Identifying disease-associated signaling pathways through a novel effector gene analysis

Zhenshen Bao[1], Bing Zhang[1], Li Li[2], Qinyu Ge[1], Wanjun Gu[1] and Yunfei Bai[1]

[1] State Key Laboratory of Bioelectronics, School of Biological Sciences and Medical Engineering, Southeast University, Nanjing, Jiangsu, China
[2] Department of Respiratory Medicine, Zhongda Hospital, School of Medicine, Southeast University, Nanjing, Jiangsu, China

ABSTRACT

Background: Signaling pathway analysis methods are commonly used to explain biological behaviors of disease cells. Effector genes typically decide functional attributes (associated with biological behaviors of disease cells) by abnormal signals they received. The signals that the effector genes receive can be quite different in normal vs. disease conditions. However, most of current signaling pathway analysis methods do not take these signal variations into consideration.
Methods: In this study, we developed a novel signaling pathway analysis method called signaling pathway functional attributes analysis (SPFA) method. This method analyzes the signal variations that effector genes received between two conditions (normal and disease) in different signaling pathways.
Results: We compared the SPFA method to seven other methods across 33 Gene Expression Omnibus datasets using three measurements: the median rank of target pathways, the median *p*-value of target pathways, and the percentages of significant pathways. The results confirmed that SPFA was the top-ranking method in terms of median rank of target pathways and the fourth best method in terms of median *p*-value of target pathways. SPFA's percentage of significant pathways was modest, indicating a good false positive rate and false negative rate. Overall, SPFA was comparable to the other methods. Our results also suggested that the signal variations calculated by SPFA could help identify abnormal functional attributes and parts of pathways. The SPFA R code and functions can be accessed at https://github.com/ZhenshenBao/SPFA.

Subjects Bioinformatics, Cell Biology, Computational Biology, Computational Science
Keywords Signaling pathway analysis, Functional attributes, Cell behaviors, SPFA, Effector genes

INTRODUCTION

Recently developed high-throughput functional genomics technologies have generated large amounts of experimental disease data and detected new biological information. Challenge for biologists is understanding the biological behaviors of disease cells using both newly generated disease data and existing biological knowledge. Signaling pathway analysis methods are used to better understand the biological behaviors of disease cells.

Submitted 19 February 2020
Accepted 20 July 2020
Published 14 August 2020

Corresponding author
Yunfei Bai, whitecf@seu.edu.cn

Academic editor
Jun Chen

Additional Information and Declarations can be found on page 20

DOI 10.7717/peerj.9695

Copyright
2020 Bao et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

How to cite this article Bao Z, Zhang B, Li L, Ge Q, Gu W, Bai Y. 2020. Identifying disease-associated signaling pathways through a novel effector gene analysis. **PeerJ 8:e9695 DOI 10.7717/peerj.9695**

The understanding of biological behaviors of disease cells benefits to understand the pathological scenery and treatment. Over-representation analysis (ORA) based methods were initially presented as signaling pathway analysis methods to help biologists identify over-represented pathways from a list of relevant genes produced from experimental data. ORA-based methods merely count the number of differentially expressed genes in specific functional category gene sets such as the Gene Ontology (GO) (*Blake et al., 2013*), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (*Kanehisa et al., 2016*), BioCarta (*Nishimura, 2001*), and Reactome (*Joshitope et al., 2005*). Then they determine significance of the overlaps via statistical tests such as Fisher's exact test. Many tools are based on this method including Onto-Express (*Draghici et al., 2003*; *Khatri et al., 2002*), Fisher (*Khatri, Sirota & Butte, 2012*), and the Gene Ontology Enrichment Analysis Software Toolkit (GOEAST) (*Zheng & Wang, 2008*). However, ORA-based methods only take into account large changes in individual genes that significantly affect pathways and they do not account for smaller changes in sets of functionally-related genes (i.e., pathways) capable of significant effects. Functional class scoring (FCS) based methods have been used to avoid this limitation of ORA-based methods. FCS-based methods take into account the coordinated gene expression changes in pathways, such as gene set enrichment analysis (GSEA) (*Subramanian et al., 2005*), gene set analysis (GSA) (*Efron & Tibshirani, 2006*), and mean-rank gene set enrichment tests (MRGSE) (*Liu et al., 2008*). However, ORA-based and FCS-based methods are both limited because they do not take into account the complex interactions between genes or the complex topology of pathways. To overcome this limitation, pathway-topology-based methods were proposed. Pathway-topology-based methods integrate the complex interactions between genes using pathway topology information, specifically KEGG signaling pathway information.

