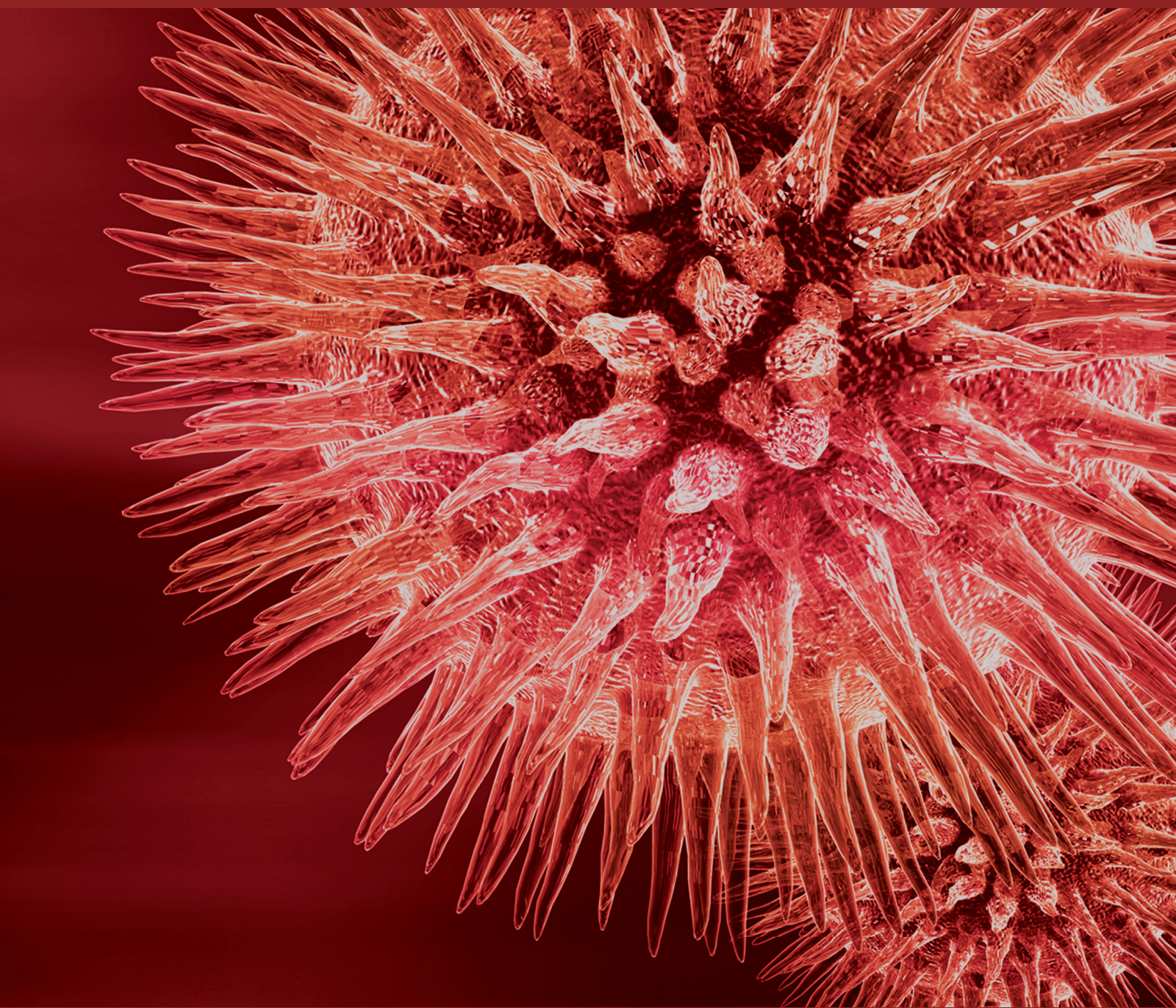
BioMed Research International

Computational Molecular Networks and Network Pharmacology

Lead Guest Editor: Kang Ning
Guest Editors: Xingming Zhao, Ansgar Poetsch, Weihua Chen, and Jialiang Yang



Computational Molecular Networks and Network Pharmacology

BioMed Research International

Computational Molecular Networks and Network Pharmacology

Lead Guest Editor: Kang Ning
Guest Editors: Xingming Zhao, Ansgar Poetsch, Weihua Chen, and Jialiang Yang

Contents

Hindawi
BioMed Research International
Volume 2017, Article ID 7573904, 1 page
https://doi.org/10.1155/2017/7573904

Editorial

Computational Molecular Networks and Network Pharmacology

Kang Ning,[1] Xinming Zhao,[2] Ansgar Poetsch,[3,4] Wei-Hua Chen,[1] and Jialiang Yang[5]

[1]*Key Laboratory of Molecular Biophysics of the Ministry of Education, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China*
[2]*Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China*
[3]*Plant Biochemistry, Ruhr-University Bochum, Bochum, Germany*
[4]*School of Biomedical and Healthcare Sciences, Plymouth University, Plymouth, UK*
[5]*Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York City, NY 10029, USA*

Correspondence should be addressed to Kang Ning; ningkang@hust.edu.cn

Received 9 October 2017; Accepted 11 October 2017; Published 8 November 2017

For biomedical research relying on systems biology approaches, two major subdomains might have profound impacts: (1) the molecular network for understanding the principles of regulations at multiple levels and (2) network pharmacology for investigating the effect of small molecules on gene dynamics. In this issue, we lay emphasis on analytical method development for molecular networks and network pharmacology for a broad area of applications. Any computational methods towards better interpretation of molecular networks, as well as those methods related to network pharmacology, would be desired. This special issue has been affiliated with the workshop "Molecular Networks and Network Pharmacology" on the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), which was held on December 15–18, 2016, in Shenzhen, Guangdong, China.

In this special issue, we have received 28 papers, out of which 17 have been accepted for publication. These papers could be categorized into four types: (1) computational approaches for biological network analysis, (2) statistical approaches for biological network analysis, (3) network pharmacology studies focusing on cancer therapy, and (4) network pharmacology studies focusing on Traditional Chinese Medicine.

We would like to lay emphasis on three works which are of special interest for this special issue. Firstly, the work by K. Huang et al. from Ohio State University on single-cell gene expression patterns, which has proposed the virtual analytical approach for single-cell sorting analysis, has advanced our knowledge on single-cell categorization as well as heterogeneities among single cells. Secondly, in the work done by J. Gao et al. from Beijing University of Chemical Technology, deep learning approach (Convolutional Neural Networks (CNN)) has been applied on calling structural variations based on low coverage data. Thirdly, the work by Y. Ga et al. from Tibetan Traditional Medical College on network pharmacology of a Traditional Chinese Medicine (i.e., TCM, which is Wuwei-Ganlu-Yaoyu-Keli) has shown us a nice example of how network pharmacology could be applied for new principles of TCM.

As our editorial team all can agree, both areas of computational molecular networks and network pharmacology, and more importantly their interplay, have represented a rapidly growing interdisciplinary research field. In this field, both computational and experimental approaches have been used for network modeling, based on which better understanding and applications of existing drugs (western drugs and TCM) could be explored. With the advancement of deep learning and the urgent need for drug development, we believe that the topics included in this special issue would be of great interest for those working in related areas, as well as for general audience. Thus, we hope that the readers will enjoy this special issue.

Kang Ning
Xinming Zhao
Ansgar Poetsch
Wei-Hua Chen
Jialiang Yang

Hindawi
BioMed Research International
Volume 2017, Article ID 9829175, 8 pages
https://doi.org/10.1155/2017/9829175

Research Article

RNA-seq Based Transcription Characterization of Fusion Breakpoints as a Potential Estimator for Its Oncogenic Potential

Jian-lei Gu,[1,2,3] Morris Chukhman,[4] Yao Lu,[1,2] Cong Liu,[1,2,4] Shi-yi Liu,[2] and Hui Lu[1,2,3,4]

[1]*Shanghai Institute of Medical Genetics, Shanghai Children's Hospital, Shanghai Jiao Tong University, Shanghai 200040, China*
[2]*Department of Bioinformatics, SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai 200240, China*
[3]*Key Laboratory of Molecular Embryology, Ministry of Health and Shanghai Key Laboratory of Embryo and Reproduction Engineering, Shanghai 200040, China*
[4]*Department of Bioengineering, Bioinformatics Program, University of Illinois at Chicago, Chicago, IL 60607, USA*

Correspondence should be addressed to Hui Lu; huilu.bioinfo@gmail.com

Received 21 December 2016; Accepted 23 August 2017; Published 17 October 2017

Academic Editor: Ansgar Poetsch

Based on high-throughput sequencing technology, the detection of gene fusions is no longer a big challenge but estimating the oncogenic potential of fusion genes remains challenging. Recent studies successfully applied machine learning methods and gene structural and functional features of fusion mutation to predict their oncogenic potentials. However, the transcription characterizations features of fusion genes have not yet been studied. In this study, based on the clonal evolution theory, we hypothesized that a fusion gene is more likely to be an oncogenic genomic alteration, if the neoplastic cells harboring this fusion mutation have larger clonal size than other neoplastic cells in a tumor. We proposed a novel method, called iFCR (internal Fusion Clone Ratio), given an estimation of oncogenic potential for fusion mutations. We have evaluated the iFCR method in three public cancer transcriptome sequencing datasets; the results demonstrated that the fusion mutations occurring in tumor samples have higher internal fusion clone ratio than normal samples. And the most frequent prostate cancer fusion mutation, TMPRSS2-ERG, appears to have a remarkably higher iFCR value in all three independent patients. The preliminary results suggest that the internal fusion clone ratio might potentially advantage current fusion mutation oncogenic potential prediction methods.

1. Introduction

Chromosomal rearrangement events often lead to gene fusion mutation and result in a hybrid fusion gene, consisting of two separate fusion parents (genes) [1, 2]. Gene fusion is an important class of genetic alterations in human cancers; it causes about 20% of human cancers [3]. In the last decades, a large number of important fusion mutations have been recognized [3], including the first identified "Philadelphia chromosome" BCR-ABL gene fusion in chronic myelogenous leukemia [4], the important biomarker of synovial sarcomas, SYT–SSX gene fusion [5], and the most studied fusion TMPRSS2-ERG in prostate cancer [6]. However, distinguishing oncogenic fusion mutations, whose functions are critical for cancer initiation, progression, and metastasis, remains a big challenge. Traditionally, a fusion event is considered as an oncogenic mutation if it occurs more frequently in cancer patients (i.e., high recurrent rate) [2, 7]. However, this strategy is expensive and time-consuming to conduct experiments for many patients. Moreover, this method has limited power to predict the oncogenic potential of novel and rare fusion mutations for a certain patient, and thus its application in the era of precise medicine is limited.

Currently, several studies have attempted to predict the oncogenic potential for fusion mutations. Shugay et al. implemented 24 structural and functional features of known oncogenic fusion genes and then predict the oncogenic potential for novel fusion genes by a SVM (Support Vector Machine) classifier [8]. Wang et al. developed an algorithm to nominate biologically important fusion mutations by integrating various molecular interactions, pathways, and functional annotations [9]. Wu and his colleagues used a molecular network based method to prioritize oncogenic fusion genes [10]. These machine learning based methods

all relied on sequence structural and functional features of fusion genes. However, due to the incompleteness of included features under investigation, these methods could be biased. Moreover, the transcription characterizations of fusion genes were ignored by these methods.

It is widely accepted that tumor has heterogeneous cell composition, which can be viewed from Darwin's evolutionary perspective as a heterogeneous population of neoplastic cells [11]. The mutation-endowed genetic alteration in cancer reflects the "survival" fitness of neoplastic cells. The neoplastic clones harboring "driver" mutations could be expanded during the progression of cancers. Thus the dynamic changes of specific clonal size also might reflect the oncogenic potentials of specific mutations [11, 12]. Based on this concept, we hypothesized that a fusion gene is more likely to be an oncogenic mutation if the neoplastic subclone harboring this fusion mutation has a larger population size, compared to other clones. And if we could estimate the clonal size of neoplastic cells, harboring a certain fusion gene, it might be helpful to predict the oncogenic potential of fusion mutations in tumor sample.

To achieve this goal, there are two fundamental questions that need to be answered: (1) if there is only transcriptome sequencing data, how can we estimate the relative subclone size in a mixture tumor sample? (2) Does this estimator have enough power to distinguish "oncogenic" fusion genes from "passenger" background? The best way to infer detecting subclonal heterogeneity is to analyze somatic DNA alterations by exome or genomic sequencing. However, if we only have RNA-seq data available, we proposed a new transcript-based method, named iFCR, to estimate the relative subclone size of neoplastic cells, harboring a certain fusion mutation. Public glioblastoma single-cell sequencing data was used to test this assumption. To address the second problem, we applied iFCR to two public datasets, including a breast cancer cell line dataset and a primary prostate tumors (with adjacent normal tissues) dataset, where the breast cancer cell lines, with homogeneous cell compositions, was used to simulate the early-stage "oncogenic" fusion mutations in primary tumor samples. In the following context, we will describe this new method in detail and then demonstrate the results of applying this estimator to two datasets.

2. Results

2.1. The Estimation of Relative Clone Size by iFCR. Traditionally, the reconstruction of subclone structure is based on in situ hybridization method [13, 14] or DNA sequencing technology [15, 16]. However, gene fusion studies used transcriptome sequencing technology and merely accompanied genome sequencing data in the same sample. In order to estimate the subclone structure based on transcriptome sequencing data, we make a simple assumption that fusion genes and their parent genes have similar expression level among neoplastic cells in the same sample. Based on this assumption, the proportion of subclone size could be represented as the ratio of expression level between chimeric transcripts and their corresponding normal parent's transcripts. This ratio, defined as iFCR, reflects the subclone proportion of specific chimeric subclones in the heterogeneous neoplastic cells. However, as Figure 1 shows, a gene fusion mutation is the juxtaposition of two separate genes; the breakpoint region is the only different part between chimeric transcript and their parents' transcripts. To represent the relative quantities of chimeric transcripts, the sequencing reads that aligned onto the breakpoint of a chimeric transcript and represented the number of chimeric transcripts were called fusion reads in this study. Correspondingly, the sequencing reads that aligned onto the breakpoint of their parents' transcripts and represented the number of normal parents' transcripts were called overlapping reads. In this work, we directly used the number of fusion reads from original published articles, and a realignment procedure was designed and performed to retrieve these overlapping reads. The details of this procedure are described in Methods.

To test this method, we used a single-cell sequencing study of glioblastoma dataset (SRP042161) and detected the fusion mutations of each single-cell sequencing library for each tumor sample. So we calculated the heterogeneity of fusion clones in two ways: (1) by summing the reads supporting the fusions and their parents in of the single-cells, we were able to calculate their $iFCR^{average}$ value. (2) As another calculation, for each tumor sample, we counted the number of cells each fusion was identified in and the number of cells that the parent genes in that fusion had nonzero transcript counts in and calculated a "real" ratio of the number of fusion clones and normal clones that is calculated from the cell counts rather than the transcripts counts. As Figure 3 shows, the log of the iFCR value linearly correlated with the "real" ratio of number of fusion cells and normal cells.

Theoretically, breast cancer cell lines, consisting of homogeneous cells, should have lower heterogeneity, while higher heterogeneity should be expected in primary tumors and their adjacent normal tissue. To evaluate whether iFCR could capture this pattern, we compared the read distributions between two datasets. There are 62% (25/40) and 54% (20/37) chimeric transcripts that have reads mapped to full-length transcripts of both parents' genes in prostate tumors and normal, respectively. This proportion reduced to 30% (7/23) in breast cancer cell lines. As Figure 2 shows, in most breakpoints of breast cancer cell lines, more reads could be mapped to chimeric transcripts (i.e., fusion reads) than parents' genes (i.e., overlapped reads), while this ratio is reversed in primary prostate tumors and their adjacent normal tissues. Specifically, in adjacent normal tissues (green), all the chimeric transcripts carry less reads in chimeric transcripts than those in parents' genes. These results suggest iFCR might be a useful ratio in estimating tumor heterogeneity.

2.2. iFCR Distribution Is Correlated with Recurrent Rate in the Prostate Tumor Dataset. The original prostate cancer study indicated [7] that the 14 prostate cancer samples harbored 38 tumor-specific chimeric transcripts, of which

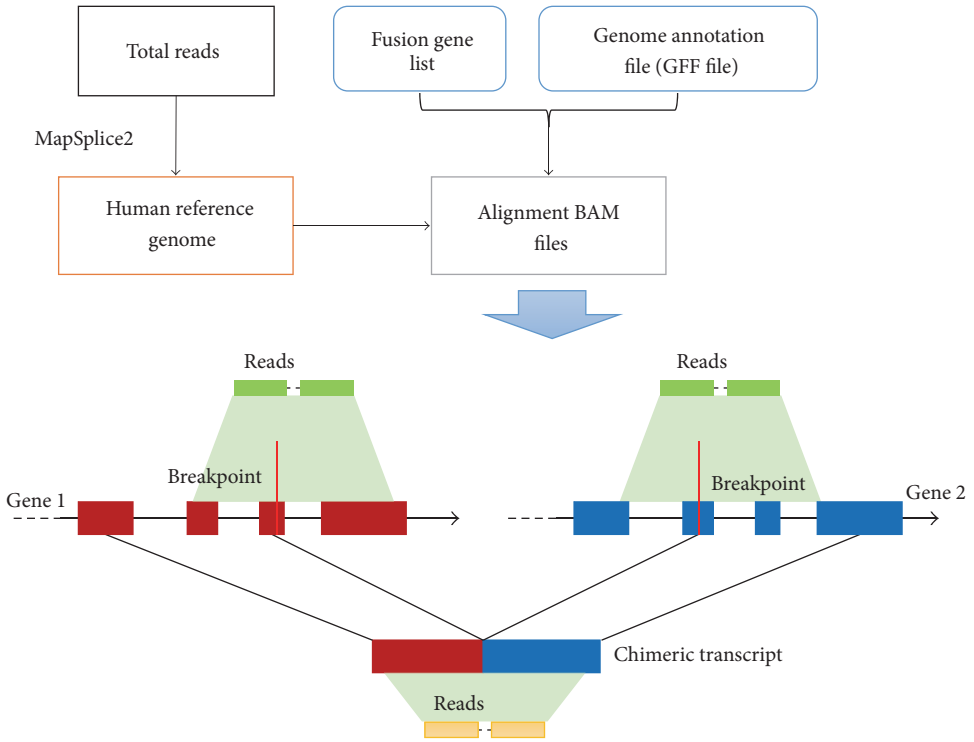


FIGURE 1: The diagram of realignment procedure to identify the overlapping reads of parents' transcripts. The RNA-seq data realigned to the corresponding reference genome, and the genome annotation file (GTF) and fusion mutations were used to retrieve these overlapping reads of parent genes. The red boxes represent the exonic sequences from Gene 1 and blue boxes are from Gene 2. The sequencing reads aligned onto the breakpoint of Gene 1 and Gene 2 were called overlapping reads; the sequencing reads aligned onto the breakpoint of chimeric transcript were called fusion reads. Breakpoints could occur in exonic region, intronic region, and UTR region.

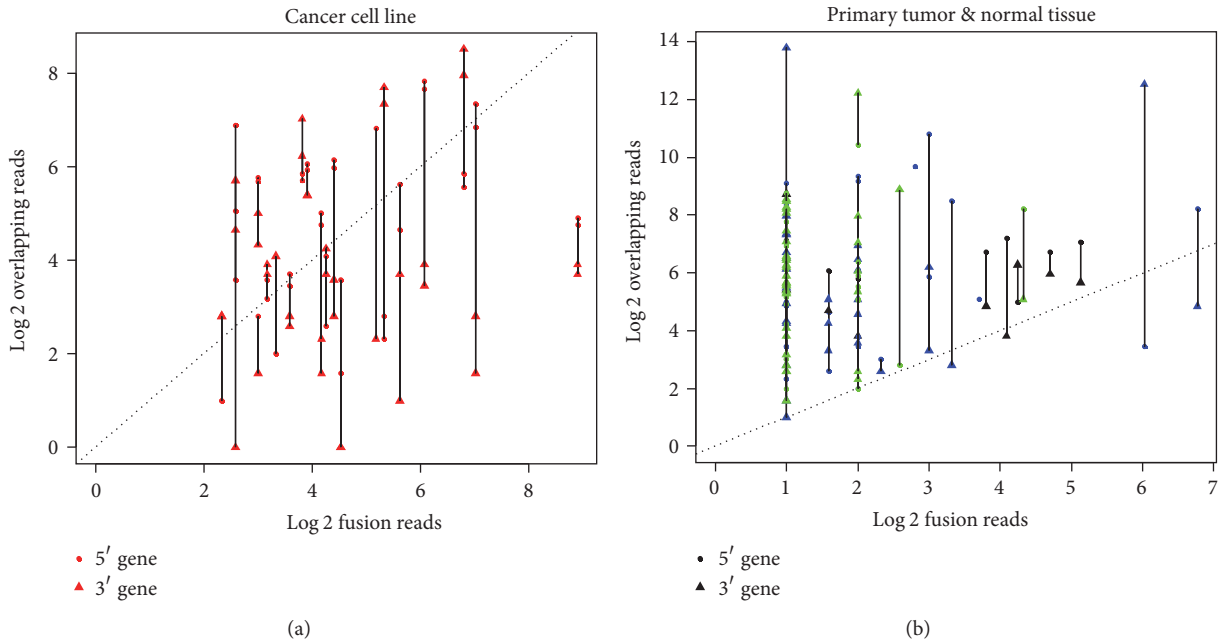


FIGURE 2: The dot-scatter plot comparison of the number of fusion reads and the overlapping reads. Diagram (a) shows fusion mutations in cell line samples. Diagram (b) shows fusion mutations in primary tumor and normal tissue samples. The green, blue, and black dots in diagram (b) represent normal counterparts, tumor, and recurrent fusion mutations, respectively. Dots give an observational estimate that fusion mutations occurring in tumors and normal tissues have more normal overlapping reads than in cancer cell lines.

at least 5 are recurrent transcripts in Chinese population, including TMPRSS2-ERG, USP9Y-TTTY15, CTAGE5-KHDRBS3, RAD50-PDLIM4, and SDK1-AMACR. Specifically, the TMPRSS2-ERG is the most studied chimeric transcript in prostate cancer [17, 18]. As shown in Figure 4, we further divided the fusion mutations reported in primary prostate tumors into three groups based on their recurrence rates, which is the gold-standard for oncogenic potential evaluation in current studies [2, 7]. The first group consists of TMPRSS2-ERG, which was detected in three prostate cancer patients (i.e., TMPRSS2-ERG group). The second groups includes 5 recurrent fusion mutations reported in 8 prostate tumor samples (i.e., recurrent group). And the rest of tumor fusion mutations were included in the third group (i.e., tumor group). We also included fusion mutations reported in breast cancer cell lines (i.e., cell line group) and adjacent normal tissues as the "positive control" and "negative control," respectively. After calculating iFCR for each group, we compared iFCR distribution across the different groups. Figures 4 and 5 showed $iFCR^{average}$ values across five groups. Our results indicate that $iFCR^{average}$ values are well correlated with recurrence rate of fusion mutation. As expected, the $iFCR^{average}$ values are higher in breast cancer cell lines than those in other groups, and the $iFCR^{average}$ values in adjacent normal tissues are the lowest. As the highest recurrent chimeric transcript, the group harboring TMPRSS2-ERG transcripts shows the highest $iFCR^{average}$ values among primary prostate tumor groups. And then the $iFCR^{average}$ values of the recurrent group are higher than the nonrecurrent group. The other two indicators, $iFCR^{max}$ and $iFCR^{min}$, show the same positive correlation trend with the recurrent rate of chimeric transcripts (see Supplement Figure 1 and Supplement Figure 2 in Supplementary Material available online at https://doi.org/10.1155/2017/9829175).

2.3. Novel Putative Oncogenic Fusion Mutations in the Prostate Cancer Dataset. The nonrecurrent fusion between exon 8 of ZC3H6 and exon 2 of LRP1B was present at a high iFCR value (0.38 for iFCR-average). The ZC3H6-LRP1B fusion was only detected in patient #13 and has not been previously reported, but its high iFCR value and the LRP1B did not seem to have any overlapping reads, indicating that it may play an important role in patient #13. The fusion mutation UPF3A-CDC16 was also identified in both the tumor and the adjacent normal tissue of the same patient (#9); did the iFCR value of this fusion mutation change between tumor and its adjacent normal tissue? We then compared the iFCR value of UPF3A-CDC16 in both tumor sample and its corresponding normal adjacent tissue. Interestingly, though it is a nonrecurrent chimeric transcript, the iFCR value of UPF3A-CDC16 was increased dramatically in tumor samples, from 0.06 in normal tissue to 0.33. This raises the possibility that this nonrecurrent chimeric transcript was under positive selection pressure and the clone harboring this specific transcript has been enriched during the progression of cancer in patient #9. However, more studies are required to clarify this mechanism for this observation.

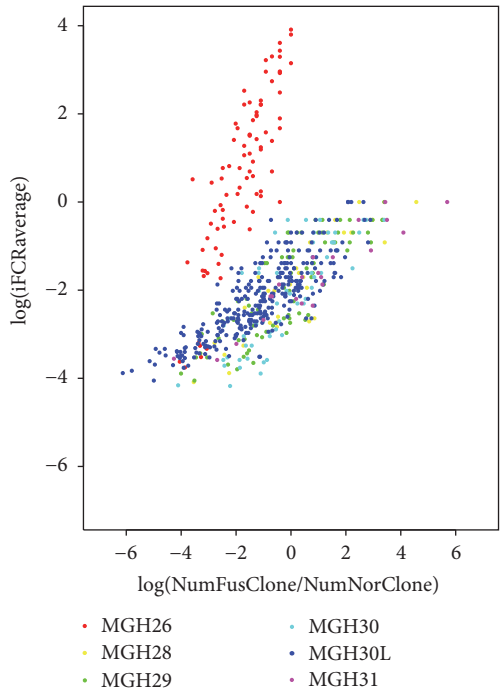


Figure 3: The logged iFCR value linearly correlated with the "real" ratio of number of fusion cells and normal cells. The x-axis is the ratio of number of fusion cells and normal cells for a certain fusion mutation. The y-axis is the $iFCR^{average}$ value for a certain fusion mutation. Each dot represents the iFCR value and the ratio of number of fusion cells and normal cells for a certain fusion mutation. The dots with different colors represented the different sequencing libraries from 5 individual tumor samples (MGH26, MGH28, MGH29, MGH30, and MGH31). MGH31L is sequenced by long reads (100 bp).

3. Discussion

Since the discovery of gene fusion 50 years ago, over 358 oncogenic chimeric transcripts were recognized [3]. With advances in NGS and bioinformatics technology, identification of hybrid fusion gene is no longer a challenge. To date, one of the main challenges in gene fusion study is to help oncologists and physicians to identify oncogenic fusion genes from noisy "background" genomic aberrations.

It is widely accepted that subclone genetic heterogeneity is a common characteristic of tumors, with both spatial and temporal heterogeneity of primary tumors observed [11, 12]. The clonal evolution theory suggested that the survival ability of neoplastic cells could be inferred by comparing subclone diversity or architecture at different time points. Based on clonal evolution theory of cancer, a fusion gene is more likely to be a survival (oncogenic) aberration if the subclone harboring this specific mutation has larger clonal proportion in a heterogeneous tumor sample. And the subclone harboring specific "survival" genomic aberration could be positively selected and enriched during the cancer progression.

Traditionally, the subclone structure is recovered by in situ hybridization methods [13, 14] or computation methods

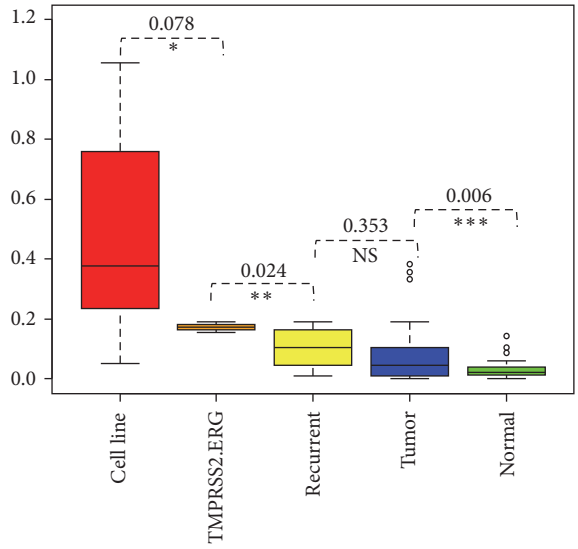


FIGURE 4: The boxplot comparison of iFCR^{average} values among four groups. The x-axis represented five different groups: C: cell lines group, TMPRSS2-ERG group, R: recurrent group, T: tumor groups, and N: normal tissue group. The y-axis is the iFCR value. The iFCR values in breast cancer cell lines are remarkably higher than other groups, and the iFCR values of tumor are remarkably higher than their normal counterparts. T-test was used to evaluate the statistical significance (p value) among different groups. NS is nonsignificance; *significance at 10% level, **significance at 5% level, and ***significance at 1% level.

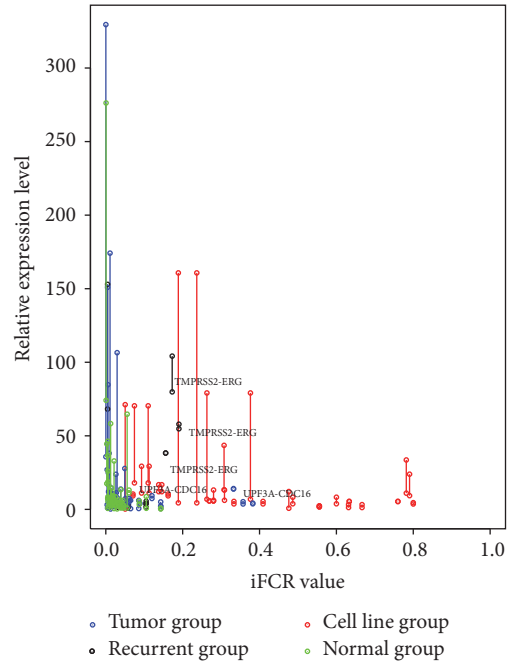


FIGURE 5: The relative RPKM expression level (y-axis) and iFCR value (x-axis) of parent's genes. Compared to primary tumors and normal tissues, fusion mutations occurring in breast cancer cell lines tend to be having higher iFCR value. The most frequent prostate cancer fusion mutation TMPRSS2-ERG appears as higher expression level and iFCR value in all three independent patients, and it is closed to the fusion mutation of cancer cell lines. A nonrecurrent tumor fusion mutation UPF3A-CDC16 from patient #9 is increased from 0.06 to 0.33 in the normal counterpart to its tumor sample.

based on DNA sequencing data [15, 16]. However, gene fusion studies often lack paired genome sequencing data. Here we proposed a novel method to estimate the subclone structure of fusion mutation based on transcriptome sequencing data only. We acknowledge that the assumption that expression in wild-type cells and tumor is similar might be flawed. However, our results suggested that iFCR could potentially reflect the tumor heterogeneity.

But the quantification of chimeric transcripts remains computationally challenging. Because the short sequencing reads from chimeric transcripts are almost the same as their parents' transcripts, it is very difficult to distinguish a chimeric transcript from their parents' transcripts. Current methods [19–26] identify chimeric transcripts by identification of fusion reads, that is (as shown in Figure 1), the short sequencing reads aligned onto the breakpoint of two parents' genes. These fusion reads are the only sequencing reads that can be used as evidence to support the occurrence of a certain chimeric transcript. In this work, we used the number of fusion reads to infer the expression level of chimeric transcripts and the number of overlapping reads to infer the expression level of parents' genes.

In this work, we tested our method on two public RNA-seq datasets and took a comparison of the chimeric subclone divergence among primary prostate tumors, normal prostate tissues, and breast cancer cell lines. As shown in Figure 1, we retrieved overlapping reads for parent genes through a realignment procedure. We compared the number of fusion reads and overlapping reads for fusion mutations among cancer cell lines, primary tumor samples, and normal tissues. The fusion mutations from primary tumors and normal tissues tended to have more overlapping reads (Figure 2). This is consistent with the fact that the somatic fusion mutations in primary tumor and normal tissue have relatively smaller subclone size than cancer cell lines, and the numbers of mutated clones might reflect the oncogenic potential of this fusion gene in a patient's tumor.

The nonrecurrent fusion between exon 8 of ZC3H6 and exon 2 of LRP1B was present at a high iFCR value (0.38 for iFCR-average). The ZC3H6-LRP1B fusion was only detected in patient #13 and has not been previously reported, but its high iFCR value and the LRP1B did not seem to have any overlapping reads, indicating that it may play an important role in patient #13. Previous studies have demonstrated that LRP1B is a potential tumor suppressor gene and downregulated expression of LRP1B proposed to be involved in multiple primary cancers [27, 28]. The deletion of LRP1B also has been associated with chemotherapy resistance in high-grade cancers [29]. These results indicate that the silencing of LRP1B may be a driver event.

Moreover, we found a very interesting fusion gene UPF3A-CDC16 in patient 9, whose iFCR value was increased from 0.06 to 0.33 in the tumor sample, compared with its

adjacent normal tissue. This result indicated that UPF3A-CDC16 might be enriched during cancer progression. A previous study has suggested that *CDC16* is an important gene which involved cell reproduction [30]. One possible oncogenic mechanism is that the proportional increase of UPF3A-CDC16 might result in the function loss of CDC16, promoting the proliferation of neoplastic cells. Although it is possible that normal tissue had contaminated tumor samples during surgical operation or experimentation, the changes in iFCR values could still reflect its differential clone size.

Gene fusion events are not only a consequence of disability of cancer genomes, it is also an important mechanism of the evolution of novel proteins, it is contributing to the transcriptome complexity in normal tissues [31, 32]. Frenkel-Morgenstern and collaborators used mass spectrometry to study the corresponding protein products of chimeric transcripts and attempted to study potential functions of these chimeric products [33]. We hypothesized that the chimera products' new biological functions may rely heavily on their quantities. The relative expression level of chimeric transcripts might also be an indicator for inferring oncogenic potential of fusion mutation. Thus, we compared the iFCR value of chimeric transcripts and expression levels of their corresponding parents' genes. As Figure 5 shows, the expression levels of most of these genes are very low. However, the fusion mutations from breast cancer cell lines (red) exhibit higher iFCR value and were located at the right part of the diagram. The fusion mutations from tumor samples appear to have various iFCR values and, interestingly, the well-studied prostate cancer fusion TMPRSS2-ERG was closed to the fusions of cancer cell lines and appears to have higher iFCR values and expression levels in all three independent patients. Next, we calculated the fold change of parent genes' expression levels between tumor samples and their counterpart samples and compared the fold change with those fusions' iFCR values. As Supp. Figure 3 suggested, the fusion mutation of TMPRSS2-ERG changed the expression of its parent genes, indicating that the TMPRSS2-ERG mutation plays a critical role in prostate cancer dependent upon the expression changes of TMPRSS2 and ERG genes, consistent with previous widely discussed studies [34]. However, for the rest of high iFCR fusion mutations, such as ZC3H6-LRP1B, EMB-ATG10, UPF3A-CDC16, DYRK1A-CMTM4, and CD97-EMR2, the oncogenic potential remains unclear. The oncogenic mechanism of these chimeric transcripts might be different.

The advantage of our method is that the oncogenic potential of fusion genes could be estimated using a single RNA-seq dataset, which makes it ideal for application in precise medicine. Further works could integrate gene structural/functional information of fusion gene and our method to achieve better performance. The limitation of our method is difficult to evaluate its discriminative power by computational methods (e.g., cross-validation) due to the wide chimeric transcript spectrum among different tumor data. Also previous studies suggested that fusion genes were often caused by genomic segment amplifications, and these amplifications were often associated with gene overexpression [35].

In summary, we present a new concept of inferring the oncogenic potential of novel fusion genes identified in tumor samples. Unlike the existing structure/functional based method, our method incorporated the concept of clone evolution theory and transcription characterization of fusion genes. This study also showed that the iFCR values of fusion genes in tumor samples were remarkably higher than those in normal tissues, especially in tumor cell lines. The most frequent fusion mutation in prostate cancer TMPRSS2-ERG shows higher iFCR value in all three independent patients. We also observed that a previously reported [7] fusion gene, UPF3A-CDC16, was enriched in the tumor sample and it is indicated that UPF3A-CDC16 might be playing an important role during the cancer progression in patient 9#. To the best of our knowledge, this is the first work to incorporate transcriptome sequencing data and clone evolution theory to investigate the oncogenic potential of chimeric transcripts. Our work provides a new insight into the oncogenic potential study of fusion genes.

4. Methods

4.1. Data Source. A single-cell transcriptome sequencing study of glioblastoma (SRP042161) was used to test our RNA-seq data based on clone size estimation assumption. This dataset has 658 tumor single-cell sequencing libraries from five independent patients. They are MGH26 tumor sample with 189 single-cell sequencing libraries; MGH28 tumor sample with 95 single-cell sequencing libraries; MGH29 tumor sample with 96 single-cell sequencing libraries; MGH30 tumor sample with 91 single-cell sequencing libraries; MGH30L tumor sample with 91 single-cell sequencing libraries; and MGH31 tumor sample with 96 single-cell sequencing libraries.

The public RNA sequencing (RNA-seq) data of a prostate cancer study [7] (SRA: ERP000550) and a breast cancer study [19] (SRA: SRP003186) was downloaded from NCBI Sequence Read Archive (SRA) database. Table 1 summarizes the datasets used in this study. The prostate cancer dataset was derived from 14 pairs of primary prostate cancer and their corresponding adjacent normal tissues in Chinese population. The breast cancer cell line dataset consists of 3 cell lines and 5 sequencing libraries; they are KPL-4, SK-BR-3 (two sequencing libraries), and BT-474 (two sequencing libraries). Since the MCF-7 cell line has not provided sequence of the chimeric transcripts, we excluded it from our analysis. The detailed descriptions of these datasets can be found in their original articles [7, 19]. In total, 28 paired-end RNA-seq libraries from the prostate cancer patients and 5 paired-end RNA-seq libraries from 3 distinct breast cancer cell lines were analyzed in this work.

4.2. Bioinformatics Preprocess Procedure. The fusion mutation detection procedure for single-cell sequencing libraries was conducted by FusionCatcher with default parameters [36], providing the BAM files and the information of sequencing reads which supported the chimeric transcripts.

Table 1: Summary of three validated datasets used in this study.

	Sample type	Sequencing libraries	Chimeric transcripts
[#]Single cell	Single cell	658	574
Prostate cancer, 14 individuals	Tumor samples	14	40
	Adjacent normal tissue	14	37
Breast cancer cell lines	BT-474	2	9
	KPL-4	1	3
	SK-BR-3	2	9

[#]The total number of fusion mutations in single cell dataset.

The iFCR$^{\text{average}}$ values for single-cell libraries could be calculated by summing the reads supporting the fusions and their parents in these single-cells libraries from single patient. For each tumor sample, we also counted the number of cells each fusion was identified in and the number of cells that the parent genes in that fusion had nonzero transcript counts in and calculated a "real" ratio of the number of fusion clones and normal clones that is calculated from the cell counts rather than the transcripts counts.

For prostate cancer and breast cancer dataset, our focus was to predict the oncogenic potentials of chimeric transcripts. So, we directly used packages from previously published articles to detect fusion events and retrieve information for downstream analysis. For each sample, paired-end reads were aligned to their corresponding reference genome by a transcriptome aligner MapSplice [26] with default settings. As Figure 1 shows, the sequencing reads which span the breakpoints of parent genes were called "overlapping reads" and the sequencing reads spanning the breakpoint of chimeric transcript were called "fusion reads." The overlapping reads were required to have at least 5 bp overlaps with flanking sequences in both sides of breakpoints. The number of fusion reads was directly obtained from their original publications [7, 19]. For breast cancer dataset, 24 validated fusion mutations were previously reported [19]. One fusion mutation (CSE1L-ENSG00000236127) was removed from our analysis due to the corresponding RefSeq gene symbol of ENSG00000236127 not being found in hg19. For the prostate cancer dataset, among 83 fusion mutations identified in their study [7], there are 4 (tumor samples) and 8 (adjacent normal tissues) fusion mutations that were removed in our further analysis due to the same reason. The detailed information of fusion mutations can be found in Supp. Table 1.

4.3. Internal Fusion Clone Ratio Calculation and Relevance Network Construction. In this work, we hypothesize that the ratio of the number of chimeric transcripts to the number of normal nonfusion transcripts could reflect the ratio of subclone population size. And this ratio could be estimated by the number of overlapped reads and fusion reads. The proposed subclone ratio estimator is defined as

$$\text{iFCR}^{\text{average}} = \frac{f_{a,b}}{\text{avg}\left(n_a, n_b\right)}$$

$$\text{iFCR}^{\text{max}} = \frac{f_{a,b}}{\min\left(n_a, n_b\right)}$$

$$\text{iFCR}^{\text{min}} = \frac{f_{a,b}}{\max\left(n_a, n_b\right)}. \tag{1}$$

Here, $f_{a,b}$ is the number of the fusion reads mapping to the breakpoint of gene a and gene b. n_a is the number of overlapping reads for gene a and n_b is the number of overlapping reads for gene b. Here $\text{avg}(n_a, n_b)$, $\min(n_a, n_b)$, and $\max(n_a, n_b)$ donate the relative expression of wild-type transcript from parent genes a and b using three simple combinations. And thus iFCR$^{\text{average}}$, iFCR$^{\text{max}}$, and iFCR$^{\text{min}}$ represent the average, maximum, and minimum ratio of chimeric transcripts subclones to wild-type subclones, respectively. This equation could be refined later with the number of reads replaced by RPKM (the number of reads per kilobase of gene length per million mappable reads) [7, 37].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported in part by the Shanghai Collaborative Innovative Center for Translational Medicine, National Natural Science Foundation of China (no. 31071167 and no. 31370751), National Basic Research Program of China (2014CB964703), Shanghai Municipal Commission of Health and Family Planning (Grant no. 20144Y0179), Shanghai Key Projects for Basic Scientific Research (14JC1405700), and the outstanding young grant (2015) of Shanghai Children's Hospital.

References

[1] M. T. Villanueva, "Genetics: Gene fusion power," *Nature Reviews Clinical Oncology*, vol. 9, no. 4, pp. 188-188, 2012.

[2] K. Kannan, L. Wang, J. Wang, M. M. Ittmann, W. Li, and L. Yen, "Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 22, pp. 9172–9177, 2011.

[3] F. Mitelman, B. Johansson, and F. Mertens, "The impact of translocations and gene fusions on cancer causation," *Nature Reviews Cancer*, vol. 7, no. 4, pp. 233–245, 2007.

[4] C. Nowell, "The minute chromosome (Ph1) in chronic granulocytic leukemia," *Blut Zeitschrift für die Gesamte Blutforschung*, vol. 8, no. 2, pp. 65-66, 1962.

[5] A. Kawai, J. Woodruff, J. H. Healey, M. F. Brennan, C. R. Antonescu, and M. Ladanyi, "*SYT-SSX* gene fusion as a determinant of morphology and prognosis in synovial sarcoma," *The New England Journal of Medicine*, vol. 338, no. 3, pp. 153–160, 1998.

[6] S. A. Tomlins, D. R. Rhodes, S. Perner et al., "Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer," *Science*, vol. 310, no. 5748, pp. 644–648, 2005.

[7] S. Ren, Z. Peng, J. Mao et al., "RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings," *Cell Research*, vol. 22, no. 5, pp. 806–821, 2012.

[8] M. Shugay, I. O. De Mendíbil, J. L. Vizmanos, and F. J. Novo, "Oncofuse: A computational framework for the prediction of the oncogenic potential of gene fusions," *Bioinformatics*, vol. 29, no. 20, pp. 2539–2546, 2013.

[9] X.-S. Wang, J. R. Prensner, G. Chen et al., "An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer," *Nature Biotechnology*, vol. 27, no. 11, pp. 1005–1011, 2009.

[10] C.-C. Wu, K. Kannan, S. Lin, L. Yen, and A. Milosavljevic, "Identification of cancer fusion drivers using network fusion centrality.," *Bioinformatics (Oxford, England)*, vol. 29, no. 9, pp. 1174–1181, 2013.

[11] M. Greaves and C. C. Maley, "Clonal evolution in cancer," *Nature*, vol. 481, no. 7381, pp. 306–313, 2012.

[12] L. M. F. Merlo, J. W. Pepper, B. J. Reid, and C. C. Maley, "Cancer as an evolutionary and ecological process," *Nature Reviews Cancer*, vol. 6, no. 12, pp. 924–935, 2006.

[13] K. Anderson, C. Lutz, F. W. van Delft et al., "Genetic variegation of clonal architecture and propagating cells in leukaemia," *Nature*, vol. 469, no. 7330, pp. 356–361, 2011.

[14] J. J. Keats, M. Chesi, J. B. Egan et al., "Clonal competition with alternating dominance in multiple myeloma," *Blood*, vol. 120, no. 5, pp. 1067–1076, 2012.

[15] P. Lundberg, A. Karow, R. Nienhold et al., "Clonal evolution and clinical correlates of somatic mutations in myeloproliferative neoplasms," *Blood*, vol. 123, no. 14, pp. 2220–2228, 2014.

[16] Y. Qiao, A. R. Quinlan, A. A. Jazaeri, R. G. W. Verhaak, D. A. Wheeler, and G. T. Marth, "SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization," *Genome biology*, vol. 15, no. 8, p. 443, 2014.

[17] D. G. Tandefelt, J. Boormans, K. Hermans, and J. Trapman, "ETS fusion genes in prostate cancer," *Endocrine-Related Cancer*, vol. 21, no. 3, pp. R143–R152, 2014.

[18] J. Romero Otero, B. Garcia Gomez, F. Campos Juanatey, and K. A. Touijer, "Prostate cancer biomarkers: an update," *Urologic Oncology: Seminars and Original Investigations*, vol. 32, no. 3, pp. 252–260, 2014.

[19] H. Edgren, A. Murumagi, S. Kangaspeska et al., "Identification of fusion genes in breast cancer by paired-end RNA-sequencing," *Genome Biology*, vol. 12, no. 1, R6 pages, 2011.

[20] H. Ge, K. Liu, T. Juan, F. Fang, M. Newman, and W. Hoeck, "FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution," *Bioinformatics*, vol. 27, no. 14, pp. 1922–1928, 2011.

[21] M. K. Iyer, A. M. Chinnaiyan, and C. A. Maher, "ChimeraScan: A tool for identifying chimeric transcription in sequencing data," *Bioinformatics*, vol. 27, no. 20, pp. 2903-2904, 2011.

[22] D. Kim and S. L. Salzberg, "TopHat-Fusion: an algorithm for discovery of novel fusion transcripts," *Genome Biology*, vol. 12, no. 8, R72 pages, 2011.

[23] Y. Li, J. Chien, D. I. Smith, and J. Ma, "FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq," *Bioinformatics*, vol. 27, no. 12, pp. 1708–1710, 2011.

[24] A. McPherson, F. Hormozdiari, A. Zayed, and e. a., "deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data," *PLoS Computational Biology*, vol. 7, no. 5, Article ID e1001138, e1001138, 16 pages, 2011.

[25] A. Sboner, L. Habegger, D. Pflueger et al., "FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data," *Genome Biology*, vol. 11, no. 10, article R104, 2010.

[26] K. Wang, D. Singh, Z. Zeng et al., "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery," *Nucleic acids research*, vol. 38, no. 18, e178 pages, 2010.

[27] H. Prazeres, J. Torres, F. Rodrigues et al., "Chromosomal, epigenetic and microRNA-mediated inactivation of LRP1B, a modulator of the extracellular environment of thyroid cancer cells," *Oncogene*, vol. 30, no. 11, pp. 1302–1317, 2011.

[28] S. Ni, J. Hu, Y. Duan et al., "Down expression of LRP1B promotes cell migration via RhoA/Cdc42 pathway and actin cytoskeleton remodeling in renal cell cancer," *Cancer Science*, vol. 104, no. 7, pp. 817–825, 2013.

[29] P. A. Cowin, J. George, S. Fereday et al., "LRP1B deletion in high-grade serous ovarian cancers is associated with acquired chemotherapy resistance to liposomal doxorubicin," *Cancer Research*, vol. 72, no. 16, pp. 4060–4073, 2012.

[30] K. A. Heichman and J. M. Roberts, "CDC16 Controls Initiation at Chromosome Replication Origins," *Molecular Cell*, vol. 1, no. 3, pp. 457–463, 1998.

[31] G. Parra, A. Reymond, N. Dabbouseh et al., "Tandem chimerism as a means to increase protein complexity in the human genome," *Genome Research*, vol. 16, no. 1, pp. 37–44, 2006.

[32] P. Akiva, A. Toporik, S. Edelheit et al., "Transcription-mediated gene fusion in the human genome," *Genome Research*, vol. 16, no. 1, pp. 30–36, 2006.

[33] M. Frenkel-Morgenstern, V. Lacroix, I. Ezkurdia et al., "Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts," *Genome Research*, vol. 22, no. 7, pp. 1231–1242, 2012.

[34] J. Yu, J. Yu, R.-S. Mani et al., "An Integrated Network of Androgen Receptor, Polycomb, and TMPRSS2-ERG Gene Fusions in Prostate Cancer Progression," *Cancer Cell*, vol. 17, no. 5, pp. 443–454, 2010.

[35] E. Hyman, P. Kauraniemi, S. Hautaniemi et al., "Impact of DNA amplification on gene expression patterns in breast cancer," *Cancer Research*, vol. 62, no. 21, pp. 6240–6245, 2002.

[36] N. Daniel, S. Mihaela, E. Henrik et al., "FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data," *bioRxiv*, 2014, FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data.

[37] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

Hindawi
BioMed Research International
Volume 2017, Article ID 2320932, 10 pages
https://doi.org/10.1155/2017/2320932

Research Article

Systems Study on the Antirheumatic Mechanism of Tibetan Medicated-Bath Therapy Using Wuwei-Ganlu-Yaoyu-Keli

Tianhong Wang,[1] Jian Yang,[2] Xing Chen,[3] Kehui Zhao,[1] Jing Wang,[1] Yi Zhang,[1] Jing Zhao,[4] and Yang Ga[5]

[1]*School of National Medicine, Chengdu University of TCM, Chengdu, China*
[2]*School of Pharmacy, Second Military Medical University, Shanghai, China*
[3]*Department of Mathematics, Logistical Engineering University, Chongqing, China*
[4]*Institute of Interdisciplinary Complex Research, Shanghai University of Traditional Chinese Medicine, Shanghai, China*
[5]*Tibetan Traditional Medical College, Lhasa, China*

Correspondence should be addressed to Jing Zhao; zhaojanne@gmail.com and Yang Ga; yanggala@hotmail.com

Received 9 December 2016; Revised 22 January 2017; Accepted 30 July 2017; Published 26 September 2017

Academic Editor: Kang Ning

In clinical practice at Tibetan area of China, Traditional Tibetan Medicine formula Wuwei-Ganlu-Yaoyu-Keli (WGYK) is commonly added in warm water of bath therapy to treat rheumatoid arthritis (RA). However, its mechanism of action is not well interpreted yet. In this paper, we first verify WGYK's anti-RA effect by an animal experiment. Then, based on gene expression data from microarray experiments, we apply approaches of network pharmacology to further reveal the mechanism of action for WGYK to treat RA by analyzing protein-protein interactions and pathways. This study may facilitate our understanding of anti-RA effect of WGYK from perspective of network pharmacology.

1. Introduction

Rheumatoid arthritis (RA) is a chronic inflammatory autoimmune disease, which primarily attacks the synovial joints and leads to the destruction of the cartilage and bone. It can also affect multiple organs throughout the body [1]. Currently RA cannot be completely cured. The aim of treatment is to eliminate the inflammation, assuage the pain, control disease activity, prevent joint damage, and retard disease progression, so as to enhance patient's quality of life.

Tibetan Plateau is an area where RA commonly occurs. Tibetan medicated-bath therapy is one of the efficient treatments to RA in the Tibetan Medicine [2], in which the Tibetan formula Wuwei-Ganlu-Yaoyu-Keli for the bath therapy has been included in the year 1995 version of the Tibetan Medicine Standard [3] issued by the Chinese Ministry of Health. However, there is not enough research into the mode of action of Tibetan medicated-bath therapy; thus its mechanism in the modern biomedical background is not well understood.

In this work, we conducted a systems study to explore the anti-RA mechanism of Tibetan medicated-bath therapy using Wuwei-Ganlu-Yaoyu-Keli (WGYK) as compared with oral medicine dexamethasone acetate (DMA) and topical creams Qing Peng ointment (QPO) used clinically in the treatment of RA [4, 5]. Adjuvant arthritis (AA) in rat has been widely used as an experimental model that shares some features with human RA, such as swelling, cartilage degradation, and loss of joint function [6]. Treating AA model rats by WGYK with the warm water bath, DMA through the mouth, and QPO external painting, respectively, we compared the effects of these different treatments to relieve foot swelling and applied gene chip technology to detect differentially expressed genes in synovial cells under each treatment. Then we used approaches of network pharmacology to identify signaling pathways and subnetworks influenced by the AA modeling disease and regulated by different drugs [7]. At last, we checked the overlaps between the pathways or subnetworks to deduce the effects of the drugs on the disease.

2. Materials and Methods

2.1. Reagents and Drugs. The Freund's complete adjuvant was obtained from Sigma Company. WGYK was purchased from Qizheng Tibetan Medicine Limited Company. DMA was produced by Zhejiang Xianju Pharmaceutical Limited Company. QPO was purchased from Jinhe Tibetan Medicine Limited Company.

2.2. Animal Experiment. We used 40 ± 5 days healthy male SD rats (Chengdu Dashuo Biological Technology Co. Ltd.) with a body weight of 200 ± 20 g. The experiment as follows was carried out after one week of adaptive feeding in the animal laboratory of Chengdu University of Traditional Chinese Medicine.

(1) The rats were randomly divided into seven groups, each of which included ten animals, that is, control group (C), adjuvant arthritis group (AA), dexamethasone acetate group (DMA), Qing Peng ointment group (QPO), low dose Wuwei-Ganlu-Yaoyu-Keli group (WGYKl), moderate dose Wuwei-Ganlu-Yaoyu-Keli group (WGYKm), and high dose Wuwei-Ganlu-Yaoyu-Keli group (WGYKh).

(2) The rats in the six groups AA, DMA, QPO, WGYKl, WGYKm, and WGYKh were injected with 0.1 mL of Freund's complete adjuvant in the skin of right rear toe, to establish the model of adjuvant arthritis rats. Control animals were similarly injected with normal saline.

(3) Two weeks after the injection, the rats in seven groups were treated in different way by drugs or dipping bath. Specifically, the rats in the groups of C, AA, WGYKl, WGYKm, and WGYKh were treated by bath therapy of different liquids. We fixed the rats at wood cylindrical bathtubs of diameter 30 cm and height 20 cm and soaked their legs into 2 L liquid at 40 ± 2°C. The bath therapy was conducted 30 min one day. A course of treatment lasted 7 days. We conducted 4 courses. After each course, the treatment was stopped for 2 days. The liquid used in the C and AA groups was fresh warm water, while the WGYKl, WGYKm, and WGYKh groups used medical solution of WGYK whose concentrations were 2.95 g/L, 5.90 g/L, and 11.80 g/L, respectively. In the DMA group, rat stomachs were perfused with dexamethasone at the dose of 0.15 mg/kg. Rats in the QPO group were painted with appropriate amount of Qing Peng ointment. The treatment frequency and course of treatment for groups DMA and QPO were the same as those of medicated-bath treatment groups.

2.3. Measurement of Swelling Degree of Foot. The volume of each rat's right rear foot was measured twice before making animal model, while the average of the two measures was taken as the basic value. Using the same method, we measured the swelling extent of the feet in each group of rats at the fifth day, the ninth day, the thirteenth day, and the day before the treatment. After two weeks that the rats were administered drugs, we measured the foot volumes of the inflammatory side (right) at the first, second, third, and fourth week, respectively, and calculated the primary foot swelling as follows:

ΔmL = the average foot volume after inflammation − the average foot volume before inflammation.

The data was processed with SPSS 19 statistical software. All values were expressed as means ± standard errors. Single factor analysis of variance was used to compare the difference between groups. The variance was not homogeneous, and the rank sum test was used. A P value < 0.05 was considered statistically significant.

2.4. Microarray Experiment and Significantly Expressed Genes

2.4.1. Sampling of the Synovial Tissue. The rats were sacrificed under anesthesia, along the median incision of the skin of right rear ankle joint (inflammatory side). Then we exposed about $3 \times 3\,cm^2$ region in the center of the ankle joint and stripped part of smooth and bright yellow synovial tissue. At last we randomly selected three cases of synovial tissue samples from each of the seven groups.

2.4.2. Microarray Experiment. The synovial tissues were washed by saline and ground in liquid nitrogen. Then the RNA was extracted using the Trizol reagent (Life technologies, Carlsbad, CA, US). Further, quality control and purity of isolated total RNA were performed by UV spectrophotometer (Beijing Kai'Ao company), agarose gel electrophoresis, and Agilent Bioanalyzer 2200 (Agilent Company, USA) and the qualified RNA samples were subpacked and stored at −80°C for further use. RNA profiling was performed by Guangzhou RIBOBIO Company in China. Genes whose |FoldChange| > 0.585 (i.e., log 2 value of 1.5) and $P \leq 0.05$ were considered as differentially expressed [8, 9]. Here the FoldChange is the log 2 ratio of average expression intensities between the treatment and control group.

2.5. Data of Human Gene Association Network. Gene association network links genes or encoded proteins by their functional interplays, including direct physical binding and indirect interaction such as being involved in the same cellular process. Here we utilized the human functional linkage network (FLN) constructed by Linghu et al. [10]. FLN is a densely connected weighted network composed of 21,657 genes and 22,388,609 edges, in which nodes represent genes, and there is an edge if two genes participate in a common biological process. The edge weight is a probabilistic confidence score of the linkage. We normalized the original edge weight to the interval [0, 1].

2.6. Data of FDA Approved Anti-Ra Drugs and Their Target Proteins. Four classes of drugs are used clinically for the treatment of RA. They are nonsteroidal anti-inflammatory drugs (NSAID) such as flurbiprofen, disease-modifying antirheumatic drugs (DMARDs) such as sulfasalazine, glucocorticoids such as cortisone acetate, and biological response modifiers such as etanercept and abatacept. The data of FDA approved anti-RA drugs and their targets were downloaded

from the DrugBank database version 5.01 [11], which was updated in July of 2016. We searched the DrugBank database with a keyword "rheumatoid arthritis" and extracted all of the FDA approved anti-RA drugs and their corresponding targets. In this way, 51 FDA approved anti-RA drugs and corresponding 82 protein targets were collected.

2.7. Pathway Enrichment Analysis. We used pathway enrichment analysis [12] to identify pathways significantly influenced by a group of differentially expressed genes. Hypergeometric cumulative distribution was applied to quantitatively measure whether a pathway was more enriched with the group of genes than would be expected by chance [13]. In our case, if all pathways under study include N distinct genes, in which K genes are differentially expressed genes, for a randomly chosen pathway which owns n genes, the probability that we can find i differentially expressed genes in this pathway by chance obeys hypergeometric distribution:

$$f(i) = \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}. \quad (1)$$

Then the probability of getting at least k differentially expressed genes in this pathway by chance can be represented by hypergeometric cumulative distribution defined as P value:

$$P = 1 - \sum_{i=0}^{k-1} f(i) = 1 - \sum_{i=0}^{k-1} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \quad (2)$$

Given significance level α, a P value smaller than α implies a low probability that the k differentially expressed genes appear in the pathway by chance; that is, this pathway can be regarded as significantly influenced by these genes.

2.8. Scoring Network Effect of a Group of Differentially Expressed Genes. A group of differentially expressed genes under a specific condition, such as a disease status or a drug treatment, could exert their impact on other genes through network links. For each gene i in the human gene association network FLN, we quantified the influence of differentially expressed genes by a network effect score. In general, the higher score a gene receives, the deeper and more pronounced it is affected by the disease or drug. Specifically, a node's score is defined as follows:

$$S_i = \sum_{j=1}^{n} w_j^{(v)} W_{ij}^{(e)}, \quad (3)$$

where n is the number of nodes in the network and $w_j^{(v)}$ is the weight of the node j defined as absolute value of log 2 ratio of the expression level if the corresponding gene is differentially expressed; otherwise it is zero. $W_{ij}^{(e)}$ is the linkage weight connecting the genes i and j, and it is defined as 1 when $i = j$.

2.9. Construction of Condition Specific Network. For a specific condition, such as a disease status or a drug treatment, we defined a condition specific network as a subnetwork of human gene association network consisting of nodes with high network effect scores. We sorted the effect scores under this condition decreasingly and collected certain fraction of top genes in the rank list. Then these genes and their links were extracted from human gene association network to construct the condition specific network. In this way, we constructed a network impacted by a disease or regulated by a drug, respectively.

2.10. Generating Random Counterparts of Differentially Expressed Genes Under the Treatment of a Drug. For the group of differentially expressed genes under the treatment of a drug, we randomly selected the same number of genes in the background network as a random counterpart. We assigned the values of expression level of the differentially expressed genes to the genes in the counterpart randomly. Repeating this process a sufficiently large number of times gave us a set of random counterparts of the differentially expressed genes, which we used as a random reference of gene expression levels for this drug's effect.

3. Results and Discussion

3.1. Evaluation of Drug Effects by Foot Swelling Degrees. Foot swelling degrees are used as apparent indicators of arthritis to evaluate the arthritic progression of adjuvant-induced arthritis [14]. It is known that redness and swelling of the joints usually appear at the onset of arthritis. We found that the foot volumes of the control rats were at a stable level during the experimental period; meanwhile the foot volumes of the rats in the AA, DMA, WGYKh, WGYKm, WGYKl, and QPO groups reached peak values about 2 weeks after adjuvant injection. After one week's treatment, the foot swelling degree of DMA group decreased sharply, and the WGYKh, WGYKm, WGYKl, and QPO groups decreased significantly compared with the AA group. In the treatment of the second to third week, the foot swelling of all groups showed a downward trend, in which the foot swelling degree of rats in the DMA group tended to be stable, and the foot swelling degree of rats in the WGYKm, WGYKh, WGYKl, and QPO groups continued to decline. In the fourth week, DMA and QPO group slightly rebounded (Figure 1).

This experiment suggests that all the drugs under study have significant effect of alleviating swelling in AA rats, in which DMA shows best effect but rebound appears if the treatment lasts four weeks.

3.2. Differentially Expressed Genes in the AA Model and under the Treatment of Drugs. The notable feature of RA is progressive joint damage caused by chronic synovitis. A large number of activated cytokines are found in the joints of RA patients, which are the key mediator of inflammation and play an important role in the generation of joint injury and other complications. Many researches in the past decades have revealed that cytokines such as tumor necrosis alpha (TNF-a) and interleukins 1, 6, and 15 (IL1, IL6, and IL15), as well as immune-mediated inflammatory signaling pathways such as JAK-STAT, NF-κB, and MAPK signaling pathway, are deeply involved in the occurrence and development of RA [15]. Thus

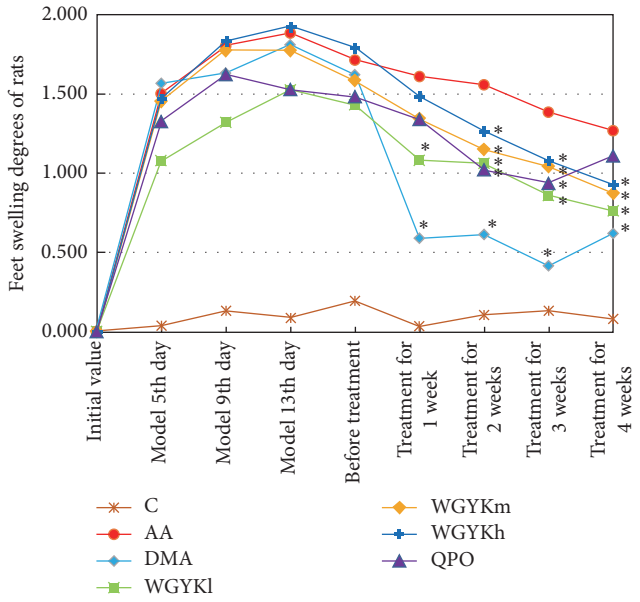


Figure 1: Changes in foot swelling degree of rats in each group: normal control (C), adjuvant arthritis (AA), dexamethasone (DMA), high dose Wuwei-Ganlu-Yaoyu-Keli (WGYKh), moderate dose Wuwei-Ganlu-Yaoyu-Keli (WGYKm), low dose Wuwei-Ganlu-Yaoyu-Keli (WGYKl), and Qing Peng ointment (QPO). Each point represents the mean ± SE. The points with black ∗ sign have significant difference (P value < 0.05) from the adjuvant arthritis group at the respective time.

we took all genes on three typical inflammatory pathways, JAK-STAT, NF-κB, and MAPK signaling pathway, as background genes to perform microarray experiment. There are totally 470 distinct genes on these three pathways. We checked the expression levels of these 470 genes in synovial tissues of 7 groups of experimental rats, that is, control group, AA model group, DMA treatment group, QPO treatment group, low dose WGYK treatment group, moderate dose WGYK treatment group, and high dose WGYK treatment group. The expression levels of all genes in AA model group were compared against the expression levels in the control group, while the expression levels of all genes in each drug treatment group were compared against the expression levels in the AA model group. In this way, we totally had six different conditions under study, that is, one condition of disease denoted as AA and five conditions of drug treatment denoted as DMA, QPO, WGYKl, WGYKm, and WGYKh, respectively. Under each condition, genes whose mean expression ratios between treatment group and control group were greater than 1.5 or less than 0.667 (1/1.5), as well as $P \leq 0.05$, were considered as differentially expressed.

In Figure 2(a) we show a comparison of the number of differentially expressed genes in the AA model and under the treatment of the five drug types. It can be seen that QPO influences the most genes, then followed by moderate dose of WGYK.

To see if the abnormally expressed genes in the disease status could be rectified by the drugs, in Figure 2(b) we show log 2 ratio values of all the 12 differentially expressed genes in the AA model and their differential expression ratios under the treatment of our drugs. It shows that, in almost all cases, the drugs could directly upregulate some genes lowly expressed in the disease status and downregulate some highly expressed genes in the disease status, with the only exception of Il2rb by QPO. This suggests the therapeutic effects of the drugs to some degree. Similar as in Figure 2(a), QPO regulates the most number of abnormal genes in disease and the moderate dose WGYK ranks the second. The moderate dose WGYK rectifies about half of the 12 abnormally expressed genes in the AA modeled disease. However, we also notice that although DMA shows the best anti-RA performance in our animal experiment, it only directly influences one abnormally expressed gene in the AA model, indicating that the drugs may exert their functions on disease by regulating other genes.

3.3. Significantly Regulated Pathways in the AA Model and under the Treatment of Drugs. To deduce the possible pathways affected by AA and the drugs, we, respectively, mapped the differentially expressed genes under different conditions onto KEGG pathways of basic biological process [16], including pathways in metabolism, organismal systems, cellular processes, environmental information processing, and genetic information processing. We conducted pathway enrichment analysis to identify the pathways significantly affected by the disease and the drugs through calculating P values for each of the pathways. Taking pathways with values of $P < 0.05$ as significantly impacted pathways, we identified pathways affected by the corresponding disease or drugs.

It comes out that only 4 pathways, that is, apoptosis, MAPK, VEGF, and T-cell receptor signaling pathway, are significantly enriched with differentially expressed genes under the condition AA, suggesting that they are affected by the AA modeling RA disease and could be dysfunctional in the disease status. Meanwhile much more pathways are significantly enriched with differentially expressed genes under most of the drug treatment conditions. In detail, the numbers of pathways affected by the treatment of QPO, WGYKm, WGYKl, WGYKh, and DMA are 25, 14, 10, 7, and 4, respectively.

The four pathways dysfunctional in the AA model have been reported to be related to the pathology of RA. Actually, earlier experiments have revealed that apoptotic pathways are defective in RA synovial tissue, resulting in "apoptosis-resistance" phenomena that few apoptotic cells can be detected in joints of RA patients [17]. MAPK signaling plays a significant role in the regulation of immune-mediated inflammatory responses and therefore it gets involved in the process of several autoimmune diseases including RA [18]. Vascular endothelial growth factor (VEGF) has been known to play angiogenic, inflammatory, and bone destructive roles in RA. In the adaptive immune response process, T-cell receptors participate in the activation of T-cells in response to foreign pathogens specifically and sensitively. Altered T-cell receptor signaling could contribute to human autoimmune arthritis, including RA [19]. Our pathway enrichment analysis shows that the five types of drugs regulate two to four of these four pathways influenced by the AA modeling

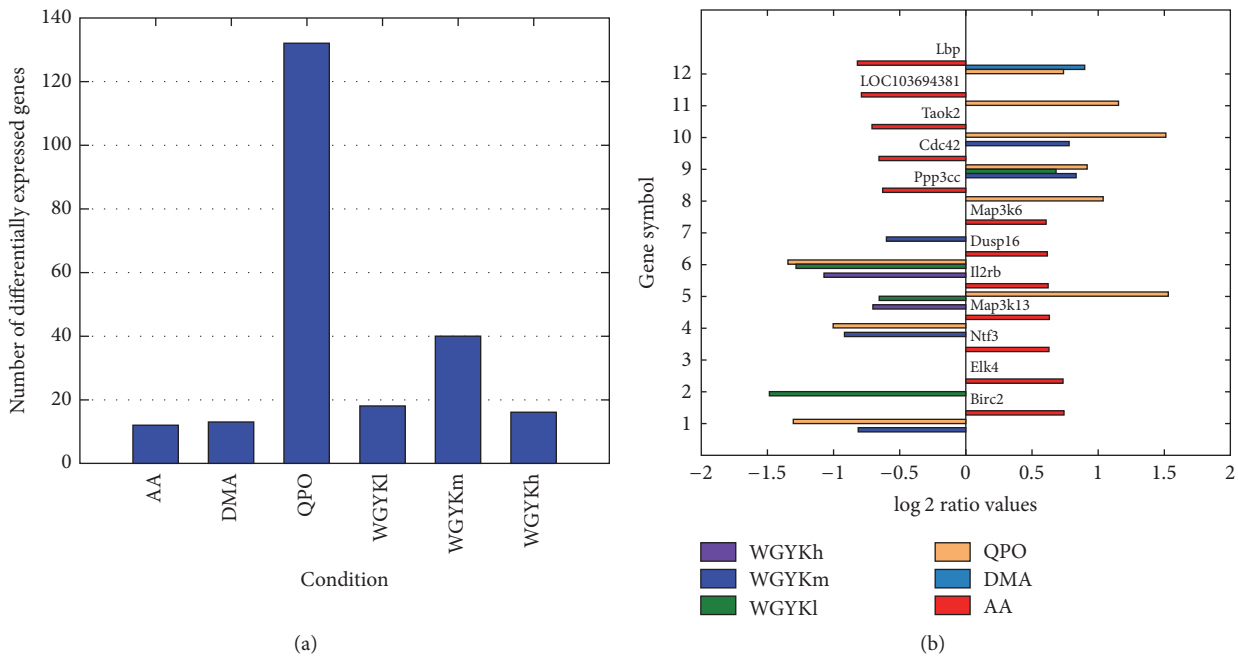


FIGURE 2: (a) Number of differentially expressed genes under different conditions. (b) Overlaps of differentially expressed genes in AA model and those under the treatment of different drugs with their corresponding log 2 ratio values.

disease. Specifically, QPO acts on all of these pathways, while moderate dose WGYK regulates three of them, suggesting anti-RA effect of these drugs.

Although only four pathways were detected to be enriched with differentially expressed genes in the AA model, much more pathways were identified being affected by the treatment of our drugs, many of which have been known being deeply involved in the initiation and progress of RA. For instance, Jak-STAT and NF-kappa B signaling pathways are proinflammatory cytokine mediated pathways related to immune-mediated inflammation and following damage of cartilage and bone in RA [18]. The osteoclast differentiation pathway maintains bone density and structure through a balance of bone resorption by osteoclasts and bone deposition by osteoblasts. Its dysfunction may disturb this balance. Both QPO and WGYKm regulate these three pathways, indicating that they are conducive to disease remission of RA.

At last, to see how moderate dose WGYK acts on the biological processes of RA, we mapped differentially expressed genes under the treatment of WGYKm, as well as pathways enriched with these genes onto the RA pathway in the KEGG database (Figure 3) [16]. It is found that WGYKm intervenes in three important pathways along the RA developing process, that is, T-cell receptor signaling pathway, VEGF signaling pathway, and osteoclast differentiation pathway. In addition, WGYKm influences two genes on the RA pathway. These results also suggest the therapeutic effect of WGYKm on RA.

3.4. Drug's Effects on AA Influenced Gene Association Network. To explore the relationships between the differentially expressed genes in different conditions and drug targets of FDA approved anti-RA drugs, we studied these genes in the context of human gene association network. We mapped all the differentially expressed rat genes onto their orthologs in human beings using Inparanoid database [20]. Thus we could check if these differentially expressed genes encode target proteins. It turns out that although 82 distinct proteins are known to be targeted by FDA approved anti-RA drugs, only three and one of their genes are differentially expressed under the treatment of QPO and moderate dose WGYK, respectively. No target genes are differentially expressed in the AA model and under the treatment of DMA and WGYK at low and high doses.

For each condition, that is, the AA model or drug treatment, we applied (3) to score the impact of its differentially expressed genes on each of the 21657 genes in the human gene association network. The higher the score is, the greater the gene is influenced by the group of differentially expressed genes. Thus a subnetwork consisting of high-score genes could be considered as a condition specific network influenced by this condition. We selected genes whose scores were top highest 1000 of the 21657 genes, that is, about top 5% of all genes in the whole network, to construct its condition specific network. In this way, we obtained one disease influenced network and five drug regulated networks, each of which owns 1000 nodes. Figure 4(a) shows how each drug regulated network overlaps with the disease influenced network. It can be seen that DMA, QPO, and moderate dose WGYK regulate significantly much more genes located at the disease influenced network than WGYK at low and high dose do, suggesting that they may have better therapeutic performance on AA.

Then, we checked how many target genes of FDA approved anti-RA drugs were included in each condition specific network. From Figure 4(b) we can see that about

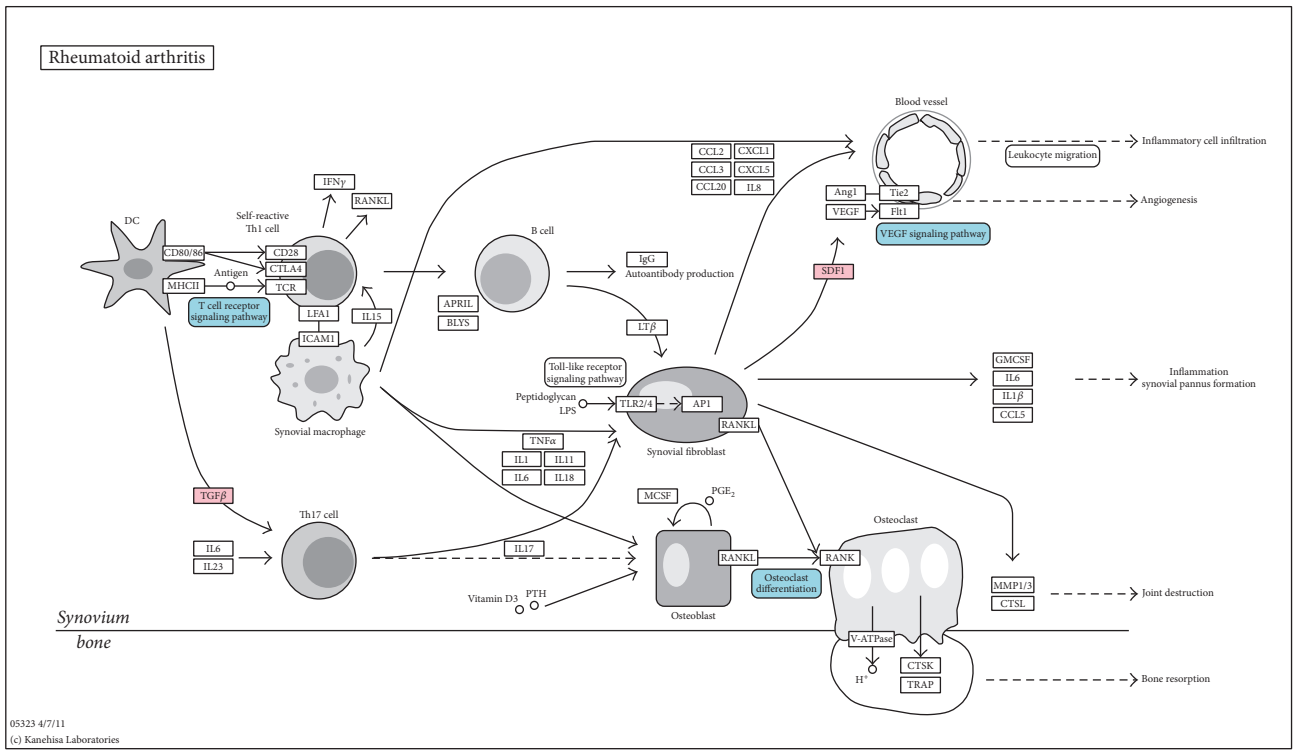


Figure 3: Regulations of WGYKm on RA pathway. Pink boxes represent differentially expressed genes under the treatment of WGYKm that appear on the RA pathway, while blue boxes represent WGYKm regulated pathways involved in the RA biological process. The original pathway map was downloaded from the KEGG database.

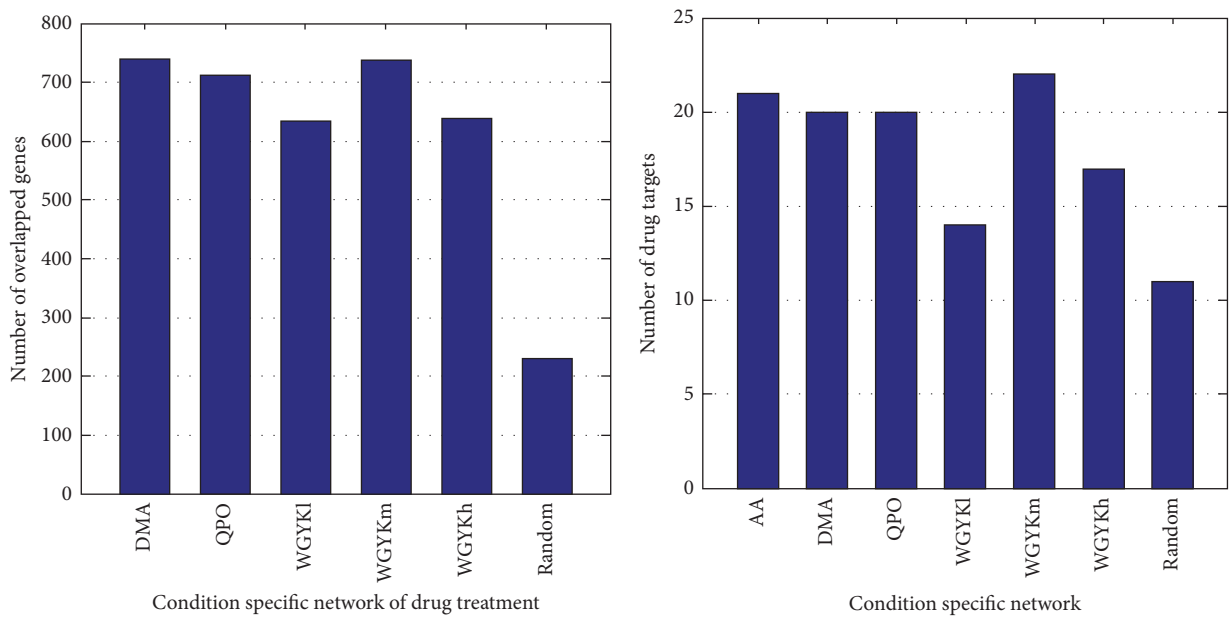


Figure 4: (a) Number of overlapped genes in different drug regulated network with the disease influenced network. (b) Number of drug targets for FDA approved anti-RA drugs included in disease influenced network and different drug regulated networks.

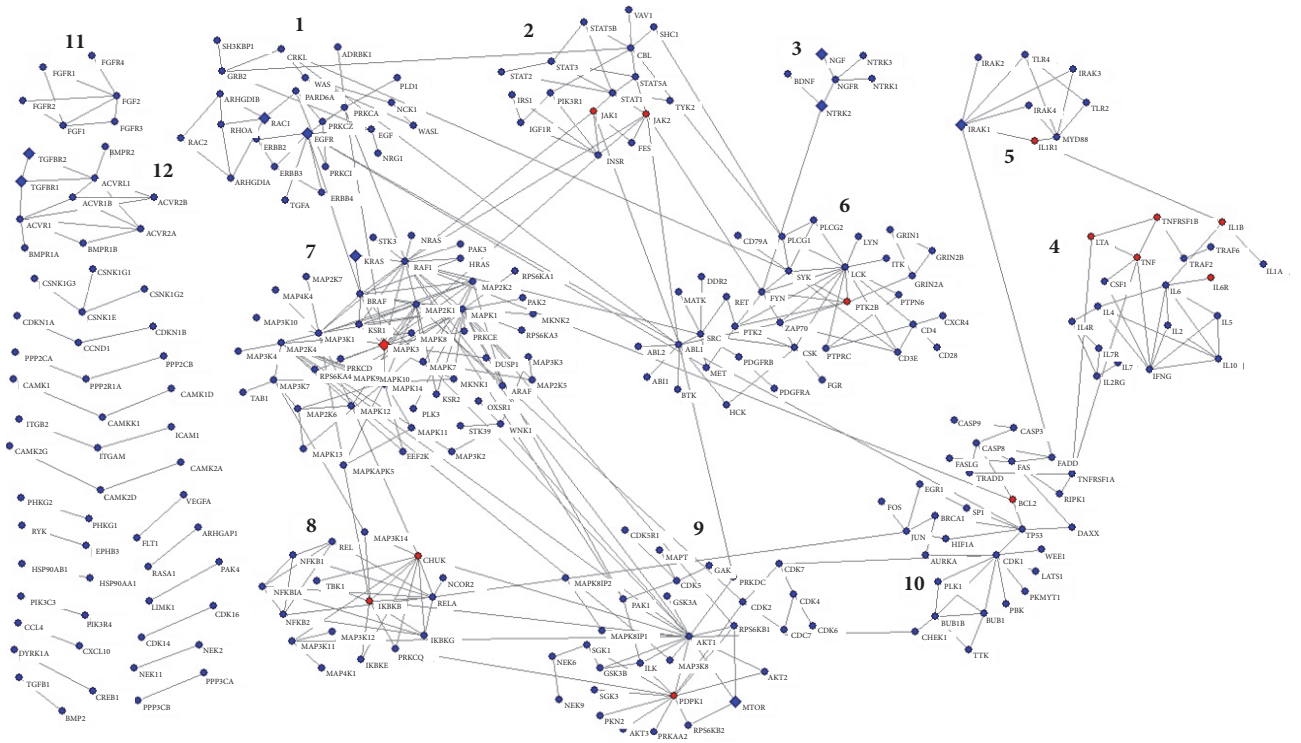


Figure 5: A gene association network regulated by AA disease and WGYKm, which includes high confidence links. The giant connected component of the network was decomposed into modules by Louvain algorithm. Diamond nodes are differentially expressed genes under the treatment of WGYKm, while red nodes are drug targets of FDA approved anti-RA drugs.

one-fourth of these target genes are included in the network at condition AA, DMA, QPO, and WGYKm, significantly more enriched than in the network at condition WGYKl and WGYKh. In addition, by checking the overlapped target genes in different condition specific networks, we found that DMA, QPO, and moderate dose WGYK regulated 16 common target genes appearing in the disease influenced network, including the most important target gene PTGS2 for nonsteroidal anti-inflammatory agents and TNF for biotech agents. These results further suggest the better performance of DMA, QPO, and moderate dose WGYK to act against AA.

We generated 100 groups of random counterpart for the 39 differentially expressed genes under the treatment of moderate dose WGYK. For each counterpart set, we repeated the two processes illustrated above. Thus we got the average number of overlapped genes affected by the counterpart genes and the AA disease, as well as the average number of drug targets for FDA approved anti-RA drugs included in the network influenced by the counterpart genes. The averages for the 100 sets of counterpart are shown in the last bars in Figures 4(a) and 4(b), respectively. These values are significantly smaller than corresponding values of the drugs, further verifying the effect of the drugs to AA.

3.5. Gene Association Network Impacted by AA Disease and WGYKm. To see how moderate dose WGYK acts on the gene association network affected by AA disease, we constructed an overlapped network of the AA affected and WGYKm regulated networks. As illustrated in the last section, both of the networks include 1000 nodes which obtained the highest impact scores from the differentially expressed genes, in which 736 nodes are overlapped. We mapped these 736 genes into the background network FLN and extracted all links between them which have confidence score larger than 0.3. Among the 22,388,609 edges of the FLN network, only 9095 have confidence score larger than 0.3, taking a percentage of around 0.04%. Thus the constructed network is a high confidence gene association network regulated by AA disease and WGYKm. This network owns 292 nodes and 448 edges. It has 22 connected clusters, in which the largest one (called giant connected component) includes 231 nodes and only two other clusters have more than 4 nodes. We will focus on these 3 connected clusters.

To investigate the biological functions of this network, we applied Louvain algorithm to decompose the giant connected component of the network into topological modules [21], so that there are much more links within modules than between modules. As shown in Figure 5, together with the other 2 larger connected clusters, this network can be considered as having 12 topological modules. Since it has been suggested that topological modules in molecular networks usually correspond to relatively independent biological functions [22], we conducted Gene Ontology (GO) enrichment analysis for each module [23]. The Gene Ontology Consortium is organized in a hierarchical way, from high level for generally descriptive terms to very low level for highly specific terms.

Table 1: Selection of the most significantly enriched and specific GO terms in the network modules.

Module	GO term (biological process)	Level	Total genes	Mapped genes
(1)	Positive regulation of interleukin-5 secretion	9	14	13
	Positive regulation of interleukin-13 secretion	9	14	13
	Positive regulation of T-helper 2 cell cytokine production	11	15	13
	Positive regulation of interleukin-10 secretion	9	15	13
(2)	JAK-STAT cascade	5	57	12
	JAK-STAT cascade involved in growth hormone signaling pathway	9	26	8
	Cytokine-mediated signaling pathway	6	406	15
(3)	Positive regulation of neuron differentiation	7	342	6
	Positive regulation of neurogenesis	7	472	6
	Positive regulation of neuron projection development	8	208	5
(4)	Positive regulation of leukocyte differentiation	6	137	19
	Positive regulation of hemopoiesis	6	166	19
	Regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	3	127	16
(5)	MyD88-dependent toll-like receptor signaling pathway	8	92	11
	Innate immune response-activating signal transduction	8	157	11
(6)	Immune response-activating cell surface receptor signaling pathway	7	350	33
	Antigen receptor-mediated signaling pathway	7	158	27
(7)	Positive regulation of MAPK cascade	10	512	39
(8)	Regulation of interleukin-12 biosynthetic process	6	16	9
	Positive regulation of type I interferon production	5	57	11
	Activation of innate immune response	7	170	13
	Positive regulation of cytokine biosynthetic process	10	87	11
	Regulation of I-kappaB kinase/NF-kappaB signaling	4	249	14
(9)	Cellular response to insulin stimulus	8	255	30
	Peripheral nervous system myelin maintenance	9	29	17
(10)	Apoptotic signaling pathway	5	310	18
(11)	Positive regulation of cell proliferation	3	964	16
(12)	Transmembrane receptor protein serine/threonine kinase signaling pathway	5	273	20
	Cellular response to growth factor stimulus	4	719	19

For each module, we selected the GO terms with the highest statistical significance and the lowest GO levels, which can represent the major and specific functions of the module (see Table 1). Table 1 suggests that most modules that WGYKm acts on participate in the regulation of immune process, including pathogen recognition, proinflammatory response and inflammatory signaling in innate immune defenses (modules (1), (2), (4), (5), (6), (7), (8), and (12)), innate immune response (modules (5) and (8)), and adaptive immune response (module (4)). WGYKm also regulates cell proliferation and differentiation by influencing modules (3), (10), and (11).

Figure 5 shows that half of the modules include differentially expressed genes under the treatment of WGYKm, while 8 modules own genes that encode drug targets for FDA approved anti-RA drugs. We list these drug targets and corresponding anti-RA drugs in Table 2. Targets of 3 classes of anti-RA drugs, nonsteroidal anti-inflammatory drugs (NSAID), disease-modifying antirheumatic drugs (DMARDs), and biotechnology agents are included in different modules of the network. This result also suggests WGYKm's effect in anti-inflammation and regulation of immune process.

4. Conclusions

This work systematically studies the anti-RA mechanism of Tibetan medicated-bath therapy using Wuwei-Ganlu-Yaoyu-Keli. First, we performed animal experiment to verify that the bath therapy by different doses of WGYK exhibited similar effect of relieving foot swelling of adjuvant arthritis model

Table 2: Genes encoding drug targets for FDA approved anti-RA drugs that appear in the network modules.

Module	Drug target	Drug	Drug class
(2)	JAK1	Tofacitinib	DMARDs
	JAK2	Tofacitinib	DMARDs
(4)	IL1B	Canakinumab	Biotech agents
	IL6R	Tocilizumab	Biotech agents
	LTA	Etanercept	Biotech agents
	TNF	Etanercept	Biotech agents
		Adalimumab	Biotech agents
		Infliximab	Biotech agents
		Golimumab	Biotech agents
		Certolizumab pegol	Biotech agents
		Chloroquine	DMARDs
	TNFRSF1B	Etanercept	Biotech agents
(5)	IL1R1	Anakinra	Biotech agents
(6)	PTK2B	Leflunomide	DMARDs
(7)	MAPK3	Sulindac	NSAIAs
(8)	CHUK	Sulfasalazine	DMARDs
	IKBKB	Sulfasalazine	DMARDs
		Auranofin	DMARDs
(9)	PDPK1	Celecoxib	NSAIAs
(10)	BCL2	Ibuprofen	NSAIAs

rats as positive contracts dexamethasone and Qing Peng ointment did. Then, based on differentially expressed genes in the disease status and under the treatment of different drugs, we investigated the effects of the bath therapy on RA in the contexts of single genes, pathways, and networks, respectively. We found that the drugs could directly upregulate some lowly expressed genes and downregulate some highly expressed genes in the AA modeling disease status, in which moderate dose WGYK rectified about half of the 12 abnormally expressed genes in the disease status. Our pathway enrichment analysis revealed that moderate dose WGYK regulated three of the four pathways influenced by the disease. It also intervened in three of the five important pathways along the RA developing process recorded in the KEGG database. By scoring the impacts of abnormally expressed genes on all genes in the human gene association network, we constructed subnetworks influenced by the disease and regulated by the drugs, respectively. It comes out that the subnetwork regulated by moderate dose WGYK has more than 70% nodes overlapped with the subnetwork influenced by the AA modeling disease. In addition, 16 common target genes of FDA approved anti-RA drugs appear in the disease influenced subnetwork and the subnetworks regulated by moderate dose WGYK, dexamethasone, and Qing Peng ointment, including the most important target gene PTGS2 for nonsteroidal anti-inflammatory agents and TNF for biotech agents. Finally, we constructed a gene association network regulated by AA disease and WGYKm, which only includes high confidence links. Our GO analysis for topological modules of this network suggests that WGYKm performs its therapeutic effect on RA by regulating immune process, as well as cell proliferation and differentiation. All these results support the anti-RA effect of Tibetan medicated-bath therapy using Wuwei-Ganlu-Yaoyu-Keli at moderate dose.

This work applies network approach to explain WGYK's antirheumatic effect. It may shed light on the study about the pharmacology of Tibetan medicated-bath therapy and promote the development of traditional medicine.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Tianhong Wang, Kehui Zhao, and Jing Wang carried out animal experiment. Jian Yang and Xing Chen collected data and conducted computation. Tianhong Wang and Jing Zhao drafted the manuscript. Yang Ga, Jing Zhao, and Yi Zhang conceived of the study, provided overall guidance for this project, and edited the manuscript. All authors have read and approved the final manuscript.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant nos. 81260672 and 61372194 and Chongqing Education Reform Project of Graduate (yjg152017). The authors thank Ms. Xiaona Shen for her help in editing the manuscript.

References

[1] I. B. McInnes and G. Schett, "The pathogenesis of rheumatoid arthritis," *The New England Journal of Medicine*, vol. 365, no. 23, pp. 2205–2219, 2011.

[2] S. Zhao, P. Geng, and J. Sang, "Immune function of rheumatoid arthritis treated by medicated-bath therapy in Tibetan medicine," *Zhongguo Zhong Xi Yi Jie He Za Zhi*, vol. 13, pp. 452-453, 1993.

[3] *Drug standard of Ministry of Health of People's Republic of China Tibetan Medicine: the first volume*, vol. 1, Pharmacopoeia Committee of the Ministry of health of People's Republic of China, 1995.

[4] H.-m. Li and B.-l. Li, "Experimental research on effect of qingpeng paste treating adjuvant arthritis in rats," *Chinese Journal of Experimental Traditional Medical Formulae*, vol. 17, no. 6, pp. 228–231, 2011.

[5] M. Verhoef, J. A. G. Van Roon, M. E. Vianen, F. P. J. G. Lafeber, and J. W. J. Bijlsma, "The immune suppressive effect of dexamethasone in rheumatoid arthritis is accompanied by upregulation of interleukin 10 and by differential changes in interferon γ and interleukin 4 production," *Annals of the Rheumatic Diseases*, vol. 58, no. 1, pp. 49–54, 1999.

[6] P. B. Jacobson, S. J. Morgan, D. M. Wilcox et al., "A new spin on an old model: In vivo evaluation of disease progression by magnetic resonance imaging with respect to standard inflammatory parameters and histopathology in the adjuvant arthritic rat," *Arthritis and Rheumatism*, vol. 42, no. 10, pp. 2060–2073, 1999.

[7] J. Zhao, P. Jiang, and W. Zhang, "Molecular networks for the study of TCM pharmacology," *Briefings in Bioinformatics*, vol. 11, no. 4, pp. 417–430, 2009.

[8] D. K. Slonim and I. Yanai, "Getting started in gene expression microarray analysis," *PLoS Computational Biology*, vol. 5, no. 10, Article ID e1000543, 2009.

[9] M. O'Mahony, "Sensory Evaluation of Food: Statistical Methods and Procedures," p. 487, CRC Press, 1986.

[10] B. Linghu, E. S. Snitkin, Z. Hu, Y. Xia, and C. DeLisi, "Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network," *Genome Biology*, vol. 10, no. 9, article R91, 2009.

[11] D. S. Wishart, C. Knox, A. C. Guo et al., "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Research*, vol. 34, pp. D668–D672, 2006.

[12] R. K. Curtis, M. Orešič, and A. Vidal-Puig, "Pathways to the analysis of microarray data," *Trends in Biotechnology*, vol. 23, no. 8, pp. 429–435, 2005.

[13] J. Zhao, G.-H. Ding, L. Tao et al., "Modular co-evolution of metabolic networks," *BMC Bioinformatics*, vol. 8, article no. 311, 2007.

[14] P. Li, G. Xie, S. Song et al., "Clinical manifestations and the main evaluation method on adjuvant-induced arthritis model in rats," *Chinese Journal of Immunology*, vol. 28, pp. 453–457, 2012.

[15] F. M. Brennan and I. B. McInnes, "Evidence that cytokines play a role in rheumatoid arthritis," *Journal of Clinical Investigation*, vol. 118, no. 11, pp. 3537–3545, 2008.

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

[17] H. Liu and R. M. Pope, "The role of apoptosis in rheumatoid arthritis," *Current Opinion in Pharmacology*, vol. 3, no. 3, pp. 317–322, 2003.

[18] C. J. Malemud, "Intracellular Signaling Pathways in Rheumatoid Arthritis," *Journal of Clinical & Cellular Immunology*, vol. 4, no. 160, 2013.

[19] S. Sakaguchi, H. Benham, A. P. Cope, and R. Thomas, "T-cell receptor signaling and the pathogenesis of autoimmune arthritis: Insights from mouse and man," *Immunology and Cell Biology*, vol. 90, no. 3, pp. 277–287, 2012.

[20] E. L. L. Sonnhammer and G. Östlund, "InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic," *Nucleic Acids Research*, vol. 43, no. 1, pp. D234–D239, 2015.

[21] V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, Article ID P10008, 2008.

[22] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[23] C. Gene Ontology, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Res*, vol. 32, no. suppl_1, pp. D258–D261, 2004.

Hindawi
BioMed Research International
Volume 2017, Article ID 3035481, 9 pages
https://doi.org/10.1155/2017/3035481

Research Article

Functional Virtual Flow Cytometry: A Visual Analytic Approach for Characterizing Single-Cell Gene Expression Patterns

Zhi Han,[1,2] Travis Johnson,[2] Jie Zhang,[2,3] Xuan Zhang,[1] and Kun Huang[2]

[1]*College of Software, Nankai University, Tianjin, China*
[2]*Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA*
[3]*The CCC Biomedical Informatics Shared Resource, The Ohio State University, Columbus, OH, USA*

Correspondence should be addressed to Kun Huang; kun.huang@osumc.edu

Received 3 March 2017; Accepted 22 May 2017; Published 17 July 2017

Academic Editor: Ansgar Poetsch

We presented a novel workflow for detecting distribution patterns in cell populations based on single-cell transcriptome study. With the fast adoption of single-cell analysis, a challenge to researchers is how to effectively extract gene features to meaningfully separate the cell population. Considering that coexpressed genes are often functionally or structurally related and the number of coexpressed modules is much smaller than the number of genes, our workflow uses gene coexpression modules as features instead of individual genes. Thus, when the coexpressed modules are summarized into eigengenes, not only can we interactively explore the distribution of cells but also we can promptly interpret the gene features. The interactive visualization is aided by a novel application of spatial statistical analysis to the scatter plots using a clustering index parameter. This parameter helps to highlight interesting 2D patterns in the scatter plot matrix (SPLOM). We demonstrated the effectiveness of the workflow using two large single-cell studies. In the Allen Brain scRNA-seq dataset, the visual analytics suggested a new hypothesis such as the involvement of glutamate metabolism in the separation of the brain cells. In a large glioblastoma study, a sample with a unique cell migration related signature was identified.

1. Background

Single-cell RNA sequencing (scRNA-seq) is becoming a powerful tool for studying heterogeneity and subtypes in cell populations. Many bioinformatics and computational tools have been developed to visualize, cluster, and categorize the cells based on their expression profiles [1, 2]. Different algorithmic approaches such as principal component analysis (PCA) or multidimensional scaling (MDS) [3], nonnegative matrix factorization [4], minimum spanning tree (MST) [5, 6], latent variable modeling [7], diffusion map [8, 9], and spline models [10] have all been applied and implemented for such purposes. Moreover, it has been shown that often the cells in a population do not always form "clusters." Instead, the cells form a continuous distribution over the space of featured genes and gene signatures [1]. Therefore, it is often of great interest to identify the interesting distribution patterns (e.g., wishbone pattern and bifurcation) which often imply important biological processes such as stem cell differentiation as well as the gene signatures that can be used to reveal such patterns.

However, this effort often leads to a "chicken-and-egg" situation. Since the patterns may not always be readily perceivable from whole genome data, methods such as PCA and MDS may not always be effective. Therefore, it often ends up in an iterative process and a subjective selection of genes of interests. Another commonly adopted workflow is to first cluster the cells based on their expression profiles and identify "gene signatures" that differentiate the clusters followed by enrichment analysis on these signature genes for potential biological functions or processes involved in the separation of the cells. Since there could be many genes involved in differential analysis, the functional enrichment signals can be diluted.

In this paper, we propose a visual analytic workflow called functional virtual flow cytometry (FVFC) for identifying functional gene groups that can effectively separate the cells using scRNA-seq data. We specifically take advantage of

gene coexpression network analysis (GCNA). GCNA aims to identify modules of genes with similar expression profiles. It has been well known that the coexpressed genes often are functionally or structurally related [11–16]. Therefore, instead of surveying all the genes, by focusing on the coexpressed gene clusters, we can directly study the cells based on functional gene groups with increased statistical power [17].

Our method is innovative in the following ways. First, it focuses on the gene modules with clear functional relationships (coexpression) and thus greatly enhances the statistical power. Secondly, only the gene modules that are "informative" among the single cells are used. Specifically we focus on the modules that show bimodal or multimodal distributions among the cells to ensure separation power of the genes on the cell population. Thirdly, we apply spatial statistical methods to detect combinations of gene modules that lead to interesting spatial patterns or separation of the cells and thus identify the gene signatures associated with the underlying biological processes. Last but not least, instead of developing this workflow as an "algorithm," we implement it as a visual analytic workflow, allowing the researchers to interactively select gene modules and cell distribution patterns of interest for further investigation. To this end, we take advantage of the SPLOM combined with various visual cues derived from spatial statistical calculation. We demonstrate our workflow using two large single-cell studies on brain and cancer, respectively.

2. Methods

2.1. Workflow. Figure 1 outlines the workflow of our approach that contains three stages. Given a set of processed scRNA-seq data, the first stage carries out the coexpression network analysis and summarization of each network module into a single "eigengene" as well as enrichment analysis to determine the function or structural relationships for each module. The second stage analyzes each eigengene to select the ones with more information content, in particular, the bimodal ones. Then scatterplots are generated for every pair of informative eigengenes. The scatterplots are further analyzed using spatial statistical parameters to determine if they form interesting patterns, specifically if there is clustering or clumping in the scatterplot, implying potential relationships between the two gene modules associated with the two eigengenes. In the final stage, the scatterplots are colored based on the spatial statistical parameters and interesting patterns are further examined with their functional relevance. Overall, this workflow provides an intuitive visual analytic approach for researchers to quickly explore the relationships among functional gene groups in single-cell populations. The details of the steps in the workflow are discussed in the following sections.

2.2. Weighted Gene Coexpression Network Analysis. The first stage in Figure 1 is to carry out gene coexpression network analysis. The detailed workflow for this stage is illustrated in Figure 2. Given a set of M genes and their expression levels over N cells, the gene expression profile can be expressed with a matrix

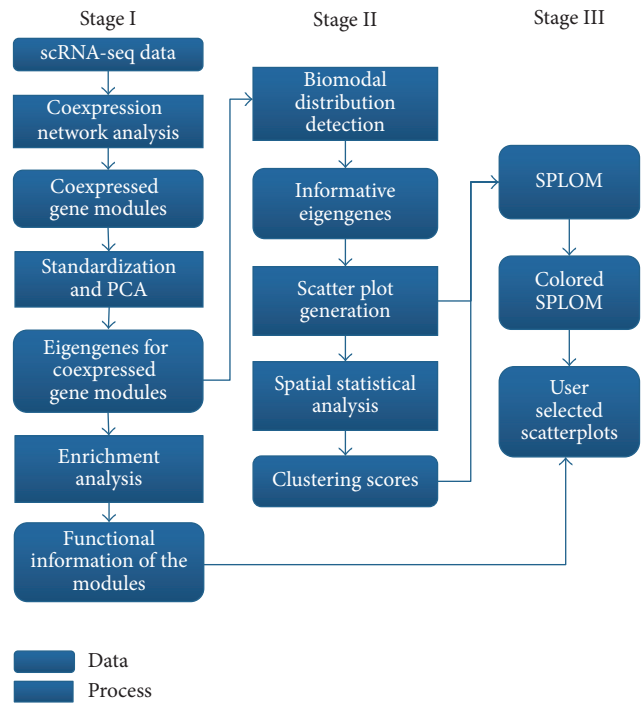


Figure 1: The workflow of the functional virtual flow cytometry system.

$$\mathbf{G} = \begin{bmatrix} g_{11} & \cdots & g_{1N} \\ \vdots & \ddots & \vdots \\ g_{M1} & \cdots & g_{MN} \end{bmatrix} = \begin{bmatrix} \mathbf{g_1} \\ \vdots \\ \mathbf{g_M} \end{bmatrix}, \quad (1)$$

where the N-dimensional row vector $\mathbf{g}_i = [g_{i1} \cdots g_{iN}]$ is the expression profile for the ith gene across the samples ($i = 1, 2, \ldots, N$). Then the pairwise correlation matrix C can be represented by

$$C = \begin{bmatrix} c_{11} & \cdots & c_{1M} \\ \vdots & \ddots & \vdots \\ c_{M1} & \cdots & c_{MM} \end{bmatrix}, \quad (2)$$

where c_{ij} is the correlation coefficient between ith gene vector $\mathbf{g}_i$ and jth gene vector $\mathbf{g}_j$. In our experiment, we use Spearman rank correlation coefficients in the pairwise correlation matrix since Gaussian distribution cannot be assumed for RNA-seq data as required by Pearson correlation.

