IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED INFORMATION SENSING AND LEARNING TECHNOLOGIES FOR DATA-CENTRIC SMART HEALTH APPLICATIONS

Received April 10, 2019, accepted May 14, 2019, date of publication May 22, 2019, date of current version June 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918150

gwSPIA: Improved Signaling Pathway Impact Analysis With Gene Weights

ZHENSHEN BAO[1], YIHUA ZHU[2], QINYU GE[1], WANJUN GU[1], XIANJUN DONG[3,4], AND YUNFEI BAI[1]

[1]State Key Laboratory of Bioelectronics, School of Biological Sciences and Medical Engineering, Southeast University, Nanjing 210096, China
[2]College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China
[3]Neurogenomics Laboratory and Precision Neurology Program, Brigham and Women's Hospital, Boston, MA 02115, USA
[4]Department of Neurology, Harvard Medical School, Boston, MA 02115, USA

Corresponding authors: Xianjun Dong (xdong@rics.bwh.harvard.edu) and Yunfei Bai (whitecf@seu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61871121, Grant 61271055, and Grant 61471112.

ABSTRACT Gene set analysis using signaling pathway has become a popular downstream analysis following differential expression analysis. From a biological point of view, only some portions of a pathway are expected to be altered; however, a few approaches using the different importance of genes in signaling pathways, which encompass the constitutive functional nonequivalent roles of genes in real pathways, have been proposed and none of them tries to associate the importance of genes with the related disease. In this paper, we developed an extended method of signaling pathway impact analysis (SPIA), called gwSPIA, by incorporating three signaling pathway-based gene weight merits that reflect the importance of genes from different aspects and attempt to associate the importance of genes with the related diseases. By applying the gwSPIA to the gene expression data sets in comparison with other seven methods in three measures, sensitivity, prioritization, and specificity, we show that the gwSPIA ranks in the second place in both sensitivity and prioritization. Furthermore, the specificity of the gwSPIA is better than SPIA, which is lower than 25%. The results also suggest that the gene weight used in the gwSPIA can reflect the association between the genes and the related diseases. The R package of the gwSPIA can be accessed from https://github.com/sterding/gwSPIA.

INDEX TERMS Differentially expressed genes, gene weights, gwSPIA, signaling pathways analysis.

I. INTRODUCTION

As the rapid development of high-throughput sequencing technology in recent years, more and more differentially expressed genes (DEGs) studies have been proposed to reveal the perturbed signaling pathways across different disease conditions, drug treatments, or developmental stages. The analysis that combines DEGs with signaling pathways has become a dominant analytical method. In such an analysis, significant signaling pathways based on DEGs are identified using statistical methods, allowing researchers and clinicians to better understand interactions between diseases and genes. Common signaling pathway databases include Kyoto Encyclopedia of Genes and Genomes (KEGG) [1]–[3], BioCarta [4], [5], and Reactome [6], [7]. The original signaling pathway analysis methods can be divided into two categories: pathway-topology based and non-pathway-topology based approaches.

The associate editor coordinating the review of this manuscript and approving it for publication was Qingxue Zhang.

Most classic signaling pathway analysis methods in the non-pathway-topology based category are designed based on either over-representation analysis (ORA) or functional class scoring (FCS). ORA-based methods, including OntoExpress [8], [9] and Gene Ontology Enrichment Analysis Software Toolkit (GOEASE) [10], merely measure the number of differential expressed genes in a specific signaling pathway and determine the significance of overlapping via statistical tests like Fisher's exact test. FCS-based methods however take into account of coordinated changes of genes expression in the specific signaling pathways, such as gene set enrichment analysis (GSEA) [11]. All these methods have common limitations that genes in a signaling pathway are treated equally, and they are without considering the complex interactions between genes.

On the other hand, pathway-topology based approaches consider the complex interaction between genes through incorporating pathway topology information, specifically the KEGG signaling pathways. SPIA is a classic pathway-topology based approach, which combines the features

of ORA or FCS method and the perturbation on a given signaling pathway [12]. Later, Li et al. improved the SPIA method by using a subgraph method to increase the accuracy of SPIA [13]. Bao et al. recently improved the SPIA by substituting $+1$ or -1 as the strength of interaction of genes with Pearson correlation coefficients and mutual information to increase the accuracy of SPIA [14]. Gene Graph Enrichment Analysis (GGEA) is another method to detect gene sets enriched consistently and coherently, based on prior knowledge derived from directed gene regulatory networks [15]. Ihnatova et al. developed a novel R package that offers seven distinct methods for topology-based pathway analysis [16]. Liu et al. proposed a topological method to find sub signaling pathways in a signaling pathway to improve the performance of pathway analysis method [17]. However, all these pathway topology-based methods still have the limitation of treating the functional roles of genes in pathways equivalently. Because genes in pathways can function inequivalently, they could have different importance in signaling pathways in different disease.

Apparently, some genes are more important than others are, if they are in a hub position in the pathway (or network) topology or have a key functional role in different disease process biology. There are existing methods weighing the importance of genes. And these methods are called gene-weight-based methods. For example, EnrichNet measures the functional association between the gene list of interest and a functional gene set using the Random Walk with Restart (RWR) algorithm [18]. Pathway Analysis with Down-weighting of Overlapping Genes (PADOG) uses the frequency of a gene present in the pathways analyzed to improve gene set analysis [19]. Functional Link Enrichment of Gene Ontology or gene sets (LEGO) takes into consideration these two types of information by incorporating network-based gene weights in ORA analysis [20]. However, methods like EnrichNet, PADOG or, LEGO are all based on the non-pathway-topology based method, not based on pathway topology. Thus, like any other non-pathway-topology based methods, these methods do not consider the complex interactions between genes in pathways. Moreover, the gene weights used by them are fixed values for different diseases. These weights cannot reflect the association between genes and diseases.

In this study, we developed an extended method of SPIA, called *gwSPIA,* by considering both the pathway topology (i.e. the interactions between genes in pathways) and the importance of genes. This method is a gene weight method, but not like the methods mentioned above this method is improved from a pathway-topology based method. It enhanced SPIA by weighing genes with various signaling pathway-based merits. Thus, gwSPIA has two advantages: on one hand, it takes the complex interactions in consideration; on the other hand, it also takes the importance of gene which can reflect the association between genes and diseases in consideration. Three types of weights are used in gwSPIA: impact factor (IF), betweenness centrality (BC), and specificity (SP). We increased the accuracy of SPIA by incorporating these weights into the SPIA method. In addition, we can analyze the relationship strength of differential expressed genes and specific phenotypes based on IF and SP. The IF value can vary a lot for different diseases which can reflect the association between genes and diseases. We applied the new method gwSPIA to 33 data sets. All these data sets are picked up based on specific disease pathways. For example, we aim for the *colorectal cancer* pathway as the target pathway in colorectal cancer data set. As an extended version of SPIA, we compare gwSPIA with SPIA and other 6 methods from three aspects: sensitivity, prioritization, and specificity. Results show that the performance of gwSPIA is always ranked in the second place in terms of comparing both sensitivity and prioritization compared with other methods. And our results demonstrate the efficacy of the gene weights used in this study.

II. METHODS

A. BENCHMARK DATA

A total 33 datasets from the KEGGdzPathwaysGEO R-package and KEGGandMetacoreDzPathwaysGEO R-packages are used as benchmark data (see Table 1) [19], [21], including various cancers and neurological diseases. These disease datasets are analyzed in the same way: First, for the probes with the same Entrez gene ID, only the probes with the highest average expression can be retained. Second, differential expression analysis is performed by fitting linear models using the empirical Bayes method as implemented in the limma R-package [22]. Each of the 33 datasets was matched with the corresponding KEGG pathway according to its name, e.g. a dataset of colorectal cancer patients is associated with the colorectal cancer pathway [23]. We call such a pathway a target pathway and its p-value and rank in the database are further evaluated. We use the following rules to identify DEGs: i) selecting more than 200 genes with FDR adjusted p-values < 0.1; ii) if not, selecting more than 200 genes with original p-values < 0.05 and log (fold change)> 1.5; iii) if not, selecting top 1% of genes ranked by p-values.

B. PATHWAYS IN KEGG DATABASE

The signaling pathways are downloaded in the KEGG. In the KEGG database, each pathway stores in an XML document (KGML format). For our method, we analyze 213 signaling pathways downloaded from the KEGG database one by one. The pathways can be transformed into networks format using KEGGgraph R-package, where genes and their interactions are nodes and edges respectively [50]. The igraph R-package is used to combine the 213 gene networks into one big gene network [51].

C. SIGNIFICANTLY ENRICHED PATHWAYS ANALYSIS

The procedure to identify significantly enriched pathways using gwSPIA include: (i) reconstructing the gene network from the signaling pathways using KEGGgraph R-package,

TABLE 1. Data sets used for assessing the proposed methods.

