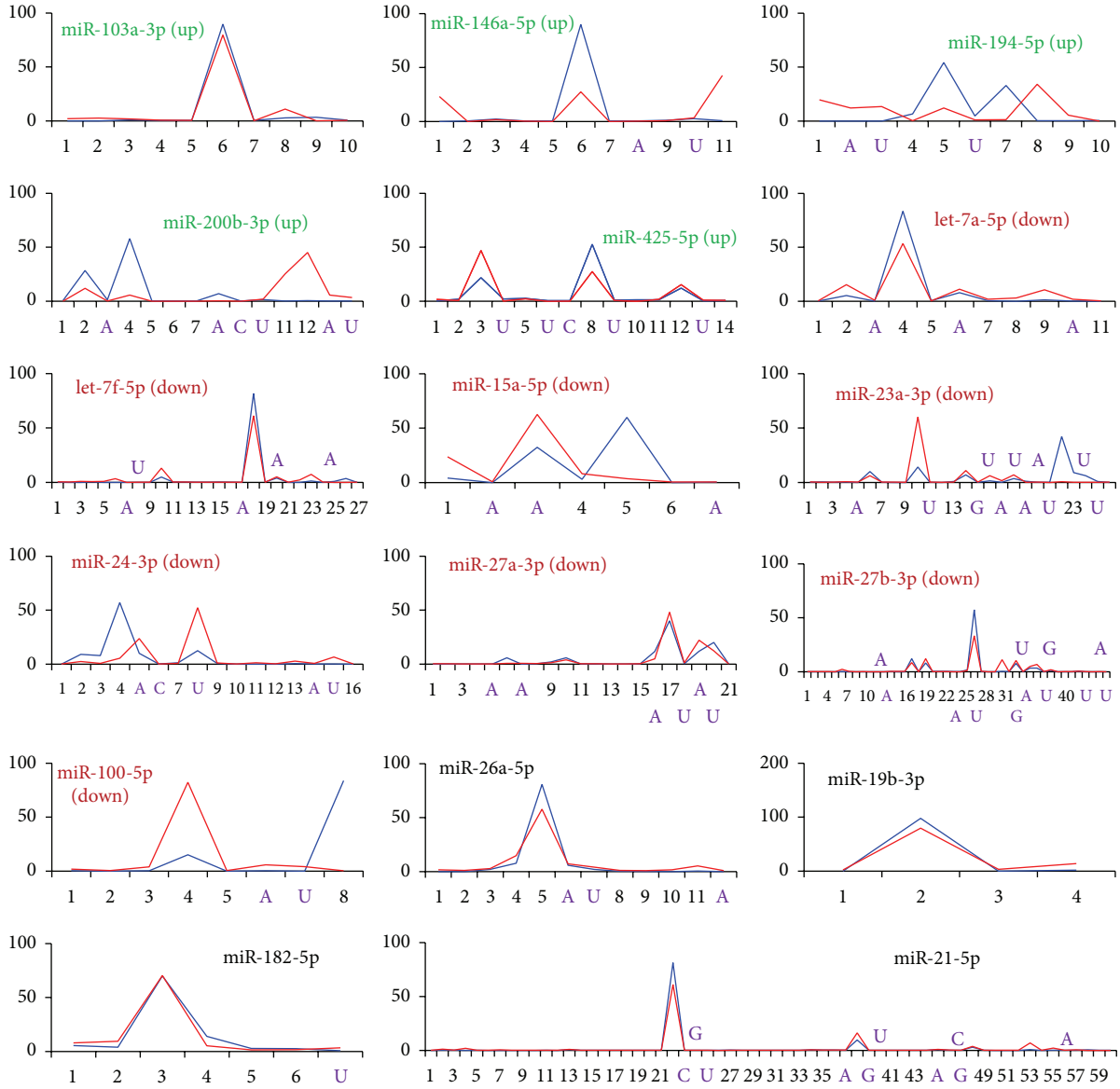


Figure 2: Various isomiR profiles. These miRNAs are the most abundantly up- (green) or downregulated (purple) miRNAs among those examined here. Stably expressed miRNAs are also shown (black). The ordinate axis indicates the relative amount of expression in a specific miRNA locus, and the horizontal axis indicates various types of isomiRs. Blue lines indicate isomiRs from HepG2 cells, and the red line indicates isomiRs from L02 cells. Some miRNAs with 3′ additional nucleotides are also highlighted in the horizontal axis. Stably expressed miRNAs show similar patterns of expression, and deregulated miRNAs show various patterns of expression in those cells. The figure only lists isomiRs (>50 for deregulated miRNAs, >100 for stably expressed miRNAs) for which normalized data was available.

on chromosomes 6 and 12 (mRNA) and chromosomes 4 and 5 (lncRNA).

Significantly deregulated mRNAs and lncRNAs were found to be prone to be located on chromosomes 1 and 2, especially upregulated species (Figures 3(a), 3(b), and 3(c)). Upregulation was found to be more common than downregulation (Figure 3). Generally, no significant distribution bias was found between sense and antisense strands (Figures 3(a) and 3(b)). Coding and noncoding RNAs indicated similar ratios of downregulated species, while they showed diversity of upregulated species (Figure 3(c)). Inconsistent locational distribution patterns were detected based on the total number of down- and upregulated mRNA and lncRNA species (Figure 3(d)).

The pathway and GO analysis of abnormal mRNA expression profiles showed various results (Figure 4 and Figure S3). Downregulated mRNAs were prone to be found in pathways of regulation of actin cytoskeleton and pathways in cancer, and upregulated mRNAs contributed to the biological processes of ribosomes and spliceosomes (Figure 4). Dominant abnormal mRNAs were collected for functional enrichment analysis. Some of them had important roles in

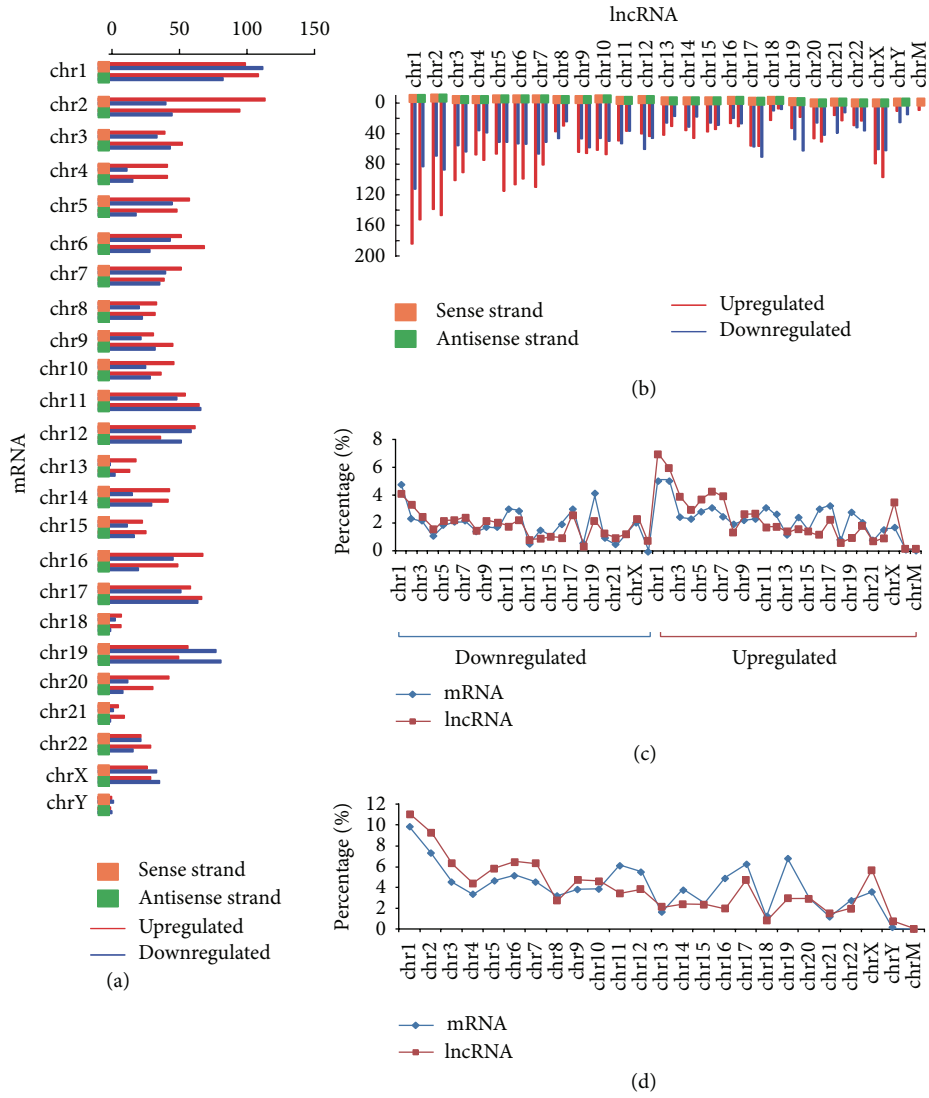


Figure 3: Distribution of aberrantly expressed mRNA and lncRNA profiles. Location distributions of aberrantly expressed (a) mRNA and (b) lncRNA profiles on human chromosomes. The number of deregulated species, including detailed up- and downregulated mRNAs and lncRNAs, is given with respect to sense and antisense strands, respectively. "chr": chromosome; "chrM": mitochondrial chromosome. (c) Distribution of upregulated and downregulated species. (d) Distribution of total deregulated mRNAs and lncRNAs.

diverse essential biological processes through involvement in the pathways, including purine metabolism, ribosomes, the cell cycle, glycolysis, and gluconeogenesis (Table S1). They also contributed to occurrence and development of some human diseases, such as Parkinson's disease and small-cell lung cancer.

3.3. ncRNA-mRNA Data Integration and Interactive Regulators in Tumorigenesis. According to functional enrichment analysis of abnormal miRNAs and mRNAs, the common pathways could be obtained using different genes (Table S2). This was mainly attributable to the selected threshold values of analyzed miRNA and mRNA species. Not all dominant deregulated miRNAs and mRNAs had direct relationships.

An analysis of miRNA-mRNA interactions showed a complex regulatory network (Figure 5). mRNAs that were regulated by at least 2 abnormally expressed miRNAs were collected. Generally, they were prone to form closed networks with close regulatory relationships. Some miRNAs, such as let-7a-5p and miR-15a-5p, were located in the central positions with multiple target mRNAs. Although small regulatory molecules were downregulated or upregulated, their target mRNAs might show consistent or inconsistent dysregulation patterns (Figure 5). Locational relationships indicated that related lncRNAs were also constructed in the regulatory network. Some miRNAs, such as miR-24-3p (mir-24-2 gene is located in BX640708), always showed consistent deregulation patterns with their host lncRNAs (Figure 5). Several mRNAs were also found to be related to nearby lncRNAs. mRNA and associated lncRNA might be located on the same strand or have a sense/antisense relationship within a specific genomic region. mRNA-lncRNA might

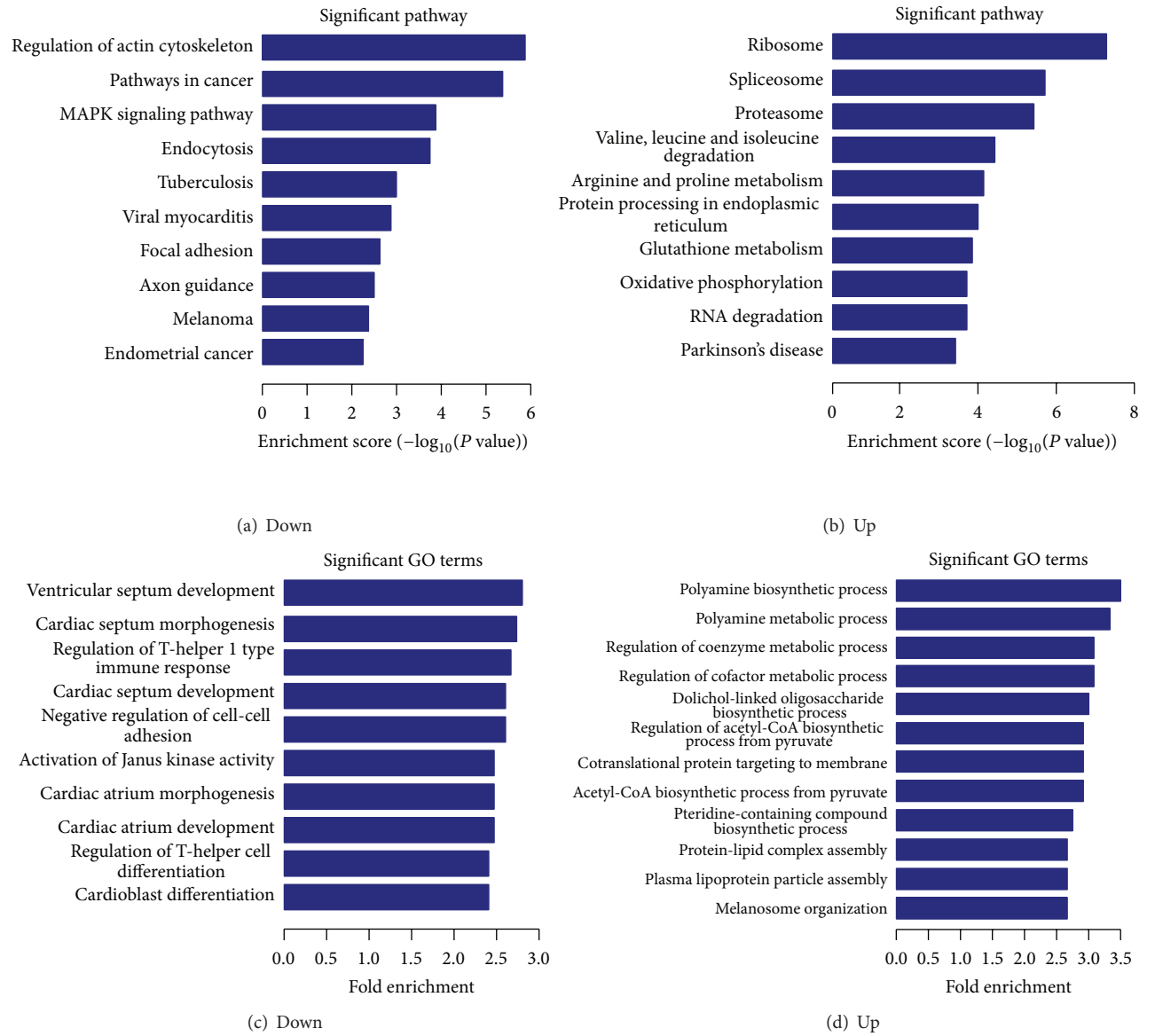


Figure 4: Analysis of significant pathways and GO terms regarding biological processes associated with abnormally expressed mRNA profiles in HepG2 cells. The *P* value denotes the significance of the pathway correlated to the conditions and GO term (the recommend *P* value cutoff is 0.05).

show consistent (APP & AP001439.2) or inconsistent (E2F2 & AL021154.3) deregulation patterns (Figure 5).

To identify the overall patterns of deregulation between mRNAs/miRNAs and lncRNAs, a comprehensive survey of their potential relationships was performed incorporating information regarding mRNAs, miRNAs, and lncRNAs. Most mRNA-lncRNA pairs had sense/antisense relationships, and miRNA-lncRNA pairs were prone to be located on the same strands (Table S3). Generally, these mRNA/miRNA-lncRNA pairs could completely or partially overlap (from the same strands) or show reverse complementarily binding (from sense/antisense strands). The mRNA and lncRNA could show the same or different deregulation patterns, but they were usually the same (Figure 6). Some pairs were up- or downregulated, and their fold change values differed (Figures 6(a) and 6(b)). These RNA molecules, both coding RNAs, which are functional molecules, and noncoding RNAs, which are regulatory molecules, were prone to be downregulated in tumor cells.

4. Discussion

Although the recorded values of differences in expression, as defined as the abundance of isomiRs, expression levels of isomiRs may have been influenced by higher sensitivity of next-generation sequencing technology, the diversity of those expression is mainly attributable to differences in the isomiR profiles and expression patterns in normal and tumor

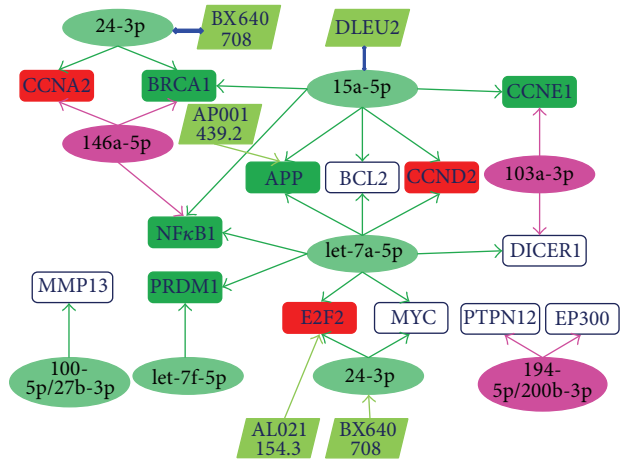


Figure 5: Interaction between deregulated ncRNA-mRNA in HepG2 cells. Each ellipse indicates up- (purple) or downregulated (light green) miRNA. Each square indicates up- (red) or downregulated (green) mRNA (stably expressed mRNAs are in white), and each tilted rectangle indicates downregulated lncRNAs (pale green). These miRNAs are shown in Table 1. A regulatory network was constructed using experimentally validated target mRNAs (each mRNA is regulated by at least 2 deregulated miRNAs). Stably expressed mRNAs are shown in white squares (*BCL2*, *MMP13*, and *PTPN12*). mRNAs not detected in HepG2 and L02 cells are also shown in white squares (*DICER1*, *EP300*, and *MYC*).