After the correlation matrix was computed, we apply a recently developed algorithm called Normalized lmQCM [15]. Compared to widely adopted gene coexpression network analysis software package WGCNA [18], this algorithm takes a network mining approach allowing overlaps between modules and also is guaranteed to have a lower bound on the density of the detected modules. The output of algorithm lmQCM is a set of gene modules $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_L$, where each module $M_k = \{i_1, i_2, \ldots, i_{N_k}\}$ is composed of a group of N_k coexpressed genes. The number of modules L and the sizes of the modules are determined by the four parameters of the

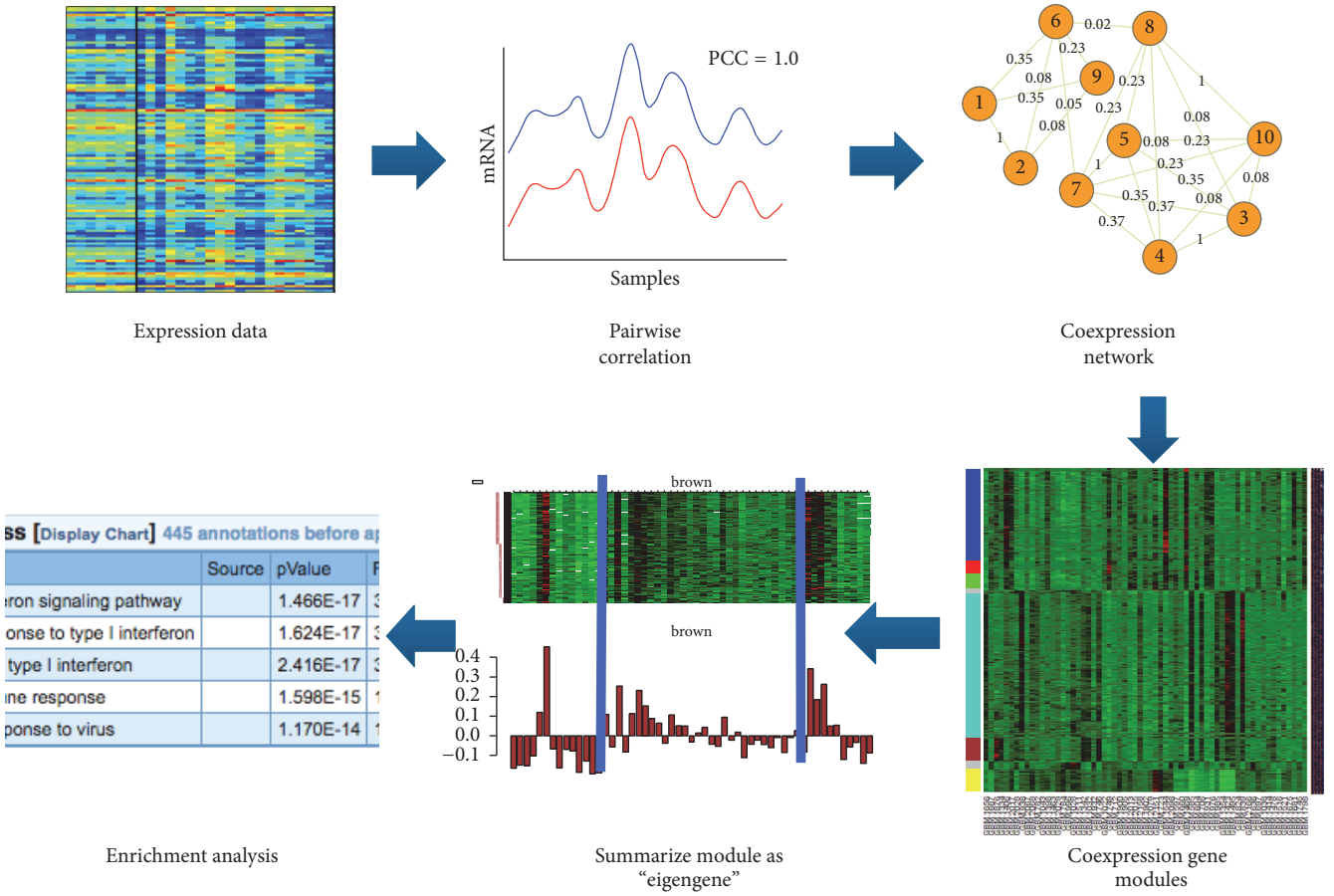


FIGURE 2: Workflow for weighted GCNA and eigengene calculation.

lmQCM algorithm. While detailed choice of parameters was discussed in [15], the most important parameter is γ, which is the threshold for the weight of the first edge of any module and thus controls the number of modules. Usually we choose γ to ensure that the maximum size of a module is not too large (i.e., less than 500 genes). In addition, we focus on gene modules with at least 10 genes so that meaningful functional enrichment analysis can be applied.

For each gene module detected by lmQCM, M_k can be represented by a gene expression matrix. If we want to compare one gene module against another, it is advantageous to take only a representative of that module rather than taking all the genes. We use PCA to reduce the gene module data meaningfully and take the first principal component as a summary of that module. This first principal component is called "eigengene" in this context. Computationally, we take the submatrix of $\mathbf{G}$ for M_k as

$$\mathbf{G_k} = \begin{bmatrix} \mathbf{g_{i_1}} \\ \vdots \\ \mathbf{g_{i_{N_k}}} \end{bmatrix} \in \Re^{N_k \times N}. \quad (3)$$

$\mathbf{G_k}$ is centralized and standardized as $\mathbf{G_k'}$ such that for each row the mean is zero and the norm is one. Let $\mathbf{G_k'} = \mathbf{USV^T}$ be the singular value decomposition of $\mathbf{G_k'}$. Then the first column of $\mathbf{V}$ (denoted as $\mathbf{v_1}$) is the "eigengene" for $\mathbf{M_k}$ up to a sign since $\mathbf{V}$ is an orthonormal matrix whose determinant is 1 or −1. Since the eigengene $\mathbf{w_k}$ should reflect the directions of the majority of genes in $\mathbf{G_k}$, its projection on the majority of the genes should be positive. Thus, if $\sum \text{sgn}(\mathbf{G_k v_1}) < 0$, then $\mathbf{w_k} = -\mathbf{v_1}$; otherwise, $\mathbf{w_k} = \mathbf{v_1}$. So each gene module detected by lmQCM corresponds to one "eigengene."

For the reported modules, enrichment analyses are carried out using NIH DAVID (https://david.ncifcrf.gov/) [19] and TOPPGene (https://toppgene.cchmc.org/enrichment.jsp) [20].

2.3. Identify Eigengenes with Bimodal or Long Tail Distribution. Before exploring pairwise relationships between gene modules with eigengenes, we identify and keep eigengenes which are "informative," that is, eigengenes whose distribution follows a bimodal or long tail distribution. Therefore, eigengenes with unimodal distribution, especially the ones with narrow sharp peak-shaped distribution, will be filtered out. To differentiate unimodal distribution with bimodal or long tail distributions, metrics such as Kurtosis, second central differences, and likelihood ratio are adopted [21–23]. Specifically, Kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable [24]. Here we use Kurtosis as a measure to filter whether the histogram of a given eigengene has a very narrow sharp peak

Table 1: The seven gene modules whose eigengenes show long tail distributions.

Eigengene #	Index	Size	Kurtosis	Enrichment/notes
1	3	38	10.7844	32 predicted genes: three genes are immunoglobulins and two are T cell receptors, acute lymphocytic leukemia ($p = 3.157e-7$)
2	6	35	5.0379	Ion transport ($p = 3.341e-7$), synapse ($p = 2.590e-7$)
3	12	18	8.5550	Glutamate decarboxylation to succinate ($p = 7.715e-7$), inhibitory synapse ($p = 7.843e-7$)
4	13	17	19.9492	Development of lower uro neuro e15.5 BladdPelvicGanglion Sox10 top-relative-expression-ranked 1000 ($1.227e-7$), six genes on chromosome X
5	28	11	4.9068	Hydrogen ion transmembrane transport ($p = 4.859e-20$), mitochondrial inner membrane ($p = 1.533e-16$)
6	48	6	3.8686	NADH metabolic process ($p = 2.960e-13$), myelin sheath ($p = 1.643e-3$), gluconeogenesis ($p = 5.401e-14$), genes upregulated in hippocampus at late postnatal stages ($p = 9.341e-10$)
7	60	5	12.5680	Mostly predicted genes

distribution. For each eigengene vector $\mathbf{w_k}$, first the histogram of the vector is computed and then Kurtosis of the histogram distribution is computed as

$$\text{Kurt}\left(\mathbf{w_k}\right) = \frac{E\left[\left(\mathbf{w_k} - \mu\right)^4\right]}{\left(E\left[\left(\mathbf{w_k} - \mu\right)^2\right]\right)^2}, \tag{4}$$

where μ is mean of $\mathbf{w_k}$. In [24], the Kurtosis value between 3 and 9 show peakness of the distribution while higher values imply sharper peak-shaped distribution. In this paper, we set the threshold for Kurtosis as user defined parameter. If Kurtosis value of histogram for a given eigengene is smaller than a given threshold, then eigengene will be kept.

2.4. Spatial Statistical Analysis of the 2D Scatterplot Using the Nearest Neighbor Distribution. In order to find the relationship between two coexpressed gene modules, we generate pairwise scatter plots for all pairs of eigengene vectors in a 2D space. For two given eigengene vectors $e_i = [e_{i1}, e_{i2}, \ldots, e_{iN}]$ and $e_j = [e_{j1}, e_{j2}, \ldots, e_{jN}]$, scatter plot is the points with coordinates $(e_{i1}, e_{j1}), (e_{i2}, e_{j2}), \ldots, (e_{iN}, e_{jN})$ in the 2D space. Then we use the nearest neighbor distance (NND) to analyze the pattern. NND for a data point is the distance to its closest neighbor. It is a spatial statistical parameter effectively used for detecting cell patterns in the space [25, 26]. Define $\overline{d_0}$ as the mean NND for all the points. Then we make 100 random simulations, each time the same number of points is created in the same region covering $(e_{i1}, e_{j1}), (e_{i2}, e_{j2}), \ldots, (e_{iN}, e_{jN})$, and the *mean NND* is calculated. Assuming that $\overline{d_E}$ is the mean of 100 randomly simulated mean NND and $\overline{\sigma}$ is the standard variation, the z-score is calculated as

$$z = \frac{\overline{d_0} - \overline{d_E}}{\overline{\sigma}}. \tag{5}$$

We call the z-score as the *clustering index* for a scatter plot.

2.5. Layout for Visualization. Once the eigengenes with long tail or bimodal distributions are detected, SPLOM is generated. Each scatterplot is then colored using the color scale based on the clustering index. User can then select plots with interesting patters for further visualization and analysis.

3. Results

3.1. Datasets and Preprocessing. We applied the above analysis to two large gene expression single-cell datasets. One dataset is RNA sequencing data of single cells isolated from mouse dorsal lateral geniculate nucleus (dLGN) of the thalamus, which is downloaded from Allen Brian Atlas (ABA) website. This data set includes 1,772 single cells collected from dLGN in adult mouse and transcriptionally profiled with RNA sequencing. The dataset contains transcription readings for 45,772 genes and transcripts. However, since many of the genes have zero readings in most cells, these genes were filtered out; specifically we removed genes with zeros in more than half of the cells. In addition, genes whose mean values are among the lowest 20% and variances are among the lowest 50% were removed. This way, 20,000 genes were retained for further analysis.

Another dataset is from a single-cell study on human glioblastoma. The dataset was downloaded from NCBI Gene Expression Omnibus (GEO) with accession number GSE57872. It contains transcriptomes from 430 single glioblastoma cells isolated from 5 individual tumors and 102 single cells from gliomasphere cells lines generated using SMART-seq [27]. Using the same preprocessing procedure, 5,948 genes were kept for further analysis in this dataset.

3.2. Analysis of the ABA Mouse Brain scRNA-Seq Data. Using the lmQCM algorithm (with $\gamma = 0.75$), 60 coexpressed gene modules with at least five genes are identified. Using a threshold 20 for the Kurtosis metric, seven eigengenes are selected. Table 1 summarizes the information for the seven

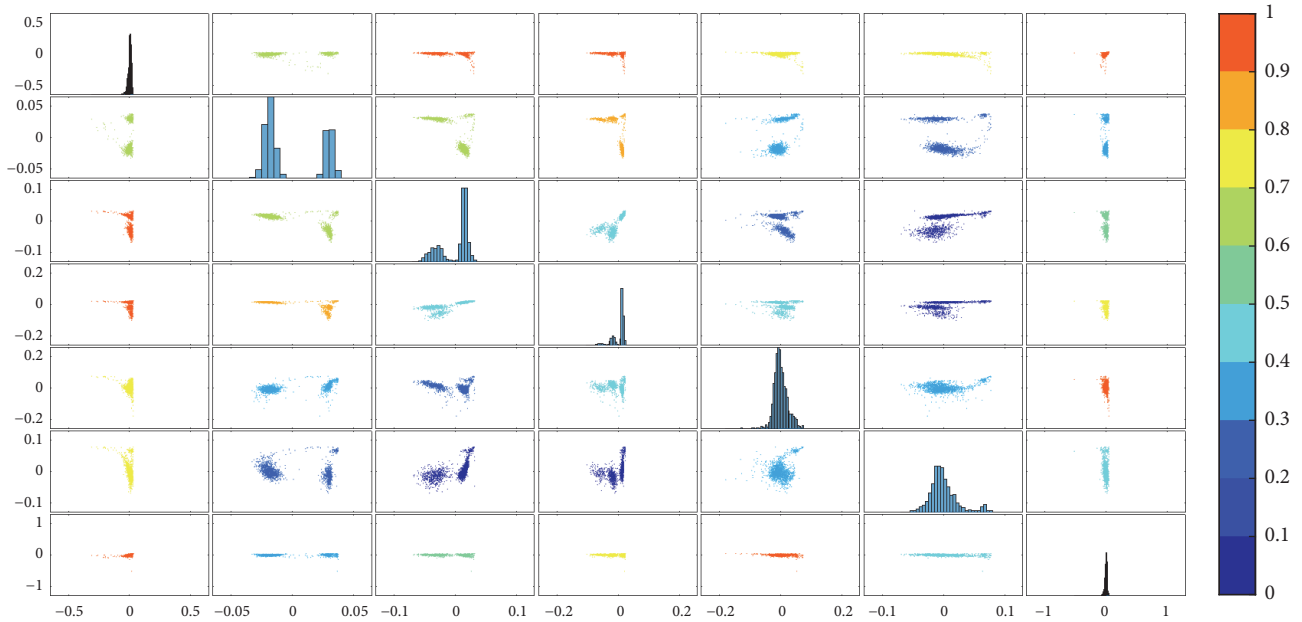


Figure 3: Colored SPLOM for the seven long tail eigengenes from the Allen Brain scRNA-seq data. The subplot in the ith row, jth column of the matrix is a scatter plot of the ith eigengene against the jth eigengene. Along the diagonal are histogram plots of each eigengene.

modules. Figure 3 shows the colored SPLOM for the seven eigengenes.

The color scheme in Figure 3 allows us to further inspect scatterplots with interesting patterns. In order to determine if these patterns are associated with specific annotations, for selected scatter plots, we further overlay the annotation information using different colors. Figure 4 shows examples when the broad subtype information about the neurons is overlaid on the scatter plots as points with different colors. It is apparent that none of the gene modules can thoroughly separate the cells based on the subtypes. Instead, some of them can separate specific subtypes. For instance, as in Figure 4(a), the cells are separated into two major clusters based on the "clustering index" as defined in the previous section, which does not fully reflect the subtypes as the blue and yellow points are not separated. Instead, the blue and yellow points are segmented in Figure 4(b) and even further away in Figure 4(c).

As in Figure 4(b), it is clear that the groups of yellow cells and cyan cells are separated from the rest groups based on eigengene #4 that is enriched with genes that are important to bladder/pelvic ganglion development and may be involved in gender development too. At the same time, it can be noted that the red group is different from the blue, cyan, and yellow groups based on eigengene #2 that is closely associated with synapse formation. In addition, according to Figure 4(c), the blue and yellow groups are separated when both eigengenes #3 and #4 are involved and eigengene #3 is closely connected with the glutamate metabolism and inhibitory synapse development. These neural functions are critical for the interpretation of the cell population clustering.

It is important to notice that the visual outcome is very different from traditional PCA based visualization. As shown in Figure 5(a), if all the genes are used for visualization of the cells using traditional PCA, there is not a clear separation of the cells except for a small group. If we limit the gene features for PCA to the ones involved only in the gene modules listed in Table 1, we can clearly see three major groups. As a control, we marked the three groups of cells in Figure 5(a) with three different colors, and we can see that there is no clear separation of the cells in Figure 5(a). However, without explicit functional grouping, it is difficult to determine which biological processes and functions are involved in such separation.

3.3. Analysis of the Human Glioblastoma Patients' Brain scRNA-Seq Data. Using the lmQCM algorithm (with $\gamma = 0.2$), 18 coexpressed gene modules with at least five genes are identified. Using a threshold of 5 for the Kurtosis metric, 16 eigengenes are selected.

Figure 6 is the SPLOM for the long tail eigengenes from the brain tumor study.

From the SPLOM, it is notable that the fourth gene module not only has an eigengene with bimodal distribution but also is involved in effective separation of the cells. While the cells are labeled by the patient and sample IDs, it is clear that some of the separation cases are closely related to the differences between different tumor samples as shown in Figure 7. In particular, eigengene #4 is key in separating the cells in the green group from the rest while other eigengenes can separate other groups (e.g., eigengene #6 separates the yellow cell group from the rest while eigengene #11 separates the red cell group). Interestingly enrichment analysis shows that this gene module for eigengene #4 is highly enriched with extracellular matrix genes (14 genes out of 36, $p = 6.304e-8$) and the cell migration process (10 genes, $p = 9.145e-5$), suggesting a particular property of the cells in the green group and it is important as the extracellular matrix and cell

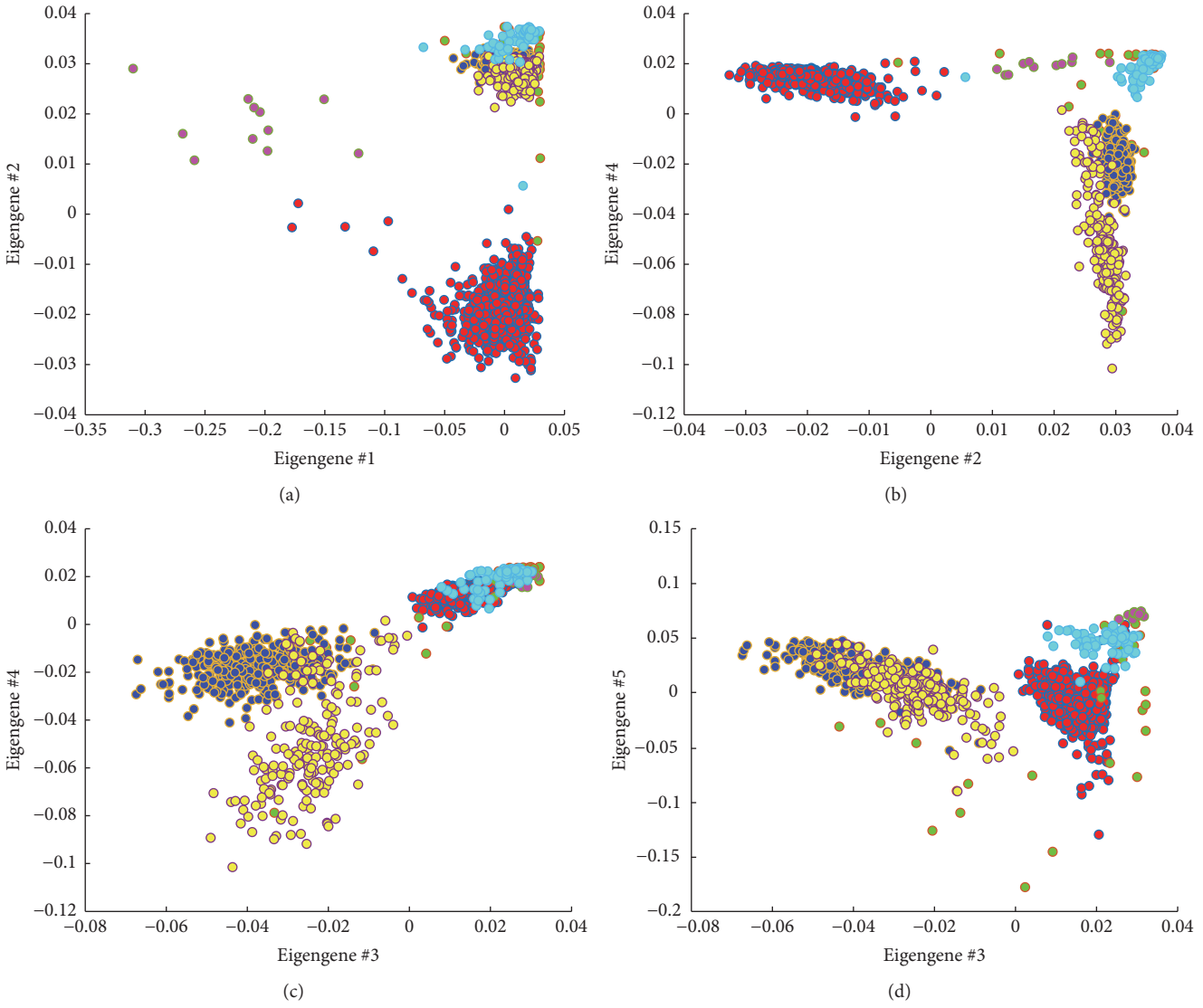


FIGURE 4: Four example scatterplots with broad classes annotated in different colors.

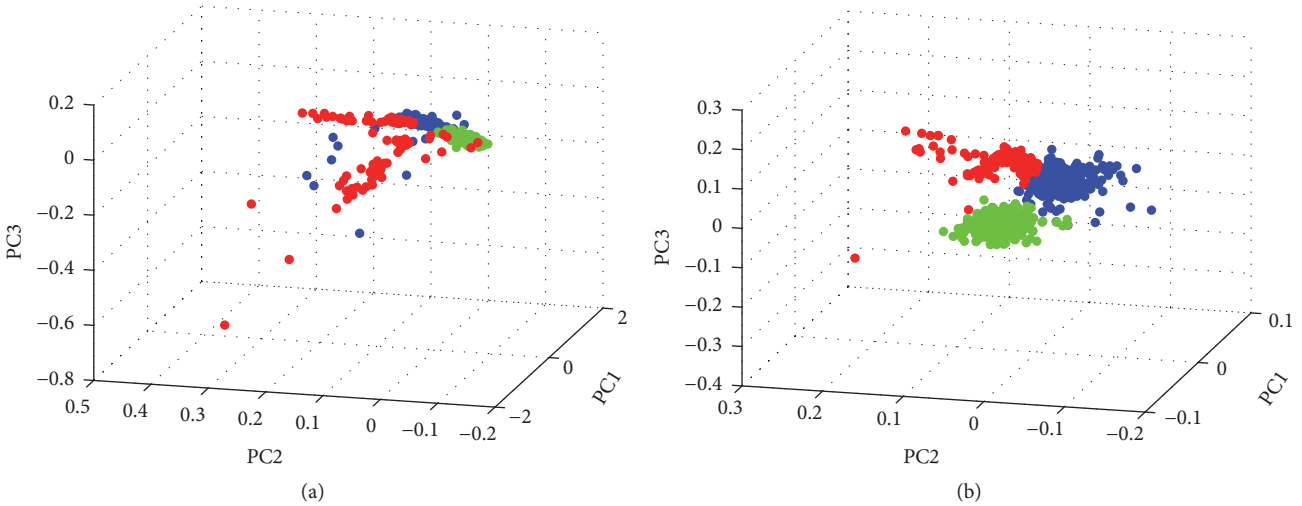


FIGURE 5: (a) The 3D plot for the first three principal components using all genes for the cells. (b) The 3D plot for the first three principal components using genes in the gene modules in Table 1 for the cells.

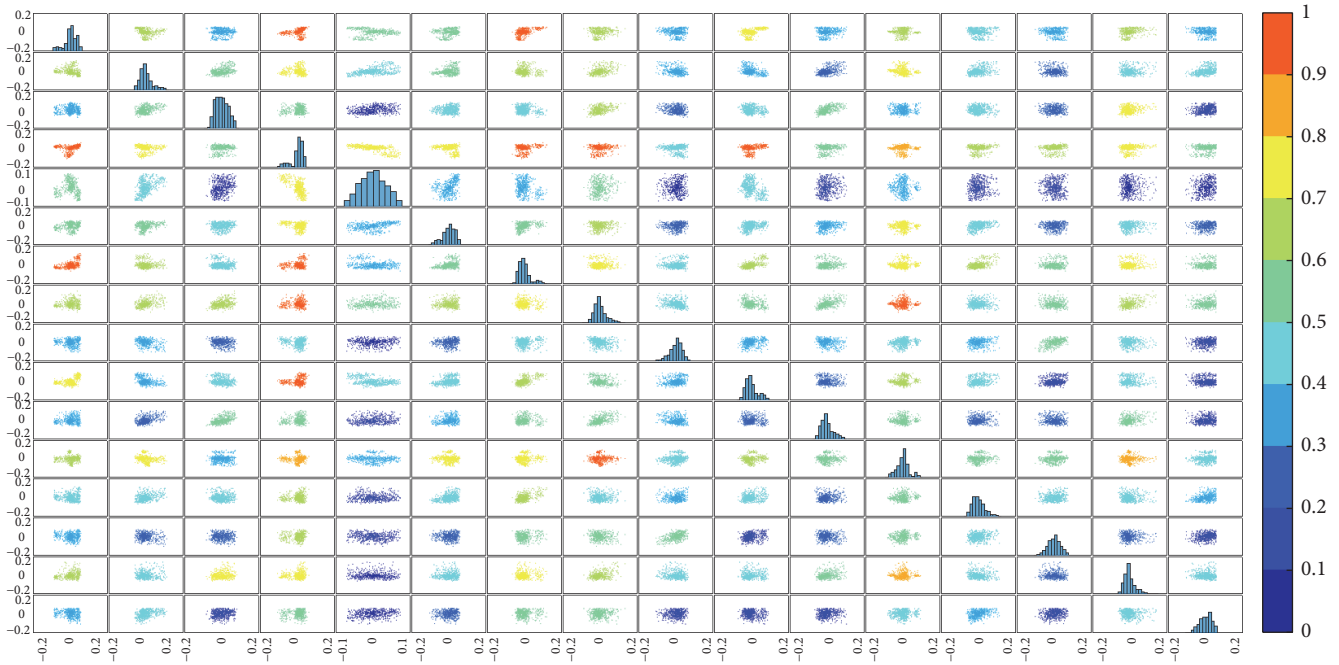


Figure 6: Colored SPLOM for the long tail eigengenes from the brain tumor study. The subplot in the ith row, jth column of the matrix is a scatter plot of the ith eigengene against the jth eigengene. Along the diagonal are histogram plots of each eigengene.

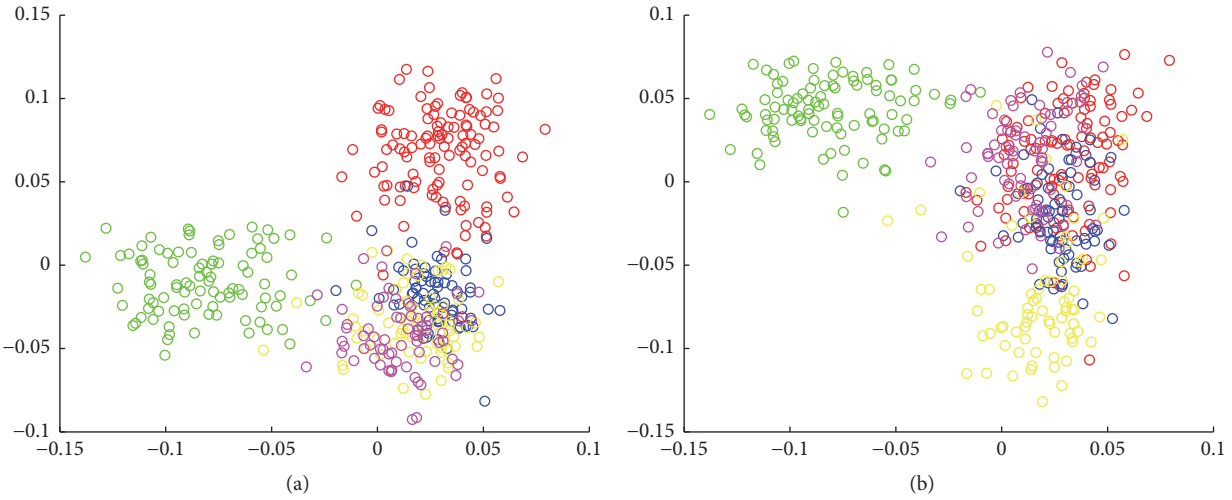


Figure 7: (a) The scatter plot between eigengene #4 (x-axis) and eigengene #11 (y-axis). (b) The scatter plot between eigengenes #4 (x-axis) and #6 (y-axis).

migration process is considered critical to the invasion of glioblastoma [28, 29].

4. Discussion and Conclusion

In this work, we presented a workflow for detecting distribution patterns in cell populations based on single-cell transcriptome study. With the fast adoption of single-cell analysis, a challenge to researchers is how to effectively extract gene features to meaningfully separate the cell population. However, this often ends up in a chicken-and-egg situation as the separation of the cells often depends on the choice of gene features, yet without a clear pattern it is difficult to determine which gene features are effective. Our workflow uses the well-developed gene coexpression network analysis to take advantage of the fact that coexpressed genes are often functionally or structurally related and the number of coexpressed modules is much smaller than the number of genes. Thus, when the coexpressed modules are summarized into eigengenes, not only can we quickly explore the distribution of cells interactively but also we can promptly interpret the gene features and generate new hypothesis.

Since the cells are separated based on different choices of the gene features, we dub the workflow as "functional virtual

flow cytometry," which achieves separation of the cells based on salient gene features. The separation of cells leads to new hypothesis such as the involvement of glutamate metabolism in the separation of the brain cells in the Allen Brain scRNA-seq data and the specific glioblastoma sample with unique cell migration related signature. While for the latter it is unclear if this observation is indeed biological or due to batch effect, our workflow quickly pointed out the pattern for researchers in deeper examination.

With the interactive visualization, additional advanced analysis can be carried out. For instance, in both Figures 4(b) and 7(b), an interesting observation is that the x- and y-axes cannot both have low values, suggesting interesting Boolean relationships between the gene groups [30]. Therefore, as our ongoing work, these analytic tools along with the workflow are being implemented in an online single-cell analytics portal.

Abbreviations

SPLOM:	Scatter plot matrix
scRNA-seq:	Single-cell RNA sequencing
PCA:	Principal component analysis
MDS:	Multidimensional scaling
FVFC:	Functional virtual flow cytometry
GCNA:	Gene coexpression network analysis
dLGN:	Dorsal lateral geniculate nucleus
ABA:	Allen Brian Atlas
GEO:	Gene Expression Omnibus.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is partially supported by Human Frontier Science Program (to Kun Huang), the NCI ITCR U01CA188547 (to Kun Huang), and the National Natural Science Foundation of China (61572265 to Zhi Han). The Ohio Supercomputer Center provided computing support.

References

[1] M. Setty, M. D. Tadmor, S. Reich-Zeliger et al., "Wishbone identifies bifurcating developmental trajectories from single-cell data," *Nature Biotechnology*, vol. 34, no. 6, pp. 637–645, 2016.

[2] S. Krishnaswamy, M. H. Spitzer, M. Mingueneau et al., "Conditional density-based analysis of T cell signaling in single-cell data," *Science*, vol. 346, no. 6213, Article ID 1250689, 2014.

[3] A. Scialdone, K. N. Natarajan, L. R. Saraiva et al., "Computational assignment of cell-cycle stage from single-cell transcriptome data," *Methods*, vol. 85, pp. 54–61, 2015.

[4] C. Shao and T. Höfer, "Robust classification of single-cell transcriptome data by nonnegative matrix factorization," *Bioinformatics*, vol. 33, no. 2, Article ID btw607, pp. 235–242, 2017.

[5] B. Anchang, T. D. P. Hart, S. C. Bendall et al., "Visualization and cellular hierarchy inference of single-cell data using SPADE," *Nature Protocols*, vol. 11, no. 7, pp. 1264–1279, 2016.

[6] P. Qiu, E. F. Simonds, S. C. Bendall et al., "Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE," *Nature Biotechnology*, vol. 29, no. 10, pp. 886–893, 2011.

[7] F. Buettner, K. N. Natarajan, F. P. Casale et al., "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells," *Nature Biotechnology*, vol. 33, no. 2, pp. 155–160, 2015.

[8] L. Haghverdi, F. Buettner, and F. J. Theis, "Diffusion maps for high-dimensional single-cell analysis of differentiation data," *Bioinformatics*, vol. 31, no. 18, pp. 2989–2998, 2015.

[9] P. Angerer, L. Haghverdi, M. Büttner, F. J. Theis, C. Marr, and F. Buettner, "Destiny: diffusion maps for large-scale single-cell data in R," *Bioinformatics*, vol. 32, no. 8, pp. 1241–1243, 2016.

[10] J. A. DiGiuseppe, M. D. Tadmor, and D. Pe'Er, "Detection of minimal residual disease in B lymphoblastic leukemia using viSNE," *Cytometry Part B - Clinical Cytometry*, vol. 88, no. 5, pp. 294–304, 2015.

[11] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, article 17, 2005.

[12] A. M. Yip and S. Horvath, "Gene network interconnectedness and the generalized topological overlap measure," *BMC Bioinformatics*, vol. 8, article 22, 2007.

[13] J. Zhang, K. Lu, Y. Xiang et al., "Weighted frequent gene co-expression network mining to identify genes involved in genome stability," *PLoS Computational Biology*, vol. 8, no. 8, Article ID e1002656, 2012.

[14] Z. Han, J. Zhang, G. Sun, G. Liu, and K. Huang, "A matrix rank based concordance index for evaluating and detecting conditional specific co-expressed gene modules," *BMC Genomics*, vol. 17, article no. 519, 2016.

[15] J. Zhang and K. Huang, "Normalized lmQCM: an algorithm for detecting weak quasi-cliques in weighted graph with applications in gene co-expression module discovery in cancers," *Cancer Informatics*, vol. 1, article 1, p. 137.

[16] Y. Xiang, J. Zhang, and K. Huang, "Mining the tissue-tissue gene co-expression network for tumor microenvironment study and biomarker prediction," *BMC genomics*, vol. 14, supplement 5, article s4, 2013.

[17] P. Langfelder and S. Horvath, "Eigengene networks for studying the relationships between co-expression modules," *BMC systems biology*, vol. 1, article 54, 2007.

[18] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, article 559, 2008.

[19] 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.

[20] 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.

[21] A. L. Muratov and O. Y. Gnedin, "Modeling the metallicity distribution of globular clusters," *Astrophysical Journal*, vol. 718, no. 2, pp. 1266–1288, 2010.

[22] J. B. HALDANE, "Simple tests for bimodality and bitangentiality," *Annals of Eugenics*, vol. 16, no. 1, pp. 359–364, 1951.

[23] H. Holzmann and S. Vollmer, "A likelihood ratio test for bimodality in two-component mixtures with application to regional income distribution in the EU," *AStA. Advances in Statistical Analysis*, vol. 92, no. 1, pp. 57–69, 2008.

[24] L. T. DeCarlo, "On the meaning and use of kurtosis," *Psychological Methods*, vol. 2, no. 3, pp. 292–307, 1997.

[25] K. N. Brown, S. Chen, Z. Han et al., "Clonal production and organization of inhibitory interneurons in the neocortex," *Science*, vol. 334, no. 6055, pp. 480–486, 2011.

[26] H.-T. Xu, Z. Han, P. Gao et al., "Distinct lineage-dependent structural and functional organization of the hippocampus," *Cell*, vol. 157, no. 7, pp. 1552–1564, 2014.

[27] A. P. Patel, I. Tirosh, J. J. Trombetta et al., "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma," *Science*, vol. 344, no. 6190, pp. 1396–1401, 2014.

[28] E. T. Sayegh, G. Kaur, O. Bloch, and A. T. Parsa, "Systematic review of protein biomarkers of invasive behavior in glioblastoma," *Molecular Neurobiology*, vol. 49, no. 3, pp. 1212–1244, 2014.

[29] S. M. Turaga and J. D. Lathia, "Adhering towards tumorigenicity: altered adhesion mechanisms in glioblastoma cancer stem cells," *CNS Oncology*, vol. 5, no. 4, pp. 251–259, 2016.

[30] D. Sahoo, D. L. Dill, A. J. Gentles, R. Tibshirani, and S. K. Plevritis, "Boolean implication networks derived from large scale, whole genome microarray datasets," *Genome Biology*, vol. 9, no. 10, article no. R157, 2008.

Hindawi
BioMed Research International
Volume 2017, Article ID 2563085, 9 pages
https://doi.org/10.1155/2017/2563085

Research Article

Diagnostic MicroRNA Biomarker Discovery for Non-Small-Cell Lung Cancer Adenocarcinoma by Integrative Bioinformatics Analysis

Yang Shao, Bin Liang, Fei Long, and Shu-Juan Jiang

Department of Respiratory Medicine, Shandong Provincial Hospital Affiliated to Shandong University, Jinan, Shandong 250021, China

Correspondence should be addressed to Shu-Juan Jiang; shujuan-jiang@163.com

Received 3 March 2017; Accepted 10 April 2017; Published 15 June 2017

Academic Editor: Xingming Zhao

Lung cancer is the leading cause of cancer death and its incidence is ranked high in men and women worldwide. Non-small-cell lung cancer (NSCLC) adenocarcinoma is one of the most frequent histological subtypes of lung cancer. The aberration profile and the molecular mechanism driving its progression are the key for precision therapy of lung cancer, while the screening of biomarkers is essential to the precision early diagnosis and treatment of the cancer. In this work, we applied a bioinformatics method to analyze the dysregulated interaction network of microRNA-mRNA in NSCLC, based on both the gene expression data and the microRNA-gene regulation network. Considering the properties of the substructure and their biological functions, we identified the putative diagnostic biomarker microRNAs, some of which have been reported on the PubMed citations while the rest, that is, miR-204-5p, miR-567, miR-454-3p, miR-338-3p, and miR-139-5p, were predicted as the putative novel microRNA biomarker for the diagnosis of NSCLC adenocarcinoma. They were further validated by functional enrichment analysis of their target genes. These findings deserve further experimental validations for future clinical application.

1. Introduction

Lung cancer is the most death causing cancer for both men and women in the United States, and it is also the most death causing cancer in men and second in women worldwide. The incidence rate is high and ranked second for both men and women in the United States [1, 2]. Non-small-cell lung cancer (NSCLC) adenocarcinoma is one of the most common histological subtypes of lung cancer [3] and it is reported that nearly 40% of the lung cancer are adenocarcinoma, and the other two subtypes of NSCLS are squamous-cell carcinoma and large-cell carcinoma [4]. The NSCLS adenocarcinoma is reported associated with the aberrations like the epidermal growth factor receptor (EGFR) mutations and anaplastic lymphoma kinase (ALK) fusion or rearrangement [3, 5], and several drugs such as gefitinib, erlotinib, and afatinib were developed for the targeting the aberrant gene products, but only few patients are ideal for the targeted treatments [6]. In addition, the patients treated with these target drugs may acquire resistance and make the treatment invalid [7]. There are also other aberrations reported associated with the NSCLC adenocarcinoma, such as mutations or fusions happen in HER2, BRAF, NF1, MEK1, RET, ROS1, and other genes. The risk factors for NSCLS adenocarcinoma may also include air pollution, gender, age, smoking, occupation, and eating habits [8–10]. For personalized diagnosis and treatment of cancer, the expression profile characterization and the key player screening [11] are the necessary steps. With the coming of aging era and the air pollution in the developing countries, the incidence of lung cancer will keep high, and the early diagnosis of lung cancer becomes very necessary. However, we still lack sensitive and precision biomarkers for the early diagnosis or the personalized therapy of the lung cancer [7, 12, 13].

MicroRNAs are endogenous small noncoding RNAs which regulate many important biological roles and theirs aberrations may have significant effects on the cancer genesis and progression, such as cell proliferation, cell cycle, apoptosis, and tumorigenesis, and therefore are good candidates for cancer diagnosis and therapy biomarkers [14–16]. Biomarker microRNA discovery could be implemented both experimentally and computationally. The former is a routine

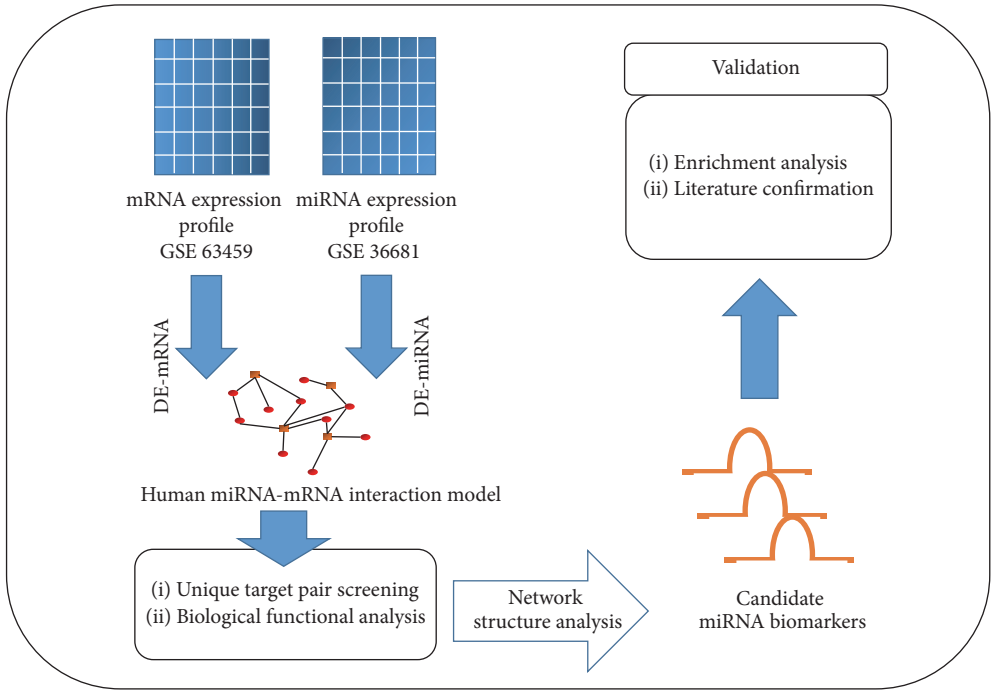


Figure 1: The procedure of data collection, identification of microRNA biomarkers with integrative analysis, and the validation of the miRNA biomarkers.

but time-consuming and costing method, since the biological systems are complex and the mechanisms are diverse. The computational methods based on integrative analysis of different omics data have more advantages, such that it could integrate diverse omics data sets and model the biological process by network construction and then understand the aberrations at the systems level [17, 18]. Furthermore, the computational methods are also cheap, less time-consuming and could be easily validated by literature mining, association analysis, and bioinformatics functional enrichment confirmation. With more and more biomedical data available and accumulated, the computational methods will be more and more powerful for the future precision medicine strategies.

At present, more and more computational and bioinformatics models are developing for biomarker discovery; some of them are machine learning based [19, 20], while others are mechanism-based [14–16, 21]. The machine learning based methods need more data to train the model and the mechanism-based models are more knowledge based, and the two types of methods complement each other and promote biomarker discovery. We here applied the previous reported mechanism-based method, which is successful in microRNA biomarker discovery, to screen novel diagnostic biomarker for NSCLC adenocarcinoma.

2. Materials and Methods

The data used in our integrative analysis include the gene expression data of NSCLC adenocarcinoma from both microRNA and messenger RNA(mRNA) and the human reference microRNA-mRNA network. We need to first collect the data and reconstruct the NSCLC adenocarcinoma specific network and then to identify putative microRNA biomarkers based on the network structure and their biological functions. The screened microRNAs need to be validated by literature mining, confirmation, and bioinformatics exploration of their associations with NSCLC adenocarcinoma. The pipeline of the whole process of this work could be seen in Figure 1.

2.1. The NSCLC Adenocarcinoma Gene Expression Data Collection and the Human Reference MicroRNA-mRNA Interactions. The gene expression and microRNA expression data for the NSCLC adenocarcinoma were extracted from the public GEO database [22]. The details of data sets are listed in Table 1. The data sets we used for the construction of NSCLC adenocarcinoma specific microRNA-mRNA network are GSE63459 and GSE36681, where the GSE63459 data set is the mRNA expression data which includes 33 NSCLC adenocarcinoma samples and 32 samples as control, and the other data set is the microRNA expression data with 47 NSCLC adenocarcinoma samples and 47 control samples. The data were first normalized and the differentially expressed mRNAs were then selected with the linear models in limma R package [23, 24]. The empirical Bayes (eBayes) method was applied to calculate the p value and other parameters. The Benjamini-Hochberg correction was used to adjust the p values. The adjusted p values less than 0.05 were regarded as significant. The human reference microRNA-mRNA interaction network was constructed based mainly on experimentally validated and the consensus predicted microRNA-mRNA interaction pairs as reported in previous studies [16, 25, 26].

Table 1: NSCLC adenocarcinoma gene expression data collected from GEO data sets.

Accession/ID	PMID	Platform	Treatment	Control	Materials	Year	mRNA/miRNA
GSE36681	22573352	GPL8179	$n = 47$	$n = 47$	tissue	2012	miRNA
GSE63459	26134223	GPL6883	$n = 33$	$n = 32$	tissue	2015	mRNA

Table 2: Literature reported lung adenocarcinoma miRNA biomarkers.

Reported miRNA	Official ID	PMID	Biomarker type	Samples	Expression level	NOG	TFP
miR-155	miR-155-5p	24190459 [27]	Diagnosis	Serum	Up	71	0.21
miR-196a-5p	miR-196a-5p	27247934 [28]	Diagnosis	Tissue	Up	7	0.19
miR-218-5p	miR-218-5p	27247934 [28]	Diagnosis	Tissue	Down	10	0.12
miR-143	miR-143-3p	24286416 [29]	Diagnosis	Blood	Down	15	0.03
miR-182	miR-182-5p	19493678 [30]	Diagnosis	Tissue	Up	9	0.19
miR-650	miR-650	23991130 [31]	Prognosis	Tissue	Up	0	0
miR-141	miR-141-3p	25746592 [32]	Prognosis	Tissue	Up	22	0.17
miR-29c	miR-29c-3p	28241836 [33]	Prognosis	Tissue	Down	5	0.15
miR-23b-3p	miR-23b-3p	28055956 [34]	Prognosis	Plasma	Up	23	0.15
miR-10b-5p	miR-10b-5p	28055956 [34]	Prognosis	Plasma	Up	9	0.15
miR-21-5p	miR-21-5p	28055956 [34]	Prognosis	Plasma	Up	38	0.13
miR-126-3p	miR-126-3p	27277197 [35]	Prognosis	Tissue	Down	6	0.12
miR-451a	miR-451a	27277197 [35]	Prognosis	Tissue	Down	4	0.17
miR-25	miR-25-3p	26687391 [36]	Prognosis	Blood	Up	9	0.17
miR-145	miR-145-5p	26687391 [36]	Prognosis	Blood	Down	36	0.11
miR-210	miR-210	26687391 [36]	Prognosis	Blood	Down	1	0.18
miR-142-3p	miR-142-3p	23410826 [37]	Prognosis	Serum	Up	12	0.14
miR-29b	miR-29b-3p	22249264 [38]	Prognosis	Tissue	Down	9	0.14
miR-590	miR-590-5p	28012926 [39]	Prognosis	Tissue	Up	14	0.16

2.2. Model Construction, Validation, and Identification of NSCLC Adenocarcinoma MicroRNA Biomarkers. The basic idea and the methods we used here are based on the models which were developed previously [26, 40–42], where two parameters are used to measure the importance of microRNAs as the potential biomarkers for a specific disease. The first is the number of genes (NOG) uniquely targeted by a certain microRNA [40, 41], and this index is reasonable to quantify the tendency to be a biomarker since the alteration of the unique interaction cannot be substituted or compensated by other microRNA-mRNA interaction pairs. The other index was proposed to quantify the transcription factor percentage (TFP) and was defined as the percentage of transcription factor (TF) genes of all the microRNA targets [26]. With these two indexes, the NSCLC adenocarcinoma specific microRNA-mRNA interaction network was constructed by mapping the detected differentially expressed microRNAs in NSCLC adenocarcinoma onto the reference human microRNA-mRNA interaction network. With the reconstructed conditional network, the abovementioned measurements, that is, the NOG and TFP, were calculated for each microRNA in the NSCLC adenocarcinoma network. MicroRNAs with significant large NOG and TFP values (Wilcoxon signed-rank test, p value < 0.05) were detected as our putative biomarkers for diagnosis of NSCLC adenocarcinoma.

To validate the bioinformatics model first, we also collected reported NSCLC adenocarcinoma associated microRNAs from PubMed citations with the searching criteria "(lung adenocarcinoma OR NSCLC adenocarcinoma) AND (miRNA OR microRNA) AND (biomarker* OR marker*)". The related PMIDs, NOGs, and TFPs of these microRNAs were also calculated.

2.3. The Literature Confirmation and Functional Validation of the Putative NSCLC Adenocarcinoma Diagnostic MicroRNA Biomarkers. For validation of the bioinformatics method and the screened microRNA biomarkers, we checked the PubMed citations and extracted the reported microRNA biomarkers for both diagnosis and prognosis of NSCLC adenocarcinoma. The identified novel microRNA biomarkers were then validated with functional enrichment analysis of the targeted genes of the microRNAs. The enrichment analysis were performed with Gene Ontology Annotations, KEGG pathway analysis, which were done by the DAVID (Database for Annotation, Visualization, and Integrated Discovery) online tool [43]. The p value threshold was set to 0.05 and the FDR adjustment was used for multiple test correction. Then we calculated the enrichment based on the hypergeometric test.

3. Results and Discussion

3.1. The Validation of Bioinformatics Methods for the MicroRNA Biomarker Discovery in NSCLC Adenocarcinoma. The microRNA biomarkers for diagnosis and prognosis of NSCLC adenocarcinoma reported in PubMed citations were collected and listed in Table 2. According to the NOGs and TFPs listed in Table 2, all the microRNAs except miR-650 are characterized with high NOGs and TFPs, and therefore the

Table 3: Predicted putative lung adenocarcinoma microRNA biomarkers.

miRNA ID	NOG	p value	TFP	p value	Whole target genes
miR-145-5p	1	$1.66E-02$	0.67	$2.98E-08$	MMP12; ZFP36; KLF4
miR-204-5p	5	$1.78E-15$	0.13	$1.05E-02$	DPYSL2; EMP1; SPDEF; LMO7; SLC1A1; ALPL; MMP9; FRAS1
miR-182-5p	2	$1.80E-08$	0.20	$4.34E-04$	CAMK2N1; ZFP36; UBE2T; LPHN2; RGS17.
miR-567	1	$1.66E-02$	0.25	$4.61E-04$	SPTBN1; DUSP1; BCHE; LPHN2
miR-141-3p	2	$1.80E-08$	0.14	$2.05E-03$	H3F3B; TCEAL2; MYH10; LHFP; LPHN2; CCL2; KLF9
miR-454-3p	2	$1.80E-08$	0.11	$4.64E-02$	DPYSL2; HOXA5; FKBP11; SRPX; EDN1; LDLR; CAV2; BMPR2; SLC2A1
miR-590-3p	5	$1.78E-15$	0.13	$2.33E-03$	SERPINE2; PLEKHC1; SPTBN1; H3F3B; TMEM47; TIMP3; COL3A1; CXCL13; ETS2; CELSR3; LPL; SMAD6; BMPR2; LPHN2; SASH1.
miR-338-3p	2	$1.80E-08$	0.25	$4.61E-04$	COL1A1; FOSB; ADAMTS1; MMP9
miR-139-5p	1	$1.66E-02$	1.00	$1.49E-08$	FOS

performance of the microRNA biomarker discovery method based on the two measurements is reasonable and can be extended and applied to the biomarker discovery in NSCLC adenocarcinoma.

Among the reported microRNA biomarkers for NSCLC adenocarcinoma, many of them played essential roles in lung carcinogenesis and their abnormal expression patterns were highly associated with the occurrence and development of NSCLC adenocarcinoma. For example, serum miR-155 was a sensitive indicator for predicting the initiation of lung adenocarcinoma, especially combining with the index of carbohydrate antigen 125. It altered the expression of downstream proteins and activated the lung carcinogenic signal [27]. Two miRNAs, that is, miR-196a-5p and miR-218-5p, were validated to be up- and downregulated from normal to adenocarcinoma tumor tissues, respectively. Their target genes were functional in lung cancer related processes by activating or inhibiting biological activities in pathways in cancer, cell cycle, transcriptional misregulation in cancer, and small-cell lung cancer [28]. The prognostic value of miRNAs for NSCLC adenocarcinoma was also comprehensively investigated. For instance, Huang et al. [31] showed that miR-650 was able to regulate the expression of Bcl-2/Bax, which would thereby contribute to the docetaxel chemoresistance of lung adenocarcinoma cells. This miRNA was a powerful indicator for predicting the chemosensitivity of lung adenocarcinoma patients to docetaxel-based chemotherapy regimen. Zhang et al. [32] analyzed the clinical potential of miR-141 and found that this miRNAs was positively correlated with the tumor size, lymph NOGe metastasis, and TNM stage of lung adenocarcinomas. Meanwhile, Liu et al. [34] screened that the upregulation of plasma exosomal miRNAs miR-23b-3p, miR-10b-5p and miR-21-5p was strongly connected with the poor overall survival of lung adenocarcinoma patients. In order to evaluate the efficacy of maintenance treatment on lung adenocarcinoma patients with negativity for epidermal growth factor receptor (EGFR) mutations or anaplastic lymphoma kinase (ALK) translocations, Shi et al. [36] designed the experiment in which patients were divided into a pemetrexed group and a control group, respectively. As a result, the expression levels of miR-25, miR-145, and miR-210 were associated with the progression-free survival time of patients in the treatment group, which highlighted the prognostic potential of these miRNAs to the pemetrexed therapy in specific lung adenocarcinoma individuals.

3.2. The Predicted MicroRNA Biomarker for Diagnosis of NSCLC Adenocarcinoma. We performed the predictions according to the pipeline shown in Figure 1. At first, 93 differentially expressed microRNAs and 331 differentially expressed genes in NSCLC adenocarcinoma were detected, respectively. Nine microRNAs were screened by Wilcoxon signed-rank test with p value < 0.05. These nine microRNAs were predicted to be biomarkers for the diagnosis of NSCLC adenocarcinoma as listed in Table 3. Their network structural features in the microRNA-mRNA interaction network were shown, which are the whole set of targeted genes, NOGs and TFPs. Four of the nine predicted microRNAs (bolded and underlined in Table 3), that is, miR-145-5p, miR-182-5p, miR-141-3p, and miR-590-3p, have been reported as biomarkers previously, the remaining five, that is, miR-204-5p, miR-567, miR-454-3p, miR-338-3p, and miR-139-5p, were recommended as novel diagnosis biomarker for NSCLC adenocarcinoma.

3.3. Functional Enrichment Validation of the Predicted MicroRNA Biomarkers. We further performed the functional enrichment analysis to investigate the roles of genes regulated by identified microRNA biomarkers through Database for Annotation, Visualization and Integrated Discovery (DAVID) online tools. This analysis was conducted in two ways: Gene Ontology (GO) analysis and KEGG pathway analysis.

In GO analysis, we did this analysis at three levels: biological process (BP), cellular component (CC), and molecular function (MF). The top 10 most significantly enriched

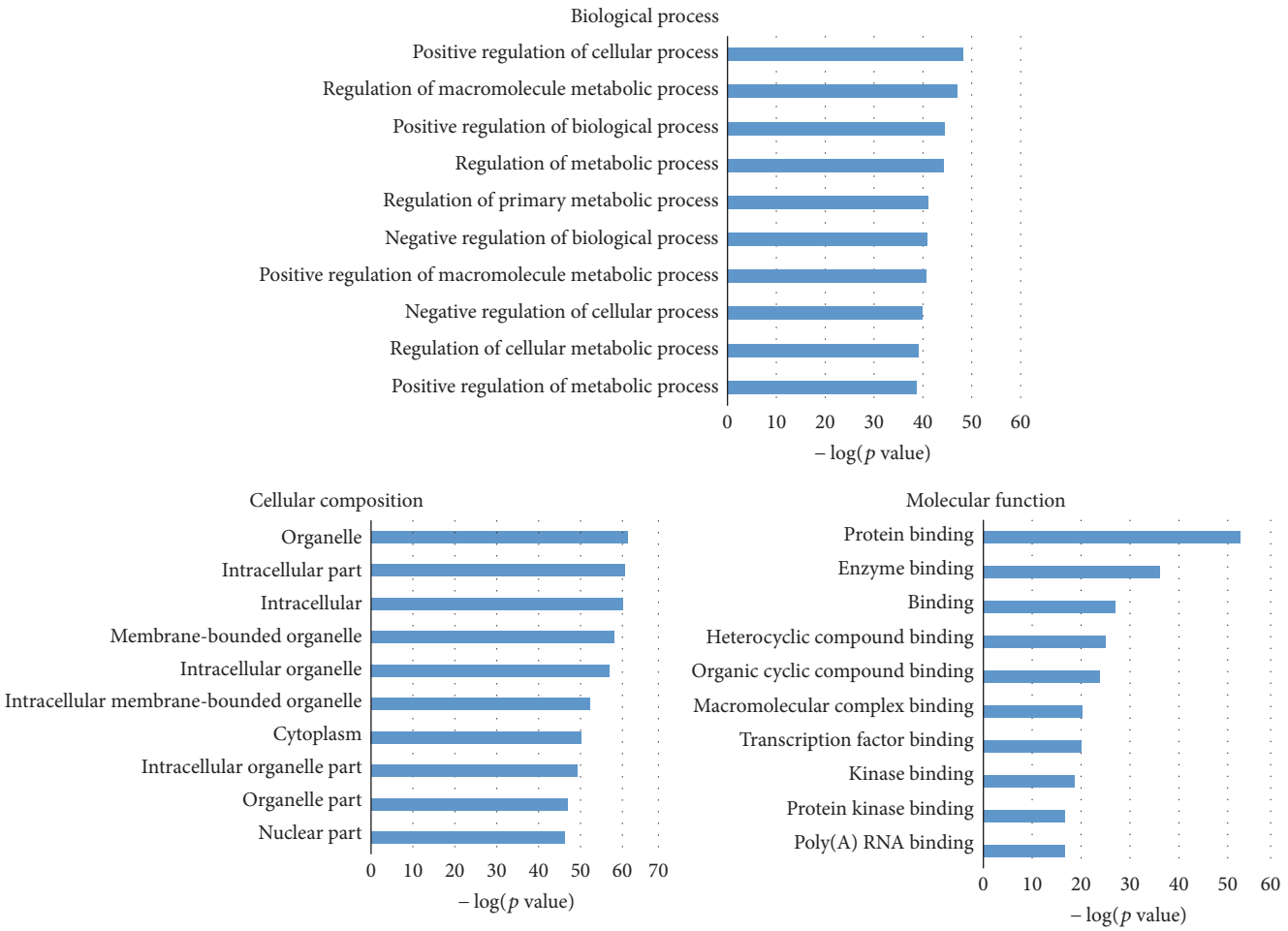


Figure 2: Gene ontology (GO) analysis for genes targeted by 9 identified microRNA biomarkers. The statistical significance value (p value) has been negative 10-based log transformed. Top 10 significantly enriched items are listed for each level.

items were shown in Figure 2. Through further literature validation, we found that most of the items have a strong relationship with lung adenocarcinoma. For example, in molecular function level, relevant research has shown that FGFR2's function could be regulated by two proteins: Grb2 and Plcγ1 under the situation of growth factor absence. These two proteins will compete for the same protein binding site [44]. FGFR2 expression could be repressed by miR-338-3p, one of the identified miRNA biomarkers. Other items, such as cytoplasm [45], membrane-bounded organelle [46], positive regulation of metabolic process, and transcription factor binding [47, 48] have already been validated through biological and clinical experiments to have an impact on the occurrence and metastasis of lung adenocarcinoma.

In KEGG pathway analysis, we totally found 72 significantly enriched pathways. Here we still selected top 10 most significantly enriched pathways for further investigation. These 10 pathways were listed in Table 4 and shown in Figure 3.

Here, we investigate the relationship between these 10 pathways and lung adenocarcinoma through literature validation and we found that most of these pathways have been demonstrated to be associated with lung adenocarcinoma.

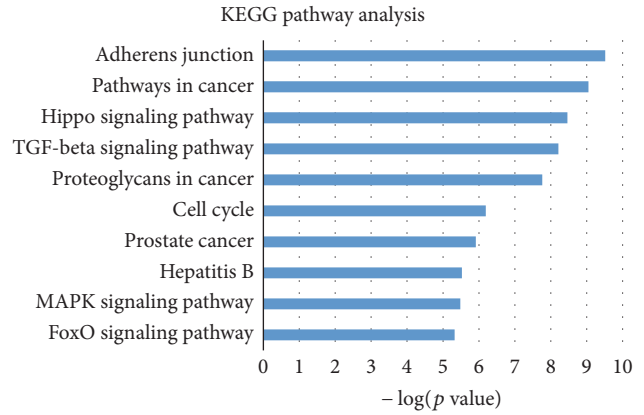


Figure 3: KEGG pathway enrichment analysis for genes targeted by 9 candidate microRNA biomarkers. The statistical significance value (p value) has been negative 10-based log transformed. The top 10 significantly enriched pathways are listed, respectively, in this figure.

Besides common pathways in lung cancer like cell cycle and MAPK signaling pathways [49, 50], there are still other pathways such as Hippo signaling pathway [51] and TGF-beta

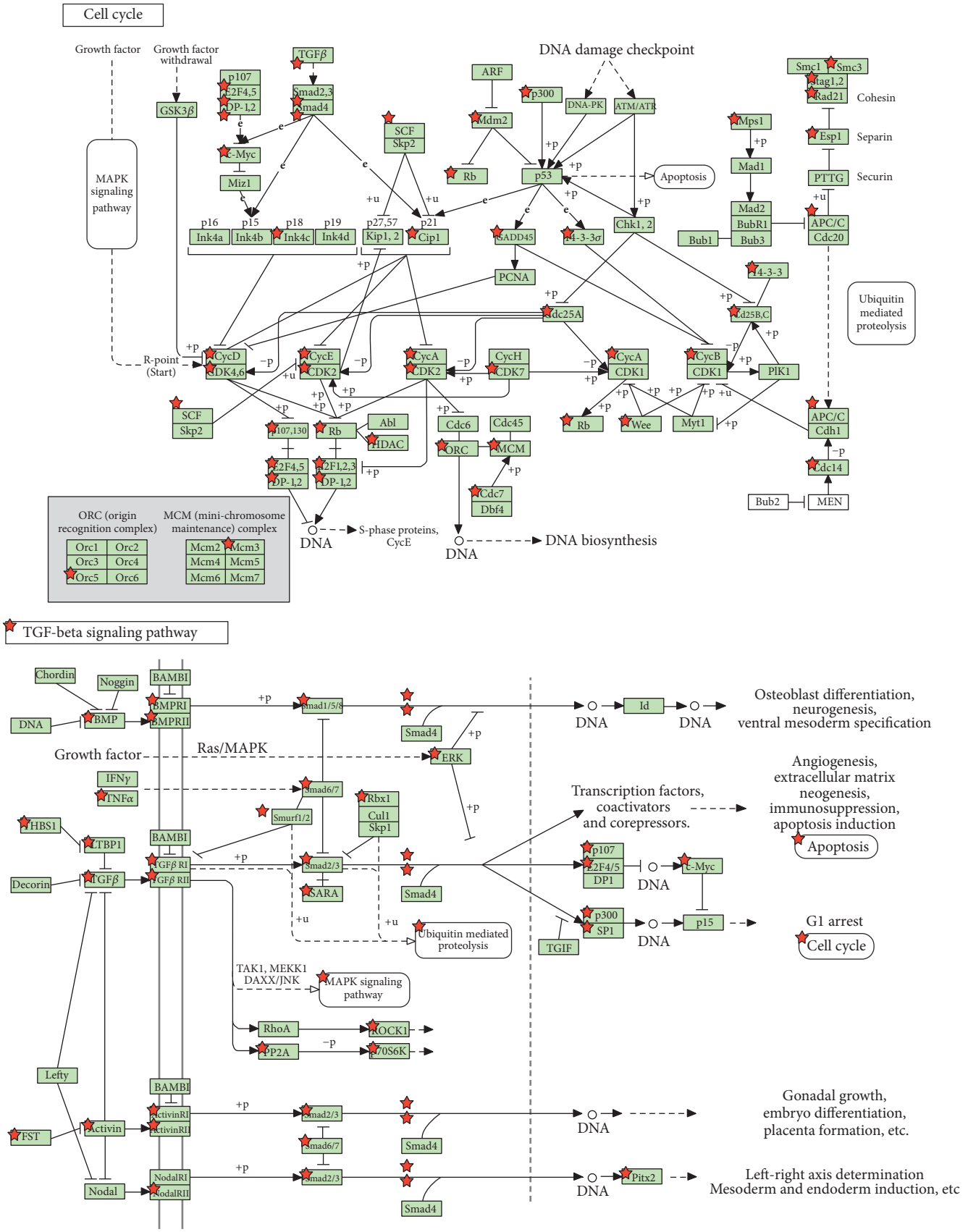


Figure 4: The pipeline of cell cycle and MAPK signaling pathway. The stars in these pipelines represent the action sites of the genes regulated by 9 candidate miRNA biomarkers.

Table 4: Top 10 significantly enriched pathways in KEGG pathway analysis.

Term	Adj. p value
Adherens junction	$3.05E-10$
Pathways in cancer	$8.91E-10$
Hippo signaling pathway	$3.43E-09$
TGF-beta signaling pathway	$6.13E-09$
Proteoglycans in cancer	$1.72E-08$
Cell cycle	$6.44E-07$
Prostate cancer	$1.22E-06$
Hepatitis B	$2.98E-06$
MAPK signaling pathway	$3.29E-06$
FoxO signaling pathway	$4.74E-06$

signaling pathway and proteoglycans in cancer are all supported by relevant research [52, 53]. The relevant pipeline of cell cycle and MAPK signaling pathways can be referred to in Figure 4.

4. Conclusions

In this research, an integrative bioinformatics model considering the network structure and biological functions of the microRNA targets was used to predict novel biomarker microRNAs for the diagnosis of NSCLC adenocarcinoma. The method was first tested with the reported microRNA biomarkers of NSCLC adenocarcinoma; then we extended the model and applied it to the microRNA and gene expression data. We detected five novel biomarker microRNAs for the diagnosis of NSCLC adenocarcinoma, including miR-204-5p, miR-567, miR-454-3p, miR-338-3p, and miR-139-5p. The novel bioinformatics microRNAs were validated with bioinformatics exploring their functions associated with NSCLC adenocarcinoma.

Abbreviations

NSCLC: Non-small-cell lung cancer
DE: Differentially expressed
TF: Transcription factor
NOG: Novel out degree
TFP: Transcription factor gene percentage
KEGG: Kyoto encyclopedia of genes and genomes
DAVID: Database for Annotation, Visualization, and Integrated Discovery.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to acknowledge the financial support of the National Natural Science Foundation of China (81370138).

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," *CA: A Cancer Journal for Clinicians*, vol. 67, no. 1, pp. 7–30, 2017.

[2] L. A. Torre, R. L. Siegel, and A. Jemal, "Lung cancer statistics," *Advances in Experimental Medicine and Biology*, vol. 893, pp. 1–19, 2016.

[3] M. Saito, K. Shiraishi, H. Kunitoh, S. Takenoshita, J. Yokota, and T. Kohno, "Gene aberrations for precision medicine against lung adenocarcinoma," *Cancer Science*, vol. 107, no. 6, pp. 713–720, 2016.

[4] Y. Zhang, H. Wang, J. Wang et al., "Global analysis of chromosome 1 genes among patients with lung adenocarcinoma, squamous carcinoma, large-cell carcinoma, small-cell carcinoma, or non-cancer," *Cancer and Metastasis Reviews*, vol. 34, no. 2, pp. 249–264, 2015.

[5] C.-W. Xu, X.-Y. Cai, Y. Shao et al., "A case of lung adenocarcinoma with a concurrent EGFR mutation and ALK rearrangement: a case report and literature review," *Molecular Medicine Reports*, vol. 12, no. 3, pp. 4370–4375, 2015.

[6] G. Roviello, "The distinctive nature of adenocarcinoma of the lung," *OncoTargets and Therapy*, vol. 8, pp. 2399–2406, 2015.

[7] P. T. Cagle, K. Raparia, and B. P. Portier, "Emerging biomarkers in personalized therapy of lung cancer," *Advances in Experimental Medicine and Biology*, vol. 890, pp. 25–36, 2016.

[8] F. L. Wang, "Analysis of risk factors for female lung adenocarcinoma in haerbin: indoor air pollution," *Zhonghua Yu Fang Yi Xue Za Zhi*, vol. 23, no. 5, pp. 270–273, 1989.

[9] L. M. Butler, J. A. Montague, W.-P. Koh, R. Wang, M. C. Yu, and J.-M. Yuan, "Fried meat intake is a risk factor for lung adenocarcinoma in a prospective cohort of Chinese men and women in Singapore," *Carcinogenesis*, vol. 34, no. 8, pp. 1794–1799, 2013.

[10] C. Paris, C. Clement-Duchene, J. M. Vignaud et al., "Relationships between lung adenocarcinoma and gender, age, smoking and occupational risk factors: a case-case study," *Lung Cancer*, vol. 68, no. 2, pp. 146–153, 2010.

[11] J. Jiang, P. Jia, Z. Zhao, and B. Shen, "Key regulators in prostate cancer identified by co-expression module analysis," *BMC Genomics*, vol. 15, no. 1, article 1015, 2014.

[12] H. C. Jeong, "Clinical aspect of MicroRNA in lung cancer," *Tuberculosis and Respiratory Diseases*, vol. 77, no. 2, pp. 60–64, 2014.

[13] M. K. Kim, S. B. Jung, J.-S. Kim et al., "Expression of microRNA miR-126 and miR-200c is associated with prognosis in patients with non-small cell lung cancer," *Virchows Archiv*, vol. 465, no. 4, pp. 463–471, 2014.

[14] S. Shen, Y. Lin, X. Yuan et al., "Biomarker microRNAs for Diagnosis, prognosis and treatment of hepatocellular carcinoma: a functional survey and comparison," *Scientific Reports*, vol. 6, article 38311, 2016.

[15] L. Shen, Y. Lin, Z. Sun, X. Yuan, L. Chen, and B. Shen, "Knowledge-guided bioinformatics model for identifying autism spectrum disorder diagnostic microrna biomarkers," *Scientific Reports*, vol. 6, article 39663, 2016.

[16] Y. Zhu, Y. Lin, W. Yan et al., "Novel biomarker micrornas for subtyping of acute coronary syndrome: a bioinformatics approach," *BioMed Research International*, vol. 2016, Article ID 4618323, 11 pages, 2016.

[17] Y. Hu, J. Li, W. Yan et al., "Identifying novel glioma associated pathways based on systems biology level meta-analysis," *BMC Systems Biology*, vol. 7, supplement 2, p. S9, 2013.

[18] B. Shen, Shen H. -B., Tian T., Lü Q., and Hu G., "Translational bioinformatics and computational systems medicine," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 375641, 2 pages, 2013.

[19] H. A. Mattison, T. Stewart, and J. Zhang, "Applying bioinformatics to proteomics: is machine learning the answer to biomarker discovery for PD and MSA?" *Movement Disorders*, vol. 27, no. 13, pp. 1595–1597, 2012.

[20] V. A. Huynh-Thu, Y. Saeys, L. Wehenkel, and P. Geurts, "Statistical interpretation of machine learning-based feature importance scores for biomarker discovery," *Bioinformatics*, vol. 28, no. 13, Article ID bts238, pp. 1766–1774, 2012.

[21] X. M. Zhao, K. Q. Liu, G. Zhu et al., "Identifying cancer-related microRNAs based on gene expression data," *Bioinformatics*, vol. 31, no. 8, pp. 1226–1234, 2015.

[22] T. Barrett, T. O. Suzek, D. B. Troup et al., "NCBI GEO: mining millions of expression profiles—database and tools," *Nucleic Acids Research*, vol. 35, pp. D760–D765, 2007.

[23] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, article 3, 2004.

[24] M. E. Ritchie, B. Phipson, D. Wu et al., "*limma* powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, 2015.

[25] J. Huang, Z. Sun, W. Yan et al., "Identification of MicroRNA as sepsis biomarker based on miRNAs regulatory network analysis," *BioMed Research International*, vol. 2014, Article ID 594350, 12 pages, 2014.

[26] W. Yan, L. Xu, Z. Sun et al., "MicroRNA biomarker identification for pediatric acute myeloid leukemia based on a novel bioinformatics model," *Oncotarget*, vol. 6, no. 28, pp. 26424–26436, 2015.

[27] F. Gao, J. Chang, H. Wang, and G. Zhang, "Potential diagnostic value of miR-155 in serum from lung adenocarcinoma patients," *Oncology Reports*, vol. 31, no. 1, pp. 351–357, 2014.

[28] F. Tian, R. Li, Z. Chen et al., "Differentially expressed miRNAs in tumor, adjacent, and normal tissues of lung adenocarcinoma," *BioMed Research International*, vol. 2016, Article ID 1428271, 2016.

[29] X. L. Zeng, S. Y. Zhang, J. F. Zheng, H. Yuan, and Y. Wang, "Altered miR-143 and miR-150 expressions in peripheral blood mononuclear cells for diagnosis of non-small cell lung cancer," *Chinese Medical Journal*, vol. 126, no. 23, pp. 4510-6, 2013 (English).

[30] W. C. Cho, A. S. Chow, and J. S. Au, "Restoration of tumour suppressor hsa-miR-145 inhibits cancer cell growth in lung adenocarcinoma patients with epidermal growth factor receptor mutation," *European Journal of Cancer*, vol. 45, no. 12, pp. 2197–2206, 2009.

[31] Huang J. Y., Cui S. Y., Chen Y. T. et al., "MicroRNA-650 was a prognostic factor in human lung adenocarcinoma and confers the docetaxel chemoresistance of lung adenocarcinoma cells via regulating Bcl-2/Bax expression," *PLoS ONE*, vol. 8, no. 8, Article ID e72615, 2013.

[32] X. Zhang, P. Li, M. Rong et al., "MicroRNA-141 is a biomarker for progression of squamous cell carcinoma and adenocarcinoma of the lung: clinical analysis of 125 patients," *The Tohoku Journal of Experimental Medicine*, vol. 235, no. 3, pp. 161–169, 2015.

[33] L. Liu, N. Bi, L. Wu et al., "MicroRNA-29c functions as a tumor suppressor by targeting VEGFA in lung adenocarcinoma," *Molecular Cancer*, vol. 16, article 50, no. 1, 2017.

[34] Q. Liu, Z. Yu, S. Yuan et al., "Circulating exosomal microRNAs as prognostic biomarkers for non-small-cell lung cancer," *Oncotarget*, vol. 8, no. 8, pp. 13048–13058, 2017.

[35] Q. Chen, H. Hu, D. Jiao et al., "miR-126-3p and miR-451a correlate with clinicopathological features of lung adenocarcinoma: The underlying molecular mechanisms," *Oncology Reports*, vol. 36, no. 2, pp. 209–917, 2016.

[36] S.-B. Shi, M. Wang, J. Tian, R. Li, C.-X. Chang, and J.-L. Qi, "MicroRNA 25, microRNA 145, and microRNA 210 as biomarkers for predicting the efficacy of maintenance treatment with pemetrexed in lung adenocarcinoma patients who are negative for epidermal growth factor receptor mutations or anaplastic lymphoma kinase translocations," *Translational Research*, vol. 170, pp. 1–7, 2016.

[37] S. Kaduthanam, S. Gade, M. Meister et al., "Serum miR-142-3p is associated with early relapse in operable lung adenocarcinoma patients," *Lung Cancer*, vol. 80, no. 2, pp. 223–227, 2013.

[38] S. I. Rothschild, M. P. Tschan, E. A. Federzoni et al., "MicroRNA-29b is involved in the Src-ID1 signaling pathway and is dysregulated in human lung adenocarcinoma," *Oncogene*, vol. 31, no. 38, pp. 4221–4232, 2012.

[39] Y. Liu, F. Wang, and P. Xu, "miR-590 accelerates lung adenocarcinoma migration and invasion through directly suppressing functional target OLFM4," *Biomedicine & Pharmacotherapy*, vol. 86, pp. 466–474, 2017.

[40] J. Zhu, S. Wang, W. Zhang et al., "Screening key microRNAs for castration-resistant prostate cancer based on miRNA/mRNA functional synergistic network," *Oncotarget*, vol. 6, no. 41, pp. 43819–43830, 2015.

[41] W. Zhang, J. Zang, X. Jing et al., "Identification of candidate miRNA biomarkers from miRNA regulatory network with application to prostate cancer," *Journal of Translational Medicine*, vol. 12, p. 12, 2014.

[42] J. Chen, D. Zhang, W. Zhang et al., "Clear cell renal cell carcinoma associated microRNA expression signatures identified by an integrated bioinformatics analysis," *Journal of Translational Medicine*, vol. 11, article 169, 2013.

[43] 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, p. P3, 2003.

[44] Z. Timsah, Z. Ahmed, C.-C. Lin et al., "Competition between Grb2 and Plcγ1 for FGFR2 regulates basal phospholipase activity and invasion," *Nature Structural and Molecular Biology*, vol. 21, no. 2, pp. 180–188, 2014.

[45] L. Li, L. Wang, K. M. Prise et al., "Akt/mTOR mediated induction of bystander effect signaling in a nucleus independent manner in irradiated human lung adenocarcinoma epithelial cells," *Oncotarget*, 2017.

[46] Z. Su, K. Wang, R. Li et al., "Overexpression of RBM5 induces autophagy in human lung adenocarcinoma cells," *World Journal of Surgical Oncology*, vol. 14, no. 1, article 57, 2016.

[47] E. Y. P. Lee, P.-L. Khong, V. H. F. Lee, W. Qian, X. Yu, and M. P. Wong, "Metabolic phenotype of stage IV lung adenocarcinoma: relationship with epidermal growth factor receptor mutation," *Clinical Nuclear Medicine*, vol. 40, no. 3, pp. e190–e195, 2015.

[48] D. D. Becker-Santos, K. L. Thu, J. C. English et al., "Developmental transcription factor NFIB is a putative target of oncofetal miRNAs and is associated with tumour aggressiveness in lung adenocarcinoma," *The Journal of Pathology*, vol. 240, no. 2, pp. 161–172, 2016.

[49] X. Wang, M. Long, K. Dong et al., "Chemotherapy agents-induced immunoresistance in lung cancer cells could be

reversed by trop-2 inhibition in vitro and in vivo by interaction with MAPK signaling pathway," *Cancer Biology & Therapy*, vol. 14, no. 12, pp. 1123–1132, 2013.

[50] G. M. Dancik and D. Theodorescu, "Robust prognostic gene expression signatures in bladder cancer and lung adenocarcinoma depend on cell cycle related genes," *PLoS ONE*, vol. 9, no. 1, Article ID e85249, 2014.

[51] J. M. Kim, D. W. Kang, L. Z. Long et al., "Differential expression of Yes-associated protein is correlated with expression of cell cycle markers and pathologic TNM staging in non-small-cell lung carcinoma," *Human Pathology*, vol. 42, no. 3, pp. 315–323, 2011.

[52] R. L. Toonkel, A. C. Borczuk, and C. A. Powell, "TGF-β signaling pathway in lung adenocarcinoma invasion," *Journal of Thoracic Oncology*, vol. 5, no. 2, pp. 153–157, 2010.

[53] Y. Kusano, Y. Yoshitomi, S. Munesue, M. Okayama, and K. Oguri, "Cooperation of syndecan-2 and syndecan-4 among cell surface heparan sulfate proteoglycans in the actin cytoskeletal organization of Lewis lung carcinoma cells," *Journal of Biochemistry*, vol. 135, no. 1, pp. 129–137, 2004.

Hindawi
BioMed Research International
Volume 2017, Article ID 1016305, 8 pages
https://doi.org/10.1155/2017/1016305

Research Article

Identification of Pharmacologically Tractable Protein Complexes in Cancer Using the R-Based Network Clustering and Visualization Program MCODER

Sungjin Kwon,[1] Hyosil Kim,[2] and Hyun Seok Kim[1,2]

[1]*Graduate Programs for Nanomedical Science, Yonsei University, Seoul, Republic of Korea*
[2]*Severance Biomedical Science Institute, Brain Korea 21 Plus Project for Medical Science, Yonsei University, College of Medicine, Seoul, Republic of Korea*

Correspondence should be addressed to Hyun Seok Kim; hsfkim@yuhs.ac

Received 3 March 2017; Revised 21 April 2017; Accepted 23 May 2017; Published 13 June 2017

Academic Editor: Xingming Zhao

Current multiomics assay platforms facilitate systematic identification of functional entities that are mappable in a biological network, and computational methods that are better able to detect densely connected clusters of signals within a biological network are considered increasingly important. One of the most famous algorithms for detecting network subclusters is Molecular Complex Detection (MCODE). MCODE, however, is limited in simultaneous analyses of multiple, large-scale data sets, since it runs on the Cytoscape platform, which requires extensive computational resources and has limited coding flexibility. In the present study, we implemented the MCODE algorithm in R programming language and developed a related package, which we called MCODER. We found the MCODER package to be particularly useful in analyzing multiple omics data sets simultaneously within the R framework. Thus, we applied MCODER to detect pharmacologically tractable protein-protein interactions selectively elevated in molecular subtypes of ovarian and colorectal tumors. In doing so, we found that a single molecular subtype representing epithelial-mesenchymal transition in both cancer types exhibited enhanced production of the collagen-integrin protein complex. These results suggest that tumors of this molecular subtype could be susceptible to pharmacological inhibition of integrin signaling.

1. Introduction

Biological functions often arise from multisubunit protein complexes, rather than a single, isolated protein [1, 2]. Many high throughput assay platforms in genomics, transcriptomics, and proteomics have become standard methods for investigating gene/protein interactions that give rise to biological functions [3]. However, because of biological and technical errors, these methods are hindered by a limited signal-to-noise ratio, rendering them vulnerable to high rates of false positives and false negatives; particularly when discovered hits represent a single gene, protein, and so forth. In this regard, codiscovery of hits for multiple subunits of a protein complex in an experimental condition helps mutually support the significance of such findings [4]. Detection of higher order clusters in a large network, however, is computationally challenging [5]. A number of algorithms have been developed over the past decade to tackle this problem, including the Markov Cluster Algorithm (MCL) [6], Molecular Complex Detection (MCODE) [7], DPClus [8], Affinity Propagation Clustering (APC) [9], Clustering based on Maximal Clique (CMC) [10], ClusterMaker [11], and Clustering with Overlapping Neighborhood Expansion (ClusterONE) [12]. Many of these algorithms have been implemented in various Cytoscape applications (CytoCluster, ClusterViz [13], and ClusterMaker [11]), as well as in java-based applications (C-DEVA [14]). Of these, as of February 2017, MCODE was the most downloaded Cytoscape application within the clustering category. MCODE discovers interconnected network clusters based on k-core score: the k-core of a particular graph (graph X) represents the maximal number of connected subgraphs of graph X, in which all nodes are connected by k (minimum number of degrees). Although Cytoscape is a java-based, open source, bioinformatics software platform with

a user-friendly graphic-user interface [15], it requires extensive computational resources due to the memory restraints of java virtual machines (Cytoscape version 3.2.1: 2 GB+ recommended). Thus, its capacity to process input networks and graphical outputs is limited. For a computationally intensive task, R may be a better-suited platform. R is the most popular open source, statistical programming language, and data analysis platform used in analysis of broad, high throughput, and multiomics data. While the platform is suitable for iterative analysis of large-scale data sets in batch mode, R-based network clustering software is rare. Herein, we describe our implementation of the MCODE algorithm in R programming language and a related package, hereinafter referred to as MCODER. The MCODER package can be easily integrated into custom R projects and provides powerful and enhanced graphical output options, compared to its Cytoscape counterpart.

The Cancer Genome Atlas projects have classified tumors into subtypes that share distinct molecular and genetic features. To do so, researchers have leveraged multiomics data sets, including global and phosphoproteomic quantification, as well as DNA- and RNA-level measurements. Nevertheless, drawing associations between these subtypes and clinically important features, such as prognosis and therapeutic options, remains important challenges. In this study, we intended to focus on these challenges in high-grade serous ovarian carcinoma (HGS-OvCa) and colorectal cancer (CRC). Currently, standard treatment for ovarian cancer involves primary cytoreductive surgery, followed by platinum-based chemotherapy. Only two targeted therapies are clinically available for ovarian cancer, including poly (ADP-ribose) polymerase inhibitors and angiogenesis inhibitors in recurrent ovarian cancer [16], although they have been shown to offer little survival benefit. The four molecular subtypes of HGS-OvCa are differentiated, immunoreactive, proliferative, and mesenchymal, according to gene content analysis within each subtype, following transcriptome-based subtype classification [17, 18]. Of these, the mesenchymal subtype displays the worst prognosis [19, 20]. Meanwhile, CRC has four consensus molecular subtypes (CMSs): CMS1, CMS2, CMS3, and CMS4. The CMS subtypes of CRC are associated with various clinical features, such as sex, tumor site, stage at diagnosis, histopathological grade, and prognosis, as well as molecular features of microsatellite status, CpG island methylator phenotype (CIMP), somatic copy number alteration (SCNA), and enrichment of particular driver mutations. The CMS1 subtype exhibits high MSI, high CIMP, strong immune activation, and frequent *BRAF* mutation and involves an intermediate prognosis, showing worse survival after relapse. The CMS2 subtype displays a high degree of chromosomal instability (high SCNA), frequent *APC* mutation, and good prognosis. Tumors of the CMS3 subtype display mixed MSI, high SCNA, frequent *KRAS* mutation, metabolic deregulation, and good prognosis. Finally, the CMS4 subtype is characterized by distinct epithelial-mesenchymal transition (EMT) signature, high SCNA, and the poorest prognosis. Notably, in both cancers, mesenchymal subtype confers the worst prognosis. To gain insights into molecular subtype-selective opportunities for targeted therapies in ovarian and colorectal cancer, HGS-OvCa [21] and CRC [22] data sets were analyzed using MCODER. Both data sets contained mass-spec-based quantitative proteomic assay results for the well-defined molecular subtypes of these cancers. In particular, we aimed to identify pharmacologically tractable protein complexes selectively elevated within the distinct molecular subtypes of both cancers.

2. Implementation

MCODER identifies the maximal subset of vertices interconnected by the minimal number of degrees (k) from an input network of nodes (genes or proteins) and edges (pairwise interactions). Although the MCODER package does not account for the direction of the edges when calculating k-core scores and when detecting subnetworks, it can indicate directions using arrows and display multiple edges between a pair of nodes, which is not supported by the original MCODE. Moreover, various graphical parameters provided by "igraph" (http://igraph.org/redirect.html) can be manipulated in MCODER, facilitating customization of the shape, size, and color of the network output. The MCODER R package requires preinstallation of two other packages, "sna" (Social Network Analysis) (https://cran.r-project.org/web/packages/sna/index.html) for calculating k-cores and "igraph" for plotting figures.

The overall workflow of the present study to identify pharmacologically tractable protein complexes is presented in Figure 1. Before running MCODER, we downloaded the STRING database (Homo sapiens, v10.0) from http://string-db.org: STRING is an archive of direct (physical) and indirect (functional) protein-protein interactions [23]. We filtered low confident interactions by applying an interaction-score cutoff (score < 0.4) to obtain 13,159 genes with 738,312 interactions. In parallel, we downloaded and preprocessed proteome data sets by selecting samples that have preassigned molecular subtypes and matched normal controls to obtain input data sets: HGS-OvCa (n = 3,329 proteins, 140 samples) and CRC (n = 3,718 proteins, 70 samples) [21, 22]. HGS-OvCa consisted of four molecular subtypes: differentiated (n = 35 samples), immunoreactive (n = 37 samples), proliferative (n = 34 samples), and mesenchymal (n = 34 samples). CRC consisted of four molecular subtypes: CMS1 (n = 14 samples), CMS2 (n = 28 samples), CMS3 (n = 9 samples), and CMS4 (n = 14 samples). To identify differentially expressed proteins (DEPs) selectively elevated in a particular molecular subtype, a one-sided t-test was conducted iteratively within a tumor (e.g., CMS1 versus CMS2, CMS3, CMS4). After preparing differentially expressed protein sets, we converted them into adjacency matrices for each set, with connection information between nodes according to the STRING database, followed by calculation of k-core values, vertex density, and vertex score. Self-loop and duplicated connections between nodes were not considered for the calculation. Clusters were detected with the following parameters: minimal k-core value = 2, haircut = TRUE, fluff = FALSE, self-loop = FALSE, node score cutoff = 0.2, depth = 20, and degree cutoff = 2. Subsequently, vertices

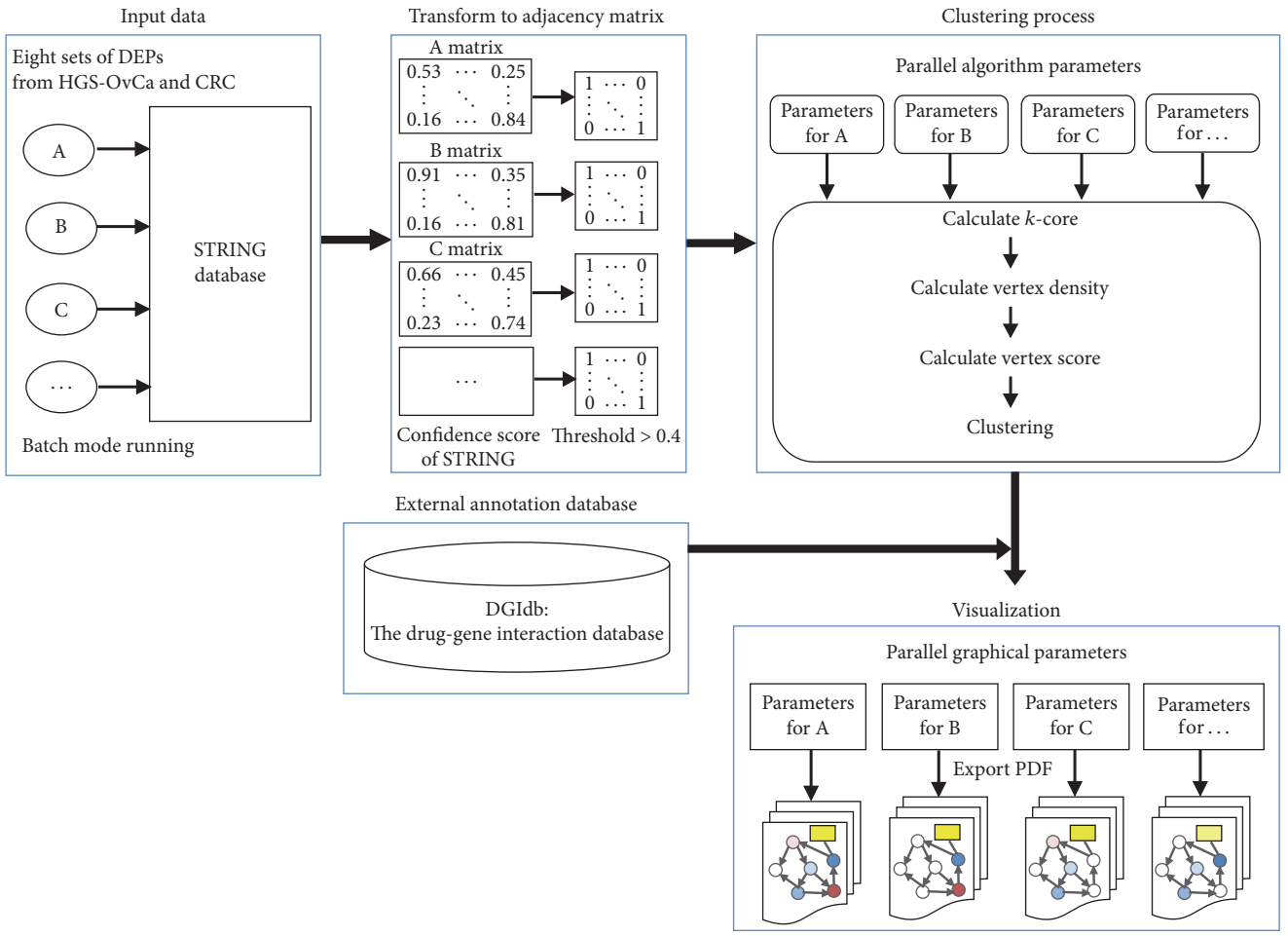


Figure 1: Workflow for detecting densely connected network clusters using MCODER. See Implementation for further details.

Table 1: Comparison of computational time and memory usage between MCODER and the MCODE Cytoscape application.

Network size	Performance	
	MCODER	Cytoscape MCODE
5K edges, 2,902 vertexes	6 s.	1 m. 14 s.
100K edges, 3,786 vertexes	11 s.	3 m. 44 s.
200K edges, 4,625 vertexes	19 s.	18 m. 47 s.
Memory usage	0.45 GB	5 GB

in the clusters were annotated according to the DGI database [24], allowing for detection of druggable DEPs.

3. Results

First, we examined the performance of MCODER (Figure 1) in comparison to the MCODE Cytoscape application, testing input networks of different sizes (Table 1). All tests were performed using MacBook Pro (Mac OS X, Late 2013, 2.4-GHz Intel Core i5, 8 GB RAM). Input data sets were prepared by random sampling of the given number of interactions from the STRING database. We found that both software packages returned identical protein complexes as an output. Meanwhile, however, MCODER in the R environment offered enhanced performance in regard to speed and memory usage in all test settings (Table 1). The MCODER installation package is available online at https://sourceforge.net/projects/mcoder.

Next, for the individual molecular subtypes, we identified selectively elevated proteins under a p value threshold of 0.01: 300 proteins for differentiated, 284 proteins for immunoreactive, 547 proteins for proliferative, and 493 proteins for mesenchymal HGS-OvCa and 236 proteins for CMS1, 284 proteins for CMS2, 134 proteins for CMS3, and 137 proteins for CMS4 subtypes of CRC (see Supplementary Data 1 in the Supplementary Material available online at https://doi.org/10.1155/2017/1016305). For each of the DEP sets, MCODER identified highly interconnected subnetworks of protein-protein interactions. For HGS-OvCa, we detected pharmacologically targetable clusters in three of the four subtypes (Supplementary Data 2). In the immunoreactive subtype, two clusters showed connections with pharmacological agents. The first cluster contained interferon-stimulated gene 15 *(ISG15)*, which is a biomarker for predicting sensitivity to irinotecan, an anticancer drug and topoisomerase I inhibitor (Figure 2(a)). Previous studies have demonstrated that *ISG15* encodes an ubiquitin-like protein conjugated to specific E3

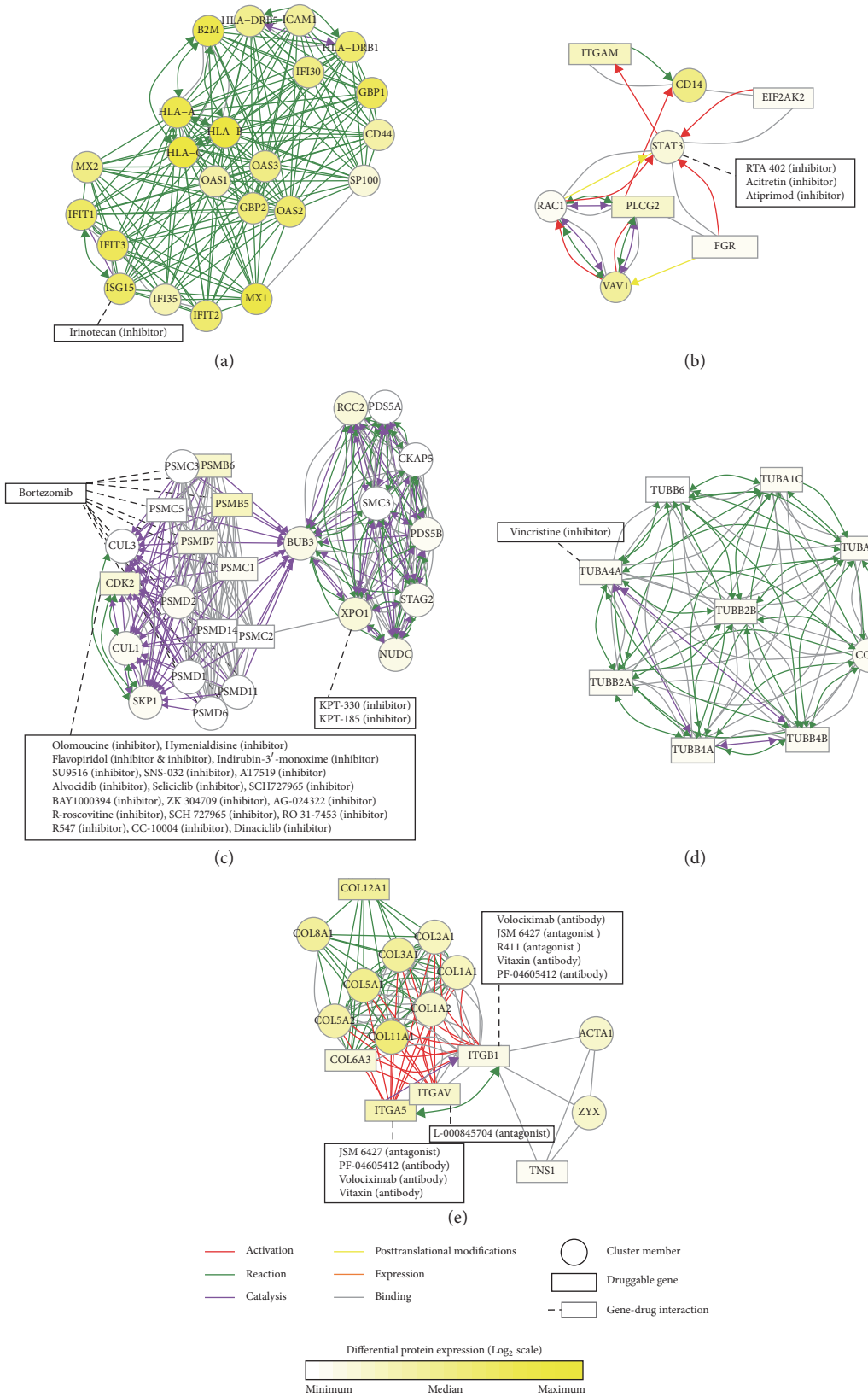


Figure 2: Pharmacologically targetable network clusters overexpressed in molecular subtypes of HGS-OvCa: (a, b) immunoreactive, (c, d) proliferative, and (e) mesenchymal subtype.

ubiquitin ligases and seems to inhibit the signaling consequences of ubiquitin/26S proteasome pathways [25]. Currently, treatments with irinotecan, in combination with bevacizumab or cisplatin, are in clinical trials for recurrent ovarian cancer [26]. Our findings suggest that selecting patients with immunoreactive features might increase response rates to irinotecan in future trials. The second cluster comprised a chemokine signaling related protein complex, including STAT3, which can be inhibited by RTA402, acitretin, and atiprimod (Figure 2(b)). Previous studies have indicated that STAT3 inhibitors, in combination with cisplatin, enhance cisplatin sensitivity in cisplatin-resistant ovarian cancer [27, 28]. Thus, a combination of irinotecan and STAT3 inhibitors might be plausible in treating ovarian cancers of immunoreactive subtype. In the proliferative subtype, two clusters displayed connections with pharmacological compounds. The CDK2-proteasome-XPO1 cluster was enriched with pharmacological options, including the proteasome inhibitor bortezomib, which is available clinically, and CDK2 and XPO1 inhibitors, which are under active clinical trials for various tumor types (Figure 2(c)) [29, 30]: XPO1 inhibitors have been used to target platinum-resistant ovarian tumors [31] and have been described as potentially inhibiting abnormal NF-kB signaling [32]. The second DEP cluster was the tubulin complex, in which TUBA4A can be targeted by vincristine to blunt mitotic chromosomal separation (Figure 2(d)). Similar to paclitaxel, a microtubule stabilizer and an antiproliferative agent [33], vincristine may be a potential agent for the treatment of ovarian cancer, particularly that of proliferative subtype. In the mesenchymal subtype, focal adhesion, endocytosis, vascular smooth muscle contraction, the PI3K-AKT signaling pathway, and so forth were identified. Of these, the integrin-collagen complex is a pharmacologically tractable target; various integrin signaling inhibitors include ITGA5 inhibitors (JSM 6427, PF-04605412, Volociximab, and Vitaxin), ITGB1 inhibitors (Volociximab, JSM 6427, R411, Vitaxin, and PF-04605412), and an ITGAV inhibitor (L-000845704) (Figure 2(e)). Integrin signaling is involved in the migration, invasion, proliferation, and survival of cancer cells [34]. Recently published studies have demonstrated that integrins participate in maintaining cancer stem cell populations and contribute to cancer progression and drug resistance [35]. Although integrin inhibitors as monotherapy agents have failed to demonstrate benefits in metastatic ovarian tumors, possibly due to compensation by other integrins [36], simultaneous targeting of integrin-FAK and c-Myc signaling has been found to synergistically disrupt tumor cell proliferation and survival in HGS-OvCa [37], supporting the notion of combinatorial targeting of integrin as a valid approach for treating ovarian cancer, particularly that of mesenchymal subtype.

For CRC, MCODER identified pharmacologically targetable protein complexes in three of the four CMSs (Supplementary Data 2). In CMS1 subtype (MSI immune), proteasome complex (similar to the HGS-OvCa proliferative subtype) and ROCK1 signaling subnetworks were found to be overexpressed (Figures 3(a)-3(b)). Bortezomib treatment has been shown to induce G2-M arrest by activation of an ataxia-telangiectasia mutated protein-cell cycle checkpoint kinase 1 pathway in colon cancer cells [38]. Combination of platelet-derived growth factor and the ROCK inhibitor Y27632 has been found to decrease the invasive potential of SW620 colon cancer cells [39]. In the CMS2 subtype (canonical), tubulin complex was found to be elevated, similar to the HGS-OvCa proliferative subtype (Figure 3(c)). This observation suggests that vincristine could have therapeutic effects on CRCs of CMS2 subtype. Alternatively, or in combination with microtubule inhibitors, Src inhibitors may also be a plausible approach for CMS2 tumors (Figure 3(d)). The CMS4 subtype of CRCs exhibits EMT activation and confers the poorest prognosis. Other study groups have formerly referred to this subtype as colon cancer subtype 3 [40] or stem-like subtype [41]. In CMS4 tumors, we found the total MAPK3 (ERK1) protein complex to be elevated, which is targetable with ERK inhibitor II (Figure 3(e)). Surprisingly, in accordance with the HGS-OvCa mesenchymal subtype, CMS4 was also characterized by elevation of the extracellular matrix collagen-integrin complex (Figure 3(f)): collagen in the extracellular matrix has indeed been found to drive EMT in CRC [42]. Thus, the collagen-integrin protein complex may work as a molecular linchpin that, when removed, could diminish the malignant potential of EMT tumors. Accordingly, we suggest that therapeutic antibodies that interrupt the signaling of integrin proteins could potentially be utilized as therapeutic options, in combination with other chemo- or targeted therapies, for this refractory subtype of colon cancer.

Finally, we sought to determine whether our findings are reproducible with other network clustering algorithms, including ClusterONE [12] and MCL [6]. Although the sizes of the detected clusters varied, all of the subclusters detected by MCODER were identified by these algorithms as well, indicating that our findings are robust across different clustering algorithms.