ID	Target pathway	GEO ID	Ref
1	Colorectal cancer	GSE4107	[24]
2	Colorectal cancer	GSE4183	[25, 26]
3	Colorectal cancer	GSE8671	[27]
4	Colorectal cancer	GSE9348	[28]
5	Colorectal cancer	GSE23878	[29]
6	Non-small cell lung cancer	GSE18842	[30]
7	Pancreatic cancer	GSE15471	[31]
8	Pancreatic cancer	GSE16515	[32]
9	Pancreatic cancer	GSE32676	[33]
10	Thyroid cancer	GSE3467	[34]
11	Thyroid cancer	GSE3678	-
12	Alzheimer's disease	GSE5281_HIP	[35]
13	Alzheimer's disease	GSE5281_EC	[35]
14	Alzheimer's disease	GSE5281_VCX	[35]
15	Alzheimer's disease	GSE1297	[36]
16	Alzheimer's disease	GSE16759	[37]
17	Chronic myeloid leukemia	GSE24739_G0	[38]
18	Chronic myeloid leukemia	GSE24739_G1	[38]
19	Acute myeloid leukemia	GSE14924_CD4	[39]
20	Acute myeloid leukemia	GSE14924_CD8	[39]
21	Acute myeloid leukemia	GSE9476	[40]
22	Dilated cardiomyopathy	GSE1145	-
23	Dilated cardiomyopathy	GSE3585	[41]
24	Endometrial cancer	GSE7305	[42]
25	Glioma	GSE19728	[43]
26	Glioma	GSE21354	[43]
27	Huntington's disease	GSE8762	[44]
28	Parkinson's disease	GSE20291	[45]
29	Parkinson's disease	GSE20164	[46]
30	Prostate cancer	GSE6956AA	[47]
31	Prostate cancer	GSE6956C	[47]
32	Renal cell carcinoma	GSE781	[48]
33	Renal cell carcinoma	GSE14762	[49]

(ii) calculating the gene weights including SP and BC from these gene networks, (iii) reconstructing the all signaling pathways into a big gene network by the genes shared in different pathways through R package igraph, (iv) projecting the DEGs onto the networks of the constructed gene networks, (v) calculating the gene weight IF based on the big gene network, and (vi) evaluating the statistical significance of each pathway. gwSPIA is implemented by using the statistical programming language R. The flow can be seen in Figure 1.

D. PATHWAY-BASED GENE WEIGHTING

Three types of weight measures are used in gwSPIA: IF, BC, and SP. IF of a gene is defined as the number of differential

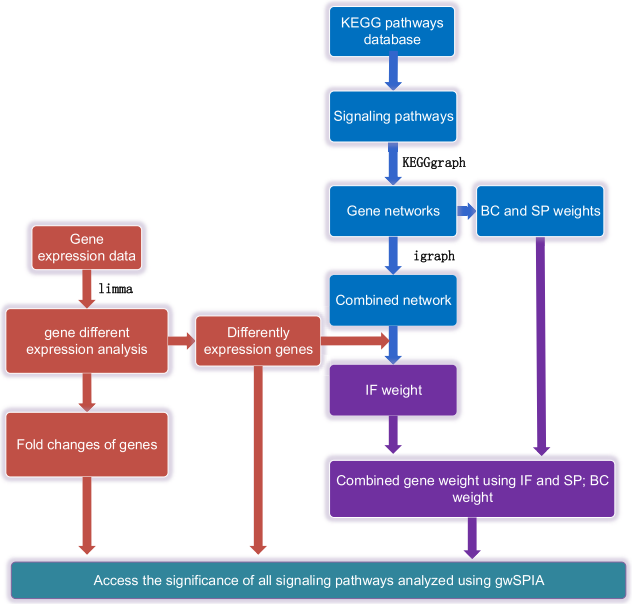


FIGURE 1. The brief workflow of gwSPIA. The operations with blue color are only done based on the topology of signaling pathways. The operations with red color are done from gene expression data. The operations with purple color are done with the topology of signaling pathways and the gene expression data. The operations with green color are done with all information.

expressed genes linked to the gene downstream directly. If the value which reflects the DEGs of the genes in the direct downstream place of a gene is bigger, we think that the gene is tended to be an important gene which may lead to the different expression of the DEGs downstream directly. Thus, the bigger IF is, the more important the gene is. The IF is an up-weighting value. Furthermore, the IF value is variational in different datasets and can reflect the association between gene importance and related diseases. In graph theory, BC is a measure of centrality in a graph based on shortest paths. Here we define BC of a gene as the number of shortest paths across it between two any genes in a given signaling pathway. The more the shortest paths cross a gene in a signaling pathway, the more important the gene is. For example, in the *Ras signaling pathway*, *Ras* is included in much more shortest paths than other genes in the pathway so changing expression of gene *Ras* might have a bigger impact on the expression of the pathway. And SP of a gene is the number of times of the gene that appears in all signaling pathways analyzed. The genes with lower SP might more important than other genes. This is just like the sight words in English [19]. Thus, the SP value is a down-weighting value [19].

Given P networks with G genes in these networks, we calculate the gene weights of each gene in the P networks. Here, for each gene G_i in P_j network, we compute its BC value as the numbers of liner paths across G_i using *igraph* R package. The bigger the BC value, the more important a gene is. If the BC values in the network P_j have s levels, and the BC value of the gene G_i in the network P_j at k level, then the BC value was normalized in such a way:

$$w_{BC} = k \quad 1, \cdots, S \tag{1}$$

We compute the SP from all 182 signaling pathways. The SP value of a gene G_i is the frequency of the gene appearing in all 182 signaling pathways. Before calculating the IF value of genes, we use the R package *igraph* to reconstruct the 182 signaling pathways into a big gene network. The IF value of a gene is defined as the number of differential expressed genes linked with the gene in the downstream. Thus, the IF value of a gene is ranged from 0 to the out-degree of the gene.

We think that the IF value is associated with the SP value. The larger the frequency of a gene appearing in the pathways analyzed is, the bigger out-degree the gene is supported to have in the big gene network contents all signaling pathways analyzed. The IF value of a gene is calculated from the out-degree of the gene. So the IF value may be influenced by the SP value. So to eliminate the effects of SP value to IF value, we calculate the weight of the gene from IF and SP value in such way:

$$w_I = \frac{IF}{SP} \tag{2}$$

The w_I is an up-weighting value. The bigger w_I of a gene is, the more important the gene is. The w_I can vary in a large range. Thus, we use the Min-Max normalization to normalize the gene weight w_I using the formula:

$$w_n = 1 + \frac{w_I - \min(w_I)}{\max(w_I) - \min(w_I)} \tag{3}$$

E. SIGNALING PATHWAY IMPACT ANALYSIS BY INCORPORATING SIGNALING PATHWAY-BASED GENE WEIGHTS (gwSPIA)

gwSPIA method combines the over-representation of pathways and the abnormal perturbation in a given pathway in the same way with the SPIA method [12]. These two aspects are captured by two independent probability values, P_{FCS} and P_{PERT}.

P_{FCS} captures the significance of the given pathway by a method improved from the PADOG method [19]. P_{FCS} is the probability to observe an average t-score in a given pathway, S_{per}, more extreme than S_{obs} just by chance. The t-score of a gene is calculated using R package limma. The average t-score of a given pathway P_j is calculated in this way:

$$S_{obs} = \frac{1}{N(P_j)} \sum_{g_i \in P_j} |T(g_i)| \cdot w_n(g_i) \tag{4}$$

In Equation (4), $T(g_i)$ dominate the t-score of a gene g_i in the given pathway P_j, $w_n(g_i)$ represent the weight w_n of the gene g_i. $N(P_j)$ is the number of genes in the given pathway P_j. Different from the original PADOG method, we multiply the $T(g_i)$ with the gene weight $w_n(g_i)$ instead of the normalized frequency of the gene g_i across all gene sets to be analyzed to improve the accuracy. When computing the random score S_{per} for the same pathway, we randomly select $N(P_j)$ t-scores from the t-scores of all genes analyzed. Then S_{per} is computed for nB=2000 times, so $P_{FCS}(P_j)$ $P_{OVER}(P_j)$ is computed in such way:

$$P_{FCS}(P_j) = \frac{\sum_{nB} I\left(S_{per}(P_j) \geq S_{obs}(P_j)\right)}{nB} \tag{5}$$

Because the t-score and the combined gene weight of a gene are all independent of fold change, the modified PADOG can be used to replace the original ORA method used in SPIA.