Signaling pathway impact analysis (SPIA), one of the most widely-used pathway-topology-based methods, considers both the number of differentially expressed genes (DEGs) in a given pathway and the topology information of that pathway (*Tarca et al., 2009*). Many improved methods based on SPIA have been proposed. *Li et al. (2015)* developed a method called sub-SPIA, which used a minimum spanning tree way to prune signaling pathways and used the SPIA method to identify significant signaling subpathways (*Li et al., 2015*). *Bao et al. (2016)* developed two SPIA-based methods called PSPIA and MSPIA. These two methods replaced +1 or −1 interaction strength in SPIA with the interaction strength of the Pearson correlation coefficients and mutual information, respectively (*Bao et al., 2016*). There are different pathway-topology methods that make use of the topological information of signaling pathways. For instance, Gene Graph Enrichment Analysis (GGEA) uses prior knowledge derived from directed gene regulatory networks (*Geistlinger et al., 2011*). *Liu, Xu & Bao (2019)* used a subgraph method to take advantage of pathway topological information (*Liu, Xu & Bao, 2019*). ROntoTools introduced a term of perturbation factor by considering the type of interactions to take the pathway topology into consideration (*Tarca et al., 2009*;

Voichita, Donato & Draghici, 2012). Sebastian-Leon et al. (2014) developed a method using topology to detect liner subpathways in a signaling pathway (Sebastian-Leon et al., 2014).

These methods still have their disadvantages. Pathway-topology-based methods do not consider the importance of genes in pathways. Gene-weight-based methods have been proposed to overcome this limitation. Pathway analysis with down-weighting of overlapping genes (PADOG) uses the frequency of a present gene in the analyzed pathways to improve gene set analysis (Tarca et al., 2012). Functional link enrichment of gene ontology or gene sets (LEGO) measures gene weights in a gene set according to its relative association with genes inside and outside the gene set in a functional association network (Dong et al., 2016). Fang et al. (2017) proposed an improved SPIA method called SPIA-IS that measured and assigned the importance as the average output degree of the gene in the pathway.

A signaling pathway is a cascade of molecular reactions that bring out the functional attributes (e.g., cell proliferation, apoptosis) associated with the biological behaviors of disease cells using effector genes. Effector genes receive signals without outputting signals to other genes in an individual signaling pathway (Sebastian-Leon et al., 2014). Diseases are always related to the abnormal signal that the effector genes receive. Therefore, the signal that the effector genes receive can be very different under disease and normal conditions. The limitation of the previously mentioned methods, including gene-weight-based methods, is that they do not consider the signal variations between disease and normal conditions.

Additionally, the functional attributes in the same signaling pathway may be very different from one another, and can sometimes be opposites. For example, there are two opposite functional attributes on the axon guidance pathway: axon repulsion and axon attraction (see the hsa04360 pathway in the KEGG dataset). We cannot determine which functional attributes contribute more to the disease using most current pathway analysis methods. Furthermore, some pathways consist of several parts, each with very different contributions. For example, the Wnt signaling pathway is significant across different diseases and can be divided into three parts. Most existing pathway analysis methods cannot determine which part of the Wnt signaling pathway most significantly contributes to a specific disease.

We propose a new method that considers the signal variations between normal and disease conditions that effector genes received in pathways: the signaling pathway functional attributes analysis (SPFA) method. SPFA calculates the gene expression changes in a given pathway using an ORA method and then combines the ORA method results with the signal variation results under two conditions (normal vs. disease). The signal variations can help identify functional attributes and abnormal pathways. We tested the capabilities of the proposed signaling pathway analysis method on a series of real datasets using three parameters. We also showed that the two types of probabilities considered in this method were indeed independent. Ultimately, we verified the usefulness of the signal variations the effector genes received under two different conditions using the SPFA method.