4. Discussion

In this study, we implemented the network clustering algorithm MCODE into the R software environment (which we called MCODER) and demonstrated that the MCODER package saves computational resources and time, making it particularly suited for analyzing multiple omics data sets. Using MCODER, we identified potential candidates for anticancer therapy in molecular subtypes of ovarian and colorectal cancer by detecting protein complexes that were selectively overexpressed therein and that could be targeted with known pharmacological agents. For HGS-OvCa, we found that irinotecan and STAT3 inhibitors may be candidates for the immunoreactive subtype, along with bortezomib, CDK2, XPO1 inhibitors, and vincristine for the proliferative subtype and integrin signaling inhibitors for the mesenchymal subtype. For CRC, we found bortezomib and ROCK inhibitors to be potential candidates for the CMS1 subtype, along with vincristine and Src inhibitors for the CMS2 subtype and ERK inhibitor II and integrin signaling inhibitors for the CMS4 subtype. Importantly, our analyses revealed that the collagen-integrin protein complex, which is pharmacologically tractable, is commonly overexpressed

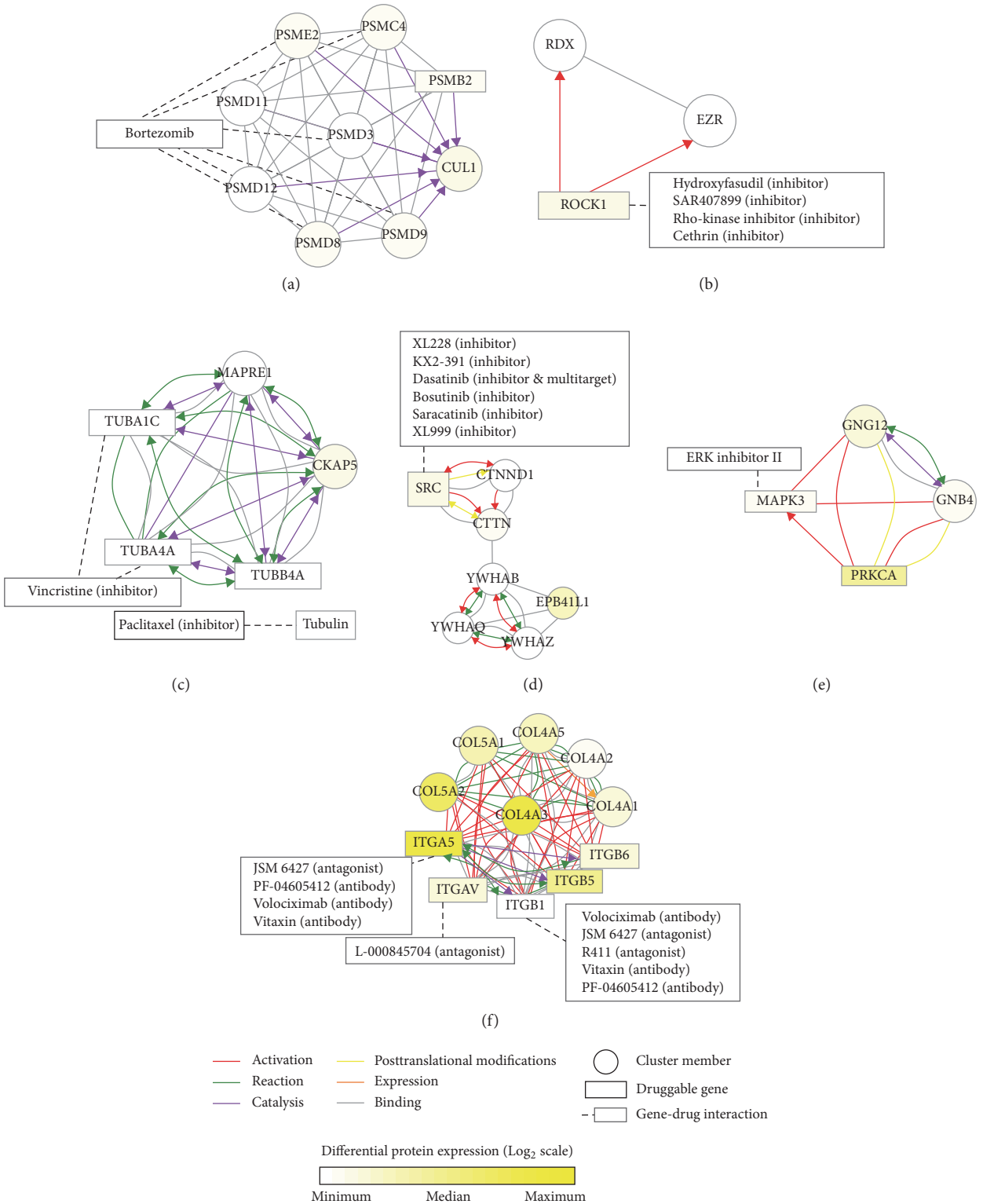


Figure 3: Pharmacologically targetable network clusters overexpressed in molecular subtypes of CRC: (a, b) CMS1, (c, d) CMS2, and (e, f) CMS4.

in EMT subtypes of both ovarian and colorectal cancers. Further studies are needed to determine whether pharmacological inhibition of collagen-integrin signaling blunts tumor growth in an in vivo model of EMT cancer.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Sungjin Kwon and Hyosil Kim contributed equally to this work.

Acknowledgments

This study was supported by grants from the National R&D Program for Cancer Control, Ministry of Health & Welfare, Republic of Korea (1420100), from the Korea Health Technology R&D project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare (HI14C1324), and from Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2014R1A1A2057232) and a faculty research grant from Yonsei University College of Medicine for 2014 Grant no. 6-2014-0066.

References

[1] T. Ideker and R. Sharan, "Protein networks in disease," *Genome Research*, vol. 18, no. 4, pp. 644–652, 2008.

[2] A. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[3] J. T. Erler and R. Linding, "Network-based drugs and biomarkers," *Journal of Pathology*, vol. 220, no. 2, pp. 290–296, 2010.

[4] K. Fortney and I. Jurisica, "Integrative computational biology for cancer research," *Human Genetics*, vol. 130, no. 4, pp. 465–481, 2011.

[5] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, pp. 163–166, 2016.

[6] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic Acids Research*, vol. 30, no. 7, pp. 1575–1584, 2002.

[7] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, no. 1, p. 2, 2003.

[8] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya, "Development and implementation of an algorithm for detection of protein complexes in large interaction networks," *BMC Bioinformatics*, vol. 7, article 207, 2006.

[9] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *American Association for the Advancement of Science. Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[10] G. Liu, L. Wong, and H. N. Chua, "Complex discovery from weighted PPI networks," *Bioinformatics*, vol. 25, no. 15, pp. 1891–1897, 2009.

[11] J. H. Morris, L. Apeltsin, A. M. Newman et al., "ClusterMaker: a multi-algorithm clustering plugin for Cytoscape," *BMC Bioinformatics*, vol. 12, article 436, 2011.

[12] 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.

[13] J. Wang, J. Zhong, G. Chen, M. Li, F.-X. Wu, and Y. Pan, "ClusterViz: A Cytoscape APP for Cluster Analysis of Biological Network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 4, pp. 815–822, 2015.

[14] M. Li, Y. Tang, X. Wu, J. Wang, F.-X. Wu, and Y. Pan, "C-DEVA: Detection, evaluation, visualization and annotation of clusters from biological networks," *BioSystems*, vol. 150, pp. 78–86, 2016.

[15] 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.

[16] S. Vaughan, J. I. Coward, R. C. Bast et al., "Rethinking ovarian cancer: recommendations for improving outcomes," *Nature Reviews Cancer*, vol. 11, no. 10, pp. 719–725, 2011.

[17] R. G. Verhaak, P. Tamayo, J. Y. Yang et al., "Prognostically relevant gene signatures of high-grade serous ovarian carcinoma," *Journal of Clinical Investigation*, vol. 123, no. 1, pp. 517–525, 2013.

[18] Cancer Genome Atlas Research Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, pp. 609–615, 2011.

[19] R. W. Tothill, A. V. Tinker, J. George et al., "Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome," *Clinical Cancer Research*, vol. 14, no. 16, pp. 5198–5208, 2008.

[20] X. Yin, X. Wang, B. Shen et al., "A VEGF-dependent gene signature enriched in mesenchymal ovarian cancer predicts patient prognosis," *Scientific Reports*, vol. 6, Article ID 31079, 2016.

[21] H. Zhang, T. Liu, and Z. Zhang, "Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer," *Cell*, vol. 166, no. 3, pp. 755–765, 2016.

[22] B. Zhang, J. Wang, and X. Wang, "Proteogenomic characterization of human colon and rectal cancer," *Nature*, vol. 513, no. 7518, pp. 382–387, 2014.

[23] D. Szklarczyk, A. Franceschini, S. Wyder et al., "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, pp. D447–D452, 2015.

[24] A. H. Wagner, A. C. Coffman, B. J. Ainscough et al., "DGIdb 2.0: mining clinically relevant drug–gene interactions," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1036–D1044, 2016.

[25] S. D. Desai, A. L. Haas, L. M. Wood et al., "Elevated expression of ISG15 in tumor cells interferes with the ubiquitin/26S proteasome pathway," *Cancer Research*, vol. 66, no. 2, pp. 921–928, 2006.

[26] F. Musa, B. Pothuri, S. V. Blank et al., "Phase II study of irinotecan in combination with bevacizumab in recurrent ovarian cancer," *Gynecologic Oncology*, vol. 144, no. 2, pp. 279–284, 2016.

[27] K. Selvendiran, S. Ahmed, A. Dayton et al., "HO-3867, a curcumin analog, sensitizes cisplatin-resistant ovarian carcinoma, leading to therapeutic synergy through STAT3 inhibition," *Cancer Biology & Therapy*, vol. 12, no. 9, pp. 837–845, 2011.

[28] Z. Duan, R. Y. Ames, M. Ryan, F. J. Hornicek, H. Mankin, and M. V. Seiden, "CDDO-Me, a synthetic triterpenoid, inhibits expression of IL-6 and Stat3 phosphorylation in multi-drug resistant

ovarian cancer cells," *Cancer Chemotherapy and Pharmacology*, vol. 63, no. 4, pp. 681–689, 2009.

[29] R. Roskoski, "Cyclin-dependent protein kinase inhibitors including palbociclib as anticancer drugs," *Pharmacological Research*, vol. 107, pp. 249–275, 2016.

[30] A. R. Abdul Razak, M. Mau-Soerensen, N. Y. Gabrail et al., "First-in-class, first-in-human phase I study of selinexor, a selective inhibitor of nuclear export, in patients with advanced solid tumors," *Journal of Clinical Oncology*, vol. 34, no. 34, pp. 4142–4150, 2016.

[31] Y. Chen, S. C. Camacho, T. R. Silvers et al., "Inhibition of the nuclear export receptor XPO1 as a therapeutic target for platinum-resistant ovarian cancer," *Clinical Cancer Research*, 2016.

[32] J. Kim, E. McMillan, H. S. Kim et al., "XPO1-dependent nuclear export is a druggable vulnerability in KRAS-mutant lung cancer," *Nature*, vol. 538, no. 7623, pp. 114–117, 2016.

[33] B. Ai, Z. Bie, S. Zhang, and A. Li, "Paclitaxel targets VEGF-mediated angiogenesis in ovarian cancer treatment," *American Journal of Cancer Research*, vol. 6, no. 8, pp. 1624–1635, 2016.

[34] J. S. Desgrosellier and D. A. Cheresh, "Integrins in cancer: biological implications and therapeutic opportunities," *Nature Reviews Cancer*, vol. 10, no. 1, pp. 9–22, 2010.

[35] L. Seguin, J. S. Desgrosellier, S. M. Weis, and D. A. Cheresh, "Integrins and cancer: regulators of cancer stemness, metastasis, and drug resistance," *Trends in Cell Biology*, vol. 25, no. 4, pp. 234–240, 2015.

[36] K. Sawada, C. Ohyagi-Hara, T. Kimura, and K.-I. Morishige, "Integrin inhibitors as a therapeutic agent for ovarian cancer," *Journal of Oncology*, Article ID 915140, 2012.

[37] B. Xu, J. Lefringhouse, Z. Liu et al., "Inhibition of the integrin/FAK signaling axis and c-Myc synergistically disrupts ovarian cancer malignancy," *Oncogenesis*, vol. 6, no. 1, article e295, 2017.

[38] Y. S. Hong, S.-W. Hong, S.-M. Kim et al., "Bortezomib induces G 2-M arrest in human colon cancer cells through ROS-inducible phosphorylation of ATM-CHK1," *International Journal of Oncology*, vol. 41, no. 1, pp. 76–82, 2012.

[39] M. de Toledo, C. Anguille, L. Roger, P. Roux, and G. Gadea, "Cooperative Anti-Invasive Effect of Cdc42/Rac1 Activation and ROCK Inhibition in SW620 Colorectal Cancer Cells with Elevated Blebbing Activity," *PLoS ONE*, vol. 7, no. 11, Article ID e48344, 2012.

[40] F. De Sousa, E. Melo, X. Wang, and M. Jansen, "Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions," *Nature Medicine*, vol. 19, no. 5, pp. 614–618, 2013.

[41] A. Sadanandam, C. A. Lyssiotis, K. Homicsko et al., "A colorectal cancer classification system that associates cellular phenotype and responses to therapy," *Nature Medicine*, vol. 19, no. 5, pp. 619–625, 2013.

[42] T. T. Vellinga, S. Den Uil, I. H. B. Rinkes et al., "Collagen-rich stroma in aggressive colon tumors induces mesenchymal gene expression and tumor cell invasion," *Oncogene*, vol. 35, no. 40, pp. 5263–5271, 2016.

Hindawi
BioMed Research International
Volume 2017, Article ID 7049406, 8 pages
https://doi.org/10.1155/2017/7049406

Review Article

Methods of MicroRNA Promoter Prediction and Transcription Factor Mediated Regulatory Network

Yuming Zhao,[1,2] Fang Wang,[3] Su Chen,[1] Jun Wan,[4] and Guohua Wang[3]

[1]*State Key Laboratory of Tree Genetics and Breeding, Northeast Forestry University, Harbin, Heilongjiang, China*
[2]*Information and Computer Engineering College, Northeast Forestry University, Harbin, Heilongjiang, China*
[3]*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang, China*
[4]*Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA*

Correspondence should be addressed to Yuming Zhao; zymyoyo@hotmail.com

Received 3 March 2017; Accepted 7 May 2017; Published 5 June 2017

Academic Editor: Weihua Chen

MicroRNAs (miRNAs) are short (~22 nucleotides) noncoding RNAs and disseminated throughout the genome, either in the intergenic regions or in the intronic sequences of protein-coding genes. MiRNAs have been proved to play important roles in regulating gene expression. Hence, understanding the transcriptional mechanism of miRNA genes is a very critical step to uncover the whole regulatory network. A number of miRNA promoter prediction models have been proposed in the past decade. This review summarized several most popular miRNA promoter prediction models which used genome sequence features, or other features, for example, histone markers, RNA Pol II binding sites, and nucleosome-free regions, achieved by high-throughput sequencing data. Some databases were described as resources for miRNA promoter information. We then performed comprehensive discussion on prediction and identification of transcription factor mediated microRNA regulatory networks.

1. Introduction

MicroRNAs (miRNAs) are small noncoding RNAs with about 22 nucleotides, which are transcribed by noncoding DNA sequences [1, 2]. MiRNAs are disseminated throughout the genome. They were found either in the intergenic regions or in the intronic sequences of protein-coding genes. It has been known that miRNAs are key elements in many species, such as human and mouse, to function in posttranscriptional gene regulation. One single miRNA can influence one-third of human genome by potentially regulating thousands of genes at the same time [3]. Similar to protein-coding genes, miRNAs were also regulated by transcription factors (TFs) at transcription level. Uncovering the transcriptional mechanisms of miRNAs themselves can help people better understand regulatory networks of gene expression.

The promoters of genes are important regions, bound by different regulatory elements to start and regulate the transcription [4, 5]. Locating the promoter regions of genes is crucial for revealing the transcriptional mechanism. It remains difficult to define miRNA promoters and understand how TFs regulate downstream miRNAs. The classical features of promoter regions, including signal, context, and structure features, can be used to recognize miRNAs from other sequences [6]. However, only a fraction of the human miRNAs have their transcription start sites (TSSs) confirmed. Insufficient knowledge of the TSSs of miRNA genes limited our ability to study the transcriptional mechanism and the regulatory function of miRNAs. While most of promoter prediction methods based on the promoters of protein-coding genes may not be suitable for miRNA genes, it is required to develop promoter prediction methods special for miRNA genes.

In recent years, more and more prediction models have been developed to identify the miRNA promoters [7–12]. These studies utilized genome sequence features or took advantage of the high-throughput sequencing technology to identify the putative promoter regions of miRNA genes. In this article, we reviewed algorithms of miRNA promoter recognition based on genome sequence features, histone

markers, RNA Pol II binding, and nucleosome-free regions achieved from high-throughput sequencing data, respectively. We performed a comparative analysis on these models and corresponding identified miRNA promoter regions. In order to better understand the regulatory mechanisms of miRNA, more and more databases have been developed to collect miRNA promoter regions by integrating different prediction models. We also evaluated several databases collecting such miRNA promoter information. In the last part, we discussed the TF-miRNA regulatory networks either predicted by computational methods or derived by high-throughput experiments.

2. A Survey on Methods for MiRNA Promoter Regions Prediction

The prediction of miRNA promoters is significant for constructing the regulatory network of TF-miRNA or miRNA gene and further understanding the regulatory function of miRNAs. Several most popular prediction approaches used traditional genome sequence features, either individual one or mixed features, whereas more and more methods adopted next-generation sequencing (NGS) data to employ the information of histone markers, RNA Pol II binding sites, and nucleosome-free regions. Below is a survey on some representative methods.

2.1. Prediction Methods Using Traditional Genome Sequence Features

2.1.1. Individual Genome Sequence Features-Based Method. At the early beginning, researchers used one single genome sequence feature, expressed sequence tags (ESTs), to predict miRNA promoter regions. ESTs technology directly originated from the human genome project to construct the genetic map of genome. Many intergenic miRNAs are transcribed as pri-miRNAs. Gu et al. successfully predicted the location of pre-miRNAs by mapping the ESTs to the long flanking sequences. They then used EST-extension method to predict the location of about tens of pri-miRNA [13]. By comparing promoters of known miRNAs and protein-coding genes, Zhou et al. discovered that the transcriptional mechanism of miRNAs was similar to that of protein-coding genes in that both miRNAs and protein-coding genes were transcribed by RNA Pol II [14]. By relying on the sequence feature of known Pol II promoters, they extracted all possible k-mers as such features and used WordSpy algorithm [21, 22] to discover sequence motifs. Then they developed a new approach, CoVote [14], to predict unknown core promoters of miRNAs. CoVote was based on the decision tree algorithm followed by training well-known Pol II promoters compared to randomly selected sequences. The method has been proved to create good predictions by being applied on four species, *C. elegans*, *H. sapiens*, *A. thaliana*, and *O. sativa*.

2.1.2. Mixed Genome Sequence Features-Based Method. As we discussed previously, early modeling of miRNA promoters focused only on individual sequence features [13, 14]. While genome sequences have plenty of different features, combining these features can improve the accuracy of miRNA promoters' prediction. Genome sequences are composed of four bases, A, C, G, and T. The different assemblies of four bases form the sequence features of genome, such as TATA box, CAAT box, and GC box. Using the TRANSFAC weight matrices of TATA box, CAAT box, and GC box, Fujita and Iba utilized an entropy-based calculation to search the promoter of miRNA genes, which was implemented in the aligned and conserved blocks that contained miRNA hairpin regions [15]. To verify this method, they predicted 59 core promoter regions for 79 miRNAs, which were conserved between human and chicken or between human and zebrafish.

Furthermore, by incorporating several different sequence features, Bhattacharyya et al. used SVM model to predict TSSs of intergenic miRNA [17]. They extracted a large number of sequencing features in their study, such as N-mer features, palindromic features, special features, and CpG island based features. Those miRNA TSSs experimentally verified in previous studies were used to design the SVM classification model. Then they used well-trained complex AMOSA-SVM model to recognize unknown miRNA TSSs.

Similar with the above approach, Marsico et al. proposed a new approach, named PROmiRNA [16], based on a semisupervised statistical model. First, the TSS clusters of pre-miRNAs were generated. Second, they normalized the TSS clusters by removing the TSS clusters overlapping with the start of other protein-coding transcripts or spanning exon regions. Third, the sequence features, including CpG density, conservation score, TATA box affinity, and normalized tag counts, were calculated around the putative TSSs regions and the random regions. The region with a higher probability of being a promoter region than being a nonpromoter region is determined as a potential promoter region.

The summary of the prediction results of these methods is shown in Table 1, including the number of the putative miRNA promoter region of every method using genome sequence features. However, these methods using genome sequence features still have limitations in different tissues and species and hence their accuracy is not high enough.

2.2. Prediction Methods Using High-Throughput Sequencing Data. With rapid development of the NGS technology, whole genome and exome sequencing provides researchers with the opportunity to deal with the complex transcriptional and regulatory problem. Many sequencing technology such as RNA-seq and ChIP-seq can obtain the detailed information of genes, TFs, histone markers, nucleosome-free regions, and so on. Nowadays, more and more high-throughput sequencing data about miRNA expression have been collected, providing the opportunity to more accurately identify the TSS of miRNA and predict miRNA promoters.

2.2.1. Histone Markers-Based Method. Histone modifications represent different chromatin states. The NGS technology, ChIP-seq, is widely used to recognize locations of histone modifications. Many previous studies have showed that H3K4me3 was enriched in miRNA promoter regions, similar to that in the promoters of protein-coding genes. Therefore,

Table 1: Putative miRNA promoter numbers using traditional sequencing data.

Method name	Number of putative human promoters	Number of putative other species promoters			References
EST-extension	41	Rat	Mouse		[13]
		517	162		
CoVote	107	*C. elegans*	*A. thaliana*	*O. sativa*	[14]
		73	95	114	
miPPRs	59	—			[15]

using the data of histone modifications becomes popular to predict miRNA promoters. Marson et al. used the ChIP-seq data of H3K4me3 containing genomic enriched loci of H3K4me3, to predict the TSSs of miRNA genes in human and mouse genomes [9]. As a consequence, almost 80% of miRNAs promoters were identified in human and mouse genome.

In 2009, Wang et al. developed a computational program, called CoreBoost_HM, which combines several DNA features with histone modification [8]. The DNA features included the core promoter elements score, density of transcription factor binding sites (TF BSs), Markovian log-likelihood ratio scores, and N-mer frequencies. The boosting algorithm was used to model these feature data to predict the core promoters of miRNA genes. This combination of DNA features and histone modification improved the accuracy of prediction of miRNA promoters.

Different types of histone markers exhibit different patterns and functions in the genome sequences. In a previous study, we used nine different histone markers to predict miRNA promoters in *Arabidopsis* [18]. These histone markers included H3K4me2, H3K4me3, H3K9Ac, H3K9me2, H3K18Ac, H3K27me1, H3K27me3, H3K36me2, and H3K36me3. The RPM (reads per million per 100 bp bin) values of these nine histone modifications were extracted from corresponding ChIP-seq experiments for each known and unknown promoter region, indicating their binding patterns on these regions, respectively. The SVM model was trained based on these datasets, by using radial basis function (RBF) as the kernel function. Finally we identified TSSs of most miRNA genes and analyzed distinct histone patterns around the predicted TSSs of miRNA genes.

2.2.2. Pol II Binding-Based Method. It is believed that most miRNA genes are also transcribed by RNA polymerase II (Pol II), just like the protein-coding genes [23, 24], although some exceptions exist [25]. The binding of Pol II on the genome sequences can be used to investigate the transcriptional mechanism of miRNA genes. In order to start the transcription, Pol II always binds in close proximity to the TSSs of genes. In other words, Pol II binding pattern may be a key element of the promoter prediction. To better make out the transcriptional mechanism of miRNAs, Corcoran et al. performed ChIP-chip experiment for Pol II [11]. Based on SVM, they developed an efficient method for predicting core promoter, called CPPP. They successfully applied these tools to predict miRNA TSSs and analyzed the transcriptional mechanisms of miRNA genes.

Using genome-wide Pol II binding patterns, Wang et al. designed a computational approach to identify the promoter regions of miRNA genes [10]. A statistical model was developed to simulate the binding patterns of Pol II around the known TSSs of highly expressed protein-coding genes. Utilizing maximum likelihood estimation, they selected the best parameters that described the binding patterns of Pol II around TSSs. According to the assumption that the Pol II distribution around the TSSs of miRNA genes is similar to that around the TSSs of protein-coding genes, the upstream regions of miRNAs were then scanned to search for the regions with similar simulated Pol II binding patterns. These regions were inferred as the putative TSSs of miRNA genes.

To predict the promoter of *Arabidopsis* miRNAs, Zhao et al. performed ChIP analysis of Pol II in *Arabidopsis* using a genome tiling microarray based on the function of Pol II [19]. Using the approach of sliding window, the Pol II binding profiles around the known TSS of 59 miRNA genes were obtained. To predict TSSs for miRNA genes, they developed a procedure with three major steps: (i) setting the loci in the upstream of the Pol II signal intensity valley as an initial start position; (ii) using motif matcher to search for TATA box around the start point; (iii) scanning the same region by using the transcription initiation motifs verified by experiments. Then different TSS was identified for each miRNA gene based on the different position of TATA box.

The previous studies have indicated that H3K4me3, Pol II, and TFs played important roles in regulating the expression of miRNA genes. While most of the above studies used just one type of feature to identify miRNA TSSs, Georgakilas et al. incorporated three different types of features, including H3K4me3 peaks, Pol II peaks, and DNaseI peaks data, to construct a method, named MicroTSS, for predicting miRNA TSSs [20]. They first utilized these three features to train three SVM models using libsvm v3.0. Then MicroTSS was developed by combining H3K4me3, Pol II, and DNaseI occupancy models together. Considering that miRNA genes had the similar expression mechanism with protein-coding genes, they applied MicroTSS acquired by protein-coding genes data to predict miRNA TSSs.

2.2.3. Nucleosome-Free Region-Based Method. It is known that TFs generally bind in nucleosome-free regions, which is proximity to the TSSs, to activate downstream genes. Promoter regions usually reveal some significant features, such as high evolutionary conservation, nucleosome-depleted regions, CpG islands, TFBS motif within regions, and specific histone modification containing H3K4me3, H3K9ac, and H3K14ac. Based on the assumption that a nucleosome-free

Table 2: The 20 common putative miRNA promoters predicted by four models.

Name	Chrom	microRNA position	Marson	X. Wang	Ozsolak	Guohua Wang
hsa-mir-200b	Chr1	1092347–1092441	1088265	1088515	1087712	1088380
hsa-mir-200a	Chr1	1093106–1093195	1088265	1088515	1087791	1088380
hsa-mir-429	Chr1	1094248–1094330	1088265	1088515	1087795	1088380
hsa-mir-92b	Chr1	153431592–153431687	153429179	153429505	153430515	153430271
hsa-mir-148a	Chr7	25956064–25956131	25955148	25957430	25957069	25957227
hsa-mir-182	Chr7	129197459–129197568	129204548	129206490	129206638	129207158
hsa-mir-96	Chr7	129201768–129201845	129204548	129206490	129206331	129207158
hsa-mir-183	Chr7	129201981–129202090	129204548	129206490	129206299	129207158
hsa-let-7a-1	Chr9	95978060–95978139	95968291	95968360	95967305	95968990
hsa-let-7f-1	Chr9	95978450–95978536	95968291	95968360	95967585	95968990
hsa-let-7d	Chr9	95980937–95981023	95968291	95968360	95967585	95968990
hsa-mir-345	Chr14	99843949–99844046	99840674	99842750	99840834	99843433
hsa-mir-484	Chr16	15644652–15644730	15643092	15644680	15643760	15644503
hsa-mir-99b	Chr19	56887677–56887746	56883717	56884440	56882679	56884486
hsa-let-7e	Chr19	56887851–56887929	56883717	56884440	56884876	56884486
hsa-mir-125a	Chr19	56888319–56888404	56883717	56884440	56883012	56884486
hsa-mir-659	Chr22	36573631–36573727	36573792	36575350	36574528	36575388
hsa-mir-545	ChrX	73423664–73423769	73426274	73428915	73428342	73428923
hsa-mir-374a	ChrX	73423846–73423917	73426274	73428915	73428487	73428923
hsa-mir-505	ChrX	138833973–138834056	138840014	138842900	138842217	138842643

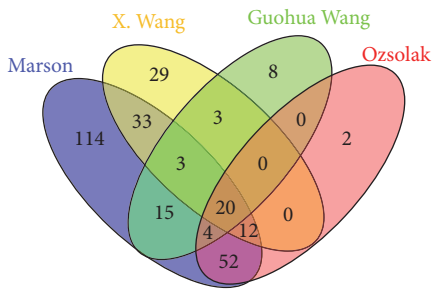


Figure 1: *Vann diagram of putative promoter regions*. The promoter was predicted by four models proposed by Marson et al., X. Wang et al., Ozsolak et al., and Guohua Wang et al., respectively.

region within the ChIP-enriched site may contain a TSS, Ozsolak et al. utilized these characteristic features to develop a scoring function to predict miRNA promoters [7]. The center of the valley with the highest score was defined as the putative TSS.

We make a comparison of four models using high-throughput sequencing data [7–10]. Figure 1 shows the Venn diagram of the prediction promoter regions of these four models. The putative miRNA promoter regions of the model used by Marson et al. are chosen as the criterion to compare with other three putative results. Since genomic coordinates of datasets of Marson et al. are based on GRCh37/hg19, the other three datasets are based on NCBI36/hg18, the liftOver program obtained from the UCSC Genome browser [26] was applied to convert genomic loci of datasets of Marson et al. into NCBI36/hg18. There are 20 common putative miRNA promoter regions predicted by these four models, shown in Table 2.

3. The Database of miRNA Promoter Construction

There were not many studies on miRNA promoter at the early stage. Moreover, most reports focused on only a few miRNAs in special species or tissues. In recent years, more and more investigations about the prediction of miRNA promoter regions have appeared. In order to make researchers have a comprehensive understanding of miRNA genes expression and functions, there are increasing numbers of databases that collect the promoter information of different miRNA and provide analysis tools to researchers.

Bhattacharyya et al. constructed a database, named miRT, which accumulated the validated miRNA TSSs of the previous studies [27]. They searched PubMed extensively to obtain the information about miRNA TSSs. The miRT database covers 670 TSS loci of 588 miRNAs with a minimum support value of one, which includes 206 inter-miRNAs and 382 intra-miRNAs. Some miRNAs may have multiple TSSs. The miRT database is available at http://www.isical.ac.in/~bioinfo_miu/miRT/miRT.php.

Chien et al. constructed the database, miRStart, a novel resource of human miRNA TSSs [12]. It systematically incorporates three significant datasets, including CAGE tags, TSSs seq data, and H3K4me3 ChIP-seq data, derived from TSS-relevant experiments to identify TSSs of miRNAs. In general, a high-confidence TSS is recommended for each miRNA genes based on a SVM training model. Through the database, users can define their preferable miRNA TSSs according to the straightforward display of experimental TSS signals. In total, miRStart involves 940 human miRNAs. Among them, 352 miRNAs are inter-miRNAs, and 588 miRNAs are intra-miRNAs. The

Table 3: The number of miRNA promoters in five databases.

Database	miRNAs	Inter-miRNAs	Intra-miRNAs	Species
miRStart	940	352	588	Human
miRT	588	206	382	Human
DIANA-miRGen	428	428	0	Human, mouse
miRGen	1189	766	423	Human, mouse
AtmiRNET	281	237	44	Arabidopsis

miRStart database is freely available at http://mirstart.mbc.nctu.edu.tw/.

Panagiotis Alexiou et al. constructed miRGen database, providing the promoter positions of miRNA genes in human and mouse, and their regulation by TFs [28]. The data are supported by experimental results. The information about microRNA coding transcripts, such as promoter regions, is supported by four literature sources: (i) Corcoran et al. [11], (ii) Landgraf et al. [29], (iii) Ozsolak et al. [7], and (iv) Marson et al. [9]. In total, there are 812 human miRNAs and 386 mouse miRNA coding transcripts' information stored in this database. Among these, 423 miRNAs are intra-miRNAs. In addition, this database shows binding sites of some TFs on the promoter regions of miRNAs and the information about SNPs. The miRGen database is freely available at http://www.microrna.gr/mirgen/.

To accurately characterize the mechanisms of miRNA transcription regulation, Georgakilas et al. constructed DIANA-miRGen v3.0 database to provide accurate TSSs of miRNA genes and the genome-wide maps of TFBSs [30]. According to their previous work [20], they used microTSS algorithm to accurately predict 276 miRNA TSSs. These accurately identified miRNA TSSs and TFBSs are stored in the database. The database DIANA-miRGen v3.0 is available at http://www.microrna.gr/mirgen.

The above databases are all about human and mouse miRNAs. To provide comprehensive information about plant miRNA genes, Chien et al. established the AtmiRNET database [31]. They used high-throughput next-generation sequencing datasets to construct SVM prediction model to predict *Arabidopsis* miRNA TSSs. This database also provides the transcriptional regulation on miRNA genes and putative miRNA-target interactions. In total, 281 *Arabidopsis* miRNA TSSs are provided in this study. Among them, 44 miRNAs are intra-miRNA, and this study used TSSs of host genes to define intra-miRNA TSSs. This database is very helpful in that users can understand the transcriptional mechanisms and regulatory functions of miRNA in *A. thaliana*. The AtmiRNET database is freely available at http://AtmiRNET.itps.ncku.edu.tw/. Table 3 shows the statistics of all five databases discussed above.

4. The Analysis of the Construction of the TF-miRNA Regulatory Networks

According to previous studies, most miRNAs are transcribed by noncoding genes, which are also regulated by related transcription factors. It remains unclear how TFs regulate miRNA genes. Constructing the regulatory network of TFs on miRNA genes is a critical step to better understand the functional mechanism of related miRNAs. In recent years, TF-miRNA network has captured increased attentions. People established such network by building computational models or utilizing NGS experiment data.

4.1. Computational Methods. Based on Pol II binding patterns around TSSs, Wang et al. developed an approach to predict inter-miRNAs promoter regions [32]. After that, they used position-specific score matrices (PSSM) to predict the TFBSs of STAT1 on genomic regions. Compared with the background promoters nonoverlapped with ChIP-enriched regions of STAT1, it is believed that STAT1 regulates this miRNA if the binding sites are more enriched in specific miRNA promoters. TargetScan was then used for microRNA target prediction to construct the feedback network of STAT1 and miRNAs.

To identify *Arabidopsis* miRNA promoters, Chien et al. established a SVM-based model [33]. First, they paired coexpressed annotated genes with specific miRNAs. By using PWMs from TRANSFAC, they adopted Match program [34] to search TFBSs motifs and defined the coTFBS as the common TFBS motifs that coincided in the promoters of a miRNA and its coexpressed genes. According to the assumption that genes with coexpression pattern may be regulated by the same TFs, the TFs with high frequency of coTFBSs are thought to regulate this miRNA. Finally, the regulatory networks about TFs and miRNAs are visualized by the Cytoscape software.

The previous related studies just provided limited regulatory network of TFs on miRNAs, which restricted the identification of novel TF-miRNA networks. Thus, Falcone et al. developed a software, named infinity, to reveal new regulatory networks of TFs and miRNAs [35]. They collected TSS positions from miRStart and extracted the promoter region sequences of miRNAs from UCSC Genome Browser. This software allows users to search the binding matrix of TFs on the defined promoter regions. This flexibility in this research offers the possibility of establishing unknown TF-miRNAs regulatory networks.

4.2. Experimental Evidence-Based Method. Most of the computation methods described above were developed based on the human or mouse genome. Nowadays, people have paid more and more attention to the expressional regulation of other species. Martinez et al. constructed miRNAs regulatory network on the *C. elegans* genome, using high-throughput sequencing technology to experimentally map transcriptional TF-miRNA interactions [36]. For constructing the

Table 4: The features used in the miRNA promoter prediction models.

Literature	EST	N-mer	TATA box	CAAT box	GC box	CpG island	Conservation	TFBS	DNase I	Histone marker	Pol II	Nucleosome
[13]	√											
[14]		√										
[15]			√	√	√							
[16]	√		√			√	√					
[17]		√				√						
[9]										√		
[8]		√						√		√		
[18]										√		
[10]											√	
[11]											√	
[19]											√	
[20]									√	√	√	
[7]						√	√	√		√		√

feedback network of miRNA-TF, they used previous algorithms, such as Pictar [37] and miRanda Targets version 4 [38], to predict the target of miRNAs on specific TFs.

In the meantime, more and more relative databases have been constructed for TF-miRNA regulatory network. For example, TSmiR, constructed by Guo et al., is a database that stores the regulatory networks of TFs and miRNAs in 12 human tissues. Those interactions were derived from the high-throughput experimental data [39]. In total, TSmiR database involves 116 TS miRNAs, 101 TFs, and 2347 TF-miRNA regulatory relations of 12 tissues and is freely available at http://bioeng.swjtu.edu.cn/TSmiR.

TFs and miRNAs are two key elements in the regulation of genes. The regulatory relations, TF-miRNA-target gene, are extremely complex, but they play an important role in pathogenic mechanism of diseases. TFmiR is a web server to collect the coregulatory networks of disease-specific TFs and miRNAs [40]. It integrates genome-wide transcriptional and posttranscriptional regulatory interactions on human diseases, by covering TF-gene, TF-miRNA, miRNA-miRNA, and miRNA gene regulatory networks. In total, TFmiR currently includes the information of almost 10000 genes, 1856 miRNAs, 3000 diseases, and more than 111000 interactions. TFmiR is freely accessible at http://service.bioinformatik.uni-saarland.de/tfmir.

5. Discussion

MiRNA has an important role in expressional mechanism of genes, while miRNA also is transcribed by DNA sequences, which is regulated by some special TFs. As we know, promoter regions control the important initiation process of transcription of genes. Accurate identification of the promoter location is significant for better constructing the regulatory networks and understanding the transcriptional mechanisms. Nowadays, plenty of researchers have focused on the prediction of miRNA promoters and have developed many methods. In this review, we summarized these algorithms by two main types, which is either based on the genome sequence features, or based on the high-throughput sequencing technology. The second types based on NGS data used one or mixed features of histone markers, RNA Pol II binding patterns, and nucleosome-free region. The methods based on genome sequence features have limitation in tissues and species which may lead to lower accuracy in different studies. With the development of NGS technology, more and more sequencing datasets will support the models using histone markers, RNA Pol II binding patterns, and nucleosome-free region. They can further improve the prediction accuracy of miRNA promoters.

Plenty of characteristic features that have been used to predict the promoter regions of miRNAs in methods were discussed in this paper, including expressed sequence tags (EST), TSSs, CpG island, TF binding sites, sequence features (N-mer), conservation, histone modification (especially H3K4me3), expression ditags, poly(A) signal, cap analysis of gene expression (CAGE) tags, familial binding profiles (FBP), nucleosome-depleted regions, and GC content (Table 4). We found that a number of models were built based on histone markers which account for the biggest proportion. It indicates that histone markers are key elements for identification of miRNA promoters, especially H3K4me3 enriched in the promoter regions [41].

It is interesting to see that different features can get common prediction for the human miRNA genes at some level after we compared four typical methods shown in Figure 1. But there is no doubt that methods using different features may result in distinct prediction patterns for miRNA promoters. It should be noticed that most of putative results have not had strong experiment evidence to support and verify. In the future, we can exploit more and more NGS data and use machine learning technology to improve the prediction accuracy by selecting appropriate combination of these features.

Benefited from miRNA promoter predictions, regulatory networks of TFs and miRNAs are being constructed. The TF-mediated miRNA regulation network is valuable to better understand the functional mechanisms of most miRNAs. As we discussed in the paper, some models were built based on the computational methods, which can be modified for

different tissues, species or diseases, by using appropriate datasets. On the other hand, other models were based on experimental methods. They aimed at one specific tissue, species, or disease according to the experimental design. These models are somehow more accurate in the construction of specific regulatory networks. It is worth integrating the computational method and experimental data to further construct dynamic regulatory networks of TFs and miRNAs.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Yuming Zhao designed the project. Yuming Zhao, Fang Wang, and Jun Wan performed the experiments and wrote the manuscript. Su Chen and Guohua Wang revised, read, and approved the final manuscript.

Acknowledgments

This work was supported by the State Key Laboratory of Tree Genetics and Breeding (Northeast Forestry University) (201207), the Fundamental Research Funds for the Central Universities (2572016CB19), and the Natural Science Foundation of China (61371179, 61601110).

References

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

[2] L. He and G. J. Hannon, "MicroRNAs: small RNAs with a big role in gene regulation," *Nature Reviews Genetics*, vol. 5, no. 7, pp. 522–531, 2004.

[3] 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.

[4] L. Weis and D. Reinberg, "Transcription by RNA polymerase II: initiator-directed formation of transcription-competent complexes," *The FASEB Journal*, vol. 6, pp. 3300–3309, 1992.

[5] S. T. Smale and J. T. Kadonaga, "The RNA polymerase II core promoter," *Annual Review of Biochemistry*, vol. 72, pp. 449–479, 2003.

[6] J. Zeng, S. Zhu, and H. Yan, "Towards accurate human promoter recognition: a review of currently used sequence features and classification methods," *Briefings in Bioinformatics*, vol. 10, no. 5, pp. 498–508, 2009.

[7] F. Ozsolak, L. L. Poling, Z. Wang et al., "Chromatin structure analyses identify miRNA promoters," *Genes & Development*, vol. 22, no. 22, pp. 3172–3183, 2008.

[8] X. Wang, Z. Xuan, X. Zhao, Y. Li, and M. Q. Zhang, "High-resolution human core-promoter prediction with CoreBoost-HM," *Genome Research*, vol. 19, no. 2, pp. 266–275, 2009.

[9] A. Marson, S. S. Levine, M. F. Cole et al., "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells," *Cell*, vol. 134, no. 3, pp. 521–533, 2008.

[10] G. Wang, Y. Wang, C. Shen et al., "RNA polymerase II binding patterns reveal genomic regions involved in microRNA gene regulation," *PLoS ONE*, vol. 5, no. 11, Article ID e13798, 2010.

[11] D. L. Corcoran, K. V. Pandit, B. Gordon, A. Bhattacharjee, N. Kaminski, and P. V. Benos, "Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data," *PLoS ONE*, vol. 4, no. 4, Article ID e5279, 2009.

[12] C.-H. Chien, Y.-M. Sun, W.-C. Chang et al., "Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data," *Nucleic Acids Research*, vol. 39, no. 21, pp. 9345–9356, 2011.

[13] J. Gu, T. He, Y. Pei et al., "Primary transcripts and expressions of mammal intergenic microRNAs detected by mapping ESTs to their flanking sequences," *Mammalian Genome*, vol. 17, no. 10, pp. 1033–1041, 2006.

[14] X. Zhou, J. Ruan, G. Wang, and W. Zhang, "Characterization and identification of microRNA core promoters in four model species," *PLoS Computational Biology*, vol. 3, no. 3, pp. 0412–0423, 2007.

[15] S. Fujita and H. Iba, "Putative promoter regions of miRNA genes involved in evolutionarily conserved regulatory systems among vertebrates," *Bioinformatics*, vol. 24, no. 3, pp. 303–308, 2008.

[16] A. Marsico, M. R. Huska, J. Lasserre et al., "PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs," *Genome Biology*, vol. 14, no. 8, article R84, 2013.

[17] M. Bhattacharyya, L. Feuerbach, T. Bhadra, T. Lengauer, and S. Bandyopadhyay, "MicroRNA transcription start site prediction with multi-objective feature selection," *Statistical Applications in Genetics and Molecular Biology*, vol. 11, no. 1, article 6, 2012.

[18] Y. Zhao, F. Wang, and L. Juan, "MicroRNA promoter identification in arabidopsis using multiple histone markers," *BioMed Research International*, vol. 2015, Article ID 861402, 10 pages, 2015.

[19] X. Zhao, H. Zhang, and L. Li, "Identification and analysis of the proximal promoters of microRNA genes in Arabidopsis," *Genomics*, vol. 101, no. 3, pp. 187–194, 2013.

[20] G. Georgakilas, I. S. Vlachos, M. D. Paraskevopoulou et al., "microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs," *Nature Communications*, vol. 5, no. 5, article 5700, 2014.

[21] G. Wang, T. Yu, and W. Zhang, "WordSpy: identifying transcription factor binding motifs by building a dictionary and learning a grammar," *Nucleic Acids Research*, vol. 33, no. W2, pp. W412–W416, 2005.

[22] G. Wang and W. Zhang, "A steganalysis-based approach to comprehensive identification and characterization of functional regulatory elements," *Genome biology*, vol. 7, no. 6, pp. 1–16, 2006.

[23] Y. Lee, M. Kim, J. Han et al., "MicroRNA genes are transcribed by RNA polymerase II," *The EMBO Journal*, vol. 23, no. 20, pp. 4051–4060, 2004.

[24] 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. 10A, pp. 1902–1910, 2004.

[25] G. M. Borchert, W. Lanier, and B. L. Davidson, "RNA polymerase III transcribes human microRNAs," *Nature Structural & Molecular Biology*, vol. 13, no. 12, pp. 1097–1101, 2006.

[26] B. Rhead, D. Karolchik, R. M. Kuhn et al., "The UCSC genome browser database: update 2010," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D613–D619, 2010.

[27] M. Bhattacharyya, M. Das, and S. Bandyopadhyay, "miRT: a database of validated transcription start sites of human MicroRNAs," *Genomics, Proteomics and Bioinformatics*, vol. 10, no. 5, pp. 310–316, 2012.

[28] P. Alexiou, T. Vergoulis, M. Gleditzsch et al., "miRGen 2.0: a database of microRNA genomic information and regulation," *Nucleic Acids Research*, vol. 38, Article ID gkp888, pp. D137–D141, 2010.

[29] P. Landgraf, M. Rusu, R. Sheridan et al., "A mammalian microRNA expression atlas based on small RNA library sequencing," *Cell*, vol. 129, no. 7, pp. 1401–1414, 2007.

[30] G. Georgakilas, I. S. Vlachos, K. Zagganas et al., "DIANA-miRGen v3.0: accurate characterization of microRNA promoters and their regulators," *Nucleic Acids Research*, vol. 44, no. D1, pp. D190–D195, 2016.

[31] C.-H. Chien, Y.-F. Chiang-Hsieh, Y.-A. Chen et al., "AtmiRNET: a web-based resource for reconstructing regulatory networks of Arabidopsis microRNAs," *Database*, vol. 2015, Article ID bav042, 2015.

[32] G. Wang, Y. Wang, M. Teng, D. Zhang, L. Li, and Y. Liu, "Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon γ-stimulated HeLa cells," *PLoS ONE*, vol. 5, no. 7, Article ID e11794, 2010.

[33] C.-H. Chien, Y.-F. Chiang-Hsieh, A.-P. Tsou, S.-L. Weng, W.-C. Chang, and H.-D. Huang, "Large-scale investigation of human TF-miRNA relations based on coexpression profiles," *BioMed Research International*, vol. 2014, Article ID 623078, 8 pages, 2014.

[34] A. E. Kel, E. Gossling, I. Reuter et al., "MATCH: a tool for searching transcription factor binding sites in DNA sequences," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3576–3579, 2003.

[35] E. Falcone, L. Grandoni, F. Garibaldi et al., "Infinity: an in-silico tool for genome-wide prediction of specific DNA matrices in miRNA genomic loci," *PLoS ONE*, vol. 11, no. 4, Article ID e0153658, 2016.

[36] N. J. Martinez, M. C. Ow, M. I. Barrasa et al., "A C. elegans genome-scale microRNA network contains composite feedback motifs with high flux capacity," *Genes and Development*, vol. 22, no. 18, pp. 2535–2549, 2008.

[37] S. Lall, D. Grün, A. Krek et al., "A genome-wide map of conserved MicroRNA targets in C. elegans," *Current Biology*, vol. 16, no. 5, pp. 460–471, 2006.

[38] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, "miRBase: microRNA sequences, targets and gene nomenclature," *Nucleic Acids Research*, vol. 34, pp. D140–D144, 2006.

[39] Z. Guo, M. Maki, R. Ding, Y. Yang, B. Zhang, and L. Xiong, "Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues," *Scientific Reports*, vol. 4, no. 22, article 5150, 2014.

[40] M. Hamed, C. Spaniol, M. Nazarieh, and V. Helms, "TFmiR: a web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks," *Nucleic Acids Research*, vol. 43, no. W1, pp. W283–W288, 2015.

[41] T. S. Mikkelsen, M. Ku, D. B. Jaffe et al., "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells," *Nature*, vol. 448, no. 7153, pp. 553–560, 2007.

Hindawi
BioMed Research International
Volume 2017, Article ID 6375059, 8 pages
https://doi.org/10.1155/2017/6375059

Research Article

CNNdel: Calling Structural Variations on Low Coverage Data Based on Convolutional Neural Networks

Jing Wang, Cheng Ling, and Jingyang Gao

Department of Computer Science and Technology, Beijing University of Chemical Technology, Beijing, China

Correspondence should be addressed to Cheng Ling; lingcheng@buct.edu.cn and Jingyang Gao; gaojy@mail.buct.edu.cn

Received 29 December 2016; Revised 3 April 2017; Accepted 12 April 2017; Published 28 May 2017

Academic Editor: Jialiang Yang

Many structural variations (SVs) detection methods have been proposed due to the popularization of next-generation sequencing (NGS). These SV calling methods use different SV-property-dependent features; however, they all suffer from poor accuracy when running on low coverage sequences. The union of results from these tools achieves fairly high sensitivity but still produces low accuracy on low coverage sequence data. That is, these methods contain many false positives. In this paper, we present CNNdel, an approach for calling deletions from paired-end reads. CNNdel gathers SV candidates reported by multiple tools and then extracts features from aligned BAM files at the positions of candidates. With labeled feature-expressed candidates as a training set, CNNdel trains convolutional neural networks (CNNs) to distinguish true unlabeled candidates from false ones. Results show that CNNdel works well with NGS reads from 26 low coverage genomes of the 1000 Genomes Project. The paper demonstrates that convolutional neural networks can automatically assign the priority of SV features and reduce the false positives efficaciously.

1. Introduction

Genomic structural variation, usually longer than 50 bp [1], is one of the most important types of genetic mutations, which potentially leads to severe diseases, cancers, and even death by breaking the structure of chromosomes. For example, the deletions in ADAM17 are linked to inflammatory skin and bowel diseases [2]. Lee et al. have shown that a variety of prenatally diagnosed congenital heart diseases are related to 22q11.2 deletions [3].

NGS [4] parallelizes the sequencing process and produces massive short reads within 400 bp, which are aligned to the reference sequence by reads mappers like Burrows-Wheeler Aligner (BWA) [5] and Bowtie2 [6]. The alignments of reads are often stored in SAM or BAM format devised by SAMtools [7]. The data mapping step filters anomalously mapped reads, which are direct evidence of SVs.

Most existing SV callers are classified into four categories [8]: (1) discordantly mapping read pairs (i.e., two reads in a pair cross the SV region, and the distance between them is inconsistent with the insert size); (2) split reads: split reads are subdivided into soft-clip reads (i.e., one of the paired reads is partially mapped) and one-end-anchored reads (i.e., one of the paired reads is unmapped); (3) read depth (i.e., the number of reads covering a region); (4) local contig assembly (i.e., assemble reads to form longer consensus sequences, which are called contigs, and then remap them to the reference genome). Many NGS-based SV detection methods have been proposed based on these four theories. These SV detection methods vary in both accuracy and sensitivity, since they utilize different properties to assess the likelihood of SVs.

Each method has its own advantages on the judgement standards of SVs. Take deletion as an example, which is the most common mutation in structural variation [9]. Pindel [10] concentrates on one-end-anchored reads. It performs poorly under low coverage. BreakDancer [11] compares insert size and the separation distance between discordant paired reads to ascertain breakpoints. SVseq2 [12] and DELLY [13] are hybrid approaches to call SVs. SVseq2 applies an enhanced split-reads mapping algorithm to identify deletions and filters the candidates with discordant read analyses. DELLY, on the contrary, uses discordant reads to find candidate SVs and then verifies the exact breakpoints by split-reads alignments. Unsatisfactorily, all these tools produce low accuracy and sensitivity on low coverage sequence datasets.

MetaSV [14], a recently proposed method, combines the results derived from many direct SV calling tools and verifies the candidates using local assembly to reduce false positives rate. Such integrated SV callers still suffer from low accuracy despite relatively high sensitivity. It is worth learning that MetaSV places higher weight to more accurate split-reads methods than discordant paired reads methods.

In this paper, we introduce a SV caller named CNNdel. CNNdel utilizes a convolutional neural network model to accomplish the false positives filter procedure. Compared with other integrated methods, CNNdel is capable of automatically assigning the weights of SV features by neuron networks and the detection accuracy on low coverage real data greatly outstrips the prior methods.

2. Background

Convolutional neural network (CNN) [15] is a typical supervised deep learning algorithm, which is widely applied in image and video recognition, such as face recognition, license plate recognition, and motion prediction in video. For example, the famous LeNet-5 network is applied to recognize handwritten characters [16].

CNN consists of multiple convolutional layers and pooling layers, following full connected networks as hidden layers and the output layer. Each neuron in convolved layers is connected to a small region of the previous layer. Convolution operation is executed with the input of the small region and a filter. The products are summed up as the value of the current neuron. Each convolved layer contains a set of feature maps. Each map has its own filter or kernel. Pooling is a form of nonlinear downsampling. For example, in max-pooling, the input matrix can be divided into nonoverlapping small regions, and for each small region, the layer outputs the maximum. Similarly, in average-pooling or mean-pooling, the layer outputs the average values of each small region.

There are many popular deep learning software frameworks such as Caffe [17], Theano [18], TensorFlow [19], and Torch [20]. The paper [21] gives a detailed presentation about these frameworks. The latest neural network library Keras [22] has attracted wide attention. Taking Theano or TensorFlow as backend, Keras models minimalist and highly modular networks. Other than frameworks that support many kinds of deep architectures, Keras is designed for convolutional networks and recurrent networks. In this paper, Keras is chosen to model a CNN classifier. To further confirm the performance of CNNdel, parameters of the CNN classifier are regulated.

3. Method

In this paper, we focus on the calling of deletions. CNNdel is not a direct SV caller like Pindel or SVseq2; it collects the results from other tools. The pipeline of CNNdel can be generalized to a 4-stage process: (1) get the union of candidates derived from four prior tools by merging duplications; (2) extract features of each candidate; (3) label each candidate by checking the SV benchmark file; and (4) supervised by the labels, use a major part of candidates to train the CNN model and validate the trained model with the remaining candidates. Figure 1 illustrates the framework of CNNdel.

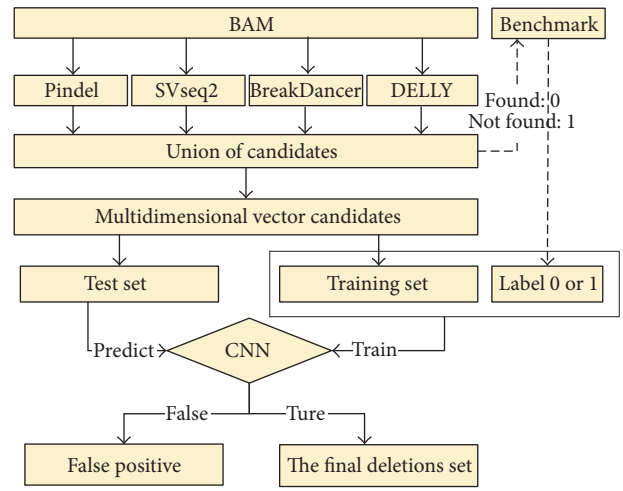


Figure 1: CNNdel pipeline. The pipeline is generalized to 4 steps: (1) get the union of candidates resulting from four tools; (2) get the feature information of each candidate; (3) label each candidate; and (4) use labeled candidates to train the CNNs and validate the trained model.

3.1. Get the Union of Candidates. In order to get as many candidates as possible, Pindel, SVseq2, BreakDancer, and DELLY are run with default parameters. When the distance between two deletions is less than 2% of the shorter deletion length, they are considered as duplications. Learning from MetaSV, when a candidate is assigned with different bounds in the merging process, the bounds given by split-reads methods are more trustworthy.

3.2. Check the Feature Information of Candidates. By checking the features which distinguish deletions from the normal sequence regions, we transform each candidate deletion into a multi-dimensional vector. Five major feature types are specified as:

(i) Feature (1) (deletion length): split-reads mapping reacts badly on overlong deletions. Longer deletions are likely to have different reads distributions with shorter deletions. It is essential to add the length of a deletion as a feature.

(ii) Feature (2~9) (consistency of mapped read pairs): discordant mapped read pair is one of the most direct lines of evidence to support the existence of a deletion. For the discordant and concordant mapped read pairs, refinement works are demanded. Both are, respectively, subdivided into two branches: (i) read mapping error (i.e., note whether the mapped reads are error-free or with mismatches, since the reads mapper BWA is designed to allow mismatches) and (ii) read mapping uniqueness (i.e., note whether a read is uniquely mapped or can be mapped to multiple positions).

Table 1: List of features to call deletions.

Feature types	Amounts
Deletion length	1
Consistency of mapped read pairs	8
Split reads analysis	24
Read depth	4
Mapping reads statistics	12

(iii) Feature (10~33) (split-reads analysis): the reads overlapping the breakpoints of the deletions can be classified into three sorts: (i) fully mapped, (ii) soft-clip (the read cannot be mapped as a whole but its prefix or suffix part can be mapped), and (iii) one-end-anchored (one read in a pair can be mapped while the other one is unmapped). These three sorts are, respectively, subdivided into three detailed branches: (iv) breakpoint positions (the reads overlap whether with the left or the right breakpoint), (v) anchor positions (the mapped one in a pair lies whether upstream or downstream of the deletion region), and (vi) reads mapping uniqueness.

(iv) Feature (34~37) (read depth): the depth in deletion regions is close to 0, since few reads can be mapped to the region. Use SAMtools to count the depths of reads within the deletion region, reads upstream of the deletion region, and reads downstream of the deletion region. The depth of a region is defined as $\sum_{i=1}^{l} \text{depth}(i)/l$, in which depth($i$) is the depth of the ith base in the region and "l" is the length of the region. The three counts are normalized to four values between 0 and 1 in preprocessing.

(v) Feature (38~49) (mapping reads statistics): the last feature type counts the depths in and around the deletions. In this type, the eligible reads in the same three regions are counted: (i) reads within the deletion region, (ii) reads upstream of the deletion region, and (iii) reads downstream of the deletion region. These reads are sorted by (iv) reads mapping error and (v) reads mapping uniqueness.

All features are listed in Table 1. Searching in and around a candidate region according to its known chromosome ID, individual ID, and start and end positions, reads which match the above conditions are counted. Before being imported into the CNN model, these 49 features are normalized into decimals between 0 and 1 in preprocessing.

In the application of CNNs on images, the local receptive fields (sliding windows) are geographically relevant to the neighboring fields. CNNs training could fail when shuffling the pixels in images. As shown in Table 1, 49 features are initially ranked according to the five types. In the Results and Discussion, we will explore the impact of the order of features on the performance of CNNdel.

3.3. Label Each Candidate. Search the deletion benchmark files to inspect whether a candidate deletion is in it. Once a deletion is confirmed, it will be labeled as 1 or 0 otherwise. Thereby, the procedure of false positives filtering can be regarded as a supervised binary-classification problem.

3.4. Use Labeled Candidates to Train the CNNs and Validate the Trained Model

(i) Layer structure: the convolved layer is abbreviated to "C." The pooling layer is abbreviated to "P." The networks' structure is usually set as "C1 + P1 + C2 + P2 + ⋯" following flattening hidden layers and an output layer.

(ii) Parameters: the convolutional neural network model is trained in a supervised way, and we optimize the weights of networks by stochastic gradient descent (SGD), which is given as

$$\theta = \theta - \alpha \nabla_\theta J\left(\theta; x^{(i)}; y^{(i)}\right). \quad (1)$$

"θ" is the weights of input features. "α" is the learning rate. The gradient $\nabla_\theta J$ gives the descent direction of weights. In gradient descent, all samples are calculated to decide the gradient, which costs massive time. To solve the problem, SGD learns the gradient on a batch-size number of samples, followed by a next round on other batch-size samples, until all samples run out. This procedure is called one epoch. Grid search method [23] is used to adjust the learning rate and batch. Smaller epoch prevents the classifying quality of CNN, while larger epoch has the risk of overfitting the model. Split a fraction of the training data as a validation set. Train only on the training set and monitor the validation error every few epochs. Early-stop method stops training as soon as the error on the validation set is higher than it was the last time it was checked.

(iii) Activation functions: activation functions are crucial factors in CNNs which bring about nonlinearity into networks. Figure 2 shows the typical activation functions. Hyperbolic tangent (Tanh) function squashes a real-valued number to the range $[-1, 1]$. It can be computed as $\text{Tanh}(x) = (e^x - e^{-x})/(e^x + e^{-x})$. Rectified linear unit (ReLU), defined as $\text{ReLU}(x) = \max(0, x)$, involves simple operations and accelerates the convergence of SGD compared to Tanh. However, ReLU sometimes frustrates the training model, since ReLU could prevent activating a neuron on data again in the weights updating procedure. Softplus function, a smooth approximation of ReLU, has the mathematical form $\text{Softplus}(x) = \log(1 + e^x)$. These rectifiers are called biological activation functions.

(iv) Input: CNNs are frequently applied in image recognition systems, in which the inputs are 2D images. Our samples are 1D text data. Thus, we regard our 1D data as 2D "images." Each 49-feature deletion can be viewed as an image with 1 by 49 "pixels." According to simple cross-validation [24], we randomly split all candidates into a training set and a test set. We use the labeled training set to train the CNN model and

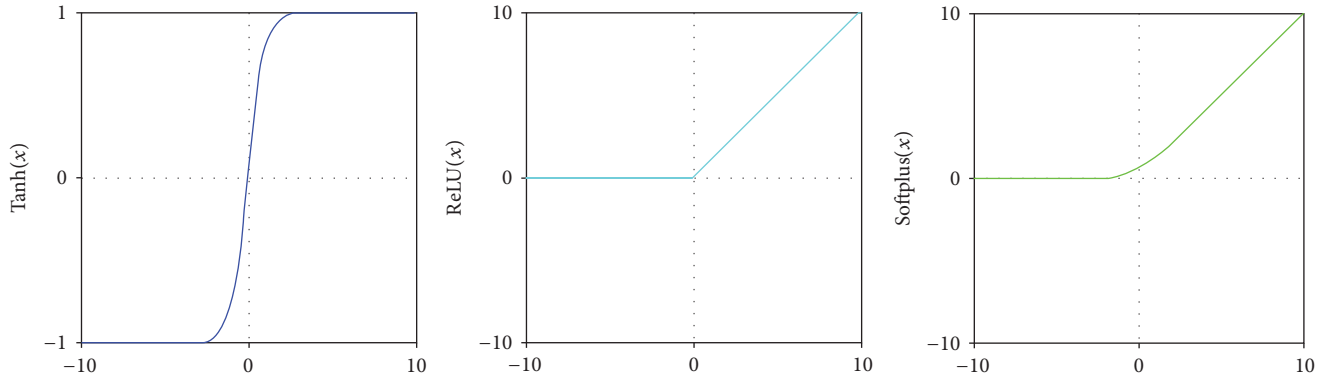


Figure 2: Typical activation functions: Tanh, ReLU, and Softplus.

validate the trained model with the test set. The size of the training set is two times the test set.

4. Results and Discussion

Throughout the following experiments, we first recommend the befitting CNN model by adjusting different parameter settings. Secondly, it is substantiated that the order of 49 features has no crash to the final performance, but shuffling the order of training candidates generates adverse effects instead. Finally, the comparisons between CNNdel and the prior tools and comparisons between CNNdel and SVM are exhibited. Taking both accuracy and sensitivity into account, the parameter F-score is used to evaluate the performance of the CNN model. F-score is specified as "2 × accuracy × sensitivity/(accuracy + sensitivity)."

4.1. Experimental Environment and Dataset. Pindel, SVseq2, BreakDancer, DELLY, and CNNdel are implemented on an Intel(R) Xeon(R) CPU E5-1620 v2 @3.70 GHz, 16 GB RAM, and 1 TB storage with average disk access speed of 164.8 MB/s. Keras runs on Python 2.7 with the backend of Theano.

The raw sequences for the experiments contain 26 samples derived from chromosome 11 and chromosome 20 from human reference hs37d5. All reads are mapped to reference sequences by mapper BWA with default parameters, with BAM files as outputs. And the BAM files are indexed by SAMtools. Benchmark files are released by 1000 Genomes Project Phase III [25]. The mean insert size and mean read length are 425 bp (range: 237–579 bp) and 79 bp, respectively. As a low coverage dataset, the average depth covers 10.6x.

Figure 3 shows the length distribution of deletion datasets. There are a total of 2138 deletions in 26 samples. Copy number variations (CNVs) [26], defined as insertions or deletions that extend to 1 kilobase (kb) or above, occupy about 30% of total SVs. Medium-length deletions take the biggest share.

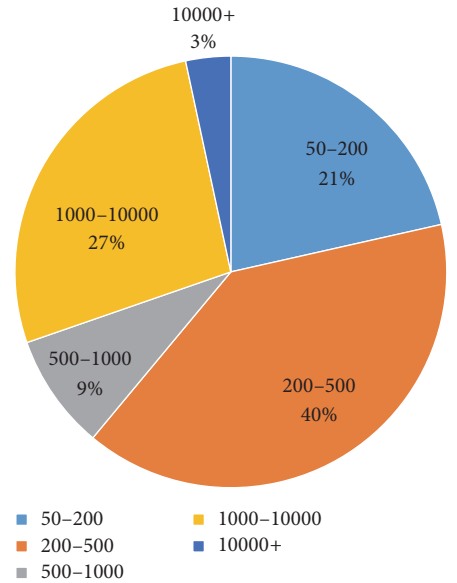


Figure 3: Length distribution of deletion datasets.

4.2. CNN Model Adjustment

4.2.1. Layer Structure. Table 2 records the efficiency and run time of different structures. Learning rate and batch in the beginning are empirically initialized as 0.1 and 64. Tanh is employed as the initial activation function. The results show the following:

(a) The efficiency differs a little in the three kinds of structures.

(b) Structures with fewer layers spend less run time.

The simplest structure "C1 + P1 + F1 + F2" performs best whether in efficiency or in run time. However, fewer layers are

Table 2: Comparisons of different structures.

Layers	A	S	F	Run time
C1 + P1 + F1 + F2	0.7001	0.7081	0.704	23
C1 + P1 + C2 + P2 + F1 + F2	0.6894	0.7069	0.6980	30
C1 + P1 + C2 + P2 + C3 + P3 + F1 + F2	0.6845	0.7124	0.6981	54

"F" means F-score. The unit of run time is seconds. "A" indicates accuracy. "S" indicates sensitivity.

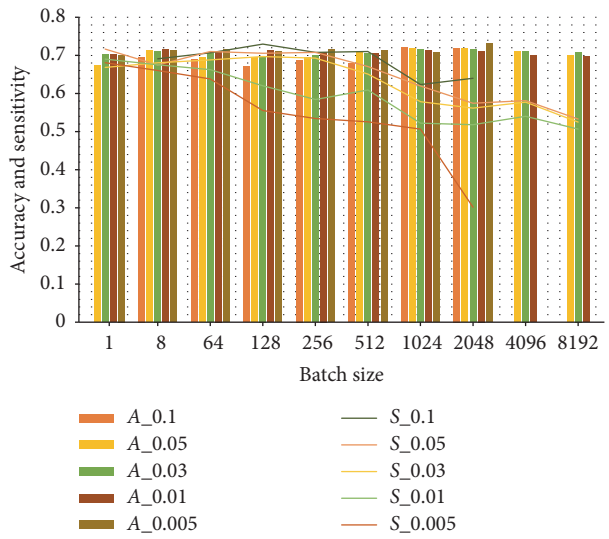


Figure 4: Grid search results of learning rate and batch. "*A*" in the legend means accuracy. "*S*" in the legend means sensitivity. The numbers after "*A*" and "*S*" are learning rates. The horizontal axis shows the batch range. The bar chart shows the accuracy, while the line chart shows the sensitivity.

accompanied by more weights, which increase the burden of memory.

4.2.2. Learning Rate and Batch. In gradient descent or full batch learning, batch size is identical to the number of all training samples. Full batch learning and online learning (batch size is equal to 1) are two extreme situations. The number of training candidates is a little higher than 8000. So, in the beginning, the batch size is between 1 and 8192. As a general rule, learning rate is usually set as 0.1 and then divided by 2 or 5. Thus, learning rates are assigned as 0.1, 0.05, 0.03, 0.01, and 0.005. The initial activation functions are still set as Tanh.

Figure 4 illustrates the grid search results of learning rate and batch. The test results show the following:

(a) The accuracies of these learning rates keep stable around 0.7 when batch size varies.

(b) The sensitivities of learning rates descend with the increase of batch size.

(c) Larger learning rate outperforms smaller ones in sensitivity.

(d) Smaller learning rates always fail to train the model under a huge batch.

Besides, smaller learning rates make the model suffer longer running time. Thus, we conclude that 0.1 is the most appropriate learning rate. In such a case, Table 3 lists the performances of different batch sizes. Too small batch size (such as 1) hinders the convergence of networks, while too large batch size cuts down the times of iterations, leading to longer time to reach a good precision. Getting rid of less accurate batch values, the range 8–512 is appropriate whether in performance or in running time. It is suggested to assign 64 as the batch size since models achieve the highest *F*-score when batch is equal to 64.

Table 3: Comparisons of different batch sizes with learning rate of 0.1.

Batch	*A*	*S*	*F*	Run time
1	0.6641	0.6574	0.6607	31
8	0.7024	0.6906	0.6964	20
64	0.7001	0.7081	**0.704**	30
128	0.6809	0.7213	0.7005	28
256	0.6907	0.7017	0.6962	24
512	0.6941	0.6915	0.6928	30
1124	0.7153	0.6237	0.6664	70
2048	0.7191	0.5988	0.6535	86
4096	0.7122	0.53	0.6077	107
8192	0.701	0.4817	0.571	113

"*F*" means *F*-score. The unit of run time is seconds. "*A*" indicates accuracy. "*S*" indicates sensitivity.

4.2.3. Activation Function. As shown in Figure 5, Tanh, ReLU, and Softplus are applied in CNNs in turn, with learning rate of 0.1 and batch assigned as 64. Each kind of model is run for 10 rounds to verify the stability of the model. The average run times are recorded as 10.2 s, 22.9 s, and 33.9 s when the models are applied with Tanh, ReLU, and Softplus, respectively. It can be concluded that, on accuracy and sensitivity, successfully trained models with Tanh and ReLU do not have significant differences. Softplus makes performance parameters abide violent fluctuation. Besides, models with ReLU and Softplus functions often die during training because they can prevent a neuron from being activated again. Thus, Tanh function stands out for its stability and decent performance.

Learning rate is confirmed as 0.1 for its speed advantage. To further confirm the reliability of the other recommended parameter settings, Table 4 displays the comparisons between combinations of layers, learning rate, activation functions, and batch sizes. The performances of the models are mainly influenced by activation functions. Networks employing Tanh functions always can achieve high accuracy and sensitivity with batch size in such a large scale (8–512).

Model mortality means the frequency with which a model fails during training. According to the highest *F*-score, the learning rate is suggested to be 0.1. It is advocated to use Tanh as activation functions and maintain "C1 + P1 + F1 + F2" as a hierarchical structure if the equipment can satisfy the memory requirements. As to the batch, we recommend smaller ones such as 8 and 64.

4.2.4. Other Tricks about the CNN Model. Other tricks that are beneficial to the models are listed:

(a) Max-pooling or mean-pooling: mean-pooling costs longer time and receives similar results to max-pooling. We suggest using max-pooling in the model.

(b) Filter size and stride size: smaller filter (e.g., we use "1 × 4" as the filter size) and small strides (e.g., 1) help improve the accuracy of the CNN model.

(c) Regularization: dropout, a simple regularization technique, is applied to prevent overfitting. Dropout rate is tested from 0.5 to 0.1, and through testing 0.3 is suggested.

Figure 5: The performances of CNN models which are applied with Tanh, ReLU, and Softplus, respectively.

Table 4: Comparisons of different model parameters.

Layers	Activation	Learning rate	Batch range	Run time	Mortality	Space requirements
C1 + P1 + F1 + F2	Tanh	0.1	8~512	Medium	Hardly	Large
	ReLU	0.1	8~512	Fast	Medium	
	Softplus	0.1	1, 8, 64	Slow	High	
C1 + P1 + C2 + P2 + F1 + F2	Tanh	0.1	8~512	Medium	Hardly	Medium
	ReLU	0.1	8~512	Fast	Medium	
	Softplus	0.1	1, 8	Slow	High	
C1 + P1 + C2 + P2 + C3 + P3 + F1 + F2	Tanh	0.1	8~512	Medium	Hardly	Small
	ReLU	0.1	8, 64	Faster	Medium	
	Softplus	0.1	1, 8	Slow	High	

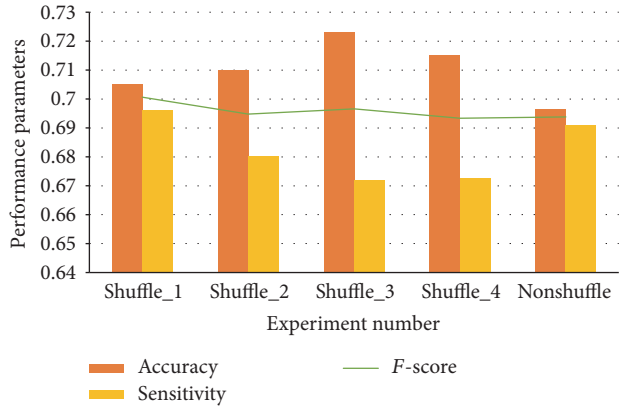


Figure 6: Impact of shuffling features.

4.3. Shuffling

4.3.1. Features Shuffling. With the experiments already shown above, the learning rate and batch are set as 0.1 and 64. 49 features are randomly shuffled. Results in Figure 6 certify that the accuracies and sensitivities of the successful trails have a little difference with the nonshuffled one.

Most detailed subclasses in the five main types are opposite, such as "mapped without error" and "mapped with mismatches" and "mapped uniquely" and "mapped to multiple positions." Most features are biologically independent of each other. Thus, shuffle has a little effect on the efficiency of models.

4.3.2. Candidates Shuffling. It is hazardous to shuffle the 49-dimensional training candidates. Ten rounds of operating results shown in Figure 7 authenticate this conjecture. In such a case, the CNN model faces frequent frustrating results.

In the preprocessing stage, candidates derived from prior tools line up in the order of coordinates. Man-made translocations happen if two deletions switch their positions. There are features like "breakpoint positions" which are related to the relative positions. According to the experimental results shown in Figure 7, candidates without shuffle are recommended.

4.4. Comparisons with the Prior Tools. 10 rounds of simple cross-validation are carried out to insure the reliability of CNNdel. In each round, gather the candidates of 9 individuals on chromosome 11 and chromosome 20 (70% of the candidates, totally 18 files) as the training set and the remaining as

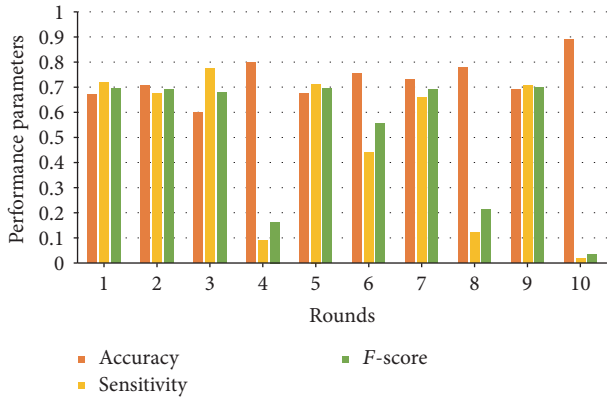


Figure 7: Impact of shuffling candidates.

Table 5: Accuracy, sensitivity, and F-score of initial tools and CNNdel.

Tools	Accuracy	Sensitivity	F-score
Pindel	0.3957	0.5315	0.4537
SVseq2	0.5573	0.5813	0.5690
BreakDancer	0.3831	0.388	0.3855
DELLY	0.4064	0.448	0.4262
Union	0.4329	**0.7906**	0.5595
CNNdel	**0.6894**	0.7069	**0.698**

test set (30% of the candidates, 8 files). The average accuracy, sensitivity, and F-score of the 10 rounds are reported. Table 5 compares the effectiveness of CNNdel with the prior tools. For Pindel, the parameters are set as "-w 0.1 -x 5." SVseq2 is run with cutoff values 3. And other tools are run with default parameters in order to get as many candidates as possible.

With handcrafted features capturing reads distribution, CNNdel outperforms all prior methods in both accuracy and sensitivity. In comparison to the union results of the tools, CNNdel removes plenty of likely false positives and achieves a higher accuracy. However, it is possible for CNNdel to misjudge nondeletion candidates as deletion, which forces the emergence of false negatives. Thus, the sensitivity suffers a little decline compared to sensitivity of the union of tools. CNNdel largely preserves the sensitivity (mean loss of 8.4%) of the test set.

Table 6 denotes the accuracies of CNNdel and the prior tools in different deletion length ranges. SVseq2 outperforms the other tools for deletions in the length of 500 bp–1000 bp. BreakDancer and DELLY operate well on CNVs within the deletion length scope of 10000 bp. Despite a minute gap with SVseq2 on deletions of 500 bp–1000 bp, CNNdel outstrips these tools from a general view, especially on deletions longer than 10000 bp.

4.5. Comparisons with SVM. Both CNNs and SVM can perform well on false positive filtering with similar F-scores. The primary dissimilarity is that CNNdel exports stable performance all along while SVM deeply relies on the parameters and waves violently when the parameters are adjusted. Therefore, it requires a considerably long run time for SVM on grid search to adjust the parameters for the sake of better results.

In SVM, the penalty factor represents the tolerance to classification error. Radial basis function (RBF), one of the commonest kernel functions, is defined as function (2), in which σ stands for the width argument. Grid searches of SVM are carried out on the penalty factor (denoted by "c") and the width argument (denoted by "g"):

$$k(\|x - x_c\|) = \exp\left(-\frac{\|x - x_c\|^2}{\sigma^2}\right). \quad (2)$$

Table 7 chooses the results of SVM when it performs best on accuracy, sensitivity, and F-score correspondingly.

CNN tends to outperform SVM significantly in accuracy when SVM receives better sensitivity. After repeated trials, SVM acquires a similar set of performance parameters.

5. Conclusion

In this paper, we propose a CNN-based method to call deletions on low coverage real data. CNNdel pipeline first collects the union of candidates derived from Pindel, BreakDancer, SVseq2, and DELLY and then finds all features of candidates. Afterwards, CNNdel searches the SV benchmark to get the labels of candidates. Finally, CNNdel trains the CNN model with labeled feature-presented candidates and filters the false positives out.

Based on the above experiments, in order to achieve better accuracy, CNNdel should be optimized by adjusting CNN model parameter settings, especially the activation functions. As a matter of fact, CNN model achieves stable accuracy and sensitivity when the structure, parameters, and activation functions vary in appropriate ranges. The impact of the order of features is also discussed. Experiments show that randomly shuffling the 49 features has a little influence on the performance of CNN. On the contrary, shuffling the order of training candidates causes severe damage to the results. The experimental results show that CNNdel outperforms other tools on low coverage real data. Not only CNN model, but also other nonlinear classification models such as SVM can remove the false positives, though it needs complex parameter regulations.

Efforts will be made to incorporate more strategies of SV detection to extract more cogent features. And CNNdel will be improved by modeling better deep learning networks. Besides, extensive experiments on genomes from patients will be conducted to realize higher clinical application value.

Disclosure

CNNdel is implemented in Python and is available at https://github.com/salarmacata/CNNdel.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Table 6: Comparison with prior tools on different deletion lengths.

Deletion length	Deletions number	Pindel	SVseq2	BreakDancer	DELLY	CNNdel
50–200	147	0.2791	0.3233	0.2222	0	**0.53**
200–500	275	0.6484	0.6738	0.3347	0.3191	**0.8**
500–1000	65	0.7576	**0.8611**	0.4920	0.5658	0.72
1000–10000	172	0.575	0.6606	0.7203	0.7181	**0.88**
10000+	24	0.09	0.375	0.5	0.2857	**0.72**

Table 7: Accuracy, sensitivity, and F-score of initial tools and CNNdel.

Tools	Accuracy	Sensitivity	F-score
SVM (c: 1 g: 3)	**0.8236**	0.4202	0.5565
SVM (c: 32 g: 0.1)	0.6901	0.7054	**0.6977**
SVM (c: 32 g: 0.01)	0.6152	**0.7789**	0.6874
CNNdel	0.6894	0.7069	**0.698**

"c" stands for penalty factor. "g" stands for σ in RBF.

Acknowledgments

The research was supported by a grant from the National Natural Science Foundation of China (no. 61472026). Cheng Ling was supported by a grant from the National Natural Science Foundation of China (no. 60602026).

References

[1] P. Stankiewicz and J. R. Lupski, "Structural variation in the human genome and its role in disease," *Annual Review of Medicine*, vol. 61, pp. 437–455, 2010.

[2] D. C. Blaydon, P. Biancheri, W.-L. Di et al., "Inflammatory skin and bowel disease linked to ADAM17 deletion," *New England Journal of Medicine*, vol. 365, no. 16, pp. 1502–1508, 2011.

[3] M. Y. Lee, H. S. Won, J. W. Baek et al., "Variety of prenatally diagnosed congenital heart disease in 22q11. 2 deletion syndrome," *Obstetrics & Gynecology Science*, vol. 57, no. 1, pp. 11–16, 2014.

[4] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nature Biotechnology*, vol. 26, no. 10, pp. 1135–1145, 2008.

[5] B. W. A. Whittlesea, "Illusions of familiarity," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 19, no. 6, pp. 1235–1253, 1993.

[6] H. Li, B. Handsaker, A. Wysoker et al., "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078-2079, 2009.

[7] C. Alkan, B. P. Coe, and E. E. Eichler, "Genome structural variation discovery and genotyping," *Nature Reviews Genetics*, vol. 12, no. 5, pp. 363–376, 2011.

[8] P. Guan and W.-K. Sung, "Structural variation detection using next-generation sequencing data: a comparative technical review," *Methods*, vol. 102, pp. 36–49, 2016.

[9] E. E. Eichler, D. A. Nickerson, D. Altshuler et al., "Completing the map of human genetic variation," *Nature*, vol. 447, no. 7141, pp. 161–165, 2007.

[10] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads," *Bioinformatics*, vol. 25, no. 21, pp. 2865–2871, 2009.

[11] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat, "Sensitive and accurate detection of copy number variants using read depth of coverage," *Genome Research*, vol. 19, no. 9, pp. 1586–1592, 2009.

[12] J. Zhang, J. Wang, and Y. Wu, "An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data," *BMC Bioinformatics*, vol. 13, no. 6, p. 1, 2012.

[13] T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel, "DELLY: Structural variant discovery by integrated paired-end and split-read analysis," *Bioinformatics*, vol. 28, no. 18, pp. i333-i339, 2012.

[14] M. Mohiyuddin, J. C. Mu, J. Li et al., "MetaSV: an accurate and integrative structural-variant caller for next generation sequencing," *Bioinformatics*, vol. 31, no. 16, pp. 2741–2744, 2015.

[15] K. Fukushima, "Analysis of the process of visual pattern recognition by the neocognitron," *Neural Networks*, vol. 2, no. 6, pp. 413–420, 1989.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

[17] Y. Jia, E. Shelhamer, J. Donahue et al., "Caffe: convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*, pp. 675–678, ACM, Orlando, Fla, USA, November 2014.

[18] B. James, B. Olivier, B. Frédéric et al., "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference*, 2010.

[19] A. Martín, A. Ashish, B. Paul et al., "TensorFlow: large-scale machine learning on heterogeneous distributed systems," Preliminary White Paper, 2015.

[20] R. Collobert, S. Bengio, and J. Marithoz, *Torch: A Modular Machine Learning Software Library*, Idiap, 2002.

[21] B. Soheil, R. Naveen, S. Lukas et al., "Comparative study of deep learning software frameworks," arXiv:1511.06435.

[22] C. François, https://keras.io/.

[23] J. A. A. Brito, F. E. McNeill, C. E. Webber, and D. R. Chettle, "Grid search: an innovative method for the estimation of the rates of lead exchange between body compartments," *Journal of Environmental Monitoring*, vol. 7, no. 3, pp. 241–247, 2005.

[24] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, Springer, NY, USA, 2000.

[25] 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[26] A. J. Iafrate, L. Feuk, M. N. Rivera et al., "Detection of large-scale variation in the human genome," *Nature Genetics*, vol. 36, no. 9, pp. 949–951, 2004.

Hindawi
BioMed Research International
Volume 2017, Article ID 8347085, 12 pages
https://doi.org/10.1155/2017/8347085

Research Article

Identification of Transcriptional Modules and Key Genes in Chickens Infected with *Salmonella enterica* Serovar Pullorum Using Integrated Coexpression Analyses

Bao-Hong Liu[1,2] and Jian-Ping Cai[1,2]

[1]*State Key Laboratory of Veterinary Etiological Biology, Key Laboratory of Veterinary Parasitology of Gansu Province, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Lanzhou, Gansu, China*
[2]*Jiangsu Co-Innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Yangzhou, China*

Correspondence should be addressed to Jian-Ping Cai; caijianping@caas.cn

Received 14 November 2016; Revised 1 March 2017; Accepted 27 March 2017; Published 26 April 2017

Academic Editor: Ansgar Poetsch

Salmonella enterica Pullorum is one of the leading causes of mortality in poultry. Understanding the molecular response in chickens in response to the infection by *S. enterica* is important in revealing the mechanisms of pathogenesis and disease progress. There have been studies on identifying genes associated with *Salmonella* infection by differential expression analysis, but the relationships among regulated genes have not been investigated. In this study, we employed weighted gene coexpression network analysis (WGCNA) and differential coexpression analysis (DCEA) to identify coexpression modules by exploring microarray data derived from chicken splenic tissues in response to the *S. enterica* infection. A total of 19 modules from 13,538 genes were associated with the Jak-STAT signaling pathway, the extracellular matrix, cytoskeleton organization, the regulation of the actin cytoskeleton, G-protein coupled receptor activity, Toll-like receptor signaling pathways, and immune system processes; among them, 14 differentially coexpressed modules (DCMs) and 2,856 differentially coexpressed genes (DCGs) were identified. The global expression of module genes between infected and uninfected chickens showed slight differences but considerable changes for global coexpression. Furthermore, DCGs were consistently linked to the hubs of the modules. These results will help prioritize candidate genes for future studies of *Salmonella* infection.

1. Introduction

Chickens are an important component in the global agricultural economy by serving as one of the primary sources of proteins for humans. However, the poultry industry has been consistently threatened by various diseases, including those caused by viral, bacterial, and parasitic infections. *Salmonella enterica* serovar Pullorum (*S.* Pullorum) is one of the most important pathogens of poultry causing severe systemic disease [1, 2]. To prevent and control *S.* Pullorum in chickens, the host responses against this pathogen have been studied for decades. Although significant advances have been made, especially in the identification of molecules and genes involved in the host immune response [3, 4] and mucosal inflammation [5, 6], as well as their differential expression during infection [7–10], the precise pathways regulating immunity to *Salmonella* infection using a systems biology approach have not been investigated. Although gene differential expression analysis (DEA) provides important information, such as identification of genes that are expressed at different times during infection, which inform our understanding of pathogenesis, identifying gene interactions using a systems biology approach greatly enhances our knowledge at the mechanistic and regulatory levels. A large amount of information regarding gene interactions is available in microarray datasets and by applying network approaches the gap between individual genes and systems can be bridged [11–13]. The modularity in biological systems allows for both the study of independent components and identification of gene relationships within modules. Modern approaches, such as weighted gene coexpression network analysis (WGCNA) [14], can identify modules with expression levels that are

highly correlated across samples and have been used to identify new candidate regulatory molecules and networks in *Salmonella*-infected pigs [15]. Differentially coexpressed modules (DCMs) can also be identified [16]. The holistic changes in modules would be reflected in transcriptional and coexpression changes for individual genes. In general, gene expression levels change during disease or infection, but some have reported that seemingly nonsignificant DEGs may also play a key role in a disease because their interactions with other genes change considerably [17]. These genes can be identified via differential coexpression analysis (DCEA), which can mine individual genes using a holistic approach [17–19]. Hence, combining the WGCNA and DCEA methods can identify interacting modules and differentially coexpressed genes (DCG) during infection, compared with controls. Here, we mined the molecular network relationships of the differential coexpression modules and genes using microarray data from spleens of *S.* Pullorum-infected and uninfected chickens using WGCNA and DCEA (Figure 1). The results complement traditional DEA and add to our understanding of the regulatory mechanisms that occur during *Salmonella* infection.

2. Materials and Methods

2.1. Microarray Data Harvesting and Processing. A comprehensive transcriptomics dataset derived from microarray analysis of spleens from chickens challenged with 10^8 CFU of *Salmonella enterica* serovar Pullorum or mock-challenged with the same volume of distilled water (controls) was obtained from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) (accession number: GSE59663). The dataset was generated with the Agilent oligo microarray chips containing 43,663 probe sets. In this study, we first streamlined the dataset by excluding 14,920 probe sets that were either unmappable to any gene IDs or mapped to multiple gene IDs. In the case of multiple probe sets mapped to one identical gene, the probe set, which is most often associated with the highest expression level, was maintained to ensure that only one probe set was left to investigate one gene. If more than one probe set was left after the above steps, their intensities were averaged. Finally, a one-to-one match between 13,538 probe sets and 13,538 genes was achieved.

Three biological replicates (chips) for each time point were available in the challenged group for these datasets. However, at each time point in the control group, only one chip was used to hybridize with the equally mixed mRNA sample containing the three control samples. We averaged the replicates for each time point, except at day 21, with the two replicates included and forming the dataset for the challenged group with 10 samples; this dataset was equivalent to the dataset of the control group. The dataset was quantile normalized by the function of normalizeQuantiles in R package limma [20].

2.2. Construction of Weighted Gene Coexpression Network and Identification of Modules. Weighted gene coexpression network analysis (WGCNA) was used to detect coexpression modules from the dataset of challenged samples [14, 21]. The R function of blockwise modules was implemented with the following parameters: power = 12, minModuleSize = 100, and networkType = "signed." Microarray data were processed as described below.

The pairwise Pearson's correlation coefficients were calculated for all the genes in the challenged groups, followed by the construction of an adjacency matrix using the power function:

$$\alpha_{ij} = \left(0.5 + 0.5 \times \mathrm{cor}\left(x_i, x_j\right)\right)^{\beta}, \quad (1)$$

where x_i and x_j were the ith and jth gene expression traits, respectively, which formed a signed weighted correlation network; and β used default value (i.e., $\beta = 12$). The topological overlap measure (TOM) was calculated as follows:

$$\mathrm{TOM}_{ij} = \frac{\sum_{u \neq i,j} \alpha_{iu}\alpha_{uj} + \alpha_{ij}}{\min\left(k.\mathrm{total}_i, k.\mathrm{total}_j\right) + 1 - \alpha_{ij}}, \quad (2)$$

where k.total is the sum of connection strengths for a gene with the other network genes. u is the other network genes.

Afterwards, 1-TOM was calculated as a biological important measure for network interconnectedness. Genes with highly similar coexpression relationships were grouped together by performing hierarchical clustering on the topological overlap. Subsequently, genes were hierarchically clustered using 1-TOM as the distance measure and modules were determined by choosing a height cutoff of 0.995 for the resulting dendrogram. Highly similar modules were identified by clustering and merged together using a dynamic tree-cutting algorithm [14]. Eigengene refers to the first principal component for a given module and could be calculated to draw a module trajectory curve [14].

2.3. Identification of Differentially Coexpressed Modules. Differentially coexpressed modules (DCMs) were identified using gene-set coexpression analysis (GSCA) that adopted the length-normalized Euclidean distance to measure the coexpression difference for the pairwise correlations between infected and control groups [16].

$$D_m = \sqrt{\frac{1}{P_m}\sum_{p=1}^{P_m}\left(r_p^c - r_p^i\right)^2}, \quad (3)$$

where P_m was the number of gene pairs from the pairwise correlation for all the module genes. r_p^c and r_p^i were the correlation coefficients for a gene pair in the control and infected groups, respectively.

The null distribution for distance was constructed by permuting samples across conditions for 10,000 times to yield gene-set specific p values. Modules with p value < 0.01 were considered as significantly differentially coexpressed.

2.4. Identification of Differentially Coexpressed Genes. The differential coexpression analysis (DCEA) was implemented by using R package DCGL, which is a useful tool to identify

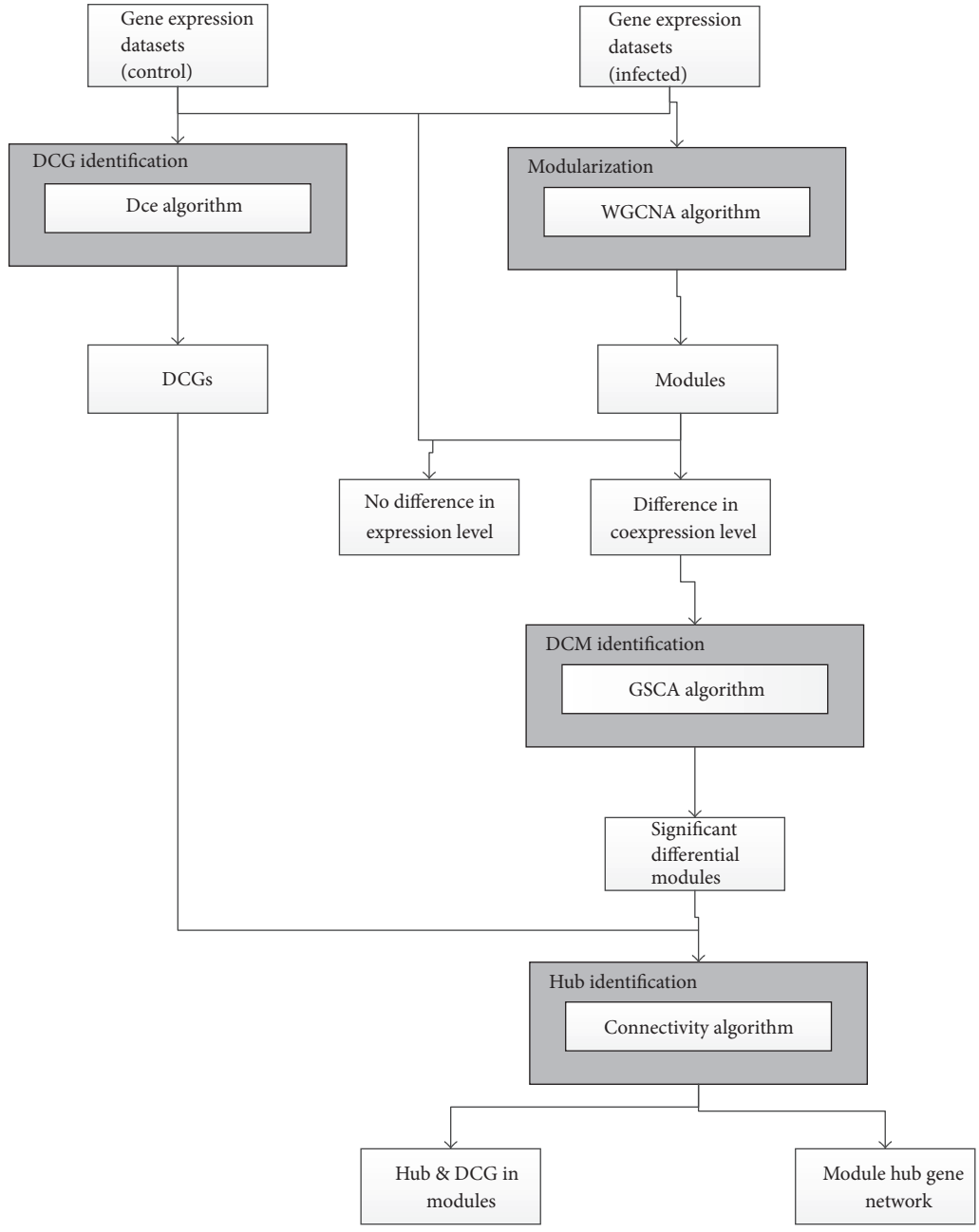


Figure 1: Workflow of the comprehensive gene coexpression network analysis.

differentially coexpressed genes (DCGs) and differentially coexpressed links (DCLs) [17–19]. The R function DCe was applied and then the p values were adjusted for a false discovery rate (FDR) using the Benjamini-Hochberg method to reduce a large amount of false positive results [22]. The genes with FDR < 0.001 were selected as DCGs.

2.5. Gene Ontology (GO) and Pathway Enrichment for Coexpression Modules. GO enrichment and KEGG pathway analyses for network modules were performed using Database for Annotation, Visualization, and Integrated Discovery (DAVID, v6.7) program using all chickens genes as the background [23, 24]. The modified Fisher's exact test with an adjustment for multiple tests by Benjamini-Hochberg method was used to identify significantly enriched terms for module genes [22].

2.6. Network Visualization. The complex network bioinformatics software Cytoscape (v3.1.1) was used to visualize the pairwise relationships between genes [25].

3. Results

3.1. Weighted Gene Coexpression Network Analysis. Using blockwiseModules R function ($\beta = 12$), a total of 19 modules ranging from 100 to 3,000 genes were recovered for the 13,538

Table 1: Module preservation and functions.

Module	Size	Z summary	Function
Lightyellow	126	5.42	Nucleus ($8.40E-4$)
Lightgreen	145	0.67	Jak-STAT signaling pathway ($9.73E-3$)
Grey60	145	7.49	Extracellular matrix ($5.10E-4$) Cytoskeleton ($7.84E-3$)
Lightcyan	155	0.52	Anchored to membrane ($4.19E-3$)
Midnightblue	159	5.78	Cytoplasm ($1.75E-4$); Organelle membrane ($4.67E-3$) Endomembrane system ($9.27E-3$)
Cyan	181	8.71	Cell adhesion molecules ($3.24E-5$) Cell adhesion ($2.75E-3$)
Salmon	248	2.14	Neuroactive ligand-receptor interaction ($2.60E-6$)
Tan	287	4.86	Lysosome ($4.80E-3$)
Greenyellow	293	6.58	Ligase activity ($3.41E-3$)
Purple	298	10.93	Proteasome complex ($6.56E-6$) Regulation of cytokine biosynthetic process ($3.83E-3$) Toll-like receptor signaling pathway ($9.52E-3$)
Magenta	357	1.82	G-Protein coupled receptor activity ($2.88E-4$)
Pink	418	13.36	Cellular macromolecular complex assembly ($1.00E-03$)
Black	599	1.97	Postsynaptic membrane ($2.35E-3$) Synapse ($2.72E-3$)
Red	648	14.89	Signal transducer activity ($2.07E-4$) Multicellular organism development ($2.70E-4$)
Green	1056	27.63	Cell cycle phase ($9.63E-13$) DNA replication ($3.67E-7$) Response to DNA damage stimulus ($1.45E-6$) DNA repair ($4.11e-6$) Cytoskeleton organization ($3.77E-4$)
Yellow	1122	32.83	Glucose catabolic process ($4.03E-5$) Glycolysis/gluconeogenesis ($1.49E-4$) Glycolysis ($1.79E-4$) Glucose metabolic process ($3.95E-4$)
Brown	1349	18.99	ABC transporters ($1.00E-03$)
Blue	2581	43.01	Immune system process ($1.31E-4$) Induction of apoptosis ($1.98E-4$) Antigen processing and presentation ($2.06E-4$) Lysosome ($3.08E-4$) Defense response to bacterium ($6.06E-3$)
Turquoise	2998	51.56	Nervous system development ($2.20E-15$) Focal adhesion ($2.04E-9$) Wnt signaling pathway ($7.56E-9$) Regulation of actin cytoskeleton ($1.64E-7$) TGF-beta signaling pathway ($4.42E-7$)

Note. The column "Size" gives the gene numbers contained in every module. "Z summary" gives the z score of module preservation. "Function" gives the module functions enriched by DAVID.

distinct genes in the *S*. Pullorum-infected group (Table 1). Each module was assigned a unique color, including gray color for the 373 unassigned genes. Genes in the same module shared the same or similar expression patterns that were catalogued by the trajectory curves (Figure 2).

Subsequent analysis using DAVID identified biological features in modules that were potentially associated with the infection by *S*. Pullorum (Figure 6 and Table 1), such as the Jak-STAT signaling pathway (module lightgreen) [26], the extracellular matrix (ECM) (module grey60) [27],

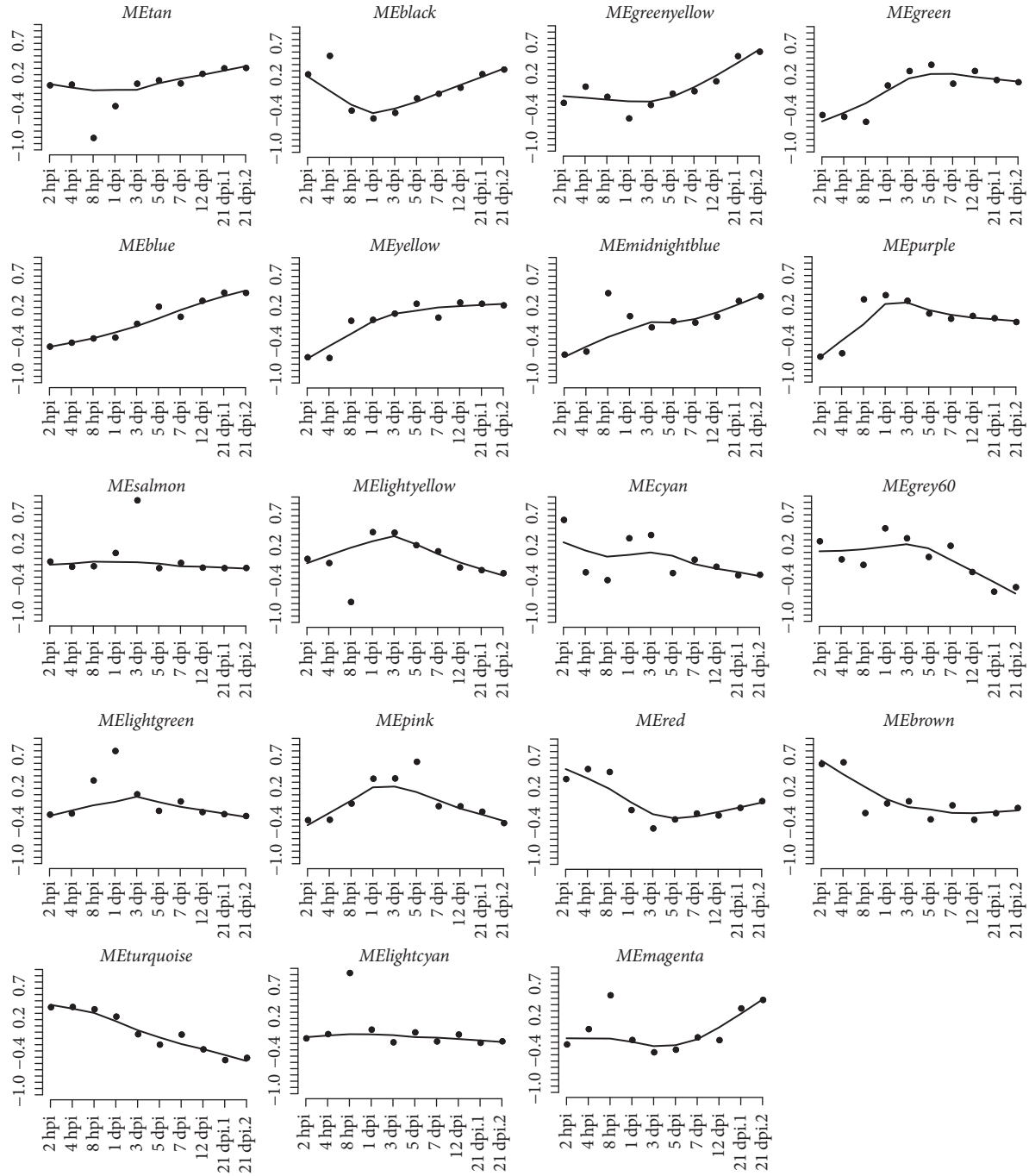


Figure 2: Module expression patterns.

cytoskeleton organization (module green), regulation of the actin cytoskeleton (module blue) [28], G-protein coupled receptor activity (module magenta), Toll-like receptor signaling pathways (module purple), and immune system processes (module blue). ECM genes and cell adhesion genes are significantly enriched in the module grey60 and cyan (FDR = $5.10e-4$ and $2.75e-3$), respectively (Table 1). The grey60 and cyan modules also displayed significant similarity in expression patterns (eigengenes' correlation = 0.76; $p = 0.01$). These observations were in congruent with those reported earlier by others on the crucial role of host cell ECM proteins and bacterial outer membrane structures in the adhesion and invasion of *Salmonella* [27].