The probability P_{PERT} is measured by propagating measured expression changes across the pathway topology and is calculated based on the amount of perturbation of genes. Formulate of calculating the perturbation of each gene (perturbation factor) is as follow:

$$PF(g_i) = w_{BC}(g_i) .\Delta E(g_i) + \sum_{m=1}^{n} \beta_{im} \frac{PF(g_m)}{N_{ds}(g_m)} \tag{6}$$

Differ from the SPIA method, the first term represents that the signed normalized measured expression change of the gene g_i (log fold-change if two conditions are compared) multiply with the normalized BC value of the gene. The second term in Equation (6) is the sum of perturbation factors of the genes g_m directly upstream of the target gene g_i, normalized by the number of downstream genes of each such gene $N_{ds}(g_m)$. The absolute value of β_{im} quantifies the strength of the interaction between genes g_i and g_j . And in this method, the β_{im} is always set as 1. Just like SPIA, gwSPIA calculates the net perturbation accumulation at the level of each gene, $Acc(g_j)$, as the difference between the perturbation factor PF of a gene and its observed weighted log fold-change:

$$Acc(g_i) = PF(g_i) - w_{BC}(g_i) .\Delta E(g_i) \tag{7}$$

The total net accumulated perturbation in the pathway is computed as $t_A = \sum Acc(g_i)$. Then the second probability, P_{PERT}, will be the probability to observe a total accumulated perturbation of the pathway, T_A, more extreme than t_A just by chance:

$$P_{PERT} = P(T_A \geq t_A | H_0) \tag{8}$$

Then gwSPIA defines a significance evaluation index P_G in the same way with SPIA, which is calculated by the following formula:

$$P_G = c - c \ln c \tag{9}$$

In the formula (9), $c = P_{OVER} \times P_{PERT}$.

When there are more than one gene sets for analysis, we use Bonferroni procedure to adjust the p-value [52]. We also report the adjusted p-value based on FDR [53]. Actually, the two weights can be used to improve other methods. Because in the calculation of the perturbation factor of a gene, N_{ds} which is the number of downstream genes of such gene is used. If we use the weight w_I to improve the PF, the function of N_{ds} and w_I will be repetitive.

F. SEVEN OTHER METHODS APPLIED TO COMPARE WITH gwSPIA

Since gwSPIA is like an enhanced version of SPIA with gene weight considered, we first compare gwSPIA with the SPIA method. Fisher [54], GSA [55], and GSEA [11] are three classic pathway analysis methods. Then Fisher, GSA, and GSEA method are also compared with gwSPIA. MRGSE and ROntoTools are also compared with gwSPIA. In the end, the gene-weight-based method PADOG is also compared with gwSPIA.

(1) SPIA: SPIA combines the GSA method instead of ORA method and the abnormal perturbation analysis. The SPIA R package developed by Tarca et al. is applied to perform SPIA [12].
(2) Fisher: Fisher is a classic method which is referred to as "2×2 table method". It uses Hypergeometric test to evaluate the significance of pathways [54].
(3) GSA: GSA uses the max mean as the statistics to test whether the genes in a particular gene set have coordinated changes [55]. We used the function "GSA" of the GSA package in R.
(4) GSEA: Gene Set Enrichment Analysis, frequently used and widely accepted method, a Kolmogorov-Smirnov statistic was used to test whether the rank of the p-value of a gene in a genome was similar to a uniform distribution [11]. The GSEA is implemented as the default set-based enrichment method of the EnrichmentBrowser [56] R package.
(5) MRGSE: tests if the ranks in a particular gene set (sorted by the t-statistics) are different from genes in the background gene list [57]. We use moderated t-test in "limma" package in R to calculate t values and "geneSetTest" function in "limma" package for ranking of enriched gene sets.
(6) ROntoTools: Pathway-Express in ROntoTools incorporates pathways topology to calculate a global probability for genes in a given pathway. The negative log value of this global probability is defined as the impact factor of a gene, and the impact factor values are then used for pathway enrichment analysis [12], [58]. This method is contained in the R package "ROntoTools" and the function "pe" was used.
(7) PADOG: the method computes a gene set score as the mean of absolute values of weighted moderated gene t-scores. The gene weights are chosen to favor genes appearing in few pathways versus genes that appear in many pathways [19]. We use the R package PADOG to analyze the data sets.

G. THE MEASURES OF SENSITIVITY, PRIORITIZATION, AND SPECIFICITY

Michaela et al. compiled many disease gene expression datasets whose KEGG target pathways were known [23] using a number of measures to evaluate the performance of different methods. We also follow these measures in the benchmark, which are including sensitivity, prioritization, and specificity. The sensitivity of a method is the median p-values of the target pathways in 33 datasets, with a lower p-value indicates higher sensitivity. In each dataset, all tested pathways are sorted by their p-value from lowest to highest. The rank percentage of the target pathway is the ratio of the rank of target pathway and the number of pathways analyzed (213 pathways in this study). The prioritization of a method is the median rank percentage of the target pathways in 33 datasets with lower rank percentages indicates higher prioritization. The specificity of a method is the ratio of significant pathways (using a significance threshold of 0.05) in the pathways analyzed.

III. RESULTS

In 2015, Michaela et al. compiled 36 disease gene expression datasets in which the target pathway of each dataset was known [23]. They used these datasets to compare 7 FCS methods. 33 datasets of 36 can acquire the results of target pathways using all 8 methods compared in this article. Thus, here, to further test the usefulness of gwSPIA in functional studies, we select 33 datasets as the benchmark. And we compare gwSPIA with SPIA, Fisher, GSA, GSEA, MRGSE, ROntoTools, and PADOG. The descriptions of these methods can be found in Methods. We follow the instructions of their respective R package to perform pathway enrichment. The usefulness of these gene weights used in this article is also proved in this section.

A. COMPARISON OF gwSPIA WITH SEVEN METHODS USING A BENCHMARK OF 33 DISEASE GENE EXPRESSION DATASETS

In total 33 gene expression datasets are used to compare gwSPIA with seven other methods. Every dataset represents a certain disease and has been linked to a defined target pathway from the KEGG database (see Table 1). Michaela et al. proposed measures to test the performance of a method, which are sensitivity (defined as the p-value of the target pathway), prioritization (defined as the rank percentage for the target pathway), and specificity (defined as the average percentage of pathways detected as significant and not significant).

Out of all eight methods being compared, gwSPIA ranks in second place in terms of sensitivity (the median p-value of the target pathways in 33 datasets) (Figure 2). And gwSPIA also ranks in second place in terms of prioritization (the median rank for target pathways in 33 datasets) (Figure 3). It is worth noting that the sensitivity measure and the prioritization measure evaluate the performance of a method from different aspects, and a method that ranks in first place based on a measure may not also rank in first place based on another measure. For instance, PADOG ranks in first place in terms of prioritization, while it ranks in fourth place in terms of sensitivity. MRGSE ranks in first place in terms of sensitivity, but it only ranks in sixth place in terms of prioritization. gwSPIA ranks in second place in both sensitivity and prioritization, suggesting that it is able

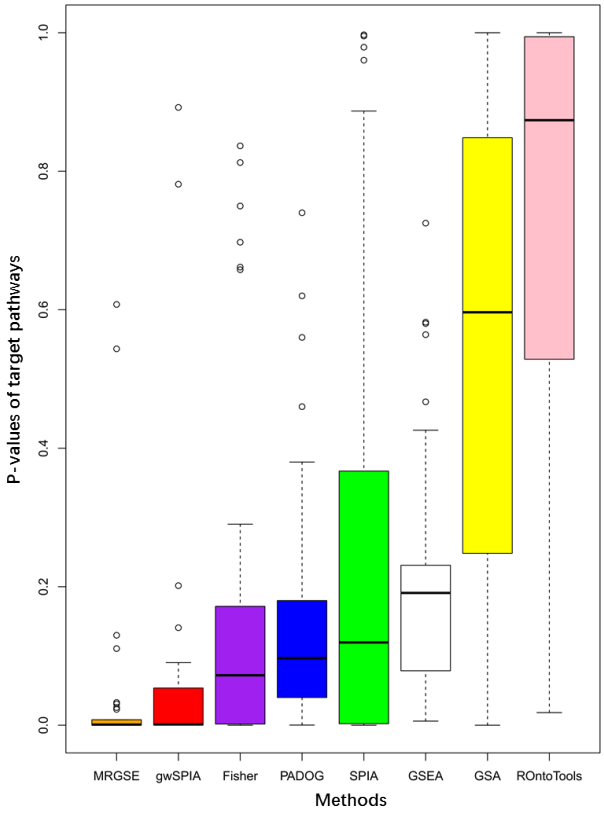


FIGURE 2. gwSPIA performs the 2nd among other methods in terms of the sensitivity of detecting target pathway. The boxplots show the distribution of the target pathways p-values.

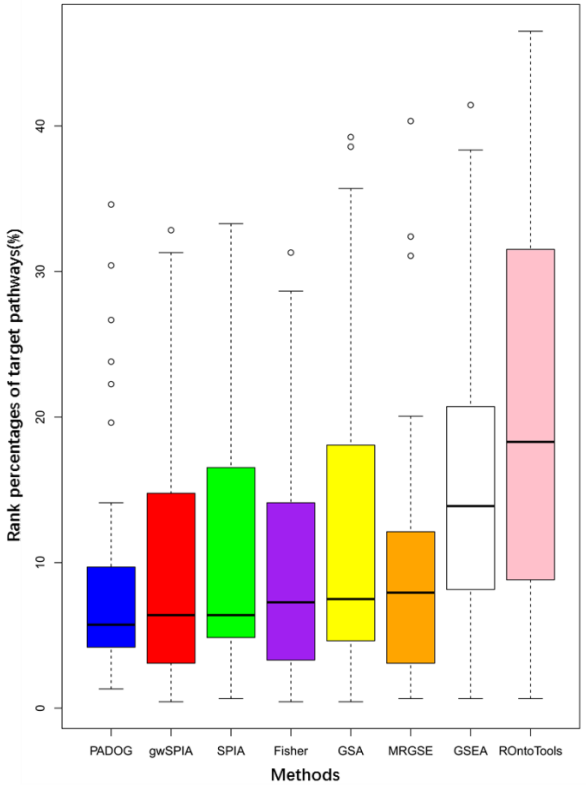


FIGURE 3. gwSPIA performs the 2nd among other methods in terms of prioritization of detecting target pathway. The boxplots show the distribution of the target pathways ranks.