MATERIALS AND METHODS

Data sources and preprocessing

Signaling pathway analysis methods require two types of input: a collection of pathways and a list of genes or gene products with accompanying expression values across different samples between the compared phenotypes. We used the KEGG signaling pathway as it is the most common manually-curated signaling pathway used for pathway analysis. We downloaded 213 signaling pathways from the KEGG PATHWAY dataset.

We acquired 33 disease gene expression datasets from the KEGGdzPathwaysGEO R-package and KEGGandMetacoreDzPathwaysGEO R-packages (Table 1) (*Tarca, Bhatti & Romero, 2013*; *Tarca et al., 2012*). Each disease gene expression dataset was matched with a corresponding disease KEGG pathway. For example, a colorectal cancer dataset was associated with the colorectal cancer pathway (*Tarca et al., 2012*). The corresponding disease KEGG pathways were called target pathways. Three rules were used to select the gene expression datasets:

1. The dataset's DEGs were available. If no DEGs were selected, other comparable methods would return null results.
2. The results of these datasets could be analyzed. Pathway analysis result p-values of 1 could not be analyzed.
3. The target pathways of these datasets were KEGG pathways since we used KEGG pathways as examples.

DEGs were selected if they contained more than 200 genes with FDR adjusted p-values < 0.05. Otherwise, we selected more than 200 genes with original p-values < 0.05 and absolute log (fold change) > 1.5. If DEGs still less than 200 genes, we selected the top 1% of genes ranked by p-values as DEGs.

SPFA algorithm design

To assess the signal variations between two conditions (normal vs. disease) that the effector genes received from upstream genes, we calculated the sum of signal variations from all upstream genes to effector genes. Given an effector gene g_e and an upstream gene g_s, the signal variation from the gene g_s to the effector gene g_e can be defined as:

$$e_{se} = \frac{\left| cor^{\text{disease}}(g_s g_e) - cor^{\text{normal}}(g_s g_e) \right|}{d_{se}} \tag{1}$$

where $cor^{\text{disease}}(g_s g_e)$ and $cor^{\text{normal}}(g_s g_e)$ refer to the Pearson correlation coefficient between the gene expression data of gene g_s and gene g_e in the disease and normal states, respectively. d_{se} is the network distance between gene g_s and gene g_e. The Pearson correlation coefficient is always used in gene co-expression networks to represent the strength of interactions between two genes. The Pearson correlation coefficient can also be used to represent the strength of an interaction between two gene expression values. Studies have shown that the genetic regulatory patterns in signaling pathways between

Table 1 Data sets used for assessing the proposed method and compared methods.

ID	Target pathway	GEO ID	References
1	Colorectal cancer	GSE4107	*Hong et al. (2007)*
2	Colorectal cancer	GSE4183	*Galamb et al. (2008)* and *Gyorffy et al. (2009)*
3	Colorectal cancer	GSE8671	*Sabates-Bellver et al. (2007)*
4	Colorectal cancer	GSE9348	*Hong et al. (2010)*
5	Colorectal cancer	GSE23878	*Uddin et al. (2011)*
6	Non-small cell lung cancer	GSE18842	*Sanchez-Palencia et al. (2010)*
7	Pancreatic cancer	GSE15471	*Badea et al. (2008)*
8	Pancreatic cancer	GSE16515	*Pei et al. (2009)*
9	Pancreatic cancer	GSE32676	*Donahue et al. (2012)*
10	Thyroid cancer	GSE3467	*He et al. (2005)*
11	Thyroid cancer	GSE3678	–
12	Alzheimer's disease	GSE5281_HIP	*Liang et al. (2007)*
13	Alzheimer's disease	GSE5281_EC	*Liang et al. (2007)*
14	Alzheimer's disease	GSE5281_VCX	*Liang et al. (2007)*
15	Alzheimer's disease	GSE1297	*Blalock et al. (2004)*
16	Alzheimer's disease	GSE16759	*Juan et al. (2010)*
17	Chronic myeloid leukemia	GSE24739_G0	*Affer et al. (2011)*
18	Chronic myeloid leukemia	GSE24739_G1	*Affer et al. (2011)*
19	Acute myeloid leukemia	GSE14924_CD4	*Le Dieu et al. (2009)*
20	Acute myeloid leukemia	GSE14924_CD8	*Le Dieu et al. (2009)*
21	Acute myeloid leukemia	GSE9476	*Stirewalt et al. (2008)*
22	Dilated cardiomyopathy	GSE1145	–
23	Dilated cardiomyopathy	GSE3585	*Barth et al. (2006)*
24	Endometrial cancer	GSE7305	*Hever et al. (2007)*
25	Glioma	GSE19728	*Liu et al. (2011)*
26	Glioma	GSE21354	*Liu et al. (2011)*
27	Huntington's disease	GSE8762	*Runne et al. (2007)*
28	Parkinson's disease	GSE20291	*Zhang et al. (2005)*
29	Parkinson's disease	GSE20164	*Zheng et al. (2010)*
30	Prostate cancer	GSE6956AA	*Wallace et al. (2008)*
31	Prostate cancer	GSE6956C	*Wallace et al. (2008)*
32	Renal cell carcinoma	GSE781	*Lenburg et al. (2003)*
33	Renal cell carcinoma	GSE14762	*Wang et al. (2009)*