3.2. Module Stability. To test the reproducibility of the identified modules, we performed a sampling test, in which we randomly selected half of the samples to calculate the new intramodule connectivity. The sampling was repeated 100 times and then the module stability was expressed as the correlation of intramodule connectivity between the original and sampled ones [29]. Most modules displayed good stability; module salmon was the least stable (Figure 3).

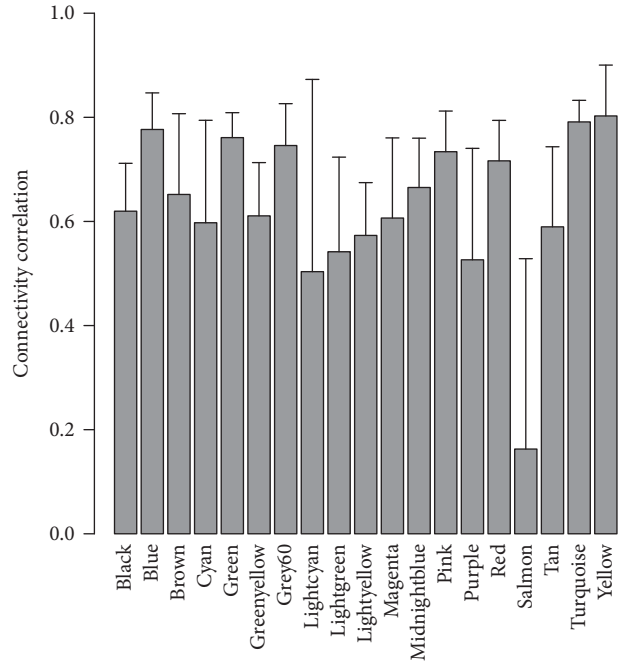


Figure 3: Correlation of intramodule connectivity for each module after 100 samplings.

3.3. Module Preservation Analysis. We investigated whether the *S.* Pullorum-infected module was preserved in the corresponding controls by testing whether the infection-associated coexpression network can be replicated in the control groups. The preservation scores for all the modules were listed in Table 1, in which Z summary scores <2, between 2 and 10, and >10 indicate no evidence, weak-to-moderate evidence, and strong evidence for module preservation, respectively. Preservation analysis provided strong evidence to support the conservation of modules turquoise, brown, blue, yellow, green, red, pink, and purple, which all contained considerably large numbers of genes, but no evidence to support the preservation of modules lightcyan, lightgreen, magenta, and black associated with the membranes, the Jak-STAT signaling pathway, G-protein coupled receptor activity, and synapses, respectively (Table 1).

3.4. Module Gene Expression and Coexpression Comparison. We compared the module genes' expression and coexpression level between the infected and control groups. The violin plot in Figure 4(a) showed that the gene expression for modules in the infection versus control groups is not significantly different, and the distribution for the expression intensities is similar. Subsequently, we compared the gene coexpression level by calculating the gene connectivity for each module. The module turquoise exhibits the largest connectivities since it includes the largest number of genes (2,998 genes). Modules blue (2,581 genes), yellow (1,122 genes), brown (1,349 genes), and green (1,056 genes), which include a considerable number of genes, display the next highest connectivities. In addition, the coexpression levels are different between modules in the two conditions. The coexpressions are strengthened in the infected state (Figure 4(b)).

3.5. Identification of Differentially Coexpressed Modules. Gene-set coexpression analysis (GSCA) revealed that 14 of the 19 modules were significantly differentially coexpressed ($p < 0.01$ by bootstrap sampling test) (Table 2). Among them, modules black ($z = 1.97$), magenta ($z = 1.82$), salmon ($z = 2.14$), and lightcyan ($z = 0.52$) were significantly differentially coexpressed. These observations were in agreement with the module preservation analysis, in which significantly differentially coexpressed modules (DCM) were only weakly preserved in the control group.

3.6. Identification of Differentially Coexpressed Genes. A total of 2,856 differentially coexpressed genes (DCG) were selected with a false discovery rate (FDR) of less than 0.001 using the DCe method in the DCGL package. And a total of 284,213 differentially coexpressed links (DCLs) were same signed, 82,619 were differently signed, and 272,491 were switched links.

Furthermore, we mapped the DCGs for each module and found that the DCMs enrich the DCGs. For example, a total of 152 DCGs appeared in the module magenta with Z summary of 1.82 ($p = 0$), 231 DCGs in module black with Z summary of 1.97 ($p = 0$), and 147 DCGs in module salmon with Z summary of 2.14 ($p = 0$). In network biology, a hub gene is a good representative of a module. We identified the hub genes for all of the modules. Table 2 gives the gene names which are not only hub genes but also DCGs in each module.

4. Discussion

We constructed a gene network for the *S.* Pullorum-infected chickens using weighted gene coexpression network analysis (WGCNA) from the data of time-series microarray. This module detection strategy utilizes the biological variability inherent in the prospective cohort study to reveal the modular organization and function of transcriptional systems. The time series expression profiles allow the study of the transcriptional regulation of these gene coexpression networks during infection. A network-based analysis provides a systems-level understanding of the relationships between members of a network by focusing on genome-wide gene modules rather than individual genes [30]. Differential expression analysis (DEA) aims to identify genes that are expressed significantly higher or lower in one group compared with another. By contrast, WGCNA is not biased toward genes with significant changes in expression. Moreover, the dimensionality of microarray data in the present study was reduced from 13,538 genes to 19 modules, which significantly increased the ability to identify concordant changes in the expression of multiple genes.

The module expression analysis showed that module salmon was the least abundant but exhibited the largest variation in gene expression causing the instability in module construction. Gene expression was the most stable within module midnightblue (Figure 4(a)). The expression distribution for module genes in different conditions (infected versus control) was the same. The coexpression level was further compared. The coexpression comparison showed significant changes for different conditions. We investigated

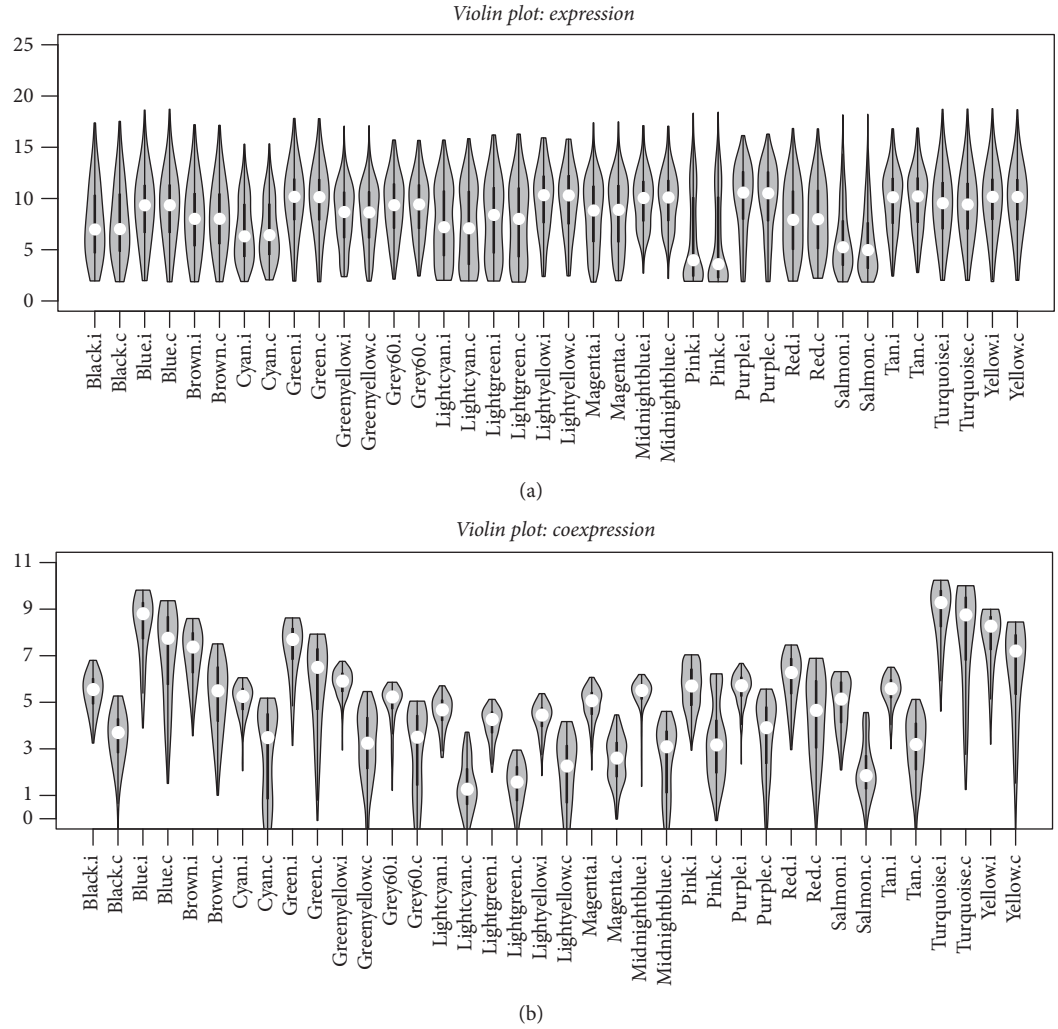


Figure 4: (a) Violin plot showing the gene expression differences between modules in the infected and control groups. (b) Violin plot showing the gene coexpression connectivity differences between modules in the infected and control groups. .i represents the module in the infected group, and .c represents the module in the control group.

the key modules and genes resulting in the differences. The GSCA analysis identified ten significantly differentially coexpressed modules (DCMs), which are in accordance with the module preservation results that are significant in coexpression differences with little evidence for preservation. Regulatory relationships among genes can be parsed as the pairwise correlations between gene expression levels, so the changes in coexpression patterns between two conditions may indicate dysfunctional regulatory systems in disease [31]. Module lightgreen associated with the Jak-STAT signaling pathway and lightcyan associated with membrane anchoring functions were the two weakest preserved modules (Table 1 and Figure 5). Thus, these modules may be associated with *S.* Pullorum infection in chickens.

Furthermore, we investigated the driven genes leading to the coexpression difference. The differential coexpression analysis (DCEA) method was applied, and 2,856 differentially coexpressed genes (DCGs) were identified. Compared to the differential expression analysis (DEA), it was found that the overlapping of DCGs with the 234 DEGs (t-test p value less than 0.01) was significant (hypergeometric test $p = 1.07e-07$), indicating that differential expression and differential coexpression are somewhat related to each other, which is consistent with a previous report [17]. However, there are many *Salmonella* infection-related genes identified by the DCEA method. The top one DCG identified is *WASF1*, which is an important gene in the *Salmonella* infection pathway and was not identified as a DEG (expression fold change: 1.08; t-test p value of 0.34). The protein encoded by *WASF1*, a member of the Wiskott-Aldrich syndrome protein family, plays a critical role downstream of *Rac*, which is a Rho family of small GTPases, in regulating the actin cytoskeleton required for membrane ruffling. This gene associates with an actin nucleation core Arp2/3 complex while enhancing actin polymerization in vitro [32]. Another gene, *CDC42* (fold change = 1.02; $p = 0.6$), a member of the Rho subfamily of actin-organizing small GTP-binding proteins, interacts with *WASF1* and is essential for *S.* Typhimurium entry into host

Table 2: Differentially coexpressed modules enriched with differentially coexpressed genes (DCGs).

Module	Size	GSCA.p	Hub and DCGs
Black	599	0	GJA8, LOC421988, MYOCD
Magenta	357	0	C10orf83, PPME1, NRTN, TMEM167B, ABCA5, C10orf58
Salmon	248	0	CPA5, PRSS2, ARHGEF19, INS, CELA2A, TCERG1L, LOC396296, LOC771434, KCNC2, XKR9
Lightcyan	155	$5.55E-16$	LOC769741, TRPA1, C10orf96, CASKIN2, GMCSF, cor6
Brown	1349	$2.11E-09$	—
Cyan	181	$1.77E-08$	RAB36, LOC415324, UPK3B, CWH43
Lightgreen	145	$4.05E-07$	CIP1, ZC3H12D, TNS3, SOCS3, LOC415844
Pink	418	$4.20E-07$	ABCG8
Lightyellow	126	$2.10E-05$	DEPDC1, CDCA2
Tan	287	$2.23E-04$	TP53INP1, CNRIP1, NEURL, SEPP1
Midnightblue	159	$1.16E-02$	LOC416257
Greenyellow	293	$2.04E-01$	INVS, ARSH
Grey60	145	$2.10E-01$	—
Purple	298	$2.51E-01$	WDR5, NIPA2
Red	648	$8.66E-01$	—
Blue	2581	1	—
Green	1056	1	—
Turquoise	2998	1	—
Yellow	1122	1	—

Note. The column "Size" gives the gene numbers contained in every module. "GSCA.p" gives the p value calculated by GSCA for modules between the two different conditions and the p value less than 0.05 indicates that the module is significantly differentially coexpressed. "Hub and DCGs" shows the genes which are DCGs in the top ten hub genes.

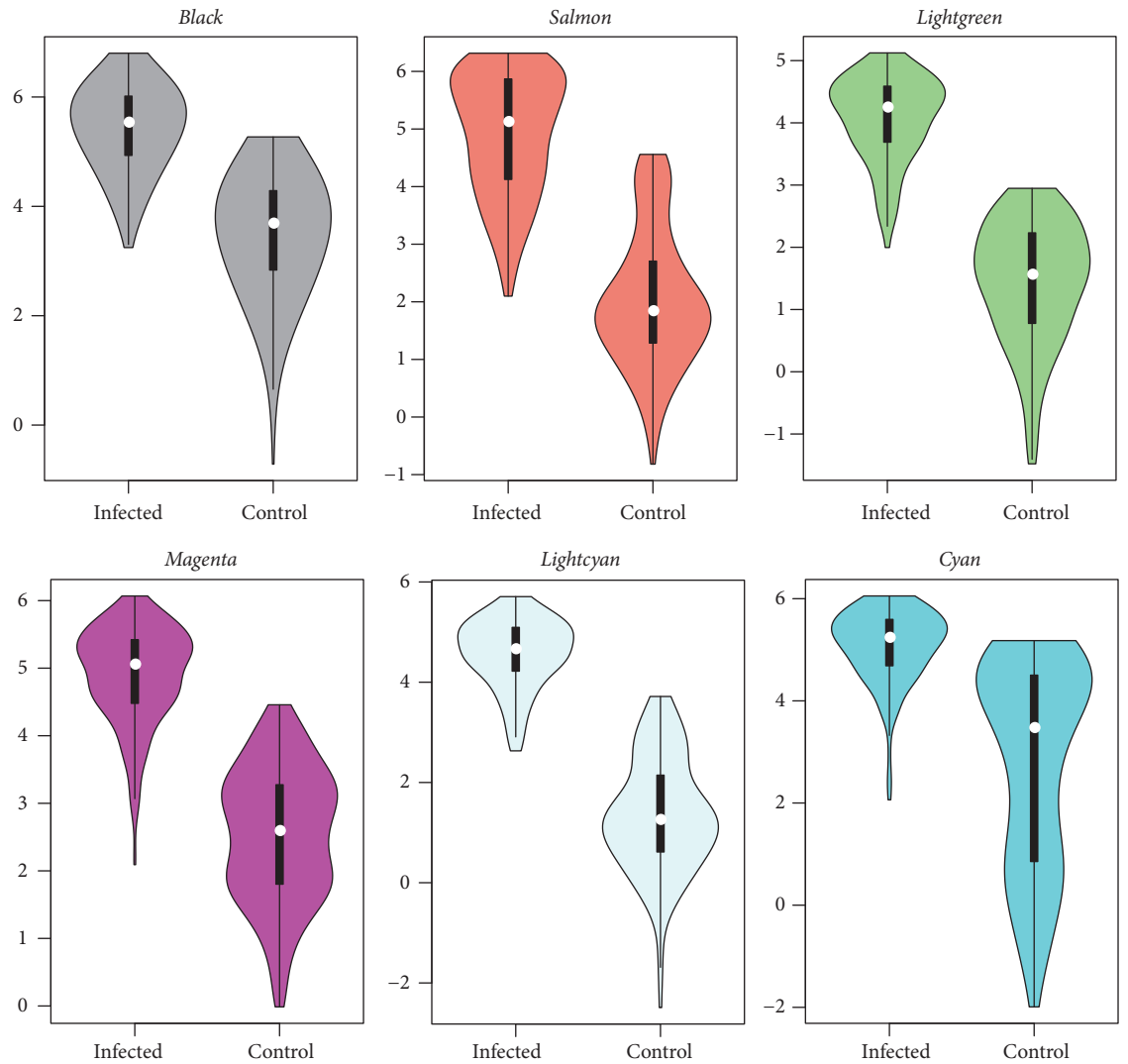


Figure 5: Violin plot of gene coexpression connectivity for significantly differentially coexpressed modules.

cells [33, 34]. *CDC42* was not selected as DCG with a false discovery rate (FDR) of $1.55E-3$, but it interacts with genes *PAK7* (fold change = 1.95; $p = 0.64$) [35, 36], *CDC42EP3* (fold change = 1.05; $p = 0.81$) [37, 38], *PAK1* (fold change = 0.97; $p = 0.67$) [39], *PARD6B* (fold change = 0.76; $p = 0.05$) [40, 41], *PARD6A* (fold change = 1.75; $p = 0.90$) [29, 40], and *IQGAP2* (fold change = 1.58; $p = 0.12$) [42], which are all identified as DCGs. Carow and Rottenberg reported that gene *SOCS3,* which was also identified as a DCG (fold change = 1.54; $p = 0.15$), is a major regulator of infection and inflammation and controls immune homeostasis in physiological and pathological conditions such as infection and autoimmunity [43]. *SOCS3* is a hub gene in the module lightgreen associated with the Jak-STAT signaling pathway, an important pathway for *Salmonella* infection [44]. It is well known that the Jak-STAT pathway can regulate cell growth, apoptosis, immunity, and inflammatory responses and because of its significance in the immune response, the Jak-STAT pathway is often exploited by pathogens [45]. In our study, we found that the Jak-STAT pathway genes were significantly enriched in the module lightgreen which is not detectable in controls. So we think that *SOCS3* and the other Jak-STAT pathway genes may together regulate the activity of the organism in infection, which leads this module to be differentially coexpressed.

The above results showed some specific subnetworks for infection, in spite of a common network existing whether in the control or infected group. We constructed two coexpression networks from the top ten hub genes' expression profiles for each module from the two different conditions. As shown in Figures 6(a) and 6(b), common core networks, including the most preserved modules between infected and control groups, were present. However, some closely interacted subnetworks seen during infection disappeared in the control. These infection-specific subnetworks included genes that are members of the Jak-STAT signaling pathway (module lightgreen); others associated with membrane anchoring (module lightcyan), neuroactive ligand-receptor interaction (module salmon), and lysosomal processing (module tan), which suggested that these subnetworks dysregulated the

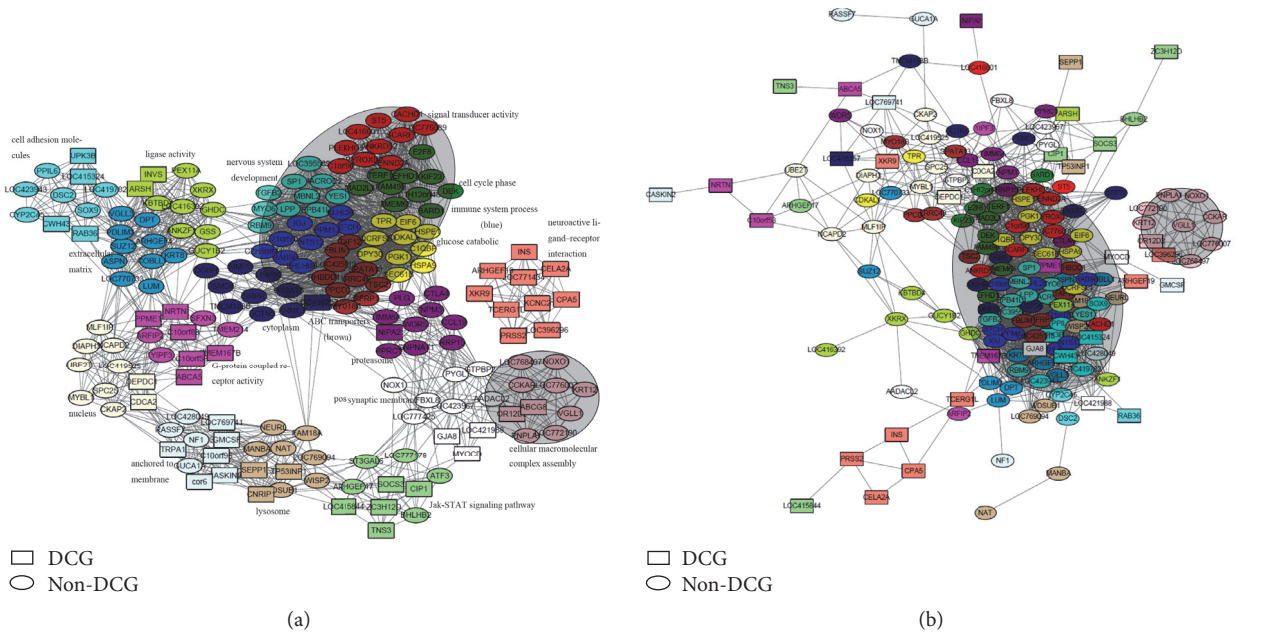


Figure 6: (a) Hub genes' network in the infected group. The node colors represent the module colors: the nodes with shape of rectangle are DCGs and the elliptical nodes are non-DCGs. (b) Hub genes' network in the control group. The node colors represent the module colors: the nodes with shape of rectangle are DCGs and the elliptical nodes are non-DCGs.

systems during infection. Although only one dataset was used here, due to the lack of published related microarray datasets, these present results advance our understanding of the cell biology and immunoregulatory pathways involved in *Salmonella* infection in the chicken host.

Abbreviations

DAVID:	Database for Annotation, Visualization, and Integrated Discovery
DCEA:	Differential coexpression analysis
DCG:	Differentially coexpressed gene
DCL:	Differentially coexpressed link
DCM:	Differentially coexpressed module
DEA:	Differential expression analysis
ECM:	Extracellular matrix
FDR:	False discovery rate
GO:	Gene Ontology
GSCA:	Gene-set coexpression analysis
TOM:	Topological overlap measure
WGCNA:	Weighted gene coexpression network analysis.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

Bao-Hong Liu contributed to the design and conception of this study, conducted computational experiments, performed and interpreted data, and drafted the manuscript. Jian-Ping Cai conceived this project and participated in its design, helped in interpreting the data, and drafted and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors thank Professor Guan Zhu from Texas A&M University and Dr. Patricia Wilkins from CDC, USA, for revising the manuscript. And they also thank the researchers from Yangzhou University for submitting the dataset to GEO. This work was supported by Fundamental Research Program of Chinese Academy of Agricultural Sciences (0032015027) and the Innovative Special Project of Agricultural Sci-Tech and by the Special Fund for Agro-Scientific Research in Public Interest (201303044-7) to Professor Cai's team.

References

[1] R. T. Khan, M. Chevenon, K. E. Yuki, and D. Malo, "Genetic dissection of the Ity3 locus identifies a role for Ncf2 co-expression modules and suggests Selp as a candidate gene underlying the Ity3.2 locus," *Frontiers in Immunology*, vol. 5, article 375, 2014.

[2] G. C. Buckle, C. L. Walker, and R. E. Black, "Typhoid fever and paratyphoid fever: systematic review to estimate global morbidity and mortality for 2010," *Journal of Global Health*, vol. 2, no. 1, Article ID 010401, 2012.

[3] R. L. Santos, "Pathobiology of Salmonella, intestinal microbiota, and the host innate immune response," *Frontiers in Immunology*, vol. 5, article 252, 2014.

[4] C. Guadarrama, T. Villaseñor, and E. Calva, "The subtleties and contrasts of the LeuO regulator in *Salmonella* Typhi:

implications in the immune response," *Frontiers in Immunology*, vol. 5, article 581, 2014.

[5] S. Patel and B. A. McCormick, "Mucosal inflammatory response to Salmonella typhimurium infection," *Frontiers in Immunology*, vol. 5, article 311, 2014.

[6] J. P. Mooney, B. P. Butler, K. L. Lokken et al., "The mucosal inflammatory response to non-typhoidal Salmonella in the intestine is blunted by IL-10 during concurrent malaria parasite infection," *Mucosal Immunology*, vol. 7, no. 6, pp. 1302–1311, 2014.

[7] M. M. Bellet, E. Deriu, J. Z. Liu et al., "Circadian clock regulates the host response to Salmonella," *Proceedings of the National Academy of Sciences*, vol. 110, no. 24, pp. 9897–9902, 2013.

[8] T.-H. Huang, J. J. Uthe, S. M. D. Bearson et al., "Distinct peripheral blood RNA responses to *Salmonella* in pigs differing in *Salmonella* shedding levels: intersection of IFNG, TLR and miRNA pathways," *PLoS ONE*, vol. 6, no. 12, Article ID e28768, 2011.

[9] A. Chaussé, O. Grépinet, E. Bottreau et al., "Susceptibility to Salmonella carrier-state: a possible Th2 response in susceptible chicks," *Veterinary Immunology and Immunopathology*, vol. 159, no. 1-2, pp. 16–28, 2014.

[10] A. Szmolka, Z. Wiener, M. E. Matulova, K. Varmuzova, and I. Rychlik, "Gene expression profiles of chicken embryo fibroblasts in response to Salmonella Enteritidis infection," *PLoS ONE*, vol. 10, no. 6, Article ID e0127708, 2015.

[11] V. van Noort, B. Snel, and M. A. Huynen, "The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model," *EMBO Reports*, vol. 5, no. 3, pp. 280–284, 2004.

[12] 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.

[13] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, "Coexpresion analysis of human genes across many microarray data sets," *Genome Research*, vol. 14, no. 6, pp. 1085–1094, 2004.

[14] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, article 17, 2005.

[15] A. Kommadath, H. Bao, A. S. Arantes et al., "Gene co-expression network analysis identifies porcine genes associated with variation in *Salmonella* shedding," *BMC Genomics*, vol. 15, no. 1, article 452, 2014.

[16] Y. Choi and C. Kendziorski, "Statistical methods for gene set co-expression analysis," *Bioinformatics*, vol. 25, no. 21, pp. 2780–2786, 2009.

[17] H. Yu, B.-H. Liu, Z.-Q. Ye, C. Li, Y.-X. Li, and Y.-Y. Li, "Link-based quantitative methods to identify differentially coexpressed genes and gene Pairs," *BMC Bioinformatics*, vol. 12, article 315, 2011.

[18] B.-H. Liu, H. Yu, K. Tu, C. Li, Y.-X. Li, and Y.-Y. Li, "DCGL: an R package for identifying differentially coexpressed genes and links from gene expression microarray data," *Bioinformatics*, vol. 26, no. 20, pp. 2637–2638, 2010.

[19] J. Yang, H. Yu, B.-H. Liu et al., "DCGL v2.0: an R package for unveiling differential regulation from differential co-expression," *PLoS ONE*, vol. 8, no. 11, Article ID e79729, 2013.

[20] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, article 3, 2004.

[21] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, article 559, 2008.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] K.-I. Uchiya and T. Nikai, "Salmonella pathogenicity island 2-dependent expression of suppressor of cytokine signaling 3 in macrophages," *Infection and Immunity*, vol. 73, no. 9, pp. 5587–5594, 2005.

[27] A. Berndt, J. Müller, L. Borsi, H. Kosmehl, U. Methner, and A. Berndt, "Reorganisation of the caecal extracellular matrix upon Salmonella infection—relation between bacterial invasiveness and expression of virulence genes," *Veterinary Microbiology*, vol. 133, no. 1-2, pp. 123–137, 2009.

[28] D. G. Guiney and M. Lesnick, "Targeting of the actin cytoskeleton during infection by *Salmonella* strains," *Clinical Immunology*, vol. 114, no. 3, pp. 248–255, 2005.

[29] W. Liu and H. Ye, "Co-expression network analysis identifies transcriptional modules in the mouse liver," *Molecular Genetics and Genomics*, vol. 289, no. 5, pp. 847–853, 2014.

[30] J. A. Miller, M. C. Oldham, and D. H. Geschwind, "A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging," *Journal of Neuroscience*, vol. 28, no. 6, pp. 1410–1420, 2008.

[31] Y. Zhai, L. M. Franco, R. L. Atmar et al., "Host transcriptional response to influenza and other acute respiratory viral infections—a prospective cohort study," *PLoS Pathogens*, vol. 11, no. 6, Article ID e1004869, 2015.

[32] Z. Chen, D. Borek, S. B. Padrick et al., "Structure and control of the actin regulatory WAVE complex," *Nature*, vol. 468, no. 7323, pp. 533–538, 2010.

[33] L.-M. Chen, S. Hobbie, and J. E. Galán, "Requirement of CDC42 for Salmonella-induced cytoskeletal and nuclear responses," *Science*, vol. 274, no. 5295, pp. 2115–2118, 1996.

[34] K. Hallstrom and B. A. McCormick, "*Salmonella* interaction with and passage through the intestinal mucosa: through the lens of the organism," *Frontiers in Microbiology*, vol. 2, p. 88, 2011.

[35] A. Pandey, I. Dan, T. Z. Kristiansen et al., "Cloning and characterization of PAK5, a novel member of mammalian p21-activated kinase-II subfamily that is predominantly expressed in brain," *Oncogene*, vol. 21, no. 24, pp. 3939–3948, 2002.

[36] C. Dan, N. Nath, M. Liberto, and A. Minden, "PAK5, a new brain-specific kinase, promotes neurite outgrowth in N1E-115 cells," *Molecular and Cellular Biology*, vol. 22, no. 2, pp. 567–577, 2002.

[37] G. Joberty, R. R. Perlungher, and I. G. Macara, "The Borgs, a new family of Cdc42 and TC10 GTPase-interacting proteins," *Molecular and Cellular Biology*, vol. 19, no. 10, pp. 6585–6597, 1999.

[38] A. S. Alberts, N. Bouquin, L. H. Johnston, and R. Treisman, "Analysis of RhoA-binding proteins reveals an interaction domain conserved in heterotrimeric G protein β subunits and the yeast response regulator protein Skn7," *Journal of Biological Chemistry*, vol. 273, no. 15, pp. 8616–8622, 1998.

[39] B. Zhang, J. Chernoff, and Y. Zheng, "Interaction of Rac1 with GTPase-activating proteins and putative effectors. A comparison with Cdc42 and RhoA," *The Journal of Biological Chemistry*, vol. 273, no. 15, pp. 8776–8782, 1998.

[40] G. Joberty, C. Petersen, L. Gao, and I. G. Macara, "The cell-polarity protein Par6 links Par3 and atypical protein kinase C to Cdc42," *Nature Cell Biology*, vol. 2, no. 8, pp. 531–539, 2000.

[41] Y. Noda, R. Takeya, S. Ohno, S. Naito, T. Ito, and H. Sumimoto, "Human homologues of the Caenorhabditis elegans cell polarity protein PAR6 as an adaptor that links the small GTPases Rac and Cdc42 to atypical protein kinase C," *Genes to Cells*, vol. 6, no. 2, pp. 107–119, 2001.

[42] S. Brill, S. Li, C. W. Lyman et al., "The Ras GTPase-activating-protein-related human protein IQGAP2 harbors a potential actin binding domain and interacts with calmodulin and Rho family GTPases," *Molecular and Cellular Biology*, vol. 16, no. 9, pp. 4869–4878, 1996.

[43] B. Carow and M. E. Rottenberg, "SOCS3, a major regulator of infection and inflammation," *Frontiers in Immunology*, vol. 5, article 58, 2014.

[44] X. Liu, R. Lu, Y. Xia, S. Wu, and J. Sun, "Eukaryotic signaling pathways targeted by *Salmonella* effector protein AvrA in intestinal infection in vivo," *BMC Microbiology*, vol. 10, p. 326, 2010.

[45] L. Yang and Y.-J. Zhang, "Antagonizing cytokine-mediated JAK-STAT signaling by porcine reproductive and respiratory syndrome virus," *Veterinary Microbiology*, 2016.

Hindawi
BioMed Research International
Volume 2017, Article ID 5073427, 14 pages
https://doi.org/10.1155/2017/5073427

Research Article

Joint $L_{1/2}$-Norm Constraint and Graph-Laplacian PCA Method for Feature Extraction

Chun-Mei Feng,[1] Ying-Lian Gao,[2] Jin-Xing Liu,[1] Juan Wang,[1] Dong-Qin Wang,[1] and Chang-Gang Wen[1]

[1]*School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China*
[2]*Library of Qufu Normal University, Qufu Normal University, Rizhao 276826, China*

Correspondence should be addressed to Ying-Lian Gao; yinliangao@126.com

Received 30 December 2016; Revised 12 February 2017; Accepted 1 March 2017; Published 2 April 2017

Academic Editor: Jialiang Yang

Principal Component Analysis (PCA) as a tool for dimensionality reduction is widely used in many areas. In the area of bioinformatics, each involved variable corresponds to a specific gene. In order to improve the robustness of PCA-based method, this paper proposes a novel graph-Laplacian PCA algorithm by adopting $L_{1/2}$ constraint ($L_{1/2}$ gLPCA) on error function for feature (gene) extraction. The error function based on $L_{1/2}$-norm helps to reduce the influence of outliers and noise. Augmented Lagrange Multipliers (ALM) method is applied to solve the subproblem. This method gets better results in feature extraction than other state-of-the-art PCA-based methods. Extensive experimental results on simulation data and gene expression data sets demonstrate that our method can get higher identification accuracies than others.

1. Introduction

With the rapid development of gene-chip and deep-sequencing technologies, a lot of gene expression data have been generated. It is possible for biologists to monitor the expression of thousands of genes with the maturation of the sequencing technology [1–3]. It is reported that a growing body of research has been used to select the feature genes from gene expression data [4–6]. Feature extraction is a typical application of gene expression data. Cancer has become a threat to human health. Modern medicine has proved all cancers are directly or indirectly related to genes. How to identify what is believed to be related to cancer has become a hotspot in the field of bioinformatics. The major bottleneck of the development of bioinformatics is how to build an effective approach to integrate and analyze the expression data [7].

One striking feature of gene expression data is the case that the number of genes is far greater than the number of samples, commonly called the high-dimension-small-sample-size problem [8]. Typically this means that expression data are always with more than thousands of genes, while the size of samples is generally less than 100. The huge expression data make them hard to analyze, but only a small size of genes can control the gene expression. More attention has been attached to the importance of feature genes by modern biologists. Correspondingly, it is especially important how to discover these genes effectively, so many dimensionality reduction approaches are proposed.

Traditional dimensionality reduction methods have been widely used. For example, Principal Component Analysis (PCA) recombines the original data which have a certain relevance into a new set of independent indicators [9–11]. However, because of the sparsity of gene regulation, the weaknesses of traditional approaches in the field of feature extraction become increasingly evident [12, 13]. With the development of deep-sequencing technique, the inadequacy of conventional methods is emerging. Within the process of feature selection on biological data, the principal components of PCA are dense, which makes it difficult to give an objective and reasonable explanation on the significance of biology. PCA-based methods have achieved good results in the application of feature extraction [3, 12]. Although this method

shows the significance of sparsity in the aspect of handling high dimensional data, there are still a lot of shortcomings in the algorithm.

(1) The high dimensionality of data poses a great challenge to the research, which is called data disaster.

(2) Facing with millions of data points, it is reasonable to consider the internal geometric structure of the data.

(3) Gene expression data usually contain a lot of outliers and noise, but the above methods cannot effectively deal with these problems.

With the development of graph theory [14] and manifold learning theory [15], the embedded structure problem has been effectively resolved. Laplacian embedding as a classical method of manifold learning has been used in machine learning and pattern recognition, whose essential idea is recovery of low dimensional manifold structure from high dimensional sampled data. The performance of feature extraction will be improved remarkably after joining Laplacian in gene expression data. In the case of maintaining the local adjacency relationship of the graph, the graph can be drawn from the high dimensional space to a low dimensional space (drawing graph). However, graph-Laplacian cannot dispose outliers.

In the field of dimensionality reduction, L_p $(0 < p < 1)$-norm was getting more and more popular to replace L_1, which was first proposed by Nie et al. [16]. Research shows that a proper value of p can achieve a more exact result for dimensionality reduction [17]. Furthermore, Xu et al. developed an simple iterative thresholding representation theory for $L_{1/2}$-norm [18], which was similar to the notable iterative soft thresholding algorithm for the solution of L_0 [19] and L_1-norm [20]. Xu et al. have shown that L_p-norm generates more better solution than L_1-norm [21]. Besides, among all regularization with p in $(0, 1/2]$, there is no obvious difference. However, when $p \in [1/2, 1)$, the smaller p is, the more effective result will be [17]. This provides a motivation to introduce $L_{1/2}$-norm constraint into original method. Since the error of each data point is calculated in the form of the square. It will also cause a lot of errors while the data contains some tiny abnormal values.

In order to solve the above problems, we propose a novel method based on $L_{1/2}$-norm constraint, graph-Laplacian PCA ($L_{1/2}$ gLPCA) which provides a good performance. In summary, the main work of this paper is as follows. (1) The error function based on $L_{1/2}$-norm is used to reduce the influence of outliers and noise. (2) Graph-Laplacian is introduced to recover low dimensional manifold structure from high dimensional sampled data.

The remainder of the paper is organized as follows. Section 2 provides some related work. We present our formulation and algorithm for $L_{1/2}$-norm constraint graph-Laplacian PCA in Section 3. We evaluate our algorithm on both simulation data and real gene expression data in Section 4. The correlations between the identified genes and cancer data are also included. The paper is concluded in Section 5.

2. Related Work

2.1. Principal Component Analysis. In the field of bioinformatics, the principal components (PCs) of PCA are used to select feature genes. Assume $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbf{R}^{m \times n}$ is the input data matrix, which contains the collection of n data column vectors and m dimension space. Traditional PCA approaches recombine the original data which have a certain relevance into a new set of independent indicators [9]. More specifically, this method reduces the input data to k-dim ($k < n$) subspace by minimizing:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & \left\| \mathbf{X} - \mathbf{U}\mathbf{V}^T \right\|_F^2 \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{1}$$

where each column of $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_k) \in \mathbf{R}^{m \times k}$ is the principal directions and $\mathbf{V}^T = (\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \mathbf{R}^{k \times n}$ is the projected data points in the new subspace.

2.2. Graph-Laplacian PCA. Since the traditional PCA has not taken into account the intrinsic geometrical structure within input data, the mutual influences among data may be missed during a research project [9]. With the increasing popularity of the manifold learning theory, people are becoming aware that the intrinsic geometrical structure is essential for modeling input data [15]. It is a well-known fact that graph-Laplacian is the fastest approach in the manifold learning method [14]. The essential idea of graph-Laplacian is to recover low dimensional manifold structure from high dimensional sampled data. PCA closely relates to K-means clustering [22]. The principal components V are also the continuous solution of the cluster indicators in the K-means clustering method. Thus, it provides a motivation to embed Laplacian to PCA whose primary purpose is clustering [23, 24]. Let symmetric weight matrix $\mathbf{W} \in \mathbf{R}^{n \times n}$ be the nearest neighbor graph where $\mathbf{W}_{ij}$ is the weight of the edge connecting vertices i and j. The value of $\mathbf{W}_{ij}$ is set as follows:

$$\mathbf{W}_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \in \mathbf{N}_k\left(\mathbf{x}_j\right) \text{ or } \mathbf{x}_j \in \mathbf{N}_k\left(\mathbf{x}_i\right), \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathbf{N}_k(\mathbf{x}_i)$ is the set of k nearest neighbors of $\mathbf{x}_i$. $\mathbf{V}^T = (\mathbf{v}_1, \ldots, \mathbf{v}_n) \in R^{k \times n}$ is supposed as the embedding coordinates of the data and $\mathbf{D} = \operatorname{diag}(\mathbf{d}_1, \ldots, \mathbf{d}_n)$ is defined as a diagonal matrix and $\mathbf{d}_i = \sum_j \mathbf{W}_{ij}$. $\mathbf{V}$ can be obtained by minimizing:

$$\begin{aligned} \min_{\mathbf{V}} \quad & \sum_{i,j=1}^{n} \left\| \mathbf{v}_i - \mathbf{v}_j \right\|^2 \mathbf{W}_{ij} = \operatorname{tr}\left(\mathbf{V}^T(\mathbf{D} - \mathbf{W})\mathbf{V}\right) \\ & = \operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right) \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{3}$$

where $\mathbf{d}_i$ is the column or row sums of $\mathbf{W}$ and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is named as Laplacian matrix. Simply put, in the case of maintaining the local adjacency relationship of the graph, the

graph can be drawn from the high dimensional space to a low dimensional space (drawing graph). In the view of the function of graph-Laplacian, Jiang et al. proposed a model named graph-Laplacian PCA (gLPCA), which incorporates graph structure encoded in $\mathbf{W}$ [23]. This model can be considered as follows:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & J = \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\right\|_F^2 + \alpha \operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right) \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{4}$$

where $\alpha \geq 0$ is a parameter adjusting the contribution of the two parts. This model has three aspects. (a) It is a data representation, where $\mathbf{X} \simeq \mathbf{U}\mathbf{V}^T$. (b) It uses $\mathbf{V}$ to embed manifold learning. (c) This model is a nonconvex problem but has a closed-form solution and can be efficient to work out.

In (4), from the perspective of data point, it can be rewritten as follows:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & J = \sum_{j=1}^{n} \left(\left\|\mathbf{X}_n - \mathbf{U}\mathbf{v}_n{}^T\right\|_F^2 + \alpha \operatorname{tr}\left(\mathbf{v}_n{}^T\mathbf{L}\mathbf{v}_n\right)\right) \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}. \end{aligned} \tag{5}$$

In this formula, the error of each data point is calculated in the form of the square. It will also cause a lot of errors while the data contains some tiny abnormal values. Thus, the author formulates a robust version using $L_{2,1}$-norm as follows:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\right\|_{2,1} + \alpha \operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right) \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{6}$$

but the major contribution of $L_{2,1}$-norm is to generate sparse on rows, in which the effect is not so obvious [3, 25].

3. Proposed Algorithm

Research shows that a proper value of p can achieve a more exact result for dimensionality reduction [17]. When $p \in [1/2, 1)$, the smaller p is, the more effective result will be [17]. Then, Xu et al. developed a simple iterative thresholding representation theory for $L_{1/2}$-norm and obtained the desired results [18]. Thus, motivated by former theory, it is reasonable and necessary to introduce $L_{1/2}$-norm on error function to reduce the impact of outliers on the data. Based on the half thresholding theory, we propose a novel method using $L_{1/2}$-norm on error function by minimizing the following problem:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\right\|_{1/2}^{1/2} + \alpha \operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right) \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{7}$$

where $L_{1/2}$-norm is defined as $\|\mathbf{A}\|_{1/2}^{1/2} = \sum_j^n \sum_j^m |\mathbf{a}_{ij}|^{1/2}$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbf{R}^{m\times n}$ is the input data matrix, and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_k) \in \mathbf{R}^{m\times k}$ and $\mathbf{V}^T = (\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \mathbf{R}^{k\times n}$ are the principal directions and the subspace of projected data, respectively. We call this model graph-Laplacian PCA based on $L_{1/2}$-norm constraint ($L_{1/2}$ gLPCA).

At first, the subproblems are solved by using the Augmented Lagrange Multipliers (ALM) method. Then, an efficient updating algorithm is presented to solve this optimization problem.

3.1. Solving the Subproblems. ALM is used to solve the subproblem. Firstly, an auxiliary variable is introduced to rewrite the formulation (4) as follows:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V},\mathbf{S}} \quad & \|\mathbf{S}\|_{1/2}^{1/2} + \alpha \operatorname{tr} \mathbf{V}^T (\mathbf{D} - \mathbf{W}) \mathbf{V}, \\ \text{s.t.} \quad & \mathbf{S} = \mathbf{X} - \mathbf{U}\mathbf{V}^T, \ \mathbf{V}^T\mathbf{V} = \mathbf{I}. \end{aligned} \tag{8}$$

The augmented Lagrangian function of (8) is defined as follows:

$$\begin{aligned} L_\mu(\mathbf{S}, \mathbf{U}, \mathbf{V}, \mathbf{\Lambda}) = {} & \|\mathbf{S}\|_{1/2}^{1/2} + \operatorname{tr} \mathbf{\Lambda}^T \left(\mathbf{S} - \mathbf{X} + \mathbf{U}\mathbf{V}^T\right) \\ & + \frac{\mu}{2} \left\|\mathbf{S} - \mathbf{X} + \mathbf{U}\mathbf{V}^T\right\|_F^2 \\ & + \alpha \operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right), \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{9}$$

where $\mathbf{\Lambda}$ is Lagrangian multipliers and μ is the step size of update. By mathematical deduction, the function of (9) can be rewritten as

$$\begin{aligned} L_\mu(\mathbf{S}, \mathbf{U}, \mathbf{V}, \mathbf{\Lambda}) = {} & \|\mathbf{S}\|_{1/2}^{1/2} + \frac{\mu}{2} \left\|\mathbf{S} - \mathbf{X} + \mathbf{U}\mathbf{V}^T + \frac{\mathbf{\Lambda}}{\mu}\right\|_F^2 \\ & + \alpha \operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right), \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}. \end{aligned} \tag{10}$$

The general approach of (10) consists of the following iterations:

$$\begin{aligned} \mathbf{S}^{k+1} &= \arg\min_{\mathbf{S}} L_\mu\left(\mathbf{S}, \mathbf{U}^k, \mathbf{V}^k, \mathbf{\Lambda}^k\right), \\ \mathbf{V}^{k+1} &= \left(\mathbf{v}_1, \ldots, \mathbf{v}_k\right), \\ \mathbf{U}^{k+1} &= \mathbf{M}\mathbf{V}^k, \\ \mathbf{\Lambda}^{k+1} &= \mathbf{\Lambda}^k + \mu\left(\mathbf{S}^{k+1} - \mathbf{X} + \mathbf{U}^k\mathbf{V}^{T^k}\right), \\ \mu^{k+1} &= \rho\mu^k. \end{aligned} \tag{11}$$

Then, the details to update each variable in (11) are given as follows.

Updating $\mathbf{S}$. At first, we solve $\mathbf{S}$ while fixing $\mathbf{U}$ and $\mathbf{V}$. The update of $\mathbf{S}$ relates the following issue:

$$\mathbf{S}^{k+1} = \arg\min_{\mathbf{S}} \|\mathbf{S}\|_{1/2}^{1/2} + \frac{\mu}{2} \left\|\mathbf{S} - \mathbf{X} + \mathbf{U}^k\mathbf{V}^{T^k} + \frac{\mathbf{\Lambda}^k}{\mu}\right\|_F^2, \tag{12}$$

which is the proximal operator of $L_{1/2}$-norm. Since this formulation is a nonconvex, nonsmooth, non-Lipschitz, and complex optimization problem; an iterative half thresholding approach is used for fast solution of $L_{1/2}$-norm and summarizes according to the following lemma [18].

Lemma 1. *The proximal operator of $L_{1/2}$-norm minimizes the following problem:*

$$\min_{\mathbf{X}\in\mathbf{R}^{m\times n}} \|\mathbf{X}-\mathbf{A}\|_F^2 + \lambda\|\mathbf{X}\|_{1/2}^{1/2}, \tag{13}$$

which is given by

$$\mathbf{X}^* = \mathbf{H}_\lambda(\mathbf{A}) = \mathbf{U}\operatorname{diag}\left(\mathbf{H}_\lambda(\sigma)\right)\mathbf{V}^T, \tag{14}$$

where $\mathbf{H}_\lambda(\sigma) \coloneqq (h_\lambda(\sigma_1), h_\lambda(\sigma_2), \ldots, h_\lambda(\sigma_n))^T$ and $h_\lambda(\sigma_i)$ is the half threshold operator and defined as follows:

$$h_\lambda(\sigma_i) = \begin{cases} \frac{2}{3}\sigma_i\left(1+\cos\left(\frac{2\pi}{3}-\frac{2}{3}\psi\lambda(\sigma_i)\right)\right), & \text{if } |\sigma_i| > \frac{\sqrt[3]{54}}{4}\lambda^{2/3} \\ 0, & \text{otherwise,} \end{cases} \tag{15}$$

where $\psi\lambda(\sigma_i) = \arccos((\lambda/8)(|\sigma_i|/3)^{-2/3})$.

Solving $\mathbf{U}$ *and* $\mathbf{V}$. Here, we solve $\mathbf{U}$ while fixing others. The update of $\mathbf{U}$ amounts to solving

$$\mathbf{U}^{k+1} = \arg\min_{\mathbf{U}} \frac{\mu}{2}\left\|\mathbf{S}^k - \mathbf{X} + \mathbf{U}^k\mathbf{V}^{T^k} + \frac{\mathbf{\Lambda}^k}{\mu}\right\|_F^2. \tag{16}$$

Letting $\mathbf{X} - \mathbf{S} - \mathbf{\Lambda}/\mu = \mathbf{M}$, (16) becomes $\mathbf{U}^{k+1} = \arg\min_{\mathbf{U}}(\mu/2)\|\mathbf{M} - \mathbf{U}\mathbf{V}^{T^k}\|_F^2$, taking partial derivatives of $\mathbf{U}$ as follows:

$$\frac{\partial J}{\partial \mathbf{U}} = -\mu\left(\mathbf{M} - \mathbf{U}\mathbf{V}^{T^k}\right)\mathbf{V}. \tag{17}$$

Setting the partial derivatives to 0, we have

$$\mathbf{U}^{k+1} = \mathbf{M}\mathbf{V}^k. \tag{18}$$

Then, we solve $\mathbf{V}$ while fixing others. Similarly, letting $\mathbf{X} - \mathbf{S} - \mathbf{\Lambda}/\mu = \mathbf{M}$, $\mathbf{U} = \mathbf{M}\mathbf{V}$, the update of $\mathbf{V}$ can be listed as follows:

$$\begin{aligned} &\mathbf{V}^{k+1} = \arg\min_{\mathbf{V}} \frac{\mu}{2}\left\|\mathbf{M} - \mathbf{M}\mathbf{V}\mathbf{V}^T\right\|_F^2 + \alpha\operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right), \\ &\text{s.t.} \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}. \end{aligned} \tag{19}$$

By some algebra, we have

$$\begin{aligned} \mathbf{V}^{k+1} &= \arg\min_{\mathbf{V}} \left\|\mathbf{M} - \mathbf{M}\mathbf{V}\mathbf{V}^T\right\|_F^2 + \frac{2\alpha}{\mu}\operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right) \\ &= \arg\min_{\mathbf{V}} \operatorname{tr}\left(\mathbf{M}\mathbf{M}^T\right) - 2\left(\sqrt{\operatorname{tr}\left(\mathbf{M}\mathbf{M}^T\right)}\right)^2 \\ &\quad + \frac{2\alpha}{\mu}\operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right) \\ &= \arg\min_{\mathbf{V}} \operatorname{tr}\mathbf{V}^T\left(-\mathbf{M}^T\mathbf{M} + \frac{2\alpha}{\mu}\mathbf{L}\right)\mathbf{V}. \end{aligned} \tag{20}$$

Therefore, (19) can be rewritten as follows:

$$\begin{aligned} &\mathbf{V}^{k+1} = \arg\min_{\mathbf{V}} \operatorname{tr}\mathbf{V}^T\left(-\mathbf{M}^T\mathbf{M} + \frac{2\alpha}{\mu}\mathbf{L}\right)\mathbf{V}, \\ &\text{s.t.} \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}. \end{aligned} \tag{21}$$

Thus, the optimal $\mathbf{V}^{k+1}$ can be obtained by calculating eigenvectors

$$\mathbf{V}^{k+1} = (\mathbf{v}_1, \ldots, \mathbf{v}_k), \tag{22}$$

which corresponds to the first k smallest eigenvalues of the matrix $G_\alpha = -\mathbf{M}^T\mathbf{M} + 2\alpha\mathbf{L}/\mu$.

Updating $\mathbf{\Lambda}$ *and* μ. The update of $\mathbf{\Lambda}$ and μ is standard:

$$\begin{aligned} \mathbf{\Lambda}^{k+1} &= \mathbf{\Lambda}^k + \mu\left(\mathbf{S}^{k+1} - \mathbf{X} + \mathbf{U}^k\mathbf{V}^{T^k}\right), \\ \mu^{k+1} &= \rho\mu^k, \end{aligned} \tag{23}$$

where $\rho > 1$ is used to update the parameter μ. Since the value of ρ is usually bigger than 1, and over a large number of experiments, we find $\rho = 1.1\sim1.5$ are good choice. We selected $\rho = 1.2$ in such practice conditions.

The complete procedure is summarized in Algorithm 1.

3.2. Properties of Algorithm. We set $\rho = 1.2$ through all our gene expression data experiments. Whereas we introduce σ_m, σ_l is the largest eigenvalue of matrix $\mathbf{M}^T\mathbf{M}$ and $\mathbf{L}$ to normalize them, respectively. Setting

$$\frac{2\alpha}{\mu} = \frac{\beta}{1-\beta}\frac{\sigma_m}{\sigma_l}, \tag{24}$$

where β is the parameter to substitute for α, (20) can be rewritten as

$$\begin{aligned} &\mathbf{V} = \arg\min_{\mathbf{V}} \operatorname{tr}\mathbf{V}^T\left[(1-\beta)\left(\mathbf{I} - \frac{\mathbf{M}^T\mathbf{M}}{\sigma_m}\right) + \frac{2\beta}{\mu}\frac{\mathbf{L}}{\sigma_l}\right]\mathbf{V}, \\ &s.t. \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}. \end{aligned} \tag{25}$$

Therefore, the solution of $\mathbf{V}$ can be expressed by the eigenvectors of G_β:

$$G_\beta = (1-\beta)\left(\mathbf{I} - \frac{\mathbf{M}^T\mathbf{M}}{\sigma_m}\right) + \frac{2\beta}{\mu}\frac{\mathbf{L}}{\sigma_l}. \tag{26}$$

It is easy to see that β should be in the range $0 \le \beta \le 1$. Without $L_{1/2}$-norm, there will be standard PCA if $\beta = 0$. Similarly, when $\beta = 1$, it reduces to Laplacian embedding.

Furthermore, we rewrite the matrix G_β as follows:

$$G_\beta = (1-\beta)\left(\mathbf{I} - \frac{\mathbf{M}^T\mathbf{M}}{\sigma_m}\right) + \frac{2\beta}{\mu}\left(\frac{\mathbf{L}}{\sigma_l} + \frac{\mathbf{e}\mathbf{e}^T}{n}\right), \tag{27}$$

where $\mathbf{e} = (1\cdots1)^T$ is an eigenvector of G_β: $G_\beta\mathbf{e} = (1-\beta)\mathbf{e}$. We have $\mathbf{M}\mathbf{e} = 0$, because $\mathbf{X}$ is centered and it is easy to

Input: Data matrix $\mathbf{X} \in \mathbf{R}^{m\times n}$;
weight matrix: $\mathbf{W} \in R^{m\times n}$;
parameters: β, ρ, k, μ.
Output: Optimized matrix: $\mathbf{U}$, $\mathbf{V}$.
Initialization: $\mathbf{S} = \mathbf{\Lambda} = 0$, $\mathbf{U} = 0$, $\mathbf{V} = 0$.
repeat
Step 1. Update $\mathbf{S}$ and fix the others by $\mathbf{S}^{k+1} = \arg\min_{\mathbf{s}} \|\mathbf{S}\|_{1/2}^{1/2} + (\mu/2)\|\mathbf{S} - \mathbf{X} + \mathbf{U}\mathbf{V}^T + \mathbf{\Lambda}/\mu\|_F^2$.
Step 2. Update $\mathbf{U}$ and fix the others by $\mathbf{U}^{k+1} = \mathbf{M}\mathbf{V}^k$.
Step 3. Update $\mathbf{V}$ and fix the others by $\mathbf{V}^{k+1} = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$.
Step 4. Update $\mathbf{\Lambda}$, μ and fix the others by
$\mathbf{\Lambda}^{k+1} = \mathbf{\Lambda}^k + \mu(\mathbf{S}^{k+1} - \mathbf{X} + \mathbf{U}^k\mathbf{V}^{T^k})$
$\mu^{k+1} = \rho\mu^k$
until converge

Algorithm 1: Procedure of $L_{1/2}$ gLPCA.

see that $\mathbf{M} = \mathbf{X} - \mathbf{S} - \mathbf{\Lambda}/\mu$ is centered. G_β is semipositive definite, because σ_m is the biggest eigenvalue of $\mathbf{M}^T\mathbf{M}$; thus $\mathbf{I} - \mathbf{M}^T\mathbf{M}/\sigma_m$ is semipositive definite. Meanwhile, it is easy to see that $\mathbf{L}$ is semipositive definite. Since G_β is a symmetric real matrix that eigenvectors are mutually orthogonal, thus $\mathbf{e}$ is orthogonal to others. Although we apply $\mathbf{e}\mathbf{e}^T/n$ in the Laplacian matrix part, the eigenvectors and eigenvalues do not change, which guarantees that the lowest k eigenvectors do not include $\mathbf{e}$.

4. Experiments

In this section, we compare our algorithm with Laplacian embedding (LE) [26], PCA [9], L_0 PCA, L_1 PCA [12], gLPCA, and RgLPCA [23] on simulation data and real gene expression data, respectively, to verify the performance of our algorithm. Among them, PCA and LE are obtained by adjusting the parameters of gLPCA $\beta = 0$ and $\beta = 1$, respectively. Since our algorithm is not sensitive to parameter mu in practice. In the first subsection, we provide the source of simulation data and experimental comparison results. The experimental results and the function of selected genes on real gene expression data with different methods are compared in the next two subsections.

4.1. Results on Simulation Data

4.1.1. Data Source. Here, we describe a method to produce simulation data. Supposing we generate the data matrix $\mathbf{A} \in \mathbf{R}^{k\times j}$, where $k = 2000$ and $j = 10$ are the number of genes and samples, respectively, the simulation data are generated as $\mathbf{A} \sim (0, \Sigma_4)$. Let $\tilde{\mathbf{v}}_1 \sim \tilde{\mathbf{v}}_4$ be four 2000-dimensional vectors; for instance, $\tilde{\mathbf{v}}_{1k} = 1$, $k = 1, \ldots, 50$, and $\tilde{\mathbf{v}}_{1k} = 0$, $k = 51, \ldots, 2000$; $\tilde{\mathbf{v}}_{2k} = 1$, $k = 51, \ldots, 100$, and $\tilde{\mathbf{v}}_{2k} = 0$, $k \neq 51, \ldots, 100$; $\tilde{\mathbf{v}}_{3k} = 1$, $k = 101, \ldots, 150$, and $\tilde{\mathbf{v}}_{3k} = 0$, $k \neq 101, \ldots, 150$; $\tilde{\mathbf{v}}_{4k} = 1$, $k = 151, \ldots, 200$, and $\tilde{\mathbf{v}}_{4k} = 1$, $k \neq 151, \ldots, 200$. Given a matrix $\mathbf{E} \sim N(0, 1)$ as a noise matrix with 2000-dimension and different Signal-to-Noise Ratio (SNR), which is added into $\tilde{v}$, the four eigenvectors of Σ_4 can be expressed as $\tilde{\mathbf{v}}_k = \tilde{\mathbf{v}}_k/\|\tilde{\mathbf{v}}_k\|$, $k = 1, 2, 3, 4$. Let the four eigenvectors dominate; the eigenvalues of $\mathbf{A}$ can be denoted as $c_1 = 400$, $c_2 = 300$, $c_3 = 200$, $c_4 = 100$, and $c_k = 1$ for $k = 5, \ldots, 2000$.

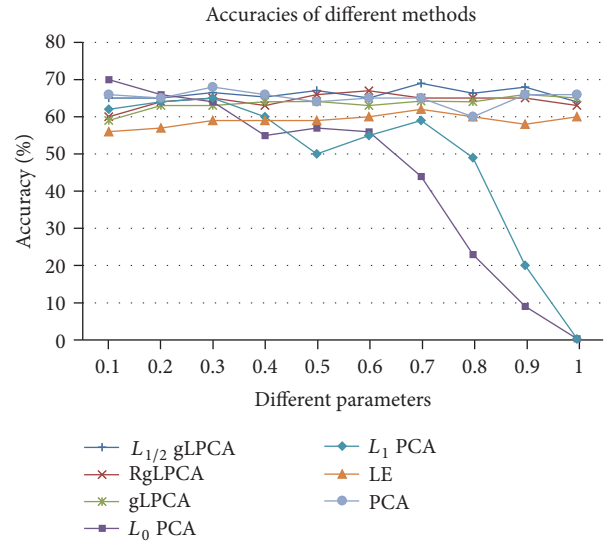


Figure 1: The accuracy of different methods on simulation data with different parameters.

4.1.2. Detailed Results on Simulation Data. In order to give more accurate experiment results, the average values of the results of 30 times are adopted. For fairness and uniformity, 200 genes are selected by the five methods with their unique parameters. Here, we show the accuracy (%) of these methods. In Figure 1, two factors as two different axes are in the figure. In Figure 2, x-axis is the number of samples. x-axis is the value of parameter μ. The accuracy is defined as follows:

$$\text{Accuracy} = \frac{1}{t}\sum_{i=1}^{t} \text{Acc}_i \times 100\%, \tag{28}$$

Table 1: The average accuracy and variance of different methods on simulation data with different parameters.

Methods	$L_{1/2}$ gLPCA	RgLPCA	gLPCA	L_0 PCA	L_1 PCA	PCA	LE
Average accuracy (%)	66.12	65.47	63.53	44.43	48.43	59.00	65.10
Variance	1.48	1.62	1.76	23.60	20.30	1.61	1.97

Table 2: The average accuracy and variance of different methods on simulation data with different numbers of samples.

Methods	$L_{1/2}$ gLPCA	RgLPCA	gLPCA	L_0 PCA	L_1 PCA	PCA	LE
Average accuracy (%)	70.25	68.25	67.90	67.30	69.20	58.62	69.60
Variance	2.58	3.84	4.41	3.52	2.23	1.79	2.50

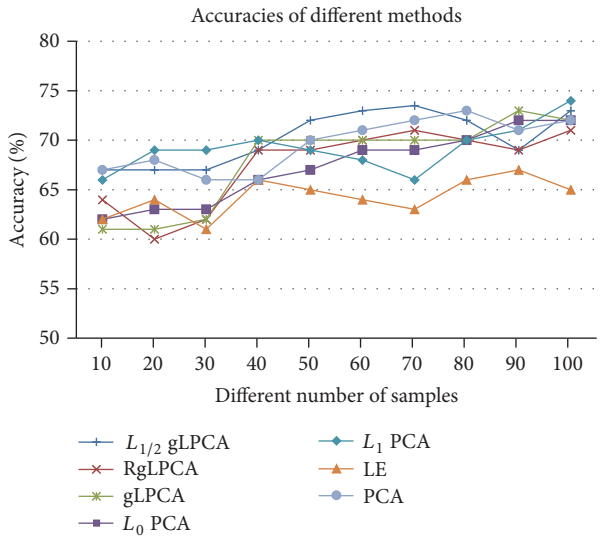


Figure 2: The accuracy of different methods on simulation data with different numbers of samples.

where t is the iterative times and Acc_i is the identification accuracy of the ith time. We define Acc as follows:

$$\mathrm{Acc} = \frac{1}{r}\sum_{j=1}^{r}\delta\left(I_j, \mathrm{map}\left(I_j\right)\right), \tag{29}$$

where r denotes the number of genes, $\delta(m,n)$ is a function that equals to 0 if $m \neq n$ and equals to 1 if $m = n$. We use the function map(I) to map the identification of labels. In Figure 1, we show the average accuracies of the seven methods with different sparse parameters while the simulation data is 2000×10 and the average accuracy with all parameters is listed in Table 1. In general, if the algorithm is more sensitive to noise and outliers, the deviation will be greater and the accuracy will be greatly reduced. It is worthy to notice that $L_{1/2}$ gLPCA works better than other six methods with higher identification accuracies. This means that our algorithm has lower sensitivity to noise and outliers. This table clearly displays the detail of the identification accuracies in different sparse parameters; our method indicates the superiority when the parameter is larger than 0.4 and the curve is more stable. The accuracy of L_0 PCA and L_1 PCA starts a precipitous decline when the parameter is larger than 0.7 and 0.8. Compared with L_0 PCA and L_1 PCA, the methods of $L_{1/2}$ gLPCA, RgLPCA, gLPCA, PCA, and LE are not sensitive to the parameter, so there is no substantial change. The stability and average accuracy of various methods can be seen from Table 1.

Furthermore, the number of samples in real gene expression data has a significant influence on the identification accuracy when we select feature gene. Based on this theory, we test different numbers of samples with their best parameters and the average values of the results of 30 times. From the results of Figure 1, we select 0.8 as the parameters of $L_{1/2}$ gLPCA, gLPCA, RgLPCA, PCA, and LE. For L_0 PCA and L_1 PCA, we do not change its parameters, since it can obtain the best result from the author's description. The details of average identification accuracies which use seven methods with different sample numbers can be seen from Figure 2. As seen in Figure 2, the accuracy of $L_{1/2}$ gLPCA is generally better than other methods and increases with the increase of the number of samples. Besides, Table 2 shows the average accuracy and variance of seven different methods on simulation data with different number of samples. From Table 2, our approach performs better than other methods, even though, in the case of a small number of samples, the accuracy is still high.

4.2. Results on Gene Expression Data. In this subsection, the features (genes) are selected by these methods and sent to ToppFun to detect the gene-set enrichment analysis, which is a type of GOTermFinder [27]. The primary role of GOTermFinder is to discover the common of large amounts of gene expression data. The analysis of GOTermFinder provides critical information for the experiment of feature extraction. It is available publicly at https://toppgene.cchmc.org/enrichment.jsp. We set P value cutoff to 0.01 through all the experiment. For fair comparison, about $L_{1/2}$ gLPCA, RgLPCA, and gLPCA, we both set $\beta = 0.5$ to control the degree of Laplacian embedding through all experiments in this paper. When $\beta = 0$, $\beta = 1$, it results in standard PCA and LE, respectively. Since our algorithm is not sensitive to parameter μ mu in practice, we set $\mu = 0.3$ through our experiment.

4.2.1. Results on ALLAML Data. The data of ALLAML as a matrix includes 38 samples and 5000 features (genes), which are publicly available at https://sites.google.com/site/feipingnie/file. It is made up of 11 types of acute myelogenous leukemia (AML) and 27 types of acute lymphoblastic

leukemia (ALL) [28]. This data contains the difference between AML and ALL, and ALL is divided into T and B cell subtypes. In this experiment, 300 genes are selected and sent to ToppFun. A series of enrichment analyses are conducted on the extracted top 500 genes corresponding to different methods. The complete experimental data have been listed as supplementary data. The P value and hit count of top nine terms about molecular function, biological process, and cellular component of ALLAML data by different methods are listed in Table 3. The P value is significance for these genes enrichment analysis in these GO terms; the smaller the P value is, the more significant these GO terms are. In this Table, the number of hits is the number of genes from input, and the P value was influenced by the number of genes from input and so on. Thus, the difference in number of hits is smaller than the difference in P value. It shows clearly that our method performs better than compared methods in 8 terms. The lower P value shows that the algorithm is less affected by noise and outliers and thus has high efficiency. If the algorithm is affected by noise and outliers significantly, the degree of gene enrichment will be reduced. Nevertheless, LE has the lowest P value in term GO: 0098552. From this table, we can see that there are 93 genes in the item of "immune response" which are selected by our method. This item can be considered as the most probable enrichment item, since it has the lowest P value. And many researches were focused on the immune status of leukemia [29–32]. Besides, 210 genes associated with leukemia are listed in an article, and 26 out of top 30 genes selected by our method can be found in this article [33]. And 30 genes selected by our method can be found in another published article [34]. The high overlap rate of these genes selected by our method with this published literature approved the effectiveness of our method.

4.2.2. Pathway Search Result on ALLAML Data. For the sake of the correlations between the selected genes and ALLAML data, the genes selected by $L_{1/2}$ gLPCA are proved based on gene-set enrichment analysis (GSEA) that is publicly available at http://software.broadinstitute.org/gsea/msigdb/annotate.jsp. We make analysis by GSEA to compute overlaps for selected genes. Figure 3 displays the pathway of hematopoietic cell lineage that has highest gene overlaps in this experiment. From Figure 3, 15 genes from our experiment are contained. Among them, HLA-DR occurs seven times. Hematopoietic cell lineage belongs to organismal systems and immune system. On the subject of acute myeloid leukemia (AML), there is consensus about the target cell within the hematopoietic stem cell hierarchy that is sensitive to leukemic transformation, or about the mechanism, that is, basic phenotypic, genotypic, and clinical heterogeneity [35]. Hematopoietic stem cell (HSC) developing from the blood-cell can undergo self-renewal and differentiate into a multilineage committed progenitor cell: one is a common lymphoid progenitor (CLP) and the other is called a common myeloid progenitor (CMP) [36]. A CLP causes the lymphoid lineage of white blood cells or leukocytes, the natural killer (NK) cells and the T and B lymphocytes. A CMP causes the myeloid lineage, which comprises the rest of the leukocytes, the erythrocytes (red blood cells), and the megakaryocytes that produce platelets important in blood clotting. Cells express a stage- and lineage-specific set of surface markers in the differentiation process. So the specific expression pattern of these genes is one way to identify the cellular stages. Related diseases include hemophilia, Bernard-Soulier syndrome, and castleman disease. In medicine, leukemia is a kind of malignant clonal disease of hematopoietic stem cells. Bone marrow transplantation is a magic weapon for the cure of leukemia, by recreating the hematopoietic system to cure leukemia. Generally speaking, when a person has problem in hematopoietic system, it might be related to leukemia [37].

4.2.3. Results on TCGA with PAAD-GE Data. As the largest public database of cancer gene information, The Cancer Genome Atlas (TCGA, https://tcgadata.nci.nih.gov/tcga/) has been producing multimodal genomics, epigenomics, and proteomics data for thousands of tumor samples across over 30 types of cancer. At the same time, as a multidimensional combination of data, five levels of data are involved, such as gene expression (GE), Protein Expression (PE), DNA Methylation (ME), DNA Copy Number (CN), and microRNA Expression (miRExp). Two disease data sets are downloaded from TCGA to be analyzed in the following two experiments. Pancreatic cancer is a type of disease that threatens human health. In this experiment, pancreatic cancer gene expression data (PAAD-GE) is analyzed by these methods. The data of PAAD-GE data as a matrix includes 180 samples and 20502 features (genes). In this subsection, we extract PAAD-GE data to complete this set of comparative experiments and 500 genes are selected and sent to ToppFun. We select top nine terms from molecular function, biological process, and cellular component by $L_{1/2}$ gLPCA and compare with other methods. The P value and hit count of these terms are listed in Table 4. It is indicated clearly in Table 4 that our method is more stable than other methods, which has lower P value in 7 terms. But in terms GO:0045047 and GO:0072599, PCA performs better than other methods. Nevertheless, $L_{1/2}$ gLPCA has the same P value with gLPCA in terms GO:0045047 and GO:0072599. 196 genes in the item of "extracellular space" are selected by our method.

4.2.4. Pathway Search Result on PAAD-GE Data. Similarly as the last experiment, we send our result to GSEA and list the highest genes overlap pathway map in Figure 4. In 1982, Ohhashi reported 4 cases with unique clinical pathological features and is different from normal pancreatic cancer cases, and these 4 cases belong to a completely new clinical type, known as "mucus production type carcinoma (mucin-producing carcinoma, M-pC)." Focal adhesion belongs to cellular processes and cellular community. More specifically, cell-matrix adhesions play important roles in biological processes including cell motility, cell proliferation, cell differentiation, regulation of gene expression, and cell survival. At the cell-extracellular matrix contact points, specialized structures are created and termed focal adhesions, where bundles of actin filaments are fixed to transmembrane receptors of the integrin family through a multimolecular complex of junctional plaque proteins. Integrin signaling is dependent on the nonreceptor tyrosine kinase activities of the FAK

Table 3: Enrichment analysis of the top 500 genes in the ALLAML data corresponding to different methods.

ID	Name	$L_{1/2}$ gLPCA		RgLPCA		gLPCA		L_0 PCA		L_1 PCA		PCA		LE	
		P-value	Hit	P-value	Hit	P-value	Hit	P-value	Hit	P-value	Hit	P-value	Hit	P-value	Hit
GO:0006955	Immune response	**1.34E − 36**	93	$2.51E-34$	91	$1.20E-34$	91	$2.45E-31$	87	$5.14E-32$	89	$4.05E-35$	91	$1.98E-35$	91
GO:0002684	Positive regulation of immune system process	**2.44E − 29**	67	$2.17E-25$	63	$1.24E-26$	64	$3.60E-28$	66	$1.56E-27$	66	$8.98E-28$	65	$3.45E-28$	66
GO:0098552	Side of membrane	$3.80E-25$	46	$5.19E-34$	45	$2.70E-22$	43	$2.23E-20$	41	$7.25E-21$	42	$2.01E-23$	44	**4.24E − 26**	47
GO:0009897	External side of plasma membrane	**1.83E − 17**	30	$6.34E-16$	29	$9.51E-14$	26	$1.14E-13$	26	$1.41E-12$	25	$1.31E-16$	29	$1.83E-14$	26
GO:0005615	Extracellular space	**2.01E − 17**	63	$8.37E-15$	60	$2.39E-14$	58	$6.12E-16$	61	$3.52E-13$	57	$2.27E-16$	61	$4.74E-16$	61
GO:0005764	Lysosome	**3.49E − 17**	38	$7.43E-16$	37	$5.46E-14$	34	$1.20E-14$	35	$9.22E-11$	30	$1.08E-15$	36	$3.49E-16$	37
GO:0009986	Cell surface	**3.58E − 17**	48	$4.82E-15$	45	$4.68E-13$	42	$6.13E-13$	42	$5.58E-12$	41	$6.58E-16$	46	$3.58E-16$	46
GO:0042277	Peptide binding	**5.03E − 14**	25	$5.92E-13$	24	$2.85E-11$	22	$3.33E-11$	22	$7.54E-08$	18	$3.09E-12$	23	$1.80E-10$	21
GO:0033218	Amide binding	**7.37E − 14**	26	$4.34E-12$	24	$3.44E-11$	23	$4.04E-11$	23	$7.36E-08$	19	$3.95E-12$	24	$2.04E-10$	22

Table 4: Enrichment analysis of the top 500 genes in the PAAD-GE data corresponding to different methods.

ID	Name	$L_{1/2}$ gLPCA		RgLPCA		gLPCA		L_0 PCA		L_1 PCA		PCA		LE	
		P value	Hit	P value	Hit	P value	Hit	P value	Hit	P value	Hit	P value	Hit	P value	Hit
GO:0005615	Extracellular space	**3.20E − 93**	196	$3.56E-80$	183	$2.18E-72$	173	$2.742E-61$	160	$7.82E-61$	161	$1.44E-58$	157	$3.20E-89$	191
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	**2.79E − 86**	67	$6.82E-75$	56	$1.37E-73$	64	$3.45E-51$	48	$8.17E-56$	51	$7.45E-82$	63	$2.76E-56$	51
GO:0070972	Protein localization to endoplasmic reticulum	**1.01E − 83**	73	$2.42E-72$	69	$6.37E-71$	68	$5.31E-48$	51	$4.63E-52$	54	$2.88E-76$	71	$4.70E-51$	53
GO:0006613	Cotranslational protein targeting to membrane	**1.86E − 82**	67	$3.48E-79$	65	$7.58E-73$	64	$3.27E-49$	48	$1.19E-53$	51	$2.04E-80$	66	$4.04E-54$	51
GO:0045047	Protein targeting to ER	**5.01E − 82**	67	$5.19E-74$	65	$2.00E-71$	64	$6.04E-49$	48	$2.33E-53$	51	**5.80E − 82**	67	$7.90E-54$	51
GO:0022626	Cytosolic ribosome	**1.34E − 81**	68	$2.13E-75$	64	$1.44E-70$	62	$8.30E-44$	47	$8.45E-48$	50	$6.34E-74$	66	$4.01E-51$	52
GO:0072599	Establishment of protein localization to endoplasmic reticulum	**2.77E − 80**	67	$8.15E-78$	66	$4.44E-70$	64	$6.44E-48$	48	$3.09E-52$	51	**3.20E − 80**	67	$1.05E-52$	51
GO:0005198	Structural molecule activity	**1.82E − 68**	126	$2.46E-65$	124	$6.14E-63$	121	$3.62E-52$	110	$5.16E-54$	113	$3.32E-65$	124	$1.03E-54$	113
GO:0044391	Ribosomal subunit	**5.14E − 64**	69	$5.18E-60$	65	$3.22E-56$	63	$2.86E-37$	49	$1.42E-40$	52	$1.58E-63$	68	$3.70E-42$	53

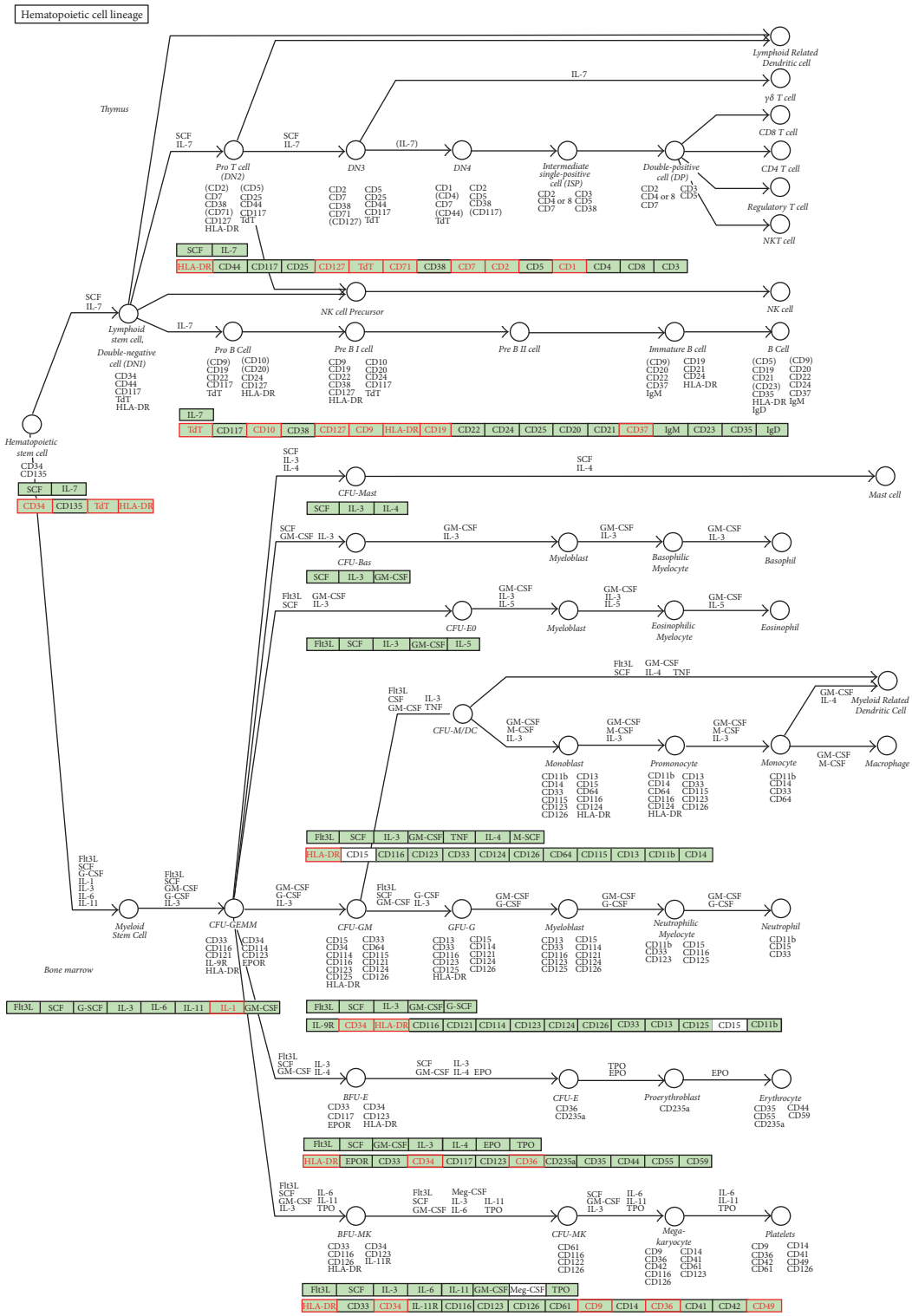


FIGURE 3: The pathway of hematopoietic cell lineage.

and src proteins as well as the adaptor protein functions of FAK, src and Shc to start downstream signaling events. Similar morphological alterations and modulation of gene expression are started by the binding of growth factors to their respective receptors, underling the considerable crosstalk between adhesion- and growth factor-mediated signaling. The early literatures have shown that there is a certain relationship between the pancreatic cancer and focal adhesion [38]. Activation of focal adhesion kinase enhances the adhesion and invasion of pancreatic cancer cells. Besides,

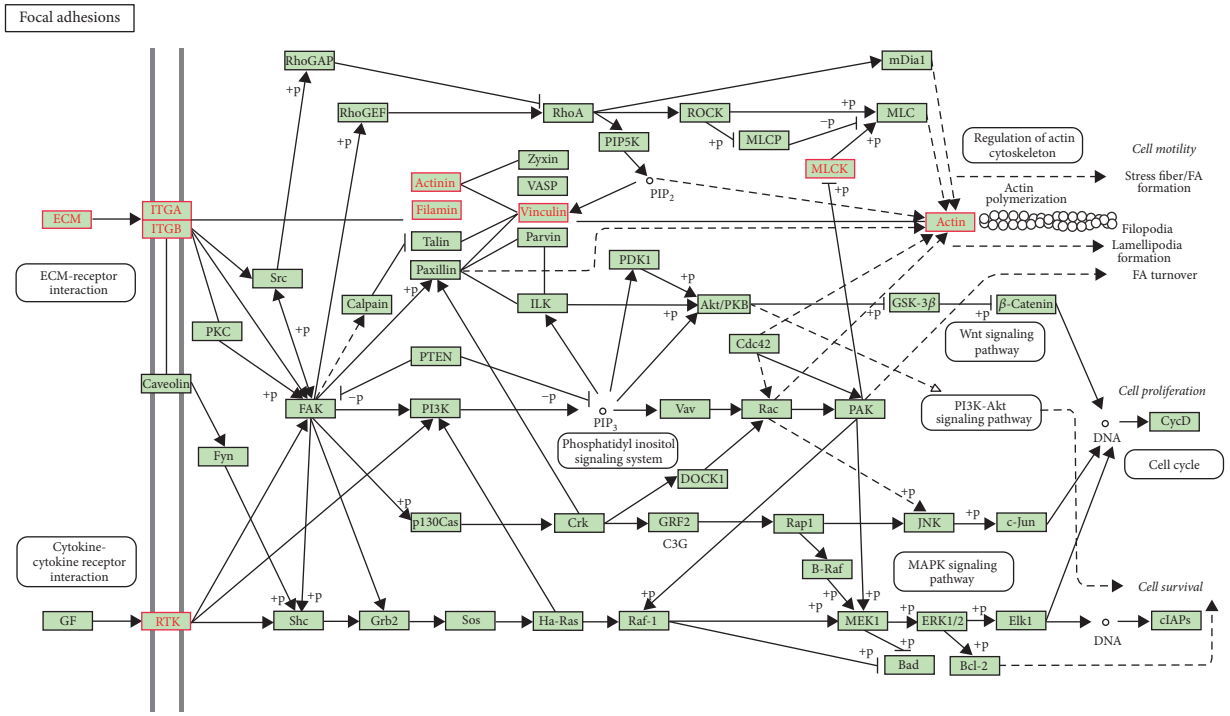


Figure 4: The pathway of focal adhesion.

Table 5: The function of top 7 extraction genes.

Gene ID	Gene name	Related GO annotations	Related diseases	Paralogous genes
5644	PRSS1	Serine-type endopeptidase activity	Trypsinogen deficiency and prss1-related hereditary pancreatitis	KLK12
5406	PNLIP	Carboxylic ester hydrolase activity and triglyceride lipase activity	Pancreatic colipase deficiency and pancreatic lipase deficiency	LPL
1357	CPA1	Metallocarboxypeptidase activity and exopeptidase activity	Borna disease and pancreatitis, hereditary	CPA3
1360	CPB1	Metallocarboxypeptidase activity and carboxypeptidase activity	Acute pancreatitis and tricuspid valve insufficiency	CPA3
63036	CELA2A	Serine-type endopeptidase activity and serine hydrolase activity	Pancreatitis, hereditary	CELA2B
5967	REG1A	Carbohydrate binding and growth factor activity	Acinar cell carcinoma and tropical calcific pancreatitis	REG3G
1056	CEL	Hydrolase activity and carboxylic ester hydrolase activity	Maturity-onset diabetes of the young, Type VIII and maturity-onset diabetes of the young	CES2

Type II diabetes mellitus is another important pathway and is widely believed to be associated with pancreatic cancer; a meta-analysis has examined this association [39].

4.2.5. Correlations between the Selected Genes and PAAD-GE Data. The function of top 7 genes selected by $L_{1/2}$ gLPCA is listed in Table 5 based on literatures and GeneCards (http://www.genecards.org/). As can be clearly seen from the table, most of these genetic lesions would likely incur pancreas-related diseases. The etiology of pancreatic cancer is not very clear; it is noted that there is a certain relationship between the incidence of chronic pancreatitis and pancreatic cancer, and we find a significant increase in the proportion of chronic pancreatitis patients with pancreatic cancer. This view is consistent with our experimental result. The clinical observation shows that abdominal pain is the most obvious symptom in the early stage of pancreatic cancer. Some literature on these genes also made a further research as follows. The gene PRSS1 variant likely affects disease susceptibility by altering expression of the primary trypsinogen gene [40]. The pancreatic lipase gene (PNLIP) is located within the genomic region of a bovine marbling quantitative trait locus. PNLIP is a positional and functional candidate for the marbling gene [41].

Table 6: The Acc and highest relevance score of these methods.

Dataset	$L_{1/2}$ gLPCA		RgLPCA		gLPCA		L_0 PCA		L_1 PCA		PCA		LE	
	Acc (%)	Relevance score	Acc (%)	Relevance score	Acc (%)	Relevance score	Acc (%)	Relevance score	Acc (%)	Relevance score	Acc (%)	Relevance score	Acc (%)	Relevance score
AMLALL	**51.33**	**55.37**	49.88	46.11	48.67	46.11	40.00	38.15	52.00	46.11	49.00	46.11	49.60	46.11
PAAD-GE	**61.60**	**85.56**	60.51	61.01	59.40	61.01	43.80	54.77	47.20	54.77	57.20	82.20	61.40	82.20

4.3. The Accuracy and Highest Relevance Score. Because ALLAML and PPAD are human disease data sets, we can find them directly from GeneCards and they are publicly available at http://www.genecards.org/.

In order to summarize the experiments on gene expression data, we compute the accuracy and highest relevance score of these methods from GeneCards and list the details in Table 6. The accuracy in Table 6 indicates the proportion of genes which are real associated with the disease in all of the genes selected by these methods. From Table 6, we observe the following. (1) Both PCA and LE commonly provide better accuracy results than L_0 PCA and L_1 PCA, demonstrating the usefulness of PCA and LE. (2) gLPCA has a good performance in some conditions and is unstable. Thus, it is necessary to reduce the effects of outliers and noise. (3) $L_{1/2}$ gLPCA and RgLPCA consistently perform better than other methods, but $L_{1/2}$ gLPCA has the highest relevance score and highest accuracy.

5. Conclusions

This paper investigates a new method of graph-Laplacian PCA ($L_{1/2}$ gLPCA) by applying $L_{1/2}$-norm constraint on the former method. $L_{1/2}$-norm constraint is applied on error function to improve the robustness of the PCA-based method. Augmented Lagrange Multipliers (ALM) method is applied to solve the optimization problem. Extensive experiments on both simulation and real gene expression data have been performed. Results on these two kinds of data show that our proposed method performs better than compared methods. Based on our proposed method, many genes have been extracted to analyze. The identified genes are demonstrated that they are closely related to the corresponding cancer data set.

In future, we will modify the model to improve sparse and robustness of the structure at the same time.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the NSFC under Grants nos. 61572284 and 61502272.

References

[1] M. J. Heller, "DNA microarray technology: devices, systems, and applications," *Annual Review of Biomedical Engineering*, vol. 4, pp. 129–153, 2002.

[2] C. K. Sarmah and S. Samarasinghe, "Microarray gene expression: a study of between-platform association of Affymetrix and cDNA arrays," *Computers in Biology and Medicine*, vol. 41, no. 10, pp. 980–986, 2011.

[3] J. Liu, D. Wang, Y. Gao et al., "A joint-L2,1-norm-constraint-based semi-supervised feature extraction for RNA-Seq data analysis," *Neurocomputing*, vol. 228, pp. 263–269, 2017.

[4] J.-X. Liu, Y. Xu, Y.-L. Gao, C.-H. Zheng, D. Wang, and Q. Zhu, "A class-information-based sparse component analysis method to identify differentially expressed genes on RNA-Seq data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 2, pp. 392–398, 2016.

[5] E. Levine, Z. Zhang, T. Kuhlman, and T. Hwa, "Quantitative characteristics of gene regulation by small RNA," *PLoS Biology*, vol. 5, no. 9, article e229, 2007.

[6] J. Liu, Y. Gao, C. Zheng, Y. Xu, and J. Yu, "Block-constraint robust principal component analysis and its application to integrated analysis of TCGA data," *IEEE Transactions on NanoBioscience*, vol. 15, no. 6, pp. 510–516, 2016.

[7] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[8] K.-J. Kim and S.-B. Cho, "Meta-classifiers for high-dimensional, small sample classification for gene expression analysis," *Pattern Analysis and Applications*, vol. 18, no. 3, pp. 553–569, 2015.

[9] B. E. Jolli, *Principal Component Analysis*, Springer, 2nd edition, 2012.

[10] D. Meng, Q. Zhao, and Z. Xu, "Improve robustness of sparse PCA by L_1-norm maximization," *Pattern Recognition*, vol. 45, no. 1, pp. 487–497, 2012.

[11] D. Meng, H. Cui, Z. Xu, and K. Jing, "Following the entire solution path of sparse principal component analysis by coordinate-pairwise algorithm," *Data and Knowledge Engineering*, vol. 88, pp. 25–36, 2013.

[12] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *Journal of Machine Learning Research*, vol. 11, pp. 517–553, 2010.

[13] Y. Zheng, B. Jeon, D. Xu, Q. M. J. Wu, and H. Zhang, "Image segmentation by generalized hierarchical fuzzy C-means algorithm," *Journal of Intelligent and Fuzzy Systems*, vol. 28, no. 2, pp. 961–973, 2015.

[14] F. R. Chung, *Spectral Graph Theory*, vol. 92, American Mathematical Society, Providence, RI, USA, 1997.

[15] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proceedings of the 15th Annual Neural Information Processing Systems Conference (NIPS '01)*, pp. 585–591, December 2001.

[16] F. Nie, H. Huang, and C. H. Ding, "Low-rank matrix recovery via efficient schatten p-norm minimization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2012.

[17] S. Jia, X. Zhang, and Q. Li, "Spectral-spatial hyperspectral image classification using regularized low-rank representation and sparse representation-based graph cuts," *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, vol. 8, pp. 2473–2484, 2015.

[18] Z. Xu, X. Chang, F. Xu, and H. Zhang, "L1/2 regularization: a thresholding representation theory and a fast solver," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 7, pp. 1013–1027, 2012.

[19] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *The Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 629–654, 2008.

[20] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[21] Z.-B. Xu, H.-L. Guo, Y. Wang, and H. Zhang, "Representative of $L_{1/2}$ regularization among L_q $(0 < q \leq 1)$ regularizations: an experimental study based on phase diagram," *Acta Automatica Sinica*, vol. 38, no. 7, pp. 1225–1228, 2012.

[22] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, p. 29, ACM, July 2004.

[23] B. Jiang, C. Ding, B. Luo, and J. Tang, "Graph-laplacian PCA: closed-form solution and robustness," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 3492–3498, June 2013.

[24] C. M. Feng, J. X. Liu, Y. L. Gao, J. Wang, D. Q. Wang, and Y. Du, "A graph-Laplacian PCA based on L1/2-norm constraint for characteristic gene selection," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '16)*, pp. 1795–1799, Shenzhen, China, 2016.

[25] S. Yang, C. Hou, F. Nie, and Y. Wu, "Unsupervised maximum margin feature selection via $L_{2,1}$-norm minimization," *Neural Computing and Applications*, vol. 21, no. 7, pp. 1791–1799, 2012.

[26] X. Xin, Z. Li, and A. K. Katsaggelos, "Laplacian embedding and key points topology verification for large scale mobile visual identification," *Signal Processing: Image Communication*, vol. 28, no. 4, pp. 323–333, 2013.

[27] 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.

[28] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 12, pp. 4164–4169, 2004.

[29] T. K. Richmond, E. Tili, M. Brown, M. Chiabai, D. Palmieri, and R. Cui, "Abstract LB-289: interaction between miR-155 and Quaking in the innate immune response and leukemia," *Cancer Research*, vol. 75, pp. 534–538, 2015.

[30] M. Essex, A. Sliski, W. D. Hardy Jr., and S. M. Cotter, "Immune response to leukemia virus and tumor-associated antigens in cats," *Cancer Research*, vol. 36, no. 2, part 2, pp. 640–645, 1976.

[31] X. Zhang, Y. Su, H. Song, Z. Yu, B. Zhang, and H. Chen, "Attenuated A20 expression of acute myeloid leukemia-derived dendritic cells increased the anti-leukemia immune response of autologous cytolytic T cells," *Leukemia Research*, vol. 38, no. 6, pp. 673–681, 2014.

[32] N. A. Gillet, M. Hamaidia, A. de Brogniez et al., "The bovine leukemia virus microRNAs permit escape from innate immune response and contribute to viral replication in the natural host," *Retrovirology*, vol. 12, supplement 1, pp. 1190–1195, 2015.

[33] M.-Y. Wu, D.-Q. Dai, X.-F. Zhang, and Y. Zhu, "Cancer subtype discovery and biomarker identification via a new robust network clustering algorithm," *PLoS ONE*, vol. 8, no. 6, Article ID e66256, 2013.

[34] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[35] D. Bonnet and J. E. Dick, "Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell," *Nature Medicine*, vol. 3, no. 7, pp. 730–737, 1997.

[36] D. Metcalf, "On hematopoietic stem cell fate," *Immunity*, vol. 26, no. 6, pp. 669–673, 2007.

[37] F. Ciceri, M. Labopin, F. Aversa et al., "A survey of fully haploidentical hematopoietic stem cell transplantation in adults with high-risk acute leukemia: a risk factor analysis of outcomes for patients in remission at transplantation," *Blood*, vol. 112, no. 9, pp. 3574–3581, 2008.

[38] H. Sawai, Y. Okada, H. Funahashi et al., "Activation of focal adhesion kinase enhances the adhesion and invasion of pancreatic cancer cells via extracellular signal-regulated kinase-1/2 signaling pathway activation," *Molecular Cancer*, vol. 4, article 37, 12 pages, 2005.

[39] R. Huxley, A. Ansary-Moghaddam, A. Berrington De González, F. Barzi, and M. Woodward, "Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies," *British Journal of Cancer*, vol. 92, no. 11, pp. 2076–2083, 2005.

[40] D. C. Whitcomb, J. LaRusch, A. M. Krasinskas et al., "Common genetic variants in the CLDN2 and PRSS1-PRSS2 loci alter risk for alcohol-related and sporadic pancreatitis," *Nature*, vol. 44, no. 12, pp. 1349–1354, 2012.

[41] Y. Muramatsu, H. Tanomura, T. Ohta, H. Kose, and T. Yamada, "Allele frequency distribution in PNLIP promoter SNP is different between high-marbled and low-marbled Japanese Black Beef Cattle," *Open Journal of Animal Sciences*, vol. 6, p. 137, 2016.

Hindawi
BioMed Research International
Volume 2017, Article ID 2954351, 9 pages
https://doi.org/10.1155/2017/2954351

Research Article

Dissect the Dynamic Molecular Circuits of Cell Cycle Control through Network Evolution Model

Yang Peng,[1] Paul Scott,[2] Ruikang Tao,[3] Hua Wang,[2] Yan Wu,[2] and Guang Peng[1,4]

[1]*Department of Clinical Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA*
[2]*Mathematical Sciences, Georgia Southern University, Statesboro, GA 30458, USA*
[3]*University of California, Santa Cruz, CA 65064, USA*
[4]*Department of Medical Oncology, Tongji Hospital, Tongji Medical College, The Huazhong University of Science and Technology, Wuhan, China*

Correspondence should be addressed to Yan Wu; yan@georgiasouthern.edu and Guang Peng; gpeng@mdanderson.org

Received 22 November 2016; Accepted 26 January 2017; Published 30 March 2017

Academic Editor: Xingming Zhao

The molecular circuits of cell cycle control serve as a key hub to integrate from endogenous and environmental signals into a robust biological decision driving cell growth and division. Dysfunctional cell cycle control is highlighted in a wide spectrum of human cancers. More importantly the mainstay anticancer treatment such as radiation therapy and chemotherapy targets the hallmark of uncontrolled cell proliferation in cancer cells by causing DNA damage, cell cycle arrest, and cell death. Given the functional importance of cell cycle control, the regulatory mechanisms that drive the cell division have been extensively investigated in a huge number of studies by conventional single-gene approaches. However the complexity of cell cycle control renders a significant barrier to understand its function at a network level. In this study, we used mathematical modeling through modern graph theory and differential equation systems. We believe our network evolution model can help us understand the dynamic cell cycle control in tumor evolution and optimizing dosing schedules for radiation therapy and chemotherapy targeting cell cycle.

1. Introduction

Cell growth and division are regulated by molecular circuits known as "cell cycle control," a coordinated protein-protein interacting network, that monitor cell proliferative signals, genome integrity, and proper timing of cell cycle transition from four different phases including S phase (DNA synthesis), M (mitosis), and two interphases (G1 and G2) between S and M phases [1]. A wide spectrum of biological pathways provides signaling inputs into molecular circuits of cell cycle control to determine how and when a single cell divides into two cells and also to ensure orderly cell cycle phase transition with high fidelity of cell duplication [1–3]. More specifically, the molecular circuits of cell cycle control are required by cells to respond to biological sensing systems including MAPK pathway, growth factor receptor pathways of EGFR, HER-2, and ErbB2-ErbB3, PI3K/AKT pathway, Wnt-β-catenin pathway, estrogen/androgen-mediated pathway, energy sensing pathway, metabolic pathway, and DNA damage response pathway (Figure 1(a)) [4–13]. Based on the signal inputs from these biological pathways, the molecular circuits of cell cycle control then generate decisive and robust signaling for cell growth. Thus, it is the key regulatory component to maintain cell homeostasis involved in a complex protein network (Figure 1(a)).

Given the functional importance of the molecular circuits of cell cycle control in integrating biological signals into cell growth decision, aberrant cell cycle control has been highlighted in the development of a variety of human diseases, particularly in human cancers, which contain a hallmark of uncontrolled cell proliferation [14–16]. For example, loss of a key cell cycle regulator p53 is found in more than 50% of human cancers [17, 18]. Overexpression of cyclin-dependent kinases (CDKs) and cyclin proteins (CCND1, CCNE1, and CCNB1) are found in many human cancers [19]. Overexpression of SKP2 which leads to reduced expression of negative cell cycle regulator CDKN1B is found in cancer cells as a bypass mechanism to escape cell cycle control [20].

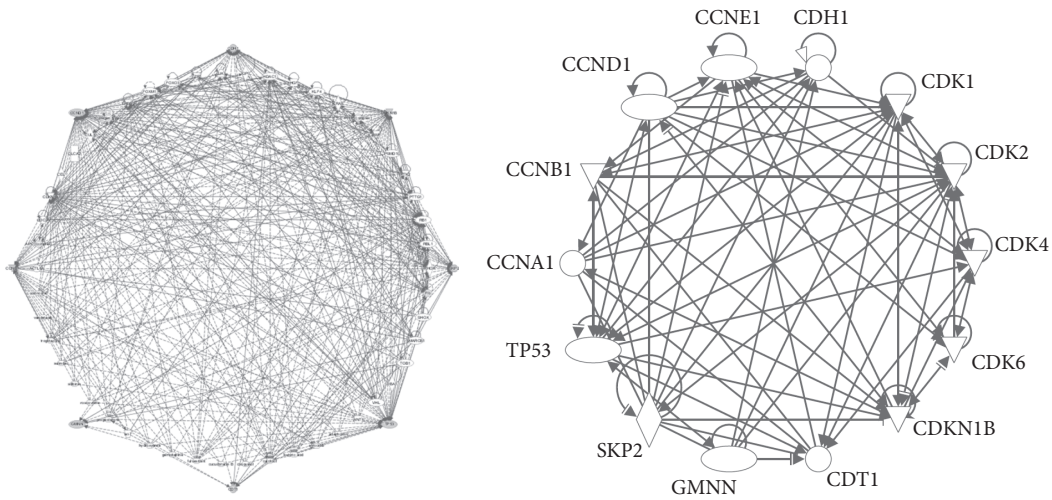


Figure 1: (a) *The complexity of cell cycle control*. IPA pathway analysis showed the regulatory protein network of cell cycle consists of a variety of biology signaling pathways and complex protein-protein interactions. (b) *A representative of protein-protein interaction network of cell cycle control*. Fourteen proteins involved in the cell cycle control were selected based on their key biological functions and relevance to human cancers.

The molecular circuits of cell cycle control are not only important for preventing the development of human cancers, but also important for determining treatment responses and toxicities to current chemotherapy and radiation therapy, most of which target cell proliferation and inhibit tumor growth [15, 21, 22]. For example, approximately 50% of cancer patients will receive radiation therapy, which induces DNA damage, arrests cell cycle progression, and leads to cell death. The efficacy of radiotherapy is largely affected by cell cycle control. Inhibition of cell cycle arrest signaling can leave cancer cells with less repair time and leads to a greater cell death to improve therapeutic responses [23]. The mainstay chemotherapeutic agents used in clinic, such as cyclophosphamide, cisplatin, 5-fluorouracil, gemcitabine, bleomycin, doxorubicin, etoposide, and topotecan/irinotecan, are targeting DNA as well [22]. They are also extremely toxic in normal tissues with the high proliferative rates such as epithelia of the gastrointestinal tract, hair follicle, and bone marrow. The selectivity of these agents between cancer and normal cells is largely determined by quantitative differences in the rates of cell division [22]. Thus, a better understanding of molecular circuits of cell cycle control will provide us with new insights into tumor evolution and anticancer treatments.

The regulatory mechanisms that control the cell cycle have been investigated in a huge number of studies. However, these studies often used conventional molecular biology approaches to dissect the function of each individual molecule. Because the molecular circuits of cell cycle control involve a variety of proteins and regulatory interactions, this biological complexity renders a great challenge in understanding the network impact of the cell cycle control by single-gene approaches. To address this challenge, a mathematical modeling of the molecular citrus of cell cycle control can be taken to simplify the complex biological circuits into a general framework for better analysis aimed at checking assumptions in addition to predicting.

Mathematical models correspond to conceptual representations that capture the essential features of the investigated process in the cell cycle and then omit details (i.e., elements that have negligible effects as well as elements that influence the explored behavior but are assumed as secondary properties) to describe its mechanisms. Mathematical models cast a process in the form of equations of a particular type to predict the system behavior and possibly suggest complementary experiments for a better understanding. The differential equations model is a continuous system in which the rates of change of the concentration of different states, such as genes, are related to the states in the state space either linearly or in a nonlinear setting [24]. Depending on the nature of the biological systems, different mathematical models based on ordinary differential equations were developed to study the progression of the system. Mamontov obtained sufficient conditions on nonautonomous ordinary differential equations that are capable of governing homeorhesis [25]. Dynamical modeling with differential equations has been shown to be effective in gaining insight of the cancer progression and response to the immunotherapy [26, 27]. The solution of the differential equations predicts the behavior of the biological system with much more details than the collected samples could reveal.