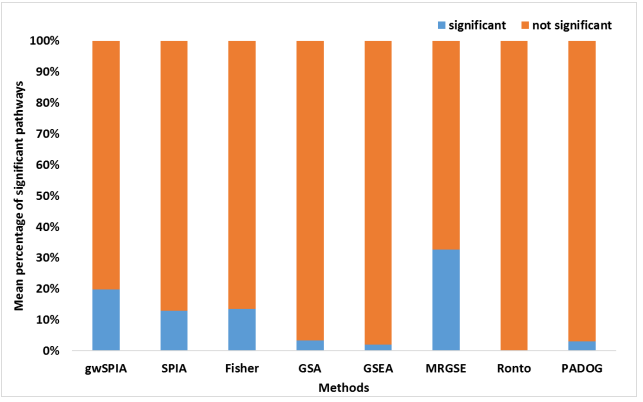


FIGURE 4. Average percentage of the pathways detected as significant and not significant by each method using the threshold of p-values <= 0.05.

to prioritize target pathways with high sensitivity. For the specificity of the eight methods, we can note that MRGSE identifies on average almost 50% of all database pathways as significantly enriched, while for GSA it is lower than 10% of the significant pathways (Figure 4). For SPIA, Fisher, and gwSPIA can identify between 20% and 35% of the significant pathways. This can demonstrate that the gwSPIA may have a similar specificity to SPIA and Fisher. For MRGSE, depending on disease, multiple pathways could be altered. However, it could be questioned whether such a number of pathways is realistic or it reflects the lack of specificity of this method [23]. For GSA, depending on disease, few pathways could be altered. This method also lacks specificity. It can infer that the specificity of gwSPIA is modest. In conclusion, the performance of gwSPIA in this benchmark well illustrates its usefulness. And this result also suggests that the use of gene weights in gwSPIA can improve the performance of SPIA. In addition, the computational efficiency of gwSPIA is similar to SPIA. This is because that gwSPIA is an extended method of SPIA.

B. DEGs ANALYSIS THROUGH THE GENE WEIGHT COMBINE IF AND SP IN TWO BENCHMARKS

In this paper, we calculate a gene weight w_I that combines IF and SP. The genes with high w_I are thought highly associated with the specific phenotype and vice versa. In this section, we analyze the top 10 DEGs ranked through the combined gene weight w_I of 2 benchmark data sets from largest to smallest, respectively. The 2 benchmark data sets are GSE4107 and GSE4183 as two examples. And we also analyze the 10 DEGs ranked in the end by the weight w_I from largest to smallest. All these genes can be seen in Table 2.

In the benchmark data set GSE4107, genes *SFRP2*, *SFRP1*, *PRNP*, *NFE2L2,* and *WWTR1* which are ranked in the first play important roles in colorectal cancer shown in MalaCards dataset). Gene *SFRP5* is an important paralog of *SFRP1*. Genes *SFRP2, SFRP1*, *SFRP4,* and *SFRP5* are in the same

TABLE 2. The top and bottom 10 DEGs ranked through the combined gene weight OF 2 BENCHMARK DATA SETS FROM largest to smallest in GSE4107 and GSE4183.

GEO	GSE4107			GSE4183		
	Gene Name	ENTREZ Gene ID	w_I	Gene Name	ENTREZ Gene ID	w_I
Top 10	*SFRP2*	6423	8	*YAP1*	10413	4
	SFRP1	6422	8	*WWTR1*	25937	3
	SFRP5	6425	8	*LATS2*	26524	3
	SFRP4	6424	8	*TEAD1*	7003	3
	PARVB	29780	7	*COL6A3*	1293	2.33
	PRNP	5621	5	*ROBO3*	64221	2
	NFE2L2	4780	4	*FST*	10468	2
	PRPH	5630	4	*AJUBA*	84962	2
	WWTR1	25937	4	*KCNK5*	8645	2
	KIF3A	11127	4	*AGAP2*	116986	2
Bottom 10	*MCU*	90550	0	*WDR61*	80349	0
	SLC27A2	11001	0	*SPIRE2*	84501	0
	STEAP1	26872	0	*CELA3B*	23436	0
	GPX7	2882	0	*NPFFR1*	64106	0
	MAN1C1	57134	0	*SPTLC2*	9517	0
	SLC30A1	7779	0	*DERL2*	51009	0
	PCGF5	84333	0	*SLC5A1*	6523	0
	ENTPD1	953	0	*CLIP1*	6249	0
	CD46	4179	0	*AMOTL1*	154810	0
	AQP8	343	0	*TRPM6*	140803	0

node in the *wnt signaling pathway*. The *wnt signaling pathway* is highly associated with colorectal cancer [59]. Gene *PARVB* is in the *Focal adhesion* pathway which also plays an important role in colorectal cancer [60]. Gene *KIF3A* only participates in the *Hedgehog signaling pathway* which is also highly associated with colorectal cancer [61]. The gene *PRPH* has no association with colorectal cancer. This may come from the noise of the data. These genes are the top 10 genes ranked by the weight from the largest to smallest. Genes *MCU*, *SLC27A2*, *STEAP1*, *GPX7*, *MAN1C1*, *SLC30A1*, *PCGF5* and *ENTPD1* are not associated with colorectal cancer as the MalaCards data set show. *CD46* and *AQP8* are associated with colorectal cancer. Gene *CD46* participates in *Complement and coagulation cascades* pathway and *Measles* which have no relationship with colorectal cancer. Gene *AQP8* is in the most downstream of the *Bile secretion* pathway. Thus, there is no gene in the downstream of gene *AQP8*. These genes are the last 10 genes ranked by the weight from the largest to smallest.

In the benchmark data set GSE4183, genes *YAP1*, *WWTR1*, *TEAD1*, and *AJUBA* are highly associated with colorectal cancer shown in the MalaCards data set. An important paralog of gene *LATS2* is *LATS1*. From the MalaCards data set, we can find that *LATS1* is associated with colorectal cancer. And *LATS2* participates in the *Hippo signaling pathway* which is highly associated with colorectal cancer [62]. *COL6A3* may be a promising biomarker or target for the prognosis and treatment of CRC [63]. An important paralog of *ROBO3* is *ROBO2*. From the MalaCards data set, we can find that *ROBO2* is associated with colorectal cancer. *ROBO3* appears in the *Axon guidance* pathway. The *Axon guidance* pathway plays an important role in colorectal cancer [64]. *FST* is in the upstream of the *TGF-beta signaling pathway* which is also an important pathway for colorectal cancer [65]. *AGAP2* is also in the upstream of the *FoxO signaling pathway* and participates in the *Endocytosis* pathway. The two pathways are both highly related to colorectal cancer [66], [67]. These genes are the top 10 genes ranked by the weight from the largest to smallest. The genes *WDR61*, *SPIRE2*, *CELA3B*, *NPFFR1*, *SPTLC2*, *DERL2*, *CLIP1*, *SLC5A1*, and *AMOTL1* have no association with colorectal cancer and are ranked in the last 10 by the weight from the largest to smallest. *TRPM6* is highly associated with colorectal cancer. However, this gene is a gene with no genes in the downstream.

From the analysis of genes with weight w_I in the two benchmark data sets, we can see that the DEGs with high w_I are supposed to highly associate with a specific phenotype. The DEGs with low w_I always have no association with specific phenotypes.