Table 3: Differentially expressed abundant mRNA and lncRNA species.

mRNA/lncRNA	Gene symbol	Chr (±)	HepG2 (Nor)	L02 (Nor)	Fold change	Up/down
mRNA	RBP4	chr10 (−)	16.41	4.90	11.51	Up
mRNA	APOA1	chr11 (−)	15.07	5.44	9.64	Up
mRNA	ALB	chr4 (+)	14.19	4.65	9.54	Up
mRNA	ID2	chr2 (+)	13.52	4.16	9.35	Up
mRNA	TFPI	chr2 (−)	13.26	4.29	8.97	Up
mRNA	SERPINA3	chr14 (+)	14.15	5.42	8.73	Up
mRNA	SRGN	chr10 (+)	5.26	13.59	−8.33	Down
mRNA	CD81	chr11 (+)	4.66	12.99	−8.33	Down
mRNA	FOLR1	chr11 (+)	5.00	14.50	−9.50	Down
mRNA	NNMT	chr11 (+)	3.74	13.28	−9.54	Down
mRNA	C11orf86	chr11 (+)	4.26	15.16	−10.90	Down
mRNA	BASP1	chr5 (+)	5.53	17.06	−11.53	Down
lncRNA	RP11-113C12.1	chr12 (−)	12.71	5.43	7.28	Up
lncRNA	D28359	chr13 (+)	13.95	6.63	7.31	Up
lncRNA	MGC12916	chr17 (+)	11.25	3.91	7.33	Up
lncRNA	ABCC6P1	chr16 (+)	12.28	4.89	7.39	Up
lncRNA	lincRNA-HEY1	chr8 (−)	12.12	4.70	7.42	Up
lncRNA	HSPEP1	chr20 (−)	13.22	5.60	7.61	Up
lncRNA	TMSL6	chr20 (−)	8.25	16.08	−7.83	Down
lncRNA	RP11-163G10.3	chr1 (−)	8.18	15.92	−7.74	Down
lncRNA	AC010907.3	chr2 (−)	7.44	15.07	−7.64	Down
lncRNA	BC106081	chr8 (−)	3.70	10.81	−7.11	Down
lncRNA	nc-HOXA11-86	chr7 (+)	3.83	10.65	−6.82	Down
lncRNA	AK054970	chr13 (+)	6.01	12.64	−6.62	Down

The table only lists the top 6 up- and downregulated mRNAs and lncRNAs based on the fold change values (log 2). These mRNAs and lncRNAs are dominantly expressed. "Chr (±)" indicates genomic location on sense or antisense strands of human chromosomes. "Nor" indicates the normalized data.

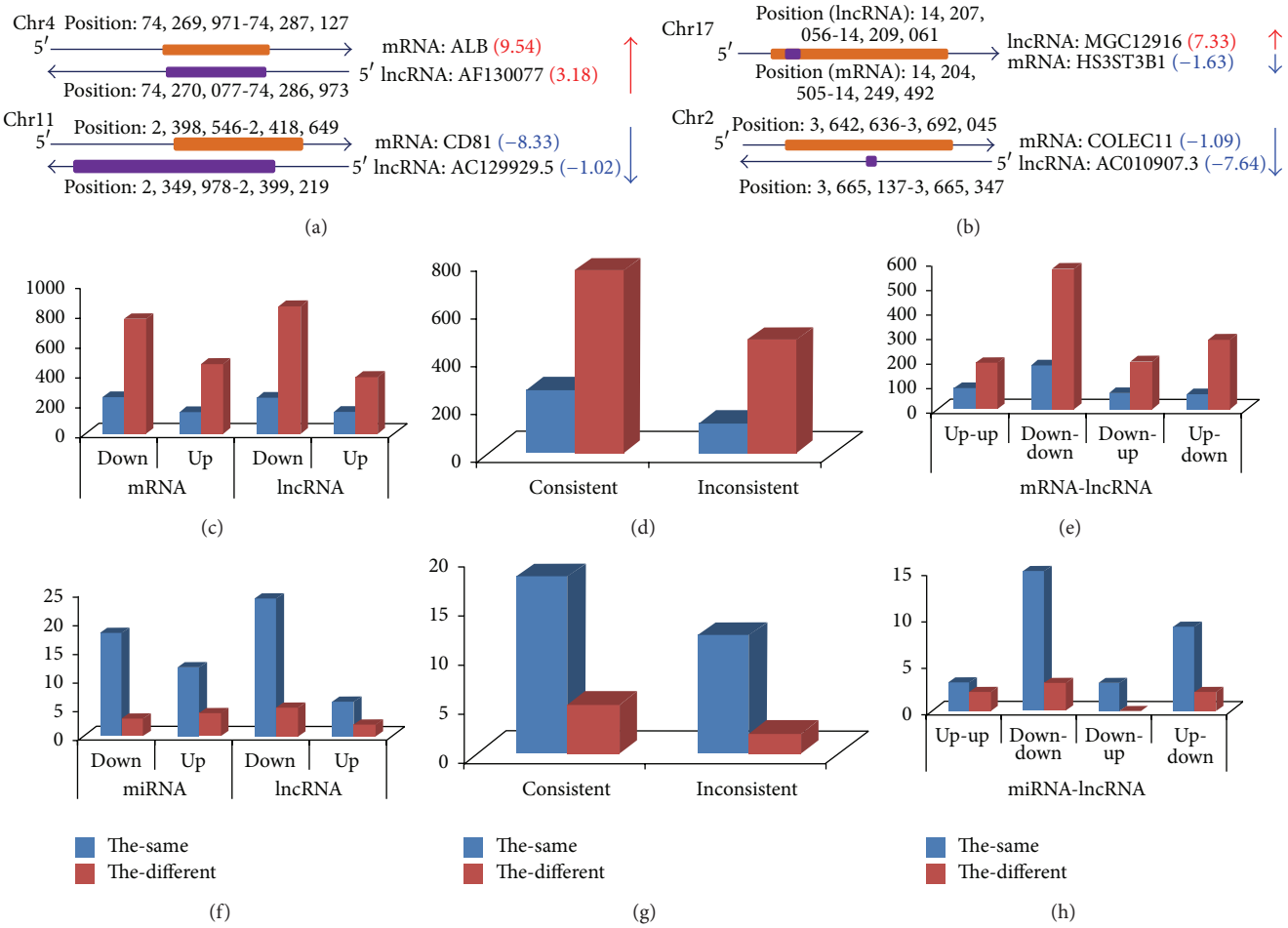


FIGURE 6: Integrative expression analysis of mRNA-lncRNA and miRNA-lncRNA based on their locations. ((a)-(b)) Schematic representation of expression of mRNA-lncRNA (Table 3). Some of these mRNA genes and lncRNA genes are always located on sense/antisense strands in specific genomic regions. Others are located in the same genomic region but are of different lengths. Some show the same deregulation patterns with different fold changes (log 2) (upregulation: red, downregulation: blue), and others show different deregulation patterns. ((c)–(e)) mRNA-lncRNA integrative analysis based on genomic location. The ordinate axis indicates the number of deregulated mRNAs or lncRNAs. The term "the-same" indicates that mRNA and lncRNA are located on the same strand. The term "the-different" indicates that mRNA and lncRNA are located on sense/antisense strands. ((f)–(h)) miRNA-lncRNA integrative analysis based on genomic location. The ordinate axis indicates the number of deregulated miRNAs or lncRNAs. The term "the-same" indicates that miRNA and lncRNA are located on the same strand. The term "the-different" indicates that miRNA and lncRNA are located on sense/antisense strands.

cells (Figure S4). Indeed, the most dominant isomiRs are not always canonical miRNA sequences. The wide range of these inconsistent sequences was found to contribute to various isomiR repertoires, leading to differences in expression between the most abundant isomiR and other isomiRs. Herein, 3′ addition is detected in many places, but it is not always present in the dominant sequences, though it may show high level of expression (Figure 2 and Figure S4). Stably expressed miRNAs always show similar isomiR patterns, and deregulated miRNA species are prone to show different isomiR repertoires (Figure 2) [34]. Generally, isomiR profiles remain stable in different tissues [31, 33, 34]. Deviant isomiR expression profiles should not be considered random events. These results strongly suggest that the isomiR repertoires and their patterns of expression might contribute to tumorigenesis through playing biological roles [34]. Collectively, the detailed isomiR repertoires might serve as markers and provide information regarding the regulatory mechanisms of small noncoding active molecules.

miRNAs are small negative regulatory molecules. They can suppress gene expression via mRNA degradation or repression of translation [4, 6]. However, integrative analysis shows both consistent and inconsistent deregulation patterns, indicating complex regulatory networks containing both noncoding RNAs and mRNAs (Figure 5). mRNAs are always regulated by multiple miRNAs, and vice versa. The dynamic expression patterns between miRNAs and mRNAs are more complex than had been believed. Even though multiple target mRNAs can be detected, miRNA may regulate specific mRNAs at specific times and at specific sites. Competitive interactions between miRNA and mRNA may be dynamic and involve complex regulatory mechanisms

in a specific microenvironment. Dominant selection may exist in miRNA and mRNA. Selective, dynamic, flexible interactions may produce robust coding-noncoding RNA regulatory networks, especially networks involving lncRNAs. The robust regulatory patterns contribute to normal biological processes. Abnormal regulatory networks may produce aberrant pathways and even disease. Specifically, let-7-5p has been experimentally validated as crucial regulatory molecule in hepatocellular carcinoma. They can negative regulate Bcl-xL expression and strengthen sorafenib-induced apoptosis [41], and can contribute to protecting human hepatocytes from oxidant injury through regulating Bach1 [42]. Common pathways can be produced through enrichment analysis of miRNAs and mRNAs (Table S2). Significantly up- and down-regulated mRNAs show various pathways and GO terms (Figure 4 and Figure S3). The large number and variety of biological roles and their positions in pathways and networks indicate that they may contribute to tumorigenesis. Assessing the actual interaction networks can be difficult *in vivo* because of dynamic expression, although high-throughput techniques can be used to track and construct whole-expression profiles.

Both dominant and deregulated miRNAs are prone to be located on specific chromosomes (Figure S1D). The bias might implicate active transcription of specific regions or chromosomes. However, abnormal species do not show distribution biases (Figure 3 and Figure S2). Among dominant aberrantly expressed miRNA, downregulation is quite common. Among mRNAs and lncRNAs, upregulation is more common. All of these findings suggest consistent or coexpression patterns shared by mRNAs and lncRNAs. These are mainly derived from original transcription from genomic DNA sequences. miRNAs and mRNAs/lncRNAs tended to show opposite deregulation patterns. Evidence suggests that miRNA can regulate lncRNA through methylation. For example, miR-29 can regulate the long noncoding gene *MEG3* in hepatocellular cancer through promoter hypermethylation [43]. Moreover, some miRNAs are encoded by exons of long noncoding transcripts [44, 45]. These miRNAs and their host gene lncRNAs may be cotranscribed and coregulated. These would include miR-31 and its host gene, lncRNA *LOC554202*, which, in triple-negative breast cancer, are regulated through promoter hypermethylation [46].

Noncoding RNA molecules, especially miRNAs and lncRNAs, are very prevalent regulatory molecules. They have been shown to play versatile roles in many biological processes. These usually involve transcriptional regulation and modulation of protein function [23]. In the present study, the nearness or separation of ncRNAs and mRNAs on the chromosome is used to perform a comprehensive analysis. Results show that mRNA-lncRNA pairs always have consistent or inconsistent deregulation patterns (Figures 5 and 6 and Table S3). Although some pairs have sense/antisense relationships, the same trends can be detected even at differences in fold change of far greater magnitude (log 2) (Figures 6(a) and 6(b)). The various fold change indicates different degrees of up- or downregulation between mRNAs and lncRNAs. This information might be used to determine the method of regulation. The two members of each mRNA-lncRNA pair can overlap completely or partially (on the same strand). Some of them can also form duplexes through reverse complementary binding (from the sense/antisense strands), which may facilitate interactions between different RNA molecules. The pronounced divergence with respect to the degree of deregulation might be attributable to different regulatory methods, although other complex mechanisms may also be involved. The mRNA and lncRNA from the same strands sometimes show opposite deregulation trends (Figure 6(b)), although the fact that they are cotranscribed from the same genomic DNA sequence with similar original expression levels. Complex negative regulatory networks, especially those involved in noncoding miRNAs and lncRNAs, contribute to the diversity of final relative expression levels *in vivo*. Abnormal regulation in the coding-noncoding RNA network may be pivotal to tumorigenesis.

miRNA-lncRNA pairs with locational relationships are also surveyed. The miRNA and lncRNA in these pairs are more prone to be located on the same strand with complete overlap than paired mRNA and lncRNA molecules are (Figure 6 and Table S3). They may show either consistent or inconsistent deregulation patterns, but consistent patterns are more common (Figure 6). The different levels of final enrichment may be attributable to degradation and regulatory mechanisms. Diversity of abnormal half-life for specific RNA molecules may cause the development of diseases. However, although molecules may have different regulatory relationships, a robust regulatory network can be detected, especially due to multiple targets of each molecule. Alternative regulatory pathways, particularly flexible candidate regulatory pathways and functional pathways, indicate that the coding-noncoding RNA regulatory network is more complex than had been believed, especially in different space and time. The possible flexible relationships between molecules in various places and at various times are crucial to determining the mechanism underlying tumorigenesis.

Abbreviations

ncRNA: Noncoding RNAs
miRNA: MicroRNA
isomiR: miRNA variant
lncRNA: Long noncoding RNA
mRNA: Messenger RNA
pre-miRNA: Precursor miRNA
RPM: Reads per million.

Conflict of Interests

The authors declare no potential conflict of interests with respect to the authorship and/or publication of this paper.

Authors' Contribution

Li Guo wrote the main paper text, Li Guo and Feng Chen designed the study, Li Guo and Yang Zhao analyzed original datasets, and Li Guo, Sheng Yang, and Hui Zhang prepared tables.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 61301251, 81072389, and 81373102), the Research Fund for the Doctoral Program of Higher Education of China (nos. 211323411002 and 20133234120009), the China Postdoctoral Science Foundation funded project (no. 2012M521100), the key Grant of the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (no. 10KJA33034), the National Natural Science Foundation of Jiangsu (no. BK20130885), the Natural Science Foundation of the Jiangsu Higher Education Institutions (nos. 12KJB310003 and 13KJB330003), the Jiangsu Planned Projects for Postdoctoral Research Funds (no. 1201022B), the Science and Technology Development Fund Key Project of Nanjing Medical University (no. 2012NJMU001), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

[1] P. Kapranov, G. St Laurent, T. Raz et al., "The majority of total nuclear-encoded non-ribosomal RNA in a human cell is "dark matter" un-annotated RNA," *BMC Biology*, vol. 8, article 149, 2010.

[2] H. Jia, M. Osak, G. K. Bogu, L. W. Stanton, R. Johnson, and L. Lipovich, "Genome-wide computational identification and manual annotation of human long noncoding RNA genes," *RNA*, vol. 16, no. 8, pp. 1478–1487, 2010.

[3] P. Kapranov, J. Cheng, S. Dike et al., "RNA maps reveal new RNA classes and a possible function for pervasive transcription," *Science*, vol. 316, no. 5830, pp. 1484–1488, 2007.

[4] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[5] D. P. Bartel and C. Z. Chen, "Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs," *Nature Reviews Genetics*, vol. 5, no. 5, pp. 396–400, 2004.

[6] V. Ambros, "The functions of animal microRNAs," *Nature*, vol. 431, no. 7006, pp. 350–355, 2004.

[7] H. W. Hwang and J. T. Mendell, "MicroRNAs in cell proliferation, cell death, and tumorigenesis," *The British Journal of Cancer*, vol. 94, no. 6, pp. 776–780, 2006.

[8] G. A. Calin, M. Ferracin, A. Cimmino et al., "A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia," *The New England Journal of Medicine*, vol. 355, p. 533, 2006.

[9] C. Caldas and J. D. Brenton, "Sizing up miRNAs as cancer genes," *Nature Medicine*, vol. 11, no. 7, pp. 712–714, 2005.

[10] R. A. Gupta, N. Shah, K. C. Wang et al., "Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis," *Nature*, vol. 464, no. 7291, pp. 1071–1076, 2010.