To establish molecular circuits of cell cycle control for mathematical modeling, we used QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, https://www.qiagenbioinformatics.com//ingenuity) to generate a 14-protein network, which involve key proteins in regulating cell cycle and with extremely high relevance in human cancers including cyclin proteins (CCND1, CCNE1, and CCN), CDKs (CDK1, 2, 4, and 6), SKP2, CDKN1B, p53, and CDH1. GMNN

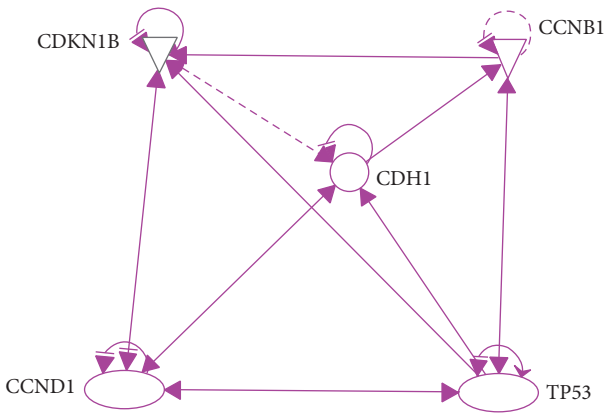


Figure 2: *A simplified network identified by IPA pathway*. Five key cell cycle regulators were selected to represent the complexity of protein-protein actions involved in the cell cycle regulation.

and CDT1, two well established markers for G2/M phase and G1 phase, are included (Figure 1(b)). To further achieve simplicity of the core 14-protein network, we used IPA analyses to represent the protein-interaction network by 5 proteins including CDKN1B, CCNB1, CDH1, CCND1, and p53, which serve as molecular hubs for the circuits of cell cycle control (Figure 2).

2. Background and Methodology

The basic introduction to modern graph theory and random walks was described in our previous publication [28]. For completeness we include some of the details here. A graph consists of nodes and edges where each edge connects two nodes. Edges in a directed graph are directed, in the sense that each edge goes from one vertex to another but not necessarily vice versa. In our model, a directed graph is constructed such that every protein is represented by a node and every protein-protein interaction is represented by a directed edge between the nodes corresponding to the proteins. For instance, if protein A regulates protein B, there is a directed edge from the vertex corresponding to A to the one corresponding to B. In addition, we add two artificial nodes: an initial node (S) and a transition node (T). The node S has directed edges to and from all existing nodes. The node T has directed edges from all existing nodes and a directed edge to S.

In a random walk, a random walker starts from any chosen node. At each step, the walker moves along the directed edges to a neighboring node with equal probabilities. That is, if a node A has directed edges to B, C, and D, the random walker, when at A, will move to each of B, C, and D with probability 1/3. At each node, the directed edge from it to the transition node serves as the chance of exiting the current network to external proteins. Also, the directed edge from the initial node serves as the chance of restarting this random walk, representing the impact from external proteins outside of this network. Clearly, the higher the probability of a node being reached from other nodes is, the more interference the corresponding protein receives from other proteins.

Suppose the directed graph has n nodes, after t steps, p_i denotes the probability of the random walker being at the ith node. The vector $P_t = (p_1, p_2, \ldots, p_n)$ is then the "state" after t steps. The sequence of P_t as t goes to infinity (i.e., the random walker keeps walking forever) forms a Markov chain. The states are also called the transition probabilities.

In order for such a Markov chain to converge, the corresponding graph must be "irreducible" [29] and "aperiodic" [30]. In simpler terms,

(I) the graph is "strongly connected"; that is, between every (ordered) pair of nodes there is a directed path;

(II) the greatest common divisor of all cycle lengths is 1.

We claim that both of these conditions are satisfied in our constructed network. First, the initial node S has a directed edge (hence, a directed path) to and from every other node in the graph. For any pair of nodes A and B, A $\rightarrow$ S $\rightarrow$ B $\rightarrow$ S $\rightarrow$ A provides the necessary directed paths from A to B and vice versa. Thus, the graph under consideration is strongly connected. Second, for any node A, the directed cycle A $\rightarrow$ T $\rightarrow$ S $\rightarrow$ A is of length 3. Also, for any directed edge $A \rightarrow B$, the directed cycle A $\rightarrow$ B $\rightarrow$ T $\rightarrow$ S $\rightarrow$ A is of length 4. Therefore, the greatest common divisor of all cycle lengths is 1.

The unique limit to which this Markov chain converges is, in other words, the unique vector P to which the transition probability P_t converges as t approaches infinity. This limit P is the unique "stationary probability" or "stationary distribution." Such a convergence indicates that if the random walking process goes on forever, the probability of each node (protein) being visited (i.e., being influenced through the network by other proteins) is a fixed value.

Using M to denote the adjacency matrix of the directed graph with edge weights corresponding to the probabilities (known as the transition matrix of this random walk), the stationary distribution can be directly determined by solving $PM = P$. In other words, an eigenvector corresponding to the eigenvalue 1, of the matrix M transposed.

We will use the vector P as a measure of how strongly each protein is performing in the network. As time changes, the variation of P with respect to each variable provides us with the necessary data to construct our differential equation model. We then use our model to predict the behavior of each protein in the network as well as the impact between the proteins.

First we construct our relation matrix from the 14-gene network in Figure 1(b), yielding Table 1 of their correlations, where an X in the ith row and jth column implies the fact that the ith gene regulates the jth gene, while Y denotes the binding relation between two genes. An entry labeled with X/Y simply means both regulation and binding relation exist between the two genes. We used the numerical labeling instead of gene names for concise presentation, while keeping the list of our labeling as in Table 2. With the addition of the transition node and restart node, we obtain a 16 by

16 adjacency matrix, from which our transition matrix is constructed as follows. Here we assume each outgoing edge from a chosen vertex is visited (by the random walker) with the same probability.

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\frac{1}{8} & 0 & 0 & 0 & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 & 0 & \frac{1}{8} & \frac{1}{8} \\
\frac{1}{10} & 0 & 0 & 0 & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & 0 & 0 & \frac{1}{10} & 0 & \frac{1}{10} & \frac{1}{10} \\
\frac{1}{13} & \frac{1}{13} & 0 & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & 0 & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & 0 & \frac{1}{13} & \frac{1}{13} \\
\frac{1}{11} & 0 & \frac{1}{11} & 0 & \frac{1}{11} & 0 & 0 & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & 0 & \frac{1}{11} & \frac{1}{11} \\
\frac{1}{9} & \frac{1}{9} & \frac{1}{9} & 0 & 0 & \frac{1}{9} & 0 & 0 & 0 & \frac{1}{9} & 0 & 0 & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\
\frac{1}{9} & \frac{1}{9} & \frac{1}{9} & 0 & \frac{1}{9} & 0 & 0 & 0 & \frac{1}{9} & \frac{1}{9} & 0 & \frac{1}{9} & 0 & 0 & \frac{1}{9} & \frac{1}{9} \\
\frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & 0 & 0 & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & 0 & 0 & \frac{1}{12} \\
\frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & 0 & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & 0 & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} \\
\frac{1}{13} & 0 & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & 0 & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & 0 & \frac{1}{13} \\
\frac{1}{11} & 0 & 0 & \frac{1}{11} & \frac{1}{11} & 0 & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & 0 & 0 & \frac{1}{11} \\
\frac{1}{8} & 0 & 0 & \frac{1}{8} & \frac{1}{8} & 0 & 0 & \frac{1}{8} & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 & \frac{1}{8} \\
\frac{1}{13} & 0 & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & 0 & 0 & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} \\
\frac{1}{9} & 0 & 0 & 0 & 0 & \frac{1}{9} & \frac{1}{9} & 0 & \frac{1}{9} & \frac{1}{9} & 0 & 0 & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\
\frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & 0 & 0 & 0 & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} \\
0 & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & \frac{1}{14} & 0
\end{bmatrix} \tag{1}$$

Starting with a probability distribution $(0, 1/14, \ldots, 1/14, 0)$ (i.e., evenly distributed among the 14 genes), repeatedly applying the transition matrix provides us the sequence of data associated with each of the five key genes CCNB1, CCND1, CDH1, CDKN1B, and TP53, whose impact on each other is shown in Figure 2. We denote their expressions by the variables x_1, x_2, x_3, x_4, and x_5, respectively. The convergence of each of these variables is shown in Figure 3. By projecting the differential of each key gene with respect to the neighboring genes (in Figure 2), we are able to model the evolution of these genes inside our original network through a system of differential equations, which enables us to predict and model the relations between individual pairs. The pseudo structure between them, together with the artificial transition and restarting points, is shown in Figure 4.

Based on the structure of the gene network and the collected data samples (i.e., the sequences of the expressions of each x_i when the hypothetical random walking process is applied), we propose the following system of differential equations that governs the evolution of the five-gene network:

$$\begin{aligned}
\dot{x}_1 &= r_{11}x_1 + r_{12}x_2 + r_{13}x_5, \\
\dot{x}_2 &= r_{21}x_1 + r_{22}x_2 + r_{23}x_3 + r_{24}x_1x_3, \\
\dot{x}_3 &= r_{31}x_2 + r_{32}x_3 + r_{33}x_4 + r_{34}x_5, \\
\dot{x}_4 &= r_{41}x_1 + r_{42}x_3 + r_{43}x_4 + r_{44}x_5, \\
\dot{x}_5 &= r_{51}x_2 + r_{52}x_3 + r_{53}x_4 + r_{54}x_5,
\end{aligned} \tag{2}$$

where x_1, x_2, x_3, x_4, and x_5 represent the genes CCNB1, TP53, CCND1, CDKN1B, and CDH1, respectively, in the gene network; r_{ij}'s are the system parameters, controlling the rate of change of their corresponding states. These parameters are computed through calibrating the system with the sampling

Table 1: Correlations between the 14 genes.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	O	O	O	O	Y	Y	Y	Y	Y	O	O	O	O	Y
2	O	O	O	Y	Y	Y	X/Y	X/Y	Y	O	O	X	O	Y
3	X	O	X/Y	X	X	O	X/Y	Y	X/Y	X/Y	X/Y	X/Y	O	Y
4	O	X/Y	O	X	O	O	Y	X/Y	X/Y	X/Y	X/Y	X	O	Y
5	Y	Y	O	O	X/Y	O	O	O	Y	O	O	Y	Y	Y
6	Y	Y	O	Y	O	O	O	Y	Y	O	Y	O	O	Y
7	X/Y	Y	X/Y	X/Y	O	O	X/Y	X/Y	X/Y	X/Y	X/Y	O	O	Y
8	Y	X/Y	Y	Y	O	X/Y	X/Y	X/Y	X/Y	X/Y	O	X/Y	Y	Y
9	O	Y	X/Y	Y	Y	X/Y	X/Y	O	X/Y	X/Y	Y	O	Y	X/Y
10	O	O	Y	X/Y	O	X/Y	Y	X/Y	X/Y	X/Y	X/Y	Y	O	O
11	O	O	Y	Y	O	O	X/Y	O	X/Y	X/Y	X/Y	O	O	O
12	O	X	X/Y	X	X/Y	X	X	X/Y	X/Y	X/Y	O	X/Y	X	O
13	O	O	O	O	X/Y	X/Y	O	Y	Y	O	O	X	Y	Y
14	Y	Y	Y	X/Y	Y	X/Y	X/Y	Y	X/Y	O	O	O	Y	X/Y

Table 2: Numerical labelling of the 14 genes.

Numerical labeling	Gene names
1	CCNA1
2	CCNB1
3	CCND1
4	CCNE1
5	CDH1
6	CDT1
7	CDKN1B
8	CDK1
9	CDK2
10	CDK4
11	CDK6
12	TP53
13	GMNN
14	SKP2

data of the state variables. The calibrating process includes a static stage followed by dynamic adaptation. In the static stage, the derivative samples are obtained from the data samples via a higher order finite difference method. The derivative samples along with the data samples are applied to (2) to optimally approximate the parameters by using the least squares method. This is carried out for each differential equation in (2). These newly computed parameters are applied to (2) for dynamic adaptation. At this stage, the dynamical system is simulated via the fourth-order Runge-Kutta method to produce the state trajectories. The state trajectories are plotted against the state samples for comparison. The discrepancies are reduced through transient-steady state compensation.

3. Results and Discussion

The simulation results of the differential equations system (2) against the collected sample data are shown in Figure 5. We compute the system parameters optimally by minimizing the errors throughout the transient and steady state stages. The numerical values of the parameters are listed in Table 3. Figure 5 shows the collection of state trajectories predicted by (2) along with the corresponding sample values. It is easy to see that the transient response and the steady state match up well with the collected samples.

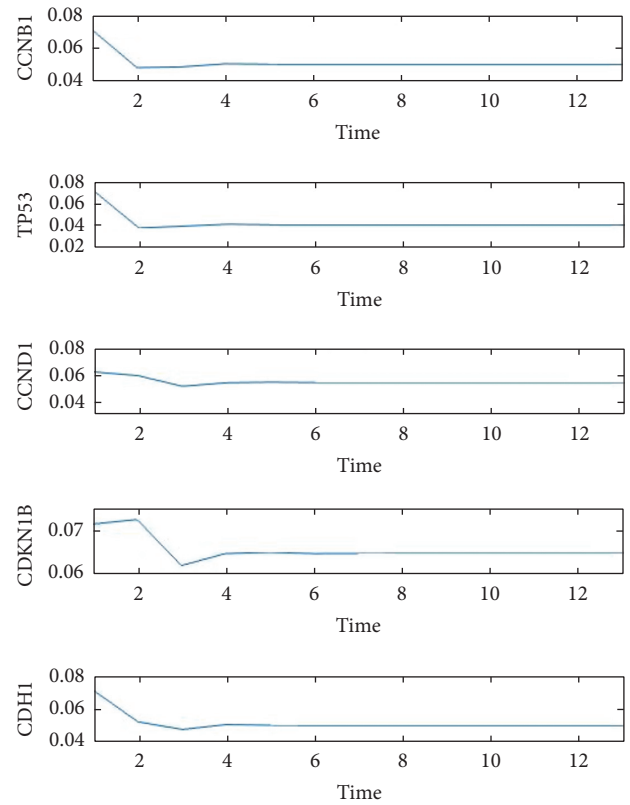


Figure 3: *Convergence of key gene expressions*. The expressions of the five key proteins CCNB1, CCND1, CDH1, CDKN1B, and TP53 are modeled through the random walk probability distribution, each converging to the stationary probability.

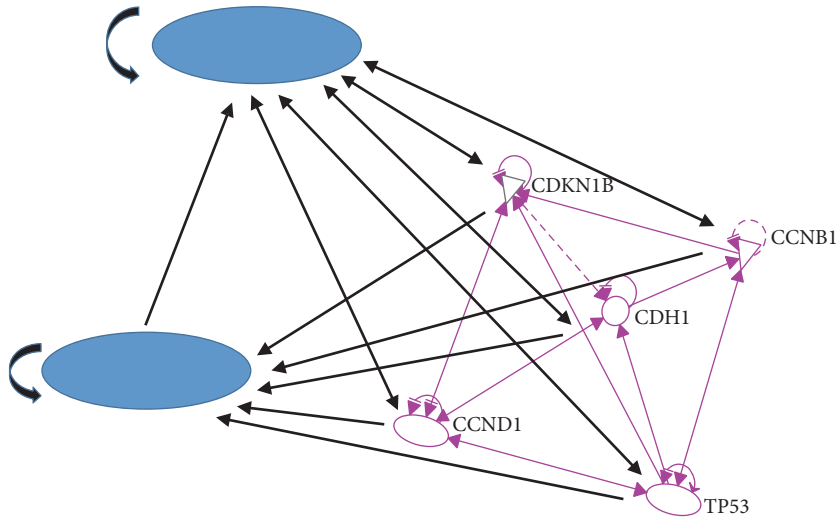


Figure 4: *Updated network with artificial nodes*. Two artificial nodes, the "transition" and "restarting" nodes, are added to the five-protein network to generate the directed graph for the random walk model.

Table 3: Determined system parameters.

r_{11}	−1.6525
r_{12}	0.56
r_{13}	1.22
r_{21}	0.21
r_{22}	−1.72
r_{23}	0.85
r_{24}	0.3919
r_{31}	0.32
r_{32}	−1.65
r_{33}	0.45
r_{34}	0.38
r_{41}	0.31
r_{42}	0.23
r_{43}	−1.62
r_{44}	0.16
r_{51}	0.24
r_{52}	0.12
r_{53}	0.11
r_{54}	−1.9

In this study, we generate a mathematical model to predict dynamics of molecular circuits of cell cycle control. Instead of studying each molecules involved in cell cycle regulation, we use a network evolution model to dissect how molecular circuits of cell cycle control function as a network to maintain homeostasis of cell growth signals.

We believe our approach can be applied to a variety of biological contexts to solve key clinical questions in cancer research. First, cancer dormancy is a stage in tumorigenesis where the cells stop dividing but survive while waiting for appropriate endogenous and environmental signals to reenter into cell cycle and proliferate again [31]. Cancer dormancy is associated with drug resistance, tumor recurrence, and metastasis. Thus our network evolution model might provide a new perspective to identify the difference at the network level of cell cycle control between dormant cancer cells and proliferative cancer cells. The results from such analyses may mechanistically explain how dormant cells achieve withdrawal and reentry of cell cycle and more importantly may identify druggable targets that can be used to develop antidormancy therapy to extend patient survival.

Secondly, this network evolution model of cell cycle control might provide us with a molecular tool to monitor dynamics of cell cycle transition, which can be used to optimize dosing schedules of cancer preventive and therapeutic drugs. Most of radiation therapy and chemotherapy are given to cancer patients with scheduled cycles. For example, typical dosing schedule of radiation therapy is 2 Gy per day, 5 days per week, for 6 weeks [23]. However, it remains open as what alternative schedules could be applied to improve treatment efficacy and reduce toxicity. Our network evolution model provides us with a possible approach to model the treatment responses of cells to radiation by monitoring the dynamics of cell cycle control, which will help us guide the experimental validations to achieve an optimized dosing schedule.

In addition to the study of cancer treatment, our model may also be helpful for designing better cancer prevention regimens for patients with high risks of cancers. For example, estrogen signaling is a major biological pathway driving cell cycle progression and cell proliferation. For women at high risks of breast cancer such as genetic predisposition or existing premalignant pathological changes, tamoxifen and raloxifene have been demonstrated by clinical trials with approved chemopreventive effects [32, 33]. It has been recommended to take these drugs for 5 years in the cancer prevention setting based on clinical experience. However this prolonged dosing schedule causes severe side effects, which lead to reluctance of women at high risks of breast cancer to take these drugs for cancer prevention [32–34].

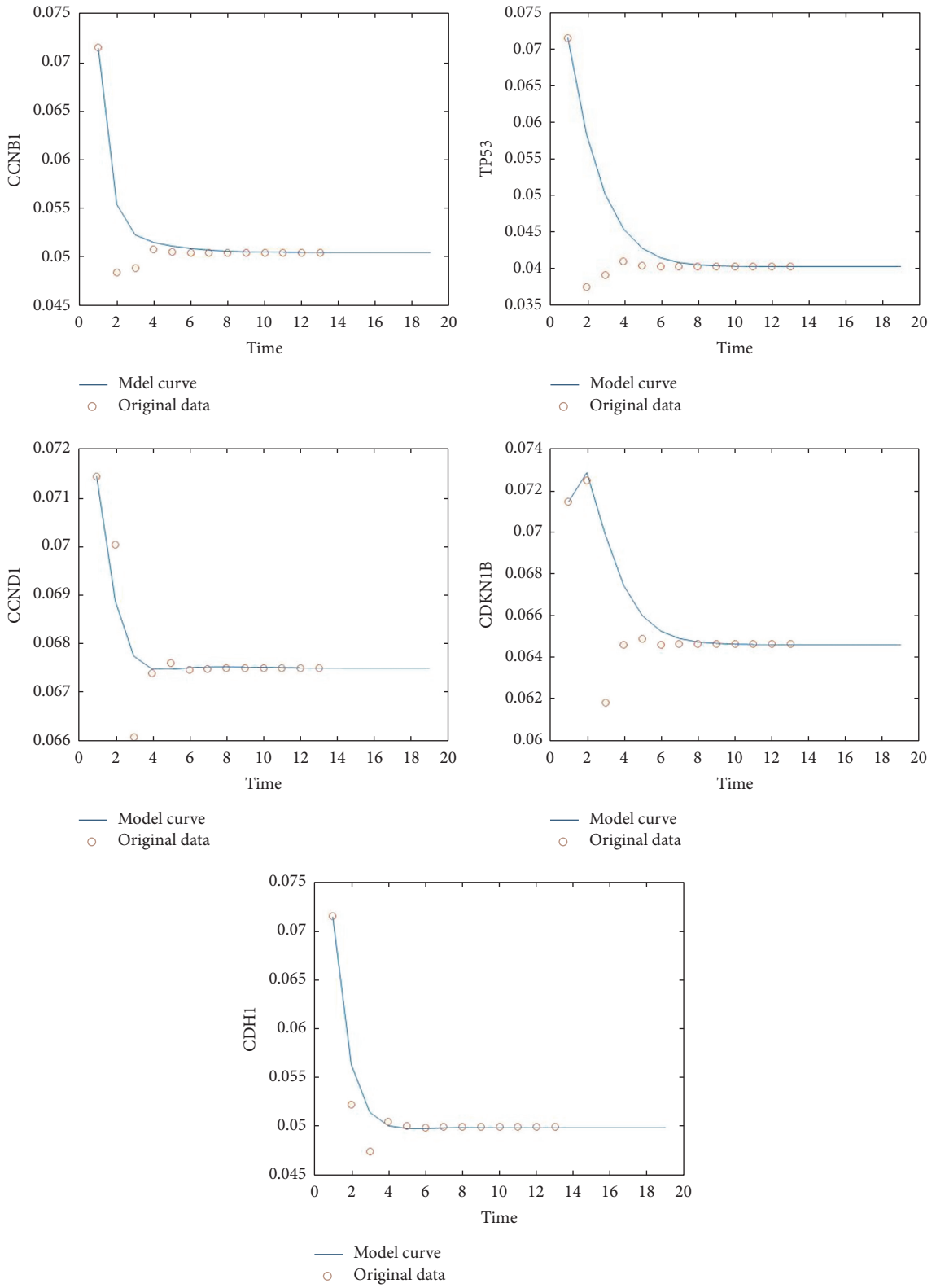


FIGURE 5: *Comparison between projected and sample data*. The state trajectories are plotted against the state samples for comparison. The discrepancies are reduced through transient-steady state compensation.

There is no method available to determine such an optimized dosing schedule with potential intermittent treatment in lieu of a 5-year continuous treatment. By applying our network evolution model, we may identify an optimal dosing schedule to conduct intermittent preventive treatment, which can reduce toxicity and improve efficacy [35]. Consistent with our findings, ordinary differential equation models have been widely used in cancer research to estimate tumor growth and anticancer treatment responses [35]. These studies demonstrate a proof of concept for using these models to simulate complex biological processes and interactions by developing simple quantitative models and also comparing experimental data.

In summary, we believe our interdisciplinary approaches may open a new avenue to study cell cycle control, which may better our understanding of tumor evolution and cancer prevention/therapy.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Yang Peng and Paul Scott contributed equally to this work.

Acknowledgments

This research is supported in part by the MD Anderson Cancer Center Support Grant CA016672 and by Susan G. Komen for the Cure Foundation CCR14300500 and NCI Grant CA181663 to Guang Peng and Simons Foundation Grant no. 245307.

References

[1] H. Harashima, N. Dissmeyer, and A. Schnittger, "Cell cycle control across the eukaryotic kingdom," *Trends in Cell Biology*, vol. 23, no. 7, pp. 345–356, 2013.

[2] E. Damiens, "Molecular events that regulate cell proliferation: an approach for the development of new anticancer drugs," *Progress in Cell Cycle Research*, vol. 4, pp. 219–233, 2000.

[3] R. J. Duronio and Y. Xiong, "Signaling pathways that control cell proliferation," *Cold Spring Harbor Perspectives in Biology*, vol. 5, no. 3, Article ID a008904, 2013.

[4] K. Al-Kuraya, H. Novotny, P. Bavi et al., "HER2, TOP2A, CCND1, EGFR and C-MYC oncogene amplification in colorectal cancer," *Journal of Clinical Pathology*, vol. 60, no. 7, pp. 768–772, 2007.

[5] F. Chang, J. T. Lee, P. M. Navolanic et al., "Involvement of PI3K/Akt pathway in cell cycle progression, apoptosis, and neoplastic transformation: A target for cancer chemotherapy," *Leukemia*, vol. 17, no. 3, pp. 590–603, 2003.

[6] D. R. Ciocca and M. A. Fanelli, "Estrogen receptors and cell proliferation in breast cancer," *Trends in Endocrinology and Metabolism*, vol. 8, no. 8, pp. 313–321, 1997.

[7] E. Cuyàs, B. Corominas-Faja, J. Joven, and J. A. Menendez, "Cell cycle regulation by the nutrient-sensing mammalian target of rapamycin (mTOR) pathway," *Methods in Molecular Biology*, vol. 1170, pp. 113–144, 2014.

[8] G. Davidson and C. Niehrs, "Emerging links between CDK cell cycle regulators and Wnt signaling," *Trends in Cell Biology*, vol. 20, no. 8, pp. 453–460, 2010.

[9] I. H. Lee and T. Finkel, "Metabolic regulation of the cell cycle," *Current Opinion in Cell Biology*, vol. 25, no. 6, pp. 724–729, 2013.

[10] V. W. Y. Lui and J. R. Grandis, "EGFR-mediated cell cycle regulation," *Anticancer Research*, vol. 22, no. 1, pp. 1–11, 2002.

[11] T. Samuel, H. O. Weber, and J. O. Funk, "Linking DNA damage to cell cycle checkpoints," *Cell Cycle*, vol. 1, no. 3, pp. 162–168, 2002.

[12] S. Toth-Fejel, J. Cheek, K. Calhoun, P. Muller, and R. F. Pommier, "Estrogen and androgen receptors as comediators of breast cancer cell proliferation: providing a new therapeutic tool," *Archives of Surgery*, vol. 139, no. 1, pp. 50–54, 2004.

[13] Z. Wei and H. T. Liu, "MAPK signal pathways in the regulation of cell proliferation in mammalian cells," *Cell Research*, vol. 12, no. 1, pp. 9–18, 2002.

[14] M. B. Kastan and J. Bartek, "Cell-cycle checkpoints and cancer," *Nature*, vol. 432, no. 7015, pp. 316–323, 2004.

[15] M. Malumbres and M. Barbacid, "Cell cycle, CDKs and cancer: a changing paradigm," *Nature Reviews Cancer*, vol. 9, no. 3, pp. 153–166, 2009.

[16] B. Zhivotovsky and S. Orrenius, "Cell cycle and cell death in disease: past, present and future," *Journal of Internal Medicine*, vol. 268, no. 5, pp. 395–409, 2010.

[17] T. Hamzehloie, M. Mojarrad, M. Hasanzadeh-Nazarabadi, and S. Shekouhi, "The role of tumor protein 53 mutations in common human cancers and targeting the murine double minute 2-P53 interaction for cancer therapy," *Iranian Journal of Medical Sciences*, vol. 37, no. 1, pp. 3–8, 2012.

[18] K. T. Bieging, S. S. Mello, and L. D. Attardi, "Unravelling mechanisms of p53-mediated tumour suppression," *Nature Reviews Cancer*, vol. 14, no. 5, pp. 359–370, 2014.

[19] A. Deshpande, P. Sicinski, and P. W. Hinds, "Cyclins and cdks in development and cancer: a perspective," *Oncogene*, vol. 24, no. 17, pp. 2909–2915, 2005.

[20] D. D. Hershko, "Oncogenic properties and prognostic implications of the ubiquitin ligase Skp2 in cancer," *Cancer*, vol. 112, no. 7, pp. 1415–1424, 2008.

[21] G. K. Schwartz and M. A. Shah, "Targeting the cell cycle: a new approach to cancer therapy," *Journal of Clinical Oncology*, vol. 23, no. 36, pp. 9408–9421, 2005.

[22] B.-B. S. Zhou and J. Bartek, "Targeting the checkpoint kinases: chemosensitization versus chemoprotection," *Nature Reviews Cancer*, vol. 4, no. 3, pp. 216–225, 2004.

[23] A. C. Begg, F. A. Stewart, and C. Vens, "Strategies to improve radiotherapy with targeted drugs," *Nature Reviews Cancer*, vol. 11, no. 4, pp. 239–253, 2011.

[24] H. V. D. Berg, *Mathematical Models of Biological Systems*, Oxford Biology, Oxford University Press, London, UK, 2011.

[25] E. Mamontov, "Modelling homeorhesis by ordinary differential equations," *Mathematical and Computer Modelling*, vol. 45, no. 5-6, pp. 694–707, 2007.

[26] L. G. de Pillis, W. Gu, and A. E. Radunskaya, "Mixed immunotherapy and chemotherapy of tumors: modeling, applications and biological interpretations," *Journal of Theoretical Biology*, vol. 238, no. 4, pp. 841–862, 2006.

[27] S. Wilson and D. Levy, "A mathematical model of the enhancement of tumor vaccine efficacy by immunotherapy," *Bulletin of Mathematical Biology*, vol. 74, no. 7, pp. 1485–1500, 2012.

[28] H. Wang and G. Peng, "Mathematical model of dynamic protein interactions regulating p53 protein stability for tumor suppression," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 358980, 2013.

[29] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1999.

[30] B. Bollabos, *Modern Graph Theory*, Springer, New York, NY, USA, 1998.

[31] B. Quesnel, "Tumor dormancy: long-term survival in a hostile environment," *Advances in Experimental Medicine and Biology*, vol. 734, pp. 181–200, 2013.

[32] S. A. Nazarali and S. A. Narod, "Tamoxifen for women at high risk of breast cancer," *Breast Cancer: Targets and Therapy*, vol. 6, pp. 29–36, 2014.

[33] L. S. Donnelly, D. G. Evans, J. Wiseman et al., "Uptake of tamoxifen in consecutive premenopausal women under surveillance in a high-risk breast cancer clinic," *British Journal of Cancer*, vol. 110, no. 7, pp. 1681–1687, 2014.

[34] A. Howell, A. S. Anderson, R. B. Clarke et al., "Risk determination and prevention of breast cancer," *Breast Cancer Research*, vol. 16, no. 5, article 446, 2014.

[35] X. Wu and S. M. Lippman, "An intermittent approach for cancer chemoprevention," *Nature Reviews Cancer*, vol. 11, no. 12, pp. 879–885, 2011.

Hindawi
BioMed Research International
Volume 2017, Article ID 5346793, 4 pages
https://doi.org/10.1155/2017/5346793

Research Article

COPAR: A ChIP-Seq Optimal Peak Analyzer

Binhua Tang,[1,2] Xihan Wang,[1] and Victor X. Jin[3]

[1]*Epigenetics & Function Group, School of Internet of Things, Hohai University, Jiangsu 213022, China*
[2]*School of Public Health & Biostatistics, Shanghai Jiao Tong University, Shanghai 200025, China*
[3]*Department of Molecular Medicine & Biostatistics, University of Texas Health Science Center, San Antonio, TX 78249, USA*

Correspondence should be addressed to Binhua Tang; bh.tang@outlook.com

Received 28 October 2016; Accepted 14 February 2017; Published 5 March 2017

Academic Editor: Xingming Zhao

Sequencing data quality and peak alignment efficiency of ChIP-sequencing profiles are directly related to the reliability and reproducibility of NGS experiments. Till now, there is no tool specifically designed for optimal peak alignment estimation and quality-related genomic feature extraction for ChIP-sequencing profiles. We developed open-sourced COPAR, a user-friendly package, to statistically investigate, quantify, and visualize the optimal peak alignment and inherent genomic features using ChIP-seq data from NGS experiments. It provides a versatile perspective for biologists to perform quality-check for high-throughput experiments and optimize their experiment design. The package COPAR can process mapped ChIP-seq read file in BED format and output statistically sound results for multiple high-throughput experiments. Together with three public ChIP-seq data sets verified with the developed package, we have deposited COPAR on GitHub under a GNU GPL license.

1. Introduction

Next-generation sequencing (NGS) integrated with ChIP technology provides a genome-wide perspective for biomedical research and clinical diagnosis applications [1–3].

Data quality and peak alignment of ChIP-sequencing profiles are directly related to the reliability and reproducibility of analysis results. For example, ChIP-seq data characterize alteration evidence for transcription factor (TF) binding activities in response to chemical or environmental stimuli, but if the ChIP-seq alignment is poorly selected, any follow-up analysis may lead to inaccurate TF binding results and inevitable loss of biological meanings [4, 5].

The mostly investigated items in ChIP-seq peak calling procedures are peak number, false discovery rate (FDR), corresponding bin-size, and other statistical thresholds selected in each analysis. Without exception, such arguments form impenetrable barriers for biologists and bioinformaticians to choose a suitable pair condition for analyzing experimental results.

And to our knowledge, few literatures or application notes focus on such topics; thus herein we propose a flexible package based on feature extraction and signal processing algorithms for solving such an argument-selection optimization problem in optimal peak alignment.

In summary, the package COPAR can quantitatively measure NGS/ChIP-seq experiment quality through global peak alignment comparison and extract genomic features based on spectrum method for in-depth analysis of ChIP-sequencing profiles.

2. Materials and Methods

2.1. Optimal Peak Alignment Estimation. For determining optimal ChIP-seq alignment, we need to analyze peak numbers under specific argument constraints. Thus we acquire optimal peak numbers by constraining specific arguments, which can be formalized as a class of optimal track analysis, illustrated as

$$\begin{aligned} \arg\max_{i} \quad & P_i, \quad i \in N \\ \text{s.t.} \quad & f_i \le \chi, \\ & b_i = \beta, \\ & p_i \le \delta, \end{aligned} \tag{1}$$

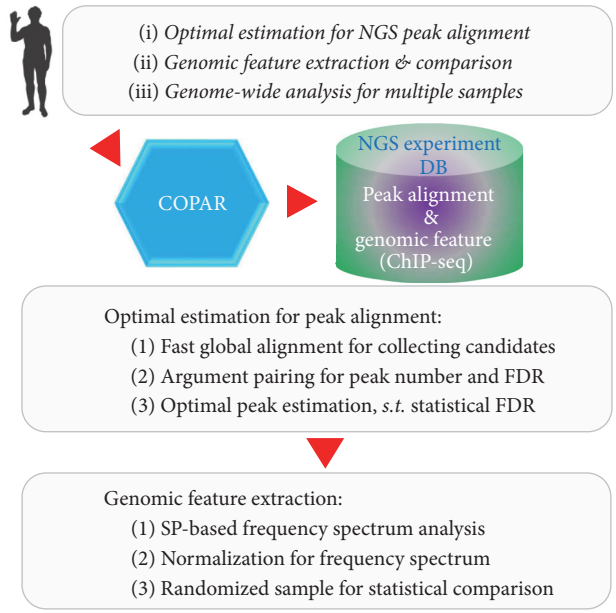


Figure 1: Flowchart for optimal peak alignment estimation and genomic feature analysis with COPAR. The package can perform optimal peak estimation based on global alignment of ChIP-seq data; then it can utilize the frequency spectrum approach for genomic feature extraction and carries out statistical comparison for multiple ChIP-seq samples.

where P_i denotes a set of optimal peak numbers under corresponding argument constraints, f_i stands for argument FDR, b_i stands for bin-size, p_i denotes p value threshold, and χ, β, and δ represent the presupposed argument values, respectively.

2.2. Spectrum-Based Genomic Feature Extraction. For a finite random variable sequence, its power spectrum is normally estimated from its autocorrelation sequence by use of discrete-time Fourier transform (DTFT), denoted as [6–8]

$$P(\omega) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} C_{xx}(n)\, e^{-jn\omega}, \tag{2}$$

where C_{xx} denotes autocorrelation sequence of a discrete signal x_n, defined as

$$C_{xx}(i,j) = \frac{E\left[\left(X_i - \mu_i\right)\left(X_j - \mu_j\right)\right]}{\sigma_i \sigma_j}, \tag{3}$$

where μ and σ stand for mean and variance, respectively.

In our study, for consideration of the ChIP-seq data characteristics, we use 128 sampling points to calculate discrete Fourier transform, with the related sampling frequency 1 KHz.

3. Results

The COPAR package was developed and open-sourced for academic biologists, and it uses built-in functions for determining optimal peak alignment candidate and extracting genomic features from ChIP-seq dataset.

The package is designed to handle BED-formatted ChIP-seq data as input [9], and it can process single ChIP-seq for optimal peak alignment and feature extraction analysis, together with the capability to perform genome-wide statistical comparison for multiple ChIP-seq samples. The analysis flowchart for the package is given in Figure 1.

It can automatically determine the optimal peak alignment with statistically meaningful FDR through fast global alignment comparison; the global comparison is subject to two statistical arguments, namely, bin-size and p value threshold.

The functionalities of our developed package are largely complementary to and extend current tools used for ChIP-seq data analysis. The optimal peak alignment estimation is shown in Figures 2(a) and 2(b); and the spectrum-based feature extraction is given in Figures 2(c) and 2(d). Figures 2(a) and 2(b) utilize heatmap to represent peak number and corresponding FDR candidate subject to each argument pair, bin-size (vertical axis), and p value threshold (horizontal axis), respectively; Figure 2(c) denotes the spectrum distribution of the global peak alignment candidate sequence, normalized with its frequency range $[0, 500]$ Hz and magnitude within $[-40, -3]$ dB; Figure 2(d) denotes the randomized case.

4. Conclusions

Based on global peak alignment, COPAR optimizes the argument selection in ChIP-seq analysis; meanwhile, COPAR utilizes the signal spectrum processing method to further extract genomic features and statistically compare multiple ChIP-seq samples for NGS high-throughput experiments.

In summary, our developed package COPAR can process mapped read file in BED format and output statistically sound

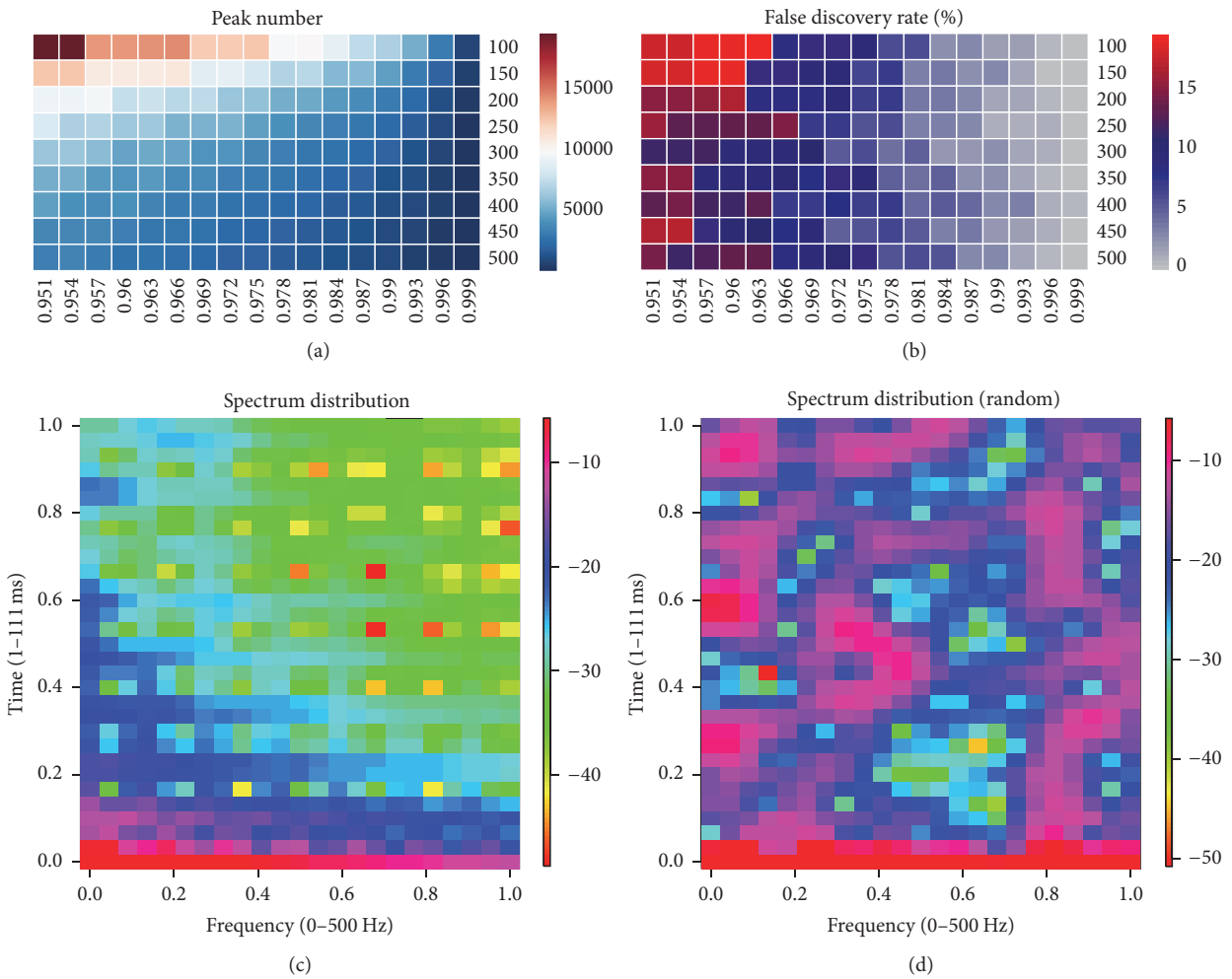


Figure 2: *Global optimal peak analysis result subject to the arguments bin-size and FDR.* (a) Global distributions for peak number candidates and (b) corresponding false discovery rate, subject to bin-size (vertical axis, from 100 through 500 bp) and p value threshold (horizontal axis, from 0.951 to 0.999), respectively; (c) genomic feature extraction based on spectrum distribution for global peak number candidates identified from COPAR; (d) spectrum distribution for the randomized sequence.

results for diverse high-throughput sequencing experiments; we further verified the package with three GEO ChIP-seq datasets as study cases, and we included the analysis results into the package manual. The developed package COPAR is currently available under a GNU GPL license from https://github.com/gladex/COPAR.

Abbreviations

NGS: Next-generation sequencing
ChIP-seq: Chromatin immunoprecipitation-sequencing
FDR: False discovery rate
TF: Transcription factor
DTFT: Discrete-time Fourier transform.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Binhua Tang and Victor X. Jin conceived the method; Binhua Tang and Xihan Wang wrote and compiled the package; Binhua Tang, Xihan Wang, and Victor X. Jin drafted and proof-checked the manuscript.

Acknowledgments

This work has been supported by the Natural Science Foundation of Jiangsu, China (BE2016655 and BK20161196), Fundamental Research Funds for China Central Universities (2016B08914), and Changzhou Science & Technology Program (CE20155050). This work made use of the resources supported by the NSFC-Guangdong Mutual Funds for Super Computing Program (2nd Phase) and the Open Cloud Consortium- (OCC-) sponsored project resource, supported in part by grants from Gordon and Betty Moore Foundation

and the National Science Foundation (USA) and major contributions from OCC members.

References

[1] E. R. Mardis, "ChIP-seq: welcome to the new frontier," *Nature Methods*, vol. 4, no. 8, pp. 613–614, 2007.

[2] G. J. Martinez and A. Rao, "Cooperative transcription factor complexes in control," *Science*, vol. 338, no. 6109, pp. 891–892, 2012.

[3] H. Kilpinen and J. C. Barrett, "How next-generation sequencing is transforming complex disease genetics," *Trends in Genetics*, vol. 29, no. 1, pp. 23–30, 2013.

[4] M. D. Chikina and O. G. Troyanskaya, "An effective statistical evaluation of chipseq dataset similarity," *Bioinformatics*, vol. 28, no. 5, pp. 607–613, 2012.

[5] T. S. Furey, "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions," *Nature Reviews Genetics*, vol. 13, no. 12, pp. 840–852, 2012.

[6] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, Upper Saddle River, NJ, USA, 3rd edition, 2010.

[7] B. Tang, H.-K. Hsu, P.-Y. Hsu et al., "Hierarchical modularity in ERα transcriptional network is associated with distinct functions and implicates clinical outcomes," *Scientific Reports*, vol. 2, article 875, 2012.

[8] S.-L. Wang, Y.-H. Zhu, W. Jia, and D.-S. Huang, "Robust classification method of tumor subtype by using correlation filters," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 2, pp. 580–591, 2012.

[9] X. Lan, R. Bonneville, J. Apostolos, W. Wu, and V. X. Jin, "W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data," *Bioinformatics*, vol. 27, no. 3, pp. 428–430, 2011.

Hindawi
BioMed Research International
Volume 2017, Article ID 4585213, 18 pages
https://doi.org/10.1155/2017/4585213

Research Article

MicroRNA Mediating Networks in Granulosa Cells Associated with Ovarian Follicular Development

Baoyun Zhang,[1] Long Chen,[1] Guangde Feng,[2] Wei Xiang,[1] Ke Zhang,[1] Mingxing Chu,[3] and Pingqing Wang[1]

[1]*Bioengineering Institute of Chongqing University, Chongqing, China*
[2]*Sichuan TQLS Animal Husbandry Science and Technology Co., Ltd., Mianyang, China*
[3]*Key Laboratory of Farm Animal Genetic Resources and Germplasm Innovation of Ministry of Agriculture, Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China*

Correspondence should be addressed to Pingqing Wang; wang_pq@21cn.com

Received 14 October 2016; Revised 21 November 2016; Accepted 23 November 2016; Published 19 February 2017

Academic Editor: Kang Ning

Ovaries, which provide a place for follicular development and oocyte maturation, are important organs in female mammals. Follicular development is complicated physiological progress mediated by various regulatory factors including microRNAs (miRNAs). To demonstrate the role of miRNAs in follicular development, this study analyzed the expression patterns of miRNAs in granulosa cells through investigating three previous datasets generated by Illumina miRNA deep sequencing. Furthermore, via bioinformatic analyses, we dissected the associated functional networks of the observed significant miRNAs, in terms of interacting with signal pathways and transcription factors. During the growth and selection of dominant follicles, 15 dysregulated miRNAs and 139 associated pathways were screened out. In comparison of different styles of follicles, 7 commonly abundant miRNAs and 195 pathways, as well as 10 differentially expressed miRNAs and 117 pathways in dominant follicles in comparison with subordinate follicles, were collected. Furthermore, SMAD2 was identified as a hub factor in regulating follicular development. The regulation of miR-26a/b on *smad2* messenger RNA has been further testified by real time PCR. In conclusion, we established functional networks which play critical roles in follicular development including pivotal miRNAs, pathways, and transcription factors, which contributed to the further investigation about miRNAs associated with mammalian follicular development.

1. Introduction

The mammalian ovary is a dynamic organ. The coordination of follicle recruitment, selection, and ovulation and the timely development are essential for a functional ovary and fertility [1]. Follicular development is a highly accurate, orchestrated, and periodic process which starts with the activation of resting follicles gradually leading to the growth and selection of dominant follicles (DFs) from small health follicles accompanied with sequential and profound differentiation of oocyte and the surrounding somatic cells [2], especially granulosa cells (GCs) [3]. Understanding the molecular mechanism of follicular development is essential for unraveling the complex synergies orchestrated during the process of forming the fertilizable ovum. In an estrous cycle, small follicular growth is characterized by 2 or 3 successive follicular waves which coincide with the luteinizing hormone- (LH-) surge waves. During the waves, a single follicle is selected, normally the largest (occasionally the largest two), which continuously grows as a DF while the others, referring to subordinate follicles (SFs), terminate development and undergo atresia [4]. Both of the two kinds of follicles are named large healthy follicles. The complex transition from primordial follicles to mature follicles is due to the functional differentiation and morphological transformation of GCs. In this crucial period, the growth of oocyte depends on the bidirectional communication between oocyte and GCs [5, 6]. Hence, the proliferation, apoptosis, and remarkable functional differentiation of GCs are significant events and required for follicle maturation [7]. Each of these development steps involves

significant changes of follicular structure and function, requiring accurate and coordinated adjustments to genes which have key roles in follicular selection, maturation, or the follicle-luteal transition. Any dysregulation in the expression of these specific genes would be critical in determining the fate of DFs or SFs [8–11]. In previous studies, transcriptome analyses have identified some genes involved in follicular growth, selection, and maturation [12–14]. However, the molecular regulatory mechanisms at differential levels are still unclear.

Recently, the posttranscriptional regulation dominated by miRNAs has attracted extensive attention. miRNAs regulate gene expression via the combination of seed sequence and the 3′-untranslated region (UTR) of target mRNAs, causing repression of translation or degradation of the target mRNAs during cell growth and differentiation [15]. The expression of miRNA in the ovary varies with cell type, function, and stage of the estrous cycle. miRNAs are involved in the formation of primordial follicles, follicular recruitment and selection, follicular atresia, oocyte-cumulus cell interaction, and GC function [1]. Profiling studies of miRNA in ovarian tissues have described the expression of miRNAs in the ovaries of various species [16–20]. Conditional knockout (cKO) of Dicer1 from follicular GCs resulted in a number of ovarian functional defects including abnormal oocyte maturation, disrupted follicular development and ovulation, increased follicular atresia, and infertility [21–23]. A single miRNA could regulate follicle development during estrus cycle via a canonical pathway [24], in which the target gene of miRNA plays an important role. Unsurprisingly, many miRNAs also could target transcription factors (TFs), such as TGF-β superfamily members, follicle stimulating hormone receptor (FSHR), and luteinizing hormone receptor (LHR), which have been confirmed to have a connection with follicular development. The abnormality of these molecules also led to dysfunction of cellular communication and dysregulation of normal follicle development and recruitment [25–29]. Moreover, several studies have demonstrated that the functions of specific miRNAs are implicated in different aspects of GC processes [30] such as proliferation [31–34], survival [35–37], terminal differentiation [38], steroidogenesis [31, 33, 39, 40], and cumulus expansion [41] in mammals. For example, miRNA-224 has been proved to be involved in transforming growth factor-beta-mediated mouse GCs proliferation and GC function by targeting *smad4* [31]. These results present evidence that miRNA might be involved in the selection of the DFs, the mechanism of which has remained largely elusive. The complex nature of miRNA target interaction, regulation, and function, however, posed challenges for functional studies. Whereas some miRNAs appear single-handedly to regulate specific signaling pathways, most miRNAs act in clusters and are fine tuners of cellular functions. In most of previous studies, results just demonstrated the function of a single miRNA in GCs. In fact, performing significant regulation of any biological process often results from complex regulating networks rather than one single miRNA. Therefore, an understanding of the mechanism of action of miRNA in follicular development requires global comprehension of the network of miRNA target interactions within the milieux of other factors that dominate follicular development. Hence, profiling studies are important in order to not only draw a spatial and temporal map of miRNA expression in different follicles, but also provide clues with regard to the function or regulation of miRNA.

In this study, according to 3 underused sequencing datasets (GSE56002, GSE55987, and GSE54692) from GCs of bovine, we attempted to systematically dissect the complex synergistic regulations of several functional miRNAs rather than a single miRNA in follicular development [42], based on the networks of miRNAs-signal pathways and miRNAs-TFs. Abundant miRNAs and differentially expressed miRNAs were identified and networks of miRNAs-signal pathways and miRNAs-TFs were constructed based on the correlation between miRNAs and predicted target genes. Moreover, by calculating the degree of every node in networks, hub players were identified to establish a centrical network, from which the critical miRNAs, pathways, and TFs in the process of follicular development were identified. The regulation of the significant miRNAs on hub genes would be further verified by RT PCR. The results might provide novel insights into revealing the potential mechanism of molecular regulation in follicular development in the context of miRNA synergistic regulatory networks.

2. Materials and Methods

2.1. Differential Expression of miRNAs in GCs and Data Normalization. A total of 64 differentially expressed miRNAs in GCs were collected from the miRNA sequencing results of Samuel Gebremedhn's research group (GSE56002) [24] and 52 ones from the study of another group, Salilew-Wondim et al. who also utilized Illumina small RNAseq (GSE55987) [47]. Gebremedhn et al. used miRDeep 2.0.0.5 software package and DESeq2 with a "hypothetical reference" to screen differentially expressed miRNAs in 6 ovarian follicle samples (3 biological replicates from DFs and others from SFs) at day 19 of the estrous cycle, in which the samples with external surface diameter $\geq$ 12 mm were recognized as DFs while the ones $\leq$ 11 mm as SFs. The study followed several cut-off criteria, adjusted p value $\leq$ 0.05, log2 fold change $\geq$ 1, and false discovery rate (FDR) $\leq$ 0.1 [24]. Similarly, Salilew-Wondim et al. employed the analogous software and methods to analyze the differential expression of miRNAs under the same criteria, which were obtained from 12 granulosa samples (three biological replicates of GCs from SFs or DFs at day 3 or day 7 of the estrous cycle) [47]. Briefly, in this study, follicles with follicular diameter of 6 mm (n = 43) were classified as SFs, while follicles with a diameter of 8–10 mm (n = 9) were considered as DFs at day 3 of the estrous cycle. On the other hand, at day 7 of the estrous cycle, follicles with a diameter $\leq$ 8 mm (n = 58) were considered as SFs while those with 9–13 mm diameter (n = 3) were categorized as DFs. Although there are some biological differences between day 3 and day 7, this did not exactly influence our analysis since what we considered was the whole process of DFs and SFs. By comparison, abundantly expressed or dysregulated miRNAs in both datasets were dissected in this study. Meanwhile, another dataset (GSE54692), which utilized microarray to

assess differentially expressed miRNAs in GCs of bovine, was investigated in the same manner, and the miRNAs meeting the same criteria mentioned above were selected [48]. In summary, after comparison of small and DFs, we firstly obtained 17 dysregulated miRNAs from the two types of follicles. Then, 57 dysregulated miRNAs were identified between large atretic follicles and DFs. Finally, a total of 15 miRNAs which played key roles in both growth and selection phase of DFs were further detected and analyzed in this paper.

2.2. miRNA Target Gene Prediction and Relevant Functional Analysis by GO and Pathway Enrichment. There were no online databases to predict the target genes of miRNAs in bovine. In this context, through sequence alignment, we confirmed that there was homology of miRNAs between human and bovine, and since the prediction of target genes was based on algorithm of sequence, this study analyzed patterns on the regulating functions of miRNAs in bovine follicular development through their target prediction along with human species. The potential target genes of differentially expressed miRNAs in GCs were acquired from the widely used online databases TargetScanHuman 7.0 (http://www.targetscan.org/) [49] with conservation (aggregate $P_{CT} > 0.8$) and context score (<−0.4 and percentile > 85%) and DIANA-microT (http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=microT_CDS/index) [50]. The aggregate P_{CT} is calculated as

$$P_{CT} = 1 - \left(\left(1 - P_{CT}\right)_{site1} \left(1 - P_{CT}\right)_{site2} \left(1 - P_{CT}\right)_{site3} \cdots\right). \quad (1)$$

In order to reduce false positives, the predicted target genes which appeared at both databases were accepted. Thus, the collection of predicted target genes of each obtained miRNA was imported into DIANA-mirPath (http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=mirpath/index), a miRNA-pathway analysis web server. Each list of canonical pathways significantly affected by differentially expressed miRNAs was made a contrast with another list and we obtained the common pathways which were as targets of dysregulated miRNAs. All the following canonical pathways were identified from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) databases.

As a comprehensive set of functional annotation tools, DAVID (the Database for Annotation, Visualization and Integrated Discovery) has been used for integrative and systematic analysis of enormous gene lists [51]. GO terms are significantly overrepresented in a set of genes from three aspects, namely, the biological process, cellular component, and molecular function. In this study, the key GO biological process terms of the predicted target genes of miRNAs were performed using DAVID with the thresholds of enrichment gene count > 2 and p value < 0.05.

2.3. miRNA-Pathway Network and miRNA-TF Network Construction. Study on pathways associated with dysregulated miRNAs was carried out using DIANA-mirPath. The results were integrated to get the intersection from those meeting p value threshold (Benjamini and Hochberg's FDR was applied with significant threshold set at p value ≤ 0.05) and microT threshold (0.8 of the score). Then the interactions between miRNA and pathways were collected to construct miRNA-pathway network. Transcription factors that can be involved in control of particular ovarian functions were screened from Animal Transcription Factor Database (http://www.bioguo.org/AnimalTFDB/BrowseAllTF.php?spe=Homo_sapiens) and Sirotkin's study [52]. In order to improve veracity, target genes of miRNAs were predicted via different software. TFs which might be regulated by miRNAs were also identified. Then the selected miRNAs-TFs database was used for constructing network by Cytoscape software under the similar protocol [53]. Furthermore, through calculating the degree of each node, hub TFs were selected to construct the core network.

2.4. Cell Culture and Transfection. A steroidogenic human granulosa-like tumor cell line (KGN) was undifferentiated and maintained the physiological characteristics of ovarian cells. The KGN cells were cultured in DMEM basic (1x)/high glucose medium (Gibco, Life Technologies, Carlsbad, CA, USA) containing 12% fetal bovine serum (FBS; Gibco, Australia) and 1% antibiotics (100 U/ml penicillin and 100 μg/ml streptomycin; Sigma) in a humidified incubator at 37.0°C with 5% CO_2. For human granulosa cells culture, the GCs were first purified as described previously [54]. GCs were cultured at a final concentration of 5×10^5 viable cells/ml culture medium, in 12-well plate at 37.0°C and 5% CO_2. After 24 h, the medium was replaced with RPMI only; then 4 μg miR-26a/b mimics, inhibitors, or miRNA expression vectors pSUPER-miR-26a/b [55] (with a negative control) were transfected using Lipofectamine 2000. The sequence of miR-26a/b mimics and miR-26a/b inhibitors, as well as their negative controls (Genepharma, Shanghai, China), are as follows:

miR-26a mimics: 5′-UUCAAGUAAUCCAGGAUAGGCU-3′

miR-26a inhibitors: 5′-AGCCUAUCCUGGAUUACUUGAA-3′

miR-26b mimics: 5′-UUCAAGUAAUUCAGGAUAGGU-3′

miR-26a inhibitors: 5′-ACCUAUCCUGAAUUACUUGAA-3′

Mimics negative control: 5′-UUCUCCGAACGUGUCACGUTT-3′

Inhibitors negative control: 5′-CAGUACUUUUGUGUAGUACAA-3′

After 8 h, the cells were subsequently treated with complete medium followed by extraction of total RNA from the cultured cells using Trizol (TaKaRa) at 48 h later.

2.5. Real Time PCR Assay. For RT PCR assays, the cDNA was synthesized from 1000 ng of purified RNA using the

Table 1: Fold changes of differentially expressed miRNAs in DFs compared with both small healthy follicles and SFs.

	miRNA ID	DFs versus SHFs		DFs versus SFs	
		Fold change	Adj. p value	Fold change	Adj. p value
Downregulated miRNAs	miR-3178	−3.60	0.002	−3.36	0.001
	miR-1275	−2.29	0.020	−2.28	0.003
	miR-625-3p	−2.21	0.006	−4.94	0.001
	miR-3621	−2.17	0.002	−2.21	0.001
	miR-483-3p	−2.17	0.002	−3.64	0.001
	miR-1469	−2.15	0.014	−5.11	0.001
	miR-498	−2.14	0.003	−3.69	0.001
	miR-4279	−2.05	0.012	−2.96	0.001
Upregulated miRNAs	miR-202	4.18	0.002	9.98	0.001
	miR-876-5p	3.79	0.002	5.99	0.001
	has-miR-876-3p	3.09	0.002	3.92	0.001
	miR-873-5p	2.70	0.003	3.53	0.001
	miR-451a	2.65	0.014	2.49	0.031
	miR-144-3p	2.35	0.009	2.02	0.032
	miR-652-3p	2.13	0.003	3.50	0.001

PrimeScript RT reagent kit (TaKaRa Bio, Inc., Otsu, Japan) following the manufacturer's instructions. RT PCR was performed in an Applied Biosystems Step One RT PCR system using a SYBR Premix Ex Taq II Kit (Takara Bio, Inc., Shiga, Japan) and RT PCR machine (Bio-Rad C1000PCR, USA). Each sample was analyzed in triplicate and the experiment was repeated three times. The primers for *smad2*, miR-26a, miR-26b, U6, and *β-actin* are designed as follows:

smad2 forward: 5′-AGAAGCAGCTCGCCAGCCAG-3′

smad2 reverse: 5′-CGGCGTGAATGGCAAGATGG-3′

miR-26a stem-loop primer: 5′-CTCAACTGGTGTCGTGGAGTCGGCAATTCAGTTGAGAGCCTATC-3′

miR-26a forward: 5′-ACACTCCAGCTGGGTTCAAGTAATCCAGGA-3′

miR-26b stem-loop primer: 5′-CTCAACTGGTGTCGTGGAGTCGGCAATTCAGTTGAGACCTATCC-3′

miR-26b forward: 5′-ACACTCCAGCTGGGTTCAAGTAATTCAGG-3′

Universal reverse primer: 5′-TGGTGTCGTGGAGTCG-3′

β-Actin forward: 5′-AAAGACCTGTACGCCAACAC-3′

β-Actin reverse: 5′-GTCATACTCCTGCTTGCTGAT-3′

U6 forward: 5′-CTCGCTTCGGCAGCACA-3′

U6 reverse: 5′-AACGCTTCACGAATTTGCGT-3′

PCR conditions were set as follows: 95°C for 30 sec, followed by 40 cycles at 95°C for 5 sec, 60°C for 35 sec, and 95°C for 15 sec, 60°C for 1 min, and 95°C for 15 sec, as described previously [31]. Expression levels of *smad2* and miR-26a/b were normalized to *β-actin* or U6 small nuclear RNA (snRNA) expression. The data was analyzed by using the comparative CT method [56].

2.6. Statistical Analysis. All statistical analyses were performed by SPSS software version 17.0 (SPSS, Inc., Chicago, IL, USA) and the results were shown as the mean ± standard deviation (SD) of at least three biological replicates. The p values were determined by a 2-sided t-test and one-way ANOVA followed by the Tukey's post hoc test which was considered as the reference of statistically significant difference.

3. Results

3.1. Differentially Expressed miRNAs Detected in Both Growth and Selection Phases of DFs. Follicular development is a complicated and elaborate process. Both the growth and selection of DFs are significant phases of the follicular development. A prerequisite for understanding follicular progression is acquiring knowledge of its component interactions. In order to investigate the detailed spatiotemporal miRNA profiles during follicular development from small healthy follicles to large healthy follicles, GEO repository (GSE54692) was analyzed in depth and the comparative results of different follicular groups (DFs versus small follicles and DFs versus large atretic follicles) demonstrated that there were 17 and 57 miRNAs differentially expressed (larger than or equal to twofold; adjusted p value ≤ 0.05) between DFs and small follicles as well as between DFs and large atretic follicles, respectively [48]. Out of them, 15 miRNAs (7 upregulated miRNAs and 8 downregulated miRNAs) were differentially expressed in DFs compared to both small follicles and large atretic follicles (Table 1). Then we used DIANA-mirPath and TargetScanHuman 7.0 to identify the target genes and affected pathways associated with these 15 differentially expressed miRNAs (S1 Table in Supplementary Material available online at https://doi.org/10.1155/2017/4585213). This

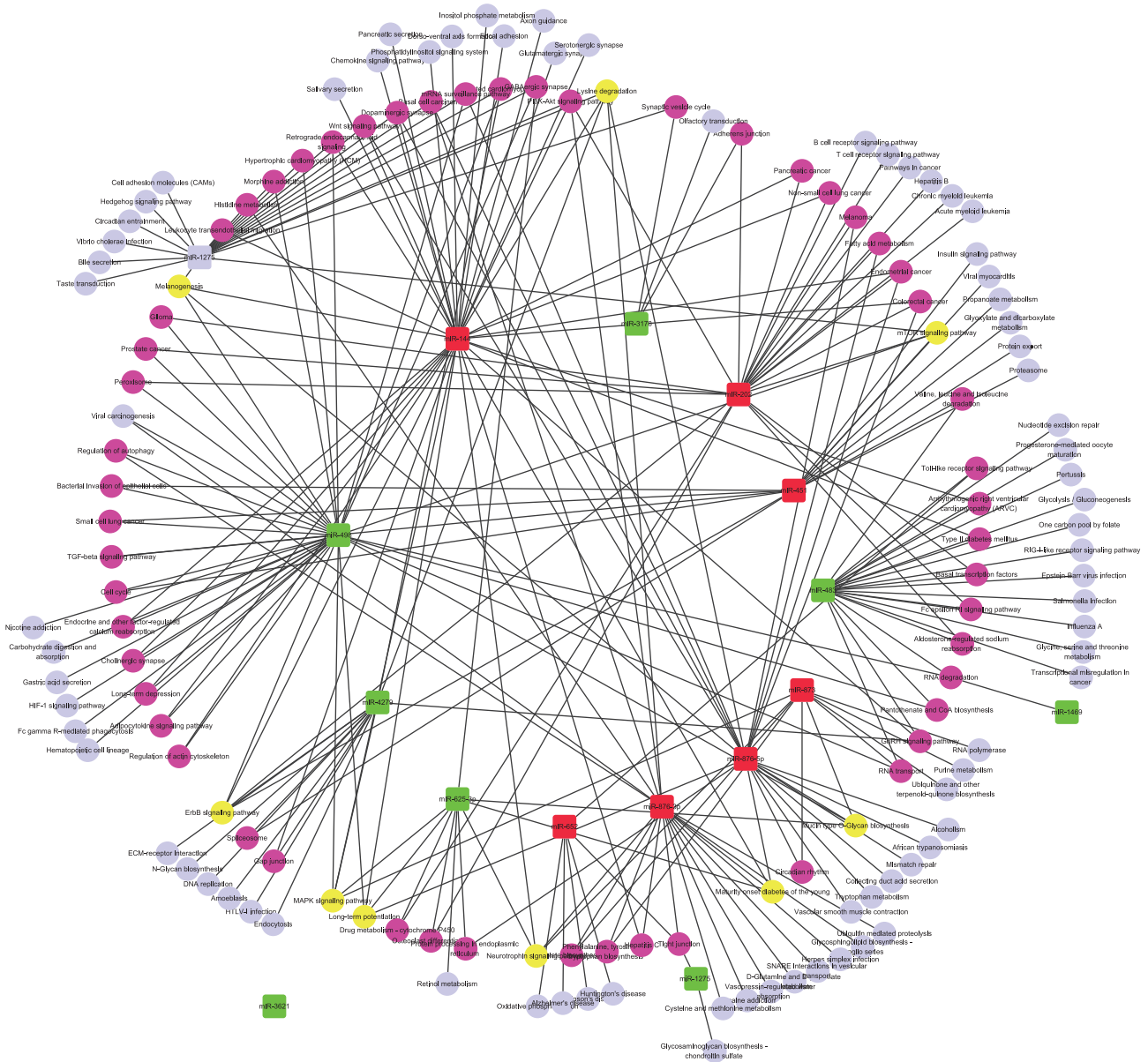


Figure 1: Network of the common differentially expressed miRNAs and their functionally associated signaling pathways in DFs comparing with both small healthy follicles and SFs. The red squares represent the upregulated miRNAs while green squares represent downregulated ones. The purple nodes represent pathways affected by two or three dysregulated miRNAs and the yellow nodes represent pathways coregulated by at least 4 dysregulated miRNAs. Other pale blue circles represent pathways affected by only one differentially expressed miRNA in DFs comparing with both small healthy follicles and SFs.

network showed interaction between 7 upregulated as well as 8 downregulated miRNAs and their identified 139 signaling pathways (Figure 1). Among them, there were 29 pathways coregulated by two or three dysregulated miRNAs. 7 pathways were identified to be influenced by at least four differentially expressed miRNAs in DFs when comparing with small healthy follicles or large atretic follicles (Figure 1). The concrete miRNAs which coregulated these 7 pathways have been showed in S2 Table. For example, MAPK signaling pathway was identified as a target of miR-873-5p, miR-144-3p, miR-625-3p, miR-498, and miR-4279, while miR-3621 was a special node without connections to any other pathway since our absent cognition about its potential function. Taking into consideration that MAPK signaling pathways play a key role in follicular function, this result suggests the significance of differentially expressed miRNAs.

3.2. Abundance of miRNAs in GCs of DFs and SFs during Estrous Cycle. Since every step of the follicular development is important and worthy of investigation, clarifying the profile of miRNAs in the two main kinds of follicles is a burning problem due to miRNAs' regulation process on gene

Table 2: (a) Summary of characteristics of miRNAs in DFs and SFs. (b) The reads of top 10 abundant miRNAs in both DFs and SFs from different samples.

(a)

Group	Sample	Number of mapped reads	Reads aligned to known miRNAs	Arithmetic mean of mapped reads	Arithmetic mean of known miRNAs	Ratio of known miRNAs
Dominant follicles	DFs1	599377	345689	663338.7	343221.7	0.5174
	DFs2	392924	255260			
	DFs3	997715	428716			
Subordinate follicles	SFs1	861596	459794	928373	467028	0.5031
	SFs2	939835	520377			
	SFs3	983688	420913			

(b)

miRNA	DFs			miRNA	SFs		
	Reads1	Reads2	Reads3		Reads1	Reads2	Reads3
miR-26a	48999	53635	42285	*miR-26a*	77730.33	37599	61123
miR-10b	28168.67	46558	14742	*miR-10b*	62390	37494	63322
miR-202	12209	4473	<1000	miR-92a	13653.33	8852	11115
let-7a-5p	10838.33	14314	10417	*let-7f*	13331	11065	11611
let-7f	9595.33	14697	8616	*miR-27b*	13194.67	12922	11789
miR-22-3p	8710.33	3626	5359	miR-99b	12241.33	18612	15768
let-7i	8695.67	13802	7269	*let-7a-5p*	11003.67	10133	10026
miR-21-5p	8695.33	<1000	36422	*let-7i*	9734.67	8944	11203
miR-27b	8476.67	15893	20250	*miR-191*	8563	8073	9995
miR-191	7700.33	15491	5147	miR-143	8397.67	3736	3906

Oblique font represents identical miRNAs both in DFs and SFs.
Reads1: the data were acquired from GSE56002.
Reads2: the data were acquired from Day 3 in GSE55987.
Reads3: the data were acquired from Day 7 in GSE55987.

expression quantity which has been demonstrated by plenty of scientific evidence [57]. As reported in the relevant GEO datasets (GSE56002 and GSE55987), 6 miRNA sequencing libraries were generated based on GCs detecting in both DFs and SFs to understand the abundance and functions of miRNAs in different follicular styles. After filtering PCR primers, low-quality reads, sequences shorter than 18 bps, and empty adaptors, the mean quality reads of the biological triplicates approximated 2.4 and 3 million in the libraries of DFs and SFs, respectively. Sequence alignment of all reads which met the criteria indicated that 663,338 and 928,373 reads in DFs and SFs were mapped to the bovine reference genome, constituting 27.6% and 31.4% of the total quality reads obtained, respectively. Furthermore, 343,221 reads in DFs and 467,028 in SFs were similar to known bovine miRNAs reported in miRbase release 20 (Table 2(a)).

Based on the sequencing results, several miRNAs were commonly abundant in both datasets. Comparing the top 10 abundantly expressed miRNAs in each group, miR-26a, miR-10b, let-7 families (let-7a-5p, let-7f, and let-7i), miR-27b, and miR-191 were consistently observed in both DFs and SFs (Table 2(b)). To sum up, the network indicated that these abundantly expressed miRNAs might play key roles in maintaining the normal physiological function in DFs as well as SFs.

3.3. Canonical Pathways Associated with the Top 7 Abundantly Expressed miRNAs in GCs of DFs and SFs. In order to understand the functional involvement of the 7 common miRNAs in follicular development, target genes of each miRNA were predicted and the relational canonical pathways were performed by mirPath. As a result, a total of 195 canonical pathways were affected by the predicted target genes of highly expressed miRNAs (S3 Table). There were 44 pathways regulated by at least 2 different miRNAs among which 16 pathways were coregulated by more than 3 miRNAs (Figure 2). Thereinto, MAPK is an important signaling pathway in cell proliferation and may function in granulosa cell death and follicular atresia [58]. It was comodulated by all the 7 miRNAs including miR-26a, miR-10b, let-7 families (let-7a-5p, let-7f, and let-7i), miR-27b, and miR-191. To understand the regulation mechanism of miRNAs in the MAPK signaling pathway, target genes for the synergic miRNAs were identified. Deserved to be mentioned, the let-7 families regulate the 6 genes and miR-27b targets 8 genes in MAPK pathway (Table 3). These target genes spread in the different positions of MAPK signaling to influence various functional modules and further establish interactions with other pathways such as Wnt signaling pathway (Figure 3). Therefore, it speculated that these miRNAs regulate some processes of follicular development through modulating MAPK signaling pathway.

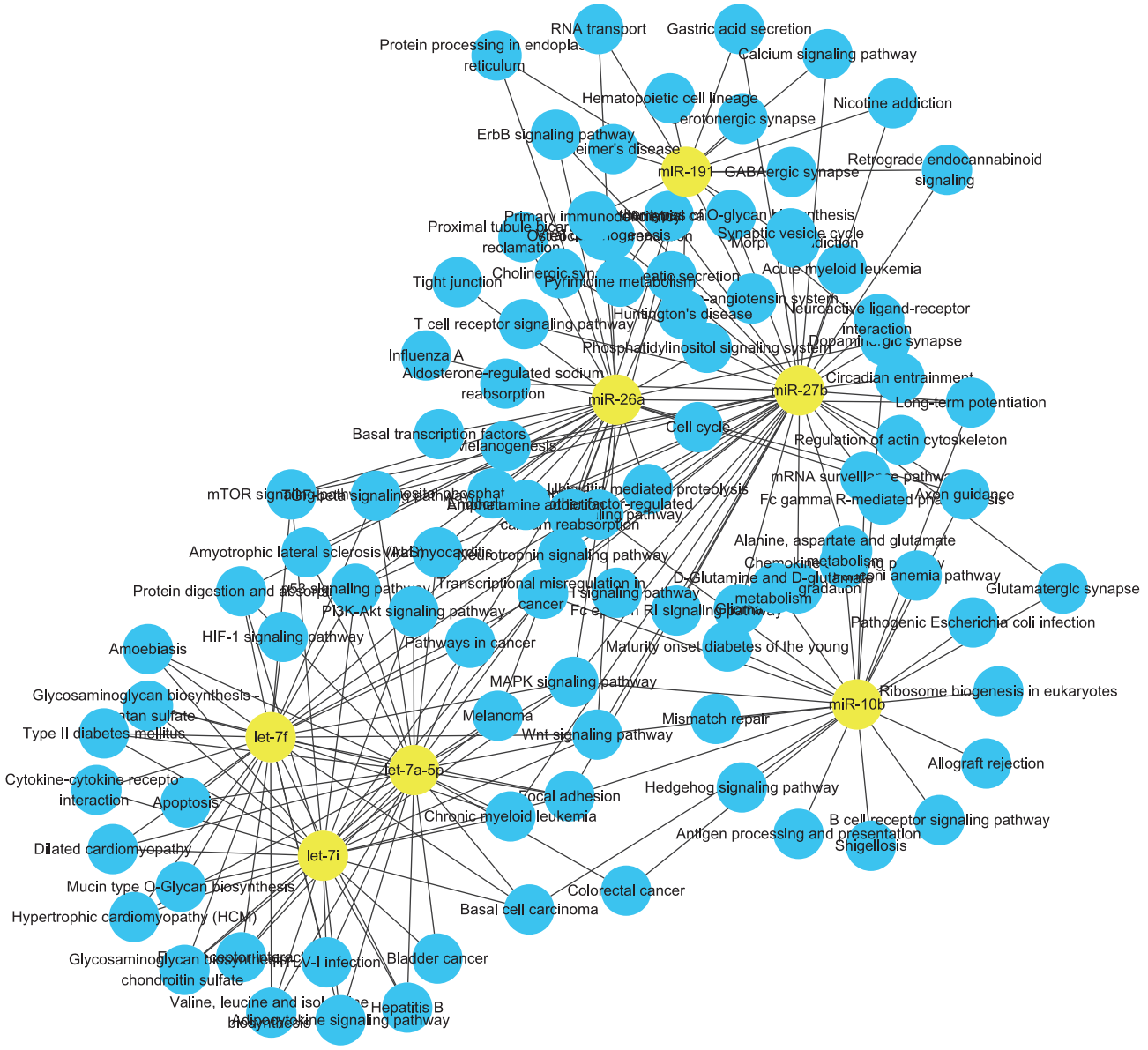


Figure 2: Network of abundantly expressed miRNAs and correlative pathways in both DFs and SFs. Yellow nodes represent abundantly expressed miRNAs in common while blue nodes represent the pathways affected by miRNAs.

Table 3: Target genes regulated by top 7 abundant miRNAs involved in MAPK signaling pathway.

miRNA ID	Target genes
miR-26a	TAB2, NLK, MEF2C, PAK1
miR-10b	MAX, MEF2C, RAP1A
let-7a-5p	FLNA, NGF, FAS, MEF2C, CASP3, RASGRP1
let-7f	FLNA, NGF, FAS, MEF2C, CASP3, RASGRP1
let-7i	FLNA, NGF, FAS, MEF2C, CASP3, RASGRP1
miR-27b	EGFR, TAB2, NLK, NF1, FAS, STMN1, MEF2C, GRB2
miR-191	BDNF

3.4. Canonical Pathways Affected by Differentially Expressed miRNAs in GCs of DFs Compared with SFs. According to the comparison of the two sequencing datasets from GEO (GSE56002 and GSE55987), there were 10 miRNAs significantly differentially expressed in GCs of DFs compared with SFs, of which 5 matured miRNAs, namely, miR-183, miR-34c, miR-708, miR-21-3p, and miR-221, were significantly upregulated in DFs while the others including miR-409a, miR-335, miR-449a, miR-214, and miR-224 were significantly lower expressed (S4 Table). The overall log2 fold change values ranged from −16.4 (miR-409a) up to 7.03 (miR-183). We also used DIANA-mirPath v2.0 software and TargetScanHuman 7.0 to identify target genes and regulatory pathways of differentially expressed miRNAs were compared and analyzed

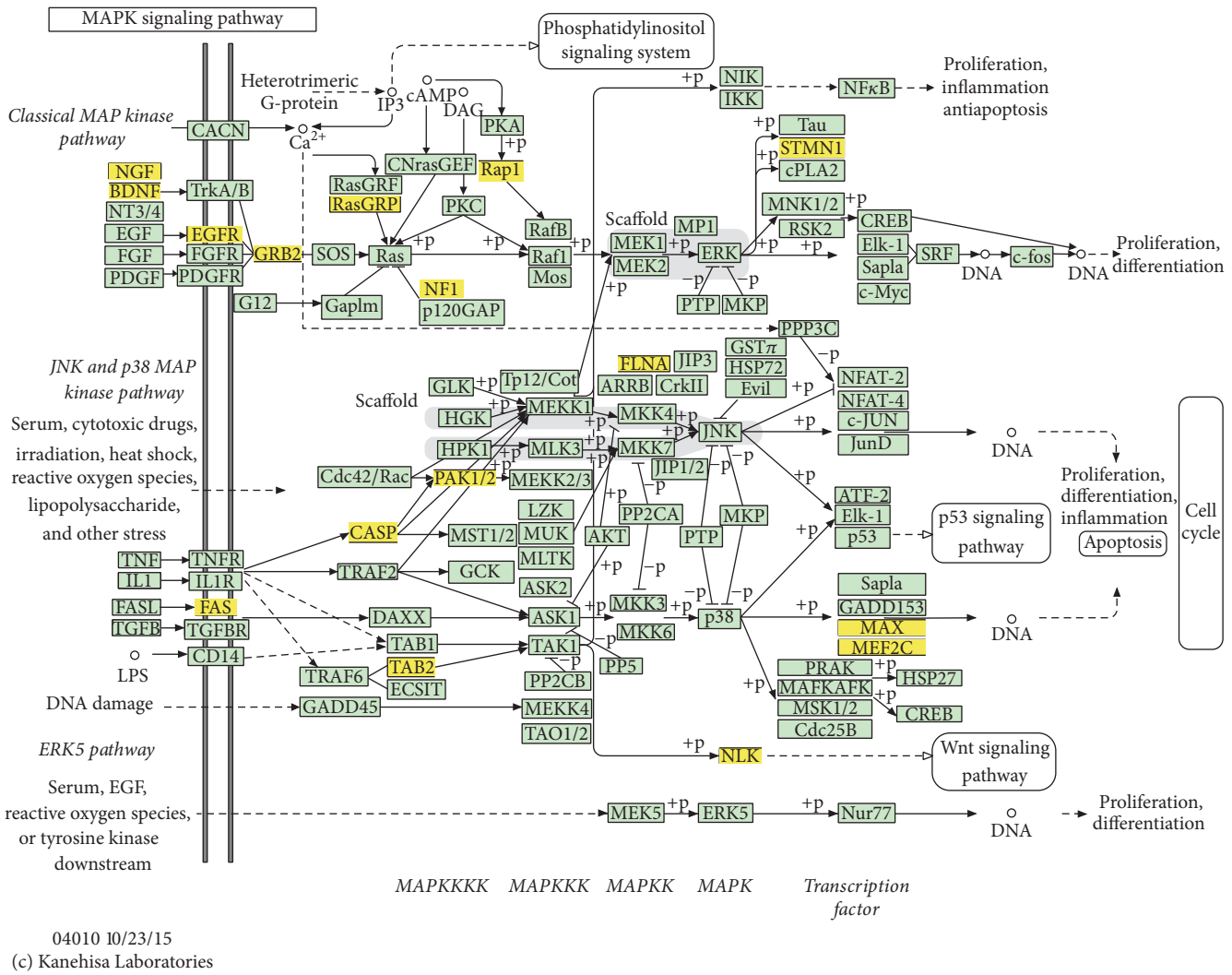


Figure 3: MAPK signaling pathway. This map of MAPK signaling pathway was obtained based on KEGG. The yellow boxes represent target genes which were regulated by the 7 critical miRNAs identified by previous analysis.

using the same method (S5 Table). There were 31 pathways coregulated by two or three dysregulated miRNAs, such as Adherens junction and PI3K-Akt signaling, respectively (Figure 4), while there were 16 pathways coregulated by at least 4 dysregulated miRNAs such as TGF-β signaling pathway. TGF-β signaling pathway has been demonstrated to play a crucial role in the local modulation of cell-cell communication [59]. Thus, TGF-β signaling pathway might exhibit regulatory function for human GCs in ovarian physiology. Specifically, considering that the influences of miRNAs on signaling pathways were achieved by manipulating relevant genes involved in signaling pathways, genes in the TGF-β signaling pathway regulated by miRNAs or in a complex miRNAs' combination were illustrated in Table 4. For example, miR-335 could regulate 7 genes, namely, DCN, FST, THBS1, RBL1, ACVR2A, LTBP1, and BMPR2, while THBS1 was also regulated by miR-708, where the two miRNAs had different regulated directions in DFs (Figure 5). Because of the synergetic regulation of several significant miRNAs, TGF-β signaling pathway might be closely related to follicular development. Besides, a single differentially expressed miRNA might regulate several pathways simultaneously. All of these complex relations between miRNAs and their modulating pathways thus form a functional network related to the different periods of estrous cycle.

3.5. GO Functional Enrichment Analysis of the Pivotal miRNAs. Based on the above findings, GO analysis seems to be meaningful with respect to the function of miRNAs on follicular development process, especially the miRNAs with high abundance and significantly differentially expressed miRNAs, which could be recognized as the pivotal miRNAs. After screening the target gene prediction of the pivotal miRNAs with the thresholds of certain P_{CT} and context score, 696 genes were identified as the functional target genes. By employing the DAVID software, a total of 260 GO function items were enriched. Among them, 206 were enriched in biological process (BP), 20 in cellular component (CC), and the

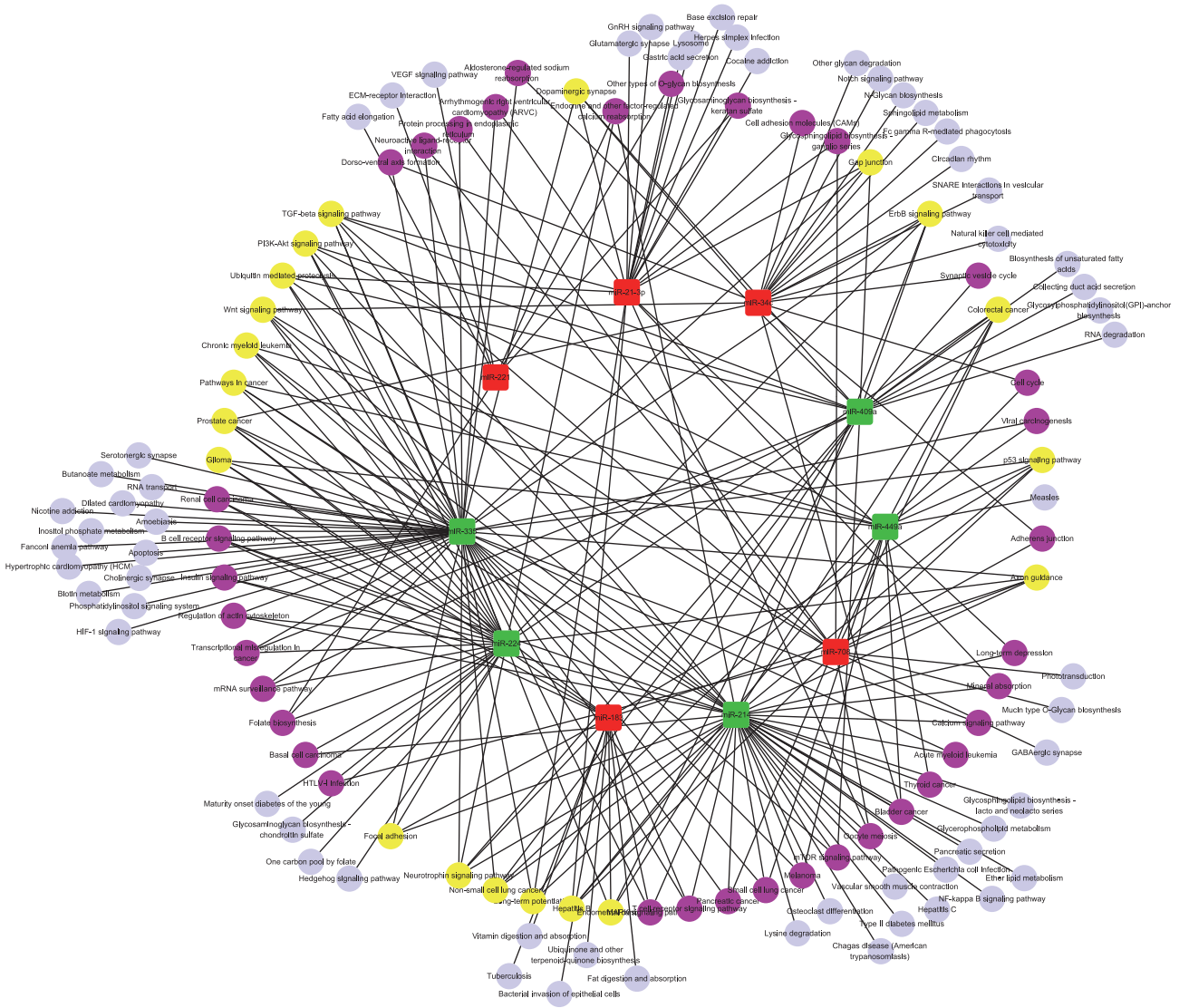


Figure 4: Network of signal pathways and their respective dysregulated miRNAs in DFs. The red squares represent the upregulated miRNAs while the green squares represent downregulated ones. The purple nodes represent pathways affected by two or three dysregulated miRNAs and the yellow nodes represent pathways coregulated by at least 4 dysregulated miRNAs. Other pale blue circles represent pathways affected by only one differentially expressed miRNA in DFs comparing with SFs.

Table 4: Target genes regulated by differentially expressed miRNAs in TGF-β signaling pathway.

miRNA ID	Target genes
miR-708	THBS1
miR-21-3p	TGIF1, SP1
miR-409a	PITX2, BMPR2
miR-335	DCN, FST, THBS1, RBL1, ACVR2A, LTBP1, BMPR2
miR-224	SMAD4, LTBP1, BMPR2
miR-214	CHRD, ACVR2A

others in molecular function (MF). The top 5 of the significant GO terms (with the smallest p value) in each category were shown in Table 5. Interestingly, in BP, the top 5 functions are correlative with embryonic development. It suggested that these miRNAs and target genes participated in not only follicular development but also embryonic development.

3.6. The miRNA-TFs Coregulatory Network in DFs and SFs. Based on the hypothesis that the complexity of the eukaryotic transcriptional regulation machinery reflects a multitude of responses and that regulatory axes involving miRNAs and TFs are not isolated instances, the predicted TFs were incorporated along with the differentially expressed miRNAs in a transcriptional network. Corresponding to the total predicted genes of the pivotal miRNAs in follicular development, a total of 31 transcription factors involved in control of particular ovarian functions retrieved from published scientific reports were collected. Then the regulating network between

Table 5: The top 5 GO terms enriched by the pivotal miRNAs.

Category[1]	Term	Description	Count[2]	p value[3]
BP	GO:0048598	Embryonic morphogenesis	34	$1.7E-10$
	GO:0035113	Embryonic appendage morphogenesis	16	$3.8E-8$
	GO:0030326	Embryonic limb morphogenesis	16	$3.8E-8$
	GO:0043009	Chordata embryonic development	31	$6.3E-8$
	GO:0009792	Embryonic development ending in birth or egg hatching	31	$7.7E-8$
CC	GO:0005667	Transcription factor complex	19	$2.00E-05$
	GO:0044451	Nucleoplasm part	32	$1.40E-04$
	GO:0005654	Nucleoplasm	44	$1.70E-04$
	GO:0005626	Insoluble fraction	41	$4.50E-04$
	GO:0005624	Membrane fraction	39	$8.20E-04$
MF	GO:0003700	Transcription factor activity	53	$2.60E-05$
	GO:0016563	Transcription activator activity	28	$9.60E-05$
	GO:0030528	Transcription regulator activity	71	$9.90E-05$
	GO:0043565	Sequence-specific DNA binding	35	$2.90E-04$
	GO:0003705	RNA polymerase II transcription factor activity,enhancer binding	7	$9.40E-04$

Notes. [1]GO function category; [2]the number of target genes involved in the GO terms. [3]p values have been adjusted using the Benjamini–Hochberg method. BP, biological process; CC, cellular component; MF, molecular function; GO: Gene Ontology.

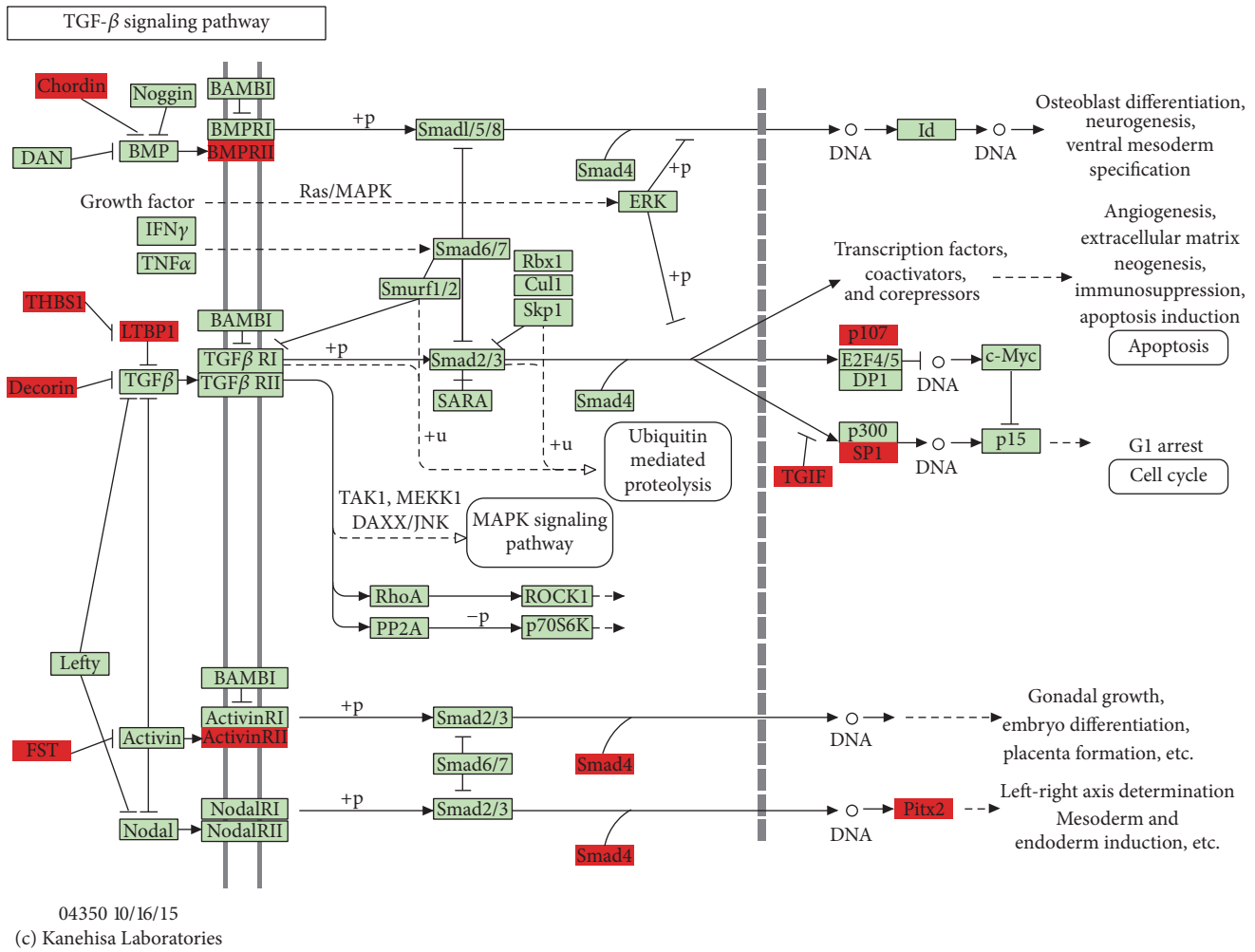


Figure 5: TGF-β signaling pathway. This map of TGF-β signaling pathway was based on KEGG where the red boxes represent target genes regulated by significantly differentially expressed miRNAs through predicting results.

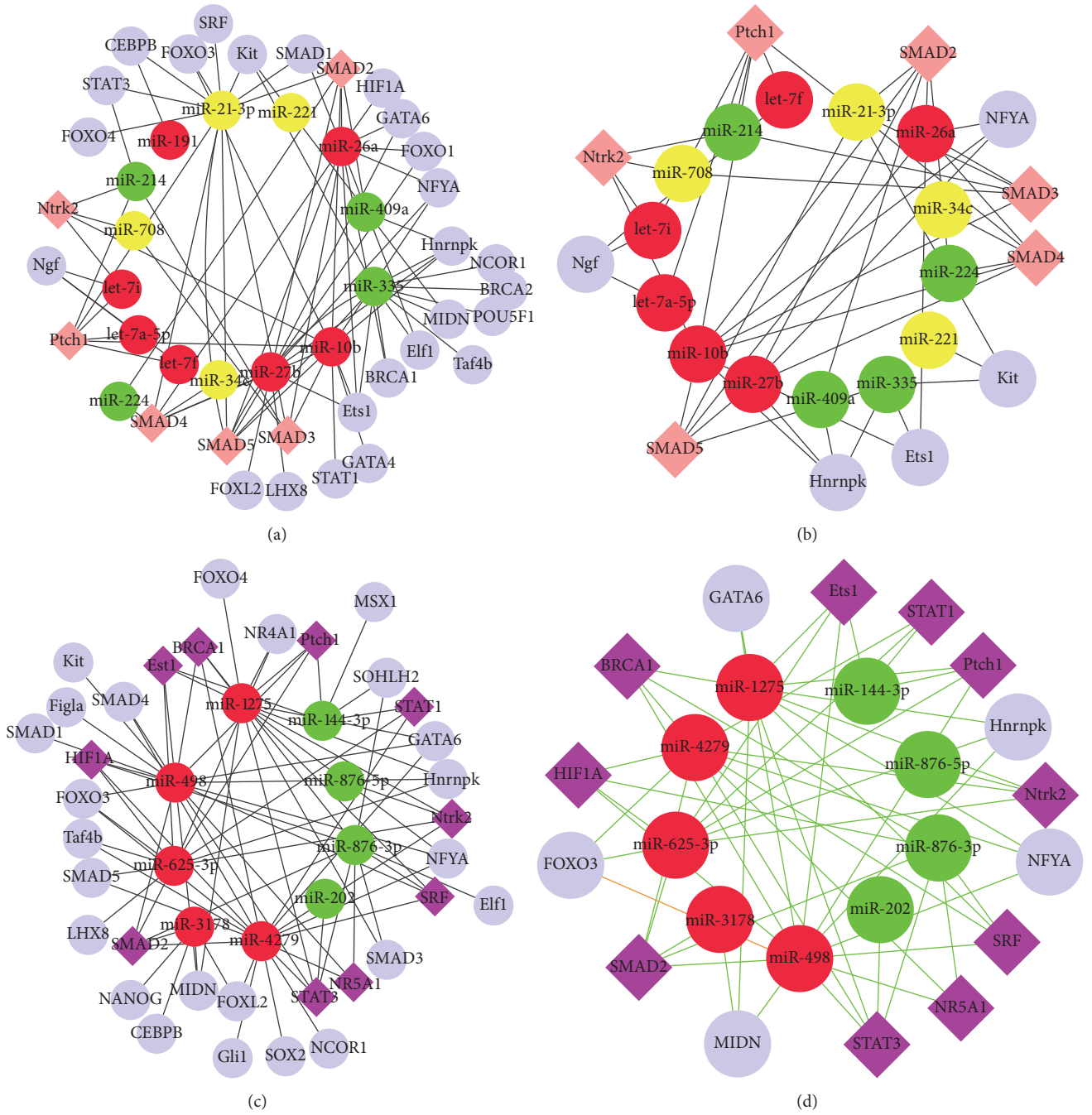


FIGURE 6: Network of miRNAs and TFs. (a) Network of TFs and pivotal miRNAs confirmed in comparison of DFs and SFs. (b) The core network shrank by extracting the hub TFs from (a). The red nodes represent abundantly expressed miRNAs both in DFs and SFs. The yellow nodes represent upregulated and the green nodes represent downregulated miRNAs. The pink rhombic nodes represent TFs affected by differentially expressed miRNAs. (c) Network of TFs and critical miRNAs confirmed in comparison of DFs versus SFs as well as DFs versus small follicles. (d) The relevant core network constructed by the relations between the hub TFs and miRNAs with correspondence to (c). The red nodes represent upregulated miRNAs and green nodes represent downregulated miRNAs. The purple rhombic nodes represent TFs affected by both differentially expressed miRNAs. The orange lines display the connections of TFs and miRNA which had evidenced basis [43–46].

14 miRNAs and 30 TFs was extracted (Figure 6(a)). Some TFs, referring to SMAD2, SMAD3, SMAD4, SMAD5, and Ptch1, have been affected by the most miRNAs with respect to their highest degrees in the network. Since the critical significance of the TFs modulated by multiple miRNAs, the hub TFs recognized as carrying degree larger than 2 (S6 Table) were determined, and a shrunk miRNA-TF regulatory network only covering miRNAs and the hub TFs was subsequently

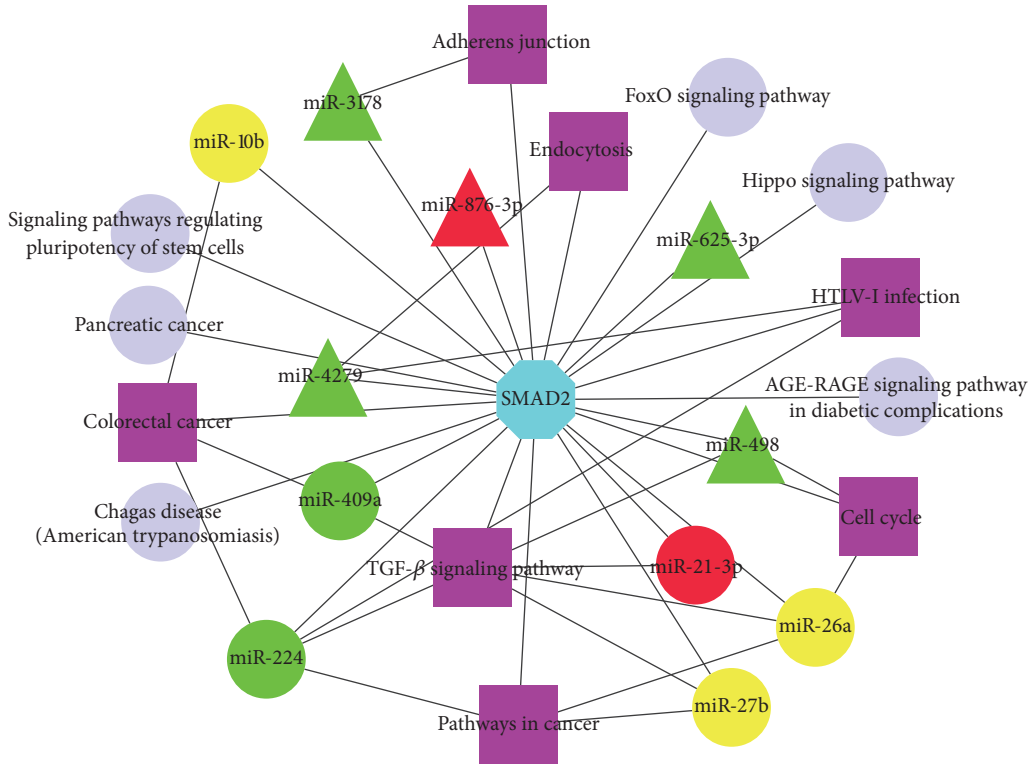


Figure 7: Network of SMAD2 and relevant pathways regulated by identified miRNAs. The yellow circular nodes represent highly abundant miRNAs both in DFs and SFs. Red circular nodes represent the upregulated miRNAs in DFs and the green circular nodes represent the downregulated miRNAs in DFs. The red and green triangular nodes represent up- and downregulated miRNAs in DFs when compared with either small healthy follicles or SFs, respectively. The purple square nodes represent pathways regulated by these identified miRNAs and SMAD2 simultaneously.

generated (Figure 6(b)). Similarly, a network consisting of 35 TFs and several dysregulated miRNAs obtained from the comparison of DFs versus small healthy follicles and DFs versus SFs was constructed (Figure 6(c)) in which STAT3 and SMAD2 showed prominent performance in terms of their highest connections with miRNAs (5 miRNAs for each). Moreover, 21 TFs were regulated by downregulated miRNAs alone while only 3 TFs were regulated by upregulated miRNAs alone. Similarly, the relevant shrunk network covering the 11 hub TFs (with degree larger than 2, S7 Table) has been constructed to represent the core miRNA-TF regulatory relations (Figure 6(d)). SMAD2 might play a key role in follicular development because of its participation in both the core miRNA-TF networks as shown in Figures 6(b) and 6(d). The network of predicted TFs and miRNAs indicated the possible transcriptional regulation in follicular development.

SMAD superfamily including SMAD1, SMAD2, SMAD3, SMAD4, and SMAD5 have been demonstrated to have functional correlations with ovarian follicular growth and selection [60–62]. SMAD2 could be regulated by all the identified important miRNAs from different detecting groups in our study. We found that TGF-β signaling pathway would be regulated by several miRNAs related to follicular development such as miRNA-21-3p while it also was mediated by SMAD superfamily [63]. There is a group of miRNAs that could coregulate both some signaling pathways (such as TGF-β signaling pathway) and SAMD2. Interestingly, in these identified pathways regulated by miRNAs, SMAD2 frequently affects the activity of these pathways as a key functional component (Figure 7). In conclusion, SMAD2 could be considered as a bridge, connecting predecessors and successors and affecting pathway activity in response to environmental signals. The result also suggested that, in ovarian follicular development, different regulatory elements functioned synergistically in networks rather than working alone.