C. THE USEFULNESS OF THE COMBINED GENE WEIGHT

We enrich the genes which ranked in the top 500 by the value of gene weight w_I ranked from high to low. The result showed that 25 signaling pathways are enriched more than 10 genes using 5 colorectal cancer datasets of 33 benchmark (Figure 5). Most of these pathways are highly associated with colorectal cancer according to scientific articles. The pathway *Chemical carcinogenesis* is highly associated with colorectal cancer [68]. Phosphatidylinositol 3-Kinase is important

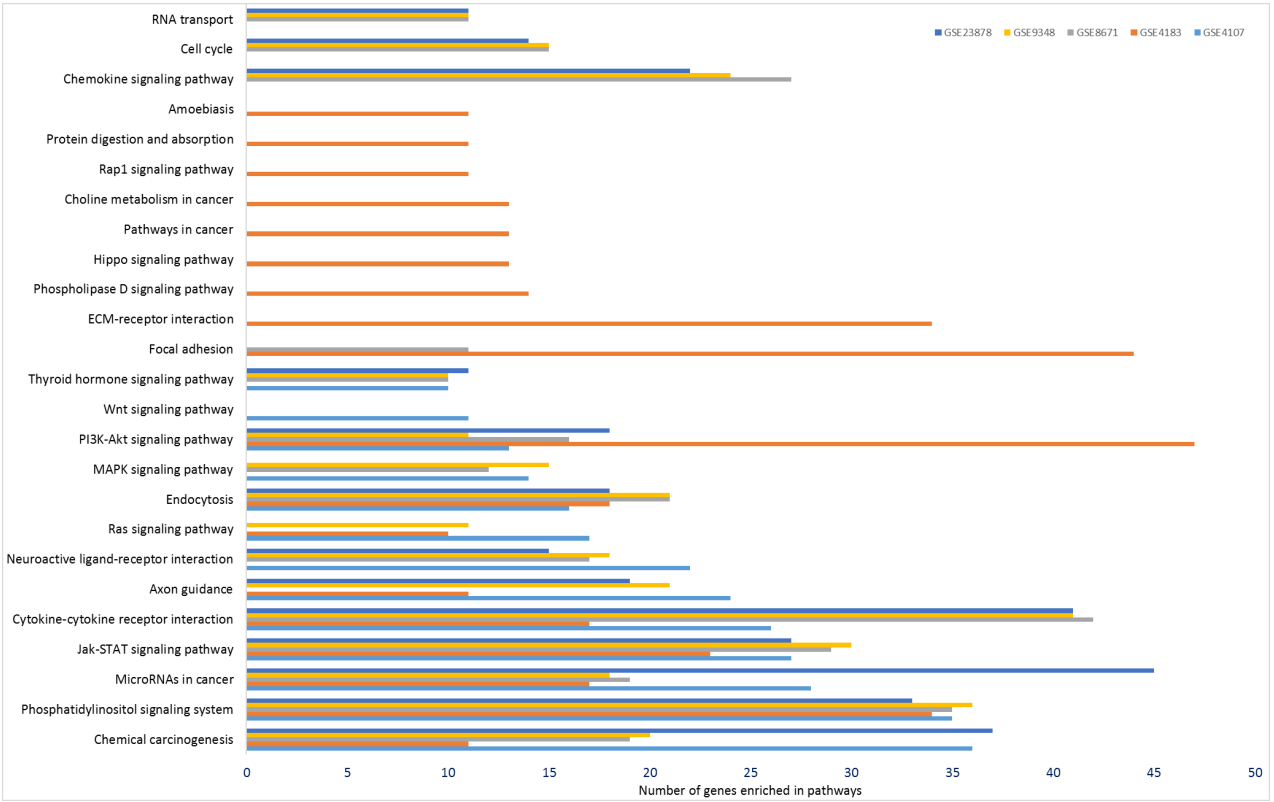


FIGURE 5. The distribution of signaling pathways that contain more than 10 genes with a high combined gene weight ranked in the top 500 genes by the value of the combined gene weight.

in the management of metastatic colorectal cancer [69]. Thus, the signaling pathway *Phosphatidylinositol signaling system* may associate with colorectal cancer. Obviously, pathway *MicroRNAs in cancer* is highly associated with cancer. The *Jak-STAT signaling pathway* is a biological target with therapeutic implications [70]. The proliferation and migration of colorectal cancer cells can be inhibited by the inhibition of cytokine receptors [71]. Thus, pathway *Cytokine-cytokine receptor interaction* may associate with colorectal cancer. SLIT2 axon guidance can suppress the growth of colorectal cancer cells [72]. It indicates that the *Axon guidance* pathway is related to colorectal cancer. *HOXA3* promotes tumor growth of human colorectal cancer through activating the *Ras signaling pathway* [73]. The pathway *Endocytosis* obviously plays an important role in colorectal cancer [74]. *Focal adhesion*, *Pathways in cancer*, *Cell cycle*, *ECM-receptor interaction*, *MAPK signaling pathway*, *PI3K-Akt signaling pathway*, *Choline metabolism in cancer*, *Phospholipase D signaling pathway*, *Wnt signaling pathway*, and *RNA transport* are highly associated with cancer including colorectal cancer [75]–[80]. Activated thyroid hormone promotes differentiation and chemotherapeutic sensitization of colorectal cancer stem cells [81]. So the *Thyroid hormone signaling pathway* is related to colorectal cancer. *RASAL2* promotes tumor progression through the *LATS2/YAP1* axis of *Hippo signaling pathway* in colorectal cancer [82]. For the *Rap1 signaling pathway*, Rap1B is a target of miR-139

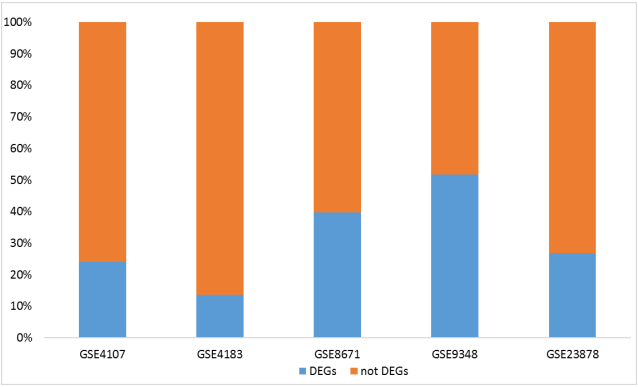


FIGURE 6. The percentages of the 500 genes detected as different expression and not different expression in 5 colorectal cancer datasets.

to suppress human colorectal cancer cell proliferation [83]. *Chemokine signaling pathway* plays an important role in hepatic metastasis of colorectal cancer [84]. Pathways like *Protein digestion and absorption* and *Amoebiasis* are not associated with colorectal cancer. Thus, they only enriched more than 10 genes in GSE4183.

The percentages of the 500 genes detected as different expression and not different expression in 5 colorectal cancer datasets are also shown in the results (Figure 6). The results show that most of the top 500 genes are not DEGs in the 5 colorectal cancer datasets. And it is worth noting that the pathways enriched more than 10 genes are mostly associated

with colorectal cancer. In conclusion, the combined gene weight used in this study is useful.

D. BC VALUES OF GENES IN PATHWAYS

The genes in a signaling pathway with high betweenness centrality always play roles as switches in the signaling pathways, such as Ras in the *Ras signaling pathway*. From the *Ras signaling pathway* in the KEGG data set, we can see that *Ras* is in the key position on the *Ras signaling pathway*. And the BC value of node *Ras* in the pathway is 112, which is the highest. *Rap1* in the *Rap1 signaling pathway* also plays an important role. The BC value of node *Rap1* in the pathway is 196 which is also the highest. In the *Notch signaling pathway*, the BC value of node *Notch* is 51 which is the second largest in the pathway. The highest BC value in this pathway is 52. *PI3K-Akt signaling pathway* contains a node named *AKT* with BC value 366 which is the highest in the pathway. All these nodes in these pathways are very important. The pathways are named by these nodes. Thus, we consider that the genes in a pathway with high BC values play important roles in the pathway.

IV. DISCUSSION

The major limitations of traditional non-pathway-topology based methods are that they ignore the functional non-equivalence roles of genes and the complex interactions between genes. And the major limitation of pathway-topology based methods is that they ignore the functional non-equivalence roles of genes. Gene weight dominated the importance of genes that collect the information of the functional non-equivalence roles of genes from the pathways can be used to overview the common limitation of non-pathway-topology based methods and pathway-topology based methods as mentioned above. In fact, methods incorporated the functional non-equivalence roles of genes are presented by some studies in recent years. For instance, EnrichNet measures the functional association between the interesting gene list and a gene set using a Random Walk with Restart (RWR) algorithm [18]. PADOG uses the value that the frequency of a gene appears in pathways analyzed to improve gene set analysis [19]. However, these methods do not consider the complex interactions between genes. These methods only incorporate gene weights with non-pathway-topology based methods. And the gene weight used in these methods cannot reflect the association between genes' importance and disease. Therefore, we propose a new method combined the pathway-based weights of genes IF, SP and BC with a pathway-topology based method SPIA, called gwSPIA. For evaluating our method is applied to 33 differentially expressed gene datasets.

To test the performance of gwSPIA, we compare gwSPIA with original SPIA and other 6 methods in three aspects: sensitivity, prioritization, and specificity following Michaela et al. proposed in 2015. Results show that gwSPIA always ranks in the top compared with the other 7 methods in terms of sensitivity and prioritization. And the specificity of gwSPIA is similar to SPIA and Fisher. Thus, the gene weight used in gwSPIA can help to improve the performance of SPIA.

The method gwSPIA is a gene-weight-based method. The gene weight used in gwSPIA can not only help to improve the performance of SPIA but also can help users to filter DEGs. In this study, the gene weights w_I which combined IF and SP value can help to filter DEGs. Two colorectal cancer data sets are used to illustrate the ability of gene weights w_I . The DEGs with high w_I are supposed to highly associate with a specific phenotype. And the DEGs with low w_I always have no association with specific phenotypes. We also enrich the genes which ranked in the top 500 by the value of gene weight ranked from high to low using the data from the 5 colorectal cancer datasets. The result showed that the combined gene weight w_I can also be used in the genes which are not DEGs. And the genes in a given pathway with high BC values are supposed to play important roles in the given pathway. These results all dominate the usefulness of the gene weights used in this article.