genes are different under normal and disease conditions (*Jung, 2018*). If the genetic regulatory pattern between the two genes changes, the signal transmitted between the two genes will be very different. Thus, we used the Pearson correlation coefficient to calculate the signal variations that the effector genes received from their upstream genes. However, if an upstream gene does not directly transmit a signal, the signal may be attenuated. Therefore, we used the network distance d_{se} between gene g_s and gene g_e as a penalty coefficient.

For each effector gene g_i in a given pathway, the accumulated signal variations between normal and disease conditions that the upstream genes received (total s genes in the upstream of the gene g_i) were calculated using the formula:

$$\mathrm{ASV}(g_i) = \sum_{j=1}^{s} e_{ij} \tag{2}$$

The accumulated signal variation $\mathrm{ASV}(g_i)$ of the effector gene g_i in a pathway can help us distinguish among the functional disease attributes. Effector genes with high $\mathrm{ASV}(g_i)$ demonstrate that these functional attributes significantly contribute to their corresponding diseases.

For a given signaling pathway, the total accumulated signal variation ASV can be defined as:

$$\mathrm{ASV} = \sum_{i=1}^{k} \mathrm{ASV}(g_i) \tag{3}$$

where k is the total number of effector genes in the given pathway.

Ultimately, the probability P_{sd} used to measure the signal variations between two conditions (normal vs. disease) that those effector genes received from genes upstream in a given signaling pathway P_x is based on the pathway's $\mathrm{ASV}(\mathrm{P}_x)$. The same number of genes as the one observed on the given signaling pathway are randomly selected from all genes (random gene IDs) and have any possible expression data in all samples in the range of the experimenter. Therefore, the observed signal variations were obtained by permuting the gene IDs 2000 times. $\mathrm{ASV}_{per}(\mathrm{P}_x)$was the total accumulated signal variation of the given pathway P_x obtained in the *per*-th time. The probability $P_{sd}(\mathrm{P}_x)$ of the given pathway was calculated as:

$$P_{sd}(\mathrm{P}_x) = \frac{\sum I(\mathrm{ASV}_{per}(\mathrm{P}_x) \geq \mathrm{ASV}(\mathrm{P}_x))}{2000} \tag{4}$$

where I is a function that returns 1 when the argument is true and 0 when the argument is false.

The probability P_{sd} does not measure the gene differential expression in a given pathway. Thus, we had to combine the probability P_{sd} with the probability P_{de} which can measure the total gene differential expression in a given signaling pathway. The probability P_{de} of a given pathway P_x can be calculated through the following hypergeometric test:

$$P_{de}(\mathrm{P}_x) = 1 - \frac{\binom{t}{r}\binom{m-t}{n-r}}{\binom{m}{n}} \tag{5}$$

where the whole genome contains a total of m genes, n genes are the number of DEGs in the m genes, and the given pathway contains t genes and r DEGs.