3.7. The Regulatory Impact of miR-26a/b on the Expression of smad2 in KGN Cells. As the above results of this study, we deduced that some miRNAs may influence follicular development process through implementing regulation on some genes and associated signaling pathways, for instance, the TF, *smad2*. However, it was primary prediction based on the importance on its functional networks, and few evidences for the prediction were provided by experimental researches. A related research reported that miR-26b could induce apoptosis in GCs by targeting SMAD4, both directly and indirectly through USP9X, which regulated the ubiquitination of SMAD4 [63, 64]. Since miR-26b and miR-26a are highly homologous and in many species of mammals the seed sequences are also highly conservative (Figure 8(a)), we conducted the investigation of the regulatory function

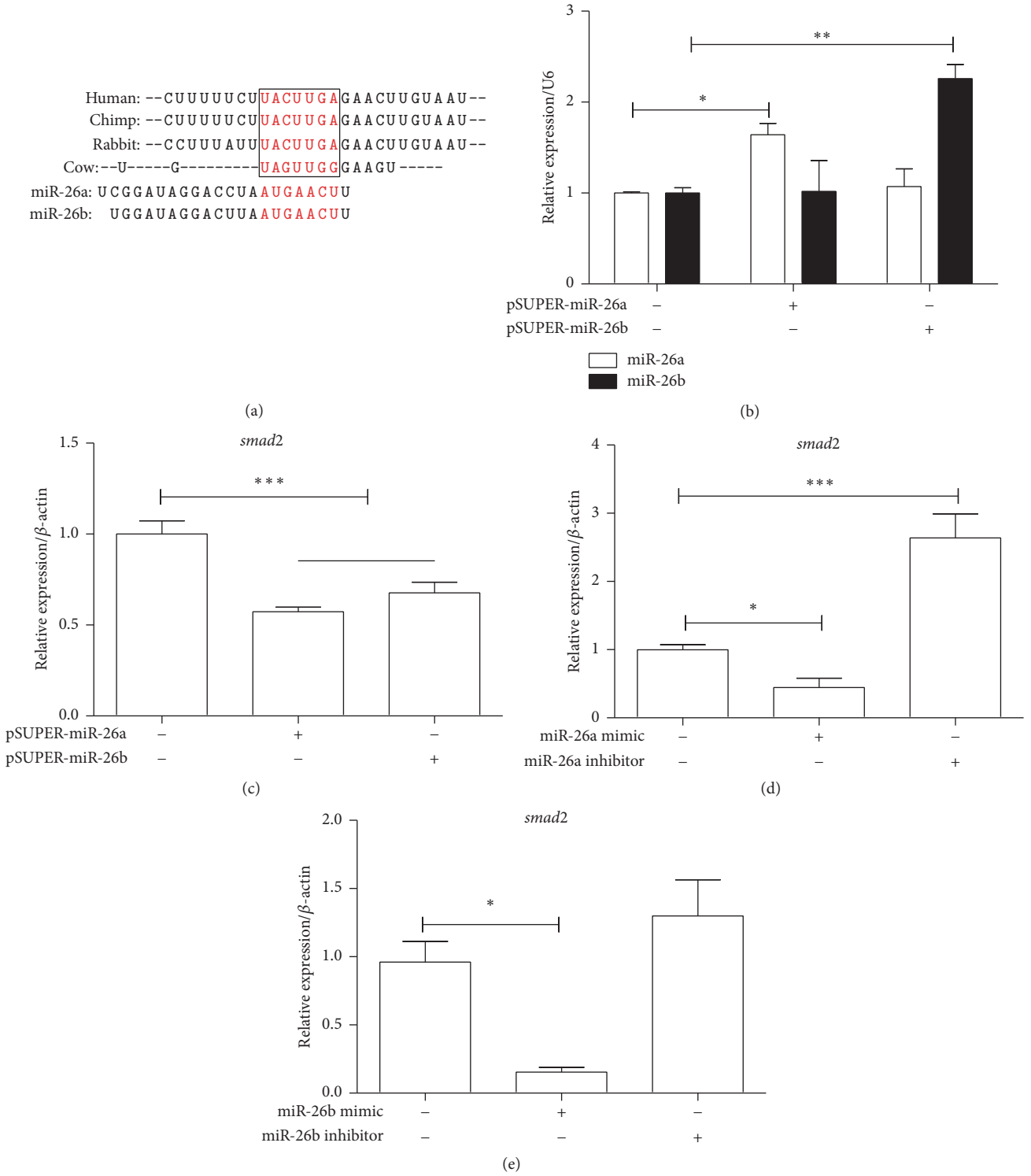


FIGURE 8: The effect of miR-26a/b on the regulating *smad2* mRNA in KGN cells. (a) A sketch map describes predicted seed sequences of miR-26a/b in the conservative 3′-UTR of *smad2* in several mammal species. (b) The transfection efficiency of overexpression vectors of miR-26a/b (pSUPER-miR-26a and pSUPER-miR-26b) was validated by RT PCR. (c, d, and e) The effects of pSUPER-miR-26a/b, miR-26a/b mimics, and miR-26a/b inhibitors on the *smad2* mRNA level were identified in KGN cells, respectively. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$; all experiments were independently repeated three times.

of the miR-26a/b on *smad2*. The miRNA overexpression vectors, miR-26a/b mimics, and miR-26a/b inhibitors were transfected into KGN cells successively. The results showed that overexpression of miR-26a/b in KGN cells (Figure 8(b)) led to the significant suppression of *smad2* (Figure 8(c)). Cells transfected with miR-26a/b mimics displayed a significantly decreased expression of *smad2*. In contrast, the miR-26a inhibitor transfected cells had an obviously increased tendency of *smad2* expression compared with NC transfected cells (Figures 8(d) and 8(e)). It suggested that miR-26a/b could regulate the expression of *smad2* in KGN cell line. However, whether miR-26a/b could directly target SMAD2 needs further research. Thus, this experiment probably provides preliminarily supportive evidence on our speculative efforts. In this context, other important regulating relations among miRNAs as well as genes and signaling pathways related to follicular development, which were identified by our analysis, may be worth studying in depth.

4. Discussion

Follicular development during the estrous cycle is an extraordinarily complex and synergistic process which might be regulated accurately by plenty of regulatory factors such as miRNAs and TFs. The aim of this study was to analyze the function of significant miRNAs macroscopically and associated functional networks related to follicular development via bioinformatic investigation. It is worth noting that miRNAs are largely conserved in different species of mammal. Furthermore, studies of miRNAs in ovarian tissues had confirmed the similar expression patterns in ovaries of various species, including humans [16], mice [17, 57, 65], pigs [18], sheep [66], goats [19], and cows [20, 67–69]. Through these references and related data, miRNAs which were discovered in transcriptome of bovine are also subsistent in humans and mice. Therefore, it is feasible that TargetScanHuman and DIANA, which are developed mainly for human, also could be used to predict target genes of differentially expressed miRNAs in bovine because of sequence conservation and the relevant algorithm basis. miRNA deep sequencing quantifies the relative abundance of miRNAs by their frequencies in terms of read counts. Highly abundant miRNAs have higher likelihood of higher read counts compared to miRNAs with lower abundance [70]. In GSE56002, from all reads that met the quality control criteria, 343,221 reads in DFs and 467,028 in SFs were found to be similar to known miRNAs reported in miRbase [24]. Among the detected miRNAs which appeared in both GSE56002 and GSE55987, lots of miRNAs such as isoforms of let-7 family were commonly expressed with high levels in both DFs and SFs. It implied that this kind of miRNAs might not regulate selection of DFs but maintain normal physiological functions in GCs of both DFs and SFs during the estrous cycle. The previous studies had demonstrated that the let-7 family could regulate steroidogenesis and expression of steroidogenesis-related genes [71]. In fact, steroidogenesis is a crucial biology process for maintaining normal physiology function in both DFs and SFs. Not only let-7 family but also miR-191 and miR-26a present abundant expression which could increase the proliferation of GCs [72] and regulate gonad development partially through its target on *exm2* [73]. Specifically, miR-26a showed abundant expression in both DFs and SFs, while the reads of miR-26a screened in SFs were actually more than in DFs. The results indicated that miRNA-26a might play an important role in discrepant functions between SFs and DFs. Furthermore, the canonical pathways, such as PI3K-Akt signaling pathway, MAPK signaling pathway, TGF-β signaling pathway, Wnt signaling pathways, and Axon guidance, which were enriched from abundant miRNAs also related to cell metabolism and signal transduction in normal follicular development. It demonstrated that the functions of these regulated pathways were of equal importance in different follicular types. Some research articles have suggested EGFR, MEF2C, and BDNF in MAPK signaling pathway were regulated by miR-27b and miR-191, respectively [74–77]. Meanwhile, it has been demonstrated that let-7 family might be involved in the estrous cycle through MAPK signaling pathway [78–80].

Abundantly expressed miRNAs might have key roles in maintaining the normal status of follicles while differentially expressed miRNAs would exhibit more regulatory functions in follicular growth, selection, and the fate of DFs and SFs. A total of 10 miRNAs were dysregulated between DFs and SFs and the result might provide valuable insights into their potential roles in folliculogenesis in a stage-dependent manner. Thus, it could be assumed that several signal pathways influenced by differentially expressed miRNAs were associated with growth and selection of follicles. Previous studies reported that increased cell apoptosis would be observed in GCs transfected with the miR-21 inhibitor. Furthermore, the critical roles of miR-21 in regulating follicular development and preventing GCs apoptosis in DFs had also been verified [35], while miR-183, upregulated in DFs, might induce cell apoptosis [72]. It suggested that in follicular development, multiple miRNAs, pathways, or TFs work simultaneously, though they might show some opposite effects. In addition, SMAD4 which played a key role in TGF-β pathway would be targeted by miR-224 [31], while miR-214/199a and miR-335 could affect TGF-β pathway in different manners [81, 82]. TGF-β superfamily members have been implicated in regulating GC proliferation [83] and terminal differentiation [84] that are critical for normal ovarian follicular development. Furthermore, by GO functional enrichment, we found that the miRNA target genes played key role in follicular development, which can also effect the embryonic development. It is well known that follicular development and embryonic development are both two important biological processes for reproduction. The roles of some miRNA target genes in both these processes were probably similar. For instance, the genes in Wnt signaling pathway were regulated by same miRNAs in both follicular development and embryonic development [46].

miRNAs participate in follicular development not only in one single manner; there is another primary mode to command maturation of follicles related to critical transcription factors. Similar to the significance of pathway regulation, TFs also could influence the regulatory network in mediating follicular development. According to the bioinformatic

analysis, SMAD2 might be coregulated by miR-27b, miR-10b, and miR-26a which were expressed abundantly in both follicle types. miR-21 and miR-409a with crosscurrent in DFs also could regulate *smad2*. Interestingly, among these miRNAs, miR-21 was demonstrated to be regulated via SMAD2/3 signaling [85]. It suggested that SMAD2 interacted with some miRNAs and played key roles in follicular development. miR-26a and miR-26b are highly homologous and the seed regions are highly conservative in several species. In our study, we concluded that miR-26a/b could decrease the mRNA expression of *smad2* through RT PCR results. Nevertheless, the detailed regulatory mechanisms need to be further studied. Because of the abundance of miR-26, both in DFs and in SFs, *smad2* expression may be in a low level and the correlation pathways may be in an inactive status. Moreover, *smad2* was also coregulated by miR-498, miR-3178, miR-4279, and miR-625-3p, which were downregulated, and miR-876-3p that was upregulated in DFs compared with both small follicles and SFs. Although the complex correlation of SMAD2 and significant miRNAs made this TF a hub in the whole gene regulatory network of follicular development, deficient evidence about the functional relations of SMAD2 and miRNAs was obviously scarce to confirm this description. The functions of SMAD2-miRNAs network in follicular development need further study. Besides SMAD2, other members of SMAD superfamily such as SMAD3 or SMAD4 were also demonstrated to be associated with miRNAs and signaling pathways [31, 86], especially TGF-β signaling pathway which had illustrated affecting ovary development [87]. For example, miR-21 could regulate the expression of SMAD7 and TGF-β1 [88]. The accurately regulatory network composed of important miRNAs and hub TFs would provide a valuable perspective for understanding the formation of DFs during estrous cycle.

Precisely because of the miscellaneous connection among miRNAs, signaling pathways, and TFs, the regulatory network in final maturation of follicle and preparation for the subsequent follicle-luteal transition would be more thorough. Our coregulatory network will help to draw the dynamic changes of miRNAs and associated regulatory modules across a wide range of follicular developmental stages. However, there were also some deficiencies in this study. The work was not based on the relationship of miRNAs and hormones which play a key role in the whole estrous cycle. Several miRNAs were demonstrated to be regulated by hormones, such as luteinizing hormone (LH), hCG (human chorionic gonadotropin), and follicle stimulating hormone (FSH) [31, 35, 43, 89] which displayed important controls on ovarian functions. It suggested that miRNAs may regulate ovarian functions associated with changes of hormones content. Moreover, although the significant miRNAs and their influencing pathways or TFs were identified, the functions of differentially expressed miRNAs in follicular development were still unclear. Despite these limitations, this study still predicted possible details about molecular regulation network and identified probable key regulators in the process of follicular development, which provided a novel idea to understand the regulation of follicular development deeply, as well as diagnosis and treatment for infertility.

5. Conclusions

In this study, via bioinformatic analysis, we concluded that during the growth and selections of DFs lots of miRNAs in GCs which are abundantly expressed (let-7) or dysregulated (miR-21-3p) take part in biology processes by forming networks with TFs (such as SMAD superfamily) and pathways (TGF-β signaling pathway). By GO functional enrichment analysis of the pivotal miRNAs target genes, 260 GO function items in BP, CC, and MF were enriched. SMAD2, regulated by miR-26a/b, has been demonstrated by RT PCR as one of these kinds of TFs may play a key role in several pathways as a participator. It suggested that miR-26a/b-SMAD2-TGF-β signaling pathway might play a significant role in follicular development as an axis. Having said that, we wish to emphasize that the functions of other pivotal miRNAs, TFs, and pathways need more studies at macro level.

Disclosure

Baoyun Zhang and Long Chen are co-first authors.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 31372287), Ministry of Agriculture Transgenic Major Projects of China (2014ZX0800952B), the Agricultural Science and Technology Innovation Program of China (ASTIP-IAS13), and National Biological Breeding Capacity Building and Industrialization Projects (2014-2573). The authors want to take this opportunity to acknowledge all the projects for their support and participation. They also would like to acknowledge all sequencing datasets in this study.

References

[1] S. W. Maalouf, W. S. Liu, and J. L. Pate, "MicroRNA in ovarian function," *Cell and Tissue Research*, vol. 363, no. 1, pp. 7–18, 2016.

[2] P. G. Knight and C. Glister, "Potential local regulatory functions of inhibins, activins and follistatin in the ovary," *Reproduction*, vol. 121, no. 4, pp. 503–512, 2001.

[3] F. Tu, Z. X. Pan, Y. Yao et al., "miR-34a targets the inhibin beta B gene, promoting granulosa cell apoptosis in the porcine ovary," *Genetics and Molecular Research*, vol. 13, no. 2, pp. 2504–2512, 2014.

[4] O. J. Ginther, M. C. Wiltbank, P. M. Fricke, J. R. Gibbons, and K. Kot, "Selection of the dominant follicle in cattle," *Biology of Reproduction*, vol. 55, no. 6, pp. 1187–1194, 1996.

[5] R. Buccione, A. C. Schroeder, and J. J. Eppig, "Interactions between somatic cells and germ cells throughout mammalian oogenesis," *Biology of Reproduction*, vol. 43, no. 4, pp. 543–547, 1990.

[6] J. M. J. Aerts and P. E. J. Bols, "Ovarian follicular dynamics: a review with emphasis on the bovine species. Part I: folliculogenesis and pre-antral follicle development," *Reproduction in Domestic Animals*, vol. 45, no. 1, pp. 171–179, 2010.

[7] H. Yada, K. Hosokawa, K. Tajima, Y. Hasegawa, and F. Kotsuji, "Role of ovarian theca and granulosa cell interaction in hormone production and cell growth during the bovine follicular maturation process," *Biology of Reproduction*, vol. 61, no. 6, pp. 1480–1486, 1999.

[8] M. J. Canty, M. P. Boland, A. C. O. Evans, and M. A. Crowe, "Alterations in follicular IGFBP mRNA expression and follicular fluid IGFBP concentrations during the first follicle wave in beef heifers," *Animal Reproduction Science*, vol. 93, no. 3-4, pp. 199–217, 2006.

[9] A. C. O. Evans, J. L. H. Ireland, M. E. Winn et al., "Identification of genes involved in apoptosis and dominant follicle development during follicular waves in cattle," *Biology of Reproduction*, vol. 70, no. 5, pp. 1475–1484, 2004.

[10] T. Fayad, V. Lévesque, J. Sirois, D. W. Silversides, and J. G. Lussier, "Gene expression profiling of differentially expressed genes in granulosa cells of bovine dominant follicles using suppression subtractive hybridization," *Biology of Reproduction*, vol. 70, no. 2, pp. 523–533, 2004.

[11] M. Mihm, P. J. Baker, L. M. Fleming, A. M. Monteiro, and P. J. O'Shaughnessy, "Differentiation of the bovine dominant follicle from the cohort upregulates mRNA expression for new tissue development genes," *Reproduction*, vol. 135, no. 2, pp. 253–265, 2008.

[12] I. Gilbert, C. Robert, S. Dieleman, P. Blondin, and M.-A. Sirard, "Transcriptional effect of the LH surge in bovine granulosa cells during the peri-ovulation period," *Reproduction*, vol. 141, no. 2, pp. 193–205, 2011.

[13] J. U. Rao, K. B. Shah, J. Puttaiah, and M. Rudraiah, "Gene expression profiling of preovulatory follicle in the buffalo cow: effects of increased IGF-I concentration on periovulatory events," *PLoS ONE*, vol. 6, no. 6, Article ID e20754, 2011.

[14] L. K. Christenson, S. Gunewardena, X. Hong, M. Spitschak, A. Baufeld, and J. Vanselow, "Research resource: preovulatory LH surge effects on follicular theca and granulosa transcriptomes," *Molecular Endocrinology*, vol. 27, no. 7, pp. 1153–1171, 2013.

[15] E. Huntzinger and E. Izaurralde, "Gene silencing by microRNAs: contributions of translational repression and mRNA decay," *Nature Reviews Genetics*, vol. 12, no. 2, pp. 99–110, 2011.

[16] Y. Liang, D. Ridzon, L. Wong, and C. Chen, "Characterization of microRNA expression profiles in normal human tissues," *BMC Genomics*, vol. 8, article 166, 2007.

[17] S. Ro, R. Song, C. Park, H. Zheng, K. M. Sanders, and W. Yan, "Cloning and expression profiling of small RNAs expressed in the mouse ovary," *RNA*, vol. 13, no. 12, pp. 2366–2380, 2007.

[18] M. Li, Y. Liu, T. Wang et al., "Repertoire of porcine microRNAs in adult ovary and testis by deep sequencing," *International Journal of Biological Sciences*, vol. 7, no. 7, pp. 1045–1055, 2011.

[19] Y.-H. Ling, C.-H. Ren, X.-F. Guo et al., "Identification and characterization of microRNAs in the ovaries of multiple and uniparous goats (Capra hircus) during follicular phase," *BMC Genomics*, vol. 15, article 339, 2014.

[20] M. Hossain, N. Ghanem, M. Hoelker et al., "Identification and characterization of miRNAs expressed in the bovine ovary," *BMC Genomics*, vol. 10, article 1471, p. 443, 2009.

[21] X. Hong, L. J. Luense, L. K. McGinnis, W. B. Nothnick, and L. K. Christenson, "Dicer1 is essential for female fertility and normal development of the female reproductive system," *Endocrinology*, vol. 149, no. 12, pp. 6207–6212, 2008.

[22] A. K. Nagaraja, C. Andreu-Vieyra, H. L. Franco et al., "Deletion of dicer in somatic cells of the female reproductive tract causes sterility," *Molecular Endocrinology*, vol. 22, no. 10, pp. 2336–2352, 2008.

[23] G. Gonzalez and R. R. Behringer, "Dicer is required for female reproductive tract development and fertility in the mouse," *Molecular Reproduction and Development*, vol. 76, no. 7, pp. 678–688, 2009.

[24] S. Gebremedhn, D. Salilew-Wondim, I. Ahmad et al., "MicroRNA expression profile in bovine granulosa cells of preovulatory dominant and subordinate follicles during the late follicular phase of the estrous cycle," *PLoS ONE*, vol. 10, no. 5, Article ID e0125912, 2015.

[25] B. Bao and H. A. Garverick, "Expression of steroidogenic enzyme and gonadotropin receptor genes in bovine follicles during ovarian follicular waves: a review," *Journal of Animal Science*, vol. 76, no. 7, pp. 1903–1921, 1998.

[26] U. A. Vitf, M. Hayashi, C. Klein, and A. J. W. Hsueh, "Growth differentiation factor-9 stimulates proliferation but suppresses the follicle-stimulating hormone-induced differentiation of cultured granulosa cells from small antral and preovulatory rat follicles," *Biology of Reproduction*, vol. 62, no. 2, pp. 370–377, 2000.

[27] L. J. Spicer, N. B. Schreiber, D. V. Lagaly, P. Y. Aad, L. B. Douthit, and J. A. Grado-Ahuir, "Effect of resistin on granulosa and theca cell function in cattle," *Animal Reproduction Science*, vol. 124, no. 1-2, pp. 19–27, 2011.

[28] K.-G. Hayashi, K. Ushizawa, M. Hosoe, and T. Takahashi, "Differential genome-wide gene expression profiling of bovine largest and second-largest follicles: identification of genes associated with growth of dominant follicles," *Reproductive Biology and Endocrinology*, vol. 8, article 11, 2010.

[29] B. Sisco, L. J. Hagemann, A. N. Shelling, and P. L. Pfeffer, "Isolation of genes differentially expressed in dominant and subordinate bovine follicles," *Endocrinology*, vol. 144, no. 9, pp. 3904–3913, 2003.

[30] F. X. Donadeu, S. N. Schauer, and S. D. Sontakke, "Involvement of miRNAs in ovarian follicular and luteal development," *The Journal of Endocrinology*, vol. 215, no. 3, pp. 323–334, 2012.

[31] G. Yao, M. Yin, J. Lian et al., "MicroRNA-224 is involved in transforming growth factor-β-mediated mouse granulosa cell proliferation and granulosa cell function by targeting Smad4," *Molecular Endocrinology*, vol. 24, no. 3, pp. 540–551, 2010.

[32] G. Yan, L. Zhang, T. Fang et al., "MicroRNA-145 suppresses mouse granulosa cell proliferation by targeting activin receptor IB," *FEBS Letters*, vol. 586, no. 19, pp. 3263–3270, 2012.

[33] A. Dai, H. Sun, T. Fang et al., "MicroRNA-133b stimulates ovarian estradiol synthesis by targeting Foxl2," *FEBS Letters*, vol. 587, no. 15, pp. 2474–2482, 2013.

[34] Q. Zhang, H. Sun, Y. Jiang et al., "MicroRNA-181a suppresses mouse granulosa cell proliferation by targeting activin receptor IIA," *PLoS ONE*, vol. 8, no. 3, Article ID e59667, 2013.

[35] M. Z. Carletti, S. D. Fiedler, and L. K. Christenson, "MicroRNA 21 blocks apoptosis in mouse periovulatory granulosa cells," *Biology of Reproduction*, vol. 83, no. 2, pp. 286–295, 2010.

[36] F. Lin, R. Li, Z. X. Pan et al., "miR-26b promotes granulosa cell apoptosis by targeting ATM during follicular atresia in porcine ovary," *PLoS ONE*, vol. 7, no. 6, Article ID e38640, 2012.

[37] X. Yang, Y. Zhou, S. Peng et al., "Differentially expressed plasma microRNAs in premature ovarian failure patients and the potential regulatory function of mir-23a in granulosa cell apoptosis," *Reproduction*, vol. 144, no. 2, pp. 235–244, 2012.

[38] Y. Kitahara, K. Nakamura, K. Kogure, and T. Minegishi, "Role of *microRNA-136-3p* on the expression of luteinizing hormone-human chorionic gonadotropin receptor mRNA in rat ovaries," *Biology of Reproduction*, vol. 89, no. 5, article 114, 2013.

[39] S. Xu, K. Linher-Melville, B. B. Yang, D. Wu, and J. Li, "Micro-RNA378 (miR-378) regulates ovarian estradiol production by targeting aromatase," *Endocrinology*, vol. 152, no. 10, pp. 3941–3951, 2011.

[40] M. Yin, M. Lü, G. Yao et al., "Transactivation of microRNA-383 by steroidogenic factor-1 promotes estradiol release from mouse ovarian granulosa cells by targeting RBMS1," *Molecular Endocrinology*, vol. 26, no. 7, pp. 1129–1143, 2012.

[41] G. Yao, M. Liang, N. Liang et al., "MicroRNA-224 is involved in the regulation of mouse cumulus expansion by targeting Ptx3," *Molecular and Cellular Endocrinology*, vol. 382, no. 1, pp. 244–253, 2014.

[42] X. Li, J. Xu, Y. Li et al., "Dissection of the potential characteristic of miRNA-miRNA functional synergistic regulations," *Molecular BioSystems*, vol. 9, no. 2, pp. 217–224, 2013.

[43] M. Yin, X. Wang, G. Yao et al., "Transactivation of microRNA-320 by microRNA-383 regulates granulosa cell functions by targeting E2F1 and SF-1 proteins," *Journal of Biological Chemistry*, vol. 289, no. 26, pp. 18239–18257, 2014.

[44] Y. Wang, J. Ren, Y. Gao et al., "MicroRNA-224 targets sMAD family member 4 to promote cell proliferation and negatively influence patient survival," *PLoS ONE*, vol. 8, no. 7, Article ID e68744, 2013.

[45] C. M. M. Gits, P. F. van Kuijk, M. B. E. Jonkers et al., "MiR-17-92 and miR-221/222 cluster members target KIT and ETV1 in human gastrointestinal stromal tumours," *British Journal of Cancer*, vol. 109, no. 6, pp. 1625–1635, 2013.

[46] R. Feng, Q. Sang, Y. Zhu et al., "MiRNA-320 in the human follicular fluid is associated with embryo quality in vivo and affects mouse embryonic development in vitro," *Scientific Reports*, vol. 5, article 8689, 2015.

[47] D. Salilew-Wondim, I. Ahmad, S. Gebremedhn et al., "The expression pattern of microRNAs in granulosa cells of subordinate and dominant follicles during the early luteal phase of the bovine estrous cycle," *PLoS ONE*, vol. 9, no. 9, Article ID e106795, 2014.

[48] S. D. Sontakke, B. T. Mohammed, A. S. McNeilly, and F. X. Donadeu, "Characterization of microRNAs differentially expressed during bovine follicle development," *Reproduction*, vol. 148, no. 3, pp. 271–283, 2014.

[49] 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.

[50] M. Kiriakidou, P. T. Nelson, A. Kouranov et al., "A combined computational-experimental approach predicts human microRNA targets," *Genes & Development*, vol. 18, no. 10, pp. 1165–1178, 2004.

[51] 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.

[52] A. V. Sirotkin, "Transcription factors and ovarian functions," *Journal of Cellular Physiology*, vol. 225, no. 1, pp. 20–26, 2010.

[53] 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.

[54] V. Nordhoff, B. Sonntag, D. von Tils et al., "Effects of the FSH receptor gene polymorphism p.N680S on cAMP and steroid production in cultured primary human granulosa cells," *Reproductive BioMedicine Online*, vol. 23, no. 2, pp. 196–203, 2011.

[55] D. Li, P. Yang, H. Li et al., "MicroRNA-1 inhibits proliferation of hepatocarcinoma cells by targeting endothelin-1," *Life Sciences*, vol. 91, no. 11-12, pp. 440–447, 2012.

[56] K. J. Livak and T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method," *Methods*, vol. 25, no. 4, pp. 402–408, 2001.

[57] H. W. Ahn, R. D. Morin, H. Zhao et al., "MicroRNA transcriptome in the newborn mouse ovaries determined by massive parallel sequencing," *Molecular Human Reproduction*, vol. 16, no. 7, pp. 463–471, 2010.

[58] B. G. Gasperin, M. T. Rovani, R. Ferreira et al., "Functional status of STAT3 and MAPK3/1 signaling pathways in granulosa cells during bovine follicular deviation," *Theriogenology*, vol. 83, no. 3, pp. 353–359, 2015.

[59] S. Roy and T. B. Kornberg, "Paracrine signaling mediated at cell-cell contacts," *BioEssays*, vol. 37, no. 1, pp. 25–33, 2015.

[60] N. Kaivo-Oja, L. A. Jeffery, O. Ritvos, and D. G. Mottershead, "Smad signalling in the ovary," *Reproductive Biology and Endocrinology*, vol. 4, article 21, 2006.

[61] S. A. Pangas, X. Li, E. J. Robertson, and M. M. Matzuk, "Premature luteinization and cumulus cell defects in ovarian-specific Smad4 knockout mice," *Molecular Endocrinology*, vol. 20, no. 6, pp. 1406–1422, 2006.

[62] Q. Li, S. A. Pangas, C. J. Jorgez, J. M. Graff, M. Weinstein, and M. M. Matzuk, "Redundant roles of SMAD2 and SMAD3 in ovarian granulosa cells in vivo," *Molecular and Cellular Biology*, vol. 28, no. 23, pp. 7001–7011, 2008.

[63] J. Liu, X. Du, J. Zhou, Z. Pan, H. Liu, and Q. Li, "MicroRNA-26b functions as a proapoptotic factor in porcine follicular granulosa cells by targeting Sma-and Mad-related protein 4," *Biology of Reproduction*, vol. 91, no. 6, article 146, 2014.

[64] G. Shen, Y. Lin, X. Yang, J. Zhang, Z. Xu, and H. Jia, "MicroRNA-26b inhibits epithelial-mesenchymal transition in hepatocellular carcinoma by targeting USP9X," *BMC Cancer*, vol. 14, article 393, 2014.

[65] T. Mishima, T. Takizawa, S.-S. Luo et al., "MicroRNA (miRNA) cloning analysis reveals sex differences in miRNA expression profiles between adult mouse testis and ovary," *Reproduction*, vol. 136, no. 6, pp. 811–822, 2008.

[66] R. Di, J. He, S. Song et al., "Characterization and comparative profiling of ovarian microRNAs during ovine anestrus and the breeding season," *BMC genomics*, vol. 15, p. 899, 2014.

[67] S. K. Tripurani, C. Xiao, M. Salem, and J. Yao, "Cloning and analysis of fetal ovary microRNAs in cattle," *Animal Reproduction Science*, vol. 120, no. 1-4, pp. 16–22, 2010.

[68] J. Huang, Z. Ju, Q. Li et al., "Solexa sequencing of novel and differentially expressed microRNAs in testicular and ovarian tissues in Holstein Cattle," *International Journal of Biological Sciences*, vol. 7, no. 7, pp. 1016–1026, 2011.

[69] J. R. Miles, T. G. McDaneld, R. T. Wiedmann et al., "MicroRNA expression profile in bovine cumulus-oocyte complexes: possible role of let-7 and miR-106a in the development of bovine

oocytes," *Animal Reproduction Science*, vol. 130, no. 1-2, pp. 16–26, 2012.

[70] P. A. C. 't Hoen, Y. Ariyurek, H. H. Thygesen et al., "Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms," *Nucleic Acids Research*, vol. 36, no. 21, article e141, 2008.

[71] A. V. Sirotkin, D. Ovcharenko, R. Grossmann, M. Lauková, and M. Mlynček, "Identification of microRNAs controlling human ovarian cell steroidogenesis via a genome-scale screen," *Journal of Cellular Physiology*, vol. 219, no. 2, pp. 415–420, 2009.

[72] A. V. Sirotkin, M. Lauková, D. Ovcharenko, P. Brenaut, and M. Mlynček, "Identification of microRNAs controlling human ovarian cell proliferation and apoptosis," *Journal of Cellular Physiology*, vol. 223, no. 1, pp. 49–56, 2010.

[73] C. Yin, J. Zhang, Z. Shi, W. Sun, H. Zhang, and Y. Fu, "Identification and expression of the target gene emx2 of miR-26a and miR-26b in Paralichthys olivaceus," *Gene*, vol. 570, no. 2, pp. 205–212, 2015.

[74] T. Chiyomaru, N. Seki, S. Inoguchi et al., "Dual regulation of receptor tyrosine kinase genes EGFR and c-Met by the tumor-suppressive microRNA-23b/27b cluster in bladder cancer," *International Journal of Oncology*, vol. 46, no. 2, pp. 487–496, 2015.

[75] A. Chinchilla, E. Lozano, H. Daimi et al., "MicroRNA profiling during mouse ventricular maturation: a role for miR-27 modulating Mef2c expression," *Cardiovascular Research*, vol. 89, no. 1, pp. 98–108, 2011.

[76] K. Varendi, A. Kumar, M.-A. Härma, and J.-O. Andressoo, "MiR-1, miR-10b, miR-155, and miR-191 are novel regulators of BDNF," *Cellular and Molecular Life Sciences*, vol. 71, no. 22, pp. 4443–4456, 2014.

[77] N. Nagpal, H. M. Ahmad, B. Molparia, and R. Kulshreshtha, "MicroRNA-191, an estrogen-responsive microRNA, functions as an oncogenic regulator in human breast cancer," *Carcinogenesis*, vol. 34, no. 8, pp. 1889–1899, 2013.

[78] S. E. Palma-Vera, S. Sharbati, and R. Einspanier, "Identification of miRNAs in bovine endometrium through RNAseq and prediction of regulated pathways," *Reproduction in Domestic Animals*, vol. 50, no. 5, pp. 800–806, 2015.

[79] J. C. M. Ricarte-Filho, C. S. Fuziwara, A. S. Yamashita, E. Rezende, M. J. Da-Silva, and E. T. Kimura, "Effects of let-7 microRNA on cell growth and differentiation of papillary thyroid cancer," *Translational Oncology*, vol. 2, no. 4, pp. 236–241, 2009.

[80] S. Vimalraj and N. Selvamurugan, "MicroRNAs expression and their regulatory networks during mesenchymal stem cells differentiation toward osteoblasts," *International Journal of Biological Macromolecules*, vol. 66, pp. 194–202, 2014.

[81] J. Lynch, J. Fay, M. Meehan et al., "MiRNA-335 suppresses neuroblastoma cell invasiveness by direct targeting of multiple genes from the non-canonical TGF-β signalling pathway," *Carcinogenesis*, vol. 33, no. 5, pp. 976–985, 2012.

[82] T. Suzuki, K. Mizutani, A. Minami et al., "Suppression of the TGF-β1-induced protein expression of SNAI1 and N-cadherin by miR-199a," *Genes to Cells*, vol. 19, no. 9, pp. 667–675, 2014.

[83] A. N. Hirshfield, "Development of follicles in the mammalian ovary," *International Review of Cytology*, vol. 124, pp. 43–101, 1991.

[84] A. J. Hsueh, E. Y. Adashi, P. B. Jones, and T. H. Welsh Jr., "Hormonal regulation of the differentiation of cultured ovarian granulosa cells," *Endocrine Reviews*, vol. 5, no. 1, pp. 76–127, 1984.

[85] Q. Song, L. Zhong, C. Chen et al., "MIR-21 synergizes with BMP9 in osteogenic differentiationby activating the BMP9/Smad signaling pathway in murine multilineage cells," *International Journal of Molecular Medicine*, vol. 36, no. 6, pp. 1497–1506, 2015.

[86] H. Liang, C. Xu, Z. Pan et al., "The antifibrotic effects and mechanisms of microRNA-26a action in idiopathic pulmonary fibrosis," *Molecular Therapy*, vol. 22, no. 6, pp. 1122–1133, 2014.

[87] C. Yu, J.-J. Zhou, and H.-Y. Fan, "Studying the functions of TGF-β signaling in the ovary," *Methods in Molecular Biology*, vol. 1344, pp. 301–311, 2016.

[88] S. Li, Q. Fan, S. He, T. Tang, Y. Liao, and J. Xie, "MicroRNA-21 negatively regulates treg cells through a TGF-β1/smad-independent pathway in patients with coronary heart disease," *Cellular Physiology and Biochemistry*, vol. 37, no. 3, pp. 866–878, 2015.

[89] S. D. Fiedler, M. Z. Carletti, X. Hong, and L. K. Christenson, "Hormonal regulation of microRNA expression in periovulatory mouse mural granulosa cells," *Biology of Reproduction*, vol. 79, no. 6, pp. 1030–1037, 2008.

Hindawi
BioMed Research International
Volume 2017, Article ID 5741948, 15 pages
https://doi.org/10.1155/2017/5741948

Research Article

Identification of Candidate Genes Related to Inflammatory Bowel Disease Using Minimum Redundancy Maximum Relevance, Incremental Feature Selection, and the Shortest-Path Approach

Fei Yuan,[1] Yu-Hang Zhang,[2] Xiang-Yin Kong,[2] and Yu-Dong Cai[3]

[1]*Department of Science & Technology, Binzhou Medical University Hospital, Binzhou 256603, Shandong, China*
[2]*Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[3]*School of Life Sciences, Shanghai University, Shanghai 200444, China*

Correspondence should be addressed to Fei Yuan; snowhawkyrf@outlook.com

Received 26 August 2016; Accepted 11 January 2017; Published 14 February 2017

Academic Editor: Mikihiro Fujiya

Identification of disease genes is a hot topic in biomedicine and genomics. However, it is a challenging problem because of the complexity of diseases. Inflammatory bowel disease (IBD) is an idiopathic disease caused by a dysregulated immune response to host intestinal microflora. It has been proven to be associated with the development of intestinal malignancies. Although the specific pathological characteristics and genetic background of IBD have been partially revealed, it is still an overdetermined disease and the blueprint of all genetic variants still needs to be improved. In this study, a novel computational method was built to identify genes related to IBD. Samples from two subtypes of IBD (ulcerative colitis and Crohn's disease) and normal samples were employed. By analyzing the gene expression profiles of these samples using minimum redundancy maximum relevance and incremental feature selection, 21 genes were obtained that could effectively distinguish samples from the two subtypes of IBD and the normal samples. Then, the shortest-path approach was used to search for an additional 20 genes in a large network constructed using protein-protein interactions based on the above-mentioned 21 genes. Analyses of the 41 genes obtained indicate that they are closely associated with this disease.

1. Introduction

Inflammatory bowel disease (IBD) is a common systemic disease that involves the intestinal tissue [1]. It usually refers to chronic conditions that lead to intestinal inflammation and lesions. With the gradual development of inflammation, the intestinal walls become swollen, inflamed, and ulcerogenic [2]. Due to such lesions, several classical symptoms have been considered to be diagnostic indicators. Abdominal pain or cramping, diarrhea multiple times per day, and bloody stools are all classical symptoms of IBD [3]. Such severe symptoms are induced by violent unhealthy inflammation reactions and lesions in the intestinal tissue. Additionally, several complications outside the digestive tract may also be induced by IBD. Mouth sores and skin problems have both been reported in IBD patients [4]. Furthermore, arthritis is also related to IBD, as well as eye problems [5, 6].

As we have mentioned above, several complications have been identified in IBD patients. Such severe complications and related chronic characteristics strongly increase the risk of death [4–6]. In 2013 alone, thousands of people in the world died from IBD [7]. Additionally, IBD has been proven to be associated with colorectal cancer, with a high mortality. Apart from the risk of death, IBD is a lifetime disease, and life with IBD can be quite challenging. The complications associated with IBD and disease relapse severely impact the quality of life [8]. Therefore, the prevention, diagnosis, and treatment of IBD are quite crucial. It is known that IBD is a widespread disease that can develop at any stage of life. However, the disease usually initiates during the teenage years or the early

adulthood of the patients [8]. As we mentioned above, genetic factors participate in the initiation and progression of IBD [9, 10]. Therefore, people with a family history of IBD are at least ten times more likely to suffer from it. Racial factors also contribute to the morbidity of IBD [11].

Although IBD is a very severe and widespread disease, the essential mechanism behind the disease has not been demonstrated clearly. Most people believe that some types of exogenous materials trigger the initiation of inflammation [12, 13]. However, genetic factors may also contribute to the progression of such disease. Several specific genes have been linked to IBD. Pathogenic genes such as IL23R and IL12B play a crucial role in the intestinal immune system, which may induce the initiation of IBD [14, 15]. Several transcriptional factors also contribute to disease progression. The transcription factor NKX2-3 regulates the correct localization of lymphocytes and may further contribute to the immune response in intestinal tissue that induces IBD [16]. Several genes such as ZNF365 and PTGER4 show diversity in different subtypes of IBD and contribute to IBD through their respective methods and pathways [17, 18].

As mentioned above, IBD has several subtypes. Basically, there are two main clinical classifications of IBD: Crohn's disease and ulcerative colitis [17]. Both classifications share the basic symptoms of IBD. However, Crohn's disease can occur anywhere along the digestive tract and typically appears as "skip lesions" between healthy areas [19], while the other type, ulcerative colitis, only involves the colon and rectum. Inflammation and ulcers typically affect only the innermost lining in these areas, with more superficial lesions than those with Crohn's disease [20]. Apart from the differences in clinical symptoms, genetic diversity is also observed between Crohn's disease and ulcerative colitis. Although they share most of the disease-causing genes, genes like ATG16L1, PTGER4, IRGM, and NOD2 have been proven to be specifically related to Crohn's disease but independent with ulcerative colitis [18, 21–23]. The roles of genes such as SLC22A5, ZNF365, and PTPN2 in ulcerative colitis are still unclear, even though they have been proven to be strongly related to Crohn's disease [24, 25].

Because genetic factors have been shown to be related to IBD and its specific subtypes, we developed a new computational method to screen differential expressing genes among different clusters based on a database for Crohn's disease and ulcerative colitis. From the Gene Expression Omnibus (GEO), we obtained the gene expression profiles (information often used to deduce and understand gene functions) for 59 Crohn's disease, 26 ulcerative colitis, and 42 normal samples. Each sample was represented using the expression levels of 12,754 genes. Two feature selection methods, minimum redundancy maximum relevance (mRMR) and incremental feature selection (IFS) [26], and a basic machine learning algorithm, sequential minimal optimization (SMO) [27, 28], were adopted to analyze the gene expression profiles and extract 21 promising candidate genes that could be used to distinguish the samples from the two subtypes of IBD and the normal samples; that is, they may be related to IBD. Furthermore, based on these 21 genes, the shortest-path (SP) approach was employed to identify additional 20 genes in a network constructed using protein-protein interaction (PPI) information. It was concluded that the 41 (21 + 20) genes obtained are closely associated with IBD and can be used to clearly distinguish healthy people from those who have IBD and to identify the subtypes of IBD.

2. Materials and Methods

2.1. Dataset. We downloaded the gene expression profiles of 59 Crohn's disease, 26 ulcerative colitis, and 42 normal samples from GEO under accession number GSE3365 [29]. The expression levels of 12,754 genes were measured using an Affymetrix Human Genome U133A Array. The gene expression profiles were quantile normalized. Each sample was represented using the expression levels of 12,754 genes; that is, each sample was encoded into a 12754-D vector. These features/genes were analyzed to identify the genes that can best discriminate the samples from these three different classes.

2.2. mRMR Method. It is known that some genes can effectively help us discriminate the samples from the three different classes mentioned in Section 2.1, while others offer few or no contributions. To identify these genes, the mRMR method, proposed by Peng et al. [26], was adopted to analyze the gene expression data. The mRMR method employed two criteria, Max-Relevance and Min-Redundancy, to analyze the features. Using the Max-Relevance criterion, the MaxRel feature list can be obtained, in which features are sorted by measuring the relevance between them and sample class labels. Features with high relevance receive high ranks, whereas those with low relevance receive low ranks. It is clearly seen that the rank of a feature in the MaxRel feature list indicates its single contribution to classification. Furthermore, another list, namely, the mRMR feature list, was created using both Max-Relevance and Min-Redundancy criteria. The rank of a feature in this list is determined using the relevance between it and sample class labels and the redundancies between it and the features listed before it. The MaxRel feature list and mRMR feature list in this study were formulated as follows:

MaxRel features list is

$$F_{\text{MaxRel}} = \left[f_1^M, f_2^M, \ldots, f_N^M\right]; \quad (1)$$

mRMR features list is

$$F_{\text{mRMR}} = \left[f_1^m, f_2^m, \ldots, f_N^m\right], \quad (2)$$

where N represents the total number of features. Many investigators have used the mRMR method to analyze various complicated biological systems [30, 31], and it is deemed to be a useful tool for extracting important information from a complicated system. Readers can refer to Peng et al.'s paper [26] for the detailed procedures and principle of this method.

2.3. Prediction Engine. SMO is a type of support vector machine that uses Platt's sequential minimal optimization

algorithm to train and optimize the support vector classifier. The kernels can be polynomial or Gaussian [27, 28]. For implementing our method, we employed the classifier SMO implemented in Weka [32] as the prediction engine.

2.4. Tenfold Cross-Validation. Tenfold cross-validation [33] is a type of cross-validation method that is widely used to examine the performance of a classifier on a given dataset. The given dataset is randomly and equally divided into ten partitions. Samples in each partition are singled out in turn as the test data, while other samples are used to train the classifier. Compared to the jackknife test [34, 35], another popular cross-validation method, this method involves a lower amount of computational time and always yields similar results. Thus, it was used in this study for evaluating the performance of the current prediction engine.

2.5. IFS Method. Using the mRMR method, features/genes were sorted and listed in the MaxRel feature list and mRMR feature list. Because the MaxRel feature list sorted features/genes by only measuring their own contributions to classification, the combination of some features/genes with high ranks in this list is not always an optimal combination for classification. The mRMR feature list is more appropriate for this purpose because it further considers the redundancies between features. The IFS method uses the mRMR feature list and the SMO prediction engine to extract the optimal combination of features/genes as biomarkers. First, according to the mRMR feature list $F_{\text{mRMR}} = [f_1^m, f_2^m, \ldots, f_N^m]$, we constructed N feature set, denoted by $F_1, F_2, \ldots, F_N$, where $F_i = \{f_1^m, f_2^m, \ldots, f_i^m\}$; that is, F_i contained the top i features in the mRMR feature list. Second, for each F_i, SMO was executed on the dataset, in which samples were represented using features in F_i, with its performance evaluated by tenfold cross-validation. Finally, we counted the total prediction accuracy and accuracies for each class. The feature set yielding the highest total prediction accuracy was deemed to be the optimal gene set (G_{optimal}) for IBD, as features in this set may be significant for IBD.

2.6. Network Construction from PPI Information. The optimal gene set G_{optimal} containing some genes closely related to IBD can be obtained using the mRMR and IFS methods. To further mine for other related genes, we constructed a large network from the PPI data and searched for additional candidate genes in the network.

To construct the network, we downloaded the file "protein.links.v9.1.txt.gz" containing the PPI information from STRING (Search Tool for the Retrieval of Interacting Genes/Proteins, version 9.1, http://www.string-db.org/), from which the human PPI data were extracted by identifying lines starting with "9606." A total of 2,425,314 human PPIs involving 20,770 proteins represented using Ensembl IDs were obtained. According to STRING (http://string-db.org/) [36, 37], these PPIs are derived from the following sources: (i) genomic context, (ii) high-throughput experiments, (iii) (conserved) coexpression, and (iv) previous knowledge. Thus, the obtained PPIs contained actual PPIs validated using experiments and predicted PPIs, suggesting that they can be used to widely measure the physical and functional relationships between proteins. Each PPI contained two proteins represented using Ensembl IDs and one score that indicates the strength of the interaction with a range between 150 and 999. The constructed network had 20,770 proteins as nodes. Two nodes were adjacent if and only if the corresponding proteins comprise an interaction that is contained in the 2,425,314 human PPIs. Furthermore, the interaction score was also added to the network. Each edge was assigned a weight defined to be 1,000 minus the corresponding interaction score.

2.7. SP Approach for Searching for Additional Candidates. Network method is an important type of approaches for investigation of disease genes, such as methods based on guilt-by-association (GBA) [38–40] and Random Walk with Restart (RWR) [41–43]. This section proposed another network method for identifying novel disease genes.

It has been elaborated in some previous studies [44–46] that two proteins in an interaction are more likely to share similar functions. It can be induced that the interactive proteins of the proteins encoded by genes in G_{optimal} are also related to IBD. Furthermore, if we consider a series of proteins $p_1, p_2, \ldots, p_s$ such that the consecutive proteins comprise a PPI with a high score and p_1, p_s are proteins encoded by genes in G_{optimal}, $p_2, p_3, \ldots, p_{s-1}$ may also be related to IBD. From the construction of the network mentioned in Section 2.6, the corresponding nodes of $p_1, p_2, \ldots, p_s$ may comprise a shortest path connecting p_1 and p_s. Therefore, for any two genes in G_{optimal}, we searched the shortest path connecting these two genes, thereby collecting a number of shortest paths. Because the endpoints of these paths represented proteins encoded by genes in G_{optimal}, genes on these paths may be related to IBD. Thus, we extracted inner nodes on the obtained shortest paths and their corresponding genes can be obtained. To identify novel genes related to IBD, genes in G_{optimal} were excluded from the obtained genes. The remaining genes were called shortest-path genes for convenience. To identify these shortest-path genes, a measurement, namely, the betweenness [47], was recorded for each shortest-path gene, and it was defined to be the number of shortest paths containing the shortest-path gene.

Because some nodes occupied general hubs in the constructed network, the corresponding genes may always be selected even if we searched for the shortest path connecting any pair of randomly selected genes; some of these genes may be selected as the shortest-path genes obtained as described above. In fact, they have few or no associations with IBD. Thus, a permutation test is necessary to control for this type of gene. The procedures used are as follows:

(1) Randomly produce 1,000 gene sets, say $G_1, G_2, \ldots, G_{1000}$, where the size of each set is the same as that of G_{optimal}.

(2) For each G_i, search for all the shortest paths connecting any pair of genes in G_i and count the betweenness of the shortest-path gene based on these paths.

(3) A total of 1,000 betweenness scores on 1,000 randomly produced gene sets can be obtained for each

shortest-path gene. After comparing the betweenness on $G_{optimal}$, we calculate another measurement, the permutation FDR, for each shortest-path gene, which is defined to be "the number of betweenness scores on randomly produced gene sets that was larger than that on $G_{optimal}$"/1000.

(4) Because it is implied that shortest-path genes with high permutation FDRs are general hubs in the network and not specific to IBD, those with permutation FDRs larger than or equal to 0.05 are excluded. The remaining genes are termed candidate genes.

To select genes with core relationships with IBD from the candidate genes, the human PPIs and their interaction scores were directly used. For each candidate gene g, we checked the scores of the interactions between g and genes in $G_{optimal}$ and selected the maximum value among them as the maximum interaction score of g. If a candidate gene has a high maximum interaction score, this suggests that it is highly related to at least one gene in $G_{optimal}$, indicating that it is more likely to be related to IBD. As 900 is set to be the threshold of the highest confidence cutoff in STRING, we also set 900 as the threshold for the maximum interaction score; that is, genes with maximum interaction scores no less than 900 were finally selected as the candidate genes in this study.

3. Results and Discussion

3.1. Results of the mRMR and IFS Methods. The mRMR method was executed on the dataset containing 59 Crohn's disease, 26 ulcerative colitis, and 42 normal samples, and each sample was represented using the expression levels of 12,754 genes, thereby yielding the MaxRel feature list and mRMR feature list, which are provided in Supplementary Material I in Supplementary Material available online at https://doi.org/10.1155/2017/5741948.

To extract the optimal gene sets for discriminating the samples from two subtypes of IBD and normal samples, the IFS method was used with the mRMR feature list obtained using the mRMR method and SMO as the prediction engine. To reduce computational time and account for the fact that genes with important contributions for discriminating samples from two subtypes of IBD and normal samples are few in number, we only investigated the first 2,000 feature sets. According to the procedures of the IFS method, each feature set can yield four accuracies: three accuracies for three classes and the total prediction accuracy. All of these are provided in Supplementary Material II. Furthermore, an IFS curve was plotted by representing the total prediction accuracy along y-axis and the size of the feature set, that is, the number of features participating in the classification, along x-axis, as shown in Figure 1. It can be seen that the highest total prediction accuracy was 97.64% using the 1170th feature set. The corresponding accuracies for the three classes were 100%, 92.31%, and 97.62%, respectively. Although the accuracies were quite good, the involved features/genes were too many in number, which is not realistic. By carefully checking the IFS curve shown in Figure 1, we observe a sharp increasing trend with more and more features participating in the classification at the beginning of the curve with a rather high total prediction accuracy (93.70%) using the 21st feature set. Then the curve is unstable; increasing trends and decreasing trend occur in succession. Thus, we believe that the first 21 features in the mRMR feature list are more important for discriminating the samples from two subtypes of IBD and normal samples than others and set the optimal gene set $G_{optimal}$ to be the 21st feature set. These 21 genes are listed in Table 1. The associations between these 21 genes and IBD are elaborated in Section 3.4. However, some important IBD-related genes may not be omitted using the mRMR and IFS methods. Based on these genes, the SP approach was applied to discover additional genes related to IBD, which is described in the following sections.

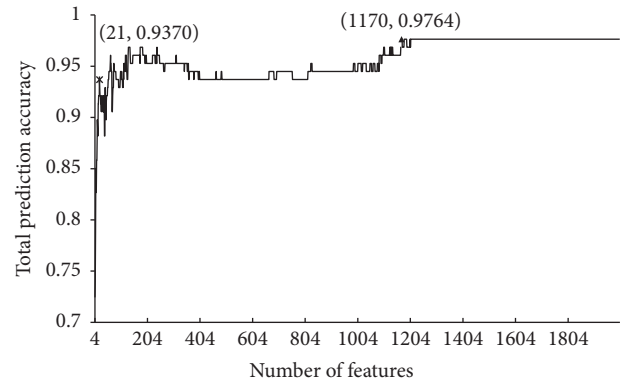


Figure 1: IFS curve. y-axis represents the total prediction accuracy, and x-axis represents the number of features participating in the classification.

3.2. Shortest-Path Genes. As mentioned in Section 3.1, 21 genes were obtained and deemed to be important for discriminating the samples from the two subtypes of IBD and the normal samples. To further identify more candidate genes, we constructed a large network, as described in Section 2.6. These 21 genes were mapped to 20 genes in the network. We searched for all shortest paths connecting any pair of 20 genes, resulting in 190 paths. The graph of these 190 paths is shown in Figure 2, where we can see that there are 110 Ensembl genes on these paths other than the 21 genes obtained in Section 3.1. By mapping to their gene symbols, we obtained 107 shortest-path genes. These genes and their betweenness are listed in Supplementary Material III.

3.3. Additional Candidate Genes. According to Section 2.7, a permutation test was executed to exclude general genes in the network. The obtained permutation FDRs of 107 shortest-path genes are also provided in Supplementary Material III. By setting the threshold of the permutation FDR to be 0.05, 57 candidate genes were obtained, which are listed in Supplementary Material IV.

To select the core genes among the 57 candidate genes, the maximum interaction score of each candidate gene was calculated. These values are also provided in Supplementary Material IV. The threshold of the maximum interaction score

Table 1: Twenty-one important genes for IBD obtained using the mRMR and IFS methods.

GO term/KEGG pathway ID	Description	Rank[a]	Gene symbol	Description
hsa04660	T cell receptor signaling pathway	3	CD247	CD247 molecule
		19	CD4	CD4 molecule
GO: 0001775	Cell activation	5	PF4	Platelet factor 4
GO: 0045449	Regulation of transcription	1	ZNF207	Zinc finger protein 207
		8	EGR3	Early growth response 3
		4	SLTM	SAFB-like, transcription modulator
		14	CNOT8	CCR4-NOT transcription complex, subunit 8
		13	TH1L (NELFCD)	Negative elongation factor complex member C/D
		9	HMGB1	High mobility group box 1
		12	UBE2I	Ubiquitin-conjugating enzyme E2I
GO: 0007243	Protein kinase cascade	2	MARK2	MAP/microtubule affinity-regulating kinase 2
GO: 0031226	Intrinsic to plasma membrane	6	FOLR1	Folate receptor 1 (adult)
		18	SLC22A4	Solute carrier family 22 (organic cation/zwitterion transporter), member 4
		10	LEPROT	Leptin receptor overlapping transcript
		16	CLEC1B	C-type lectin domain family 1, member B
GO: 0006915	Apoptosis	21	RHOT2	ras homolog family member T2
		7	BLCAP	Bladder cancer associated protein
Non-grouped genes		20	ANXA11	Annexin A11
		15	OGT	O-linked N-acetylglucosamine (GlcNAc) transferase
		17	USPL1	Ubiquitin specific peptidase like 1
		11	HIST1H2AC	Histone cluster 1, H2ac

a: this column indicates the ranks of related features in the mRMR feature list.

was set to 900, resulting in 20 candidate genes that are listed in Table 2.

3.4. Analysis of Candidate Genes. Based on feature analysis of 59 Crohn's disease, 26 ulcerative colitis, and 42 normal samples, we obtained 21 genes, listed in Table 1, which may be related to IBD and can help distinguish healthy people from those who have two subtypes of IBD. Furthermore, according to the above 21 genes and the SP approach, we obtained additional 20 candidate genes, listed in Table 2. These genes are also thought to be related to IBD. This section provides some evidence for this claim.

We combined two candidate gene sets and analyzed the biological meaning behind them using Functional Annotation Bioinformatics Microarray Analysis (DAVID) (version 6.7, https://david.ncifcrf.gov/) [48]. The obtained results are provided in Supplementary Material V. According to the results yielded by DAVID, crucial gene ontology (GO) terms and KEGG pathways like hsa04660 (T cell receptor signaling pathway), GO: 0001775 (cell activation), and GO: 0045449 (regulation of transcription) were screened out to be enriched by 41 candidate genes. In addition, the results also gave clues for clustering 41 candidate genes into some groups, which provided convenience for analyzing candidate genes.

3.4.1. Candidate Genes Contributing to T Cell Receptor Signaling Pathway (hsa04660). As mentioned above, IBD is a severe disease induced by inflammation reactions [1]. Considering the core regulatory role of T cells in immune system, it is quite reasonable that various candidate genes contribute to such pathway. Based on SP approach, we identified a specific gene *FOS*. It is also a tumor-associated gene, which encodes a leucine zipper protein that can dimerize with proteins of the JUN family, thereby forming the transcription factor complex AP-1 [49]. Related to crucial pathways such as NF-kB and MAPK, FOS is quite significant in inflammation initiation, especially in the digestive tract [50, 51]. Another calcium-associated gene *PLCG1* was also discovered. PLC1 participates in the intracellular transduction of receptor-mediated tyrosine kinase activators and may participate in the inflammation reaction through a specific function [52, 53]. *LCK* (also known as p56lck) is another predicted IBD-related gene that encodes a functional tyrosine kinase. Similar to ZAP70, LCK also regulates the metabolism and maturation of T cells and may further regulate the inflammation process [54, 55]. In terms of IBD, LCK has been reported to be associated with ulcerative colitis but not with Crohn's disease [56]. Our predicted gene *ZAP70* is a protein tyrosine kinase participating in the development and activation of T cells [57, 58].

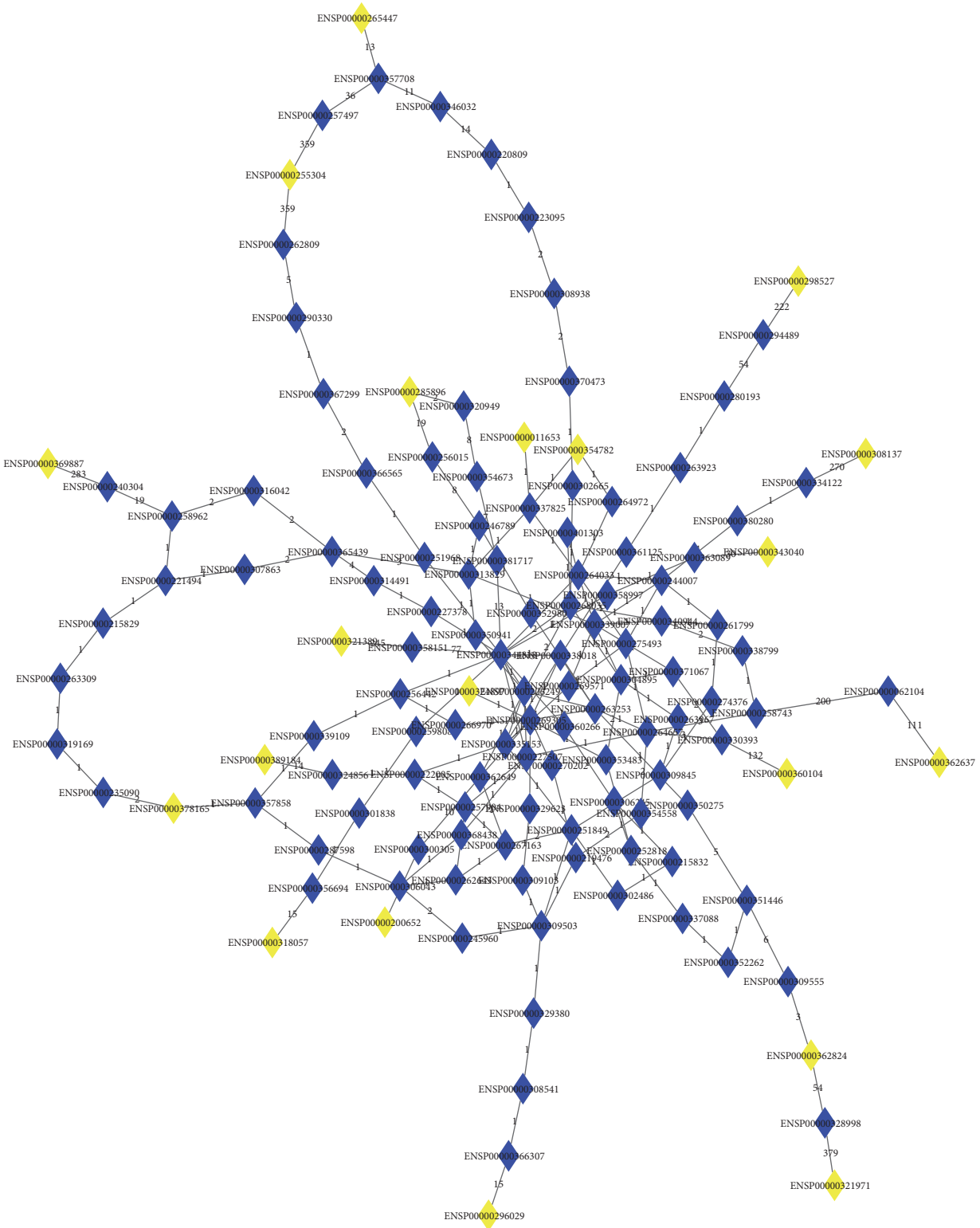


Figure 2: The graph consisting of 190 shortest paths connecting any two genes in the optimal gene set. The yellow diamonds represent genes in the optimal gene set. The blue diamonds represent shortest-path genes. The numbers on the edges represent the edge weights in the network.

Table 2: Twenty candidate genes obtained by SP approach.

GO term/KEGG pathway ID	Description	Gene symbol	Ensembl ID	Description	Betweenness	Permutation FDR	Maximum interaction score	Most related gene in the optimal gene set
hsa04660	T cell receptor signaling pathway	FOS	ENSP00000306245	FBJ murine osteosarcoma viral oncogene homolog	20	0.035	950	CD4
		PLCG1	ENSP00000244007	Phospholipase C, gamma 1	19	0.022	927	CD4
		LCK	ENSP00000337825	LCK protooncogene, Src family tyrosine kinase	21	0.02	999	CD4
		ZAP70	ENSP00000264972	Zeta-chain (TCR) associated protein kinase 70 kDa	16	0.016	999	CD247
GO: 0001775	Cell activation	YWHAZ	ENSP00000309503	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta	19	0.009	962	MARK2
		TLR4	ENSP00000363089	Toll-like receptor 4	19	<0.001	970	HMGB1
		F2	ENSP00000308541	coagulation factor II (thrombin)	19	<0.001	953	PF4
GO: 0045449	Regulation of transcription	HCFC1	ENSP00000309555	Host cell factor C1	36	<0.001	997	OGT
		CNOT1	ENSP00000320949	CCR4-NOT transcription complex, subunit 1	6	0.004	998	CNOT8
		CNOT4	ENSP00000354673	CCR4-NOT transcription complex, subunit 4	6	0.008	987	CNOT8
		TRAK1	ENSP00000328998	Trafficking protein, kinesin binding 1	19	<0.001	946	OGT
		HDAC1	ENSP00000362649	Histone deacetylase 1	19	0.031	967	UBE2I
		BTG1	ENSP00000256015	B-cell translocation gene 1, anti-proliferative	13	0.006	981	CNOT8
		RUNX1	ENSP00000300305	Runt-related transcription factor 1	19	<0.001	927	SLC22A4
GO: 0031226	Intrinsic to plasma membrane	THBD	ENSP00000366307	Thrombomodulin	19	<0.001	985	PF4
		FASLG	ENSP00000356694	Fas ligand (TNF superfamily, member 6)	19	<0.001	985	EGR3
GO: 0042127	Regulation of cell proliferation	STK11	ENSP00000324856	Serine/threonine kinase 11	19	<0.001	986	MARK2
		S100A6	ENSP00000357708	S100 calcium binding protein A6	19	<0.001	987	ANXA11
		SERPINE1	ENSP00000223095	Serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1	18	0.034	933	PF4
		VEGFC	ENSP00000280193	Vascular endothelial growth factor C	19	<0.001	919	PF4

ZAP70 has been reported to be associated with a specific subtype of IBD, Crohn's disease, but not ulcerative colitis [59]. Therefore, the expression level of ZAP70 can be a useful biomarker for distinguishing different subtypes of IBD.

While based on the mRMR and IFS method, we also identified a group of candidate genes. Among them, *CD247* (rank 3 in the mRMR feature list) and *CD4* (rank 19 in the mRMR feature list) are both crucial genes for T cells and have been confirmed to further regulate the inflammation reaction [60, 61]. Our predicted gene CD4 is characteristically expressed in IBD [62]. However, CD4 has also been reported as a differentially expressed gene in Crohn's disease and ulcerative colitis, and it may further serve as a new biomarker for distinguishing these two diseases [63]. Based on our functional clustering, various screened and predicted genes also are enriched in a similar GO term, GO: 0042101 which describes T cell receptor complex as a cellular component, validating the enrichment of T cell receptor signaling pathway of our screened out IBD associated genes.

3.4.2. Candidate Genes Contributing to Cell Activation (GO: 0001775). In Table 2, a highly conserved monooxygenase-associated protein *YWHAZ* was identified as a functional protein in intestinal bowel disease. Such gene is a crucial housekeeping gene that has been proven to be a suitable normalizer for bowel inflammation and cancer [60]. As a functional factor of innate immune response, which is also crucial in intestinal tissues, *TLR4* is predicted to be associated with IBD. TLR4 has been reported as a crucial factor in the innate immune barrier of the intestine [61]. Such factors can be activated by specific factors (FFA, etc.) and further induce the initiation of IBD [64, 65]. Another thrombin-associated gene, *F2* (coagulation factor II) was also identified by the SP approach. F2 and THBD are both coagulation-associated genes. The coagulation process is reported to be associated with Crohn's disease but not with ulcerative colitis, which reflects the differences between various subtypes of IBD [66]. There are two major subtypes of IBD: Crohn's disease and ulcerative colitis. Some candidate genes yielded by mRMR and IFS methods may distinguish these two subtypes. *PF4* (rank 5 in the mRMR feature list), a crucial diagnostic biomarker for IBD, has been clearly reported to be overexpressed in Crohn's disease and thought that it does not play a clear role in ulcerative colitis [67, 68]. PF4 can also separate IBD from normal inflammation, which is crucial for diagnosis [69].

3.4.3. Candidate Genes Contributing to Regulation of Transcription (GO: 0045449). Among the 41 candidate genes, quite a lot of genes contribute to the regulation of transcription, implying the complicated endogenous pathological factors of IBD on multiple levels. Based on mRMR and IFS methods, the candidate gene *ZNF207* (rank 1 in the mRMR feature list), which is a specific microtubule-associated zinc finger protein, may regulate the inflammation of IBD [70]. As a regulator of mitotic chromosome alignment, ZNF207 has been reported to be related to another type of inflammation disorder, chronic obstructive pulmonary disease (COPD). Since both COPD and IBD are localized inflammation involving the mucosal tissue, ZNF207 as our candidate gene may also contribute to inflammatory bowel disease [71]. As a T cell regulator, *EGR3* (rank 8 in the mRMR feature list) was also identified. As a member of the EGR family, EGR3 may be a crucial transcriptional factor for T cells, with high similarity with EGR2 [72]. *SLTM* (rank 4 in the mRMR feature list) acts as a general inhibitor of transcription that eventually leads to apoptosis via the regulation of telomere [73]. Because IBD is associated with abnormal cell death, SLTM may participate in IBD through the regulation of the apoptosis of intestinal cells [74]. *CNOT8* (rank 14 in the mRMR feature list) is a significant predicted gene that interacts with BTG, the regulator of the cell cycle, especially in B cells [75]. Therefore, CNOT8 may indirectly participate in the intestinal inflammation reaction [75, 76]. *TH1L* (rank 13 in the mRMR feature list), as a negative elongation factor complex member C/D (NELFCD), promotes the proliferation of intestinal cells and has been proved to induce carcinoma progression [77]. As a regulator of B cells, *HMGB1* (rank 9 in the mRMR feature list) and its homolog HMGB2 constitute a complex that is differentially expressed in Crohn's disease and ulcerative colitis [78, 79]. Such a complex has also been reported as a new marker of IBD and may be a sensitive marker of mucosal inflammation [80]. As we have mentioned above, our predicted gene CD4 is characteristically expressed in IBD [62]. However, CD4 has also been reported as a differentially expressed gene in Crohn's disease and ulcerative colitis, and it may further serve as a new biomarker for distinguishing these two diseases [63]. *UBE2I* (rank 12 in the mRMR feature list) also regulates the proliferation of intestinal cells [81]. Unlike FOLR1, which we will analyze below, UBE2I is a major part of the SUMO ligases and further promotes the proliferation of intestinal cells via multiple means even under pathological conditions [81, 82].

For the candidate genes obtained by the SP approach, *HCFC1* is a functional nuclear activator. As a unique cleavage signal, it has been reported to be associated with cell cycle regulation and may have a specific function in tumorigenesis [83, 84]. As a part of the CCR4-NOT complex, *CNOT1* is a crucial immune associated gene that is a major cellular mRNA deadenylase and has been reported to participate in several processes related to immune reactions [85]. Regulated by the CCR4-NOT complex, a crucial microRNA, miR155, has been reported to be directly associated with inflammation, which may further reveal the tight connection between CNOT1 and the inflammation reaction [86, 87]. Such functional genes may also participate in the initiation of inflammation and tumors. *CNOT4* is also a part of the CCR-NOT complex, and CNOT4 may act similarly to CNOT1 and contribute to the regulation of the immune reaction [85]. As a functional factor of innate immune response, which is also crucial in intestinal tissues, *TRAK1* is a regulatory gene that may be related to endosome-to-lysosome trafficking and EGF-EGFR interaction [88]. Such an EGF-EGFR interaction is definitely associated with the initiation of bowel inflammation [89]. The candidate gene *HDAC1* regulates the acetylation of specific genes and further participates in the regulation of corresponding functions [90]. Gene acetylation and deacetylation are functional regulatory methods for

cell metabolism, which have been identified in IBD [91–93]. Therefore, HDAC1 may play a regulatory role in the initiation and progression of intestinal bowel diseases. *BTG1* is a functional regulatory gene associated with cell growth and differentiation. Similar to FASLG, it also regulates the apoptosis of specific target cells and may further regulate specific cytokines associated with inflammation such as IFN-γ [94]. Histone deacetylase is commonly used to modify the epigenetic status and regulate gene expression [95]. *RUNX1*, known as runt-related transcription factor 1, is quite crucial in the development of normal hematopoiesis as a part of CBF (core binding factor). Associated with T cell function and TGF-β, RUNX1 has been proven to be quite crucial in inflammation initiation [96, 97]. Considering the strong relationship between IBD and immune reaction, RUNX1, which regulates the function of T cells, may also participate in the initiation of IBD [98].

3.4.4. Candidate Genes Contributing to Protein Kinase Cascade (GO: 0007243). Four functional genes have been clustered into such group. Genes like F2, ZAP70, and TLR4 have already been analyzed above. The gene *MARK2* (rank 2 in the mRMR feature list) may also contribute to the initiation and progression of IBD by interfering with the protein kinase cascade. Inflammation is a basic pathological process regulated by the immune system [99]. Therefore, the immune system plays an irreplaceable role in IBD [100]. Several predicted genes have been confirmed to be associated with the immune system and participate in the immune reaction. MARK2 is a serine/threonine-protein kinase that is the major regulator of cell polarity in epithelial cells, including intestinal epithelial cells. Since immune cells in intestinal system have been proven to be regulated by such gene, the abnormal expression and effect of MARK2 may contribute to the unusual activation of focal inflammatory reaction in the digestive system, which may further promote IBD [101].

3.4.5. Candidate Genes Contributing to Intrinsic to Plasma Membrane (GO: 0031226). Among the candidate genes obtained by the SP approach, *THBD* is an endothelial-specific type I membrane receptor that binds thrombin [102]. As a specific protein in coagulation mechanisms, this receptor has also been reported as a potential inflammation mediator and may have a specific function in IBD [103, 104]. We also predicted a specific member of the TNF family, *FASLG*, as a candidate gene. FASLG has been proven to be involved in the induction of apoptosis triggered by binding to FAS [105]. Members of the TNF family have been widely reported to participate in IBDs by regulating the apoptosis of specific local cells [106, 107].

For candidate genes listed in Table 1, *FOLR1* (rank 6 in the mRMR feature list), the folate receptor, participates in intestinal inflammation via the regulation of folate. Folate is associated with cell apoptosis in bowel tissues and has been reported to be crucial in colonic epithelial cell proliferation implying its potential role in inflammatory bowel diseases [108, 109]. *SLC22A4* (rank 18 in the mRMR feature list) is a homolog of SLC22A5, which has been reported to be crucial in Crohn's disease and is also overexpressed in this disease [110]. However, just like SLC22A5, SLC22A4 has not been confirmed to be overexpressed in ulcerative colitis [111, 112]. The genes mentioned above can distinguish IBD subtypes at the genetic level and may serve as new markers for the classification of inflammation in intestinal tissues. As a receptor of significant biological signals, *LEPROT* (rank 10 in the mRMR feature list) encodes a crucial receptor of GH and has been reported to be associated with the initiation of inflammation in the intestine in mice [113]. IBD has been regarded to be the result of immune systematic disorders and autoimmune reactions [1, 100]. Another gene, *CLEC1B* (rank 16 in the mRMR feature list), also participates in the development of IBD via the regulation of the intestinal immune system, especially the proliferation of NK cells and the formation of lymph nodes [114]. Apart from NK cells, activated cell is also a major part of the immune system and has been shown to be related to IBD [115, 116].

3.4.6. Candidate Genes Contributing to Apoptosis (GO: 0006915). Some candidate genes have been confirmed to participate in the apoptosis processes during the pathological processes of IBD. Apart from genes like SLTM, LCK, F2, FASLG, and BLCAP which we have just analyzed above, the candidate gene *RHOT2* (rank 21 in the mRMR feature list), a mitochondrial GTPase involved in mitochondrial trafficking, has been proven to be crucial regulator of Ca^{2+} in T cells. Thus, RHOT2 may also contribute to IBD [117]. IBD is a common disease involving the digestive system, especially the intestinal tissue [1]. However, IBD has also been shown to be associated with carcinoma in the digestive system, especially colorectal cancer [100]. Several of our predicted genes are also involved in tumor initiation, where cells may have mutated in precancerous lesions, including severe IBD. Most of these genes are related to cell proliferation. *BLCAP* (rank 7 in the mRMR feature list), which was first reported in bladder cancer, regulates the proliferation of cells that are quite common in intestinal tissue of IBD patients [118].

3.4.7. Candidate Genes Contributing to Regulation of Cell Proliferation (GO: 0042127). Among the candidate genes listed in Table 2, several have been confirmed to contribute to cell proliferation, implying the potential role of that during IBD initiation and progression. *STK11*, a functional serine/threonine kinase, regulates the polarity of cells and may participate in tumor suppression [119]. NF-kB is a crucial transcriptional factor that participates in the inflammation process [120]. STK11 (also known as LKB1) directly regulates the function of NF-kB and is definitely associated with inflammation [121]. STK11 also regulates the proliferation and maturation of intestinal cells, which indirectly reflects the regulatory function of STK11 in intestinal tissues. The calcium binding protein *S100A6* is also on our predicted list, and it is located in the cytoplasm and nucleus of a wide range of cells. S100A6 regulates the progression of the cell cycle and the differentiation of specific cells [122]. Considering the tight relationship between IBD and cancer, some of our predicted genes are also associated with tumor initiation [123]. We

also predict as a candidate gene a serine proteinase inhibitor *SERPINE1*, which encodes the principal inhibitor of tissue plasminogen activator (tPA) and urokinase (uPA). Tissue plasminogen activator and urokinase are both associated with inflammation and the process of wound healing [124]. SERPINE1 and proteins in the downstream of its specific pathway have also been reported to be directly associated with IBD as a functional regulator [125, 126]. As a candidate gene, we also predicted an angiogenesis-associated gene *VEGFC*, which regulates angiogenesis and endothelial cell growth [127, 128]. VEGFC has been reported to participate in several intestinal disorders including IBD and some specific digestive tract cancers [129, 130].

3.4.8. Other Candidate Genes. Four candidate genes obtained by mRMR and IFS methods were not clustered into any above group. The candidate gene *ANXA11* (rank 20 in the mRMR feature list) is a predicted gene that regulates the autoimmune reaction. Such a gene has been reported to be related to several autoimmune disorders and may further participate in intestinal inflammation [131]. As a part of the MLL complex, *OGT* (rank 15 in the mRMR feature list) regulates the cell cycle of intestinal cells, including immune cells [132]. Therefore, the abnormity of the OGT gene may induce IBD in various downstream pathways. Another candidate gene, *USPL1* (rank 17 in the mRMR feature list), also participates in intestinal inflammation reaction via the SUMO complex [133]. Large-scale mapping of human protein-protein interactions by mass spectrometry revealed several genes associated with inflammation, especially in the intestine [76]. The last gene, *HIST1H2AC* (rank 11 in the mRMR feature list), is also a candidate gene for cancer. Such gene was first reported in breast cancer and regulates the proliferation of tissue cells, similar to BLCAP [134].

3.5. Comparison of Other Methods. To indicate the effectiveness of the proposed method and the reliability of the obtained genes, we compared our method with other methods. Before making the comparison, 77 validated IBD-related genes were retrieved from [135], which are provided in Supplementary Material VI. These genes were used to test the results yielded by our method and other methods.

DisGeNET (Verison 4.0) [136] is a discovery platform that collects gene-disease associations from several public data sources and the literature. Here, it was used to search IBD-related genes. The obtained material is provided in Supplementary Material VII, from which we extracted 100 genes with high confidence (score > 0.1) as the predicted genes of this method. DAVID 6.7 (https://david.ncifcrf.gov/) [48] was employed again to analyze the biological meanings behind the validated genes, predicted genes by our method, and predicted genes by DisGeNET. The enriched gene ontology (GO) terms and KEGG pathways for three gene lists are listed in Supplementary Material V. It can be observed that 209 GO terms and KEGG pathways were enriched by 77 validated genes, while, for predicted genes by our method and DisGeNET, we obtained 154 and 314 GO terms and KEGG pathways, respectively. For the 154 GO terms and KEGG pathways enriched by 41 predicted genes of our method, 51 (51/154 = 33.12%) were also enriched by 77 validated genes, while there were 117 (117/314 = 37.26%) GO terms and KEGG pathways enriched by both 77 validated genes and 100 predicted genes of DisGeNET.

At a first glance, the performance of the DisGeNET is superior to our method. However, our method still has its advantages. According to our method, 21 genes were extracted by analyzing the gene expression profiles using mRMR, IFS, and SMO methods. In fact, these genes can only help us to distinguish two subtypes of IBD (rather than all subtypes of IBD) and normal samples. Thus, they are parts of IBD-related genes even if they are really IBD-related genes. 20 additional candidate genes were further obtained based on these genes, thereby accessing 41 predicted genes. These 41 predicted genes, in fact, are deemed to be related to two subtypes of IBD rather other all IBD subtypes. On the other hand, 100 predicted genes yielded by DisGeNET considered all subtypes of IBD. It is an important reason why DisGeNET gave the better performance. However, the performance of our method is only slightly lower than that of DisGeNET. Therefore, we believe that the proposed method is still quite effective and the obtained genes can be important and reliable materials for the investigation of IBD.

4. Conclusions

This contribution provides a novel computational method to identify genes related to IBD, which consists of two main steps: (1) analyzing the gene expression profiles and extracting important genes for IBD and (2) applying the shortest-path approach to the network constructed using protein-protein interactions and identifying additional related genes. By analyzing the obtained genes, it is concluded that they have special relationships with IBD, implying that our method is effective. It is also believed that our method has potential applicability for the investigation of other diseases.

Competing Interests

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

Acknowledgments

This paper is supported by National Natural Science Foundation of China (31371335).

References

[1] D. J. Mulder, A. J. Noble, C. J. Justinich, and J. M. Duffin, "A tale of two diseases: the history of inflammatory bowel disease," *Journal of Crohn's and Colitis*, vol. 8, no. 5, pp. 341–348, 2014.

[2] G. Barbara, C. Cremon, and V. Stanghellini, "Inflammatory bowel disease and irritable bowel syndrome: similarities and differences," *Current Opinion in Gastroenterology*, vol. 30, no. 4, pp. 352–358, 2014.

[3] G. Bassotti, E. Antonelli, V. Villanacci, M. Salemme, M. Coppola, and V. Annese, "Gastrointestinal motility disorders in inflammatory bowel diseases," *World Journal of Gastroenterology*, vol. 20, no. 1, pp. 37–44, 2014.

[4] M. Zippi, C. Corrado, R. Pica et al., "Extraintestinal manifestations in a large series of Italian inflammatory bowel disease patients," *World Journal of Gastroenterology*, vol. 20, no. 46, pp. 17463–17467, 2014.

[5] A. Marineata, E. Rezus, C. Mihai, and C. C. Prelipcean, "Extra intestinal manifestations and complications in inflammatory bowel disease," *Revista Medico-Chirurgicala a Societatii de Medici si Naturalisti din Iasi*, vol. 118, no. 2, pp. 279–288, 2014.

[6] P. L. Lakatos, L. Lakatos, L. S. Kiss, L. Peyrin-Biroulet, A. Schoepfer, and S. Vavricka, "Treatment of extraintestinal manifestations in inflammatory bowel disease," *Digestion*, vol. 86, no. 1, pp. 28–35, 2012.

[7] S. Singh, I. J. Kullo, D. S. Pardi, and E. V. Loftus Jr., "Epidemiology, risk factors and management of cardiovascular diseases in IBD," *Nature Reviews Gastroenterology & Hepatology*, vol. 12, no. 1, pp. 26–35, 2015.

[8] J. Ruel, D. Ruane, S. Mehandru, C. Gower-Rousseau, and J.-F. Colombel, "IBD across the age spectrum—is it the same disease?" *Nature Reviews Gastroenterology and Hepatology*, vol. 11, no. 2, pp. 88–98, 2014.

[9] E. Jaźwińska-Tarnawska, I. Jęśkowiak, E. Waszczuk et al., "Genetic polymorphism of ABCB1 gene (C3435T) in patients with inflammatory bowel diseases. Is there any gender dependency?" *Pharmacological Reports*, vol. 67, no. 2, pp. 294–298, 2015.

[10] C. Jakobsen, I. Cleynen, P. S. Andersen et al., "Genetic susceptibility and genotype-phenotype association in 588 Danish children with inflammatory bowel disease," *Journal of Crohn's & Colitis*, vol. 8, no. 7, pp. 678–685, 2014.

[11] G. C. Nguyen, C. A. Chong, and R. Y. Chong, "National estimates of the burden of inflammatory bowel disease among racial and ethnic groups in the United States," *Journal of Crohn's and Colitis*, vol. 8, no. 4, pp. 288–295, 2014.

[12] Y. Zhang and Y. Y. Li, "Inflammatory bowel disease: pathogenesis," *World Journal of Gastroenterology*, vol. 20, no. 1, pp. 91–99, 2014.

[13] P. Flanagan, B. J. Campbell, and J. M. Rhodes, "Bacteria in the pathogenesis of inflammatory bowel disease," *Biochemical Society Transactions*, vol. 39, no. 4, pp. 1067–1072, 2011.

[14] C. W. Lees, J. C. Barrett, M. Parkes, and J. Satsangi, "New IBD genetics: common pathways with other diseases," *Gut*, vol. 60, no. 12, pp. 1739–1753, 2011.

[15] J. Glas, J. Seiderer, J. Wagner et al., "Analysis of IL12B gene variants in inflammatory bowel disease," *PLOS ONE*, vol. 7, no. 3, Article ID e34349, 2012.

[16] G. John, J. P. Hegarty, W. Yu et al., "NKX2-3 variant rs11190140 is associated with IBD and alters binding of NFAT," *Molecular Genetics and Metabolism*, vol. 104, no. 1-2, pp. 174–179, 2011.

[17] K. L. VanDussen, T. C. Liu, D. Li et al., "Genetic variants synthesize to produce paneth cell phenotypes that define subtypes of Crohn's disease," *Gastroenterology*, vol. 146, no. 1, pp. 200–209, 2014.

[18] K. Kabashima, T. Saji, T. Murata et al., "The prostaglandin receptor EP4 suppresses colitis, mucosal damage and CD4 cell activation in the gut," *Journal of Clinical Investigation*, vol. 109, no. 7, pp. 883–893, 2002.

[19] G. Patman, "Crohn's disease: suppression of p21Rac1 signalling contributes to skip-lesion phenotype in Crohn's disease," *Nature Reviews Gastroenterology & Hepatology*, vol. 11, no. 6, article 332, 2014.

[20] J. J. Rumessen, "Ultrastructure of interstitial cells of Cajal at the colonic submuscular border in patients with ulcerative colitis," *Gastroenterology*, vol. 111, no. 6, pp. 1447–1455, 1996.

[21] J. Hampe, A. Franke, P. Rosenstiel et al., "A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1," *Nature Genetics*, vol. 39, no. 2, pp. 207–211, 2007.

[22] M. Parkes, J. C. Barrett, N. J. Prescott et al., "Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility," *Nature Genetics*, vol. 39, no. 7, pp. 830–832, 2007.

[23] T. Hisamatsu, M. Suzuki, H.-C. Reinecker, W. J. Nadeau, B. A. McCormick, and D. K. Podolsky, "CARD15/NOD2 functions as an antibacterial factor in human intestinal epithelial cells," *Gastroenterology*, vol. 124, no. 4, pp. 993–1000, 2003.

[24] E. Leung, J. Hong, A. G. Fraser, T. R. Merriman, P. Vishnu, and G. W. Krissansen, "Polymorphisms in the organic cation transporter genes SLC22A4 and SLC22A5 and Crohn's disease in a New Zealand Caucasian cohort," *Immunology and Cell Biology*, vol. 84, no. 2, pp. 233–236, 2006.

[25] D. F. McCole, "Regulation of epithelial barrier function by the inflammatory bowel disease candidate gene, PTPN2," *Annals of the New York Academy of Sciences*, vol. 1257, no. 1, pp. 108–114, 2012.

[26] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[27] J. Platt, *Fast Training of Support Vector Machines Using Sequential Minimal Optimization*, MIT Press, Cambridge, Mass, USA, 1998.

[28] 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.

[29] M. E. Burczynski, R. L. Peterson, N. C. Twine et al., "Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells," *The Journal of Molecular Diagnostics*, vol. 8, no. 1, pp. 51–61, 2006.

[30] L. Chen, C. Chu, and K. Feng, "Predicting the types of metabolic pathway of compounds using molecular fragments and sequential minimal optimization," *Combinatorial Chemistry & High Throughput Screening*, vol. 19, no. 2, pp. 136–143, 2016.

[31] H. Mohabatkar, M. Mohammad Beigi, and A. Esmaeili, "Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine," *Journal of Theoretical Biology*, vol. 281, no. 1, pp. 18–23, 2011.

[32] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2005.

[33] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the International Joint Conference on Artificial Intelligence*, Lawrence Erlbaum Associates Ltd, Quebec, Canada, 1995.

[34] 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.

[35] 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.

[36] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel, "STRING: a database of predicted functional associations between proteins," *Nucleic Acids Research*, vol. 31, no. 1, pp. 258–261, 2003.

[37] A. Franceschini, D. Szklarczyk, S. Frankild et al., "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research*, vol. 41, no. 1, pp. D808–D815, 2013.

[38] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner, "Predicting disease genes using protein-protein interactions," *Journal of Medical Genetics*, vol. 43, no. 8, pp. 691–698, 2006.

[39] M. Krauthammer, C. A. Kaufmann, T. C. Gilliam, and A. Rzhetsky, "Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's desease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 42, pp. 15148–15153, 2004.

[40] L. Franke, H. van Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes," *The American Journal of Human Genetics*, vol. 78, no. 6, pp. 1011–1025, 2006.

[41] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the Interactome for Prioritization of Candidate Disease Genes," *The American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

[42] R. Jiang, M. Gan, and P. He, "Constructing a gene semantic similarity network for the inference of disease genes," *BMC Systems Biology*, vol. 5, supplement 2, article no. S2, 2011.

[43] H. Shi, J. Xu, G. Zhang et al., "Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes," *BMC Systems Biology*, vol. 7, no. 1, article 101, 2013.

[44] 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.

[45] L. Chen, Z. H. Xing, T. Huang, Y. Shu, G. Huang, and H.-P. Li, "Application of the shortest path algorithm for the discovery of breast cancer-related genes," *Current Bioinformatics*, vol. 11, no. 1, pp. 51–58, 2016.

[46] T. Gui, X. Dong, R. Li, Y. Li, and Z. Wang, "Identification of hepatocellular carcinoma-related genes with a machine learning and network analysis," *Journal of Computational Biology*, vol. 22, no. 1, pp. 63–71, 2015.

[47] M. Kitsak, S. Havlin, G. Paul, M. Riccaboni, F. Pammolli, and H. E. Stanley, "Betweenness centrality of fractal and nonfractal scale-free model networks and tests on real networks," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 75, no. 5, part 2, Article ID 056115, 2007.

[48] 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.

[49] C. Li, H. Li, S. Wang et al., "The c-Fos and c-Jun from *Litopenaeus vannamei* play opposite roles in *Vibrio parahaemolyticus* and white spot syndrome virus infection," *Developmental and Comparative Immunology*, vol. 52, no. 1, pp. 26–36, 2015.

[50] D. Thummuri, M. K. Jeengar, S. Shrivastava et al., "Thymoquinone prevents RANKL-induced osteoclastogenesis activation and osteolysis in an in vivo model of inflammation by suppressing NF-KB and MAPK Signalling," *Pharmacological Research*, vol. 99, pp. 63–73, 2015.

[51] M. G. Welch, K. G. Margolis, Z. Li, and M. D. Gershon, "Oxytocin regulates gastrointestinal motility, inflammation, macromolecular permeability, and mucosal maintenance in mice," *American Journal of Physiology—Gastrointestinal and Liver Physiology*, vol. 307, no. 8, pp. G848–G862, 2014.

[52] C. Auesukaree, H. Tochio, M. Shirakawa, Y. Kaneko, and S. Harashima, "Plc1p, Arg82p, and Kcs1p, enzymes involved in inositol pyrophosphate synthesis, are essential for phosphate regulation and polyphosphate accumulation in Saccharomyces cerevisiae," *Journal of Biological Chemistry*, vol. 280, no. 26, pp. 25127–25133, 2005.

[53] D. Engelberg, R. Perlman, and A. Levitzki, "Transmembrane signaling in *Saccharomyces cerevisiae* as a model for signaling in metazoans: state of the art after 25 years," *Cellular Signalling*, vol. 26, no. 12, pp. 2865–2878, 2014.

[54] Y. J. Chiang and R. J. Hodes, "Regulation of T cell development by c-Cbl: essential role of Lck," *International Immunology*, vol. 27, no. 5, pp. 245–251, 2015.

[55] M. D. Perron, S. Chowdhury, I. Aubry, E. Purisima, M. L. Tremblay, and H. U. Saragovi, "Allosteric noncompetitive small molecule selective inhibitors of CD45 tyrosine phosphatase suppress T-cell receptor signals and inflammation in vivo," *Molecular Pharmacology*, vol. 85, no. 4, pp. 553–563, 2014.

[56] L. S. Toy, X. Y. Yio, A. Lin, S. Honig, and L. Mayer, "Defective expression of gp180, a novel CD8 ligand on intestinal epithelial cells, in inflammatory bowel disease," *Journal of Clinical Investigation*, vol. 100, no. 8, pp. 2062–2071, 1997.

[57] Z. Liao, L. Zhou, C. Wang et al., "Characteristics of TCRζ, ZAP-70, and FcεRIγ Gene Expression in Patients with T- and NK/T-Cell Lymphoma," *DNA and Cell Biology*, vol. 34, no. 3, pp. 201–207, 2015.

[58] C. Sinclair, M. Ono, and B. Seddon, "A Zap70-dependent feedback circuit is essential for efficient selection of CD4 lineage thymocytes," *Immunology and Cell Biology*, vol. 93, no. 4, pp. 406–416, 2015.

[59] D. Bouzid, H. Fourati, A. Amouri et al., "Association of ZAP70 and PTPN6, but not BANK1 or CLEC2D, with inflammatory bowel disease in the tunisian population," *Genetic Testing and Molecular Biomarkers*, vol. 17, no. 4, pp. 321–326, 2013.

[60] M. Krzystek-Korpacka, D. Diakowska, J. Bania, and A. Gamian, "Expression stability of common housekeeping genes is differently affected by bowel inflammation and cancer: implications for finding suitable normalizers for inflammatory bowel disease studies," *Inflammatory Bowel Diseases*, vol. 20, no. 7, pp. 1147–1156, 2014.

[61] W. Wang, T. Xia, and X. Yu, "Wogonin suppresses inflammatory response and maintains intestinal barrier function via TLR4-MyD88-TAK1-mediated NF-κB pathway in vitro," *Inflammation Research*, vol. 64, no. 6, pp. 423–431, 2015.

[62] C. S. De Almeida, V. Andrade-Oliveira, N. O. S. Câmara, J. F. Jacysyn, and E. L. Faquim-Mauro, "Crotoxin from Crotalus durissus terrificus is able to down-modulate the acute intestinal inflammation in mice," *PLoS ONE*, vol. 10, no. 4, Article ID e0121427, 2015.

[63] G. Brandhorst, S. Weigand, C. Eberle et al., "$CD4^+$ immune response as a potential biomarker of patient reported inflammatory bowel disease (IBD) activity," *Clinica Chimica Acta*, vol. 421, pp. 31–33, 2013.

[64] R. A. Gupta, M. N. Motiwala, N. G. Dumore, K. R. Danao, and A. B. Ganjare, "Effect of piperine on inhibition of FFA induced TLR4 mediated inflammation and amelioration of acetic acid induced ulcerative colitis in mice," *Journal of Ethnopharmacology*, vol. 164, pp. 239–246, 2015.

[65] A. T. Cao, S. Yao, A. T. Stefka et al., "TLR4 regulates IFN-γ and IL-17 production by both thymic and induced Foxp3+ Tregs during intestinal inflammation," *Journal of Leukocyte Biology*, vol. 96, no. 5, pp. 895–905, 2014.

[66] D. Kohoutova, M. Pecka, M. Cihak, J. Cyrany, J. Maly, and J. Bures, "Prevalence of hypercoagulable disorders in inflammatory bowel disease," *Scandinavian Journal of Gastroenterology*, vol. 49, no. 3, pp. 287–294, 2014.

[67] T. Bennike, S. Birkelund, A. Stensballe, and V. Andersen, "Biomarkers in inflammatory bowel diseases: current status and proteomics identification strategies," *World Journal of Gastroenterology*, vol. 20, no. 12, pp. 3231–3244, 2014.

[68] M.-A. Meuwis, M. Fillet, L. Lutteri et al., "Proteomics for prediction and characterization of response to infliximab in Crohn's disease: a pilot study," *Clinical Biochemistry*, vol. 41, no. 12, pp. 960–967, 2008.

[69] M.-A. Meuwis, M. Fillet, P. Geurts et al., "Biomarker discovery for inflammatory bowel disease, using proteomic serum profiling," *Biochemical Pharmacology*, vol. 73, no. 9, pp. 1422–1433, 2007.

[70] H. Jiang, X. He, S. Wang et al., "A microtubule-associated zinc finger protein, BuGZ, regulates mitotic chromosome alignment by ensuring Bub3 stability and kinetochore targeting," *Developmental Cell*, vol. 28, no. 3, pp. 268–281, 2014.

[71] S. Bhattacharya, S. Srisuma, D. L. DeMeo et al., "Molecular biomarkers for quantitative and discrete COPD phenotypes," *American Journal of Respiratory Cell and Molecular Biology*, vol. 40, no. 3, pp. 359–367, 2009.

[72] T. Okamura, K. Fujio, M. Shibuya et al., "$CD4^+CD25^-LAG3^+$ regulatory T cells controlled by the transcription factor Egr-2," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 33, pp. 13974–13979, 2009.

[73] R. J. Giannone, H. W. McDonald, G. B. Hurst, R.-F. Shen, Y. Wang, and Y. Liu, "The protein network surrounding the human telomere repeat binding factors TRF1, TRF2, and POT1," *PLoS ONE*, vol. 5, no. 8, Article ID e12407, 2010.

[74] M. Zemljic, B. Pejkovic, I. Krajnc, and S. Lipovsek, "Biological pathways involved in the development of infammatory bowel disease," *Wiener Klinische Wochenschrift*, vol. 126, no. 19-20, pp. 626–633, 2014.

[75] Y. Du, P. Liu, W. Zang et al., "BTG3 upregulation induces cell apoptosis and suppresses invasion in esophageal adenocarcinoma," *Molecular and Cellular Biochemistry*, vol. 404, no. 1-2, pp. 31–38, 2015.

[76] 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 no. 89, 2007.

[77] B. Carvalho, C. Postma, S. Mongera et al., "Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression," *Gut*, vol. 58, no. 1, pp. 79–89, 2009.

[78] M. McDonnell, Y. Liang, A. Noronha et al., "Systemic toll-like receptor ligands modify B-cell responses in human inflammatory bowel disease," *Inflammatory Bowel Diseases*, vol. 17, no. 1, pp. 298–307, 2011.

[79] R. Vitali, L. Stronati, A. Negroni et al., "Fecal HMGB1 is a novel marker of intestinal mucosal inflammation in pediatric inflammatory bowel disease," *The American Journal of Gastroenterology*, vol. 106, no. 11, pp. 2029–2040, 2011.

[80] H. Takaishi, T. Kanai, A. Nakazawa et al., "Anti-high mobility group box 1 and box 2 non-histone chromosomal proteins (HMGB1/HMGB2) antibodies and anti-Saccharomyces cerevisiae antibodies (ASCA): accuracy in differentially diagnosing UC and CD and correlation with inflammatory bowel disease phenotype," *Journal of Gastroenterology*, vol. 47, no. 9, pp. 969–977, 2012.

[81] M. D. Demarque, K. Nacerddine, H. Neyretkahn et al., "Sumoylation by Ubc9 regulates the stem cell compartment and structure and function of the intestinal epithelium in mice," *Gastroenterology*, vol. 140, no. 1, pp. 286–296, 2011.

[82] N. S. Belaguli, M. Zhang, A.-H. Garcia, and D. H. Berger, "PIAS1 Is a GATA4 SUMO ligase that regulates GATA4-dependent intestinal promoters independent of SUMO ligase activity and GATA4 sumoylation," *PLoS ONE*, vol. 7, no. 4, Article ID e35717, 2012.

[83] P. Zhou, Z. Wang, X. Yuan et al., "Mixed Lineage Leukemia 5 (MLL5) protein regulates cell cycle progression and E2F1-responsive gene expression via association with Host Cell Factor-1 (HCF-1)," *The Journal of Biological Chemistry*, vol. 288, no. 24, pp. 17532–17543, 2013.

[84] Y. J. Machida, Y. Machida, A. A. Vashisht, J. A. Wohlschlegel, and A. Dutta, "The deubiquitinating enzyme BAP1 regulates cell growth via interaction with HCF-1," *The Journal of Biological Chemistry*, vol. 284, no. 49, pp. 34179–34188, 2009.

[85] C. Chapat and L. Corbo, "Novel roles of the CCR4-NOT complex," *Wiley Interdisciplinary Reviews: RNA*, vol. 5, no. 6, pp. 883–901, 2014.

[86] A. S. Prabowo, J. van Scheppingen, A. M. Iyer et al., "Differential expression and clinical significance of three inflammation-related microRNAs in gangliogliomas," *Journal of Neuroinflammation*, vol. 12, no. 1, article 97, 2015.

[87] T. S. Elton, H. Selemon, S. M. Elton, and N. L. Parinandi, "Regulation of the MIR155 host gene in physiological and pathological processes," *Gene*, vol. 532, no. 1, pp. 1–12, 2013.

[88] O. Loss and F. A. Stephenson, "Localization of the kinesin adaptor proteins trafficking kinesin proteins 1 and 2 in primary cultures of hippocampal pyramidal and cortical neurons," *Journal of Neuroscience Research*, vol. 93, no. 7, pp. 1056–1066, 2015.

[89] R. A. Isidro, M. L. Cruz, A. A. Isidro et al., "Immunohistochemical expression of SP-NK-1R-EGFR pathway and VDR in colonic inflammation and neoplasia," *World Journal of Gastroenterology*, vol. 21, no. 6, pp. 1749–1758, 2015.

[90] P.-J. Chen, C. Huang, X.-M. Meng, and J. Li, "Epigenetic modifications by histone deacetylases: biological implications and therapeutic potential in liver fibrosis," *Biochimie*, vol. 116, pp. 61–69, 2015.

[91] C. Felice, A. Lewis, A. Armuzzi, J. O. Lindsay, and A. Silver, "Review article: selective histone deacetylase isoforms as potential therapeutic targets in inflammatory bowel diseases," *Alimentary Pharmacology and Therapeutics*, vol. 41, no. 1, pp. 26–38, 2015.

[92] I. A. Lee, A. Kamba, D. Low, and E. Mizoguchi, "Novel methylxanthine derivative-mediated anti-inflammatory effects in inflammatory bowel disease," *World Journal of Gastroenterology*, vol. 20, no. 5, pp. 1127–1138, 2014.

[93] S. Garcia-Maurino, A. Alcaide, and C. Dominguez, "Pharmacological control of autophagy: therapeutic perspectives in inflammatory bowel disease and colorectal cancer," *Current Pharmaceutical Design*, vol. 18, no. 26, pp. 3853–3873, 2012.

[94] H. Lee, S. Cha, M.-S. Lee, G. J. Cho, W. S. Choi, and K. Suk, "Role of antiproliferative B cell translocation gene-1 as an apoptotic sensitizer in activation-induced cell death of brain microglia," *Journal of Immunology*, vol. 171, no. 11, pp. 5802–5811, 2003.

[95] C. A. Hamm and F. F. Costa, "Epigenomes as therapeutic targets," *Pharmacology & Therapeutics*, vol. 151, pp. 72–86, 2015.

[96] H. Liu, A. T. Cao, T. Feng et al., "TGF-β converts Th1 cells into Th17 cells through stimulation of Runx1 expression," *European Journal of Immunology*, vol. 45, no. 4, pp. 1010–1018, 2015.

[97] W. F. Wong, K. Kohu, A. Nakamura et al., "Runx1 deficiency in $CD4^+$ T cells causes fatal autoimmune inflammatory lung disease due to spontaneous hyperactivation of cells," *Journal of Immunology*, vol. 188, no. 11, pp. 5408–5420, 2012.

[98] G. P. Christophi, R. Rong, P. G. Holtzapple, P. T. Massa, and S. K. Landas, "Immune markers and differential signaling networks in ulcerative colitis and Crohn's disease," *Inflammatory Bowel Diseases*, vol. 18, no. 12, pp. 2342–2356, 2012.

[99] D. Salisbury and U. Bronas, "Inflammation and immune system contribution to the etiology of atherosclerosis: mechanisms and methods of assessment," *Nursing Research*, vol. 63, no. 5, pp. 375–385, 2014.

[100] D. Dunkin, S. Mehandru, and J.-F. Colombel, "Immune cell therapy in IBD," *Digestive Diseases*, vol. 32, supplement 1, pp. 61–66, 2014.