However, the gene weights IF, SP and BC cannot completely explain the importance of genes in pathways. The IF value is calculated from the DEGs in specific data sets. Thus, the redundancy of data sets can also influence the IF value. And the genes at the end of the network which contain all signaling pathways in the analysis do not all have high association with specific phenotypes. However, the IF values of these genes are zero. Weights that can completely explain the importance of genes in pathways need to be excavated to improve the limitation of proposed methods. Thus, some pathways are not associated with colorectal cancer when enriching the top 500 genes. Nevertheless, the results of gwSPIA and the analysis of gene weights in gwSPIA already make gwSPIA a useful method for large-scale functional genomics studies. In the further, we may use a gene weight to quantify the importance of a gene in the aspect that the number of articles which report the gene is associated with a specific disease. We think that this gene weight can overcome the limitation of the weight used in this article.

V. CONCLUSION

In this study, we developed a new method based on signaling pathway impact analysis combined with consideration of the importance of genes. And we show that this method outperforms better than SPIA and popular standard signaling pathway analysis methods in identifying altered signaling pathways. Furthermore, we show that the weighted genes importance could help us to detect the false positive genes in differentially expressed genes. We also show that the roles which we used to weight genes' importance are scientific.

REFERENCES

[1] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.
[2] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Res.*, vol. 44, pp. D457–D462, Jan. 2016.

[3] *KEGG*. Accessed: May 2019. [Online]. Available: https://www.kegg.jp/
[4] D. Nishimura, "BioCarta," *Biotech Softw. Internet Rep.*, vol. 2, no. 3, pp. 117–120, Jul. 2001.
[5] *BioCarta*. Accessed: May 2019. [Online]. Available: https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways
[6] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein, "Reactome: A knowledgebase of biological pathways," *Nucleic Acids Res.*, vol. 33, pp. D428–D432, Jan. 2005.
[7] *Reactome*. Accessed: May 2019. [Online]. Available: https://www.reactome.org/
[8] S. Drăghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz, "Global functional profiling of gene expression," *Genomics*, vol. 81, no. 2, pp. 98–104, Feb. 2003.
[9] P. Khatri, S. Drăghici, G. C. Ostermeier, and S. A. Krawetz, "Profiling gene expression using onto-express," *Genomics*, vol. 79, no. 2, pp. 266–270, Feb. 2002.
[10] Q. Zheng and X. J. Wang, "GOEAST: A Web-based software toolkit for gene ontology enrichment analysis," *Nucleic Acids Res.*, vol. 36, pp. W358–W363, Jul. 2008.
[11] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Nat. Acad. Sci. USA*, vol. 102, pp. 15545–15550, Sep. 2005.
[12] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J. S. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero, "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, Jan. 2009.
[13] X. Li, L. Shen, X. Shang, and W. Liu, "Subpathway analysis based on signaling-pathway impact analysis of signaling pathway," *PLoS ONE*, vol. 10, Jul. 2015, Art. no. e0132813.
[14] Z. Bao, X. Li, X. Zan, L. Shen, R. Ma, and W. Liu, "Signalling pathway impact analysis based on the strength of interaction between genes," *IET Syst. Biol.*, vol. 10, no. 4, pp. 147–152, Aug. 2016.
[15] L. Geistlinger, G. Csaba, R. Küffner, N. Mulder, and R. Zimmer, "From sets to graphs: Towards a realistic enrichment analysis of transcriptomic systems," *Bioinformatics*, vol. 27, no. 13, pp. i366–i373, Jul. 2011.
[16] I. Ihnatova and E. Budinska, "ToPASeq: An R package for topology-based pathway analysis of microarray and RNA-Seq data," *BMC Bioinformatics*, vol. 16, p. 350, Dec. 2015.
[17] W. Liu, P. Xu, and Z. Bao, "Understanding the mechanisms of cancers based on function sub-pathways," *Comput. Biol. Chem.*, vol. 78, pp. 491–496, Feb. 2019.
[18] E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, and A. Valencia, "EnrichNet: Network-based gene set enrichment analysis," *Bioinformatics*, vol. 28, no. 18, pp. i451–i457, Sep. 2012.
[19] A. L. Tarca, S. Draghici, G. Bhatti, and R. Romero, "Down-weighting overlapping genes improves gene set analysis," *BMC Bioinformatics*, vol. 13, p. 136, Jun. 2012.
[20] X. Dong, Y. Hao, X. Wang, and W. Tian, "LEGO: A novel method for gene set over-representation analysis by incorporating network-based gene weights," *Sci. Rep.*, vol. 6, Jan. 2016, Art. no. 18871.
[21] A. L. Tarca, G. Bhatti, and R. Romero, "A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity," *PLoS ONE*, vol. 8, no. 11, Nov. 2013, Art. no. e79217.
[22] G. K. Smyth, *Limma: Linear Models for Microarray Data*. New York, NY, USA: Springer, 2005.
[23] M. Bayerlová, K. Jung, F. Kramer, F. Klemm, A. Bleckmann, and T. Beißbarth, "Comparative study on gene set and pathway topology-based enrichment methods," *BMC Bioinformatics*, vol. 16, p. 334, Oct. 2015.
[24] Y. Hong, K. S. Ho, K. W. Eu, and P. Y. Cheah, "A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: Implication for tumorigenesis," *Clin. Cancer Res.*, vol. 13, pp. 1107–1114, Feb. 2007.
[25] B. Gyorffy, B. Molnar, H. Lage, Z. Szallasi, and A. C. Eklund, "Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples," *PLoS ONE*, vol. 4, no. 5, p. e5645, May 2009.
[26] O. Galamb, B. Györffy, F. Sipos, S. Spisák, A. M. Németh, P. Miheller, Z. Tulassay, E. Dinya, and B. Molnár, "Inflammation, adenoma and cancer: Objective classification of colon biopsy specimens with gene expression signature," *Disease Markers*, vol. 25, no. 1, pp. 1–6, 2008.
[27] J. Sabates-Bellver, L. G. Van der Flier, M. de Palo, E. Cattaneo, C. Maake, H. Rehrauer, E. Laczko, M. A. Kurowski, J. M. Bujnicki, M. Menigatti, J. Luz, T. V. Ranalli, V. Gomes, A. Pastorelli, R. Faggiani, M. Anti, J. Jiricny, H. Clevers, and G. Marra, "Transcriptome profile of human colorectal adenomas," *Mol. Cancer Res.*, vol. 5, no. 12, pp. 1263–1275, Dec. 2007.
[28] Y. Hong, T. Downey, K. W. Eu, P. K. Koh, and P. Y. Cheah, "A 'metastasis-prone' signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics," *Clin. Exp. Metastasis*, vol. 27, no. 2, pp. 83–90, Feb. 2010.
[29] S. Uddin, M. Ahmed, A. Hussain, J. Abubaker, N. Al-Sanea, A. AbdulJabbar, L. H. Ashari, S. Alhomoud, F. Al-Dayel, Z. Jehan, P. Bavi, A. K. Siraj, and K. S. Al-Kuraya, "Genome-wide expression analysis of Middle Eastern colorectal cancer reveals FOXM1 as a novel target for cancer therapy," *Amer. J. Pathol.*, vol. 178, no. 2, pp. 537–547, Feb. 2011.
[30] A. Sanchez-Palencia, M. Gomez-Morales, J. A. Gomez-Capilla, V. Pedraza, L. Boyero, R. Rosell, and M. E. Fárez-Vidal, "Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer," *Int. J. Cancer*, vol. 129, no. 2, pp. 355–364, Jul. 2011.
[31] L. Badea, V. Herlea, S. O. Dima, T. Dumitrascu, and I. Popescu, "Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia," *Hepatogastroenterology*, vol. 55, no. 88, pp. 2016–2027, Nov./Dec. 2008.
[32] H. Pei *et al.*, "FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt," *Cancer Cell*, vol. 16, no. 3, Sep. 2009, pp. 259–266.
[33] T. R. Donahue, L. Li, B. L. Fridley, G. D. Jenkins, K. R. Kalari, W. Lingle, G. Petersen, Z. Lou, and L. Wang, "Integrative survival-based molecular profiling of human pancreatic cancer," *Clin. Cancer Res.*, vol. 18, no. 5, pp. 1352–1363, Mar. 2012.
[34] H. He *et al.*, "The role of microRNA genes in papillary thyroid carcinoma," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 52, pp. 19075–19080, Dec. 2005.
[35] W. S. Liang, L. M. Tran, R. Hill, Y. Li, A. Kovochich, J. H. Calvopina, S. G. Patel, N. Wu, A. Hindoyan, J. J. Farrell, X. Li, D. W. Dawson, and H. Wu, "Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain," *Physiol. Genomics*, vol. 28, no. 3, pp. 311–322, 2007.
[36] E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield, "Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 7, pp. 2173–2178, Feb. 2004.
[37] J. Nunez-Iglesias, C.-C. Liu, T. E. Morgan, C. E. Finch, and Z. X. J. Zhou, "Joint genome-wide profiling of miRNA and mRNA expression in Alzheimer's disease cortex reveals altered miRNA regulation," *PLoS ONE*, vol. 5, p. e8898, Feb. 2010.
[38] M. Affer, D. C. Taussig, A. G. Ramsay, R. Mitter, F. Miraki-Moud, R. Fatah, A. M. Lee, T. A. Lister, and J. G. Gribben, "Gene expression differences between enriched normal and chronic myelogenous leukemia quiescent stem/progenitor cells and correlations with biological abnormalities," *J. Oncol.*, vol. 2011, Feb. 2011, Art. no. 798592.
[39] R. Le Dieu *et al.*, "Peripheral blood T cells in acute myeloid leukemia (AML) patients at diagnosis have abnormal phenotype and genotype and form defective immune synapses with AML blasts," *Blood*, vol. 114, no. 18, pp. 3909–3916, Oct. 2009.
[40] D. L. Stirewalt, S. Meshinchi, K. J. Kopecky, W. Fan, E. L. Pogosova-Agadjanyan, J. H. Engel, M. R. Cronk, K. S. Dorcy, A. R. McQuary, D. Hockenbery, B. Wood, S. Heimfeld, and J. P. Radich, "Identification of genes with abnormal expression changes in acute myeloid leukemia," *Genes Chromosomes Cancer*, vol. 47, no. 1, pp. 8–20, Jan. 2008.
[41] A. S. Barth, R. Kuner, A. Buness, M. Ruschhaupt, S. Merk, L. Zwermann, S. Kääb, E. Kreuzer, G. Steinbeck, U. Mansmann, A. Poustka, M. Nabauer, and H. Sültmann, "Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies," *J. Amer. College Cardiol.*, vol. 48, no. 8, pp. 1610–1617, Oct. 2006.
[42] A. Hever, R. B. Roth, P. Hevezi, M. E. Marin, J. A. Acosta, H. Acosta, J. Rojas, Herrera R., D. Grigoriadis, E. White, P. J. Conlon, R. A. Maki, and A. Zlotnik, "Human endometriosis is associated with plasma cells and overexpression of B lymphocyte stimulator," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 30, pp. 12451–12456, Jul. 2007.