[11] K. V. Morris and P. K. Vogt, "Long antisense non-coding RNAs and their role in transcription and oncogenesis," *Cell Cycle*, vol. 9, no. 13, pp. 2544–2547, 2010.

[12] K. C. Wang and H. Y. Chang, "Molecular mechanisms of long noncoding RNAs," *Molecular Cell*, vol. 43, no. 6, pp. 904–914, 2011.

[13] M. Huarte, M. Guttman, D. Feldser et al., "A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response," *Cell*, vol. 142, no. 3, pp. 409–419, 2010.

[14] A. M. Khalil, M. Guttman, M. Huarte et al., "Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 28, pp. 11667–11672, 2009.

[15] P. Carninci, T. Kasukawa, S. Katayama et al., "The transcriptional landscape of the mammalian genome," *Science*, vol. 309, no. 5740, pp. 1559–1563, 2005.

[16] E. Pasmant, I. Laurendeau, D. Héron, M. Vidaud, D. Vidaud, and I. Bièche, "Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF," *Cancer Research*, vol. 67, no. 8, pp. 3963–3969, 2007.

[17] Y. Kotake, T. Nakagawa, K. Kitagawa et al., "Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15 INK4B tumor suppressor gene," *Oncogene*, vol. 30, no. 16, pp. 1956–1962, 2011.

[18] S. Tochitani and Y. Hayashizaki, "Nkx2.2 antisense RNA overexpression enhanced oligodendrocytic differentiation," *Biochemical and Biophysical Research Communications*, vol. 372, no. 4, pp. 691–696, 2008.

[19] J. Feng, C. Bi, B. S. Clark, R. Mady, P. Shah, and J. D. Kohtz, "The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator," *Genes and Development*, vol. 20, no. 11, pp. 1470–1484, 2006.

[20] W. Yu, D. Gius, P. Onyango et al., "Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA," *Nature*, vol. 451, no. 7175, pp. 202–206, 2008.

[21] X. Wang, S. Arai, X. Song et al., "Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription," *Nature*, vol. 454, no. 7200, pp. 126–130, 2008.

[22] I. Shamovsky, M. Ivannikov, E. S. Kandel, D. Gershon, and E. Nudler, "RNA-mediated response to heat shock in mammalian cells," *Nature*, vol. 440, no. 7083, pp. 556–560, 2006.

[23] A. T. Willingham, A. P. Orth, S. Batalov et al., "A strategy for probing the function of noncoding RNAs finds a repressor of NFAT," *Science*, vol. 309, no. 5740, pp. 1570–1573, 2005.

[24] Z. Yang, Q. Zhu, K. Luo, and Q. Zhou, "The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription," *Nature*, vol. 414, no. 6861, pp. 317–322, 2001.

[25] V. T. Nguyen, T. Kiss, A. A. Michels, and O. Bensaude, "7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes," *Nature*, vol. 414, no. 6861, pp. 322–325, 2001.

[26] L. Guo, Y. Zhao, S. Yang, H. Zhang, and F. Chen, "Integrative analysis of miRNA-mRNA and miRNA-miRNA interactions," *BioMed Research International*, vol. 2014, Article ID 907420, 8 pages, 2014.

[27] A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data," *Nucleic Acids Research*, vol. 39, no. 1, pp. D152–D157, 2011.

[28] F. Kuchenbauer, R. D. Morin, B. Argiropoulos et al., "In-depth characterization of the microRNA transcriptome in a leukemia

progression model," *Genome Research*, vol. 18, no. 11, pp. 1787–1797, 2008.

[29] R. D. Morin, M. D. O'Connor, M. Griffith et al., "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells," *Genome Research*, vol. 18, no. 4, pp. 610–621, 2008.

[30] J. G. Ruby, C. Jan, C. Player et al., "Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*," *Cell*, vol. 127, no. 6, pp. 1193–1207, 2006.

[31] S. L. Fernandez-Valverde, R. J. Taft, and J. S. Mattick, "Dynamic isomiR regulation in *Drosophila* development," *RNA*, vol. 16, no. 10, pp. 1881–1888, 2010.

[32] L. W. Lee, S. Zhang, A. Etheridge et al., "Complexity of the microRNA repertoire revealed by next-generation sequencing," *RNA*, vol. 16, no. 11, pp. 2170–2180, 2010.

[33] A. M. Burroughs, Y. Ando, M. J. L. de Hoon et al., "A comprehensive survey of $3'$ animal miRNA modification events and a possible role for $3'$ adenylation in modulating miRNA targeting effectiveness," *Genome Research*, vol. 20, no. 10, pp. 1398–1410, 2010.

[34] L. Guo, Q. Yang, J. Lu et al., "A comprehensive survey of miRNA repertoire and $3'$ addition events in the placentas of patients with pre-eclampsia from high-throughput sequencing," *PLoS ONE*, vol. 6, no. 6, Article ID e21072, 2011.

[35] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[36] D. Y. Chiang, P. O. Brown, and M. B. Eisen, "Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles," *Bioinformatics*, vol. 17, no. 1, pp. S49–S55, 2001.

[37] S.-D. Hsu, F. Lin, W. Wu et al., "miRTarBase: a database curates experimentally validated microRNA-target interactions," *Nucleic Acids Research*, vol. 39, no. 1, pp. D163–D169, 2011.

[38] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, "Human microRNA targets," *PLoS Biology*, vol. 2, no. 11, article e363, 2004.

[39] M. E. Smoot, K. Ono, J. Ruscheinski, P. Wang, and T. Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics*, vol. 27, no. 3, pp. 431–432, 2011.

[40] L. Guo, H. Li, J. Lu et al., "Tracking miRNA precursor metabolic products and processing sites through completely analyzing high-throughput sequencing data," *Molecular Biology Reports*, vol. 39, no. 2, pp. 2031–2038, 2012.

[41] S. Shimizu, T. Takehara, H. Hikita et al., "The let-7 family of microRNAs inhibits Bcl-xL expression and potentiates sorafenib-induced apoptosis in human hepatocellular carcinoma," *Journal of Hepatology*, vol. 52, no. 5, pp. 698–704, 2010.

[42] W. Hou, Q. Tian, N. M. Steuerwald, L. W. Schrum, and H. L. Bonkovsky, "The let-7 microRNA enhances heme oxygenase-1 by suppressing Bach1 and attenuates oxidant injury in human hepatocytes," *Biochimica et Biophysica Acta—Gene Regulatory Mechanisms*, vol. 1819, no. 11-12, pp. 1113–1122, 2012.

[43] C. Braconi, T. Kogure, N. Valeri et al., "microRNA-29 can regulate expression of the long non-coding RNA gene MEG3 in hepatocellular cancer," *Oncogene*, vol. 30, no. 47, pp. 4750–4756, 2011.

[44] A. Rodriguez, S. Griffiths-Jones, J. L. Ashurst, and A. Bradley, "Identification of mammalian microRNA host genes and transcription units," *Genome Research*, vol. 14, no. 10 A, pp. 1902–1910, 2004.

[45] V. A. Erdmann, M. Szymanski, A. Hochberg, N. De Groot, and J. Barciszewski, "Non-coding, mRNA-like RNAs database Y2K," *Nucleic Acids Research*, vol. 28, no. 1, pp. 197–200, 2000.

[46] K. Augoff, B. McCue, E. F. Plow, and K. Sossey-Alaoui, "MiR-31 and its host gene lncRNA LOC554202 are regulated by promoter hypermethylation in triple-negative breast cancer," *Molecular Cancer*, vol. 11, article 5, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 404578, 9 pages
http://dx.doi.org/10.1155/2014/404578

Research Article

Gleditsia sinensis: Transcriptome Sequencing, Construction, and Application of Its Protein-Protein Interaction Network

Liucun Zhu,[1] Ying Zhang,[2] Wenna Guo,[1] and Qiang Wang[3]

[1] *Institute of System Biology, Shanghai University, Shanghai 200444, China*
[2] *Yangzhou Breeding Biological Agriculture Technology Co. Ltd., Yangzhou 225200, China*
[3] *State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing 210093, China*

Correspondence should be addressed to Qiang Wang; wangq@nju.edu.cn

Received 11 March 2014; Accepted 21 April 2014; Published 27 May 2014

Academic Editor: Lei Chen

Gleditsia sinensis is a genus of deciduous tree in the family *Caesalpinioideae*, native to China, and is of great economic importance. However, despite its economic value, gene sequence information is strongly lacking. In the present study, transcriptome sequencing of *G. sinensis* was performed resulting in approximately 75.5 million clean reads assembled into 142155 unique transcripts generating 58583 unigenes. The average length of the unigenes was 900 bp, with an N50 of 549 bp. The obtained unigene sequences were then compared to four protein databases to include NCBI nonredundant protein (NRDB), Swiss-prot, Kyoto Encyclopedia of Genes and Genomes (KEGG), and the Cluster of Orthologous Groups (COG). Using BLAST procedure, 31385 unigenes (53.6%) were generated to have functional annotations. Additionally, sequence homologies between identified unigenes and genes of known species in a protein-protein interaction (PPI) network facilitated *G. sinensis* PPI network construction. Based on this network construction, new stress resistance genes (including cold, drought, and high salinity) were predicted. The present study is the first investigation of genome-wide gene expression in *G. sinensis* with the results providing a basis for future functional genomic studies relating to this species.

1. Introduction

Gleditsia sinensis is a genus of deciduous tree in the family Caesalpinioideae, native to China. *G. sinensis* usually grows 15–30 m tall and is of economic and medicinal importance. The fruits of *G. sinensis* can serve as medicine, food, health products, cosmetics, and natural raw materials for cleaning products [1]; the seeds can be used as appetizing medicine [2, 3] and contain an important vegetable gum (guar gum) [2]; the thorns (soap-pin) contain flavonoid glycosides, phenols, and amino acids and have a high economic value [3]. However, up to October 2013, only 17 nucleotide sequences and eight protein sequences of *G. sinensis* were available in the NCBI database. This brings to question why such an economically valuable organism has been so understudied, making it important to generate more genetic sequence information to further study *G. sinensis*.

Plants are exposed to continuously changing environmental conditions under natural conditions. Various environmental stresses, such as heat, cold, drought, and high salinity, are major factors in affecting plant development, growth, and productivity [4–6]. The stress-induced transcriptomic responses of plants revealed that many molecular mechanisms had been evolved to help plants to adapt and survive under the harmful stresses. Usually, there is an initial activation/sensory stage followed by a physiological stage when the plant perceived and responded to the abiotic stress [6, 7]. Previous work in a variety of stresses has been studied in *Arabidopsis thaliana* [8, 9]. Compared to *Arabidopsis thaliana*, there is little known about how trees respond to abiotic stress. In recent years, the emergence of next-generation sequencing technology has provided a fast and effective approach to generation of transcriptome data of nonmodel organisms lacking a complete genome sequence [10, 11]. Compared with whole-genome sequencing, RNA sequencing (RNA-seq) is of low cost and high throughput, becoming an important part of functional genome research [12]. It provided an efficient way of identifying the expression level and new members of the genes [13, 14].

In the present study, the Illumina HiSeq 2000 platform was used for transcriptome sequencing in four tissue types collected from *G. sinensis*. A total of 7632619288 reads assembled into 58583 unigenes and their functional annotations were obtained. In addition, a protein-protein interaction (PPI) network comprising genes expressed in *G. sinensis* was constructed and utilized to identify some new potential drought, freezing, and salinity tolerance genes. These findings will provide a solid foundation for further investigation of the functional genomics of *G. sinensis*.

2. Materials and Method

2.1. RNA Preparation, Sequencing, and Assembly. The *G. sinensis* specimen used in the present study was wild-grown from the Jiangsu Province, China. Total RNA was extracted from four tissues: tender shoots, young leaves, flower buds, and round thorns, using TRIzol Reagent with qualification and quantification evaluated by Agilent 2100 Bioanalyzer Nanochips and NanoDrop 2000 Spectrophotometer. And then it was processed and used for Illumina sequencing [15].

Raw read sequences are uploaded in the Short Read Archive database from National Center for Biotechnology Information (NCBI) with the accession number SRR 1012862.

We used SeqPreq (https://github.com/jstjohn/SeqPrep) to remove sequencing adapters and then used sickle [16] to trim low-quality sequences with default parameters. After cleaning reads, all of the high-quality reads were used in assembling by Trinity (trinityrnaseq_r2013-02-25) [17, 18]. The k-mer was counted by jellyfish and the min_contig_length was set as 300. Then, RSEM [19] was used to measure expression levels of every unique transcript. Units of TPM (transcripts per million) were used to report results. After counting fraction of each isoform, length × isoform percent was defined as a standard to identify unigenes.

2.2. Functional Annotation. Unigenes larger than 300 bp were subjected to functional annotation for predicting putative gene descriptions, conserved domains, gene ontology (GO) terms, and association with metabolic pathways. First, BLAST procedure (ftp://ftp.ncbi.nih.gov/blast/executables/release/2.2.18/) was used to compare all unigenes against protein sequence databases to include NRDB, Swissprot, and Clusters of Orthologous Groups (COG) with an E-value $< 1.0E-6$. Based on BLAST best match results, gene function and protein-related information were predicted, with Blast2GO software [20] used for gene ontology (GO) annotation in terms of molecular function, cellular component, and biological process. After all GO annotations were obtained, Web Gene Ontology Annotation Plot (WEGO) software was used to construct comparative plots based on unigene classifications. Unigenes were also subjected to sequence comparison using the COG database for gene function prediction [21]. For Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation [22], the BLASTX software was used for sequence comparison of unigenes within the KEGG database and completed on the KASS website [23] (http://www.genome.jp/tools/kaas/). Following KEGG annotation, each gene acquired a KEGG Orthology (KO) number representing a node in a certain reference metabolic pathway in KEGG.

2.3. Construction and Analysis of PPI Networks. First, known PPI networks and protein sequences from six species (*Arabidopsis thaliana*, *Arabidopsis lyrata*, *Oryza sativa* subsp. *japonica*, *Brachypodium distachyon*, *Populus trichocarpa*, and *Sorghum bicolor*) were downloaded from the String database [24]. Then using the TBLASTN software, protein sequences from the downloaded PPIs were compared with *G. sinensis* unigenes to identify homologous sequences with an E-value $< 1.0E-6$. The criterion of candidate interacting genes of the network was the TBLASTN hits with identity >50% and covering query gene >80%. If two unigenes from *G. sinensis* corresponded to two homologous proteins in the known networks, the encoded proteins were considered to interact with each other. Concluding network construction, each node in the network was assigned to a degree k, which is the number of connected neighboring nodes. The degree distribution of giant network branches was computed using the formula $P(k) = N(k)/N$ where N is the number of nodes and $N(k)$ denotes the number of nodes with degree k [25].

Stress resistance genes and protein sequences of *Arabidopsis thaliana* proteins related to salinity, drought, and freezing tolerance were downloaded from the stress responsive transcription factor database [26–28] (STIFDB; http://caps.ncbs.res.in/stifdb2/) and compared with the *G. sinensis* unigene library to search for homologous sequences potentially possessing the same functions in *G. sinensis*. Next, the PPI network was used to predict novel drought, freezing, and salinity tolerance genes. The specific predictive criterions are as follows. If a protein in the *G. sinensis* network connects directly with the homologous sequences of over four known stress resistance genes, with no homology among these genes, then that gene was predicted to be a potential stress resistance gene.