[101] J. B. Hurov, T. S. Stappenbeck, C. M. Zmasek et al., "Immune system dysfunction and autoimmune disease in mice lacking Emk (Par-1) protein kinase," *Molecular and Cellular Biology*, vol. 21, no. 9, pp. 3206–3219, 2001.

[102] Y. Miwa, S. Yazaki, M. Iwamoto et al., "Functional difference between membrane-bound and soluble human thrombomodulin," *Transplantation*, vol. 99, no. 4, pp. 702–709, 2015.

[103] M. C. Soult, Y. Dobrydneva, K. H. Wahab, L. D. Britt, and C. J. Sullivan, "Outer membrane vesicles alter inflammation and coagulation mediators," *Journal of Surgical Research*, vol. 192, no. 1, pp. 134–142, 2014.

[104] J. Pekow, U. Dougherty, Y. Huang et al., "Gene signature distinguishes patients with chronic ulcerative colitis harboring remote neoplastic lesions," *Inflammatory Bowel Diseases*, vol. 19, no. 3, pp. 461–470, 2013.

[105] M. Lettau, M. Paulsen, D. Kabelitz, and O. Janssen, "FasL expression and reverse signalling," *Results and Problems in Cell Differentiation*, vol. 49, pp. 49–61, 2009.

[106] T. J. Ślebioda and Z. Kmieć, "Tumour necrosis factor superfamily members in the pathogenesis of inflammatory bowel disease," *Mediators of Inflammation*, vol. 2014, Article ID 325129, 15 pages, 2014.

[107] W. Ben Aleya, I. Sfar, L. Mouelhi et al., "Association of Fas/Apo1 gene promoter (-670 A/G) polymorphism in Tunisian patients with IBD," *World Journal of Gastroenterology*, vol. 15, no. 29, pp. 3643–3648, 2009.

[108] C. V. Antunes, A. E. Hallack Neto, C. R. Nascimento et al., "Anemia in Inflammatory bowel disease outpatients: prevalence, risk factors, and etiology," *BioMed Research International*, vol. 2015, Article ID 728925, 7 pages, 2015.

[109] J. W. Crott, Z. Liu, M. K. Keyes et al., "Moderate folate depletion modulates the expression of selected genes involved in cell cycle, intracellular signaling and folate uptake in human colonic epithelial cell lines," *Journal of Nutritional Biochemistry*, vol. 19, no. 5, pp. 328–335, 2008.

[110] L. Pochini, M. Scalise, M. Galluccio, G. Pani, K. A. Siminovitch, and C. Indiveri, "The human OCTN1 (SLC22A4) reconstituted in liposomes catalyzes acetylcholine transport which is defective in the mutant L503F associated to the Crohn's disease," *Biochimica et Biophysica Acta—Biomembranes*, vol. 1818, no. 3, pp. 559–565, 2012.

[111] K. Repnik and U. Potočnik, "Haplotype in the IBD5 region is associated with refractory Crohn's disease in Slovenian patients and modulates expression of the SLC22A5 gene," *Journal of Gastroenterology*, vol. 46, no. 9, pp. 1081–1091, 2011.

[112] P. Sarlos, D. Varszegi, V. Csongei et al., "Susceptibility to ulcerative colitis in Hungarian patients determined by gene-gene interactions," *World Journal of Gastroenterology*, vol. 20, no. 1, pp. 219–227, 2014.

[113] M. E. Gove, D. H. Rhodes, M. Pini et al., "Role of leptin receptor-induced STAT3 signaling in modulation of intestinal and hepatic inflammation in mice," *Journal of Leukocyte Biology*, vol. 85, no. 3, pp. 491–496, 2008.

[114] C. Benezech, S. Nayar, B. A. Finney et al., "CLEC-2 is required for development and maintenance of lymph nodes," *Blood*, vol. 123, no. 20, pp. 3200–3207, 2014.

[115] J. Shin, I. Yoon, J. Lim et al., "CD4+VEGFR1HIGH T cell as a novel Treg subset regulates inflammatory bowel disease in lymphopenic mice," *Cellular and Molecular Immunology*, vol. 12, no. 5, pp. 592–603, 2015.

[116] S. Maeda, K. Ohno, A. Fujiwara-Igarashi, K. Uchida, and H. Tsujimoto, "Changes in Foxp3-positive regulatory T cell number in the intestine of dogs with idiopathic inflammatory bowel disease and intestinal lymphoma," *Veterinary Pathology*, vol. 53, no. 1, pp. 102–112, 2016.

[117] A. Di Sabatino, L. Rovedatti, R. Kaur et al., "Targeting gut T cell Ca2+ release-activated Ca2+ channels inhibits T cell cytokine production and T-box transcription factor T-bet in inflammatory bowel disease," *Journal of Immunology*, vol. 183, no. 5, pp. 3454–3462, 2009.

[118] I. Gromova, P. Gromov, N. Kroman et al., "Immunoexpression analysis and prognostic value of BLCAP in breast cancer," *PLoS ONE*, vol. 7, no. 9, Article ID e45967, 2012.

[119] A. K.-F. Lo, K.-W. Lo, C.-W. Ko, L. S. Young, and C. W. Dawson, "Inhibition of the LKB1-AMPK pathway by the Epstein-Barr virus-encoded LMP1 promotes proliferation and transformation of human nasopharyngeal epithelial cells," *The Journal of Pathology*, vol. 230, no. 3, pp. 336–346, 2013.

[120] L. Verstrepen and R. Beyaert, "Receptor proximal kinases in NF-κB signaling as potential therapeutic targets in cancer and inflammation," *Biochemical Pharmacology*, vol. 92, no. 4, pp. 519–529, 2014.

[121] Z. Liu, W. Zhang, M. Zhang, H. Zhu, C. Moriasi, and M. Zou, "Liver kinase B1 suppresses lipopolysaccharide-induced nuclear factor κB (NF-κB) activation in macrophages," *Journal of Biological Chemistry*, vol. 290, no. 4, pp. 2312–2320, 2015.

[122] F. Q. Calvo, M. Fillet, D. De Seny et al., "Biomarker discovery in asthma-related inflammation and remodeling," *Proteomics*, vol. 9, no. 8, pp. 2163–2170, 2009.

[123] L. Beaugerie, "IBD and increased risk of cancer: what is the reality?" *La Revue de l'Infirmière*, vol. 63, no. 199, p. 28, 2014.

[124] M. C. Montesinos, A. Desai-Merchant, and B. N. Cronstein, "Promotion of wound healing by an agonist of adenosine A2A receptor is dependent on tissue plasminogen activator," *Inflammation*, vol. 38, no. 6, pp. 2036–2041, 2015.

[125] Z. Shaghaghi, M. Bonyadi, M. H. Somi, and M. Khoshbaten, "Association of plasminogen activator inhibitor-1 gene polymorphism with inflammatory bowel disease in Iranian Azeri Turkish patients," *Saudi Journal of Gastroenterology*, vol. 20, no. 1, pp. 54–58, 2014.

[126] I. E. Koutroubakis, A. Sfiridaki, G. Tsiolakidou, C. Coucoutsi, A. Theodoropoulou, and E. A. Kouroumalis, "Plasma thrombin-activatable fibrinolysis inhibitor and plasminogen activator inhibitor-1 levels in inflammatory bowel disease," *European Journal of Gastroenterology and Hepatology*, vol. 20, no. 9, pp. 912–916, 2008.

[127] E. Balboa-Beltran, M. J. Fernández-Seara, A. Pérez-Muñuzuri et al., "A novel stop mutation in the vascular endothelial growth factor-C gene (VEGFC) results in Milroy-like disease," *Journal of Medical Genetics*, vol. 51, no. 7, pp. 475–478, 2014.

[128] L. Le Guen, T. Karpanen, D. Schulte et al., "Ccbe1 regulates Vegfc-mediated induction of Vegfr3 signaling during embryonic lymphangiogenesis," *Development*, vol. 141, no. 6, pp. 1239–1249, 2014.

[129] C. Tacconi, C. Correale, A. Gandelli et al., "Vascular endothelial growth factor C disrupts the endothelial lymphatic barrier to promote colorectal cancer invasion," *Gastroenterology*, vol. 148, no. 7, pp. 1438.e8–1451.e8, 2015.

[130] S. D'Alessio, C. Correale, C. Tacconi et al., "VEGF-C-dependent stimulation of lymphatic function ameliorates experimental inflammatory bowel disease," *Journal of Clinical Investigation*, vol. 124, no. 9, pp. 3863–3878, 2014.

[131] C. S. Jorgensen, G. Levantino, G. Houen et al., "Determination of autoantibodies to annexin XI in systemic autoimmune diseases," *Lupus*, vol. 9, no. 7, pp. 515–520, 2000.

[132] M. Heuser, D. B. Yap, M. Leung et al., "Loss of MII5 results in pleiotropic hematopoietic defects, reduced neutrophil immune function, and extreme sensitivity to DNA demethylation," *Blood*, vol. 113, no. 7, pp. 1432–1443, 2009.

[133] S. Schulz, G. Chachami, L. Kozaczkiewicz et al., "Ubiquitin-specific protease-like 1 (USPL1) is a SUMO isopeptidase with essential, non-catalytic functions," *EMBO Reports*, vol. 13, no. 10, pp. 930–938, 2012.

[134] J. Pärssinen, E.-L. Alarmo, S. Khan, R. Karhu, M. Vihinen, and A. Kallioniemi, "Identification of differentially expressed genes after PPM1D silencing in breast cancer," *Cancer Letters*, vol. 259, no. 1, pp. 61–70, 2008.

[135] J. Z. Liu, S. van Sommeren, H. Huang et al., "Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations," *Nature Genetics*, vol. 47, no. 9, pp. 979–986, 2015.

[136] J. Piñero, N. Queralt-Rosinach, À. Bravo et al., "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes," *Database*, vol. 2015, Article ID bav028, 2015.

Hindawi
BioMed Research International
Volume 2017, Article ID 3859582, 7 pages
https://doi.org/10.1155/2017/3859582

Research Article

Cancer-Related Triplets of mRNA-lncRNA-miRNA Revealed by Integrative Network in Uterine Corpus Endometrial Carcinoma

Chenglin Liu,[1] Yu-Hang Zhang,[2] Qinfang Deng,[3] Yixue Li,[1,4,5] Tao Huang,[2] Songwen Zhou,[3] and Yu-Dong Cai[6]

[1]*School of Life Sciences and Biotechnology, Shanghai Jiaotong University, Shanghai 200240, China*
[2]*Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[3]*Department of Medical Oncology, Shanghai Pulmonary Hospital, Cancer Institute, Tongji University Medical School, Shanghai 200433, China*
[4]*Key Lab of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[5]*Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai 200433, China*
[6]*School of Life Sciences, Shanghai University, Shanghai 200444, China*

Correspondence should be addressed to Tao Huang; tohuangtao@126.com, Songwen Zhou; zhou_songwen@126.com, and Yu-Dong Cai; cai_yud@126.com

Received 2 September 2016; Revised 28 September 2016; Accepted 22 November 2016; Published 8 February 2017

Academic Editor: Kazuhisa Nishizawa

The regulation of transcriptome expression level is a complex process involving multiple-level interactions among molecules such as protein coding RNA (mRNA), long noncoding RNA (lncRNA), and microRNA (miRNA), which are essential for the transcriptome stability and maintenance and regulation of body homeostasis. The availability of multilevel expression data enables a comprehensive view of the regulatory network. In this study, we analyzed the coding and noncoding gene expression profiles of 301 patients with uterine corpus endometrial carcinoma (UCEC). A new method was proposed to construct a genome-wide integrative network based on variance inflation factor (VIF) regression method. The cross-regulation relations of mRNA, lncRNA, and miRNA were then selected based on clique-searching algorithm from the network, when any two molecules of the three were shown as interacting according to the integrative network. Such relation, which we call the mRNA-lncRNA-miRNA triplet, demonstrated the complexity in transcriptome regulation process. Finally, six UCEC-related triplets were selected in which the mRNA participates in endometrial carcinoma pathway, such as CDH1 and TP53. The multi-type RNAs are proved to be cross-regulated as to each of the six triplets according to literature. All the triplets demonstrated the association with the initiation and progression of UCEC. Our method provides a comprehensive strategy for the investigation of transcriptome regulation mechanism.

1. Introduction

Uterine corpus endometrial carcinoma (UCEC) develops from the cells of the inner lining of the uterus, which is one of the most common female genital cancer threatening the health of women all over the world [1, 2]. Only counting 2012, approximately 320,000 women have been diagnosed and about 76,000 people have died of UCEC, according to incomplete statistics [3]. Most commonly, UCEC occurs in postmenopausal women, due to the unstable level of estrogen after menopause [4]. Smoking, high blood pressure, and being overweight also indirectly relate to uterus diseases via various regulation mechanisms [5–7]. In addition, genetic disorders also contribute to the development of UCEC and associate it with other diseases such as Lynch syndrome and colon cancer [8, 9]. A potential inherited tendency shows in UCEC with an increased risk in women with a family history of endometrial cancer [10]. Clinical diagnosis is according to the symptoms such as postmenopausal vaginal bleeding, enlarged uterus, low abdominal pain, and pelvic cramping [11–13].

The understanding of regulatory network could help to investigate its mechanism and benefit the diagnosis and treatment of UCEC.

The multi-type-molecular regulatory network, especially the interaction network between coding RNAs (mRNAs) and noncoding RNAs, has gained many interests in recent years. Previous reports have revealed that the microRNAs (miRNAs) which are small size noncoding RNAs of about 22 nucleotides and long noncoding RNAs (lncRNAs) which contain more than 200 nucleotides cross-regulate their expression levels and comodulate the expression of mRNAs. On the other hand, mRNAs also affect the expression of noncoding RNAs in specific ways [14, 15]. For example, the long intergenic noncoding RNA lincRNA-p21 has been reported to be downregulated by miRNA let-7. The binding of lincRNA-p21 to JUNB and CTNNB1 mRNAs results in the repression of JunB and β-catenin translation [16]. Another experiment has shown that the depletion of lncRNA highly upregulated in liver cancer (HULC) results in significant deregulation of several genes involved in liver cancer. This lncRNA is upregulated by CREB mRNA which is underregulated by miR372 [17]. Such interaction, which we call the mRNA-lncRNA-miRNA triplet, is essential for the maintenance and regulation of body homeostasis. The aberrance of any of its molecules may influence the stability of multilevel expression and affect the tumorigenesis accordingly.

Recently, the availability of large scaled multilevel expression data provides an opportunity to obtain the comprehensive map of the multi-type-molecular regulatory network. The Cancer Genome Atlas (TCGA) database [18–20], especially the TCGA long noncoding RNAs website, provides the whole-genome profiling of 301 UCEC patients including the expression levels of mRNA, lncRNAs, and miRNAs. Such multidimensional resources allow us to investigate the mRNA-lncRNA-miRNA interactions, understand the transcriptional characteristic of UCEC, and dig deeper into the essential genetic alterations, transcriptional regulations, and posttranscriptional mechanisms throughout its initiation and progression [19].

Here, a new method is built to systematically investigate the mRNA-lncRNA-miRNA interactions in UCEC based on the patient expression profiles downloaded from TCGA long noncoding RNA website. An integrative network of mRNAs, lncRNAs, and miRNAs is constructed using an accurate and extremely efficient algorithm, the variance inflation factor (VIF) regression method. Many mRNA-lncRNA-miRNA triplets, which depict the cross-regulation relations among mRNA, lncRNA, and miRNA, are detected by searching all cliques (that is complete subgraphs with all vertices adjacent to each other) consisting of these three elements. The clique searching problem is a fundamental topic in computer science, which is very important in clustering analysis based on density and grid of data elements [21], and many solutions have been proposed to improve the searching performance. At last, the detected triplets are screened for their biological functions, and three of them are determined as UCEC-related triplets according to KEGG database and published literature. All in all, the proposed algorithm can find out disease-associated transcriptional RNA (miRNAs, lncRNAs, and mRNAs) interactions and may contribute to reveal the potential posttranscriptional regulatory mechanisms of UCEC.

Table 1: Number of genes in raw data and constructed network of 301 UCEC patients.

	mRNA	lncRNA	miRNA	Total
Raw data	20,462	10,419	742	31,623
Network	14,229	4,601	268	19,098
Triplet list	736	1,799	227	2,762

2. Material and Methods

2.1. Datasets. The expression data of UCEC are obtained from TCGA long noncoding RNAs website (http://larssonlab.org/tcga-lncrnas/datasets.php), including the profiles of 20,462 protein coding genes, 10,419 lncRNA genes, and 742 miRNAs from 301 UCEC patients, as shown in Table 1. Specifically, the miRNA expression data are selected from the profiles of noncoding genes by their gene symbols. The expression levels are given as reads per kilobase per million (RPKM) values. The zero values of the expression data are set as the minimum nonzero RPKM of their corresponding sample for the allowance of log transformation. UCEC-related pathway information is adopted from KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.kegg.jp/) database with entry ID "hsa05213". The pathway involves 52 genes including tumor protein coding gene TP53 and cadherin protein coding gene CDH1.

2.2. Integrative Network Construction. The interaction network is built by firstly determine the key factors (mRNA, lncRNA, and miRNA) affecting the expression level of each RNA molecule, respectively, and integrating them into a complete network after that. Hence, each RNA molecule is regarded as the dependent variable in one linear regression model, while all others are treated as the independent variables. In summary, 20,462 + 10,419 + 742 = 31,623 regression models are built where each one is based on 31,622 RNA expression features, and the integrative network is constructed after that.

Due to the large dataset in each regression model with far more features than observations (31,622 versus 301), an efficient regression and feature selection method, the variance inflation factor (VIF) regression algorithm [22] is utilized to select the optimal regulator set that is most related to each target RNA. The algorithm is designed to find the optimal β that can minimize the l_0 penalized sum of squared errors,

$$\arg\min_{\beta} \left\{ \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \lambda_0 \left\| \boldsymbol{\beta} \right\|_{l_0} \right\}, \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$ are n observations and $\mathbf{X} = (\mathbf{x}_1, \ldots \mathbf{x}_p)$ are p predictors, $p \gg n$, $\|\boldsymbol{\beta}\|_{l_0} = \sum_{i=1}^{p} I_{\{\beta_i \neq 0\}}$. Instead of searching over all 2^p subsets for the best β, this algorithm evaluates the marginal correlations of each candidate predictors with the target factor using a small presampled set of data and searches the optimal subset by

including t-statistic correction procedure when adding or removing one variable at a time. The method has shown great efficiency but is also accurate compared to other methods such as LASSO and has comparable accuracy even compared to the most accurate but slowest regression method FoBa. The construction of the VIF regression models is based on program from http://cran.r-project.org/web/packages/VIF/.

Next, the goodness-of-fit for linear regression models is assessed by the adjust coefficient of determination (denoted as adjust R^2). The statistic measures how well the regression line approximates the real data points and can compare the regression model containing different number of regulators. In this paper, the regression model is retained only if it surpasses the adjust R^2 cutoff of 0.8. Regulation relations between the target RNA and its regulators are obtained from the retained regression models. These relations are further integrated as a comprehensive map for the mRNA-lncRNA-miRNA interaction. Note that the constructed network is an undirected graph. The edges are constructed if the two factors are connected by arcs of any direction. It is because the regression model can only identify the regulators of the target based on their gene expression associations but cannot determine if the regulators induce the perturbation of the target or vice versa without prior biological knowledge.

2.3. The mRNA-lncRNA-miRNA Triplet Detection. The detection of mRNA-lncRNA-miRNA triplets from the integrative network is a typical clique problem in computer science. Clique problem tries to search all complete subgraphs with all vertices connected to each other. Here, the size of subgraph is set as three, and the vertices of each subgraph are restricted to contain all of the three RNA types. The subgraphs, called mRNA-lncRNA-miRNA triplets, describe the relations of mRNA, lncRNA, and miRNA with each two of them coregulated according to the VIF regression model. The detection procedure is fulfilled by the "cliques" function in R package *igraph* [23].

Next, the UCEC-related triplets are further screened out if its mRNA participates in the hsa05213 pathway (endometrial cancer, Homo sapiens) according to KEGG database. These triplets are further analyzed for their interactions and biological functions as to UCEC according to literature.

3. Results and Discussion

3.1. Structure of the Integrative Network. The whole-genome integrative network of mRNA, lncRNA, and miRNA is constructed based on their cross-regulation relations using VIF regression. Totally, 19,098 factors are included in the integrative network, composed of 14229 coding mRNAs, 4,601 lncRNA, and 268 miRNA. On the network, each RNA is regulated by an average of 30 factors. The protein coding gene-gene interactions dominate the integrative network, as the expression levels of most coding genes are largely affected by only the coding mRNAs. Noncoding RNAs tend to have more interactions with noncoding RNAs instead of coding RNAs, which implies the extensive cross-talk of noncoding RNAs in their regulation of transcriptome and posttranscriptome. The details of the integrative network can be referred in Supplementary Material S1 available online at https://doi.org/10.1155/2017/3859582.

3.2. Candidate mRNA-lncRNA-miRNA Triplets of UCEC. By restricting the vertices types of clique problem to have all three RNA types, 14,416 mRNA-lncRNA-miRNA triplets are detected from the integrative network. These triplets involve 736 coding mRNAs, 1,799 lncRNAs, and 227 miRNAs, and provide a comprehensive map for the mRNA-lncRNA-miRNA interaction. The relatively small number of coding mRNAs compared to the noncoding RNAs indicates that many coding genes are coregulated by multiple lncRNAs and miRNAs. Extensive cross-talks exist in the regulatory process of noncoding RNAs, which also explain the complexity of transcriptome regulation process. The list of the detected triplets can be found in Supplementary Material S2.

Next, the triplet is considered as UCEC-related if its mRNA participates in hsa05213 endometrial cancer pathway. Note that the mRNA-lncRNA-miRNA cross-interaction is a very special interaction case that the genes in the selected triplet have little chance to be enriched in the pathway. However, studies have shown that the mutations in a pathway are mutual exclusive, and only one functional gene mutation is enough to perturb the pathway [24–26]. Hence, any triplet having overlapped genes with the pathway may contribute to the progression of cancer. Here, six triplets related to hsa05213 are detected and are retained for further analysis, as shown in Figure 1. The mRNA, lncRNA, and miRNA are labeled as red, green, and yellow, and the cross-regulation relations are shown as an undirected 3-vertex graph. Four of the six triplets, as shown in the first graph in Figure 1, involve the same mRNA and miRNA, but different lncRNAs, that is, mRNA CDH1-lncRNA (RP4-591L5.1, CTA.929C8.5.1, U47924.27.1, and AP006285.7.1)-miRNA miR128-1. The other two triplets are mRNA CDH1-lncRNA AP006285.7.1-miRNA miR126 and mRNA TP53-lncRNA CTD-2008N3.1.1-miRNA miR203, respectively.

3.3. Interaction and Biological Function of UCEC-Related Genes and Triplets. First, we focus on the mRNA-miRNA interaction and biological functions of the first set of triplets in Figure 1, which involves CDH1 and miR128-1, as mentioned above. CDH1 encodes a classical cadherin from cadherin superfamily, which is a calcium-dependent cell adhesion regulatory protein [27, 28]. CDH1 contributes the cell adhesion, mobility, and proliferation in specific microenvironment, especially in tumor [29]. As for UCEC, CDH1 contributes the initiation and invasion of endometrial cancer through its specific role in epithelial-mesenchymal transition (EMT) [30–32]. Additionally, miR128-1 has been proved to interact with the expression product of CDH1 cadherin and participate in the regulation of EMT in prostate cancer stem during the tumorigenesis [33, 34]. Apart from that, miR128-1 also participates in the regulation of progression and EMT in glioblastoma [35]. In fact, miR128-1 interacts with CDH1 coding protein cadherin via a specific upstream protein Bmi1 which is the direct target of miR128-1 [34, 35].

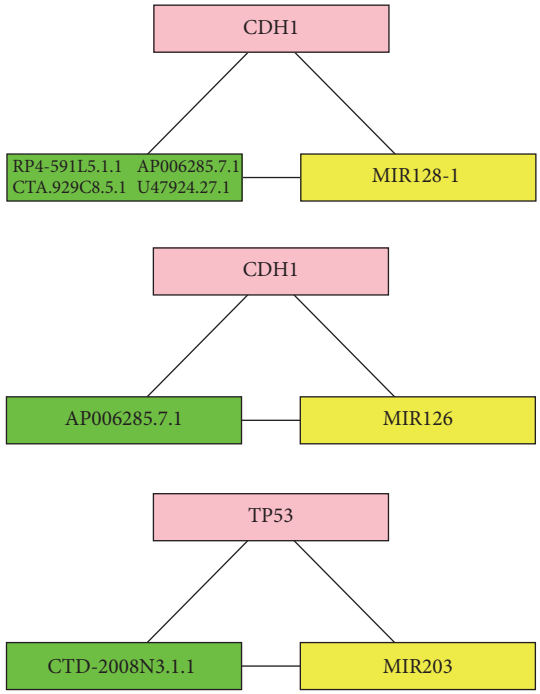


Figure 1: UCEC-related triples screened out from the integrative network. The mRNA, lncRNA, and miRNA are labeled as red, green, and yellow. The cross-regulation relations are described as an undirected 3-vertex graph. The four triplets including the same mRNA and miRNA but different lncRNAs are presented in the top figure. The other two triplets are presented in the following.

Next, we consider the four lncRNAs in the triplets. The first lncRNA RP4-591L5.1 is a crucial lncRNA which binds a specific miRNA miR218. MiR218 contributes the cellular chemosensitivity, migration, and invasion, which may further influence the cadherin regulation and associate with the function of miR128-1 [36, 37]. Another lncRNA, CTA.929C8.5.1, also called lnc-CRYBA4-7:1, has been predicted to be interacted with miR4268. MiR4268 is a rare miRNA with a special 3D structure. It has been proved to participate in the maintenance of stemness and may activate the initiation process of tumor in specific environment [38]. Additionally, as the stemness of cancer cells is associated with tumor migration and has specific relationship with the process of EMT [39, 40], the interaction of CTA.929C8.5.1 with miR4268 may affect the stemness of tumor cells and further have a specific influence on EMT, which explains its potential relationship with miR218 and gene CDH1 in the triplet. The third lncRNA U47924.27.1, also named lnc-PTPN6-1:1, is a unique lncRNA that is the target of several functional miRNAs, such as miR139, and may participate in the initiation of several tumors especially in hematopoietic malignancy [41]. Therefore, it is reasonable that, in UCEC, such lncRNA may play a similar way to interact with the miRNA and lncRNAs mentioned above and contribute to the tumor initiation and progression. Apart from that, U47924.27.1 is associated with PTEN, while CDH1 coding protein cadherin is also associated with PTEN cascades. Hence, this lncRNA may also have interaction with CDH1 [42]. The last lncRNA AP006285.7.1, also annotated as lnc-KRTAP5-4-1:1, has been proved to be associated with miR513a through sequence analysis. MiR513a may play its specific role in various cancer types and mainly regulate the proliferation of cells and contributes to the modeling of inflammation environment [43, 44]. Such regulatory functions may participate in EMT process and further promote the migration of tumor cells, which further explains the interaction of AP006285.7.1 with other factors in the triplet. In summary, all elements in the triplet contribute cohesively to the initiation and progression of UCEC and may exert influence on EMT process.

As to the triplet CDH1-lncRNA AP006285.7.1-miRNA miR126, we only investigate the functions of miR126 since the other two have been mentioned above. MiR126 has been proved as a predictive and diagnosis marker of esophageal cancer [45]. It is also associated with cell adhesion and migration and may contribute to the cadherin regulation in a similar way with other miRNAs (miR99a, miR200, etc.) [46, 47]. Therefore, miR126, lncRNA AP006285.7.1, and CDH1 can be clustered together because of their specific function and contribution to UCEC. This triplet focuses more on the cell adhesion instead of EMT progression and concentrates on the progression and migration process of the tumorigenesis of UCEC.

The next triplet is mRNA TP53-lncRNA CTD-2008N3.1.1-miRNA miR203. TP53 is the most famous tumor suppressor gene which generally contributes to every common type of tumor including endometrial cancer [48, 49]. As a multifunctional gene, TP53 also interacts with several crucial miRNAs (miR181, miR34a, miR520g, etc.) especially in various tumor tissues [50–53]. Consistent with our screen triplet, the interaction of TP53 and miR203 has been proved by several publications [54, 55]. Such interaction is quite crucial for certain kind of tumor especially for colon cancer [55]. As to lncRNA CTD-2008N3.1.1, which is also called lnc-CTD-2012 M11.2.1-1:1, it has been reported to associate with several miRNAs using computational prediction, which may have its specific way to interact with TP53 and miR203 in UCEC [56–58]. Additionally, CTD-2008N3.1.1 interacts with miR331 which participates in the tumorigenesis of various tumor types [56, 59–61]. Furthermore, CTD-2008N3.1.1 is a specific lncRNA originating from CTD sequence [62]. Since TP53 has been reported to be associated with several CTD structures in different tumor types, such screened lncRNA may also interact with TP53 and have its specific function in the process of UCEC initiation and progression [63, 64].

4. Conclusion

In summary, all of the selected triplets have been partially or fully confirmed to be associated with tumorigenesis especially in UCEC. Moreover, some of our preliminarily screened genes which are not included in these six triplets are also proved to be UCEC-associated and may have their specific function in the process of tumorigenesis. For example, miR204 is a specific miRNA in non-small cell lung cancer, which can specifically regulate the metastasis of tumor

cells [65]. Such miRNA may have similar function in UCEC. Another miRNA, miR320, has been reported as a functional regulatory miRNA in stage I endometrioid endometrial carcinoma [66]. All in all, based on the expression profile of UCEC, our proposed method can cluster miRNAs, lncRNAs, and coding genes into functional interacted groups. Such an algorithm can also be applied to other cancer types and benefit the deeper understanding of the transcriptome regulatory mechanisms, and the cross-talk of multilevel RNAs such as miRNAs and lncRNAs. Additionally, the transcriptional level regulation network prediction helps to reveal the posttranscriptional regulation in tumors and other severe diseases.

Disclosure

Chenglin Liu, Yu-Hang Zhang, and Qinfang Deng are co-first authors.

Competing Interests

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

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant no. 81301994), the Natural Science Foundation of Shanghai Science and Technique Committee (Grant no. 13ZR1434700), Shanghai Sailing Program, and the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (Grant no. 2016245).

References

[1] S. Lockwood, *Contemporary Issues in Women's Cancers*, Jones and Bartlett Publishers, Sudbury, Mass, USA, 2009.

[2] A. Keen and E. Lennan, *Women's Cancers*, John Wiley & Sons, West Sussex, UK, 2011.

[3] L. Ross, "2012 International Cancer Education Conference Proceedings," *Journal of Cancer Education*, vol. 30, supplement 1, pp. 1–96, 2015.

[4] Y. Chen, Q. Huang, Q. Chen et al., "The inflammation and estrogen metabolism impacts of polychlorinated biphenyls on endometrial cancer cells," *Toxicology in Vitro*, vol. 29, no. 2, pp. 308–313, 2015.

[5] M. Al-Zoughool, L. Dossus, R. Kaaks et al., "Risk of endometrial cancer in relationship to cigarette smoking: results from the EPIC study," *International Journal of Cancer*, vol. 121, no. 12, pp. 2741–2747, 2007.

[6] M. L. McCarroll, S. Armbruster, R. J. Pohle-Krauza et al., "Feasibility of a lifestyle intervention for overweight/obese endometrial and breast cancer survivors using an interactive mobile application," *Gynecologic Oncology*, vol. 137, no. 3, pp. 508–515, 2015.

[7] Y. Zhang, H. Liu, S. Yang, J. Zhang, L. Qian, and X. Chen, "Overweight, obesity and endometrial cancer risk: results from a systematic review and meta-analysis," *International Journal of Biological Markers*, vol. 29, no. 1, pp. e21–e29, 2014.

[8] J. T. Rabban, S. M. Calkins, A. N. Karnezis et al., "Association of tumor morphology with mismatch-repair protein status in older endometrial cancer patients: implications for universal versus selective screening strategies for lynch syndrome," *American Journal of Surgical Pathology*, vol. 38, no. 6, pp. 793–800, 2014.

[9] J. Y. Lee, H. J. Kim, E. H. Lee, H. W. Lee, J. Kim, and M. K. Kim, "One case of endometrial cancer occurrence: over 10 years after colon cancer in Lynch family," *Obstetrics & Gynecology Science*, vol. 56, no. 6, pp. 408–411, 2013.

[10] A. Dutt, H. B. Salvesen, H. Greulich, W. R. Sellers, R. Beroukhim, and M. Meyerson, "Somatic mutations are present in all members of the AKT family in endometrial carcinoma," *British Journal of Cancer*, vol. 101, no. 7, pp. 1218–1219, 2009.

[11] M. M. El Behery, M. H. Huda, E. M. Kamal, and A. E. Shehatta, "Diagnostic accuracy of uterine fluid lactate dehydrogenase isoenzyme activity profile and vaginal ultrasound in detecting endometrial cancer in women with postmenopausal bleeding," *Archives of Gynecology and Obstetrics*, vol. 281, no. 4, pp. 717–721, 2010.

[12] C. Neppe, R. Land, and A. Obermair, "Wrigley forceps to deliver a bulky uterus following a total laparoscopic hysterectomy for endometrial cancer," *Australian and New Zealand Journal of Obstetrics and Gynaecology*, vol. 45, no. 5, pp. 444–445, 2005.

[13] K. A. Donovan, A. R. Boyington, P. L. Judson, and J. F. Wyman, "Bladder and bowel symptoms in cervical and endometrial cancer survivors," *Psycho-Oncology*, vol. 23, no. 6, pp. 672–678, 2014.

[14] E. S. Martens-Uzunova, R. Böttcher, C. M. Croce, G. Jenster, T. Visakorpi, and G. A. Calin, "Long noncoding RNA in prostate, bladder, and kidney cancer," *European Urology*, vol. 65, no. 6, pp. 1140–1151, 2014.

[15] A. O. Ribeiro, C. R. G. Schoof, A. Izzotti, L. V. Pereira, and L. R. Vasques, "MicroRNAs: modulators of cell identity, and their applications in tissue engineering," *MicroRNA*, vol. 3, no. 1, pp. 45–53, 2014.

[16] J.-H. Yoon, K. Abdelmohsen, and M. Gorospe, "Functional interactions among microRNAs and long noncoding RNAs," *Seminars in Cell and Developmental Biology*, vol. 34, pp. 9–14, 2014.

[17] J.-H. Yoon, K. Abdelmohsen, S. Srikantan et al., "LincRNA-p21 suppresses target mRNA translation," *Molecular Cell*, vol. 47, no. 4, pp. 648–655, 2012.

[18] J. N. Weinstein, E. A. Collisson, G. B. Mills et al., "The Cancer Genome Atlas Pan-Cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[19] Cancer Genome Atlas Research Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, pp. 609–615, 2011.

[20] J. N. Weinstein, E. A. Collisson, G. B. Mills et al., "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[21] S. Luccioli, E. Ben-Jacob, A. Barzilai, P. Bonifazi, and A. Torcini, "Clique of functional hubs orchestrates population bursts in developmentally regulated neural networks," *PLoS Computational Biology*, vol. 10, no. 9, Article ID e1003823, 2014.

[22] D. Lin, D. P. Foster, and L. H. Ungar, "VIF regression: a fast regression algorithm for large data," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 232–247, 2011.

[23] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal Complex Systems*, Article ID 1695, 2006.

[24] J. Zhao, S. Zhang, L.-Y. Wu, and X.-S. Zhang, "Efficient methods for identifying mutated driver pathways in cancer," *Bioinformatics*, vol. 28, no. 22, pp. 2940–2947, 2012.

[25] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control," *Nature Medicine*, vol. 10, no. 8, pp. 789–799, 2004.

[26] C.-H. Yeang, F. McCormick, and A. J. Levine, "Combinatorial patterns of somatic gene mutations in cancer," *The FASEB Journal*, vol. 22, no. 8, pp. 2605–2622, 2008.

[27] M. Silies and C. Klämbt, "APC/CFzr/Cdh1-dependent regulation of cell adhesion controls glial migration in the Drosophila PNS," *Nature Neuroscience*, vol. 13, no. 11, pp. 1357–1364, 2010.

[28] Q.-W. Deng, B.-S. He, Y.-Q. Pan et al., "Roles of E-cadherin (CDH1) genetic variations in cancer risk: a meta-analysis," *Asian Pacific Journal of Cancer Prevention*, vol. 15, no. 8, pp. 3705–3713, 2014.

[29] Z. R. Qian, T. Sano, K. Yoshimoto et al., "Tumor-specific downregulation and methylation of the CDH13 (H-cadherin) and CDH1 (E-cadherin) genes correlate with aggressiveness of human pituitary adenomas," *Modern Pathology*, vol. 20, no. 12, pp. 1269–1277, 2007.

[30] E. Płuciennik, M. Nowakowska, K. Pospiech et al., "The role of WWOX tumor suppressor gene in the regulation of EMT process via regulation of CDH1-ZEB1-VIM expression in endometrial cancer," *International Journal of Oncology*, vol. 46, no. 6, pp. 2639–2648, 2015.

[31] S.-M. Hsiao, M.-W. Chen, C.-A. Chen et al., "The H3K9 methyltransferase G9a represses E-cadherin and is associated with myometrial invasion in endometrial cancer," *Annals of Surgical Oncology*, vol. 22, pp. 1556–1565, 2015.

[32] K. Banno, M. Yanokura, M. Iida, K. Masuda, and D. Aoki, "Carcinogenic mechanisms of endometrial cancer: involvement of genetics and epigenetics," *Journal of Obstetrics and Gynaecology Research*, vol. 40, no. 8, pp. 1957–1967, 2014.

[33] Y. Zhang, T. Chao, R. Li et al., "MicroRNA-128 inhibits glioma cells proliferation by targeting transcription factor E2F3a," *Journal of Molecular Medicine*, vol. 87, no. 1, pp. 43–51, 2009.

[34] R. Nanta, D. Kumar, D. Meeker et al., "NVP-LDE-225 (Erismodegib) inhibits epithelial-mesenchymal transition and human prostate cancer stem cell growth in NOD/SCID IL2Rγ null mice by regulating Bmi-1 and microRNA-128," *Oncogenesis*, vol. 2, article no. e42, 2013.

[35] J. Fu, M. Rodova, R. Nanta et al., "NPV-LDE-225 (Erismodegib) inhibits epithelial mesenchymal transition and self-renewal of glioblastoma initiating cells by regulating miR-21, miR-128, and miR-200," *Neuro-Oncology*, vol. 15, no. 6, pp. 691–706, 2013.

[36] R. Dong, H. Qiu, G. Du, Y. Wang, J. Yu, and C. Mao, "Restoration of microRNA-218 increases cellular chemosensitivity to cervical cancer by inhibiting cell-cycle progression," *Molecular Medicine Reports*, vol. 10, no. 6, pp. 3289–3295, 2014.

[37] B. Peng, D. Li, M. Qin et al., "MicroRNA218 inhibits glioma migration and invasion via inhibiting glioma-associated oncogene homolog 1 expression at N terminus," *Tumor Biology*, vol. 35, no. 4, pp. 3831–3837, 2014.

[38] L. A. Goff, J. Davila, M. R. Swerdel et al., "Ago2 immunoprecipitation identifies predicted MicroRNAs in human embryonic stem cells and neural precursors," *PLoS ONE*, vol. 4, no. 9, Article ID e7192, 2009.

[39] T. Saito and K. Mimori, "Cancer stemness and circulating tumor cells," *Nihon rinsho. Japanese journal of clinical medicine*, vol. 73, no. 5, pp. 806–810, 2015.

[40] A. W. Fender, J. M. Nutter, T. L. Fitzgerald, F. E. Bertrand, and G. Sigounas, "Notch-1 promotes stemness and epithelial to mesenchymal transition in colorectal cancer," *Journal of Cellular Biochemistry*, vol. 116, no. 11, pp. 2517–2527, 2015.

[41] M. F. Alemdehy, J. R. Haanstra, H. W. J. de Looper et al., "Interstrand cross-link induced miR139-3p and miR199a-3p have opposite roles in hematopoietic cell expansion and leukemic transformation," *Blood*, vol. 125, no. 25, pp. 3937–3948, 2015.

[42] Y. Chen, Y. Sun, L. Chen et al., "MiRNA-200c increases the sensitivity of breast cancer cells to doxorubicin through the suppression of E-cadherin-mediated PTEN/Akt signaling," *Molecular Medicine Reports*, vol. 7, no. 5, pp. 1579–1584, 2013.

[43] X. Chen, G. Zhao, F. Wang et al., "Upregulation of miR-513b inhibits cell proliferation, migration, and promotes apoptosis by targeting high mobility group-box 3 protein in gastric cancer," *Tumor Biology*, vol. 35, no. 11, pp. 11081–11089, 2014.

[44] A.-Y. Gong, R. Zhou, G. Hu et al., "MicroRNA-513 regulates B7-H1 translation and is involved in IFN-γ-induced B7-H1 expression in cholangiocytes," *Journal of Immunology*, vol. 182, no. 3, pp. 1325–1333, 2009.

[45] N. S. Sakai, E. Samia-Aly, M. Barbera, and R. C. Fitzgerald, "A review of the current understanding and clinical utility of miRNAs in esophageal cancer," *Seminars in Cancer Biology*, vol. 23, no. 6, pp. 512–521, 2013.

[46] D. Li, X. Li, W. Cao, Y. Qi, and X. Yang, "Antagonism of microRNA-99a promotes cell invasion and down-regulates E-cadherin expression in pancreatic cancer cells by regulating mammalian target of rapamycin," *Acta Histochemica*, vol. 116, no. 5, pp. 723–729, 2014.

[47] L. Romero-Pérez, M. Á. López-García, J. Díaz-Martín et al., "ZEB1 overexpression associated with E-cadherin and microRNA-200 downregulation is characteristic of undifferentiated endometrial carcinoma," *Modern Pathology*, vol. 26, no. 11, pp. 1514–1524, 2013.

[48] B. Leroy, M. Anderson, and T. Soussi, "TP53 mutations in human cancer: database reassessment and prospects for the next decade," *Human Mutation*, vol. 35, no. 6, pp. 672–688, 2014.

[49] M. Chmelarova, S. Kos, E. Dvorakova et al., "Importance of promoter methylation of GATA4 and TP53 genes in endometrioid carcinoma of endometrium," *Clinical Chemistry and Laboratory Medicine*, vol. 52, no. 8, pp. 1229–1234, 2014.

[50] F. Ganci, A. Sacconi, N. B. Ben-Moshe et al., "Expression of TP53 mutation-associated microRNAs predicts clinical outcome in head and neck squamous cell carcinoma patients," *Annals of Oncology*, vol. 24, no. 12, pp. 3082–3088, 2013.

[51] Y. K. Cheah, R. W. Cheng, S. K. Yeap, C. H. Khoo, and H. S. See, "Analysis of TP53 gene expression and p53 level of human hypopharyngeal FaDu (HTB-43) head and neck cancer cell line after microRNA-181a inhibition," *Genetics and Molecular Research*, vol. 13, no. 1, pp. 1679–1683, 2014.

[52] A. Dufour, G. Palermo, E. Zellmeier et al., "Inactivation of TP53 correlates with disease progression and low miR-34a expression in previously treated chronic lymphocytic leukemia patients," *Blood*, vol. 121, no. 18, pp. 3650–3657, 2013.

[53] Y. Zhang, L. Geng, G. Talmon, and J. Wang, "MicroRNA-520g confers drug resistance by regulating p21 expression in colorectal cancer," *Journal of Biological Chemistry*, vol. 290, no. 10, pp. 6215–6225, 2015.

[54] D. J. McKenna, S. S. McDade, D. Patel, and D. J. McCance, "MicroRNA 203 expression in keratinocytes is dependent on regulation of p53 levels by E6," *Journal of Virology*, vol. 84, no. 20, pp. 10644–10652, 2010.

[55] J. A. Li, Y. Chen, J. Zhao, F. Kong, and Y. Zhang, "miR-203 reverses chemoresistance in p53-mutated colon cancer cells through downregulation of Akt2 expression," *Cancer Letters*, vol. 304, no. 1, pp. 52–59, 2011.

[56] N. Léveillé, C. A. Melo, K. Rooijers et al., "Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA," *Nature Communications*, vol. 6, article no. 6520, 2015.

[57] Y. Sánchez, V. Segura, O. Marín-Béjar et al., "Genome-wide analysis of the human p53 transcriptional network unveils a lncRNA tumour suppressor signature," *Nature communications*, vol. 5, article 5812, p. 5812, 2014.

[58] A. Zhang, M. Xu, and Y.-Y. Mo, "Role of the lncRNA-p53 regulatory network in cancer," *Journal of Molecular Cell Biology*, vol. 6, no. 3, pp. 181–191, 2014.

[59] X.-H. Liu, M. Sun, F.-Q. Nie et al., "Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer," *Molecular Cancer*, vol. 13, no. 1, article 92, 2014.

[60] M. R. Epis, K. M. Giles, P. A. Candy, R. J. Webster, and P. J. Leedman, "MiR-331-3p regulates expression of neuropilin-2 in glioblastoma," *Journal of Neuro-Oncology*, vol. 116, no. 1, pp. 67–75, 2014.

[61] M. R. Epis, A. Barker, K. M. Giles, D. J. Beveridge, and P. J. Leedman, "The RNA-binding protein HuR opposes the repression of ERBB-2 gene expression by microRNA miR-331-3p in prostate cancer cells," *The Journal of Biological Chemistry*, vol. 286, no. 48, pp. 41442–41454, 2011.

[62] J.-H. Yoon, K. Abdelmohsen, and M. Gorospe, "Posttranscriptional gene regulation by long noncoding RNA," *Journal of Molecular Biology*, vol. 425, no. 19, pp. 3723–3730, 2013.

[63] T. Terakawa, H. Kenzaki, and S. Takada, "P53 searches on DNA by rotation-uncoupled sliding at C-terminal tails and restricted hopping of core domains," *Journal of the American Chemical Society*, vol. 134, no. 35, pp. 14555–14562, 2012.

[64] S. K. Mungamuri, S. Wang, J. J. Manfredi, W. Gu, and S. A. Aaronson, "Ash2L enables P53-dependent apoptosis by favoring stable transcription pre-initiation complex formation on its pro-apoptotic target promoters," *Oncogene*, vol. 34, no. 19, pp. 2461–2470, 2015.

[65] L. Shi, B. Zhang, X. Sun et al., "MiR-204 inhibits human NSCLC metastasis through suppression of NUAK1," *British Journal of Cancer*, vol. 111, no. 12, pp. 2316–2327, 2014.

[66] H. Xiong, Q. Li, S. Liu et al., "Integrated microRNA and mRNA transcriptome sequencing reveals the potential roles of miRNAs in stage I endometrioid endometrial carcinoma," *PLOS ONE*, vol. 9, no. 10, Article ID e110163, 2014.

Hindawi
BioMed Research International
Volume 2017, Article ID 6132436, 13 pages
https://doi.org/10.1155/2017/6132436

Research Article

Identifying and Analyzing Novel Epilepsy-Related Genes Using Random Walk with Restart Algorithm

Wei Guo,[1] Dong-Mei Shang,[1] Jing-Hui Cao,[2] Kaiyan Feng,[3] Yi-Chun He,[2] Yang Jiang,[4] ShaoPeng Wang,[5] and Yu-Fei Gao[2]

[1]*Department of Outpatient, China-Japan Union Hospital of Jilin University, Changchun 130033, China*
[2]*Department of Neurosurgery, China-Japan Union Hospital of Jilin University, Changchun 130033, China*
[3]*Department of Computer Science, Guangdong AIB Polytechnic, Guangzhou 510507, China*
[4]*Department of Surgery, China-Japan Union Hospital of Jilin University, Changchun 130033, China*
[5]*School of Life Sciences, Shanghai University, Shanghai 200444, China*

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

Received 23 October 2016; Accepted 15 January 2017; Published 1 February 2017

Academic Editor: Ansgar Poetsch

As a pathological condition, epilepsy is caused by abnormal neuronal discharge in brain which will temporarily disrupt the cerebral functions. Epilepsy is a chronic disease which occurs in all ages and would seriously affect patients' personal lives. Thus, it is highly required to develop effective medicines or instruments to treat the disease. Identifying epilepsy-related genes is essential in order to understand and treat the disease because the corresponding proteins encoded by the epilepsy-related genes are candidates of the potential drug targets. In this study, a pioneering computational workflow was proposed to predict novel epilepsy-related genes using the random walk with restart (RWR) algorithm. As reported in the literature RWR algorithm often produces a number of false positive genes, and in this study a permutation test and functional association tests were implemented to filter the genes identified by RWR algorithm, which greatly reduce the number of suspected genes and result in only thirty-three novel epilepsy genes. Finally, these novel genes were analyzed based upon some recently published literatures. Our findings implicate that all novel genes were closely related to epilepsy. It is believed that the proposed workflow can also be applied to identify genes related to other diseases and deepen our understanding of the mechanisms of these diseases.

1. Introduction

As a classical neurological condition that may suddenly interrupt normal life activities and result in physical injury, epilepsy has been widely used to describe a group of epileptic seizure associated diseases [1, 2]. Epileptic seizures are the typical symptoms of the disease, which is the consequence of a disruption of the electrical communications between neurons [3]. As a common neurological disease, epilepsies are found all over the world and affect people of all ages [4]. Only in America, more than one hundred thousand incident cases are diagnosed as epilepsy per year, seriously threatening their mental and physical health [5, 6]. Nowadays, the clinical study on epilepsy has been progressively deepening and typical diagnosis routines of epilepsy have been set up. Considering that epileptic seizures induced by disruption of electrical communications between neurons are the typical symptoms of epilepsy, the long-term paroxysmal epileptic seizures might suggest the initiation and progression of such neurological disease [7, 8]. Generally, it is quite necessary to turn to the doctor for help after someone has more than twice abnormal seizures excluding those with known medical conditions. Considering epilepsy has many subtypes induced by different pathogenic factors, resulting in different complications with similar early seizures symptoms, the diagnosis of epilepsy contributes not only to a confirmation of epilepsy but also to classification of the epilepsy from which the patients suffer into its subtype [9, 10]. The diagnosis of epilepsy can be divided into two main procedures: medical history taking and instrumental inspections. Previous experiments have confirmed that typical family history and multiple medical conditions may lead to neurological abnormalities,

which may contribute to the initiation and progression of epilepsy [11–13]. Therefore, the first step of the diagnosis of epilepsy is to inquire about the medical history of the patients and their respective family. However, a definitive diagnosis of epilepsy is performed by the following instrumental inspections. As we have mentioned above, epilepsy is referred to as a group of neurological diseases induced by abnormal electrical communication between neurons [4]. Therefore, the measurement of electrical impulses in brain by an electroencephalogram (EEG) test has been regarded as one of the golden standards for epilepsy diagnosis [14, 15]. Apart from EGG, magnetic resonance spectroscopy (MRS), positron emission tomography (PET), and magnetic resonance imaging (MRI) have also been widely used to diagnose epilepsy [16–18].

Although great progresses have been made to diagnose epilepsy, the therapeutic methods to treat epilepsy are still quite limited and they mainly contribute to the symptomatic relief. There are two main functional methods to relieve the seizure symptoms of epilepsy: certain nutrient intakes and vagus nerve stimulations by surgery [19–22]. The therapeutic nutrients that have been confirmed to contribute to the relief of epilepsy include folic acid, melatonin, and vitamins (large doses) [19]. However, such a treatment cannot provide a permanent cure but tries to temporarily relieve the symptoms. The vagus nerve stimulation, as the most effective treatment for epilepsy, necessitates implantation of a pacemaker-like device in the patient's body to stimulate the vagus nerve, relieving the symptoms with few side-effects [23]. These treatments do not have an in-depth consideration of the pathogenic factors that cause the disease but mainly concentrate on the relief of the seizure symptoms. To develop more effective curing methods, it is quite fundamental to understand and reveal the pathogenesis of epilepsy.

As it is known, traditionally epilepsy is referred to a group of diseases characterized by similar symptoms (i.e., epileptic seizures), but not by its pathogenesis. And the underlying pathogenesis of epilepsy may be quite complicated. In the past, due to technological constraints, the pathogenesis of epilepsy is largely unknown. It has been reported that the occurrence of some epilepsy cases turned out to exhibit some degree of familial aggregation, not only implicating the significance of history taking, but also suggesting that the genetic background contributes to the disease [24, 25]. According to some clinical data, it is without any doubt that if a sibling suffers from epilepsy, the brothers and sisters who have similar genetic background inherited from their patients are at higher risk of epilepsy comparing to those who do not [26, 27]. However, the detailed pathogenesis cannot be clearly revealed. Recently, with the development of sequencing technologies, some epilepsy associated genes with either pathological mutations or copy number variants have been identified [28, 29]. Among these genes, a group of functional genes, the sodium channel protein family, encode the core sodium channel in the nerve system [30]; for example, genes like SCN1A and SCN8A encoding core component of the sodium channel have been confirmed to be associated with the progression of hereditary epilepsies [31–33]. Therefore, specific genes may play a definitive role during the initiation and progression of the epilepsy, as hereditary epilepsies are deemed to have a certain genetic background.

Although specific genes have been strongly suggested to be associated with epilepsy, however, it is quite hard to identify the core regulatory genes related to epilepsy by time-consuming experimental methods such as the western blot [34, 35]. Here, based on some known epilepsy-related genes, we presented a new computational workflow to search out potential genes of interest. The random walk with restart (RWR) algorithm was employed in our workflow to search possible novel genes in a protein-protein interaction (PPI) network. Compared to the network methods based on guilt-by-association [36] which only consider the neighbors of known genes [37–39], the RWR algorithm can inspect the whole network to make extensive decisions; that is, methods based on guilt-by-association used part of the network, while the RWR algorithm can utilize the whole network. The brief procedures of our workflow was described as follows. Firstly, the RWR algorithm was executed on a PPI network using validated epilepsy associated genes as seed nodes and genes receiving high probabilities were selected as possible candidate genes. Then, these possible genes were screened by a permutation test, followed by functional association tests, resulting in thirty-three novel epilepsy-related genes. Further analysis indicates that all genes obtained may directly or indirectly contribute to the initiation and progression of epilepsy. To the best of our knowledge, this is the first study attempting to identify core regulatory factors of epilepsy using computational methods. These newly found genes may reveal the underlying mechanisms of epilepsy, and the approach may be extended to solve the similar problems of other complex diseases.

2. Materials and Methods

2.1. Epilepsy Related Genes. 499 genes related to epilepsy were retrieved from EpilepsyGene (http://61.152.91.49/EpilepsyGene/download.php) [40], a genetic resource for genes and their mutations related to epilepsy. The epilepsy genes in EpilepsyGene database were collected by searching the PubMed database (https://www.ncbi.nlm.nih.gov/pubmed). Because our method was executed on a PPI network (referred to in Section 2.2), all 499 genes were linked to their Ensembl IDs. Those without Ensembl IDs or those Ensembl IDs do not occur in the PPI network were excluded. Finally, 470 genes with their Ensembl IDs were obtained for investigation in this study. The detailed information of these genes is provided in Supplementary Material S1 in Supplementary Material available online at https://doi.org/10.1155/2017/6132436.

2.2. PPI Network. The interactions between proteins within and outside the cells provide useful information about their activities, properties, and functions. Two proteins that can interact with each other produce a PPI (protein-protein interaction), which often share similar functions or involve in the same biological processes. The PPI network comprised of large amounts of PPIs representing proteins' complicated interaction relationships and remote functional relationships

in signaling pathways, such as the proteins involved in regulation and catalysis activity in glycolysis and tricarboxylic acid cycle [41–43]. Some computational predictors and workflows employed PPIs to predict protein functions [44–46] and search for novel genes related to a variety of diseases [47–51]. Therefore, PPIs could be useful to infer novel epilepsy-related genes based on the validated 470 epilepsy genes mentioned in Section 2.1.

To obtain the PPI information and construct a PPI network, the human PPI information was retrieved from STRING (Version 10.0, http://string-db.org/) [52], a well-known public database collecting known and predicted protein-protein interactions. In the current version, it covers 9,643,763 proteins from 2,031 organisms. Interactions reported in STRING are derived from the following five sources: (I) Genomic Context Predictions; (II) High-throughput Lab Experiments; (III) (Conserved) Co-Expression; (IV) Automated Text mining; (V) Previous Knowledge in Databases. The human PPIs are collected in the file "9606.protein.links.v10.txt.gz" that can be accessed from the download page of STRING using "Homo sapiens" as a restriction to the data. Accordingly, we obtained 4,274,001 human PPIs covering 19,247 proteins. Because the 4,274,001 human PPIs include not only direct (physical) but also indirect (functional) interactions between proteins, these PPIs can offer relatively more information about the novel genes related to epilepsy.

For each PPI, there are two Ensembl IDs representing two proteins and a score ranging from 150 to 999 that indicates the strength of the interaction. A larger score assigned to a PPI indicates that the two proteins are more likely to interact with each other. For proteins p_a and p_b, their interaction score was denoted as $S(p_a, p_b)$. In the network, the 19,247 proteins were denoted as the nodes and two proteins were connected by an edge if and only if they can form a PPI. Thus, there were 4,274,001 edges in the network, and each edge represented a PPI. In addition, the interaction score was added to the network as the weight of the corresponding edge. For convenience, the PPI network was denoted as G in the following sections.

2.3. RWR Algorithm. As a ranking algorithm, the RWR algorithm simulated a walker starting from a seed node or several seed nodes and randomly moved on the network G [53]. In this study, 470 Ensembl IDs of epilepsy genes were set as the seed nodes. Based on them, we aim to mine some potential genes functionally related to epilepsy. In the beginning of the algorithm, an initialization vector P_0 was constructed with 19,247 components in it and each component was a score rating the probability of each node being a potential epilepsy-related gene. The probability scores of 470 Ensembl IDs that represented validated epilepsy genes in P_0 were set to 1/470 (0.0021) and other components were set to zeros. If the vector P_i was the probability vector after the RWR algorithm was executed ith round, then the iteration equation can be formulated as follows:

$$P_{i+1} = (1 - r) A^T P_i + r P_0, \quad (1)$$

where A was the column-wise normalized adjacency matrix and r was the probability that it returned to the start nodes, which was set to 0.8 in this study. When probability vector P_{i+1} and P_i satisfy the inequality $\|P_{i+1} - P_i\|_{L_1} < 1E - 06$, the iteration stopped and P_{i+1} was output as the results of the RWR algorithm.

According to the probability vector yielded by RWR algorithm, each node (gene) in the network was given a number representing the probability of it being a novel epilepsy gene. Genes with larger values are more likely to be epilepsy-related genes. Threshold $1E - 05$ was adopted in this study; that is, genes receiving scores larger than $1E - 05$ were selected from the network G, because it filtered out a large portion of genes and there remained enough genes for further analysis. For convenience, the obtained genes were called RWR genes.

2.4. Filtering Methods. After RWR algorithm was executed on the network, many RWR genes could be selected. However, there are likely many false positive genes among them as elaborated in our previous study [44]. These genes are not special to the epilepsy and should be excluded. In this section, a two-step filtering method was proposed to screen out false positive genes.

2.4.1. Permutation Test. The structure of network G can influence the output of RWR algorithm, which may lead to the false selection of some RWR genes. For example, a node with a degree higher than average degree of the network G may receive a larger score by RWR algorithm even if it was not related to epilepsy. To mine this type of nodes in the network, a permutation test was applied on the network. Firstly, 1,000 Ensembl ID sets, denoted as $S_1, S_2, \ldots, S_{1000}$, were randomly produced and each set contained 470 random Ensembl IDs. Secondly, for each set, the RWR algorithm was executed on the network G using the 470 Ensembl IDs in this set as seed nodes, thereby yielding a probability for each RWR gene. Thirdly, for each RWR gene g, a measurement, namely, permutation FDR, was calculated based on the following equation:

$$\mathrm{FDR}(g) = \frac{\Theta}{1000}, \quad (2)$$

where Θ was the number of randomly produced sets in which the score of gene g was larger than the score computed by the validated epilepsy related genes. According to (2), the higher permutation FDR an RWR gene had, the less possible the gene was an epilepsy related gene. Because 0.05 was widely used as a common cutoff in statistical test, it was also set to be the threshold of permutation FDR in this study. Therefore, the RWR genes with permutation FDRs less than 0.05 were selected and called candidate genes for further analysis.

2.4.2. Functional Association Test. Among the candidate genes, some of them were functionally highly associated with epilepsy while others weakly associated with it. To select essential genes among them, a functional association test that consisted of two selection schemes was proposed.

It is known that proteins in a PPI with a higher interaction score are more likely to share similar functions. Among the candidate genes, those having strong associations with validated epilepsy-related genes could be the most likely novel epilepsy genes. If a candidate gene has strong associations with exact one validated epilepsy-related gene and has weak or no associations with other epilepsy-related genes, it may still be a novel epilepsy-related gene. In view of this, we believe that using the associations between a candidate gene and its most related epilepsy-related gene is more proper to indicate its associations with epilepsy. Accordingly, an interaction measurement called maximum interaction score (MIS) was calculated for each candidate gene g, which can be defined as

$$\begin{aligned}&\mathrm{MIS}(g)\\&\quad=\max\left\{S\left(g,g'\right):\ g'\ \text{is an epilepsy-related gene}\right\}.\end{aligned}\tag{3}$$

Candidate genes with large MISs mean that it is highly possible that they can directly interact with at least one validated epilepsy gene and may cause the symptoms of epilepsy. In STRING, the value 900 is set to be the cutoff to achieve a highest confidence. Thus, it was also set to be the threshold of MIS; that is, candidate genes with MISs less than 900 were discarded.

Besides, genes related to epilepsy may share some common gene ontology (GO) [54] terms and often occurred in the same Kyoto Encyclopedia of Genes and Genomes (KEGG) [55] pathways. Thus, candidate genes sharing same or similar GO terms and KEGG pathways with validated epilepsy genes are more likely to be the genes related to epilepsy. The enrichment theory of GO terms (KEGG pathways) [56–58] was used to quantitatively measure the relationship between a gene and GO terms (KEGG pathways). For a gene g, let us denote the set containing g and its direct neighbor genes in the PPI network reported in STRING by $H(g)$. Then, the relationship between g and one GO term or KEGG pathway can be encoded into a numeric value as follows:

$$S_{\mathrm{GO}}(g,G)=-\log_{10}\left(\sum_{k=m}^{n}\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\right),\tag{4}$$

where G represented a GO term or a KEGG pathway, N was the total number of genes in humans, M was the number of genes annotated to G, n was the number of genes in $H(g)$, and m was the number of genes that are in $H(g)$ and annotated by G. The values for all GO terms and KEGG pathways can be collected into a vector $ES(g)$. The similarity of two genes g and g' on GO terms and KEGG pathways can be measured by the proximity of the two vectors $ES(g)$ and $ES(g')$ as follows:

$$\Gamma\left(g,g'\right)=\frac{ES(g)\cdot ES\left(g'\right)}{\left\|ES(g)\right\|\cdot\left\|ES\left(g'\right)\right\|}.\tag{5}$$

It is clear that if the resultant number of (5) is large, g and g' are similar on GO terms and KEGG pathways, implicating a strong relationship between them. With similar arguments on MIS, for each candidate gene g, the maximum function score (MFS) was calculated as follows:

$$\begin{aligned}&\mathrm{MFS}(g)\\&\quad=\max\left\{\Gamma\left(g,g'\right):\ g'\ \text{is an epilepsy-related gene}\right\}.\end{aligned}\tag{6}$$

A candidate gene receiving a high MFS means it shares relatively more GO terms and KEGG pathways with at least one validated epilepsy gene. In this study, we tried 0.9 as the threshold of MFS; that is, candidate genes with MFSs larger than 0.9 were selected.

In short, candidate genes resulting from permutation test with MISs larger than or equal to 900 and MFSs larger than 0.9 were selected. For convenience, they were named as core candidate genes.

3. Results

An outline for the procedure of the method, including RWR algorithm and filtering methods described in Sections 2.3 and 2.4, by a flowchart is illustrated in Figure 1. This section would show the detailed results yielded by the method.

As described in Section 2.3, the RWR algorithm was executed on the PPI network G, in which the 470 Ensembl IDs were used as seed nodes. A probability vector can be obtained, in which each composition represents the probability score of the corresponding node (gene) being a novel epilepsy-related gene. Genes with probabilities larger than $1E-05$ were selected, producing 6,886 RWR genes, which are listed in Supplementary Material S2.

For the 6,886 RWR genes derived from RWR algorithm, a permutation test was applied on them to screen out RWR genes that are not special for epilepsy. The permutation FDR was calculated for each RWR gene, which is provided in Supplementary Material S2. Value 0.05 was set to be the threshold of permutation FDR, thereby producing 980 candidate genes, which are listed in Supplementary Material S3.

To further select genes that are functionally related to epilepsy, a functional association test was applied to the 980 candidate genes. As described in Section 2.4.2, for each candidate gene, we calculated its MIS (cf. (3)) and MFS (cf. (6)). Values 900 and 0.9 were used as the threshold of MIS and MFS, respectively. And finally thirty-three core candidate genes were obtained. These genes were deemed to be closely related to epilepsy and are listed in Table 1. According to some recent published literature as discussed in Section 4, these core candidate genes, which had similar functions with the validated genes, are highly likely to be the novel epilepsy genes.

4. Discussions

For a long time, epilepsy has been regarded as complicated neurological diseases with various pathogenesis. Based on clinical data, the occurrence of epilepsy has shown conspicuous familial aggregation characteristics, implying that genetic background features (such as mutations, copy

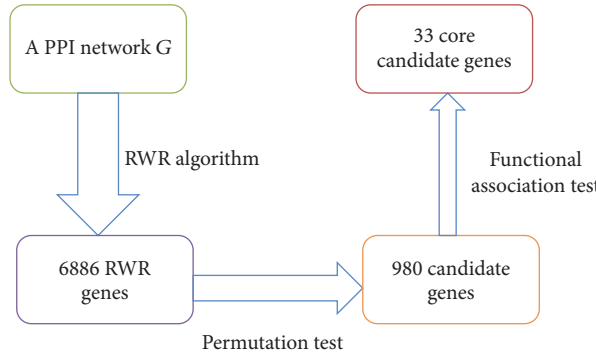


Figure 1: The flowchart of RWR algorithm and filtering methods for identifying core candidate genes.

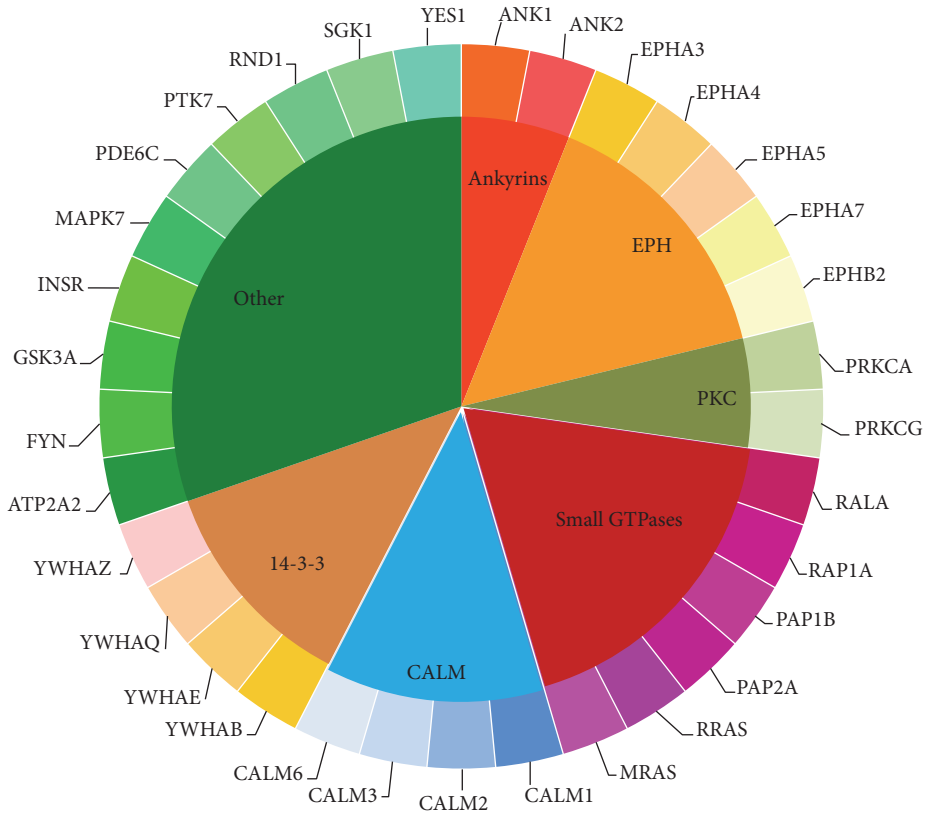


Figure 2: The distribution of the thirty-three core candidate genes according to their protein families.

number variants of genes) may play an irreplaceable role for epilepsy [59–61]. Recent publications have also confirmed such implication. Various epilepsy associated genes have been identified [62–66]. However, it is quite expensive and time-consuming to identify epilepsy associated genes with experiments. Based on our computational method, we identified thirty-three candidate epilepsy associated genes, listed in Table 1. According to some recent publications, all these core candidates show specific relationship with the initiation and progression of epilepsy, validating the effectiveness of our computational method. According to the gene families of these thirty-three core candidate genes, we classified them into seven clusters, shown in Figure 2, and analyzed them accordingly.

Ankyrins. In our prediction list, two of the candidate genes, *ANK1* and *ANK2,* turned out to be the functional members of the ankyrins. As we all know, ankyrins are a group of connexin that link the integral membrane proteins to the cytoskeleton, which have been widely reported that they contribute to cell proliferation, motility, and the maintenance of specialized membrane domains [67–69]. In human bodies, ankyrins have been confirmed to bind to the voltage-gated potassium channel subunits KCNQ2 and KCNQ3, regulating

Table 1: Thirty-three core candidate genes identified by our method.

Gene symbol	Ensembl ID	Probability	Permutation FDR	MIS	MFS
ANK2	ENSP00000349588	$4.53E-05$	<0.001	990	0.997
ANK1	ENSP00000265709	$3.52E-05$	0.034	995	0.995
EPHA7	ENSP00000358309	$3.28E-05$	0.025	906	0.988
EPHA5	ENSP00000273854	$3.46E-05$	0.032	906	0.988
PRIKCG	ENSP00000263431	$4.75E-05$	0.025	905	0.987
PTK7	ENSP00000230419	$3.10E-05$	0.023	943	0.987
EPHA3	ENSP00000337451	$4.40E-05$	0.017	912	0.986
PDE6C	ENSP00000360502	$3.61E-05$	0.032	900	0.980
EPHA4	ENSP00000281821	$4.16E-05$	0.008	990	0.980
YWHAQ	ENSP00000238081	$5.50E-05$	0.004	999	0.978
GSK3A	ENSP00000222330	$6.40E-05$	0.017	977	0.976
CALM1	ENSP00000349467	$6.55E-05$	0.023	966	0.976
EPHB2	ENSP00000363763	$4.39E-05$	0.026	908	0.975
PRIKCA	ENSP00000408695	$7.02E-05$	0.039	992	0.975
CALM2	ENSP00000272298	$5.52E-05$	0.048	985	0.974
YWHAE	ENSP00000264335	$1.63E-04$	0.009	999	0.973
YWHAB	ENSP00000300161	$5.56E-05$	0.047	999	0.971
ATP2A2	ENSP00000440045	$5.03E-05$	0.005	908	0.970
YES1	ENSP00000324740	$6.54E-05$	0.002	967	0.969
CALML3	ENSP00000315299	$3.96E-05$	0.027	906	0.968
SGK1	ENSP00000356832	$4.90E-05$	0.034	999	0.966
CALML6	ENSP00000304643	$4.44E-05$	0.01	909	0.965
YWHAZ	ENSP00000309503	$1.86E-04$	0.002	999	0.958
MAPK7	ENSP00000311005	$7.15E-05$	0.031	999	0.956
RRAS	ENSP00000246792	$5.94E-05$	0.004	951	0.941
RAP2A	ENSP00000245304	$4.97E-05$	0.021	940	0.937
RALA	ENSP00000005257	$6.57E-05$	0.006	981	0.932
RAP1B	ENSP00000250559	$5.28E-05$	0.009	972	0.931
MRAS	ENSP00000289104	$4.70E-05$	0.022	932	0.931
INSR	ENSP00000303830	$7.27E-05$	0.024	996	0.927
PAP1A	ENSP00000348786	$6.49E-05$	0.014	995	0.925
RND1	ENSP00000308461	$5.70E-05$	0.002	996	0.903
FYN	ENSP00000346671	$1.02E-04$	<0.001	999	0.900

their normal functions [70]. Considering that KCNQ2 and KCNQ3 turned out to directly contributing to the initiation and progression of epilepsy, our predicted genes ANK1 and ANK2 as the functional components of ankyrins may very probably be epilepsy associated genes, validating our prediction [71].

EPH Subfamily. Apart from the Ankyrin protein family, another group of proteins, the EPH subfamily, have also been identified to contribute to epilepsy. In our prediction list, five genes can be classified in such subfamily: *EPHA3*, *EPHA4*, *EPHA5*, *EPHA7,* and *EPHB2*. Such five genes all encode the receptors for the erythropoietin-producing hepatoma amplified sequences (EPH), acting as the tyrosine-protein kinase receptor [72, 73]. In mouse model, it has been confirmed that the activation of EPH receptor associated genes, like EPHB3, contributes to the onset of epilepsy [74]. Considering the functional similarity and underlying correlations, it is quite reasonable that our predicted genes of EPH receptor family may also contribute to such processes [75]. Apart from that, another publication confirmed that, by stimulating NMDA receptor activity, ERK activates the progression of epilepsy [76]. During the activation, various genes of our predicted EPH family have been identified to promote such biological processes, validating the crucial role of EPH family including our predicted genes EPHA3, EPHA4, EPHA5, EPHA7, and EPHB2 during epilepsy.

Protein Kinases. Two of our predicted candidates, *PRKCA* and *PRKCG,* can be clustered into another functional family, the family of serine- and threonine-specific protein kinases. Such two genes turn out to be functional components of the protein kinase C, a core member of the protein family we mentioned above [77–79]. The protein kinase C associated signaling pathway has been widely reported to be associated with epilepsy and may be a candidate therapeutic target for such

disease [80]. As two major components for such pathway, our predicted genes PRKCA and PRKCG may definitely contribute to such disease. Apart from such evidence, a specific mutation of PRKCG (SCA-14) has been reported to be associated with a typical movement disorder, which can be called Ramsay Hunt phenotype [81]. Considering that such disorder has been widely identified in epilepsy patients, such mutation may be functionally related to the progression of epilepsy, validating our prediction [82, 83]. As for MAPK7, such gene has been widely regarded as a multifunctional gene that involves various biological processes including proliferation, differentiation, transcription regulation, and development [84]. MAPK7 has been reported to interact with a specific protein Aquaporin 4 (AQP4) in human beings [85]. Since AQP4 has been confirmed to accumulate in neuron cell during epilepsy and contribute to the pathological processes, it is quite reasonable to summarize that, as a functional related protein of AQP4, our predicted gene MAPK7 very probably contributes to epilepsy, validating the accuracy and efficacy of our prediction [85]. The next gene is also a crucial kinase for human beings, the *PTK7*. Although, different from other proteins from protein tyrosine kinase family, PTK7 lacks detectable catalytic tyrosine kinase activity, it has still been reported to contribute to the functional Wnt signaling pathway and regulate the cellular polarity and adhesion [86]. Though no direct relationship between PTK7 and epilepsy has been confirmed, recent publications reported that PTK7 may participate in the metabolism of antiepileptic drugs (AED), suggesting that there may remain uncovered interactions between PTK7 and epilepsy, validating our prediction [87]. *GSK3A*, as a multifunctional Ser/Thr protein kinase, has been reported to contribute to glycogen synthesis and transcriptional regulation [88, 89]. Such gene has been reported to be quite essential for the development and maturation of cortical neurons [90]. Considering that cortical neurons, especially the migration of neurons, are quite significant for epilepsy, our predicted gene GSK3A may very probably be epilepsy associated gene [91, 92].

Small GTPases. Apart from PKC associated genes, there are also six genes (*RALA*, *RAP1A*, *RAP1B*, *RAP2A*, *RRAS,* and *MRAS*) that can be clustered into the famous Ras family of small GTPases. Based on recent publications, various small GTPases have been identified to contribute to the progression of epilepsy, including Cdc42, RAB39B [93, 94]. As for our predicted candidates, it has been confirmed that, during the pathological processes of epilepsy, the normal function of RAP1A and its related Ras signaling pathway has been regulated and altered by microRNAs, implying the potential role of Ras signaling pathway during epilepsy [95]. As for RAP1B and RAP2A, RAP1B has been confirmed to be specifically activated in nerve system and contributes to the regeneration of neuronal connectivity, the dysfunction of which turns out to be one of the pathological factors for epilepsy, validating the regulatory role of RAP1B during such disease [96]. RAP2A has been validated and confirmed to contribute to the childhood absence epilepsy, a specific subtype of epilepsy, and may be related with a specific glioma inducing epilepsy associated symptoms, validating our prediction of epilepsy associated genes [97, 98]. As for RRAS and MRAS, considering the functional similarity of MRAS and MAPK, the detailed analysis of such genes can be seen below, while the inner relationship between RRAS and epilepsy has also been revealed in mouse model, validating our prediction [99, 100]. RALA, encoding a functional small GTPase belonging to Ras family, has been confirmed to mediate the transmembrane signaling by the occupancy of functional receptors [101]. Such gene has been definitely confirmed to be associated with epilepsy by regulating the drug resistance of such disease [102]. Another gene, *RND1,* encodes a small GTPase, which does not belong to Ras family but to Rho GTPase family. In response to various extracellular signaling, the protein encoded by such gene turns out to regulate the actin cytoskeleton [103, 104]. In intractable epilepsy, a clinical subtype of epilepsy which is hard to cure, recent publications confirmed the expression of such gene in the central nerve system of the patients, implying that such gene may definitely contribute to the progression and prognosis of such disease [105, 106]. MAPK7 and MRAS are two proliferation-associated genes in our candidate epilepsy associated gene list. Among them, *MRAS* turn out to contribute to Ras signaling pathway, which has been confirmed above to be associated with the progression of epilepsy [95]. What is more, as for MRAS itself, it has been reported that such gene may contribute to the development of brain in early stage and the abnormal activation of such gene may induce epilepsy-like syndrome, validating our prediction [107].

Calmodulin (CALM) Family. Apart from that, such genes may also contribute to the specific seizure like features during the progression of epilepsy, suggesting its core regulatory role [108]. Four genes (*CALM1*, *CALM2*, *CALM3,* and *CALM6*) of the functional calmodulin (CALM) family have also been predicted to contribute to epilepsy. Genes of calmodulin family mainly act as a calcium binding protein that participate in cell cycle and proliferation associated biological processes [109, 110]. Considering that epilepsy has been confirmed to be associated with abnormal calcium ion transportation, it is quite reasonable to speculate that our predicted genes of CALM family may contribute to epilepsy which has also been confirmed by recent publications [111, 112].

14-3-3 Family of Proteins. The remaining group of functional proteins, including *YWHAB*, *YWHAE*, *YWHAQ*, and *YWHAZ*, that contribute to epilepsy turn out to be encoded by the so-called 14-3-3 family of proteins. Such family of proteins contribute to the signaling transduction by binding to phosphoserine-containing proteins [113]. As we have mentioned above, epilepsy has been confirmed to be associated with the protein kinase C signaling pathway [81]. Recent publications identified that, during the progression of epilepsy, our predicted candidates, proteins of 14-3-3 family, may interact with protein kinase C and further promote the progression, implying the functional role of such genes [114]. Apart from that, such five genes that we sorted out have also been directly identified in epilepsy cases. Take YWHAB as an example. Such gene has been identified to contribute to

the regeneration of neurons after physical or chemical injury. The dysfunction of such gene may be related to the initiation of epilepsy in certain pathological conditions [115]. Therefore, such four genes in our prediction list have all been confirmed to participate in epilepsy associated biological processes, validating the accuracy and efficacy of our prediction.

Other Crucial Genes. Apart from such genes, there still remain six genes with no clear family enrichment in our prediction list that may also contribute to epilepsy in their respective ways. Among such genes, *ATP2A2* turns out to encode a significant intracellular pump located in the sarcoplasmic or endoplasmic reticula [116, 117]. Epilepsy has been confirmed to be a specific complication of various diseases, including Darier's disease [118, 119]. Our predicted gene ATP2A2 has been identified in clinical cases of Darier's disease and may directly contribute to epilepsy associated symptoms, validating the efficacy and accuracy of our prediction [120]. Another gene *INSR* turns out to be the functional receptor for a core endogenous hormone insulin, which further activates the downstream of insulin signaling pathway [121, 122]. As for the underlying relationship between INSR and epilepsy, it has been reported that a group of specific mutations in INSR: INSR H1085H C>T, G972R has been confirmed to specifically occur in the epilepsy patients in Han Chinese, validating the specific role of INSR during the progression of epilepsy [123]. Another gene *FYN* is also a candidate oncogene just like genes from Ras family as we have mentioned above [124, 125]. Recent publications reveal the potential relationship between our predicted gene FYN and amygdala kindling, a specific phenotype associated with epileptogenesis [126]. Participating in mTOR signaling pathway, though the detailed mechanism of the pathology of FYN initiated epilepsy has not been fully revealed, our predicted gene FYN may definitely be an epilepsy associated gene [126–128].

Phosphodiesterase 6C, encoded by our predicted gene *PDE6C,* has been confirmed to contribute to pyrimidine metabolism and phototransduction [129, 130]. Autosomal-dominant cerebellar ataxia is a specific symptom of epilepsy in human beings [131]. It has been reported that a specific mutation, p.Arg95His, of our predicted gene, PDE6C, may be associated with the autosomal-dominant cerebellar [132]. Considering the inner linkage between autosomal-dominant cerebellar ataxia and epilepsy in human beings, our predicted gene PDE6C may also contribute to the progression of epilepsy, validating our prediction [131]. As we have mentioned above, quite a lot of kinases have been reported to contribute to the abnormal biological process during epilepsy. The remaining two genes *SGK1* and *YES1* have also been suggested to contribute to the progression of epilepsy. SGK1 turns out to encode a serum/glucocorticoid regulated kinase, which further contributes to the regulation of cellular stress response [133, 134]. It has been confirmed that our predicted gene SGK1 is upregulated by a functional cellular component, aldosterone [134]. It has been confirmed that aldosterone has strong chemical and biological effects on epileptic seizures, implying that our predicted gene SGK1 may contribute to the regulation of the typical symptoms of epilepsy, the seizures [135]. Therefore, SGK1 very probably is a functional epilepsy associated gene. The last gene, YES1, also encodes a small GTPase that has been widely regarded as a tumor associated gene [136]. Although no direct report confirms the relationship between YES1 and epilepsy, considering the core regulatory role of small GTPases for epilepsy and that it has been confirmed that YES1 is expressed in brain and central nerve system, it is quite reasonable for us to believe that YES may be a functional epilepsy associated gene [137].

5. Conclusion

Based on our newly developed computational method, we have identified thirty-three novel genes that may contribute to the initiation and progression of epilepsy. According to the comprehensive analyses on these genes, they are strongly suspected to be either directly or indirectly related to epilepsy, validating the effectiveness of our method. In summary, this method can not only contribute to the identification of potential epilepsy associated genes but also provide a new tool to investigate the underlying mechanisms of the pathological processes of epilepsy.

Competing Interests

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

Authors' Contributions

Wei Guo and Dong-Mei Shang contributed equally to this work.

Acknowledgments

This paper is supported by Norman Bethune Program of Jilin University (2015218), Science and Technology Department of Jilin Province (20160414007GH, 20160414047GH), Education Department of Jilin Province (2015509, 2016449), and Development and Reform Commission of Jilin Province (2015Y032).

References

[1] S. Wiebe, "Brain surgery for epilepsy," *The Lancet*, vol. 362, pp. s48–s49, 2003.

[2] S. Baxendale, "Epilepsy at the movies: possession to presidential assassination," *Lancet Neurology*, vol. 2, no. 12, pp. 764–770, 2003.

[3] F. M. Liu, S. Dai, A. Napoli et al., "Epileptic seizures are induced by intracerebral ablation of astrocytes in the brain, a novel model for dissecting the interaction of neurons with glial cells," *Journal of NeuroVirology*, vol. 21, pp. S41–S42, 2015.

[4] N. K. Sethi, "Psychogenic non-epileptic seizures—the age matters," *Clinical Neurology and Neurosurgery*, vol. 120, p. 142, 2014.

[5] R. Zepeda, K. A. Gleason, E. J. Bubrick, D. J. Pallin, and B. A. Dworetzky, "Disparities of epilepsy care in the emergency department," *Epilepsia*, vol. 50, pp. 307–307, 2009.

[6] J. G. Burneo, N. Jette, W. Theodore et al., "Disparities in epilepsy: report of a systematic review by the North American Commission of the international league against epilepsy," *Epilepsia*, vol. 50, no. 10, pp. 2285–2295, 2009.

[7] K. Sugai, E. Nakagawa, H. Komaki, H. Sakuma, Y. Saito, and M. Sasaki, "Pharmacotherapy for childhood nonidio-pathic partial epilepsies based on seizure symptoms: retrospective and prospective studies," *Epilepsia*, vol. 50, pp. 113–113, 2009.

[8] H. Choi, M. R. Winawer, S. Kalachikov, T. A. Pedley, W. A. Hauser, and R. Ottman, "Classification of partial seizure symptoms in genetic studies of the epilepsies," *Neurology*, vol. 66, no. 11, pp. 1648–1653, 2006.

[9] B. Aktekin, "Up-to-date critical review of the classification of epilepsies and epileptic seizures," *Noropsikiyatri Arsivi*, vol. 52, no. 2, pp. 109–110, 2015.

[10] S.-H. Lee, J. S. Lim, J.-K. Kim, J. Yang, and Y. Lee, "Classification of normal and epileptic seizure EEG signals using wavelet transform, phase-space reconstruction, and Euclidean distance," *Computer Methods and Programs in Biomedicine*, vol. 116, no. 1, pp. 10–25, 2014.

[11] M. Jose and S. V. Thomas, "Family history of congenital malformations does not increase the risk of fetal malformations in women with epilepsy," *Epilepsia*, vol. 56, pp. 31–31, 2015.

[12] C. Alonso-Cerezo, I. Herrera-Peco, V. Fernández-Millares et al., "Family history of epilepsy resistant to treatment," *Revista de Neurologia*, vol. 52, no. 9, pp. 522–526, 2011.

[13] U. C. Wieshmann, "Family history of epilepsy in epilepsy and other neurological conditions," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 75, pp. 799–800, 2004.

[14] Y. Bahbiti, F. Moutaouakil, A. Ouichou, A. El Hessni, B. Benazzouz, and A. Mesfioui, "Epilepsy: electroencephalogram and brain maturation," *Epilepsia*, vol. 54, pp. 156–156, 2013.

[15] J. Liao, L. Song, and Y. Chen, "Seizures captured with video-electroencephalogram in infants with epilepsy," *Epilepsia*, vol. 53, p. 132, 2012.

[16] G. D. Wang, Z. Y. Dai, W. G. Song et al., "Grey matter anomalies in drug-naïve childhood absence epilepsy: a voxel-based morphometry study with MRI at 3.0 T," *Epilepsy Research*, vol. 124, pp. 63–66, 2016.

[17] L. Lipatova and T. Kapustina, "Functional neuroimaging using the method 1H MRS in epilepsy," *European Journal of Neurology*, vol. 22, supplement 1, p. 632, 2015.

[18] J. Peter, S. Houshmand, T. J. Werner, D. Rubello, and A. Alavi, "Novel assessment of global metabolism by 18F-FDG-PET for localizing affected lobe in temporal lobe epilepsy," *Nuclear Medicine Communications*, vol. 37, no. 8, pp. 882–887, 2016.

[19] M. Kinney and J. Morrow, "Vitamin K is important for epilepsy in pregnancy," *British Medical Journal*, vol. 354, article i3929, 2016.

[20] F. M. Snoeijen-Schouwenaars, K. C. Van Deursen, I. Y. Tan, P. Verschuure, and M. H. Majoie, "Vitamin D supplementation in children with epilepsy and intellectual disability," *Pediatric Neurology*, vol. 52, no. 2, pp. 160–164, 2015.

[21] C. Bodin, S. Aubert, G. Daquin et al., "Responders to vagus nerve stimulation (VNS) in refractory epilepsy have reduced interictal cortical synchronicity on scalp EEG," *Epilepsy Research*, vol. 113, pp. 98–103, 2015.

[22] S. Mannino, G. Colicchio, R. Di Bonaventura et al., "Patients/caregivers satisfaction following vagal nerve stimulation (Vns) for drug-resistant epilepsies," *Epilepsia*, vol. 55, no. 1, pp. 103–104, 2014.

[23] A. Cukiert, J. Burattini, and C. Cukiert, "Vagus nerve stimulation (Vns) in refractory epilepsy," *Epilepsia*, vol. 54, p. 84, 2013.

[24] P. Fabera, H. Krijtova, M. Tomasek et al., "Familial temporal lobe epilepsy due to focal cortical dysplasia type IIIa," *Seizure*, vol. 31, pp. 120–123, 2015.

[25] A. Chentouf, A. Dahdouh, M. Guipponi et al., "Familial epilepsy in Algeria: clinical features and inheritance profiles," *Seizure*, vol. 31, pp. 12–18, 2015.

[26] A. Hames and R. Appleton, "Living with a brother or sister with epilepsy: siblings' experiences," *Seizure*, vol. 18, no. 10, pp. 699–701, 2009.

[27] E. Kurča, M. Grofik, P. Kučera, and P. Varsik, "Familial occurrence of adrenocortical insufficiency in two brothers with allgrove syndrome. A case report of 4A (Allgrove) syndrome with epilepsy and a new AAAs gene mutation," *Neuroendocrinology Letters*, vol. 26, no. 5, pp. 499–502, 2005.

[28] R. H. Purcell, L. A. Papale, C. D. Makinson et al., "Effects of an epilepsy-causing mutation in the SCN1A sodium channel gene on cocaine-induced seizure susceptibility in mice," *Psychopharmacology*, vol. 228, no. 2, pp. 263–270, 2013.

[29] A. Escayg and A. L. Goldin, "Sodium channel SCN1A and epilepsy: mutations and mechanisms," *Epilepsia*, vol. 51, no. 9, pp. 1650–1658, 2010.

[30] T. H. Rhodes, C. G. Vanoye, and A. L. George, "Functional characterization of SCN1A sodium channel mutations associated with familial epilepsy," *Biophysical Journal*, vol. 88, p. 378a, 2005.

[31] L. Baum, B. S. Haerian, H.-K. Ng et al., "Case-control association study of polymorphisms in the voltage-gated sodium channel genes SCN1A, SCN2A, SCN3A, SCN1B, and SCN2B and epilepsy," *Human Genetics*, vol. 133, no. 5, pp. 651–659, 2014.

[32] A. J. Barela, S. P. Waddy, J. G. Lickfett et al., "An epilepsy mutation in the sodium channel SCN1A that decreases channel excitability," *Journal of Neuroscience*, vol. 26, no. 10, pp. 2714–2723, 2006.

[33] C. Lossin, T. H. Rhodes, R. R. Desai et al., "Epilepsy-associated dysfunction in the voltage-gated neuronal sodium channel SCN1A," *Journal of Neuroscience*, vol. 23, no. 36, pp. 11289–11295, 2003.

[34] A. B. Holt and T. I. Netoff, "Computational modeling of epilepsy for an experimental neurologist," *Experimental Neurology*, vol. 244, pp. 75–86, 2013.

[35] R. A. Stefanescu, R. G. Shivakeshavan, and S. S. Talathi, "Computational models of epilepsy," *Seizure*, vol. 21, no. 10, pp. 748–759, 2012.

[36] S. Oliver, "Guilt-by-association goes global," *Nature*, vol. 403, no. 6770, pp. 601–603, 2000.

[37] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner, "Predicting disease genes using protein-protein interactions," *Journal of Medical Genetics*, vol. 43, no. 8, pp. 691–698, 2006.

[38] M. Krauthammer, C. A. Kaufmann, T. C. Gilliam, and A. Rzhetsky, "Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 42, pp. 15148–15153, 2004.

[39] L. Franke, H. Van Bakel, L. Fokkens, E. D. De Jong, M. Egmont-Petersen, and C. Wijmenga, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes," *American Journal of Human Genetics*, vol. 78, no. 6, pp. 1011–1025, 2006.

[40] X. Ran, J. Li, Q. Shao et al., "EpilepsyGene: a genetic resource for genes and mutations related to epilepsy," *Nucleic Acids Research*, vol. 43, no. 1, pp. D893–D899, 2015.

[41] C. Depré, M. H. Rider, and L. Hue, "Mechanisms of control of heart glycolysis," *European Journal of Biochemistry*, vol. 258, no. 2, pp. 277–290, 1998.

[42] L. Hue and M. H. Rider, "Role of fructose 2,6-bisphosphate in the control of glycolysis in mammalian tissues," *The Biochemical Journal*, vol. 245, no. 2, pp. 313–324, 1987.

[43] R. G. Hansford and D. Zorov, "Role of mitochondrial calcium transport in the control of substrate oxidation," *Molecular and Cellular Biochemistry*, vol. 184, no. 1-2, pp. 359–369, 1998.

[44] L. Chen, Y.-H. Zhang, T. Huang, and Y.-D. Cai, "Identifying novel protein phenotype annotations by hybridizing protein-protein interactions and protein sequence similarities," *Molecular genetics and genomics: MGG*, vol. 291, no. 2, pp. 913–934, 2016.

[45] 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.

[46] 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.

[47] T. Gui, X. Dong, R. Li, Y. Li, and Z. Wang, "Identification of hepatocellular carcinoma-related genes with a machine learning and network analysis," *Journal of Computational Biology*, vol. 22, no. 1, pp. 63–71, 2015.

[48] J. Zhang, J. Yang, T. Huang, Y. Shu, and L. Chen, "Identification of novel proliferative diabetic retinopathy related genes on protein–protein interaction network," *Neurocomputing*, vol. 217, pp. 63–72, 2016.

[49] L. Chen, T. Huang, Y.-H. Zhang, Y. Jiang, M. Zheng, and Y.-D. Cai, "Identification of novel candidate drivers connecting different dysfunctional levels for lung adenocarcinoma using protein-protein interactions and a shortest path approach," *Scientific Reports*, vol. 6, Article ID 29849, 2016.

[50] L. Chen, J. Yang, T. Huang, X. Kong, L. Lu, and Y.-D. Cai, "Mining for novel tumor suppressor genes using a shortest path approach," *Journal of Biomolecular Structure and Dynamics*, vol. 34, no. 3, pp. 664–675, 2016.

[51] L. Chen, Z. Xing, T. Huang, Y. Shu, G. Huang, and H.-P. Li, "Application of the shortest path algorithm for the discovery of breast cancer-related genes," *Current Bioinformatics*, vol. 11, no. 1, pp. 51–58, 2016.

[52] D. Szklarczyk, A. Franceschini, S. Wyder et al., "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, no. 1, pp. D447–D452, 2015.

[53] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *The American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

[54] "Gene ontology consortium: going forward," *Nucleic Acids Research*, vol. 43, no. D1, pp. D1049–D1056, 2015.

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

[56] J. Yang, L. Chen, X. Kong, T. Huang, and Y.-D. Cai, "Analysis of tumor suppressor genes based on gene ontology and the KEGG pathway," *PLoS ONE*, vol. 9, no. 9, Article ID e107202, 2014.

[57] J. Zhang, Z. Xing, M. Ma et al., "Gene ontology and KEGG enrichment analyses of genes related to age-related macular degeneration," *BioMed Research International*, vol. 2014, Article ID 450386, 10 pages, 2014.

[58] L. Chen, Y.-H. Zhang, M. Zheng, T. Huang, and Y.-D. Cai, "Identification of compound–protein interactions through the analysis of gene ontology, KEGG enrichment for proteins and molecular fragments of compounds," *Molecular Genetics and Genomics*, vol. 291, no. 6, pp. 2065–2079, 2016.

[59] R. Fjaer, E. Brodtkorb, A.-M. Øye et al., "Generalized epilepsy in a family with basal ganglia calcifications and mutations in SLC20A2 and CHRNB2," *European Journal of Medical Genetics*, vol. 58, no. 11, pp. 624–628, 2015.

[60] S. Partemi, S. Cestèle, M. Pezzella et al., "Loss-of-function *KCNH2* mutation in a family with long QT syndrome, epilepsy, and sudden death," *Epilepsia*, vol. 54, no. 8, pp. e112–e116, 2013.

[61] B. Berghuis, E. H. Brilstra, D. Lindhout, S. Baulac, G. J. de Haan, and M. van Kempen, "Hyperactive behavior in a family with autosomal dominant lateral temporal lobe epilepsy caused by a mutation in the LGI1/epitempin gene," *Epilepsy and Behavior*, vol. 28, no. 1, pp. 41–46, 2013.

[62] W. Bi, I. A. Glass, D. M. Muzny et al., "Whole exome sequencing identifies the first STRADA point mutation in a patient with polyhydramnios, megalencephaly, and symptomatic epilepsy syndrome (PMSE)," *American Journal of Medical Genetics A*, vol. 170, no. 8, pp. 2181–2185, 2016.

[63] M. Gal, D. Magen, Y. Zahran et al., "A novel homozygous splice site mutation in NALCN identified in siblings with cachexia, strabismus, severe intellectual disability, epilepsy and abnormal respiratory rhythm," *European Journal of Medical Genetics*, vol. 59, no. 4, pp. 204–209, 2016.

[64] G. Li, R. Shi, J. Wu et al., "Association of the hERG mutation with long-QT syndrome type 2, syncope and epilepsy," *Molecular Medicine Reports*, vol. 13, no. 3, pp. 2467–2475, 2016.

[65] M. G. Sweeney, S. R. Hammans, L. W. Duchen et al., "Mitochondrial DNA mutation underlying Leigh's syndrome: clinical, pathological, biochemical, and genetic studies of a patient presenting with progressive myoclonic epilepsy," *Journal of the Neurological Sciences*, vol. 121, no. 1, pp. 57–65, 1994.

[66] S. R. Hammans, M. G. Sweeney, M. Brockington et al., "The mitochondrial-DNA transfer Rna(Lys) a-]G(8344) mutation and the syndrome of myoclonic epilepsy with ragged-red fibers (Merrf)—relationship of clinical phenotype to proportion of mutant mitochondrial-DNA," *Brain*, vol. 116, pp. 617–632, 1993.

[67] K. J. Chang, T. S. Ho, K. Susuki et al., "Paranodal ankyrins: enigmatic glial anchors?" *Journal of Neurochemistry*, vol. 125, p. 198, 2013.

[68] A. Armani, E. Giacomello, S. Galli, D. Rossi, and V. Sorrentino, "Muscle-specific ankyrins and the organization of the sarcoplasmic reticulum in striated muscle cells," *Biophysical Journal 86(1): 222a*, vol. 86, no. 1, p. 222a, 2004.

[69] P. J. Mohler, A. O. Gramolini, and V. Bennett, "Ankyrins," *Journal of Cell Science*, vol. 115, no. 8, pp. 1565–1566, 2002.

[70] Z. Pan, T. Kao, Z. Horvath et al., "A common ankyrin-G-based mechanism retains KCNQ and Na V channels at electrically active domains of the axon," *Journal of Neuroscience*, vol. 26, no. 10, pp. 2599–2613, 2006.

[71] S. R. Cunha and P. J. Mohler, "Ankyrin protein networks in membrane formation and stabilization," *Journal of Cellular and Molecular Medicine*, vol. 13, no. 11-12, pp. 4364–4376, 2009.

[72] J. Chen, W. Song, and K. Amato, "Eph receptor tyrosine kinases in cancer stem cells," *Cytokine & Growth Factor Reviews*, vol. 26, no. 1, pp. 1–6, 2015.

[73] O. Eriksson, M. Ramström, K. Hörnaeus, J. Bergquist, D. Mokhtari, and A. Siegbahn, "The Eph tyrosine kinase receptors EphB2 and EphA2 are novel proteolytic substrates of tissue factor/coagulation factor VIIa," *Journal of Biological Chemistry*, vol. 289, no. 47, pp. 32379–32391, 2014.

[74] H. Huang, R. H. Li, J. X. Yuan et al., "Up-regulated ephrinB3/EphB3 expression in intractable temporal lobe epilepsy patients and pilocarpine induced experimental epilepsy rat model," *Brain Research*, vol. 1639, pp. 1–12, 2016.

[75] B. Hock, B. Böhme, T. Karn et al., "PDZ-domain-mediated interaction of the Eph-related receptor tyrosine kinase EphB3 and the ras-binding protein AF6 depends on the kinase activity of the receptor," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 17, pp. 9779–9784, 1998.

[76] A. S. Nateri, G. Raivich, C. Gebhardt et al., "ERK activation causes epilepsy by stimulating NMDA receptor activity," *EMBO Journal*, vol. 26, no. 23, pp. 4891–4901, 2007.

[77] G. Sánchez-Fernández, S. Cabezudo, Á. Caballero et al., "Protein kinase C ζ interacts with a novel binding region of Gαq to act as a functional effector," *Journal of Biological Chemistry*, vol. 291, no. 18, pp. 9513–9525, 2016.

[78] A. M. F. Liu and Y. H. Wong, "G16-mediated activation of nuclear factor κB by the adenosine A1 receptor involves c-Src, protein kinase C, and ERK signaling," *Journal of Biological Chemistry*, vol. 279, no. 51, pp. 53196–53204, 2004.

[79] C. J. Doering, A. E. Kisilevsky, Z.-P. Feng et al., "A single Gβ subunit locus controls cross-talk between protein kinase C and G protein regulation of N-type calcium channels," *The Journal of Biological Chemistry*, vol. 279, no. 28, pp. 29709–29717, 2004.

[80] Z. Gajda, R. Török, Z. Horváth et al., "Protein kinase inhibitor as a potential candidate for epilepsy treatment," *Epilepsia*, vol. 52, no. 3, pp. 579–588, 2011.

[81] J. E. Visser, B. R. Bloem, and B. P. C. Van De Warrenburg, "PRKCG mutation (SCA-14) causing a Ramsay Hunt phenotype," *Movement Disorders*, vol. 22, no. 7, pp. 1024–1026, 2007.

[82] M.-C. Hsiao, C.-Y. Liu, Y.-Y. Yang, C.-S. Lu, and E.-K. Yeh, "Progressive myoclonic epilepsies syndrome (Ramsay Hunt syndrome) with mental disorder: report of two cases," *Psychiatry and Clinical Neurosciences*, vol. 53, no. 5, pp. 575–578, 1999.

[83] T. D. Bird and C. M. Shaw, "Progressive myoclonus and epilepsy with dentatorubral degeneration: a clinicopathological study of the Ramsay Hunt syndrome," *Journal of Neurology Neurosurgery and Psychiatry*, vol. 41, no. 2, pp. 140–149, 1978.

[84] S. Javaid, J. Zhang, G. A. Smolen et al., "MAPK7 regulates emt features and modulates the generation of CTCs," *Molecular Cancer Research*, vol. 13, no. 5, pp. 934–943, 2015.

[85] T. S. Lee, T. Eid, S. Mane et al., "Aquaporin-4 is increased in the sclerotic hippocampus in human temporal lobe epilepsy," *Acta Neuropathologica*, vol. 108, no. 6, pp. 493–502, 2004.

[86] V. S. Golubkov, N. L. Prigozhina, Y. Zhang et al., "Protein-Tyrosine Pseudokinase 7 (PTK7) directs cancer cell motility and metastasis," *The Journal of Biological Chemistry*, vol. 289, no. 35, pp. 24238–24239, 2014.

[87] M. E. Ross, "Gene-environment interactions, folate metabolism and the embryonic nervous system," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 2, no. 4, pp. 471–480, 2010.

[88] M. Bouskila, M. F. Hirshman, J. Jensen, L. J. Goodyear, and K. Sakamoto, "Insulin promotes glycogen synthesis in the absence of GSK3 phosphorylation in skeletal muscle," *American Journal of Physiology - Endocrinology and Metabolism*, vol. 294, no. 1, pp. E28–E35, 2008.

[89] V. Matys, O. V. Kel-Margoulis, E. Fricke et al., "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes," *Nucleic Acids Research*, vol. 34, pp. D108–D110, 2006.

[90] M. Morgan-Smith, Y. Wu, X. Zhu, J. Pringle, and W. D. Snider, "GSK-3 signaling in developing cortical neurons is essential for radial migration and dendritic orientation," *eLife*, vol. 3, p. e02663, 2014.