[43] Z. Liu, Z. Yao, C. Li, Y. Lu, and C. Gao, "Gene expression profiling in human high-grade astrocytomas," *Comparative Funct. Genomics*, vol. 2011, Art. no. 245137, Aug. 2011.
[44] H. Runne, A. Kuhn, E. J. Wild, W. Pratyaksha, M. Kristiansen, J. D. Isaacs, E. Régulier, M. Delorenzi, S. J. Tabrizi, and R. Luthi-Carter, "Analysis of potential transcriptomic biomarkers for Huntington's disease in peripheral blood," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 36, pp. 14424–14429, Sep. 2007.
[45] Y. Zhang, M. James, F. A. Middleton, and R. L. Davis, "Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms," *Amer. J. Med. Genet. B, Neuropsychiatric Genet.*, vol. 137, no. 1, pp. 5–16, Aug. 2005.
[46] B. Zheng *et al.*, "PGC-1α, a potential therapeutic target for early intervention in Parkinson's disease," *Sci. Transl. Med.*, vol. 2, no. 52, Oct. 2010, Art. no. 52ra73.
[47] T. A. Wallace *et al.*, "Tumor immunobiological differences in prostate cancer between African-American and European-American men," *Cancer Res.*, vol. 68, no. 3, pp. 927–936, Feb. 2008.
[48] M. E. Lenburg, L. S. Liou, N. P. Gerry, G. M. Frampton, H. T. Cohen, and M. F. Christman, "Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data," *BMC Cancer*, vol. 3, p. 31, Nov. 2003.
[49] Y. Wang, O. Roche, M. S. Yan, G. Finak, A. J. Evans, J. L. Metcalf, B. E. Hast, S. C. Hanna, B. Wondergem, K. A Furge, M. S. Irwin, W. Y. Kim, B. T. Teh, S. Grinstein, M. Park, P. A. Marsden, and M. Ohh, "Regulation of endocytosis via the oxygen-sensing pathway," *Nat. Med.*, vol. 15, no. 3, pp. 319–324, Mar. 2009.
[50] J. D. Zhang and S. Wiemann, "KEGGgraph: A graph approach to KEGG PATHWAY in R and bioconductor," *Bioinformatics*, vol. 25, no. 11, pp. 1470–1471, Jun. 2009.
[51] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *Interjournal Complex Syst.*, vol. 1695, 2006.
[52] Y. Hochberg, "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, Dec. 1988.
[53] S. R. Narum, "Beyond bonferroni: Less conservative analyses for conservation genetics," *Conservation Genet.*, vol. 7, no. 5, pp. 783–787, Oct. 2006.
[54] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: Current approaches and outstanding challenges," *PLoS Comput. Biol.*, vol. 8, Feb. 2012, Art. no. e1002375.
[55] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," *Ann. Appl. Statist.*, vol. 1, no. 1, pp. 107–129, Jun. 2006.
[56] L. Geistlinger, G. Csaba, and R. Zimmer, "Bioconductor's enrichmentbrowser: Seamless navigation through combined results of set-& network-based enrichment analysis," *BMC Bioinformatics*, vol. 17, p. 45, Jan. 2016.
[57] J. Michaud, K. M. Simpson, R. Escher, K. Buchet-Poyau, T. Beissbarth, C. Carmichael, M. E., Ritchie F. Schütz, P. Cannon, M. Liu, X. Shen, Y. Ito, W. H. Raskind, M. S. Horwitz, M. Osato, D. R. Turner, T. P. Speed, M. Kavallaris, G. K. Smyth, and H. S. Scott, "Integrative analysis of RUNX1 downstream pathways and target genes," *BMC Genomics*, vol. 9, p. 363, Jul. 2008.
[58] C. Voichita, M. Donato, and S. Drăghici, "Incorporating gene significance in the impact analysis of signaling pathways," in *Proc. 11th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2012, pp. 126–131.
[59] J. Ungerback, N. Elander, J. Grünberg, M. Sigvardsson, and P. Söderkvist, "The Notch-2 gene is regulated by Wnt signaling in cultured colorectal cancer cells," *PLOS ONE*, vol. 6, Mar. 2011, Art. no. e17957.
[60] M. Heffler, V. M. Golubovskaya, K. M. Dunn, and W. Cance, "Focal adhesion kinase autophosphorylation inhibition decreases colon cancer cell growth and enhances the efficacy of chemotherapy," *Cancer Biol. Ther.*, vol. 14, no. 8, pp. 761–772, Aug. 2013.
[61] G. Chatel, C. Ganeff, N. Boussif, L. Delacroix, A. Briquet, G. Nolens, and R. Winkler, "Hedgehog signaling pathway is inactive in colorectal cancer cell lines," *Int. J. Cancer*, vol. 121, no. 12, pp. 2622–2627, Dec. 2007.
[62] X. Wang, D. Sun, J. Tai, S. Chen, M. Yu, D. Ren, and L. Wang, "TFAP2C promotes stemness and chemotherapeutic resistance in colorectal cancer via inactivating hippo signaling pathway," *J. Exp. Clin. Cancer Res.*, vol. 37, no. 1, p. 27, Feb. 2018.
[63] J. Qiao, C. Y. Fang, S. X. Chen, X. Q. Wang, S. J. Cui, X. H. Liu, Y. H. Jiang, J. Wang, Y. Zhang, P. Y. Yang, and F. Liu, "Stroma derived COL6A3 is a potential prognosis marker of colorectal carcinoma revealed by quantitative proteomics," *Oncotarget*, vol. 6, pp. 29929–29946, Oct. 2015.
[64] M. Duman-Scheel, "Deleted in colorectal cancer (DCC) pathfinding: Axon guidance gene finally turned tumor suppressor," *Current Drug Targets*, vol. 13, no. 11, pp. 1445–1453, Oct. 2012.
[65] J. L. Ku, S. H. Park, K. A. Yoon, Y. K. Shin, K. H. Kim, J. S. Choi, H. C. Kang, I. J. Kim, I. O. Han, and J. G. Park, "Genetic alterations of the TGF-beta signaling pathway in colorectal cancer cell lines: A novel mutation in Smad3 associated with the inactivation of TGF-beta-induced transcriptional activation," *Cancer Lett.*, vol. 247, no. 2, pp. 283–292, Mar. 2007.
[66] Z. Yang, S. Liu, M. Zhu, H. Zhang, J. Wang, Q. Xu, K. Lin, X. Zhou, M. Tao, and C. Li, "PS341 inhibits hepatocellular and colorectal cancer cells through the FOXO3/CTNNB1 signaling pathway," *Sci. Rep.*, vol. 6, Feb. 2016, Art. no. 22090.
[67] G. P. Marszalowicz, A. E. Snook, M. S. Magee, D. Merlino, L. D. Berman-Booty, and S. A. Waldman, "GUCY2C lysosomotropic endocytosis delivers immunotoxin therapy to metastatic colorectal cancer," *Oncotarget*, vol. 5, no. 19, pp. 9460–9471, Oct. 2014.
[68] L. L. Marchand, "The role of chemical carcinogens and their biotransformation in colorectal cancer," in *Genetics of Colorectal Cancer*, J. D. Potter and N. M. Lindor Eds. New York, NY, USA: Springer, 2009, pp. 261–276.
[69] A. K. Coutinho, G. Prolla, and W. Rui, "BRAF, KRAS, and Phosphatidylinositol 3-kinase in the management of metastatic colorectal cancer," *Current Colorectal Cancer Rep.*, vol. 9, no. 1, pp. 57–67, Mar. 2013.
[70] J. P. Spano, G. Milano, C. Rixe, and R. Fagard, "JAK/STAT signalling pathway in colorectal cancer: A new biological target with therapeutic implications," *Eur. J. Cancer*, vol. 42, no. 16, pp. 2668–2670, Nov. 2006.
[71] S. M. Johnson, X. Wang, and B. M. Evers, "Triptolide inhibits proliferation and migration of colon cancer cells by inhibition of cell cycle regulators and cytokine receptors," *J. Surgical Res.*, vol. 168, no. 2, pp. 197–205, Jun. 2011.
[72] A. Dallol, D. Morton, E. R. Maher, and F. Latif, "SLIT2 axon guidance molecule is frequently inactivated in colorectal cancer and suppresses growth of colorectal carcinoma cells," *Cancer Res.*, vol. 63, no. 5, pp. 1054–1058, Mar. 2003.
[73] X. Zhang, G. Liu, L. Ding, T. Jiang, S. Shao, Y. Gao, and Y. Lu, "HOXA3 promotes tumor growth of human colon cancer through activating EGFR/Ras/Raf/MEK/ERK signaling pathway," *J. Cellular Biochem.*, vol. 119, no. 3, pp. 2864–2874, Mar. 2017.
[74] U. K. Roy, N. S. Rial, K. L. Kachel, and E. W. Gerner, "Activated K-RAS increases polyamine uptake in human colon cancer cells through modulation of caveolar endocytosis," *Mol. Carcinogenesis*, vol. 47, no. 7, pp. 538–553, Jul. 2010.
[75] V. Grossi, A. Peserico, T. Tezil, and C. Simone, "p38alpha MAPK pathway: A key factor in colorectal cancer therapy and chemoresistance," *World J. Gastroenterol.*, vol. 20, no. 29, pp. 9744–9758, Aug. 2014.
[76] T. Zhang, Y. Ma, J. Fang, C. Liu, and L. Chen, "A deregulated PI3K-AKT signaling pathway in patients with colorectal cancer," *J. Gastrointest. Cancer*, vol. 50, no. 1, pp. 35–41, Mar. 2017.
[77] E. A. Jansson, A. Are, G. Greicius, I. C. Kuo, D. Kelly, V. Arulampalam, and S. Pettersson "The Wnt/beta-catenin signaling pathway targets PPARgamma activity in colon cancer cells," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 5, pp. 1460–1465, Feb. 2005.
[78] A. Albasri, W. Fadhil, J. H. Scholefield, L. G. Durrant, and M. Ilyas, "Nuclear expression of phosphorylated focal adhesion kinase is associated with poor prognosis in human colorectal cancer," *Anticancer Res.*, vol. 34, pp. 3969–3974, Aug. 2014.
[79] B. L. Adamsen, K. L. Kravik, O. P. Clausen, and P. M. De Angelis, "Apoptosis, cell cycle progression and gene expression in TP53-depleted HCT116 colon cancer cells in response to short-term 5-fluorouracil treatment," *Int. J. Oncol*, vol. 31, no. 6, pp. 1491–1500, Dec. 2007.
[80] R. C. Bruntz, C. W. Lindsley, and H. A. Brown, "Phospholipase D signaling pathways and phosphatidic acid as therapeutic targets in cancer," *Pharmacological Rev.*, vol. 66, no. 4, pp. 1033–1079, Oct. 2014.
[81] V. Catalano, M. Dentice, R. Ambrosio, C. Luongo, R. Carollo, A. Benfante, M. Todaro, G. Stassi, and D. Salvatore, "Activated thyroid hormone promotes differentiation and chemotherapeutic sensitization of colorectal cancer stem cells by regulating Wnt and BMP4 signaling," *Cancer Res.*, vol. 76, no. 5, pp. 1237–1244, Mar. 2016.
[82] Y. Pan, J. H. M. Tong, R. W. M. Lung, W. Kang, J. S. H. Kwan, W. P. Chak, K. Y. Tin, L. Y. Chung, F. Wu, S. S. M. Ng, T. W. C. Mak, J. Yu, K. W. Lo, A. W. H. Chan, and K. F. To, "RASAL2 promotes tumor progression through LATS2/YAP1 axis of hippo signaling pathway in colorectal cancer," *Mol. Cancer*, vol. 17, p. 102, Jul. 2018.