3. Result

3.1. Paired-End Sequencing and De Novo Assembly of G. sinensis Transcriptome. Genes are differentially expressed in different tissue types. In an effort to broaden the obtained gene expression profile in *G. sinensis*, total RNA was extracted from four different tissues (tender shoots, young leaves, flower buds, and round thorns) and mixed in equal parts for sequencing using the Illumina platform. This resulted in a total of 75.6 million high-quality clean reads containing 7632619288 nucleotides (nt) and an average read length of 101 bp (Table 1). Due to a lack of *G. sinensis* whole-genome sequence, Trinity software was used to assemble all high-quality clean reads into transcripts *de novo*. From all of the clean reads, a total of 142155 unique transcripts with an average length of 1537 bp were obtained with a N50 of 1202 bp and the majority of unique transcripts (31818) between 100 and 500 bp (Figure 1(a)). Then after removing redundant sequences, 58583 unigenes were obtained, with an average length of 900 bp. The lengths of the unigenes varied from

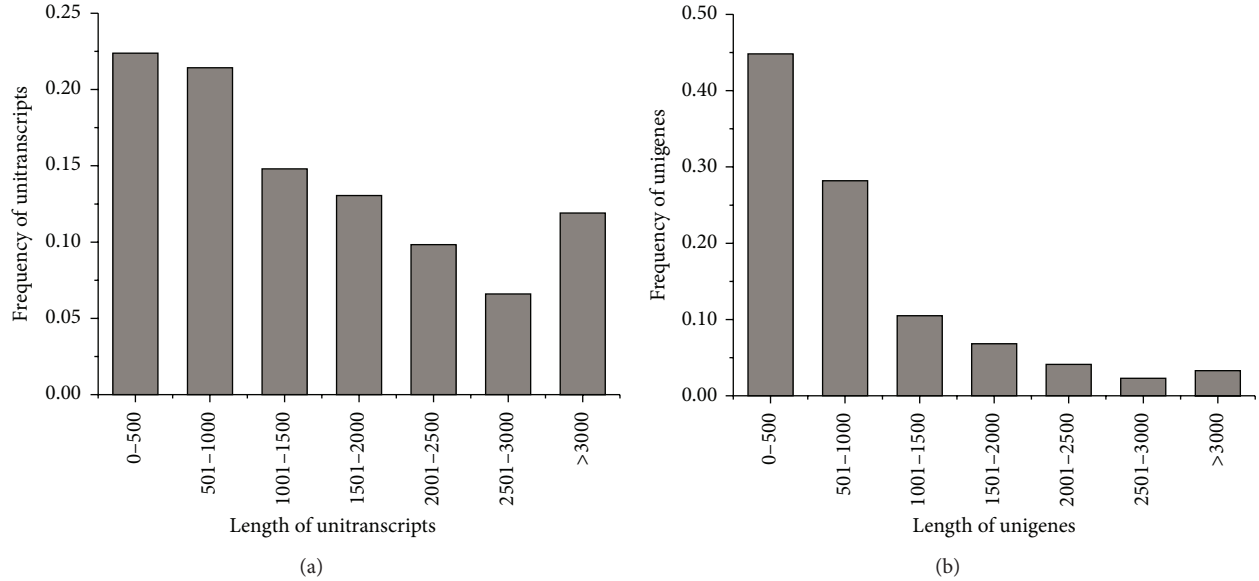


Figure 1: Overview of the *G. sinensis* transcriptome assembly. The size distribution of the UTs (a) and unigenes (b) produced from *de novo* assembly of reads by Trinity.

Table 1: Summary of sequence assembly by Trinity after Illumina sequencing.

	Number	Mean size (bp)	N50 size (bp)	Total nucleotides (bp)
Read	75570488	101	101	7632619288
Unique transcript	142155	1537	1202	218503453
Unigene	58583	900	549	52719022

300 bp to over 3000 bp and the length distribution of unigenes was shown in Figure 1(b).

3.2. Functional Annotation and Classification of G. sinensis Transcriptome. Nonredundant database (NRDB) built in NCBI contains large amounts of protein information. For annotation of the *G. sinensis* transcriptome, all unigenes were compared against the NRDB using BLASTX (an *E*-value cut-off of $1e^{-6}$) to reveal 31100 unigenes with sequence homology. Among them, 45.1% of unigenes have the best matches mapping soybean.

The distribution of *E*-values based on sequence homology showed that 61.1% of unigenes had high homology (smaller than $1.0e-50$) with the *E*-values of the other matches varying from $1.0E-50$ to $1.0E-6$ (Figure 2(a)). The similarity distribution showed the majority of unigenes (93.0%) with homologous sequences having similarities between 40% and 100%, with only 7% of the unigenes with homologous sequences having similarities less than 40% (Figure 2(b)).

Swiss-prot, an annotated protein sequence database maintained by the European Bioinformatics Institute (EBI), was also employed for unigene comparison which revealed 22157 of 58583 unigenes (37.8%) with sequence homology at an *E*-value threshold of $\leq 1.0E-6$ (Table 2). Almost half of these unigenes (49.2%) had an *E*-value $\leq 1.0E-50$, and the remaining had *E*-values between $1.0E-50$ and $1.0E-6$ which showed a slightly different result compared to the NRDB query (Figure 2(c)). In Swiss-prot comparison, 73.4% of the unigenes had sequence homology in the range of 40% to 100%, with only 26.6% having homology <40% (Figure 2(d)). In short, when combining the results from Swiss-prot and NRDB comparisons, a total of 31131 unigenes were confirmed to have homologous sequences.

Table 2: Summary of annotation of *G. sinensis* unigenes.

Category	Number	Percentage
NR annotated unigene	31100	53.09%
Swiss-prot	22157	37.82%
GO classified unigene	15264	26.06%
COG classified unigene	6413	10.95%
KEGG classified unigene	2914	4.97%

3.3. Classification by Gene Ontology (GO) Annotation. Gene ontology (GO) terms were utilized to assign gene function classifications to each unigene based on sequence similarity comparisons from NRDB. Among the 58583 unigenes identified in *G. sinensis*, 15264 were categorized into at least one of the three main GO categories which could be further subdivided into 51 subcategories (Table 2, Figure 3, and Additional File 1; see Supplementary Material available online at http://dx.doi.org/10.1155/2014/404578). The number of unigenes in cellular component, biological process, and

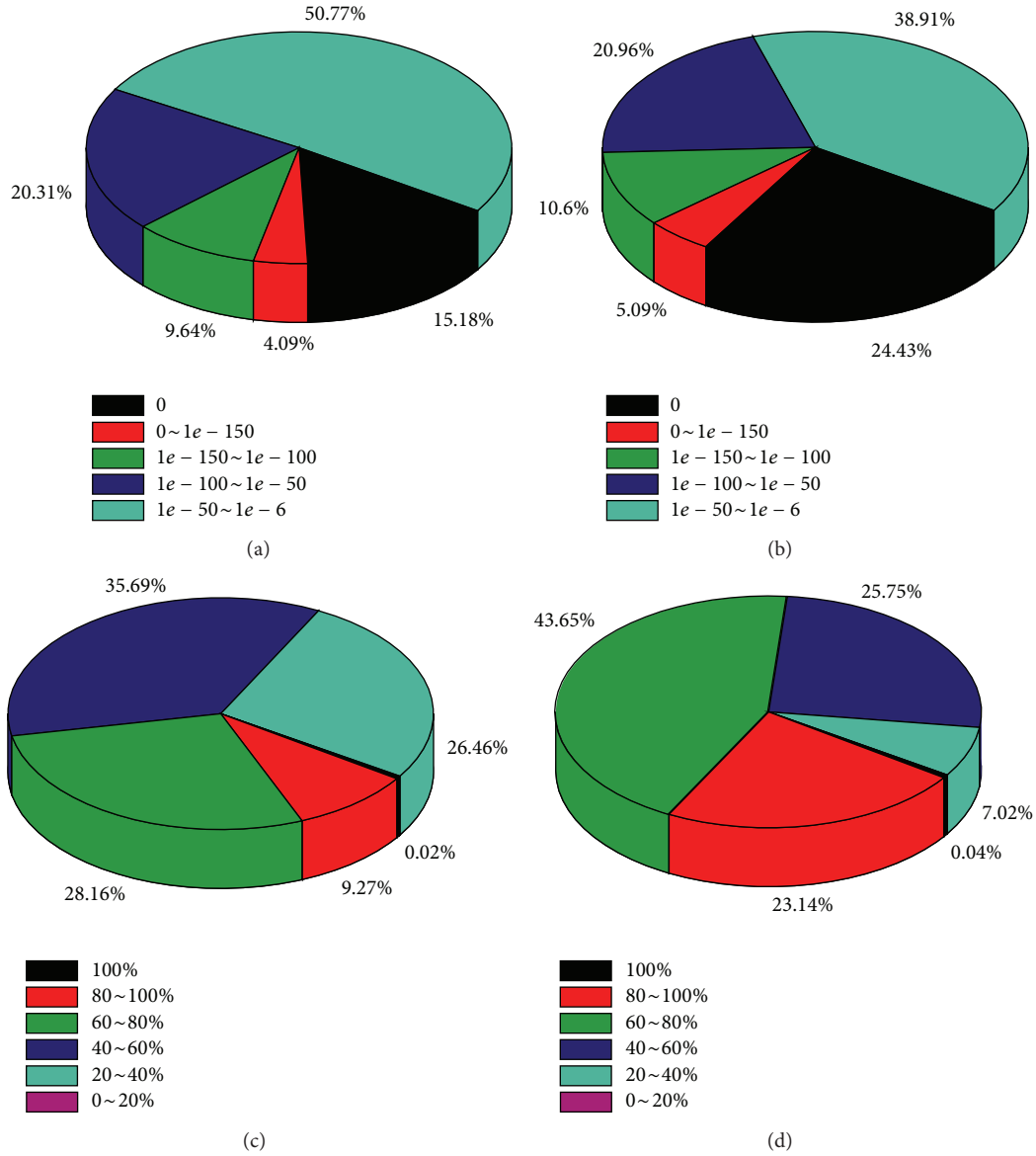


FIGURE 2: Unigene homology searches against NR and Swiss-prot databases shown by pie graphs. *E*-value proportional frequency distribution of BLAST hits against the NR database (a) and Swiss-prot database (c). Proportional frequency distribution of UTs similarities against the NR database (b) and Swiss-prot database (d) based on the best BLAST hits (*E*-value ≤ $1.0E - 6$).

molecular function was 11942 (20.4%), 11683 (19.9%), and 11129 (19.0%), respectively. The subcategory at the "cell," "cell part," "cellular process," "organelle," "metabolic process," "binding," and "catalytic activity" level included the highest number of unigenes relative to other subcategories, with the biological processes category also showing large numbers relating to cellular processes (9353) and metabolic processes (9039). This suggested an abundance of metabolic activities occurring at the time of sampling.

3.4. Functional Classification by KEGG. To further predict the metabolic pathways of *G. sinensis*, *Arabidopsis thaliana* and *Oryza sativa* were used as references and each assembled unigene was annotated in KAAS to obtain the corresponding enzyme commission (EC) numbers. KEGG is considered a basic platform for systematic functional analysis based on constructed networks comprising gene products. KEGG analysis mapped a total of 2914 unigenes to 307 metabolic pathways encompassing five KEGG categories, including metabolism, genetic information processing, environmental information processing, cellular processes, and organismal systems (Additional File 2). The "metabolic" pathways show the highest representation (1357 members, 46.6%), followed by ribosome (180 members, 6.2%, ko03010), biosynthesis of amino acids (147 members, 5.0%, ko01230), carbon metabolism (133 members, 4.6%, ko01200), plant hormone signal transduction (129 members, 4.4%, ko4075), spliceosome (127 members, 4.3%, ko03040), protein processing in

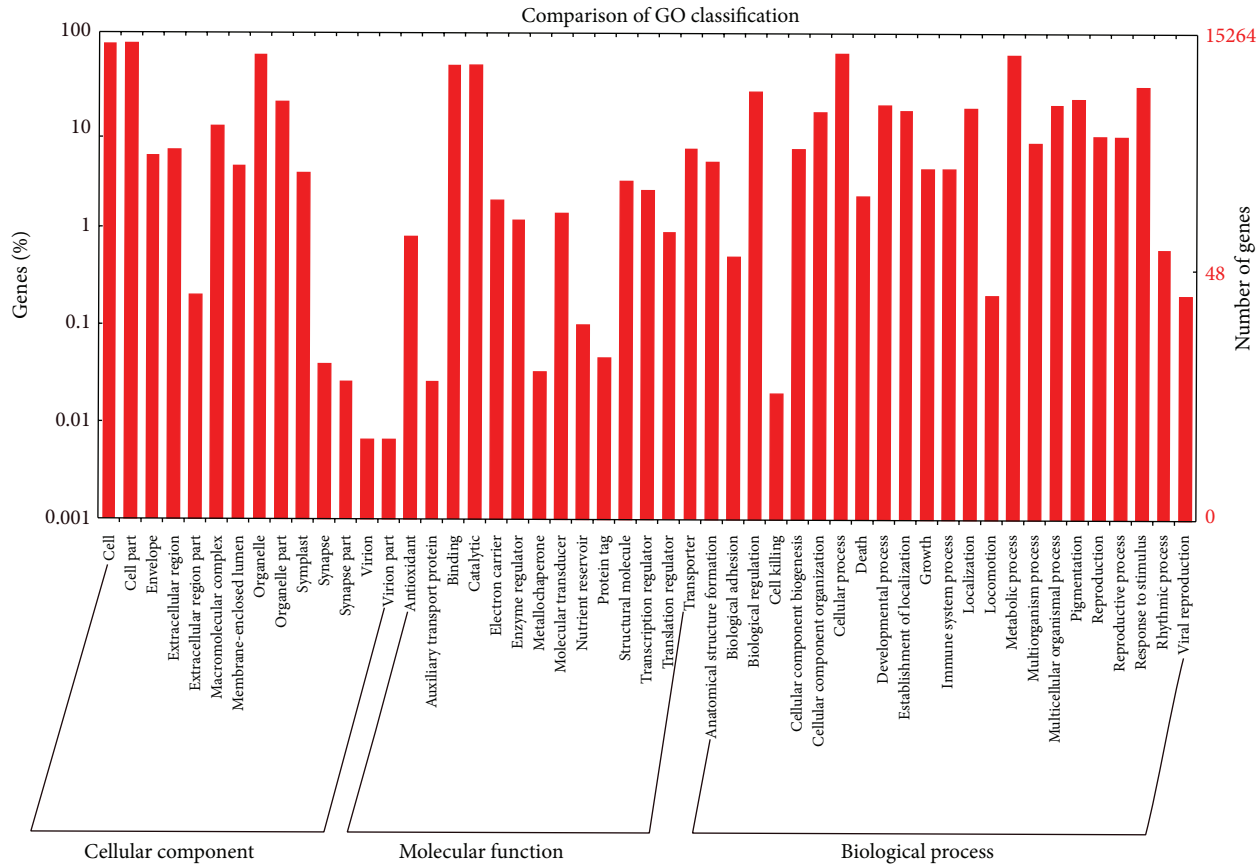


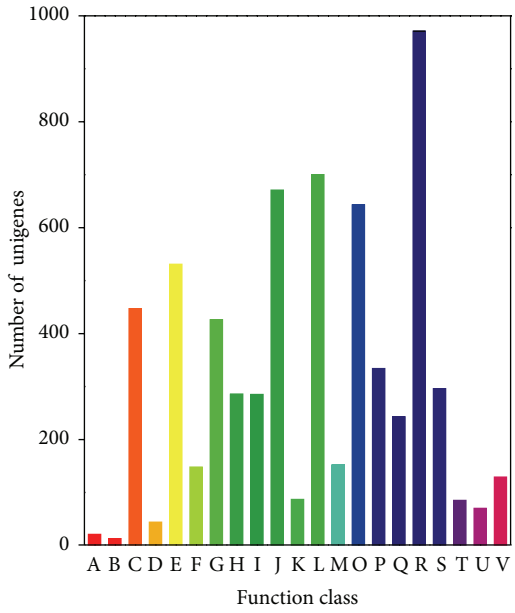
FIGURE 3: Gene ontology classification of the *G. sinensis* transcriptome. Gene ontology (GO) term assignments to *G. sinensis* unigenes based on significant hits against the NR database are summarized into three main GO categories (biological process, cellular component, and molecular function) and 51 subcategories.

the endoplasmic reticulum (125 members, 4.3%, ko04141), purine metabolism (109 members, 3.7%, ko00230), and RNA transport (100 members, 3.5%, ko03013).

3.5. Classification by COG. Cluster of Orthologous Groups (COG) database containing proteins encoded by 66 complete genomes of bacteria, algae, and eukaryotes was constructed according to phylogenetic relationships. COG is a very useful tool for predicting the functionality of the individual proteins or all proteins in a new genome. In the present study, all obtained unigenes were compared with proteins in COG and classified into appropriate COG clusters. The results identified 6413 genes displaying significant sequence homology with COG database proteins, accounting for 11.0% of all unigenes. Some unigenes were shown to have multiple COG functions with a total of 6582 functional annotations noted and grouped into 21 COG function sets with E-values $\leq 1.0E-6$ (Table 2, Figure 4, and Additional File 3). The five sets including the largest number of unigenes were (1) "general function prediction only" (14.8%); (2) "replication, recombination, and repair" (10.6%); (3) "translation, ribosomal structure, and biogenesis" (10.2%); (4) "posttranslational modification, protein turnover, and chaperones" (9.8%); and (5) "amino acid transport and metabolism" (8.1%). The "RNA processing and modification" and "chromatin structure and dynamics" sets contained the least numbers of unigenes, 21 and 13, respectively.

In summary, 31385 unigenes were annotated using NR, Swiss-prot, COG, and KEGG databases with E-values $\leq 1.0E-6$ deemed significant. Among these unigenes, 1433 showed BLAST match results in all four public databases demonstrating a strong functional annotation. These annotations provide valuable resources for further study of the specific metabolic activities, gene structures, and functions and pathways of *G. sinensis*.