[91] D. Sattarova, M. I. Sigatullina, S. S. Shamansurov, and N. A. Mirsaidova, "Outcome of epilepsy surgery in focal cortical dysplasia," *European Journal of Neurology*, vol. 19, p. 605, 2012.

[92] F. Tanaka, H. Otsubo, W. C. Gaetz et al., "FDG PET and MEG evaluation of focal cortical dysplasia: comparison with the results of intracranial invasive EEG and epilepsy surgery," *Journal of Nuclear Medicine*, vol. 41, p. 220, 2000.

[93] M. Giannandrea, V. Bianchi, M. L. Mignogna et al., "Mutations in the small GTPase gene RAB39B are responsible for X-linked mental retardation associated with autism, epilepsy, and macrocephaly," *American Journal of Human Genetics*, vol. 86, no. 2, pp. 185–195, 2010.

[94] Y. Zhang, J. Liu, G. Luan, and X. Wang, "Inhibition of the small GTPase Cdc42 in regulation of epileptic-seizure in rats," *Neuroscience*, vol. 289, pp. 381–391, 2015.

[95] A. Kretschmann, B. Danis, L. Andonovic et al., "Different MicroRNA profiles in chronic epilepsy versus acute seizure mouse models," *Journal of Molecular Neuroscience*, vol. 55, no. 2, pp. 466–479, 2015.

[96] T. Nakamura, S. Yasuda, H. Nagai et al., "Longest neurite-specific activation of Rap1B in hippocampal neurons contributes to polarity formation through RalA and Nore1A in addition to PI3-kinase," *Genes to Cells*, vol. 18, no. 11, pp. 1020–1031, 2013.

[97] H. Kano, T. Takayama, Y. Midorikawa, and H. Nagase, "Promoter hypomethylation of RAR-related orphan receptor α 1 is correlated with unfavorable clinicopathological features in patients with colorectal cancer," *BioScience Trends*, vol. 10, no. 3, pp. 202–209, 2016.

[98] L. Wang, W. Zhan, S. Xie et al., "Over-expression of Rap2a inhibits glioma migration and invasion by down-regulating p-AKT," *Cell Biology International*, vol. 38, no. 3, pp. 326–334, 2014.

[99] Q. Zhu, L. Wang, Z. Xiao et al., "Decreased expression of Ras-GRF1 in the brain tissue of the intractable epilepsy patients and experimental rats," *Brain Research*, vol. 1493, pp. 99–109, 2013.

[100] M. Adachi, Y. Abe, Y. Aoki, and Y. Matsubara, "Epilepsy in RAS/MAPK syndrome: two cases of cardio-facio-cutaneous syndrome with epileptic encephalopathy and a literature review," *Seizure*, vol. 21, no. 1, pp. 55–60, 2012.

[101] Z. Lu, A. Hornia, T. Joseph et al., "Phospholipase D and RalA cooperate with the epidermal growth factor receptor to transform 3Y1 rat fibroblasts," *Molecular and Cellular Biology*, vol. 20, no. 2, pp. 462–467, 2000.

[102] E. Manguoglu, S. Akdeniz, N. Dündar et al., "RLIP76 gene variants are not associated with drug response in turkish epilepsy patients," *Balkan Journal of Medical Genetics*, vol. 14, no. 1, pp. 25–30, 2011.

[103] C. D. Nobes, I. Lauritzen, M.-G. Mattei, S. Paris, A. Hall, and P. Chardin, "A new member of the Rho family, Rnd1,

promotes disassembly of actin filament structures and loss of cell adhesion," *Journal of Cell Biology*, vol. 141, no. 1, pp. 187–197, 1998.

[104] Y. Ishikawa, H. Katoh, and M. Negishi, "A role of Rnd1 GTPase in dendritic spine formation in hippocampal neurons," *Journal of Neuroscience*, vol. 23, no. 35, pp. 11065–11072, 2003.

[105] C. J. Wingard, V. Chintalgattu, G. Harris, R. Nolan, J. Narron, and L. C. Katwa, "Testosterone-dependent expression of RhoA, ROCK I, ROCK II and Rnd1 in rat corpus cavernosum," *The FASEB Journal*, vol. 19, p. A123, 2005.

[106] S. M. Zanata, I. Hovatta, B. Rohm, and A. W. Püschel, "Antagonistic effects of Rnd1 and RhoD GTPases regulate receptor activity in semaphorin 3A-induced cytoskeletal collapse," *Journal of Neuroscience*, vol. 22, no. 2, pp. 471–477, 2002.

[107] R. A. Teixeira, V. A. Zanardi, L. M. Li, S. L. M. Santos, and F. Cendes, "Epilepsy and destructive brain insults in early life: a topographical classification on the basis of MRI findings," *Seizure*, vol. 13, no. 6, pp. 383–391, 2004.

[108] Y.-S. Bae, W. Chung, K. Han et al., "Down-regulation of RalBP1 expression reduces seizure threshold and synaptic inhibition in mice," *Biochemical and Biophysical Research Communications*, vol. 433, no. 2, pp. 175–180, 2013.

[109] M. C. Lee, S. S. Ban, Y.-J. Woo, and S. U. Kim, "Calcium/calmodulin kinase II activity of hippocampus in kainate-induced epilepsy," *Journal of Korean Medical Science*, vol. 16, no. 5, pp. 643–648, 2001.

[110] L. S. Butler, A. J. Silva, A. Abeliovich, Y. Watanabe, S. Tonegawa, and J. O. Mcnamara, "Limbic epilepsy in transgenic mice carrying a Ca2+/calmodulin-dependent kinase II α-subunit mutation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 15, pp. 6852–6855, 1995.

[111] P. Ambrosino, A. Alaimo, S. Bartollino et al., "Epilepsy-causing mutations in Kv7.2 C-terminus affect binding and functional modulation by calmodulin," *Biochimica et Biophysica Acta—Molecular Basis of Disease*, vol. 1852, no. 9, pp. 1856–1866, 2015.

[112] S. B. Churn, L. D. Kochan, and R. J. Delorenzo, "Chronic inhibition of Ca2+/calmodulin kinase II activity in the pilocarpine model of epilepsy," *Brain Research*, vol. 875, no. 1-2, pp. 66–77, 2000.

[113] A. J. Muslin, J. W. Tanner, P. M. Allen, and A. S. Shaw, "Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine," *Cell*, vol. 84, no. 6, pp. 889–897, 1996.

[114] Y. S. Kim, M. Y. Choi, Y. H. Kim et al., "Protein kinase Cdelta is associated with 14-3-3 phosphorylation in seizure-induced neuronal death," *Epilepsy Research*, vol. 92, no. 1, pp. 30–40, 2010.

[115] S. Shinoda, S. L. Skradski, T. Araki et al., "Formation of a tumour necrosis factor receptor 1 molecular scaffolding complex and activation of apoptosis signal-regulating kinase 1 during seizure-induced neuronal death," *European Journal of Neuroscience*, vol. 17, no. 10, pp. 2065–2076, 2003.

[116] E. A. Knopp, C. Saraceni, J. Moss, J. M. McNiff, and K. A. Choate, "Somatic ATP2A2 mutation in a case of papular acantholytic dyskeratosis: mosaic Darier disease," *Journal of Cutaneous Pathology*, vol. 42, pp. 853–857, 2015.

[117] J. Dhitavat, L. Dode, N. Leslie, S. Burge, and A. Hovnanian, "Effects of mutations in ATP2A2 on calcium transport across sarco/endoplasmic reticulum (ER) membrane," *Journal of Investigative Dermatology*, vol. 117, pp. 774–774, 2001.

[118] A. Takagi, M. Kamijo, and S. Ikeda, "Darier disease," *Journal of Dermatology*, vol. 43, no. 3, pp. 275–279, 2016.

[119] R. P. Dodiuk-Gad, E. Cohen-Barak, M. Khayat et al., "Darier disease in Israel: combined evaluation of genetic and neuropsychiatric aspects," *British Journal of Dermatology*, vol. 174, no. 3, pp. 562–568, 2016.

[120] N. J. O. Jacobsen, I. Lyons, B. Hoogendoorn et al., "ATP2A2 mutations in Darier's disease and their relationship to neuropsychiatric phenotypes," *Human Molecular Genetics*, vol. 8, no. 9, pp. 1631–1636, 1999.

[121] K. Karimi, T. Mahmoudi, N. Karimi et al., "Is there an association between variants in candidate insulin pathway genes IGF-I, IGFBP-3, INSR, and IRS2 and risk of colorectal cancer in the Iranian Population?" *Asian Pacific Journal of Cancer Prevention*, vol. 14, no. 9, pp. 5011–5016, 2013.

[122] S. Pechlivanis, B. Pardini, J. L. Bermejo et al., "Insulin pathway related genes and risk of colorectal cancer: INSR promoter polymorphism shows a protective effect," *Endocrine-Related Cancer*, vol. 14, no. 3, pp. 733–740, 2007.

[123] F. Che, Q. Fu, X. Li et al., "Association of insulin receptor H1085H C>T, insulin receptor substrate 1 G972R and insulin receptor substrate 2 1057G/A polymorphisms with refractory temporal lobe epilepsy in Han Chinese," *Seizure*, vol. 25, pp. 178–180, 2015.

[124] Q. Wang, J. Qian, F. Wang, and Z. Ma, "Cellular prion protein accelerates colorectal cancer metastasis via the Fyn-SP1-SATB1 axis," *Oncology Reports*, vol. 28, no. 6, pp. 2029–2034, 2012.

[125] A. Strom, S. Diecke, G. Hunsmann, and A. W. Stuke, "Cellular prion protein promotes glucose uptake through the Fyn-HIF-2 alpha-Glut1 pathway to support colorectal cancer cell survival," *Cancer Science*, vol. 103, pp. 606–606, 2011.

[126] D. P. Cain, S. G. N. Grant, D. Saucier, E. L. Hargreaves, and E. R. Kandel, "Fyn tyrosine kinase is required for normal amygdala kindling," *Epilepsy Research*, vol. 22, no. 2, pp. 107–114, 1995.

[127] X. Yang, C. Marshall, T. Dentchev et al., "A topical PI3K/mTOR inhibitor induces regression of squamous cell carcinomas in K14-Fyn Y528F mice," *Journal of Investigative Dermatology*, vol. 132, p. S25, 2012.

[128] Y. Bermudez, S. P. Stratton, G. T. Bowden et al., "Abstract 3673: expression profile of phosphorylated proteins from the mTOR and Fyn/RSK2 signaling pathways in solar UV-induced skin carcinogenesis," *Cancer Research*, vol. 71, no. 8 supplement, pp. 3673–3673, 2011.

[129] B. Chang, T. Grau, S. Dangel et al., "A homologous genetic basis of the murine cpfl1 mutant and human achromatopsia linked to mutations in the PDE6C gene," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 46, pp. 19581–19586, 2009.

[130] S. E. Martinez, C. C. Heikaus, R. E. Klevit, and J. A. Beavo, "The structure of the GAF a domain from phosphodiesterase 6C reveals determinants of cGMP binding, a conserved binding surface, and a large cGMP-dependent conformational change," *Journal of Biological Chemistry*, vol. 283, no. 38, pp. 25913–25919, 2008.

[131] A. Filla, G. De Michele, S. Cocozza et al., "Early onset autosomal dominant dementia with ataxia, extrapyramidal features, and epilepsy," *Neurology*, vol. 58, no. 6, pp. 922–928, 2002.

[132] M. Coutelier, I. Blesneac, A. Monteil et al., "A recurrent mutation in CACNA1G alters Cav3.1 T-type calcium-channel conduction and causes autosomal-dominant cerebellar ataxia," *American Journal of Human Genetics*, vol. 97, no. 5, pp. 726–737, 2015.

[133] H. Wang, D. Xu, M. F. Toh, A. C. Pao, and G. You, "Serum- and glucocorticoid-inducible kinase SGK2 regulates human organic

anion transporters 4 via ubiquitin ligase Nedd4-2," *Biochemical Pharmacology*, vol. 102, pp. 120–129, 2016.

[134] B. Friedrich, Y. Feng, P. Cohen et al., "The serine/threonine kinases SGK2 and SGK3 are potent stimulators of the epithelial Na+ channel α,β,γ-ENaC," *Pflugers Archiv European Journal of Physiology*, vol. 445, no. 6, pp. 693–696, 2003.

[135] M. J. Seller and R. G. Spector, "Effect of aldosterone and cortisol on leptazol-induced seizures in rats," *British Journal of Pharmacology and Chemotherapy*, vol. 19, no. 2, pp. 271–273, 1962.

[136] W. Tan, S.-G. Lim, and T. M. C. Tan, "Up-regulation of microRNA-210 inhibits proliferation of hepatocellular carcinoma cells by targeting Yes1," *World Journal of Gastroenterology*, vol. 21, no. 46, pp. 13030–13041, 2015.

[137] P. R. Patel, H. Sun, S. Q. Li et al., "Identification of potent Yes1 kinase inhibitors using a library screening approach," *Bioorganic & Medicinal Chemistry Letters*, vol. 23, no. 15, pp. 4398–4403, 2013.

Hindawi
BioMed Research International
Volume 2017, Article ID 7259097, 7 pages
https://doi.org/10.1155/2017/7259097

Research Article

Gastric Cancer Associated Genes Identified by an Integrative Analysis of Gene Expression Data

Bing Jiang,[1] Shuwen Li,[2] Zhi Jiang,[3] and Ping Shao[1]

[1]*Department of Spleen and Stomach Diseases, Hospital of Traditional Chinese Medicine, Yixing, Jiangsu 214200, China*
[2]*Department of Gastroenterology, The First Affiliated Hospital of Soochow University, Suzhou, Jiangsu 215006, China*
[3]*Department of Biochemistry and Molecular Biology, School of Medicine, Soochow University, Suzhou, Jiangsu 215123, China*

Correspondence should be addressed to Ping Shao; shaoping_yixing@163.com

Received 20 December 2016; Revised 29 December 2016; Accepted 4 January 2017; Published 23 January 2017

Academic Editor: Xingming Zhao

Gastric cancer is one of the most severe complex diseases with high morbidity and mortality in the world. The molecular mechanisms and risk factors for this disease are still not clear since the cancer heterogeneity caused by different genetic and environmental factors. With more and more expression data accumulated nowadays, we can perform integrative analysis for these data to understand the complexity of gastric cancer and to identify consensus players for the heterogeneous cancer. In the present work, we screened the published gene expression data and analyzed them with integrative tool, combined with pathway and gene ontology enrichment investigation. We identified several consensus differentially expressed genes and these genes were further confirmed with literature mining; at last, two genes, that is, immunoglobulin J chain and C-X-C motif chemokine ligand 17, were screened as novel gastric cancer associated genes. Experimental validation is proposed to further confirm this finding.

1. Introduction

Gastric cancer (GC) is one of the most severe cancers in the world with high incidence and low survival rate. According to the global cancer statistics report in 2012, GC has been the fifth most common cancer in the world, which causes more than seven hundred thousand deaths each year [1]. Usually, the number of GC patients in men is twice more than that in women and Eastern Asia, especially Korea, Japan, and China, has the highest incidence rate. Although relevant reports revealed that the age-standardized incidence rate of gastric cancer is decreasing in Japan and Korea in last few years [2, 3], the number of new cases is still increasing due to the aging of the population. The pathogenesis of gastric cancer is very complex and remains unclear. Recent basic studies mainly focus on three main factors: environmental factors, *Helicobacter pylori* (*H. pylori*) infection, and gene expression dysregulation [4, 5]. Previous studies have demonstrated the unhealthy lifestyle, such as excessive diet, can raise the risk of gastric cancer [5–7]. Processed meat intakes will increase the risk of gastric non-cardia cancer in *H. pylori* antibody-positive individuals while fresh fruits and vegetables consumption will protect individuals against GC. Also, in molecular level, several host genetic factors might play a key role in GC, such as IL-1β, IL-10, TFF2, and CDH1 [8–10].

With relevant studies deepening, the size of research data is becoming larger and larger. Hundreds of gene expression profiles and diagnostic targets are uploaded into various gene expression databases. These data can be further integrated to the understanding of the complexity of the diseases, such as the cancer heterogeneity, high level consensus [11–13], biomarker discovery [14, 15], and the key players in the cancer genesis and progress [16]. In this study, we used meta-analysis approach for analysis of multiple transcriptomic datasets. We hope to integrate different gene expression data collected from GC patients and normal controls to figure out robust candidates in genes, pathways, and functions, setting the foundation for personalized treatment of gastric cancer.

The method we used here was named INMEX (integrative meta-analysis of expression data) program [17]. Data procession and screening were performed in order to make sure all the datasets we uploaded into the program were in a consistent format. Due to the existence of outliers and

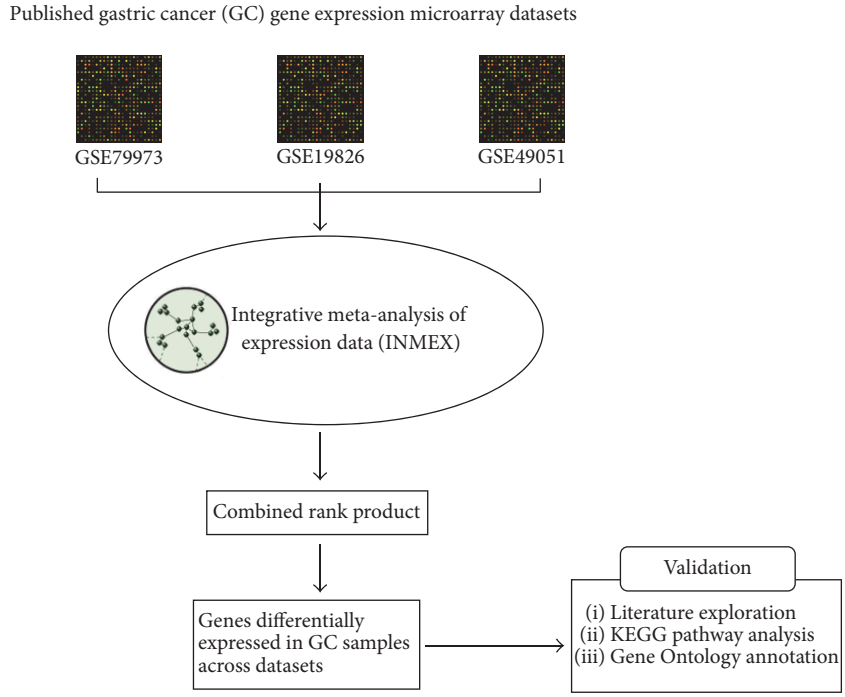


Figure 1: The pipeline of the whole analysis in this study.

variations in microarray data, a combining rank orders algorithm based on RankProd package [18] was used here to carry out the meta-analysis.

2. Materials and Methods

The pipeline of this whole analysis in the present study is shown in Figure 1. We first extract the microarray gene expression data from the GEO database, then we integrate analyzed the expression data with a meta-analysis tool INMEX, and then we further screen and validate the meta-analysis results with literature analysis and bioinformatics functional analysis.

2.1. Dataset Collection and Data Screening. We used keywords "gastric cancer," with two filters: (a) organism: *Homo sapiens* and (b) type: expression profiling by array, in searching for the gene expression profiles in Gene Expression Omnibus (GEO) database. We explored the searching results by setting four inclusion criteria: (1) datasets published after 2010; (2) case-control studies; (3) sample numbers more than 20; (4) high similarity in sample background information (i.e., sources, patients' race and location, disease status, and platforms). Datasets meeting these criteria were selected for further analysis.

2.2. Meta-Analysis for Selected Datasets. Based on the expression data we collected from each qualified microarray study, a global meta-analysis for identifying differentially expressed (DE) genes in gastric genes was conducted in this study. Here, we selected a web-based tool named INMEX (integrative meta-analysis of expression data, http://www.networkanalyst.ca/) for meta-analysis.

We firstly upload the normalized gene expression datasets into INMEX. Then we processed and annotated the datasets to adjust the data format and class labels into the consistent style. After the integrity check, we selected combining rank orders method, which is based on the RankProd package, to carry out the meta-analysis. The number of permutation tests in this method was 20 times.

2.3. Functional Enrichment Analysis of DE Genes. Functional enrichment analysis of these DE genes was further performed by INMEX program in two approaches: Geno Ontology and pathway analysis. In GO annotation, we set a p value threshold of 0.05 to identify the significantly enriched items. In pathway analysis, KEGG pathway database was used here for pathway enrichment analysis. A p value threshold of 0.05 was also set for identification of significantly enriched pathways.

3. Results

3.1. Characteristics of Datasets Included in This Meta-Analysis. The datasets selection strategy and the screening results are presented in Figure 2. Through GEO datasets searching, a total of 1722 studies were retrieved. 1618 irrelevant studies were excluded, among which 1605 studies were not expression profiling by microarray technologies and 13 studies were animal studies. The remaining 104 studies were included for full-text review. Studies without case-control matches were then excluded. Due to the platform limitation, we further excluded those studies whose microarray platforms are not available in INMEX program. After several rounds of screening, a final list of 3 microarray datasets [19, 20] was selected for meta-analysis.

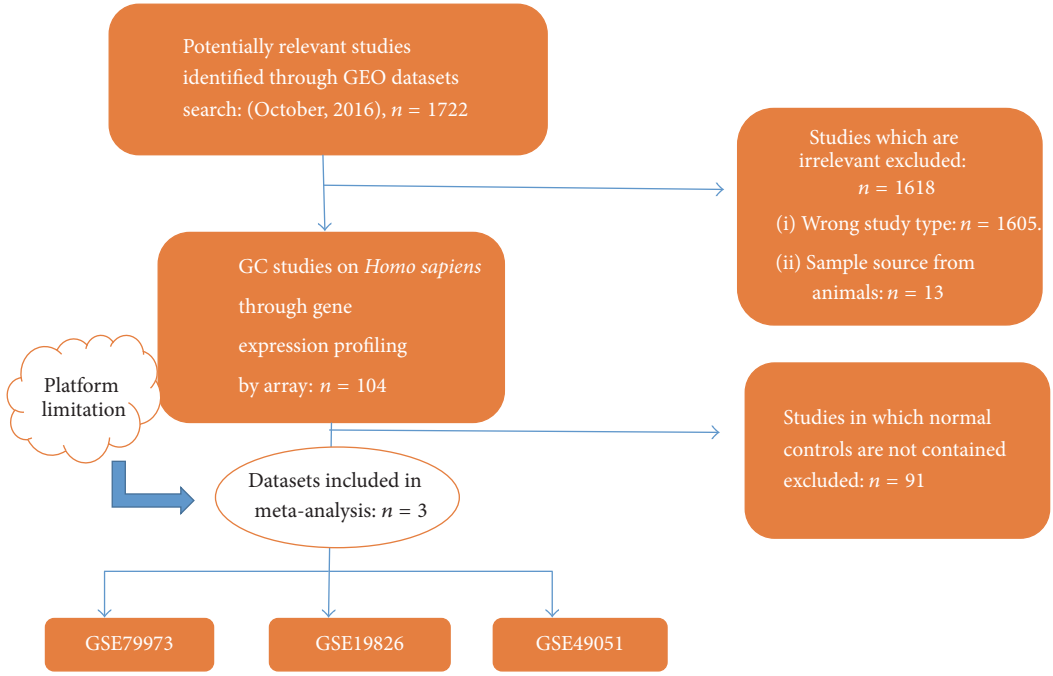


Figure 2: Datasets selection strategy and results.

Table 1: Datasets selected in this meta-analysis.

Accession/ID	Platform	GC	Control	Materials	Year	Race	Region
GSE79973	GPL570	$n = 10$	$n = 10$	Gastric tissues	2016	Chinese	Hangzhou
GSE19826	GPL570	$n = 12$	$n = 12$	Gastric tissue	2010	Chinese	Shanghai
GSE49051	GPL10332	$n = 3$	$n = 3$	Gastric tissue	2013	Chinese	Shanghai

These 3 datasets (GSE79973, GSE19826, and GSE49051) contain totally 25 cases and 25 controls. The number of cases and controls of each dataset is well matched. All the datasets were collected from Chinese hospitals and sample sources are consistent. The detailed information of these 3 datasets is listed in Table 1.

3.2. Results of Meta-Analysis. This study is performed based on combining rank orders. DE genes with p value < 0.05 were selected. Totally 1153 DE genes were got through this meta-analysis. The detailed DE gene information was listed in Table S1 (see Supplementary Material available online at https://doi.org/10.1155/2017/7259097). All of these DE genes are those identified to be differentially expressed in these three datasets rather than in individual samples. Among the 1153 DE genes, 787 genes were downregulated and 366 genes were upregulated.

The top 10 most significantly upregulated genes and top 10 most downregulated genes were listed in Tables 2 and 3. Genes with the smallest combined rank product (RP) in upregulated DE gene list and downregulated DE gene list are COL6A3 (combinedRP = 59.02) and PGC (combinedRP = 22.38), respectively.

3.3. Functional Enrichment Analysis Results. Functional enrichment analysis was carried out for further study of these DE genes. Gene Oncology (GO) analysis and KEGG pathway analysis were the two approaches we conducted here. In GO analysis, we did the analysis at three levels: biological process (BP), cellular component (CC), and molecular function (MF). The top 10 most significantly enriched terms (adj. p value < 0.05) were selected, respectively. The histograms of these terms were shown in Figure 3. Most of the DE genes are well mapped onto gastric cancer associated process of biological factors. In KEGG pathway analysis, we also selected top 10 most significantly enriched pathways, as shown in Figure 4. All of the selected items were taken into literature validation for further investigation.

4. Discussion

In this study, we have used publicly available microarray datasets to identify genes that are differentially expressed in tumor tissues from people with GC comparing to people without GC. The aim of our study is to derive additional information from the combining datasets that are unlikely to be established from individual studies in isolation through combining the data from three separate gene expression datasets in a meta-analysis. Generally, we found this is to be the case. Through PubMed literature mining, we found 8 of 10 of downregulated genes and all the upregulated genes

Table 2: Top 10 most significantly downregulated DE genes in gastric cancer.

EntrezID	Gene full name	Gene symbol	CombinedRP	AveLogFC
5225	Progastricsin	PGC	22.38	−14454.48
57016	Aldo-keto reductase family 1 member B10	AKR1B10	22.86	−6705.56
9992	Potassium voltage-gated channel subfamily E regulatory subunit 2	KCNE2	36.08	−4314.33
284340	C-X-C motif chemokine ligand 17	CXCL17	49.18	−3880.57
135656	Diffuse panbronchiolitis critical region 1	DPCR1	55.94	−1892.04
51208	Claudin 18	CLDN18	57.35	−3435.26
3512	Immunoglobulin J polypeptide, linker protein for immunoglobulin alpha, and mu polypeptides	IGJ	62.81	−25978.05
1510	Cathepsin E	CTSE	64.51	−4631.28
340547	V-set and immunoglobulin domain containing 1	VSIG1	69.66	−1752.54
4499	Metallothionein 1M	MT1M	100.58	−6787.14

Table 3: Top 10 most significantly upregulated DE genes in gastric cancer.

EntrezID	Gene full name	Gene symbol	CombinedRP	AveLogFC
1293	Collagen type VI alpha 3 chain	COL6A3	59.02	3600.96
1278	Collagen type I alpha 2 chain	COL1A2	62.06	3576.21
10562	Olfactomedin 4	OLFM4	150.67	3542.76
7058	Thrombospondin 2	THBS2	163.66	24.03
115908	Diffuse panbronchiolitis critical region 1	CTHRC1	174.61	1204.56
4680	Collagen triple helix repeat containing 1	CEACAM6	203.78	2542.02
3624	Inhibin beta A subunit	INHBA	219.12	368.69
1290	Collagen type V alpha 2 chain	COL5A2	230.72	1064.68
54829	Asporin	ASPN	255.15	237.05
1366	Claudin 7	CLDN7	288.09	356.71

have been reported to be associated with gastric cancer by biological and clinical experiment validation. For example, downregulated gene with smallest combinedRP in this study is Progastricsin (PGC). Many researchers have found it plays a key role in gastric cancer and the PGC polymorphism could serve as one of the diagnosis biomarkers for GC [21–23]. Also, in a recent research, Li et al. found, in mitogen-activated protein kinase activator with WD40 repeats (MAWD) and MAWD-binding protein (MAWBP) downregulated GC cells, the expression level of PGC was lower than that in control samples [24]. In upregulated genes, collagen VI $\alpha 3$ (COL6A3) is the gene with smallest combinedRP. Relevant research has found the expression level of COL6A3 was significantly higher in GC patients [25, 26], which also could serve as a diagnosis biomarker for GC. Other DE genes, such as COL1A2 [26], OLFM4 [27], THBS2 [28], CEACAM6 [29], CTSE [30], AKR1B10 [31], and KCNE2 [32], also have been reported to be differentially expressed in GC patients comparing to controls.

Interestingly, in the top 10 downregulated DE genes, 2 genes (IGJ and CXCL17) have not been reported to have a direct association with GC. For IGJ, Tvarijonaviciute et al. have observed that, in obese dogs, the amount of IGJ proteins was decreased [33]. Relevant research has revealed that obesity will increase the risk of gastric cancer [34]. For CXCL17, it is reported that overexpression of CXCL17 has a strong connection with colon cancer and hepatocellular carcinoma [35, 36]. The existence of gene interaction reveals the association between GC and these two cancers [37, 38]. Because there are still no specific experiments on these two genes and GC, further biological and clinical research are needed.

To further investigate the functional mechanisms of these DE genes, we performed GO analysis and KEGG pathway analysis. We finally get 102 significantly enriched terms (p value < 0.05) in biological process level, 157 in cellular component level, and 31 in molecular function level. As shown above, the top 10 significantly enriched terms were all reported to be associated with GC. For example, in extracellular matrix, there exists extracellular matrix protein 1 (ECM1). ECM1 plays a key role in lymphangiogenesis [39], which could be an inducement of cancer invasion and metastasis. Aberrant expression of ECM1 was found in GC samples in a recent study [40]. Also, in translational elongation process, relevant genes, such as translation elongation factor EEF1B2, were upregulated in the poor prognosis samples [41]. All the top 10 terms in BP, CC, and MF have been reported to have an association with GC.

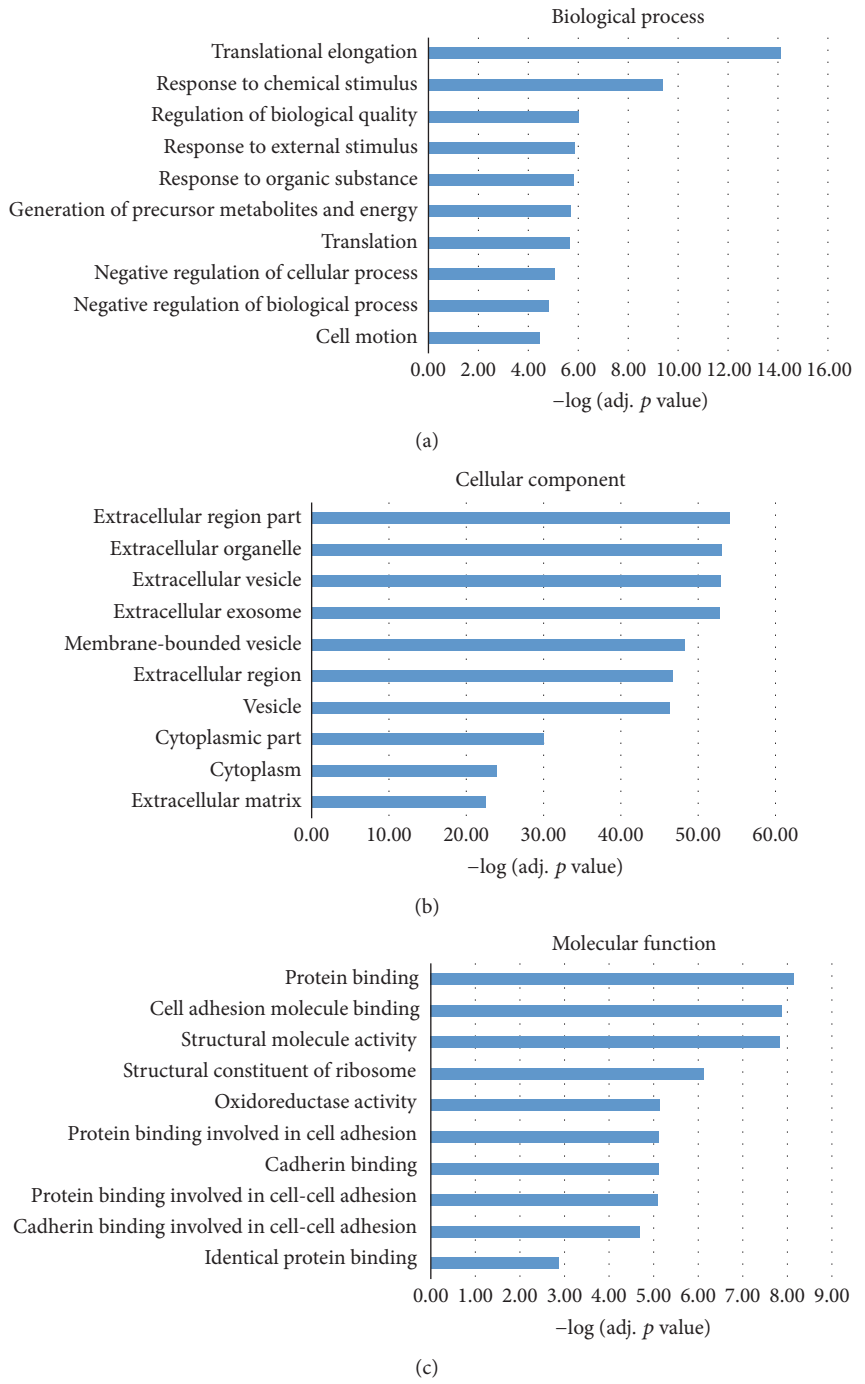


Figure 3: Gene Oncology (GO) annotation for the DE genes in gastric cancer. Here the GO annotation was used at three levels: biological process, cellular component, and molecular function. (a), (b), and (c) represent the top 10 most significantly enriched GO terms for these DE genes, respectively. All the adjusted statistical significance value (p value) of the terms was negative 10-based transformed.

In KEGG pathway analysis, the most significantly enriched pathway is Ribosome. Genes such as RPL11, RPL23, RPS6, and MRPS21 were enriched on this pathway. Ribosomal protein family (PRL/RPS) has been demonstrated to have a strong connection with GC. For example, a recent study revealed that GLTSCR2 regulates the MDM2-TP53 pathway through RPL11, playing a key role in GC progression [42]. A previous study has observed that reducing the phosphorylation of RPS6 could have an influence on the sensitivity to MEK inhibition in gastric cancer cells [43]. Another important pathway in GC is glycolysis/gluconeogenesis pathway. Reports revealed that microRNA-133b could silence PKM-splicer PTBP1, leading the inhibition of growth of human gastric cancer cells [44]. Hu and Chen also found that SIRT3 can strengthen glycolysis in SIRT3-expressing GC cells. Other pathways, like ECM-receptor interaction and

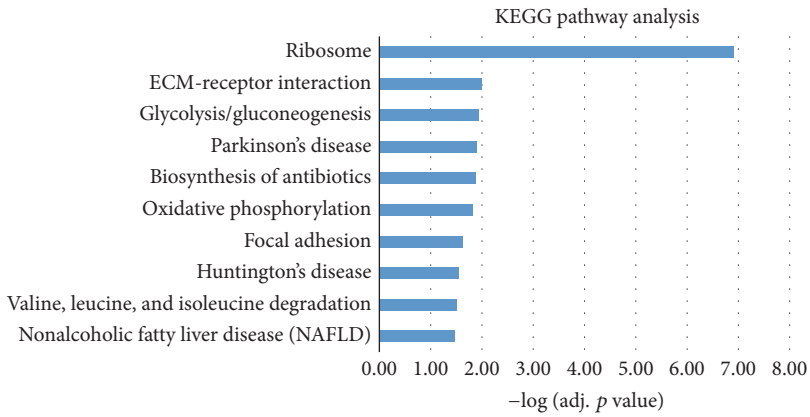


Figure 4: The top 10 most significantly enriched pathways in KEGG pathway analysis for the DE genes in gastric cancer. The adjusted statistical significance value (p value) was negative 10-based log transformed.

metabolism of xenobiotics by cytochrome P450, have been validated to be associated with GC through bioinformatics approaches based protein-protein interaction networks analysis [45].

5. Conclusions

To summarize, our research provides novel angels in pathogenesis of gastric cancer. We identified consistently DE genes in gastric cancer through INMEX meta-analysis tools. Top 10 of upregulated and downregulated genes could potentially serve as diagnosis biomarker. GO annotation and KEGG pathway analysis demonstrated those candidates have a strong relationship with gastric cancer. Moreover, we identified 2 novel GC associated genes, IGJ and CXCL17, which have never been reported to be associated with GC before. Further experimental validation should be conducted in order to understand the mechanism of these two genes on gastric cancer.

Competing Interests

The authors declare that there is no conflict of interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 31470821). Thanks are due to Mr. Shen Li in Suzhou Eastern Science, Technology and Culture Co., Ltd., Suzhou, Jiangsu 215123, China, for the help with data preparation and analysis.

References

[1] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[2] M. Inoue and S. Tsugane, "Epidemiology of gastric cancer in Japan," *Postgraduate Medical Journal*, vol. 81, no. 957, pp. 419–424, 2005.

[3] H. Shin, Y. Won, K. Jung et al., "Nationwide cancer incidence in Korea, 1999~2001; first result using the national cancer incidence database," *Cancer Research and Treatment*, vol. 37, no. 6, pp. 325–331, 2005.

[4] D. M. Parkin, "The global health burden of infection-associated cancers in the year 2002," *International Journal of Cancer*, vol. 118, no. 12, pp. 3030–3044, 2006.

[5] S. Tsugane and S. Sasazuki, "Diet and the risk of gastric cancer: review of epidemiological evidence," *Gastric Cancer*, vol. 10, no. 2, pp. 75–83, 2007.

[6] C. A. Gonzalez, L. Lujan-Barroso, H. B. Bueno-De-Mesquita et al., "Fruit and vegetable intake and the risk of gastric adenocarcinoma: a reanalysis of the european prospective investigation into cancer and nutrition (EPIC-EURGAST) study after a longer follow-up," *International Journal of Cancer*, vol. 131, no. 12, pp. 2910–2919, 2012.

[7] C. A. Gonzalez, P. Jakszyn, G. Pera et al., "Meat intake and risk of stomach and esophageal adenocarcinoma within the European Prospective Investigation Into Cancer and Nutrition (EPIC)," *Journal of the National Cancer Institute*, vol. 98, no. 5, pp. 345–354, 2006.

[8] J. Bornschein, T. Rokkas, M. Selgrad, and P. Malfertheiner, "Gastric cancer: clinical aspects, epidemiology and molecular background," *Helicobacter*, vol. 16, supplement 1, pp. 45–52, 2011.

[9] H. Kim, J. W. Eun, H. Lee et al., "Gene expression changes in patient-matched gastric normal mucosa, adenomas, and carcinomas," *Experimental and Molecular Pathology*, vol. 90, no. 2, pp. 201–209, 2011.

[10] A. Thiel and A. Ristimäki, "Gastric cancer: basic aspects," *Helicobacter*, vol. 17, no. 1, pp. 26–29, 2012.

[11] Y. Wang, J. Chen, Q. Li et al., "Identifying novel prostate cancer associated pathways based on integrative microarray data analysis," *Computational Biology and Chemistry*, vol. 35, no. 3, pp. 151–158, 2011.

[12] Y. Tang, W. Yan, J. Chen, C. Luo, A. Kaipia, and B. Shen, "Identification of novel microRNA regulatory pathways associated with heterogeneous prostate cancer," *BMC Systems Biology*, vol. 7, supplement 3, article S6, 2013.

[13] Y. Hu, J. Li, W. Yan et al., "Identifying novel glioma associated pathways based on systems biology level meta-analysis," *BMC Systems Biology*, vol. 7, supplement 2, p. S9, 2013.

[14] Y. Li, W. Vongsangnak, L. Chen, and B. Shen, "Integrative analysis reveals disease-associated genes and biomarkers for prostate cancer progression," *BMC Medical Genomics*, vol. 7, no. 1, article S3, 2014.

[15] Y. Zhu, Q. Peng, Y. Lin et al., "Identification of biomarker microRNAs for predicting the response of colorectal cancer to neoadjuvant chemoradiotherapy based on microRNA regulatory network," *Oncotarget*, 2016.

[16] J. Jiang, P. Jia, Z. Zhao, and B. Shen, "Key regulators in prostate cancer identified by co-expression module analysis," *BMC Genomics*, vol. 15, no. 1, article no. 1015, 2014.

[17] J. Xia, C. D. Fjell, M. L. Mayer, O. M. Pena, D. S. Wishart, and R. E. W. Hancock, "INMEX—a web-based tool for integrative meta-analysis of expression data," *Nucleic Acids Research*, vol. 41, pp. W63–W70, 2013.

[18] F. Hong, R. Breitling, C. W. McEntee, B. S. Wittner, J. L. Nemhauser, and J. Chory, "RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis," *Bioinformatics*, vol. 22, no. 22, pp. 2825–2827, 2006.

[19] Q. Wang, Y.-G. Wen, D.-P. Li et al., "Upregulated INHBA expression is associated with poor survival in gastric cancer," *Medical Oncology*, vol. 29, no. 1, pp. 77–83, 2012.

[20] T. Sun, W. Du, H. Xiong et al., "TMEFF2 deregulation contributes to gastric carcinogenesis and indicates poor survival outcome," *Clinical Cancer Research*, vol. 20, no. 17, pp. 4689–4704, 2014.

[21] H.-J. Liu, X.-L. Guo, M. Dong, L. Wang, and Y. Yuan, "Association between pepsinogen C gene polymorphism and genetic predisposition to gastric cancer," *World Journal of Gastroenterology*, vol. 9, no. 1, pp. 50–53, 2003.

[22] A. L. Pinto-Correia, H. Sousa, M. Fragoso et al., "Gastric cancer in a Caucasian population: role of pepsinogen C genetic variants," *World Journal of Gastroenterology*, vol. 12, no. 31, pp. 5033–5036, 2006.

[23] L.-P. Sun, X.-L. Guo, Y. Zhang et al., "Impact of pepsinogen C polymorphism on individual susceptibility to gastric cancer and its precancerous conditions in a Northeast Chinese population," *Journal of Cancer Research and Clinical Oncology*, vol. 135, no. 8, pp. 1033–1039, 2009.

[24] D. Li, J. Zhang, Y. Xi et al., "Mitogen-activated protein kinase activator with WD40 repeats (MAWD) and MAWD-binding protein induce cell differentiation in gastric cancer," *BMC Cancer*, vol. 15, no. 1, article 637, 2015.

[25] X. Xie, X. Liu, Q. Zhang, and J. Yu, "Overexpression of collagen VI $\alpha 3$ in gastric cancer," *Oncology Letters*, vol. 7, pp. 1537–1543, 2014.

[26] H. Sun, "Identification of key genes associated with gastric cancer based on DNA microarray data," *Oncology Letters*, vol. 11, no. 1, pp. 525–530, 2016.

[27] X. Ran, X. Xu, Y. Yang et al., "A quantitative proteomics study on olfactomedin 4 in the development of gastric cancer," *International Journal of Oncology*, vol. 47, no. 5, pp. 1932–1944, 2015.

[28] X. Lin, D. Hu, G. Chen et al., "Associations of THBS2 and THBS4 polymorphisms to gastric cancer in a Southeast Chinese population," *Cancer Genetics*, vol. 209, no. 5, pp. 215–222, 2016.

[29] R. K. Roy, M. M. Hoppe, S. Srivastava et al., "CEACAM6 is upregulated by *Helicobacter pylori* CagA and is a biomarker for early gastric cancer," *Oncotarget*, vol. 7, no. 34, pp. 55290–55301, 2016.

[30] M. Konno-Shimizu, N. Yamamichi, K.-I. Inada et al., "Cathepsin E is a marker of gastric differentiation and signet-ring cell carcinoma of stomach: a novel suggestion on gastric tumorigenesis," *PLoS ONE*, vol. 8, no. 2, Article ID e56766, 2013.

[31] H. B. Yao, Y. Xu, L. Chen et al., "AKR1B10, a good prognostic indicator in gastric cancer," *European Journal of Surgical Oncology*, vol. 40, no. 3, pp. 318–324, 2014.

[32] P. Yanglin, Z. Lina, L. Zhiguo et al., "KCNE2, a down-regulated gene identified by in silico analysis, suppressed proliferation of gastric cancer cells," *Cancer Letters*, vol. 246, no. 1-2, pp. 129–138, 2007.

[33] A. Tvarijonaviciute, J. J. Ceron, C. de Torre et al., "Obese dogs with and without obesity-related metabolic dysfunction—a proteomic approach," *BMC Veterinary Research*, vol. 12, no. 1, article 211, 2016.

[34] M. Song, J. Choi, J. J. Yang et al., "Obesity at adolescence and gastric cancer risk," *Cancer Causes & Control*, vol. 26, no. 2, pp. 247–256, 2015.

[35] L. Ohlsson, M.-L. Hammarström, G. Lindmark, S. Hammarström, and B. Sitohy, "Ectopic expression of the chemokine CXCL17 in colon cancer cells," *British Journal of Cancer*, vol. 114, no. 6, pp. 697–703, 2016.

[36] L. Li, J. Yan, J. Xu et al., "CXCL17 expression predicts poor prognosis and correlates with adverse immune infiltration in hepatocellular carcinoma," *PLoS ONE*, vol. 9, no. 10, Article ID e110064, 2014.

[37] L. Wang, L. Lin, X. Chen et al., "Metastasis-associated in colon cancer-1 promotes vasculogenic mimicry in gastric cancer by upregulating TWIST1/2," *Oncotarget*, vol. 6, no. 13, pp. 11492–11506, 2015.

[38] L. Liu, C. Zhou, L. Zhou et al., "Functional FEN1 genetic variants contribute to risk of hepatocellular carcinoma, esophageal cancer, gastric cancer and colorectal cancer," *Carcinogenesis*, vol. 33, no. 1, pp. 119–123, 2012.

[39] T. Uchida, H. Hayashi, M. Inaoki, T. Miyamoto, and W. Fujimoto, "A failure of mucocutaneous lymphangiogenesis may underlie the clinical features of lipoid proteinosis," *British Journal of Dermatology*, vol. 156, no. 1, pp. 152–157, 2007.

[40] Q. Wu, X. Li, H. Yang, C. Lu, J. You, and Z. Zhang, "Extracellular matrix protein 1 is correlated to carcinogenesis and lymphatic metastasis of human gastric cancer," *World Journal of Surgical Oncology*, vol. 12, no. 1, article 132, 2014.

[41] C. H. Kwon, H. J. Park, Y. R. Choi et al., "PSMB8 and PBK as potential gastric cancer subtype-specific biomarkers associated with prognosis," *Oncotarget*, vol. 7, no. 16, pp. 21454–21468, 2016.

[42] R. Uchi, R. Kogo, K. Kawahara et al., "PICT1 regulates TP53 via RPL11 and is involved in gastric cancer progression," *British Journal of Cancer*, vol. 109, no. 8, pp. 2199–2206, 2013.

[43] Y. Hirashita, Y. Tsukamoto, K. Yanagihara et al., "Reduced phosphorylation of ribosomal protein S6 is associated with sensitivity to MEK inhibition in gastric cancer cells," *Cancer Science*, vol. 107, no. 12, pp. 1919–1928, 2016.

[44] T. Sugiyama, K. Taniguchi, N. Matsuhashi et al., "MiR-133b inhibits growth of human gastric cancer cells by silencing pyruvate kinase muscle-splicer polypyrimidine tract-binding protein 1," *Cancer Science*, vol. 107, no. 12, pp. 1767–1775, 2016.

[45] K. Hu and F. Chen, "Identification of significant pathways in gastric cancer based on protein-protein interaction networks and cluster analysis," *Genetics and Molecular Biology*, vol. 35, no. 3, pp. 701–708, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 4618323, 11 pages
http://dx.doi.org/10.1155/2016/4618323

Research Article

Novel Biomarker MicroRNAs for Subtyping of Acute Coronary Syndrome: A Bioinformatics Approach

Yujie Zhu,[1,2,3] Yuxin Lin,[1] Wenying Yan,[1] Zhandong Sun,[1] Zhi Jiang,[4] Bairong Shen,[1] Xiaoqian Jiang,[2] and Jingjing Shi[5]

[1]*Center for Systems Biology, Soochow University, Suzhou 215006, China*
[2]*Biomedical Informatics Division, UC San Diego, La Jolla, CA, USA*
[3]*Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School, Nanjing, Jiangsu 210008, China*
[4]*School of Medicine, Soochow University, Suzhou 215123, China*
[5]*Department of Cardiovascular Internal Medicine, Wuxi Third People's Hospital, Wuxi 214041, China*

Correspondence should be addressed to Jingjing Shi; jjshi_wuxi3yuan@163.com

Received 20 October 2016; Accepted 27 October 2016

Academic Editor: Xingming Zhao

Acute coronary syndrome (ACS) is a life-threatening disease that affects more than half a million people in United States. We currently lack molecular biomarkers to distinguish the unstable angina (UA) and acute myocardial infarction (AMI), which are the two subtypes of ACS. MicroRNAs play significant roles in biological processes and serve as good candidates for biomarkers. In this work, we collected microRNA datasets from the Gene Expression Omnibus database and identified specific microRNAs in different subtypes and universal microRNAs in all subtypes based on our novel network-based bioinformatics approach. These microRNAs were studied for ACS association by pathway enrichment analysis of their target genes. AMI and UA were associated with 27 and 26 microRNAs, respectively, nine of them were detected for both AMI and UA, and five from each subtype had been reported previously. The remaining 22 and 21 microRNAs are novel microRNA biomarkers for AMI and UA, respectively. The findings are then supported by pathway enrichment analysis of the targets of these microRNAs. These novel microRNAs deserve further validation and will be helpful for personalized ACS diagnosis.

1. Introduction

Acute coronary syndrome (ACS) is caused by decreased blood flow in the coronary arteries arising from thrombus formation and possible coronary vasospasm, which may further lead to heart muscle dysfunction or even death [1]. In 2010, it was estimated that the number of hospital discharges with ACS was 625,000 in the United States, and secondary discharge diagnoses showed the number of inpatient hospital discharges was 1,141,000 for ACS [2]. The death toll range from ACS is the same as sepsis [3]. ACS is not only one of the severest diseases but also an economic burden to society, costing Americans more than 150 billion dollars annually [4]. The two subtypes of ACS are unstable angina (UA) (38%) and acute myocardial infarction (AMI), including ST-elevation myocardial infarction (30%) and non-ST-elevation myocardial infarction (25%) [5].

Over the past few decades, patients were usually checked with an initial evaluation and assessed with a risk score or prediction algorithms considering clinical history, physical examination, and other indices [6–9]. Additional tests have been added to these assessments including electrocardiogram [10], coronary computed tomographic angiography [11], muscle and brain fraction of creatine kinase [12], or blood tests such as troponin I or T [13]. However, the current methods are insufficient for a highly sensitive and specific diagnosis, especially in distinguishing AMI from UA. In addition, there is a phenomenon called "silent" myocardial infarction, which is estimated to occur in around 64% of cases, in which patients do not have chest pain or other symptoms [14]. It

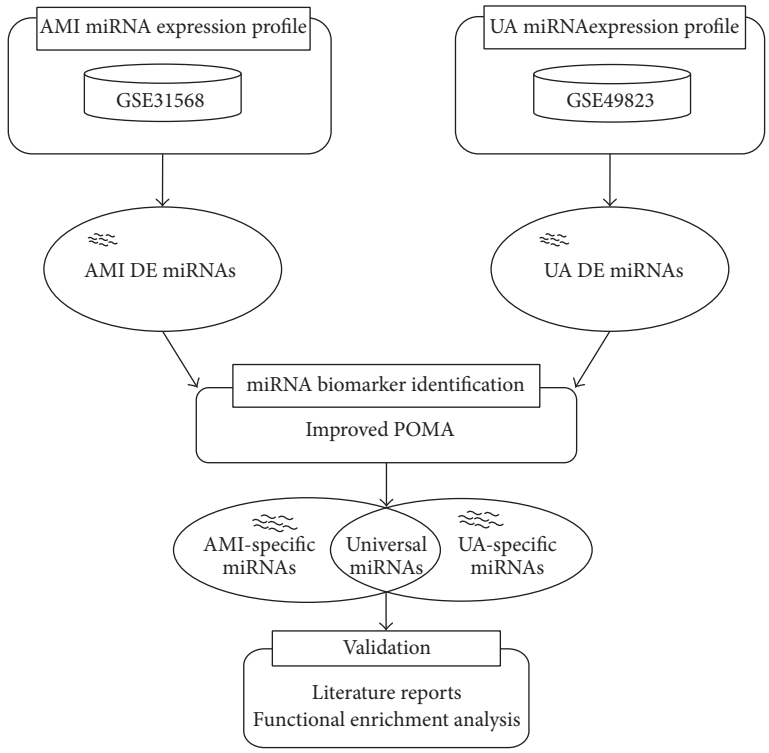


Figure 1: Schematic diagram for the identification of candidate miRNA biomarkers in acute myocardial infarction (AMI) and unstable angina (UA). Here, "DE" is the abbreviation of "differentially expressed."

Table 1: Summary of the miRNA datasets used in this study.

Subtype	GEO accession	Platform	Number of probes	Number of samples (control/disease)
AMI	GSE 31568	Febit Homo Sapiens miRBase 13.0	866	91 (70/21)
UA	GSE 49823	TaqMan® Human MiRNA Array v3.0 A/B	768	26 (13/13)

is therefore urgent to discover more effective biomarkers to precisely diagnosis the subtypes of ACS.

MicroRNAs (miRNAs) are a class of small noncoding RNAs with the posttranscriptional role of regulating about 60% of human protein-coding genes [15]. Currently there are more than 2500 mature human miRNAs listed in miRBase (release 21) [16]. They play functions in a wide variety of biological processes such as cell proliferation [17, 18], development [19], and apoptosis [20], which contribute to various physiological and pathological conditions, including cardiovascular diseases such as the acute coronary syndrome [21, 22].

Until now, very few studies have looked at the two ACS subtypes, AMI and UA, in terms of similarities and differences, and in particular miRNA expression levels have not been well studied. To better understand the disease pathogenesis of these two subtypes, we applied an in-house regulatory model termed improved Pipeline of Outlier MicroRNA Analysis (POMA) [23, 24] to identify miRNAs specific to each subtype or shared by both subtypes (see Figure 1). The model focused on miRNAs' independent regulatory power and the biological functions of their targets. Two measures, novel out degree (NOD) and transcription factor percentage (TFP) of genes, were defined, where NOD was equivalent to the number of genes that were uniquely targeted by a single miRNA and TFP represented the percentage of all transcription factor (TF) genes that were targeted. According to the statistical evidences described in our previous work, miRNAs with larger NOD and TFP values were more likely to be candidate biomarkers and represented biomarker miRNAs that had strong abilities to regulate genes independently and, meanwhile, regulate more TF genes. The application of biomarker discovery for prostate cancer [23, 25], sepsis [26], clear cell renal cell carcinoma [27], and pediatric acute myeloid leukemia [24] demonstrated its great predictive power.

2. Materials and Methods

2.1. Dataset Collection. The miRNA expression datasets (GSE31568 and GSE49823) were downloaded from Gene Expression Omnibus (GEO) [28]. GSE31568 contained 454 samples and we extracted 70 controls and 21 AMI samples [29], and GSE49823 contained 13 controls and 13 UA samples. The details of the two datasets are listed in Table 1. We then

identified differentially expressed (DE) miRNAs based on linear models in Limma R package [30, 31]; the empirical Bayes (eBayes) method was performed to calculate the p value and other parameters. The Benjamini-Hochberg method was applied to adjust and correct p values. The adjusted p value <0.05 was chosen as the cut-off criteria.

We also collected the reported miRNAs for AMI and UA from PubMed by the search criteria "(Acute Myocardial Infarction OR AMI) AND (miRNA OR microRNA) AND (biomarker* OR marker*)" and "(Unstable Angina OR UA) AND (miRNA OR microRNA) AND (biomarker* OR marker*)". We only took published reports from the past five years and all of the samples were extracted with human data in consideration. The information of biomarkers including miRNA ID, biomarker type, expression pattern, study design, publication date, and PMID are summarized in Tables S1 and S2, in Supplementary Material available online at http://dx.doi.org/10.1155/2016/4618323.

2.2. Prediction of Putative miRNA Biomarkers for AMI and UA. Based on two significantly DE miRNA sets, we employed improved POMA to predict miRNA biomarkers for AMI and UA [24]. In the pipeline, two important measures NOD and TFP were defined. NOD is the number of genes uniquely targeted by a certain miRNA and TFP is the percentage of TF genes of all targets of the miRNA. The main idea of the improved POMA model is that miRNAs with larger NOD values and targeting more TF genes are more likely to be biomarkers. The POMA and improved POMA methodologies were elaborated in our previous studies [23, 24].

Using this pipeline, the AMI- and UA-specific miRNA-mRNA networks were constructed by mapping relevant DE miRNAs onto human miRNA-mRNA network (reference network). Then, NOD and TFP were measured for each miRNA in the condition-specific network of AMI and UA, respectively. Finally, miRNAs with significantly large NOD and TFP values (Wilcoxon signed-rank test, p value <0.05) were selected as candidate biomarkers.

We calculated the percentage of reported AMI/UA biomarker miRNAs in the whole predicted set and defined it as the prediction precision for evaluating the accuracy of our model.

2.3. Functional Enrichment Analysis of the Target Genes of Candidate miRNA Biomarkers. We performed functional enrichment analysis of the genes uniquely regulated by candidate biomarker miRNAs from the two condition-specific miRNA-mRNA networks by MetaCore™ software. The significantly enriched pathways and diseases ontologies were ranked by p value (<0.05), which was calculated by hypergeometric test. FDR adjustment was used for multiple test correction.

3. Results

3.1. Identification of Candidate miRNA Biomarkers for AMI and UA. Based on AMI and UA miRNA expression datasets, we identified 292 and 182 deregulated miRNAs in AMI and UA, respectively. Employing our in-house model improved POMA [24], and a total of 27 miRNAs for AMI and 26 miRNAs for UA were screened (see Figure 2(a), Wilcoxon signed-rank test, p value < 0.05). These miRNAs were predicted to be candidate biomarkers for the two subtypes of ACS by our model. The substructural characteristics of these biomarker miRNAs in the miRNA regulatory network, including the number of whole targets (termed N), NOD, and TFP values, are listed in Table 2.

As listed in Table 2, nine miRNA biomarkers were shared by both AMI and UA subtypes, indicating that these miRNAs (miR-126, miR-142-3p, miR-145, miR-204, miR-340*, miR-346, miR-34a, miR-93, and let-7g) could be universal biomarkers for both AMI and UA. The remaining 18 and 17 miRNAs could be putative biomarkers specific for AMI and UA, respectively.

3.2. Literature-Based Validation of Identified miRNA Biomarkers. We collected AMI- and UA-specific miRNA biomarkers by analysis of citations in PubMed, as shown in Figure 2(b). Altogether, 30 miRNAs have been reported to be biomarkers for AMI and 25 of them are diagnostic. Two miRNAs (miR-155 and miR-380*) [32] and a cluster of miR-16/27a/101/150 [33] were reported to be prognostic indicators. Two miRNAs (miR-208b and miR-133a) were reported to be valuable for both diagnosis and prognosis in AMI (see Table S1).

For UA, 15 miRNAs have been reported to be biomarkers, 13 of them were diagnostic, including a cluster of three miRNAs (miR-132/150/186) [34], and two were reported to be effective for both diagnosis and prognosis (miR-133a and miR-208b) [35] (see Table S2). We then compared literature reported miRNAs with ones we identified and found five that were the same in the AMI set (prediction precision: 18.5%): miR-155, miR-34a, miR-27a, miR-101, and miR-126 (see Figure 2(c)). Among them, miR-155 expression was increased approximately 4-fold in patients with a high-risk of cardiac death after discharge and could be a biomarker for cardiac death in post-AMI patients [32]; miR-34a was investigated for its role as a p53 responsive miRNA and confirmed as predicator for the risk of heart failure after AMI [36]. Elevated miR-27a expression was included in the panel of prognostic miRNAs for outcome after AMI; downregulation of miR-101 was also included in this panel. However, miR-101 was also reported to be upregulated in another study [33].

There were also five biomarker miRNAs (miR-106b, miR-25, miR-590-5p, miR-132, and miR-126) for UA found from our analysis and the reported list (see Figure 2(d), prediction precision: 19.2%). Among them, miR-106b, miR-25, and miR-590-5p were upregulated when compared with the control group [37]. The significantly elevated expression levels of the miR-106b/25 cluster and miR-21/590-5p family could be used as an indicator of coronary artery disease. A panel that consisted of miR-132, miR-150, and miR-186 showed the highest discriminatory power (AUC = 0.91) [34]. miR-126 was a unique biomarker that was found both in our analysis and in previous studies for both AMI and UA. However,

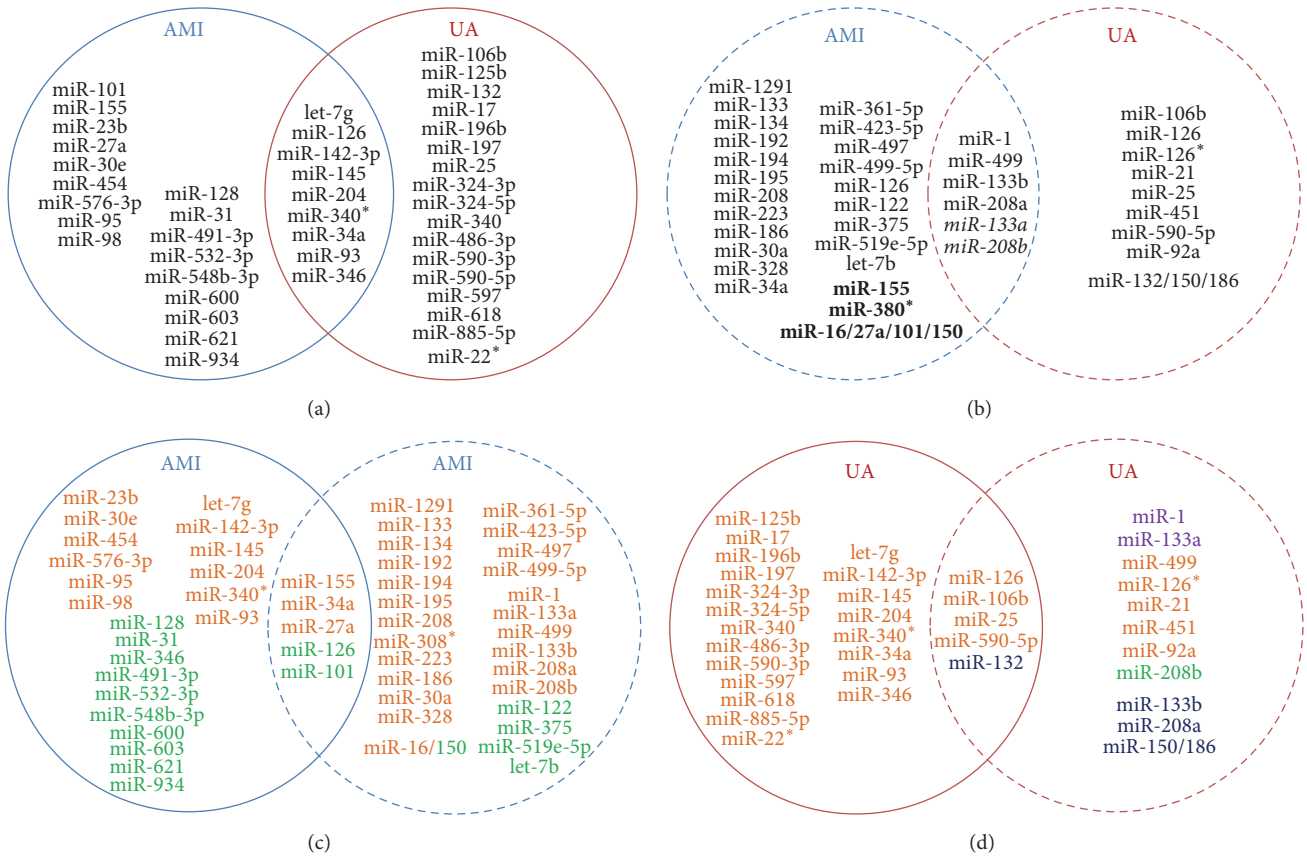


Figure 2: The Venn diagram of miRNA biomarkers for acute myocardial infarction (AMI) and unstable angina (UA). Circles in blue and red represent miRNAs for AMI and UA, respectively. Solid and dashed lines represent identified and literature reported miRNAs, respectively. (a) Candidate miRNA biomarkers identified by our model. (b) Biomarker miRNAs collected from published literature. IDs in bold mean they were prognostic, those in italic meant they were functional for both diagnosis and prognosis, and the remaining ones were reported to be diagnostic markers. (c) Comparison of AMI miRNA biomarkers identified by our model and published literature. IDs in orange and green represent up- and downregulated miRNAs, respectively. (d) Comparison of UA miRNA biomarkers identified by our model and published literature. IDs in orange and green represent up- and downregulated miRNAs, respectively. miRNAs that had both up- and downexpression patterns are colored in purple, and those with unclear expression patterns are colored in dark blue.

miR-126 was upregulated in the AMI dataset while it was reported to be downregulated in the literature [22]. In UA, the regulation pattern of the overlapping miRNAs was found to be consistent between our study and the previous reported work [38].

3.3. Functional Enrichment Analysis of Target Genes of Candidate miRNA Biomarkers. We further explored the roles of uniquely regulated genes of the identified miRNAs in AMI and UA by functional enrichment analysis using the MetaCore software [39–44]. In pathway analysis, we found 35 significantly enriched pathways in AMI and 18 in UA (p value < 0.05 and FDR < 0.05; see Figures 3(a) and 3(b)). There were nine pathways significantly enriched by the targets of candidate miRNA biomarkers for both AMI and UA (see Tables S3 and S4).

In general, the significantly enriched pathways were grouped into immune response, development, cell adhesion, signal transduction, apoptosis, and survival, and others as shown in Figures 3(c) and 3(d). In AMI, pathways in immune response (34%) and development (26%) account for 60% of the pathways. In UA, immune response and developmental pathways also play a role, with 11% and 17% of the miRNA-regulated pathways belonging to these categories, respectively. Besides these two, apoptosis and survival pathways accounted for a combined 22% of all miRNA targets.

We then evaluated the relevance of these pathways in AMI and UA by searching PubMed for published papers describing the role of constituent network objects of pathways in AMI and UA. As shown in Table S3, 28 of the 35 AMI pathways were reported to be involved with AMI and 10 of them are in the group of immune responses, such as CD40 signaling [45, 46]. Many interleukin (IL) factors were also reported in immune response pathways related to AMI such as IL-9 [47], IL-10 [48], IL-17 [49], IL-18 [50], and IL-33 [51]. In the development group, there were five pathways related to AMI, including WNT [52], G-CSF [53], SDF-1 [54], NF-kB [55], PEDF [56], and VEGF [57].

In the 18 pathways found in UA, 12 had been reported previously to relate to UA. The most important pathways were

Table 2: The identified miRNA biomarker candidates for acute myocardial infarction (AMI) and unstable angina (UA).

AMI					UA				
miRNA ID	*N*	NOD	TF (TFP)	Pathways (percentage)	miRNA ID	*N*	NOD	TF (TFP)	Pathways (percentage)
miR-155	185	64	39 (0.21)	16 (0.46)	miR-197	151	32	24 (0.16)	2 (0.11)
miR-30e	356	32	56 (0.16)	5 (0.14)	miR-125b	109	30	20 (0.18)	4 (0.22)
miR-98	329	24	62 (0.19)	2 (0.06)	miR-590-3p	255	18	44 (0.17)	2 (0.11)
miR-23b	211	18	34 (0.16)	5 (0.14)	miR-22*	158	16	32 (0.20)	0
miR-204	198	15	40 (0.20)	10 (0.29)	miR-204	198	15	40 (0.20)	2 (0.11)
miR-34a	80	14	15 (0.19)	5 (0.14)	miR-34a	80	14	15 (0.19)	6 (0.33)
let-7g	199	13	34 (0.17)	24 (0.69)	miR-486-3p	152	13	23 (0.15)	2 (0.11)
miR-576-3p	133	13	23 (0.17)	0	let-7g	199	13	34 (0.17)	13 (0.72)
miR-346	31	13	5 (0.16)	5 (0.14)	miR-346	31	13	5 (0.16)	4 (0.22)
miR-454	298	13	43 (0.14)	0	miR-340	256	11	37 (0.14)	1 (0.06)
miR-532-3p	112	12	18 (0.16)	5 (0.14)	miR-340*	256	11	37 (0.14)	1 (0.06)
miR-145	55	11	11 (0.20)	8 (0.23)	miR-196b	165	11	27 (0.16)	5 (0.28)
miR-340*	256	11	37 (0.14)	3 (0.09)	miR-145	55	11	11 (0.20)	3 (0.17)
miR-126	34	10	5 (0.15)	29 (0.83)	miR-324-3p	84	10	12 (0.14)	1 (0.06)
miR-621	65	10	13 (0.20)	9 (0.26)	**miR-126**	34	10	5 (0.15)	13 (0.72)
miR-142-3p	87	8	18 (0.21)	3 (0.09)	**miR-106b**	376	9	61 (0.16)	1 (0.06)
miR-31	34	7	8 (0.24)	5 (0.14)	miR-885-5p	89	9	14 (0.16)	0
miR-600	127	7	23 (0.18)	1 (0.03)	**miR-132**	46	8	7 (0.15)	0
miR-491-3p	119	6	21 (0.18)	0	miR-17	80	8	13 (0.16)	5 (0.28)
miR-603	149	6	32 (0.21)	2 (0.06)	miR-597	76	8	12 (0.16)	2 (0.11)
miR-93	394	6	68 (0.17)	0	miR-142-3p	87	8	18 (0.21)	1 (0.06)
miR-934	72	5	12 (0.17)	0	**miR-25**	260	7	40 (0.15)	0
miR-27a	50	5	15 (0.30)	3 (0.09)	**miR-590-5p**	112	7	23 (0.21)	0
miR-548b-3p	103	5	18 (0.17)	0	miR-324-5p	78	7	18 (0.23)	2 (0.11)
miR-101	69	4	18 (0.26)	4 (0.11)	miR-93	394	6	68 (0.17)	0
miR-128	22	4	4 (0.18)	0	miR-618	112	6	17 (0.15)	0
miR-95	69	4	10 (0.14)	0					

Notes. The miRNAs were ranked based on their NOD values. miRNA IDs in bold have been reported in published studies and those with underlines were shared by both AMI and UA.

involved with apoptosis and survival (FAS signaling cascades [58], TNFR1 signaling pathway [59], and NGF activation of NF-kB [60]). The other correlated pathways that had been previously reported were the PPAR pathway [61] and TCR and CD28 costimulation in activation of NF-kB [62] (see Table S4).

3.4. The Percentage of Pathways Potentially Regulated by Each Biomarker miRNA. Analyzing the biological processes for each subtype revealed mechanistic relationships. Some of the pathways may result in atherosclerosis progression and atherosclerotic lesion rupture. However, some may contribute to the development of coronary collateral vessels, and some may even have their roles in inhibiting the formation of the thrombus. In order to explore the role of miRNAs in the pathways, we calculated the percentage of pathways that were regulated by miRNA in all significantly enriched pathways (see Table 2).

In AMI, miR-126 (83%), let-7g (69%), and miR-155 (46%) were the top three miRNAs that regulated more than 30% of the significantly enriched pathways (as listed in Table 2). miR-126 (72%), let-7g (72%), and miR-34a (33%) were the top three miRNAs involved in UA. Both in AMI and in UA, let-7g and miR-126 regulated more than half of the pathways, which indicated that they were functionally important to both of the subtypes. This observation is helpful for understanding the molecular mechanisms of ACS common or specific to the AMI and UA subtypes.

4. Discussion

Understanding the mechanism and identifying the biomarkers specific to AMI and UA are important for ACS diagnosis and treatments. In this study, miRNA biomarkers are identified for AMI and UA using our improved POMA model. The model enriches fragile sites in the miRNA-mRNA network, focusing on miRNAs that regulate important genes

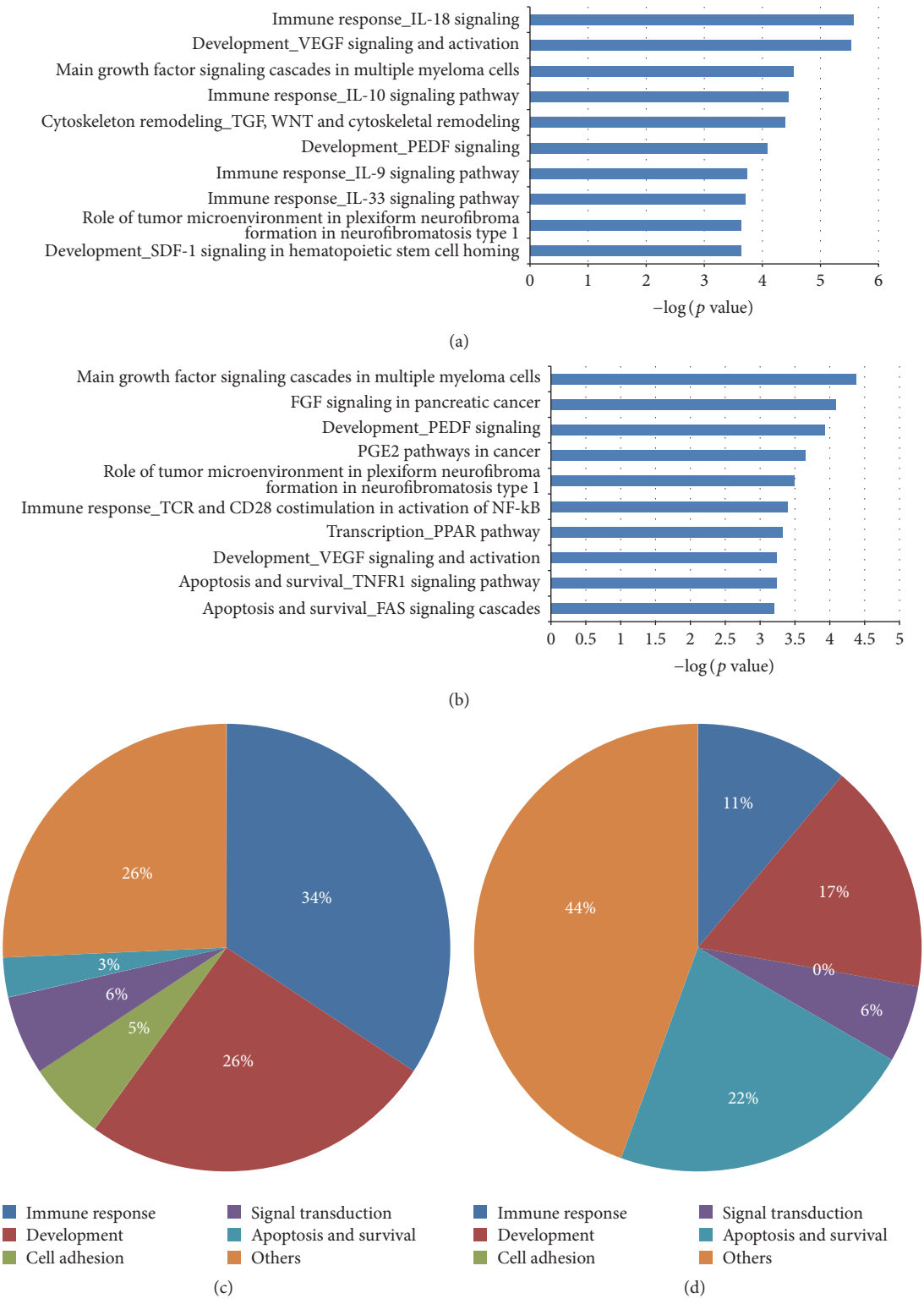


Figure 3: Significantly enriched pathways by targets of microRNA biomarkers for acute myocardial infarction (AMI) and unstable angina (UA). (a) Top 10 enriched pathways in AMI. (b) Top 10 enriched pathways in UA. (c) Pie plot of category enriched pathways in AMI. (d) Pie plot of category enriched pathways in UA.

for these disease subtypes. We defined measures NOD and TFP to quantify whether a miRNA could be a candidate biomarker. The former reflected the power of a miRNA to regulate genes independently whereas the latter indicated the potential of regulating TF genes. TF genes are chosen because TFs are important regulators and are kernels of many crucial biological processes. According to our previous studies [23, 24], biomarker miRNAs tended to have large NOD and TFP values. Based on the evidence, we identified 27 and 26 miRNAs as candidate biomarkers for AMI and UA, respectively, and nine of them were shared by the two ACS subtypes. Five AMI and five UA candidates had been previously reported as miRNA biomarkers.

In order to explore the roles of miRNAs in ACS, we performed a functional enrichment analysis for the targets of candidate miRNA biomarkers. In AMI, 35 pathways were significantly enriched and 28 (80%) have been reported to be related to AMI. Many of the pathways enriched in AMI were correlated with the immune response, and let-7g, miR-155, miR-101, miR-126, and miR-145 were closely relevant (see Table S3). A deregulated immune system is considered not only a trigger but also a factor amplifying an uncontrolled immune response in AMI [45]. CD40 signaling was reported to be upregulated in the pathogenesis of AMI patients [46]. The levels of IL-18 were upregulated in patients with AMI and the inhibition of its activity promoted cardiac function and reduced scar formation and infarct size [50, 63]. The predictive values of IL-6 and IL-10 were also shown for ST-elevation AMI [64]. IL-9 levels were significantly upregulated in patients with AMI compared with the stable angina pectoris and control groups [65]. IL-33/ST2 signaling is a mechanically activated, cardioprotective signaling system where IL-33 blocks angiotensin II- and phenylephrine-induced NF-κB activation, and soluble ST2 inhibits the antihypertrophic effects of IL-33 [66]. The ratio of IL-33/sST2 also correlated with the 6-month prognosis of AMI patients [67]. Damaged myocardial tissue is repaired and replaced by scar tissue after MI, which triggers an inflammatory cascade. Clinical studies have indicated that an excessive inflammatory reaction may evoke adverse remodeling and directly affect prognosis in patients with AMI. It has been suggested that elevated concentrations of circulating neutrophils and monocytes and enhanced extracellular matrix breakdown [68] may contribute to infarct expansion or even cardiac rupture [69]. Accumulating evidence also showed that uncontrolled immune response in AMI may result from a pleiotropic proinflammatory imbalance [70]. Accordingly, exploring the DE miRNAs and target genes within these immune cells may be promising for cell base therapies.

In UA, 12 of 18 (66%) significantly enriched pathways were reported previously and 22% of the pathways were grouped in the apoptosis and survival category. Apoptosis is programmed cell death or physiological death. The abnormalities of apoptosis may contribute to plaque rupture and ACS. Endothelial cell apoptosis differs from macrophage/monocyte apoptosis [71]; endothelial cell apoptosis results in atherosclerosis progression and atherosclerotic lesion rupture [72]. The increased expression of Fas and FasL (both in AMI and in UA) was observed on the surface of peripheral blood lymphocytes [73, 74]. Cellular apoptosis may be one of the factors involved in atherosclerosis and may play a role in the rupture of atherosclerotic plaques. Thus, we not only should get a better understanding of the whole process of programmed cell death but also need to know the contribution of antiapoptotic therapy to plaque stabilization. More importantly, miRNAs like miR-346, miR-196b, miR-126, and let-7g were functional in these pathways according to our study (see Table S4), which indicates their potential roles in the process of cell apoptosis as well as the occurrence and progression of UA.

Moreover, there were nine pathways that were enriched by targets of miRNA biomarkers in both AMI and UA. Two developmental signaling pathways, vascular endothelial growth factor (VEGF) and pigment epithelium-derived factor (PEDF), were reported to be important factors in both subtypes. VEGF, a peripheral blood cytokine, is mainly derived from platelets and granulocytes and in particular neutrophils, which play a crucial role in vascular formation in physiological and pathological conditions [75–77]. It has been reported that, in ischemic conditions, VEGF promotes the development of coronary collateral vessels, providing adequate blood supply and preventing death of cardiomyocytes [78]. Many studies have found that serum VEGF concentrations were elevated in ACS, which can be a surrogate marker of myocardial infarction [79, 80]. Serum VEGF-A was shown to be elevated after AMI, which suggested a role for the formation of coronary collateral vessels [57, 81]. VEGF-A is also a target gene of miR-126, which mapped in the pathway.

The other common developmental pathway in AMI and UA was PEDF signaling. PEDF, a 50-kDa glycoprotein, has anti-inflammatory, antioxidant, antiangiogenic, antithrombotic, antitumorigenic, and neuroprotective properties [82] and is widely expressed throughout the human body. In ACS patients, plasma PEDF concentrations were significantly lower than the control group and associated with adverse cardiac outcomes after ACS [83]. PEDF can block platelet activation and aggregation [84] through its anti-inflammatory and antioxidative properties, leading to the inhibition of the vascular inflammation and the formation of a thrombus [85].

We also calculated the percentage of pathways that were potentially regulated by the miRNAs in all significantly enriched pathways (see Tables 2, S3, and S4). A novel miRNA (let-7g), which had never been reported as an important factor in ACS before, deserves further investigation, as it participated in regulating 69% (24/35) of the pathways in AMI and 72% (13/18) in UA. About half of these enriched pathways were closely associated with the immune response, especially in AMI (see Tables S3 and S4). Notably, it was previously reported that a miRNA together with its targets was differentially regulated in E2F1-deficient mice, and the E2F1 transcription factor played important roles in the immune response to systemic *Escherichia coli* lipopolysaccharide (LPS) [86]. The conclusions demonstrated the significance of let-7g in the immune response, which may represent a latent therapeutic target for the treatment of immunological diseases as well as ACS. More clinical validations of this hypothesis will be needed in the future.

We noticed that the miRNA datasets selected for our comparison in this study were inconsistent. The AMI dataset was obtained on whole blood (GSE31568) whereas the UA was on plasma (GSE49823). As we known, the concentration of miRNAs in whole blood is higher than that in plasma; thus the prediction based on the two data sources has some limitations. It would be better if we could obtain AMI and UA samples chosen from the same source type (both were from plasma or whole blood) and compare with the same control group. Unfortunately, the miRNA expression data that could be used for analyses were quite limited. On the other hand, we considered that plasma is an important component of whole blood, where miRNAs could present in a remarkably stable form [87]. Hence further expression data analyses and clinical validations need to be done when more and better datasets are available for in-depth studies.

5. Conclusions

In this study, we applied our improved POMA model to identify miRNA biomarkers for subtyping ACS, finding 18 and 17 miRNAs to be specific biomarkers for AMI and UA, respectively. Nine miRNAs were found in both subtypes, which implied that they could be universal molecular markers for ACS. These findings were further verified by enrichment analysis and compared with previous publications. For future translational application, further experimental and clinical verifications are necessary.

Abbreviations

DE: Differentially expressed
TF: Transcription factor
POMA: Pipeline of Outlier MicroRNA Analysis
NOD: Novel out degree
TFP: Transcription factor gene percentage.

Competing Interests

The authors declare that there is no conflict of interests.

Authors' Contributions

Yujie Zhu and Yuxin Lin contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant nos. 31470821, 31400688, 81471488, and 81271378) and the Natural Science Foundation of Jiangsu Province, China (Grant no. BK20130290).

References

[1] R. A. Nishimura, C. M. Otto, R. O. Bonow et al., "2014 AHA/ACC guideline for the management of patients with valvular heart disease: executive summary: a report of the American college of cardiology/American heart association task force on practice guidelines," *Journal of the American College of Cardiology*, vol. 63, no. 22, pp. 2438–2488, 2014.

[2] D. Mozaffarian, E. J. Benjamin, A. S. Go et al., "Heart disease and stroke statistics—2015 update: a report from the American Heart Association," *Circulation*, vol. 131, no. 4, pp. e29–e322, 2015.

[3] 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.

[4] D. M. Kolansky, "Acute coronary syndromes: morbidity, mortality, and pharmacoeconomic burden," *American Journal of Managed Care*, vol. 15, no. 2, pp. S36–S41, 2009.

[5] M. Torres and S. Moayedi, "Evaluation of the acutely dyspneic elderly patient," *Clinics in Geriatric Medicine*, vol. 23, no. 2, pp. 307–325, 2007.

[6] V. Čulić, D. Eterović, D. Mirić, and N. Silić, "Symptom presentation of acute myocardial infarction: influence of sex, age, and risk factors," *American Heart Journal*, vol. 144, no. 6, pp. 1012–1017, 2002.

[7] E. M. Antman, M. Cohen, P. J. L. M. Bernink et al., "The TIMI risk score for unstable angina/non-ST elevation MI: a method for prognostication and therapeutic decision making," *Journal of the American Medical Association*, vol. 284, no. 7, pp. 835–842, 2000.

[8] B. Lee, A. M. Chang, A. C. Matsuura, S. Marcoon, and J. E. Hollander, "Comparison of cardiac risk scores in ED patients with potential acute coronary syndrome," *Critical Pathways in Cardiology*, vol. 10, no. 2, pp. 64–68, 2011.

[9] J. Sanchis, V. Bodí, J. Núñez et al., "New risk score for patients with acute chest pain, non-ST-segment deviation, and normal troponin concentrations: a comparison with the TIMI risk score," *Journal of the American College of Cardiology*, vol. 46, no. 3, pp. 443–449, 2005.

[10] D. K. Slater, M. A. Hlatky, D. B. Mark, F. E. Harrell Jr., D. B. Pryor, and R. M. Califf, "Outcome in suspected acute myocardial infarction with normal or minimally abnormal admission electrocardiographic findings," *The American Journal of Cardiology*, vol. 60, no. 10, pp. 766–770, 1987.

[11] J. A. Goldstein, K. M. Chinnaiyan, A. Abidov et al., "The CT-STAT (coronary computed tomographic angiography for systematic triage of acute chest pain patients to treatment) trial," *Journal of the American College of Cardiology*, vol. 58, no. 14, pp. 1414–1422, 2011.

[12] W. B. Gibler, L. M. Lewis, R. E. Erb et al., "Early detection of acute myocardial infarction in patients presenting with chest pain and nondiagnostic ECGs: serial CK-MB sampling in the emergency department," *Annals of Emergency Medicine*, vol. 19, no. 12, pp. 1359–1366, 1990.

[13] K. Thygesen, J. S. Alpert, and H. D. White, "Universal definition of myocardial infarction," *Journal of the American College of Cardiology*, vol. 50, no. 22, pp. 2173–2195, 2007.

[14] P. Valensi, L. Lorgis, and Y. Cottin, "Prevalence, incidence, predictive factors and prognosis of silent myocardial infarction: a review of the literature," *Archives of Cardiovascular Diseases*, vol. 104, no. 3, pp. 178–188, 2011.

[15] M. Esteller, "Non-coding RNAs in human disease," *Nature Reviews Genetics*, vol. 12, no. 12, pp. 861–874, 2011.

[16] A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data," *Nucleic Acids Research*, vol. 42, no. 1, pp. D68–D73, 2014.

[17] A. A. Dar, S. Majid, D. de Semir, M. Nosrati, V. Bezrookove, and M. Kashani-Sabet, "miRNA-205 suppresses melanoma cell proliferation and induces senescence via regulation of E2F1 protein," *The Journal of Biological Chemistry*, vol. 286, no. 19, pp. 16606–16614, 2011.

[18] X.-M. Zhao, K.-Q. Liu, G. Zhu et al., "Identifying cancer-related microRNAs based on gene expression data," *Bioinformatics*, vol. 31, no. 8, pp. 1226–1234, 2015.

[19] A. E. Kulozik, "Stay Tuned: miRNA Expression and Nonsense-Mediated Decay in Brain Development," *Molecular Cell*, vol. 42, no. 4, pp. 407–408, 2011.

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

[21] E. M. Small and E. N. Olson, "Pervasive roles of microRNAs in cardiovascular biology," *Nature*, vol. 469, no. 7330, pp. 336–342, 2011.

[22] G. Long, F. Wang, Q. Duan et al., "Human circulating microRNA-1 and microRNA-126 as potential novel indicators for acute myocardial infarction," *International Journal of Biological Sciences*, vol. 8, no. 6, pp. 811–818, 2012.

[23] W. Zhang, J. Zang, X. Jing et al., "Identification of candidate miRNA biomarkers from miRNA regulatory network with application to prostate cancer," *Journal of Translational Medicine*, vol. 12, article 66, 2014.

[24] W. Yan, L. Xu, Z. Sun et al., "MicroRNA biomarker identification for pediatric acute myeloid leukemia based on a novel bioinformatics model," *Oncotarget*, vol. 6, no. 28, pp. 26424–26436, 2015.

[25] J. Zhu, S. Wang, W. Zhang et al., "Screening key microRNAs for castration-resistant prostate cancer based on miRNA/mRNA functional synergistic network," *Oncotarget*, vol. 6, no. 41, pp. 43819–43830, 2015.

[26] J. Huang, Z. Sun, W. Yan et al., "Identification of microRNA as sepsis biomarker based on miRNAs regulatory network analysis," *BioMed Research International*, vol. 2014, Article ID 594350, 12 pages, 2014.

[27] J. Chen, D. Zhang, W. Zhang et al., "Clear cell renal cell carcinoma associated microRNA expression signatures identified by an integrated bioinformatics analysis," *Journal of Translational Medicine*, vol. 11, no. 1, article 169, 2013.

[28] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[29] A. Keller, P. Leidinger, A. Bauer et al., "Toward the blood-borne miRNome of human diseases," *Nature Methods*, vol. 8, no. 10, pp. 841–843, 2011.

[30] M. E. Ritchie, B. Phipson, D. Wu et al., "Limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, p. e47, 2015.

[31] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, article 3, 2004.

[32] S. Matsumoto, Y. Sakata, D. Nakatani et al., "A subset of circulating microRNAs are predictive for cardiac death after discharge for acute myocardial infarction," *Biochemical and Biophysical Research Communications*, vol. 427, no. 2, pp. 280–284, 2012.

[33] Y. Devaux, M. Vausort, G. P. McCann et al., "A panel of 4 microRNAs facilitates the prediction of left ventricular contractility after acute myocardial infarction," *PLoS ONE*, vol. 8, no. 8, Article ID e70644, 2013.

[34] T. Zeller, T. Keller, F. Ojeda et al., "Assessment of microRNAs in patients with unstable angina pectoris," *European Heart Journal*, vol. 35, no. 31, pp. 2106–2114, 2014.

[35] C. Widera, S. K. Gupta, J. M. Lorenzen et al., "Diagnostic and prognostic impact of six circulating microRNAs in acute coronary syndrome," *Journal of Molecular and Cellular Cardiology*, vol. 51, no. 5, pp. 872–875, 2011.

[36] S. Matsumoto, Y. Sakata, S. Suna et al., "Circulating p53-responsive MicroRNAs are predictive indicators of heart failure after acute myocardial infarction," *Circulation Research*, vol. 113, no. 3, pp. 322–326, 2013.

[37] J. Ren, J. Zhang, N. Xu et al., "Signature of circulating microRNAs as potential biomarkers in vulnerable coronary artery disease," *PLoS ONE*, vol. 8, no. 12, Article ID e80738, 2013.

[38] Y. D'Alessandra, M. C. Carena, L. Spazzafumo et al., "Diagnostic potential of plasmatic microRNA signatures in stable and unstable angina," *PLoS ONE*, vol. 8, no. 11, Article ID e80345, 2013.

[39] Y. Tang, W. Yan, J. Chen, C. Luo, A. Kaipia, and B. Shen, "Identification of novel microRNA regulatory pathways associated with heterogeneous prostate cancer," *BMC Systems Biology*, vol. 7, supplement 3, article S6, 2013.

[40] Y. Hu, J. Li, W. Yan et al., "Identifying novel glioma associated pathways based on systems biology level meta-analysis," *BMC systems biology*, vol. 7, 2, p. S9, 2013.

[41] M. Ding, H. Wang, J. Chen, B. Shen, and Z. Xu, "Identification and functional annotation of genome-wide ER-regulated genes in breast cancer based on ChIP-Seq data," *Computational and Mathematical Methods in Medicine*, vol. 2012, Article ID 568950, 2012.

[42] Y. Wang, J. Chen, Q. Li et al., "Identifying novel prostate cancer associated pathways based on integrative microarray data analysis," *Computational Biology and Chemistry*, vol. 35, no. 3, pp. 151–158, 2011.

[43] B. Liu, J. Chen, and B. Shen, "Genome-wide analysis of the transcription factor binding preference of human bi-directional promoters and functional annotation of related gene pairs," *BMC Systems Biology*, vol. 5, supplement 1, article S2, 2011.

[44] G. Liu, M. Ding, J. Chen et al., "Computational analysis of microRNA function in heart development," *Acta Biochimica et Biophysica Sinica*, vol. 42, no. 9, pp. 662–670, 2010.

[45] V. Bodi, J. Sanchis, J. Nunez et al., "Uncontrolled immune response in acute myocardial infarction. Unraveling the thread," *American Heart Journal*, vol. 156, no. 6, pp. 1065–1073, 2008.

[46] J.-C. Yan, Z.-G. Wu, X.-T. Kong, R.-Q. Zong, and L.-Z. Zhan, "Relation between upregulation of CD40 system and complex stenosis morphology in patients with acute coronary syndrome," *Acta Pharmacologica Sinica*, vol. 25, no. 2, pp. 251–256, 2004.

[47] I. Gregersen, M. Skjelland, S. Holm et al., "Increased systemic and local interleukin 9 levels in patients with carotid and coronary atherosclerosis," *PLoS ONE*, vol. 8, no. 8, Article ID e72769, 2013.

[48] B. Goswami, M. Rajappa, V. Mallika, D. K. Shukla, and S. Kumar, "TNF-α/IL-10 ratio and C-reactive protein as markers of the inflammatory response in CAD-prone North Indian patients with acute myocardial infarction," *Clinica Chimica Acta*, vol. 408, no. 1-2, pp. 14–18, 2009.

[49] X. Cheng, X. Yu, Y.-J. Ding et al., "The Th17/Treg imbalance in patients with acute coronary syndrome," *Clinical Immunology*, vol. 127, no. 1, pp. 89–97, 2008.

[50] D. Kawasaki, T. Tsujino, S. Morimoto et al., "Plasma interleukin-18 concentration: a novel marker of myocardial ischemia rather than necrosis in humans," *Coronary Artery Disease*, vol. 16, no. 7, pp. 437–441, 2005.

[51] A. M. Miller, "Role of IL-33 in inflammation and disease," *Journal of Inflammation*, vol. 8, article 22, 2011.

[52] B. Assmus, M. Iwasaki, V. Schächinger et al., "Acute myocardial infarction activates progenitor cells and increases Wnt signalling in the bone marrow," *European Heart Journal*, vol. 33, no. 15, pp. 1911–1919, 2012.

[53] S. Vandervelde, M. J. A. van Luyn, R. A. Tio, and M. C. Harmsen, "Signaling factors in stem cell-mediated repair of infarcted myocardium," *Journal of Molecular and Cellular Cardiology*, vol. 39, no. 2, pp. 363–376, 2005.

[54] K. Stellos, B. Bigalke, H. Langer et al., "Expression of stromal-cell-derived factor-1 on circulating platelets is increased in patients with acute coronary syndrome and correlates with the number of $CD34^+$ progenitor cells," *European Heart Journal*, vol. 30, no. 5, pp. 584–593, 2009.

[55] X. X. Liao, X. Li, Z. F. Ma et al., "Role of nuclear factor-κB in endothelial injury in acute myocardial infarction," *Zhongguo Wei Zhong Bing Ji Jiu Yi Xue*, vol. 20, no. 7, pp. 413–415, 2008.

[56] K. Distelmaier, C. Adlbrecht, J. Jakowitsch et al., "Proteomic profiling of acute coronary thrombosis reveals a local decrease in pigment epithelium-derived factor in acute myocardial infarction," *Clinical Science*, vol. 123, no. 2, pp. 111–119, 2012.

[57] A. Kranz, C. Rau, M. Kochs, and J. Waltenberger, "Elevation of vascular endothelial growth factor a serum levels following acute myocardial infarction. Evidence for its origin and functional significance," *Journal of Molecular and Cellular Cardiology*, vol. 32, no. 1, pp. 65–72, 2000.

[58] A. Bossowska, A. Bossowski, and B. Galar, "Analysis of apoptotic markers Fas/FasL (CD95/CD95L) expression on the lymphocytes in patients with acute coronary syndrome," *Kardiologia Polska*, vol. 65, no. 8, pp. 883–889, 2007.

[59] P. Aukrust, W. J. Sandberg, K. Otterdal et al., "Tumor necrosis factor superfamily molecules in acute coronary syndromes," *Annals of Medicine*, vol. 43, no. 2, pp. 90–103, 2011.

[60] M. E. Ritchie, "Nuclear factor-κB is selectively and markedly activated in humans with unstable angina pectoris," *Circulation*, vol. 98, no. 17, pp. 1707–1713, 1998.

[61] J. Yang, C. Liu, L. Zhang et al., "Intensive atorvastatin therapy attenuates the inflammatory responses in monocytes of patients with unstable angina undergoing percutaneous coronary intervention via peroxisome proliferator-activated receptor γ activation," *Inflammation*, vol. 38, no. 4, pp. 1415–1423, 2015.

[62] L. Cominacini, M. Anselmi, U. Garbin et al., "Enhanced plasma levels of oxidized low-density lipoprotein increase circulating nuclear factor-kappa B activation in patients with unstable angina," *Journal of the American College of Cardiology*, vol. 46, no. 5, pp. 799–806, 2005.

[63] Z. Mallat, C. Heymes, A. Corbaz et al., "Evidence for altered interleukin 18 (IL)-18 pathway in human heart failure," *The FASEB Journal*, vol. 18, no. 14, pp. 1752–1754, 2004.

[64] E. Ammirati, C. V. Cannistraci, N. A. Cristell et al., "Identification and predictive value of interleukin-6^+ interleukin-10^+ and interleukin-6^- interleukin-10^+ cytokine patterns in ST-elevation acute myocardial infarction," *Circulation Research*, vol. 111, no. 10, pp. 1336–1348, 2012.

[65] Y.-Z. Lin, B.-W. Wu, Z.-D. Lu et al., "Circulating Th22 and Th9 levels in patients with acute coronary syndrome," *Mediators of Inflammation*, vol. 2013, Article ID 635672, 2013.

[66] S. Sanada, D. Hakuno, L. J. Higgins, E. R. Schreiter, A. N. J. McKenzie, and R. T. Lee, "IL-33 and ST2 comprise a critical biomechanically induced and cardioprotective signaling system," *Journal of Clinical Investigation*, vol. 117, no. 6, pp. 1538–1549, 2007.

[67] K. Zhang, X.-C. Zhang, Y.-H. Mi, and J. Liu, "Predicting value of serum soluble ST2 and interleukin-33 for risk stratification and prognosis in patients with acute myocardial infarction," *Chinese Medical Journal*, vol. 126, no. 19, pp. 3628–3631, 2013.

[68] Y. Maekawa, T. Anzai, T. Yoshikawa et al., "Prognostic significance of peripheral monocytosis after reperfused acute myocardial infarction: a possible role for left ventricular remodeling," *Journal of the American College of Cardiology*, vol. 39, no. 2, pp. 241–246, 2002.

[69] T. Takahashi, Y. Hiasa, Y. Ohara et al., "Relationship of admission neutrophil count to microvascular injury, left ventricular dilation, and long-term outcome in patients treated with primary angioplasty for acute myocardial infarction," *Circulation Journal*, vol. 72, no. 6, pp. 867–872, 2008.

[70] X. Cheng, Y.-H. Liao, H. Ge et al., "TH1/TH2 functional imbalance after acute myocardial infarction: coronary arterial inflammation or myocardial inflammation," *Journal of Clinical Immunology*, vol. 25, no. 3, pp. 246–253, 2005.

[71] V. E. A. Stoneman and M. R. Bennett, "Role of apoptosis in atherosclerosis and its therapeutic implications," *Clinical Science*, vol. 107, no. 4, pp. 343–354, 2004.

[72] M. M. Kockx and A. G. Herman, "Apoptosis in atherogenesis: implications for plaque destabilization," *European Heart Journal*, vol. 19, pp. G23–G28, 1998.

[73] Y. Li, G. Takemura, K.-I. Kosai et al., "Critical roles for the Fas/Fas ligand system in postinfarction ventricular remodeling and heart failure," *Circulation Research*, vol. 95, no. 6, pp. 627–636, 2004.

[74] M. Shimizu, K. Fukuo, S. Nagata et al., "Increased plasma levels of the soluble form of Fas ligand in patients with acute myocardial infarction and unstable angina pectoris," *Journal of the American College of Cardiology*, vol. 39, no. 4, pp. 585–590, 2002.

[75] N. Ferrara, "Vascular endothelial growth factor," *European Journal of Cancer Part A*, vol. 32, no. 14, pp. 2413–2422, 1996.

[76] K. Harada, M. Friedman, J. J. Lopez et al., "Vascular endothelial growth factor administration in chronic myocardial ischemia," *American Journal of Physiology—Heart and Circulatory Physiology*, vol. 270, no. 5, pp. H1791–H1802, 1996.

[77] L. F. Brown, K.-T. Yeo, B. Berse et al., "Expression of vascular permeability factor (vascular endothelial growth factor) by epidermal keratinocytes during wound healing," *Journal of Experimental Medicine*, vol. 176, no. 5, pp. 1375–1379, 1992.

[78] T. Sugimoto, K. Inui, and Y. Shimazaki, "Gene therapy for myocardial angiogenesis: with direct intramuscular gene transfer of naked deoxyribonucleic acid encoding vascular endothelial growth factor and cell transplantation of vascular endothelial growth factor transfected H9c2 myoblast," *Japanese Journal of Thoracic and Cardiovascular Surgery*, vol. 51, no. 5, pp. 192–197, 2003.

[79] A. Konopka, J. Janas, W. Piotrowski, and J. Stepińska, "Concentration of vascular endothelial growth factor in patients with acute coronary syndrome," *Cytokine*, vol. 61, no. 2, pp. 664–669, 2013.

[80] C. Gui, S.-K. Li, Q.-L. Nong, F. Du, L.-G. Zhu, and Z.-Y. Zeng, "Changes of serum angiogenic factors concentrations in

patients with diabetes and unstable angina pectoris," *Cardiovascular Diabetology*, vol. 12, no. 1, article 34, 2013.

[81] Y. Wang, H. E. Johnsen, S. Mortensen et al., "Changes in circulating mesenchymal stem cells, stem cell homing factor, and vascular growth factors in patients with acute ST elevation myocardial infarction treated with primary percutaneous coronary intervention," *Heart*, vol. 92, no. 6, pp. 768–774, 2006.

[82] D. Orlic, J. Kajstura, S. Chimenti et al., "Bone marrow cells regenerate infarcted myocardium," *Nature*, vol. 410, no. 6829, pp. 701–705, 2001.

[83] J. Liu, S. Wang, J. Shi et al., "The association study of plasma levels of pigment epithelium-derived factor with acute coronary syndrome in the Chinese Han population," *Cardiology*, vol. 127, no. 1, pp. 31–37, 2014.

[84] K. Takenaka, S.-I. Yamagishi, T. Matsui et al., "Pigment epithelium-derived factor (PEDF) administration inhibits occlusive thrombus formation in rats: a possible participation of reduced intraplatelet PEDF in thrombosis of acute coronary syndromes," *Atherosclerosis*, vol. 197, no. 1, pp. 25–33, 2008.

[85] S.-I. Ueda, S.-I. Yamagishi, T. Matsui, Y. Jinnouchi, and T. Imaizumi, "Administration of pigment epithelium-derived factor inhibits left ventricular remodeling and improves cardiac function in rats with acute myocardial infarction," *American Journal of Pathology*, vol. 178, no. 2, pp. 591–598, 2011.

[86] L. A. Warg, J. L. Oakes, R. Burton et al., "The role of the E2F1 transcription factor in the innate immune response to systemic LPS," *American Journal of Physiology—Lung Cellular and Molecular Physiology*, vol. 303, no. 5, pp. L391–L400, 2012.

[87] P. S. Mitchell, R. K. Parkin, E. M. Kroh et al., "Circulating microRNAs as stable blood-based markers for cancer detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 30, pp. 10513–10518, 2008.