[83] H. Guo, X. Hu, S. Ge, G. Qian, and J. Zhang, "Regulation of RAP1B by miR-139 suppresses human colorectal carcinoma cell proliferation," *Int. J. Biochem. Cell Biol.*, vol. 44, no. 9, pp. 1465–1472, Sep. 2012.
[84] X. T. Ma *et al.*, "Role of chemokine receptor CXCR4/CXCL12 signaling pathway in hepatic metastasis of colorectal carcinoma," *World Chin. J. Digestol.*, vol. 14, no. 16, pp. 1566–1570, 2006.

ZHENSHEN BAO received the B.S. degree from the Nanjing University of Information Science and Technology, Nanjing, Jiangsu, China, in 2013, and the M.S. degree from Wenzhou University, Wenzhou, Zhejiang, China, in 2016. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Bioelectronics, School of Biological Sciences and Medical Engineering, Southeast University, Nanjing.

His research interests include gene set enrichment analysis method, regulatory networks, and miRNA

YIHUA ZHU is currently an Associate Professor with the College of Information Science and Technology, Nanjing Agricultural University, Nanjing, Jiangsu, China. He has presided over the completion of many science foundations. He has participated in many other foundations. He has published many related papers. His research interests include network information organization, bioinformatics analysis, the technology of digital library, and agricultural informationization.

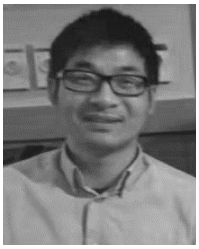
QINYU GE was born in 1978. He received the Ph.D. degree in biomedical engineering from Southeast University, Nanjing, Jiangsu, China, in 2006.

He is currently a Doctoral Supervisor with the State Key Laboratory of Bioelectronics, School of Biological Sciences and Medical Engineering, Southeast University. His research interests include high throughput sequencing sample preparation technology, maternal peripheral blood, tumor circulating nucleic acid detection technology, and site-specific labeling and conformational fluctuation of active proteins. He has presided over the completion of two National Natural Science Foundations of China. He has published more than 60 related SCI papers.

WANJUN GU received the Ph.D. degree in biomedical engineering from Southeast University, Nanjing, Jiangsu, China, in 2004.

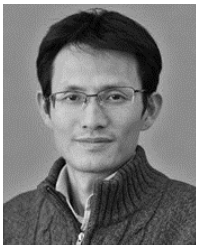
He is currently a Doctoral Supervisor with the State Key Laboratory of Bioelectronics, School of Biological Sciences and Medical Engineering, Southeast University. He has been involved in the research work in bioinformatics and computational biology, until 2000. He has completed a number of national natural science foundation projects of china and a sub-project of the Ministry of Science and Technology (863 Project) as the Project Leader.

Dr. Gu is an Editor of *Evolutionary Bioinformatics* and the *Hans Journal of Computational Biology*.



XIANJUN DONG received the B.S. and M.S. degrees from Southeast University, in 2002 and 2005, respectively, and the Ph.D. degree from the University of Bergen.

He was a Postdoctoral Fellow with the University of Massachusetts Medical School, USA, from 2010 to 2013. He is involved in scientific research at Brigham and Women's Hospital and the Harvard Medical School. He is currently an Instructor in neurology at the Harvard Medical School and the Director of computational neuroscience at the Precision Neurology Program, Brigham and Women's Hospital. He published more than 20 high-level papers in *Nature*, *Nature Neuroscience*, *Science*, *Cell*, *Genome Biology*, *Genome Research*, and other internationally renowned journals, and the cumulative number of references exceeded 10 000.



YUNFEI BAI received the B.S. and M.S. degrees from Nanjing Agricultural University, Nanjing Jiangsu, China, in 1998 and 2001, respectively, and the Ph.D. degree from Southeast University, in 2005.

He was a Lecturer with Southeast University, from 2005 to 2007, where he is currently an Associate Professor. He is in charge of the National Natural Fund Project and the 973 Project. He holds four patents. He has published 15 SCI papers. His research interests include practical biochip technology, double-stranded DNA microarray chips, DNA binding protein, high throughput DNA sequencing technology, chip sequencing, MicroRNA, and expression analysis of mRNA.