3.6. Construction of Protein-Protein Interaction Network in G. sinensis. Using TBLASTN, similarities between *G. sinensis* unigenes and genes in a PPI network consisting of six String database genomes (*Arabidopsis thaliana*, *Arabidopsis lyrata*, *Oryza sativa* subsp. *japonica*, *Brachypodium distachyon*, *Populus trichocarpa*, and *Sorghum bicolor*) facilitated the construction of a PPI network of *G. sinensis*. This network contained one giant component and 91 smaller components (Figure 5 and Figure S1). The giant component contained 1,897 nodes with 7078 links between nodes. The degree distribution of giant component conformed to $P(k) = 0.23k^{-0.91}$,



A: RNA processing and modification
B: chromatin structure and dynamics
C: energy production and conversion
D: cell cycle control, cell division, and chromosome partitioning
E: amino acid transport and metabolism
F: nucleotide transport and metabolism
G: carbohydrate transport and metabolism
H: coenzyme transport and metabolism
I: lipid transport and metabolism
J: translation, ribosomal structure, and biogenesis
K: cell wall/membrane/envelope biogenesis
L: posttranslational modification, protein turnover, and chaperones
M: inorganic ion transport and metabolism
O: secondary metabolites biosynthesis, transport, and catabolism
P: transcription
Q: replication, recombination, and repair
R: general function prediction only
S: unknown function
T: signal transduction mechanisms
U: intracellular trafficking, secretion, and vesicular transport
V: defense mechanisms

Figure 4: COG function classification of the *G. Sinensis* transcriptome. A total of 6413 unigenes with significant homologies to the COG database (E-values $\leq 1.0E-6$) were classified into 21 COG categories.

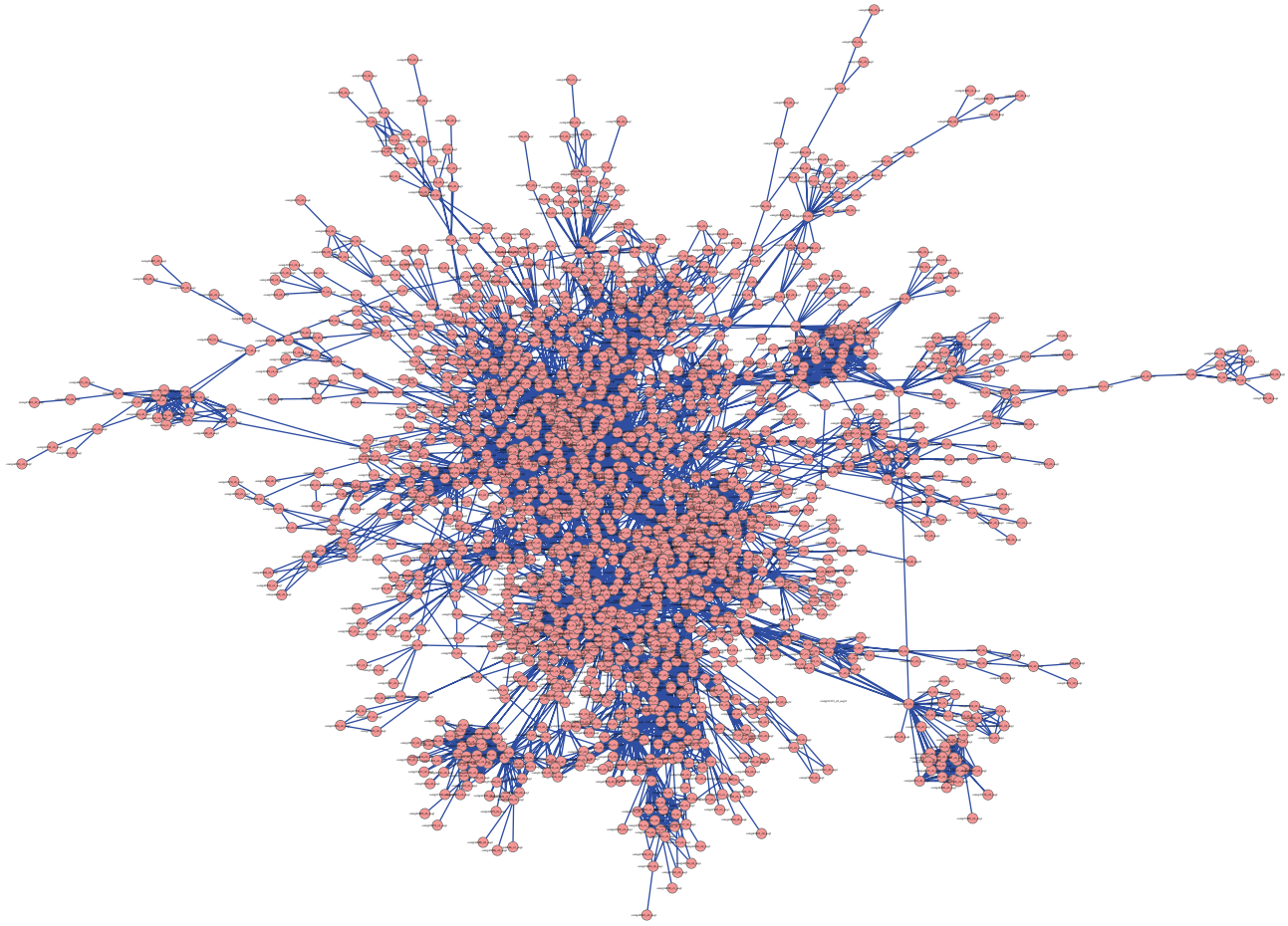


Figure 5: Illustration of the giant component of unigenes in *G. sinensis*.

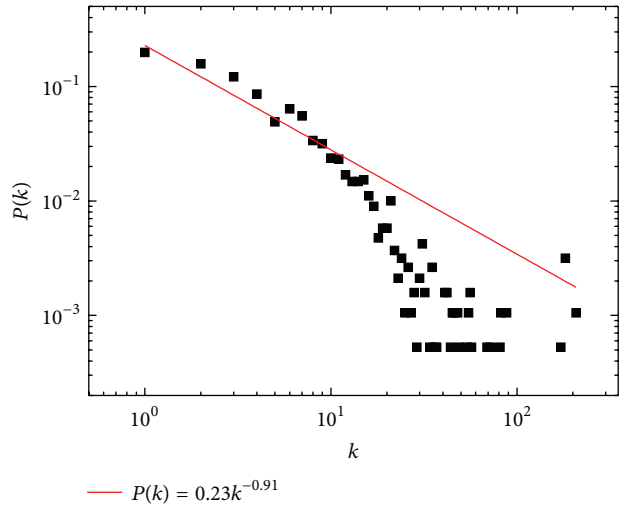


Figure 6: The topological analysis of the giant component in *G. sinensis* with 1897 nodes and 7078 edges. Log-log plots of the node degree distribution with a power-law fit (red line).

implying that the network was scale free and similar to other biological networks (Figure 6).

3.7. Prediction of Resistance Genes in G. sinensis. Resistance genes and protein sequences from *Arabidopsis thaliana* relating to freezing, drought, and salinity tolerance were downloaded from STIFDB and compared with *G. sinensis* unigenes to locate sequence homology. 435 freezing tolerance genes, 284 drought tolerance genes, and 348 salinity tolerance genes were found (Additional File 4). Based on the constructed *G. sinensis* PPI network, new freezing, drought, and salinity tolerance genes were predicted with a protein considered a potential resistance gene if interactions with over four known resistance genes were noted. The results revealed 19 new freezing tolerance genes, 11 drought tolerance genes, and 18 salinity tolerance genes (Table 3). This provides a theoretical basis for future experimental studies on resistance genes and for culturing resistant *G. sinensis* varieties.

4. Discussion

Gleditsia sinensis is a tree species of important economic and medicinal value. However, due to a lack of genomic research, molecularly based breeding of *G. sinensis* is hindered. Recently, RNA-seq technology has provided a new approach to obtaining rich sequence information to include successful applications in many plants, such as *Youngia japonica* [29], cabbage [30], tea plant [15], and citrus psyllid [31]. In the present study, RNA-seq was used for transcriptome sequencing of *G. sinensis* with 7.6 Gbp examined and 75.6 million clean reads obtained. The Trinity software was used for *de novo* assembly and a total of 58583 unigenes were obtained. Among these, 31385 unigenes were functionally annotated and shown to participate in a variety of biological processes, accounting for 53.6% of all obtained unigenes.

Table 3: Novel resistant genes relating to cold, drought, and salinity tolerance identified by network analysis.

	Cold	Drought	Salinity
comp22423_c0_seq1	Yes	Yes	Yes
comp24732_c0_seq2	Yes	No	No
comp25430_c0_seq1	Yes	No	No
comp28513_c0_seq1	Yes	No	No
comp37332_c0_seq1	Yes	Yes	Yes
comp38200_c0_seq1	Yes	No	No
comp38390_c0_seq5	No	Yes	Yes
comp39646_c0_seq2	Yes	Yes	Yes
comp39900_c0_seq1	No	Yes	Yes
comp39917_c0_seq1	No	No	Yes
comp41481_c0_seq2	Yes	No	No
comp42998_c0_seq1	Yes	No	Yes
comp43037_c0_seq1	Yes	No	Yes
comp43634_c0_seq1	No	Yes	Yes
comp45199_c0_seq1	No	Yes	Yes
comp45561_c0_seq1	Yes	No	No
comp47037_c0_seq1	No	Yes	Yes
comp47415_c0_seq1	Yes	No	Yes
comp47471_c0_seq6	No	Yes	Yes
comp47482_c0_seq2	Yes	No	Yes
comp47503_c0_seq6	Yes	No	Yes
comp47673_c0_seq3	Yes	No	No
comp47694_c0_seq2	Yes	No	No
comp48118_c0_seq4	No	Yes	Yes
comp50179_c0_seq1	Yes	Yes	Yes
comp51180_c1_seq11	Yes	No	Yes
comp81788_c0_seq1	Yes	No	No

Besides, freezing, drought, and salinity tolerance genes in *G. sinensis* were predicted by searching for homologous genes linked to resistance and using PPI networks. Currently there are many available methods for gene function prediction based on PPI. For example, George et al. assumed that genes interacting with known disease genes were also disease genes and studied third-degree interactions, yet many false positives were discovered based on nondirect interactions [32]. Xu and Li reported five topological characteristics of disease-related PPI networks, yet these characteristics were not found in the yeast two-hybrid network. These characteristics were used to predict disease-related genes, yet they were unable to explain the biological significance of these characteristics, and their results still required experimental verifications [33]. Based on these previous studies and to ensure the lowest rate of false positive predictions, the present study applied more strict conditions to include direct connections with known resistance genes and interactions with over four known resistance genes. Whether the identified genes are indeed related to resistance still needs further experimental validation and this will be the focus of our future research.

5. Conclusion

In the present study, Illumina RNA-seq and *de novo* assembly methods were applied to study the transcriptome of *G. sinensis* for the first time. A total of approximately 75.6 million reads, assembled into 58583 unigenes with an average length of 900 bp, were identified. Among these unigenes, 31385 obtained annotation from NR, Swiss-prot, COG, and KEGG databases. The results of the present study confirm that RNA-seq technology is a fast, effective method for transcriptome analysis of nonmodel plants and provides a good resource for further gene expression analysis. The constructed PPI network for *G. sinensis* when compared to known resistance genes of *Arabidopsis* predicted 18 freezing tolerance genes, 11 drought tolerance genes, and 19 salinity tolerance genes. Thus these findings provide a theoretical basis for future culturing of stress-resistant *G. sinensis* varieties.

Abbreviations

UT: Unique transcript
NR: NCBI nonredundant protein
COG: Cluster of Orthologous Groups
GO: Gene ontology
KEGG: Kyoto Encyclopedia of Genes and Genomes database
KAAS: KEGG automatic annotation server
BLAST: Basic local alignment search tool
PPI: Protein-protein network.

Conflict of Interests

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

Authors' Contribution

Liucun Zhu and Ying Zhang contributed equally to this work.

Acknowledgments

This work was supported by Grants from the Science and Technology Commission of Shanghai Municipality (12ZR1444200), Foundation for the Author of National Excellent Doctoral Dissertation of China (201134), Doctor Gathering Scheme of Jiangsu Province, and the High Performance Computing Platform of Shanghai University.

References

[1] X.-Y. Lian and Z. Zhang, "Quantitive analysis of gleditsia saponins in the fruits of *Gleditsia sinensis* Lam. by high performance liquid chromatography," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 75, pp. 41–46, 2013.

[2] H.-L. Jian, L.-W. Zhu, W.-M. Zhang, D.-F. Sun, and J.-X. Jiang, "Enzymatic production and characterization of manno-oligosaccharides from *Gleditsia sinensis* galactomannan gum," *International Journal of Biological Macromolecules*, vol. 55, pp. 282–288, 2013.

[3] J.-M. Yi, J.-S. Park, S.-M. Oh et al., "Ethanol extract of *Gleditsia sinensis* thorn suppresses angiogenesis in vitro and in vivo," *BMC Complementary and Alternative Medicine*, vol. 12, article 243, 2012.

[4] W. Wang, B. Vinocur, and A. Altman, "Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance," *Planta*, vol. 218, no. 1, pp. 1–14, 2003.

[5] J. R. Witcombe, P. A. Hollington, C. J. Howarth, S. Reader, and K. A. Steele, "Breeding for abiotic stresses for sustainable agriculture," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1492, pp. 703–716, 2008.

[6] S. Mahajan and N. Tuteja, "Cold, salinity and drought stresses: an overview," *Archives of Biochemistry and Biophysics*, vol. 444, no. 2, pp. 139–158, 2005.

[7] J.-K. Zhu, "Cell signaling under salt, water and cold stresses," *Current Opinion in Plant Biology*, vol. 4, no. 5, pp. 401–406, 2001.

[8] A. Matsui, J. Ishida, T. Morosawa et al., "Arabidopsis transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array," *Plant and Cell Physiology*, vol. 49, no. 8, pp. 1135–1149, 2008.

[9] G. Zeller, S. R. Henz, C. K. Widmer et al., "Stress-induced changes in the *Arabidopsis thaliana* transcriptome analyzed using whole-genome tiling arrays," *Plant Journal*, vol. 58, no. 6, pp. 1068–1082, 2009.

[10] X. Lin, J. Zhang, Y. Li et al., "Functional genomics of a living fossil tree, Ginkgo, based on next-generation sequencing technology," *Physiologia Plantarum*, vol. 143, no. 3, pp. 207–218, 2011.

[11] K. Tanase, C. Nishitani, H. Hirakawa et al., "Transcriptome analysis of carnation (*Dianthus caryophyllus* L.) based on next-generation sequencing technology," *BMC Genomics*, vol. 13, no. 1, article 292, 2012.

[12] S. Alsford, D. J. Turner, S. O. Obado et al., "High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome," *Genome Research*, vol. 21, no. 6, pp. 915–924, 2011.

[13] J. Halvardson, A. Zaghlool, and L. Feuk, "Exome RNA sequencing reveals rare and novel alternative transcripts," *Nucleic Acids Research*, vol. 41, no. 1, article e6, 2013.

[14] L. Xiang, Y. Li, Y. Zhu et al., "Transcriptome analysis of the Ophiocordyceps sinensis fruiting body reveals putative genes involved in fruiting body development and cordycepin biosynthesis," *Genomics*, vol. 103, no. 1, pp. 154–159, 2014.

[15] C.-Y. Shi, H. Yang, C.-L. Wei et al., "Deep sequencing of the Camellia sinensis transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds," *BMC Genomics*, vol. 12, article 131, 2011.

[16] "Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1. 21) [Software]," https://github.com/najoshi/sickle.

[17] M. G. Grabherr, B. J. Haas, M. Yassour et al., "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nature Biotechnology*, vol. 29, no. 7, pp. 644–652, 2011.

[18] B. J. Haas, A. Papanicolaou, M. Yassour et al., "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis," *Nature Protocols*, vol. 8, no. 8, pp. 1494–1512, 2013.

[19] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, article 323, 2011.

[20] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research," *Bioinformatics*, vol. 21, no. 18, pp. 3674–3676, 2005.

[21] R. L. Tatusov, D. A. Natale, I. V. Garkavtsev et al., "The COG database: new developments in phylogenetic classification of proteins from complete genomes," *Nucleic Acids Research*, vol. 29, no. 1, pp. 22–28, 2001.

[22] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The KEGG resource for deciphering the genome," *Nucleic Acids Research*, vol. 32, pp. D277–D280, 2004.

[23] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa, "KAAS: an automatic genome annotation and pathway reconstruction server," *Nucleic Acids Research*, vol. 35, no. 2, pp. W182–W185, 2007.

[24] D. Szklarczyk, A. Franceschini, M. Kuhn et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, no. 1, pp. D561–D568, 2011.

[25] R. M. Ferreira, J. L. Rybarczyk-Filho, R. J. S. Dalmolin et al., "Preferential duplication of intermodular hub genes: an evolutionary signature in eukaryotes genome networks," *PLoS ONE*, vol. 8, no. 2, Article ID e56579, 2013.

[26] M. Naika, K. Shameer, O. K. Mathew, R. Gowda, and R. Sowdhamini, "STIFDB2: an updated version of plant stress-responsive transcription factor database with additional stress signals, stress-responsive transcription factor binding sites and stress-responsive genes in arabidopsis and rice," *Plant and Cell Physiology*, vol. 54, no. 2, article e8, 2013.

[27] R. Sowdhamini, K. Shameer, S. Ambika, S. M. Varghese, N. Karaba, and M. Udayakumar, "STIFDB *Arabidopsis* stress responsive transcription factor dataBase," *International Journal of Plant Genomics*, vol. 2009, Article ID 583429, 8 pages, 2009.