The probability P_{sd} uses the Pearson correlation coefficient of the two genes' expression data, but the probability P_{de} uses the number of DEGs in a pathway. Thus, the two probabilities are independent of each other. The significance of the given pathway was calculated following the SPIA method which combines the probabilities P_{sd} and P_{de} (*Tarca et al., 2009*). The formulas are:

$$P = c - c \cdot \ln(c) \tag{6}$$

$$c = P_{sd} \times P_{de} \tag{7}$$

where c is a product of P_{de} and P_{sd}. P is the combined probability of the signaling pathway.

Significantly enriched pathway analysis using SPFA

The SPFA procedure identifies significantly enriched pathways in two steps (Fig. 1). The first step measures the total gene differential expression in the signaling pathways. DEGs need to first be identified from the gene expression datasets. Then the DEGs are mapped onto the signaling pathways. Finally, the signaling pathway p-values are calculated using a hypergeometric test.

The second step is to measure the signal variations between the two conditions (normal vs. disease) that effector genes received from upstream genes in the signaling pathways. This is completed by:

1. Finding all effector genes in each signaling pathway.
2. Ascertaining all paths from the upstream genes to the effector genes in each signaling pathway. If a path exists between the upstream genes and effector genes, an interaction must exist between them. The path's network distances are used to weight the corresponding interactions.
3. Using the Pearson correlation coefficient absolute difference values between the disease and normal samples to calculate the signal variations of the corresponding interactions.
4. Using the network distance of each interaction to decrease their signal variations.
5. Calculating the accumulation of the signal variations between the effector genes and upstream genes for each effector gene.
6. Calculating the sum of the accumulations of all effector genes in each signaling pathway.
7. Evaluating the statistical significance of each pathway based on their score.

Ultimately, the results of the two steps are combined into one p-value. We used the FDR adjust method on the combined p-value to determine the significance of each signaling pathway. The pathways with the adjusted combined p-values smaller than a threshold value were considered as significant pathways.

The distribution of effector genes in the signaling pathways

Studying the signal variations between two conditions (normal vs. disease) that the effector genes received leads to a deeper understanding of the biological behaviors of disease cells. Effector genes are widely scattered throughout the signaling pathways. If a gene has no