[28] A. S. Sundar, S. M. Varghese, K. Shameer, N. Karaba, M. Udayakumar, and R. Sowdhamini, "STIF: identification of stress-upregulated transcription factor binding sites in Arabidopsis thaliana," *Bioinformation*, vol. 2, no. 10, pp. 431–437, 2008.

[29] Y. Peng, X. Gao, R. Li, and G. Cao, "Transcriptome sequencing and De Novo analysis of Youngia japonica using the illumina platform," *PLoS ONE*, vol. 9, no. 3, Article ID e90636, 2014.

[30] N. K. Izzah, J. Lee, M. Jayakodi et al., "Transcriptome sequencing of two parental lines of cabbage (*Brassica oleracea* L. var. *capitata* L.) and construction of an EST-based genetic map," *BMC Genomics*, vol. 15, article 149, 2014.

[31] J. Reese, M. K. Christenson, N. Leng et al., "Characterization of the Asian citrus psyllid transcriptome," *Journal of Genomics*, vol. 2, pp. 54–58, 2014.

[32] R. A. George, J. Y. Liu, L. L. Feng, R. J. Bryson-Richardson, D. Fatkin, and M. A. Wouters, "Analysis of protein sequence and interaction data for candidate disease gene prediction," *Nucleic Acids Research*, vol. 34, no. 19, article e130, 2006.

[33] J. Xu and Y. Li, "Discovering disease-genes by topological features in human protein-protein interaction network," *Bioinformatics*, vol. 22, no. 22, pp. 2800–2805, 2006.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 239462, 11 pages
http://dx.doi.org/10.1155/2014/239462

Research Article

Identification of *Influenza A*/H7N9 Virus Infection-Related Human Genes Based on Shortest Paths in a Virus-Human Protein Interaction Network

Ning Zhang,[1] Min Jiang,[2] Tao Huang,[3,4] and Yu-Dong Cai[4]

[1] *Department of Biomedical Engineering, Tianjin University, Tianjin Key Lab of BME Measurement, Tianjin 300072, China*
[2] *State Key Laboratory of Medical Genomics, Institute of Health Sciences, Shanghai Jiaotong University School of Medicine and Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200025, China*
[3] *Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York City, NY, USA*
[4] *Institute of Systems Biology, Shanghai University, 99 Shangda Road, Shanghai 200444, China*

Correspondence should be addressed to Tao Huang; tohuangtao@126.com and Yu-Dong Cai; cai_yud@126.com

Received 10 March 2014; Revised 18 April 2014; Accepted 21 April 2014; Published 18 May 2014

Academic Editor: Lei Chen

The recently emerging *Influenza A*/H7N9 virus is reported to be able to infect humans and cause mortality. However, viral and host factors associated with the infection are poorly understood. It is suggested by the "guilt by association" rule that interacting proteins share the same or similar functions and hence may be involved in the same pathway. In this study, we developed a computational method to identify *Influenza A*/H7N9 virus infection-related human genes based on this rule from the shortest paths in a virus-human protein interaction network. Finally, we screened out the most significant 20 human genes, which could be the potential infection related genes, providing guidelines for further experimental validation. Analysis of the 20 genes showed that they were enriched in protein binding, saccharide or polysaccharide metabolism related pathways and oxidative phosphorylation pathways. We also compared the results with those from human rhinovirus (HRV) and respiratory syncytial virus (RSV) by the same method. It was indicated that saccharide or polysaccharide metabolism related pathways might be especially associated with the H7N9 infection. These results could shed some light on the understanding of the virus infection mechanism, providing basis for future experimental biology studies and for the development of effective strategies for H7N9 clinical therapies.

1. Introduction

Influenza is one of the most dangerous contagions worldwide and is still a serious global health threat. In the spring of 2013, a novel *Influenza A* virus subtype H7N9 (A/H7N9) broke out in China and quickly spread to other countries [1–3]. As of 11 August 2013, 136 human infections had been laboratory-confirmed, with 44 deaths.

The *Influenza A* viruses (IAVs) are classified into subtypes according to a combination of 16 hemagglutinin (HA: H1–H16) and 9 neuraminidase (NA: N1–N9) surface antigens [4]. Genomic signature and protein sequence analyses revealed that the genes of this A/H7N9 virus were of avian origin [5–7]. The six internal genes were derived from the avian *Influenza A*/H9N2 strain, whereas the haemagglutinin (HA) and neuraminidase (NA) gene segments were from viruses of domestic duck or wild birds [2, 3, 8].

Generally, most avian influenza viruses (e.g., subtypes H5N1, H9N2, H7N7, and H7N3) are of low pathogenicity [4], possibly because avian viruses are inefficient at binding to sialic acid receptors located in human upper airways [5]. However, by comparison, the novel reassortant A/H7N9 seems to cross species from poultry to human more easily [5]. The recombinant has mutations in the hemagglutinin protein, which is associated with potentially enhanced ability to bind to human-like receptors. A deletion in the viral neuraminidase stalk may be also responsible for the change in viral tropism to the respiratory tract or for enhanced

viral replication. Mammalian adaptation mutations are also observed in the polymerase basic 2 (PB2) gene of the virus [2, 9]. These are thought to be correlated with the increased virulence and the better adaptation to mammals of A/H7N9 than other avian influenza viruses [10].

No vaccine for the prevention of A/H7N9 infections is currently available [11]. Although preventing further spread of the infection is important, new drug and vaccine development are also vitally needed for the antiviral treatment. However, viral and host factors associated with the infection of this reassortant are poorly understood [5], which is an obstacle to fight against H7N9. The difficulty is increased by the unusual characteristics from hallmark mutations in the virus, differing from other avian IAVs. Therefore, it is meaningful to identify H7N9 infection-related human genes, which could be used as biomarkers for early diagnosis and targets for new drug development.

In the present study, we proposed a new method for identifying H7N9 infection-related human genes based on a protein-protein interaction (PPI) network. So far the PPI data have been widely used for gene function predictions. The "guilt by association" rule, which was first proposed by Nabieva et al. [12], suggests that interacting proteins share the same or similar functions and hence may be involved in the same pathway. This assumption can be used to identify disease-related genes from existing protein-protein interaction networks. In our previous studies, based on this assumption, we have identified genes related to other diseases, such as the ones mentioned in [13–15].

Shortest path and betweenness method are widely used to identify and analyze biomarkers on virus-host interaction networks [16–18]. If one protein is on many shortest paths between virus target genes, it has great betweenness and it can disrupt the signal transduction on the network [19, 20]. It was found that proteins with great betweenness usually have similar functions with the original seed genes [13, 21]. In this study, we used this method to identify potential host response genes to the A/H7N9 virus infection.

2. Materials and Methods

The overall procedure of our method is illustrated in Figure 1. In the following subsections, details are presented.

2.1. Dataset Construction of Target Human Proteins. The course of the *Influenza A*/H7N9 infection can be determined by comprehensive protein-protein interactions (PPIs) between the virus and its host (human). In this study, whether a human protein interacted with virus proteins was determined based on the Gene Ontology (GO) database. The Gene Ontology (GO) terms provide information about the biological process, molecular function, and cellular component of a specific protein. A human protein and a protein of H7N9 having at least 1 sharing GO term were assumed to interact with each other and the human protein was called target human protein. Since protein pairs sharing generic GO terms should be ignored, in this study, only GO terms at levels below 3 were considered. That is to say, we excluded the root GO terms ("GO:0008150: biological_process", "GO:0005575: cellular_component", "GO:0003674: molecular_function"), their children, and the children of their children terms. Based on this rule, we constructed a dataset of target human proteins. The detailed description of the procedure was presented below.

All protein sequences of the *Influenza A*/H7N9 virus were downloaded from NCBI protein database (http://www.ncbi.nlm.nih.gov/). After removing those with sequence identities >40%, only 11 proteins were left and were listed in Supporting Information S1 available online at http://dx.doi.org/10.1155/2014/239462. The Gene Ontology (GO) terms at levels below 3 of the 11 proteins were mapped by InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/) [22]. All human proteins and their Protein-GO term mappings were obtained from biomart in ENSEMBL (http://asia.ensembl.org/biomart/martview/).

Based on the rule of sharing GO terms, 3,212 target human proteins (coded by 1,023 human genes) were picked out, each of which interacted with at least 1 H7N9 virus protein. These virus-human protein pairs were provided in Supporting Information S2, together with the sharing GO terms for each pair. And we summarized the 3,212 target human proteins with their 1,023 related coding genes in Supporting Information S3.

2.2. PPI Data from STRING. STRING (Search Tool for the Retrieval of Interacting Genes) (http://string.embl.de/) [23] is an online database resource which compiles both experimental and predicted protein-protein interactions with a confidence score to quantify each interaction confidence. A weighted PPI network can be retrieved from STRING, in which proteins in the network are represented as nodes, while interactions between proteins are given as edges marked with confidence scores if they are in interaction with each other. Interacting proteins with high confidence scores in such a PPI network are more likely to share similar biological functions than noninteractive ones [23–25]. This is because the protein and its interactive neighbours may form a protein complex performing a particular function or may be involved in the same pathway.

We constructed a graph G with the PPI data from STRING (version 9.0). In such a graph, proteins were represented as nodes; however, the weight of each interaction edge was assigned a d value rather than a confidence score (s). The d value was derived from the confidence score s according to the equation $d = 1000 \times (1 - s)$. Thus, the d value can be considered as representing protein distances to each other: the smaller the distance, the higher the interaction confidence score and the more similar the functions they have.

In this study, we analyzed in such a graph every two protein interactions in the target human protein dataset.

2.3. Shortest Path Tracing. The Dijkstra algorithm [26] were used to find the shortest paths in the graph G between every two proteins in the target human protein dataset, that is, the shortest paths between each of the 3,212 proteins to all the other 3,211 proteins in the graph. The Dijkstra algorithm was

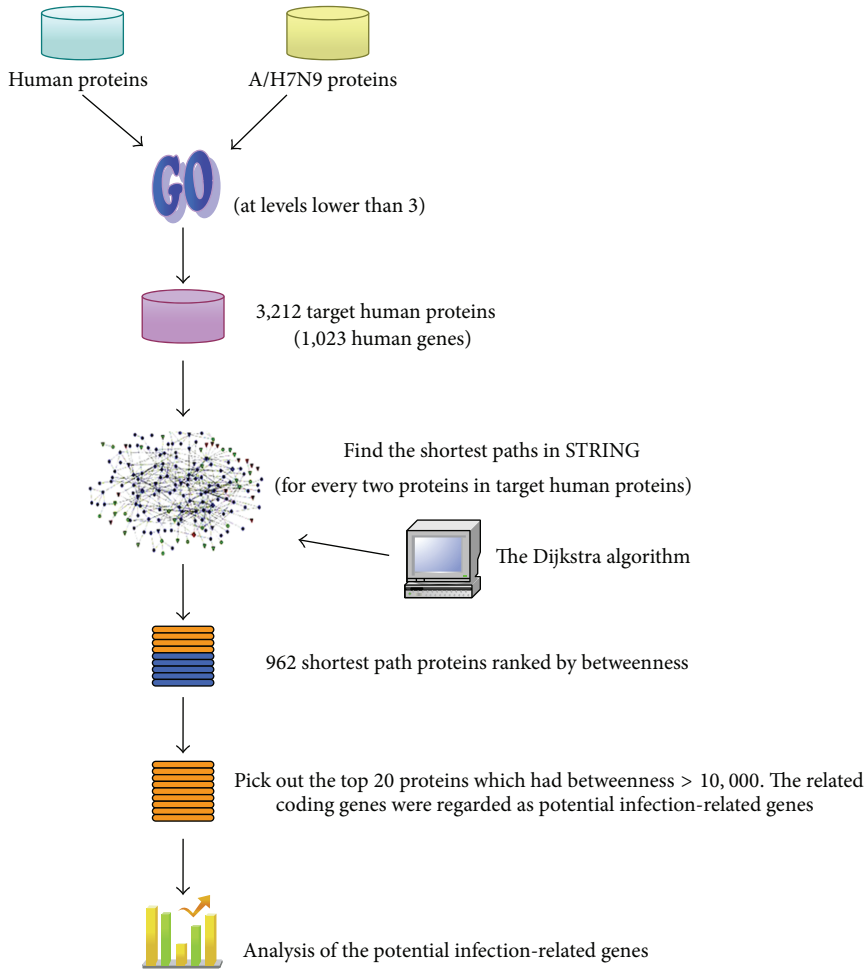


Figure 1: The flowchart of the method developed in this study to identify the *Influenza A*/H7N9 infection-related human genes. Target human proteins interacting with the *Influenza A*/H7N9 virus were obtained based on sharing GO terms. Shortest path proteins were calculated from the shortest paths between every pair of the target human proteins, by searching by the Dijkstra algorithm in the network constructed from STRING. Finally 20 shortest path proteins were screened out with betweenness >10,000, the related genes of which were considered as infection-related human genes.

implemented with R package "igraph" [27] (no parameters needed to be set in this algorithm).

Then we get all proteins existing on the shortest paths (962 proteins, called Shortest Path Proteins) and ranked these proteins according to their betweenness. Results can be found in Supporting Information S4. The top 20 proteins (20 genes) with betweenness over 10,000 were picked out and the 20 corresponding coding genes were regarded as potential H7N9 infection-related human genes.

2.4. KEGG Pathway Enrichment Analysis. The functional annotation tool DAVID [28] was used for KEGG pathway enrichment analysis (all parameters were selected as default). The enrichment P value was corrected to control family-wide false discovery rate under a certain rate (e.g., ≤ 0.05) with the Benjamin multiple testing correction method [29]. All human protein-coding genes were regarded as background during the enrichment analysis.

2.5. Comparison with Another Two Species of Viruses. To further understand the *Influenza A*/H7N9-human interaction, we compared the results of the potential H7N9 infection-related human genes obtained above with those identified from another two species of viruses: human rhinovirus (HRV) and respiratory syncytial virus (RSV), which are also causing human acute respiratory infections. The same procedure of our method presented above was performed on the two species of viruses as that on H7N9.

All protein sequences of HRV and RSV viruses were downloaded from NCBI protein database (http://www.ncbi.nlm.nih.gov/). After removing those with sequence identities >40%, the proteins left were listed in Supporting Information S5, S6, respectively. The virus-human protein pairs were also provided in Supporting Information S2, and the target human proteins with their coding genes were also summarized in Supporting Information S3. 1,904 and 9,846

Table 1: Number of proteins/genes in our datasets of the three species of viruses: *Influenza A*/H7N9 virus, human rhinovirus (HRV), and respiratory syncytial virus (RSV).

Virus	Virus proteins (sequence identity <40%)	Virus-human protein pairs	Target human proteins (coding genes)	Shortest path proteins	Betweenness threshold	Potential infection-related proteins (coding genes)
H7N9	11	9,313	3,212 (1,023)	962	>10,000	20 (20)
HRV	4	20,955	6,985 (2,028)	1,904	>47,299	11 (11)
RSV	22	40,499	36,273 (11,036)	9,846	>1,275,672	44 (44)

shortest path proteins were obtained from HRV and RSV virus, respectively, after computing shortest paths, given also in Supporting Information S4. The numbers of proteins and genes for the three species of viruses at each step were summarized in Table 1.

We selected betweenness threshold as 10,000 for the shortest path proteins of H7N9. However, the threshold should be different for the other two species of viruses since the numbers of target human proteins were different. We standardized the betweenness threshold for HRV and RSV viruses on that for H7N9 virus in this study.

Shortest paths were computed on every two proteins in target human protein dataset. Denoting the number of target human proteins as N, the number of shortest paths was C_N^2. The average threshold w was calculated on H7N9 as

$$w = \frac{10{,}000}{C^2_{N_{\text{H7N9}}}} = \frac{10{,}000}{C^2_{3212}} = 0.001939. \quad (1)$$

Then the betweenness threshold for HRV and RSV was determined by $wC^2_{N_{\text{HRV}}} = wC^2_{6985} = 47{,}299$ and $wC^2_{N_{\text{RSV}}} = wC^2_{36273} = 1{,}275{,}672$, respectively. Therefore, the top 11 proteins (11 genes) were picked out for HRV (betweenness > 47,299) while the top 44 proteins (44 genes) were picked out for RSV (betweenness > 1,275,672) from the lists in Supporting Information S4, respectively. And the corresponding 11 and 44 coding genes were regarded as potential infection-related human genes for HRV and RSV virus, respectively. The betweenness threshold and the numbers of proteins picked out were also summarized in Table 1.

3. Results and Discussion

3.1. Sharing GO Terms between H7N9 Proteins and Human Proteins. H7N9 and human proteins with at least 1 sharing GO term were considered as interacting with each other. 3,212 target human proteins were found as interacting with H7N9 proteins based on this rule. The same procedure was performed on the other two species of viruses for comparison: HRV and RSV. Types of the sharing GO terms and the share numbers of the terms could give information about the interaction between the virus and its host. Thus, a statistical analysis was made on the sharing GO terms from Supporting Information S2 for each species of virus, respectively. Results were depicted in Figure 2.