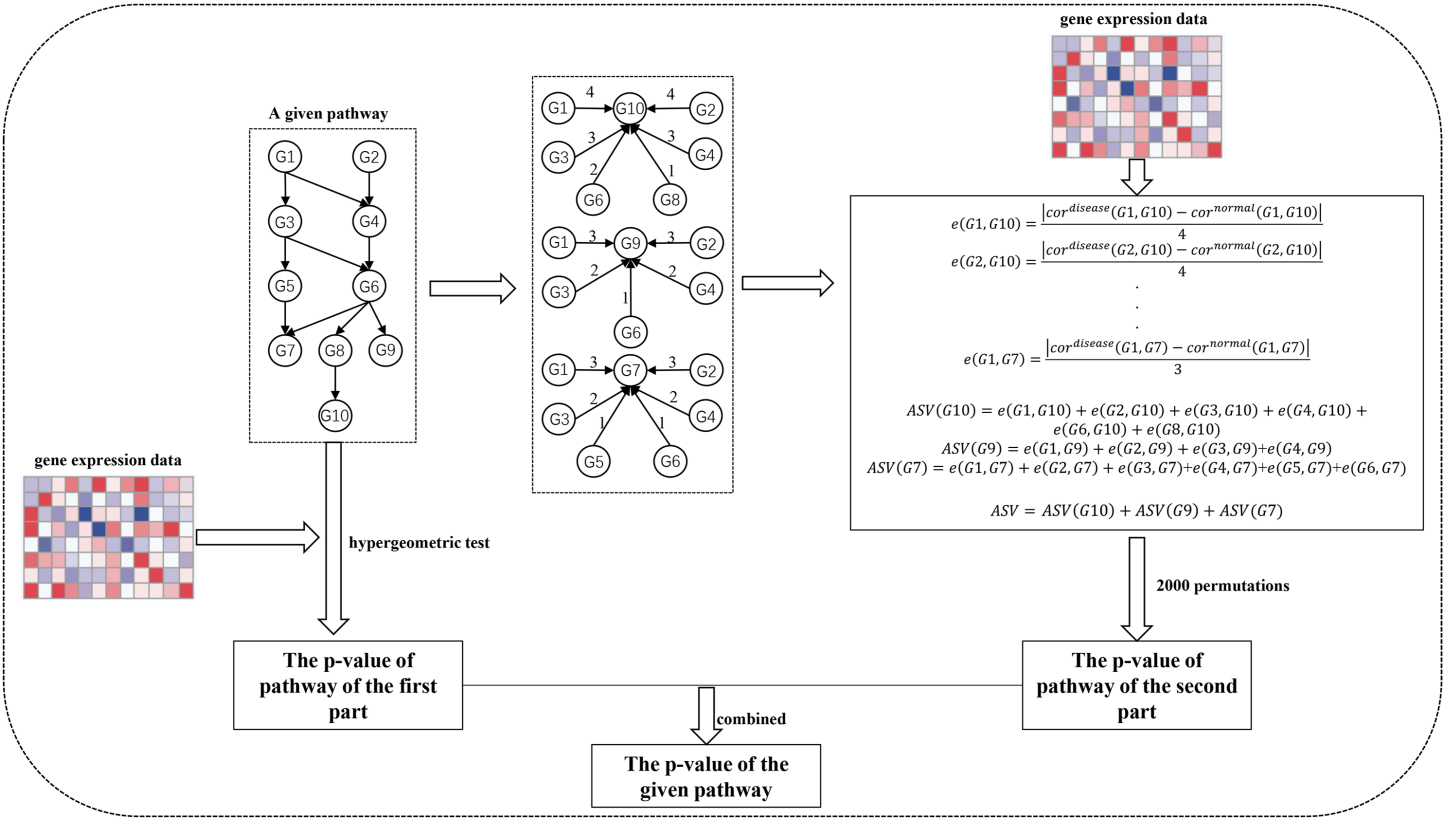

Figure 1 **The workflow of SPFA method.** The step by step to identify significant signaling pathways using SPFA.
Full-size DOI: 10.7717/peerj.9695/fig-1

signal inputs in an individual signaling pathway, the gene is not considered an effector gene. The distribution of effector genes in each signaling pathway can be seen in Fig. 2. One hundred and ninety-five of the 213 signaling pathways contained effector genes.

Comparison methods and measures

We compared seven methods to SPFA, including Fisher (*Khatri, Sirota & Butte, 2012*), GSA (*Efron & Tibshirani, 2006*), GSEA (*Subramanian et al., 2005*), MRGSE (*Liu et al., 2008*), SPIA (*Ullah, 2013*), ROnoTools (*Tarca et al., 2009*; *Voichita, Donato & Draghici, 2012*), and PADOG (*Tarca et al., 2012*). We selected these methods for their stability and prevalence; they can be compared using the same R environment as SPFA.

There is no universally accepted technique for the validation of the results of pathway analysis methods. Current pathway analysis methods use the results of a very small number of datasets based on searching corresponding published life literature. This approach has its limitations. First, the number of datasets used is small. Second, authors often search their own, leading to biased results. Third, complex biological phenomena always directly or indirectly correspond to multiple signaling pathways.

Tarca et al. (2012) compiled an objective and reproducible approach based on multiple datasets (*Tarca et al., 2012*). This approach avoided a biased literature search and required testing on a large number of different datasets (at least 10). In this work, we followed