From Figure 2, it can be seen that the sharing GO terms and their numbers were apparently different between the three species of viruses, indicating specific properties and different interactions with host during infections.

For H7N9, the term "GO:0003723|RNA binding" accounted for the most, indicating important roles of RNA binding proteins in the PPI interactions between H7N9 and human, which was consistent with the observations in *Influenza A* viruses in the literature [30–32]. As shown in Figure 2, H7N9 and HRV both fell into the significant term "GO:0003723|RNA binding," indicating that RNA binding was essential between virus-human proteins during the infection of the two viruses. However, RSV was not presented in such a term. It was possibly suggested that H7N9 and HRV had such a specificity that could be different from RSV, although all the three are RNA viruses. Several other GO terms indicated specific and important virus-human protein interactions for H7N9 infection, such as "GO:0005975|carbohydrate metabolic process," "GO:0015078|hydrogen ion transmembrane transporter activity," and "GO:0015992|proton transport."

Nevertheless, 3 terms of H7N9 were the same as those of HRV ("GO:0003723|RNA binding," "GO:0019079|viral genome replication," and "GO:0003968|RNA-directed RNA polymerase activity"), and 2 terms as RSV ("GO:0003968|RNA-directed RNA polymerase activity," "GO:0019031|viral envelope"), indicating similar processes of the infections between the three viruses.

3.2. Potential H7N9 Infection-Related Genes. The shortest paths were calculated between each pair of the 3,212 proteins. All proteins were picked out with their betweenness from the shortest paths, given in Supporting Information S4. We selected the top 20 proteins with betweenness over 10,000 and ranked them according to their betweenness. The related coding genes of the 20 proteins were also retrieved accordingly (20 genes). These were shown in Table 2. The 20 genes were regarded as potential H7N9 infection-related human genes in this study. Results of potential infection-related human genes for HRV and RSV were also listed in Table 2 by the same method as that for H7N9 for comparison. Note that the proteins (genes) listed in Table 2 were all human proteins (genes), not virus. Potential human genes found for the three viruses were also depicted in Figure 3. It was clearly seen from Figure 3 that the potential human genes found were remarkably different in H7N9 infection as compared with those in HRV and RSV, although several sharing genes existed. Thus, these 20 human genes could be closely related to the H7N9 infections. Our further analysis was based on these 20 genes.

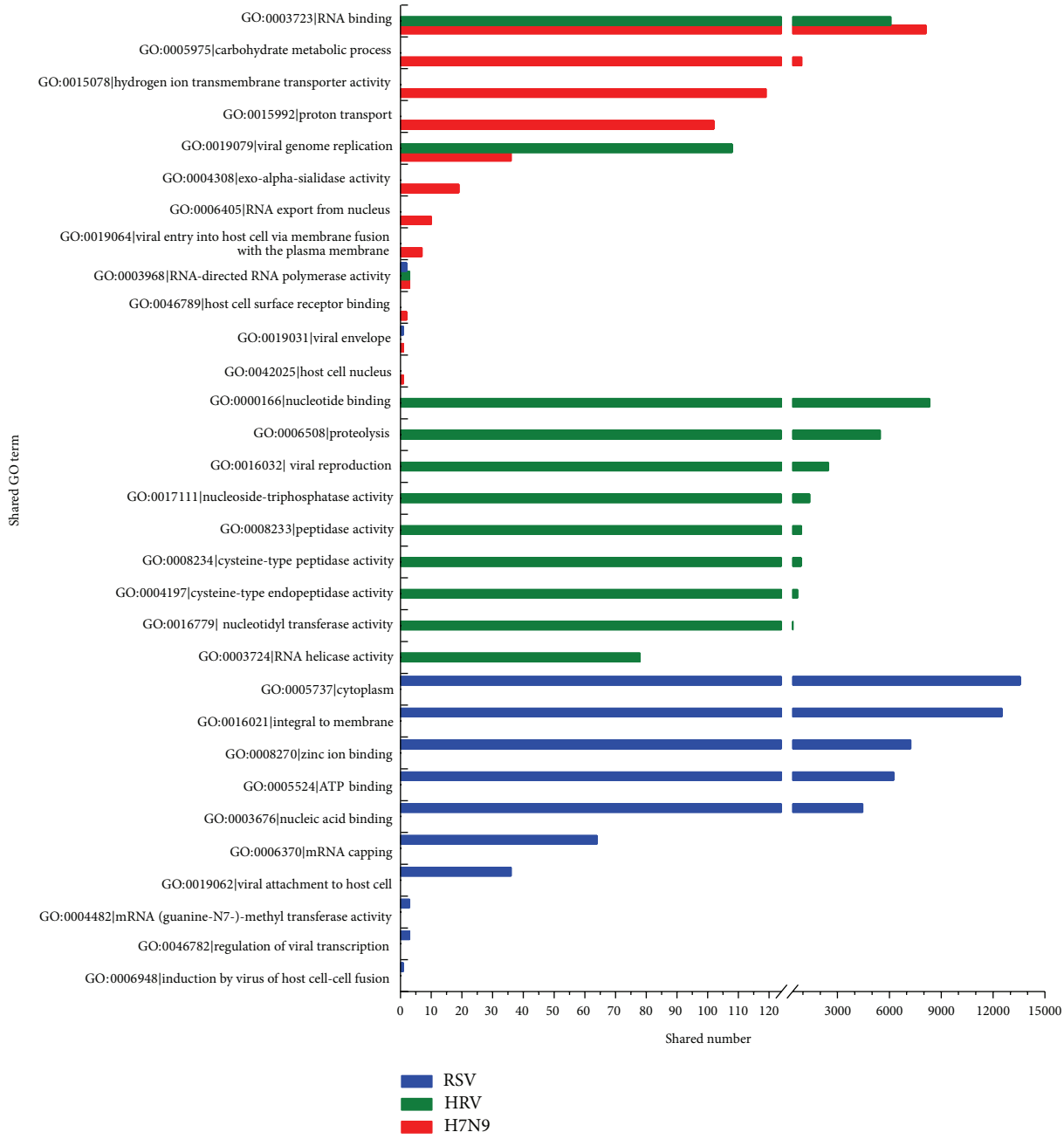


FIGURE 2: Statistical analysis of sharing GO terms between virus proteins and human proteins. All sharing GO terms and their descriptions between virus and human proteins were listed on the Y-axis. Histogram of sharing numbers showed the instances of each term used as a sharing term. The horizontal axis was truncated from 125 to 400.

The 20 human genes were submitted to the CCSB interactome database to analyze their interactions with viruses (http://interactome.dfci.harvard.edu/V_hostome/). Among them, proteins encoded by RANBP2 and GYS1were found to be related to EBV or HPV proteins, such as EBV-BVLF1, EBV-BGLF3, and HPV8-E6. These proteins could also have some relationship with H7N9 infections.

Among the 20 genes, some, such as GAPDH and NXF1, had been well documented to be relevant to H7N9 infections. However, there were also other genes with rare previous association with H7N9 infections reported or that had been only poorly characterized, such as PGK1, GYS1, YBX1, and NUP214.

GAPDH (glyceraldehyde-3-phosphate dehydrogenase) is a housekeeping gene in carbohydrate metabolism. This finding was consistent with the general agreement that GAPDH is an important gene and is widely used in the studies of host gene response to virus infections, including influenza virus infections [33–35].

NXF1 (nuclear export factor 1) is one member of a family of nuclear RNA export factor genes. It was reported that viral mRNAs of *Influenza A* virus were transported to

Table 2: The infection-related human proteins and their related coding genes for the three species of viruses calculated from shortest paths in a PPI network.

Infected by virus	Protein	Gene	Chromosome	Betweenness
Influenza A/H7N9	ENSP00000361626	YBX1	1	58376
	ENSP00000363676	RPL11	1	49155
	ENSP00000396127	RAN	12	49036
	ENSP00000362413	PGK1	X	26743
	ENSP00000229239	GAPDH	12	25345
	ENSP00000294172	NXF1	11	23550
	ENSP00000352400	NUP214	9	21849
	ENSP00000280892	EIF4E	4	20883
	ENSP00000379933	TPI1	12	19217
	ENSP00000348877	GPI	19	18885
	ENSP00000317904	GYS1	19	18349
	ENSP00000350283	BRCA1	17	16827
	ENSP00000283195	RANBP2	2	16823
	ENSP00000400591	SNRPE	1	15796
	ENSP00000265686	TCIRG1	11	14465
	ENSP00000358563	DKC1	X	13535
	ENSP00000234396	ATP6V1B1	2	13432
	ENSP00000218516	GLA	X	12471
	ENSP00000262030	ATP5B	12	11891
	ENSP00000260947	BARD1	2	11297
Human Rhinovirus (HRV)	ENSP00000344818	UBC	12	330154
	ENSP00000363676	RPL11	1	154993
	ENSP00000361626	YBX1	1	136548
	ENSP00000357879	PSMD4	1	121991
	ENSP00000337825	LCK	1	117195
	ENSP00000396127	RAN	12	116059
	ENSP00000348461	RAC1	7	111632
	ENSP00000230354	TBP	6	100485
	ENSP00000350283	BRCA1	17	65076
	ENSP00000314949	POLR2A	17	54470
	ENSP00000280892	EIF4E	4	50359
Respiratory syncytial virus (RSV)	ENSP00000269305	TP53	17	15809765
	ENSP00000344456	CTNNB1	3	5756301
	ENSP00000263253	EP300	22	5694027
	ENSP00000339007	GRB2	17	5591895
	ENSP00000275493	EGFR	7	5245421
	ENSP00000270202	AKT1	14	4663263
	ENSP00000264657	STAT3	17	4180564
	ENSP00000350941	SRC	20	3180369
	ENSP00000348461	RAC1	7	3066312
	ENSP00000221494	SF3A2	19	2994393
	ENSP00000417281	MDM2	12	2686189
	ENSP00000338345	SNCA	4	2647616
	ENSP00000206249	ESR1	6	2643164
	ENSP00000296271	RHO	3	2573058
	ENSP00000329623	BCL2	18	2541856
	ENSP00000376609	GRK5	10	2364221

Table 2: Continued.

Infected by virus	Protein	Gene	Chromosome	Betweenness
Respiratory Syncytial Virus (RSV)	ENSP00000337825	LCK	1	2306232
	ENSP00000314458	CDC42	1	2174421
	ENSP00000262613	SLC9A3R1	17	2097178
	ENSP00000355865	PARK2	6	2033100
	ENSP00000264033	CBL	11	1930392
	ENSP00000269571	ERBB2	17	1922027
	ENSP00000338018	HIF1A	14	1915325
	ENSP00000324806	GSK3B	3	1910676
	ENSP00000215832	MAPK1	22	1831541
	ENSP00000358490	CD2	1	1751073
	ENSP00000262160	SMAD2	18	1727787
	ENSP00000304903	CD2BP2	16	1714523
	ENSP00000362649	HDAC1	1	1703720
	ENSP00000353483	MAPK8	10	1702626
	ENSP00000261799	PDGFRB	5	1679113
	ENSP00000003084	CFTR	7	1662248
	ENSP00000401303	SHC1	1	1548773
	ENSP00000321656	CDC25C	5	1521621
	ENSP00000357656	FYN	6	1503978
	ENSP00000326366	PSEN1	14	1498004
	ENSP00000230354	TBP	6	1458835
	ENSP00000300093	PLK1	16	1444680
	ENSP00000350283	BRCA1	17	1389799
	ENSP00000228307	PXN	12	1358706
	ENSP00000329357	SP1	12	1347630
	ENSP00000361626	YBX1	1	1342956
	ENSP00000387662	GCG	2	1321174
	ENSP00000367207	MYC	8	1284185

the cytoplasm by the NXF1 pathway for translation of viral proteins [36]. Not surprisingly, the H7N9 virus exploited the same pathway.

YBX1 (Y box binding protein 1) has been found to be an interacting partner of genomic RNA of Hepatitis C Virus, which negatively regulates the equilibrium between viral translation/replication and particle production [37]. NUP214 (nucleoporin 214 kDa) encodes one of nucleoporins composing the nuclear pore complex (NPC), which forms a gateway regulating the flow of macromolecules between nucleus and cytoplasm. Many viruses have been reported to require these mechanisms to deliver their genomes into the host cell nucleus for replication, such as human immunodeficiency virus (HIV) [38], encephalomyocarditis virus [39], and herpes simplex virus [40]. However, reports on NUP214, YBX1 related to *Influenza A* viruses, were sparse.

Cancer-related genes were also included. BRCA1 (breast cancer 1) encodes a nuclear phosphoprotein that plays a role in maintaining genomic stability, and it also acts as a tumor suppressor. BARD1 (BRCA1 associated RING domain 1) encodes a protein which interacts with the N-terminal region of BRCA1, regulating cell growth and the products of tumor suppressor genes, and may be related to breast or ovarian cancer.

Interestingly, more genes were involved in energy pathways containing glycolysis and gluconeogenesis, such as GPI (glucose-6-phosphate isomerase), PGK1 (phosphoglycerate kinase 1), and TPI1 (triosephosphate isomerase 1). In addition, GYS1 (glycogen synthase 1) encodes a protein catalyzing the addition of glucose monomers to the growing glycogen molecule in starch and sucrose metabolism. GLA (galactosidase) encodes a glycoprotein that hydrolyses the terminal alpha-galactosyl moieties from glycolipids and glycoproteins. Therefore, it was suggested that the H7N9 infection could be probably linked to saccharide or polysaccharide metabolism related pathways. Central metabolism could be strongly affected by virus infections [41]. Janke et al. [42] also found changes in metabolism in cells infected by *Influenza A*/H1N1 virus, suggesting that fatty acid synthesis might play a crucial role for the virus replication as they acquired lipid.

ATP6V1B1 (ATPase, H+ transporting, lysosomal 56/58 kDa, V1 subunit B1) and ATP5B (ATP synthase, H+ transporting, mitochondrial F1 complex, beta polypeptide) were involved in ATP synthase and hydrolysis.

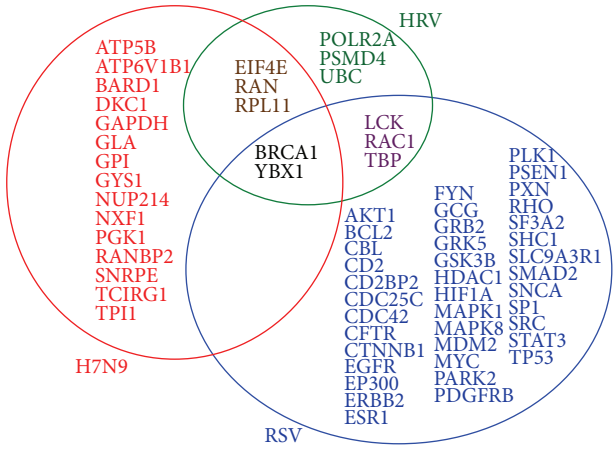


Figure 3: The potential virus infection-related human genes found based on our method for the three species of viruses. 20, 11, and 44 potential infection-related human genes were found for H7N9, HRV, and RSV, respectively. There were 5 sharing genes between those for H7N9 and HRV, and 2 sharing genes between H7N9 and RSV. Other human genes related were not all the same, indicating specific properties or particular characteristics between the infections of the three species of viruses.

From Table 2, it also can be seen that although several genes (PGK1, DKC1, was GLA) were located on Chromosome X, none on Chromosome Y was found in this study. Although earlier findings reported that H7N9 infections preferentially occurred in males, it was suggested from our findings that it may not be so significant. This was also consistent with results of Chen et al.'s work [43], in which they indicated that it did not show any statistically significant differences in clinical outcomes between genders from their logistic regression analysis.

3.3. GO Enrichment Analysis of H7N9 Infection-Related Genes. We performed GO enrichment analysis on these 20 genes. The 20 proteins encoded by the genes were mapped to GO terms on the levels below 3 from Gene Ontology. Totally 504 GO terms were obtained. GO enrichment analysis was performed on these terms. The GO terms and the number of proteins related to each GO term were shown in Table 3. The same procedure was performed on the other two species of viruses for comparison, with results shown in Table 3. Both commonness and differences of GO term enrichment between the three species of viruses existing as described in Table 3.

Form Table 3, it can be seen that 15 out of the 20 H7N9, all the 11 HRV, and 42 out of the 44 RSV infection-related proteins were involved in protein binding (GO:0005515). Protein binding played important roles in both virus infection and host immune responses [44]. This could partially explain why the novel reassortant had more enhanced ability to bind to human receptors than other avian influenza viruses [2, 10]. The recombinant proteins could also induce immune responses via protein interactions [45]. Once the host immune system activated, patients would have severe symptoms, such as cough, sputum, fever, and shortness of breath. Many related proteins of the three viruses fell into GO terms "GO:0005829 cytosol" and "GO:0005737 cytoplasm," since all the three viruses are RNA viruses and replication of RNA viruses usually takes place in cytoplasm.

These were commonness. However, differences or specific characteristics still exist in H7N9-related proteins from those of other two viruses.

Nine of these proteins were enriched in "GO:0044281 small molecule metabolic process" (45.00%) for H7N9, whereas only 1 (9.09%) and 2 (4.55%) proteins were enriched in this term for HRV and RSV, respectively. Furthermore, still many related proteins of H7N9 enriched in "GO:0005975 carbohydrate metabolic process," "GO:0006006 glucose metabolic process," "GO:0006094 gluconeogenesis," and "GO:0006096 glycolysis," differing from those cases of HRV or RSV. These specific enrichment of GO terms indicated that the H7N9 infection could be especially relevant with human saccharide or polysaccharide metabolism-related pathways.

For H7N9, 3 proteins fell into the term "GO:0015991 ATP hydrolysis coupled proton transport" and 3 proteins into "GO:0015992 proton transport," but it was not the case for HRV or RSV. Proteins involved in "GO:0005215 transporter activity" and "GO:0055085 transmembrane transport" were also different between the H7N9 infections and the other two viruses.

3.4. KEGG Pathway Enrichment Analysis. KEGG pathway enrichment analysis was also performed on the 20 genes. The KEGG pathway terms and the number of proteins belonging to each pathway term were shown in Table 4.

Only 3 pathways were retrieved. However, all the 3 pathways were specially related to H7N9; that is, none of the 3 pathways appeared in the KEGG results of the other two viruses (data not shown of the KEGG results for the other two viruses).

Form Table 4, it can be seen that 2 out of the 3 pathways were saccharide or polysaccharide metabolism-related pathways ("Glycolysis/Gluconeogenesis" and "Starch and sucrose metabolism"), suggesting that these types of pathways could play pivotal roles in the H7N9 infections. Another pathway involved was "oxidative phosphorylation." This pathway could also be important, but it may not so as the former two, since genes involved in this pathway (ATP5B, ATP6V1B1, and TCIRG1) were ranked at the bottom in the gene list in Table 2 according to betweenness.

4. Conclusion

In this study, we developed a computational method to identify *Influenza A*/H7N9 infection-related human genes based on the shortest paths in a PPI network. Finally, 20 human genes were screened out which could be the most significant, providing guidelines for further experimental validation. Among the genes, several ones such as PGK1, GYS1, YBX1, and NUP214 were previously reported with rare association with influenza virus infections or had been only poorly

Table 3: GO term enrichment analysis of the 20 potential H7N9 infection-related human proteins (data not shown of GO terms with number of related proteins below 3) and comparisons with HRV and RSV for these terms.

GO terms	H7N9		HRV		RSV	
	Number of proteins	Percentage accounting for the 20 proteins (%)	Number of proteins*	Percentage accounting for the 11 proteins (%)*	Number of proteins*	Percentage accounting for the 44 proteins (%)*
GO:0005515 protein binding	15	75.00	11	100.00	42	95.45
GO:0005829 cytosol	13	65.00	7	63.64	23	52.27
GO:0005737 cytoplasm	11	55.00	6	54.55	30	68.18
GO:0005634 nucleus	9	45.00	4	36.36	33	75.00
GO:0044281 small molecule metabolic process	9	45.00	1	9.09	2	4.55
GO:0003723 RNA binding	8	40.00	4	36.36	2	4.55
GO:0005975 carbohydrate metabolic process	8	40.00	—	—	—	—
GO:0005654 nucleoplasm	7	35.00	7	63.64	19	43.18
GO:0010467 gene expression	7	35.00	8	72.73	6	13.64
GO:0006006 glucose metabolic process	5	25.00	—	—	2	4.55
GO:0005886 plasma membrane	4	20.00	4	36.36	27	61.36
GO:0005622 intracellular	4	20.00	3	27.27	7	15.91
GO:0005643 nuclear pore	4	20.00	1	9.09	—	—
GO:0016032 viral reproduction	4	20.00	8	72.73	4	9.09
GO:0016070 RNA metabolic process	4	20.00	4	36.36	1	2.27
GO:0006094 gluconeogenesis	4	20.00	—	—	—	—
GO:0006096 glycolysis	4	20.00	—	—	—	—
GO:0055085 transmembrane transport	4	20.00	—	—	1	2.27
GO:0005625 soluble fraction	3	15.00	—	—	5	11.36
GO:0008270 zinc ion binding	3	15.00	2	18.18	11	25.00
GO:0016071 mRNA metabolic process	3	15.00	4	36.36	1	2.27
GO:0006606 protein import into nucleus	3	15.00	1	9.09	1	2.27
GO:0005524 ATP binding	3	15.00	1	9.09	14	31.82
GO:0006406 mRNA export from nucleus	3	15.00	1	9.09	—	—
GO:0008286 insulin receptor signaling pathway	3	15.00	1	9.09	4	9.09
GO:0005215 transporter activity	3	15.00	—	—	—	—
GO:0015991 ATP hydrolysis coupled proton transport	3	15.00	—	—	—	—
GO:0015992 proton transport	3	15.00	—	—	—	—
GO:0019221 cytokine-mediated signaling pathway	3	15.00	2	18.18	1	2.27

* —: no proteins having the GO term was picked out as potential infection-related proteins for the virus.

Table 4: KEGG pathway enrichment analysis of the 20 potential H7N9 infection-related human genes.

Terms	Genes	Number of genes belonging to the pathway	Percentage accounting for the 20 genes (%)	Adjusted P value (Benjamini)
Glycolysis/Gluconeogenesis	TPI1, GPI, GAPDH, and PGK1	4	20.00	$7.0E-3$
Oxidative phosphorylation	ATP5B, ATP6V1B1, and TCIRG1	3	15.00	$3.3E-1$
Starch and sucrose metabolism	GPI, GYS1	2	10.00	$5.2E-1$

characterized in the literature. Most of the 20 genes were enriched in protein binding, saccharide, or polysaccharide metabolism-related pathways and oxidative phosphorylation pathways, compared to the other two viruses HRV and RSV, suggesting direct or indirect relationship with the formation or development of the infection. These candidate genes may provide clues for further researches and experimental validations. Results from this study may shed some light on the understanding of the virus infection mechanism, providing new references for researches into the disease and for new strategies for antivirals, such as new drug and vaccine development.

Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by Grants from National Basic Research Program of China (2011CB510102, 2011CB510101) and National Natural Science Foundation of China (31371335), Innovation Program of Shanghai Municipal Education Commission (12ZZ087), the grant of "The First-class Discipline of Universities in Shanghai", National Natural Science Foundation of China (81030015, 81171342, and 81201148), Tianjin Research Program of Application Foundation and Advanced Technology (14JCQNJC09500), the National Research Foundation for the Doctoral Program of Higher Education of China (20130032120070, 20120032120073), Independent Innovation Foundation of Tianjin University (60302064, 60302069), and the E-Institutes of Shanghai Municipal Education Commission.

References

[1] Q. Li, L. Zhou, M. Zhou et al., "Preliminary report: epidemiology of the avian influenza A, (H7N9) outbreak in China," *The New England Journal of Medicine*, 2013.

[2] R. Gao, B. Cao, Y. Hu et al., "Human infection with a novel avian-origin infl uenza A, (H7N9) virus," *The New England Journal of Medicine*, vol. 368, no. 20, pp. 1888–1897, 2013.

[3] Y. Chen, W. Liang, S. Yang et al., "Human infections with the emerging avian influenza A H7N9 virus from wet market poultry: clinical analysis and characterisation of viral genome," *The Lancet*, vol. 3819881, pp. 1916–1925, 2013.

[4] A. Nagy, L. Černíková, V. Křivda, and J. Horníčková, "Digital genotyping of avian influenza viruses of H7 subtype detected in central Europe in 2007–2011," *Virus Research*, vol. 165, no. 2, pp. 126–133, 2012.

[5] Y. Hu, S. Lu, Z. Song et al., "Association between adverse clinical outcome in human disease caused by novel infl uenza A H7N9 virus and sustained viral shedding and emergence of antiviral resistance," *The Lancet*, vol. 381, no. 9885, pp. 2273–2279, 2013.

[6] Q. Liu, L. Lu, Z. Sun, G. W. Chen, Y. Wen, and S. Jiang, "Genomic signature and protein sequence analysis of a novel influenza A, (H7N9) virus that causes an outbreak in humans in China," *Microbes and Infection*, vol. 15, no. 6-7, pp. 432–439, 2013.

[7] T. Kageyama, S. Fujisaki, E. Takashita et al., "Genetic analysis of novel avian A(H7N9) influenza viruses isolated from patients in China, February to April 2013," *Eurosurveillance*, vol. 18, no. 15, pp. 20453–10467, 2013.

[8] J. P. Dudley and I. M. Mackayd, "Age-specific and sex-specific morbidity and mortality from avian influenza A(H7N9)," *Journal of Clinical Virology*, vol. 58, pp. 568–570, 2013.

[9] E. K. Subbarao, W. London, and B. R. Murphy, "A single amino acid in the PB2 gene of influenza A virus is a determinant of host range," *Journal of Virology*, vol. 67, no. 4, pp. 1761–1764, 1993.

[10] T. M. Uyeki and N. J. Cox, "Global concerns regarding novel influenza A, (H7N9) virus infections," *The New England Journal of Medicine*, vol. 368, no. 20, pp. 1862–1864, 2013.

[11] L. Mei, P. Song, Q. Tang et al., "Changes in and shortcomings of control strategies, drug stockpiles, and vaccine development during outbreaks of avian influenza A H5N1, H1N1, and H7N9 among humans," *BioScience Trends*, vol. 7, no. 2, pp. 64–76, 2013.

[12] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh, "Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps," *Bioinformatics*, vol. 21, supplement 1, pp. i302–i310, 2005.

[13] M. Jiang, Y. Chen, Y. Zhang et al., "Identification of hepatocellular carcinoma related genes with k-th shortest paths in a protein-protein interaction network," *Molecular BioSystems*, vol. 9, no. 11, pp. 2720–2728, 2013.

[14] B. Q. Li, J. Zhang, T. Huang, L. Zhang, and Y. D. Cai, "Identification of retinoblastoma related genes with shortest path in a proteineprotein interaction network," *Biochimie*, vol. 94, no. 9, pp. 1910–1917, 2012.

[15] B. Q. Li, J. You, L. Chen et al., "Identification of lung-cancer-related genes with the shortest path approach in a protein-protein interaction network," *BioMed Research International*, vol. 2013, Article ID 267375, 8 pages, 2013.

[16] T. Huang, L. Liu, Q. Liu et al., "The role of Hepatitis C Virus in the dynamic protein interaction networks of Hepatocellular cirrhosis and Carcinoma," *International Journal of Computational Biology and Drug Design*, vol. 4, no. 1, pp. 5–18, 2011.

[17] T. Huang, Z. Xu, L. Chen, Y. D. Cai, and X. Kong, "Computational analysis of HIV-1 resistance based on gene expression profiles and the virus-host interaction network," *PLoS ONE*, vol. 6, no. 3, Article ID e17291, 2011.

[18] T. Huang, J. Wang, Y.-D. Cai, H. Yu, and K.-C. Chou, "Hepatitis c virus network based classification of hepatocellular cirrhosis and carcinoma," *PLoS ONE*, vol. 7, no. 4, Article ID e34460, 2012.

[19] T. Huang, P. Wang, Z. Ye et al., "Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties," *PLoS ONE*, vol. 5, no. 7, Article ID e11900, 2010.

[20] Y. Jiang, T. Huang, L. Chen et al., "Signal propagation in protein interaction network during colorectal cancer progression," *BioMed Research International*, vol. 2013, Article ID 287019, 9 pages, 2013.

[21] B.-Q. Li, T. Huang, L. Liu, Y. D. Cai, and K. C. Chou, "Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network," *PLoS ONE*, vol. 7, no. 4, Article ID e33393, 2012.

[22] E. Quevillon, V. Silventoinen, S. Pillai et al., "InterProScan: protein domains identifier," *Nucleic Acids Research*, vol. 33, no. 2, pp. W116–W120, 2005.

[23] D. Szklarczyk, A. Franceschini, M. Kuhn et al., "The STRING database in 2011: Functional interaction networks of proteins,

globally integrated and scored," *Nucleic Acids Research*, vol. 39, no. 1, pp. D561–D568, 2011.

[24] Y. A. I. Kourmpetis, A. D. van Dijk, M. C. Bink, R. C. van Ham, and C. J. Ter Braak, "Bayesian markov random field analysis for protein function prediction based on network data," *PLoS ONE*, vol. 5, no. 2, Article ID e9293, 2010.

[25] K. L. Ng, J. S. Ciou, and C. H. Huang, "Prediction of protein functions based on function-function correlation relations," *Computers in Biology and Medicine*, vol. 40, no. 3, pp. 300–305, 2010.

[26] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[27] G. Csardi and T. Nepusz, The igraph Software Package for Complex Network Research. InterJournal Complex Systems, 2006.

[28] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[29] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

[30] J.-Y. Min and R. M. Krug, "The primary function of RNA binding by the influenza A virus NS1 protein in infected cells: inhibiting the $2'$-$5'$ oligo (A) synthetase/RNase L pathway," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 18, pp. 7100–7105, 2006.

[31] S. Chenavas, T. Crépin, B. Delmas, R. W. Ruigrok, and A. Slama-Schwok, "Influenza virus nucleoprotein: structure, RNA binding, oligomerization and antiviral drug target," *Future Microbiol*, vol. 8, pp. 1537–1545, 2013.

[32] P. L. Tsai, N. T. Chiou, S. Kuss, A. García-Sastre, K. W. Lynch, and B. M. Fontoura, "Cellular RNA binding proteins NS1-BP and hnRNP K regulate influenza A virus RNA splicing," *PLoS Pathog*, vol. 9, no. 6, Article ID e1003460, 2013.

[33] S. V. Kuchipudi, M. Tellabati, R. K. Nelli et al., "18S rRNA is a reliable normalisation gene for real time PCR based on influenza virus infected cells," *Virology Journal*, vol. 8, no. 9, article 230, 2012.

[34] M. R. Barber, J. R. Aldridge Jr., R. G. Webster, and K. E. Magor, "Association of RIG-I with innate immunity of ducks to influenza," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 13, pp. 5913–5918, 2010.

[35] L. Josset, J. Textoris, B. Loriod et al., "Gene expression signature-based screening identifies new broadly effective influenza A antivirals," *PLoS ONE*, vol. 5, no. 10, Article ID e13169, 2010.

[36] A. York and E. Fodor, "The major mRNA nuclear export NXF1 pathway is increasingly implicated in viral mRNA export and this review considers and discusses the current understanding of how influenza A virus exploits the host mRNA export pathway for replication," *RNA Biology*, vol. 10, no. 8, pp. 1274–1282, 2013.

[37] L. Chatel-Chaix, M. A. Germain, A. Motorina et al., "A host YB-1 ribonucleoprotein complex is hijacked by hepatitis C virus for the control of NS3-dependent particle production," *Journal of Virology*, vol. 87, no. 21, pp. 11704–11720, 2013.

[38] F. di Nunzio, A. Danckaert, T. Fricke et al., "Human nucleoporins promote HIV-1 docking at the nuclear pore, nuclear import and integration," *PLoS ONE*, vol. 7, no. 9, Article ID e46037, 2012.

[39] F. W. Porter, B. Brown, and A. C. Palmenberg, "Nucleoporin phosphorylation triggered by the encephalomyocarditis virus leader protein is mediated by mitogen-activated protein kinases," *Journal of Virology*, vol. 84, no. 24, pp. 12538–12548, 2010.

[40] A. M. Copeland, W. W. Newcomb, and J. C. Brown, "Herpes simplex virus replication: roles of viral proteins and nucleoporins in capsid-nucleus attachment," *Journal of Virology*, vol. 83, no. 4, pp. 1660–1668, 2009.

[41] J. B. Ritter, A. S. Wahl, S. Freund, Y. Genzel, and U. Reichl, "Metabolic effects of influenza virus infection in cultured animal cells: Intra- and extracellular metabolite profiling," *BMC Systems Biology*, vol. 4, article 61, 2010.

[42] R. Janke, Y. Genzel, M. Wetzel, and U. Reich, "Effect of influenza virus infection on key metabolic enzyme activities in MDCK cells," *BMC Proceedings*, vol. 5, supplement 8, article P129, 2011.

[43] X. Chen, Z. Yang, Y. Lu et al., "Clinical features and factors associated with outcomes of patients infected with a novel influenza A, (H7N9) virus: a preliminary study," *PLoS ONE*, vol. 8, no. 9, Article ID e73362, 2013.

[44] S. Chabierski, G. R. Makert, A. Kerzhner et al., "Antibody responses in humans infected with newly emerging strains of west nile virus in europe," *PLoS One*, vol. 8, Article ID e66507, 2013.

[45] Y. Li, L. Du, H. Qiu et al., "A recombinant protein containing highly conserved hemagglutinin residues 81–122 of influenza H5N1 induces strong humoral and mucosal immune responses," *BioScience Trends*, vol. 7, no. 3, pp. 129–137, 2013.