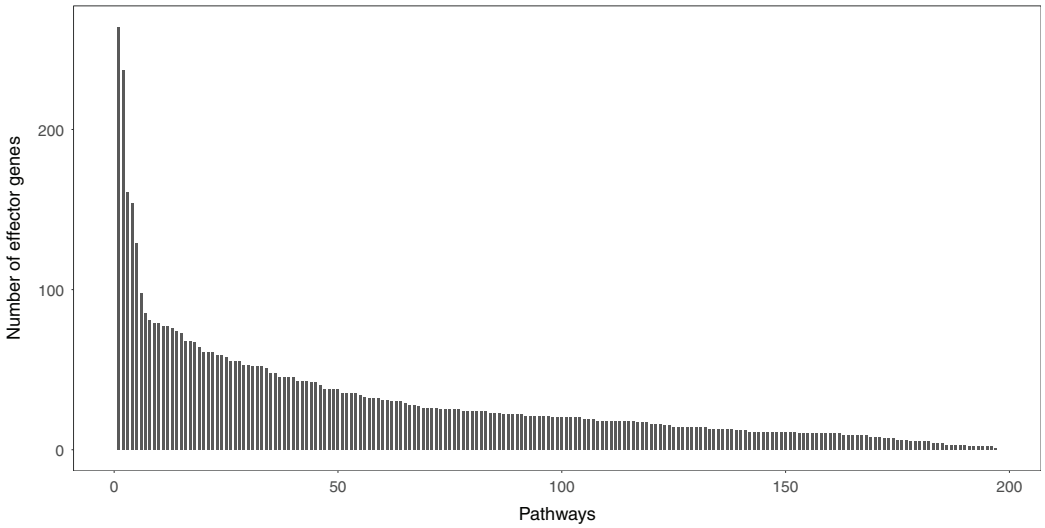

Figure 2 The distribution of the effector genes' number in each signaling pathway. A total of 195 of 213 signaling pathways contain the effector genes. Full-size DOI: 10.7717/peerj.9695/fig-2

Tarca et al. (2012) validation approach. Two measurements were compared in this validation approach. The first measurement was the median *p*-value of the 33 target pathways of the 33 disease datasets. Smaller median *p*-values meant higher sensitivity. The second measurement was the median rank of the 33 disease target pathways. The higher ranked methods were more accurate. To further validate the different pathway analysis method results, we used a third measurement: the ratio of significant pathways (using a significance threshold of 0.05 of the adjusted *p*-value) in the 33 datasets. This measured the method's ability to control false positive and false negative rates.

RESULTS

The independence between the two probabilities

The two probabilities P_{de} and P_{sd} are theoretically independent under the null hypothesis. We verified their independence by calculating the squared correlation coefficient between the two probabilities using the 33 gene expression datasets (Table 2). Our results showed that the average squared correlation coefficient of the 33 datasets was $R^2 = 0.029$. Only four of the 33 squared correlation coefficients were slightly higher than $R^2 = 0.09$. These results indicated essentially no correlation between the two probabilities.

SPFA method performance

We compared SPFA with the other seven methods using three measurements: the median *p*-value of the 33 target pathways, the median rank of the 33 target pathways, and the ratio of significant pathways. The signaling pathways with adjusted *p*-values ≤ 0.05 were significant.

When comparing the median rank of the 33 target pathways, SPFA ranked first (Fig. 3). When comparing the median *p*-value of the 33 target pathways, SPFA ranked fourth

Table 2 The squared correlation coefficients between the two probabilities using the 33 gene expression datasets. The four squared correlation coefficients which are slightly more than 0.09 are shown in bold.

GEO ID	Squared correlation between the probabilities P_{de} and P_{sd}
GSE4107	0.006928102
GSE4183	0.032207913
GSE8671	0.00011503
GSE9348	0.027441819
GSE23878	0.013047606
GSE18842	0.089945631
GSE15471	0.032082501
GSE16515	0.022817456
GSE32676	0.010161372
GSE3467	0.001098836
GSE3678	0.000879454
GSE5281_HIP	0.026379598
GSE5281_EC	0.032472155
GSE5281_VCX	0.063438794
GSE1297	0.000346566
GSE16759	0.028461474
GSE24739_G0	0.009721816
GSE24739_G1	0.022257943
GSE14924_CD4	**0.106127**
GSE14924_CD8	0.051189135
GSE9476	0.073960111
GSE1145	**0.098132151**
GSE3585	6.61523E−05
GSE7305	**0.101902794**
GSE19728	**0.094956883**
GSE21354	0.00854786
GSE8762	0.000830428
GSE20291	0.000499751
GSE20164	7.48134E—07
GSE6956AA	0.006999771
GSE6956C	0.001917359
GSE781	0.000219909
GSE14762	0.000513602
Average	0.029262658

(Fig. 4). Notably, the methods with the highest ranking in one measurement did not necessarily rank the highest in the others. This is because different measurements analyze different abilities. For example, MRGSE was first in median *p*-value but was sixth in median rank. Fisher was second in median *p*-value but ranked fourth in median rank. To better compare SPFA's performance against the other methods, we added the ranks of

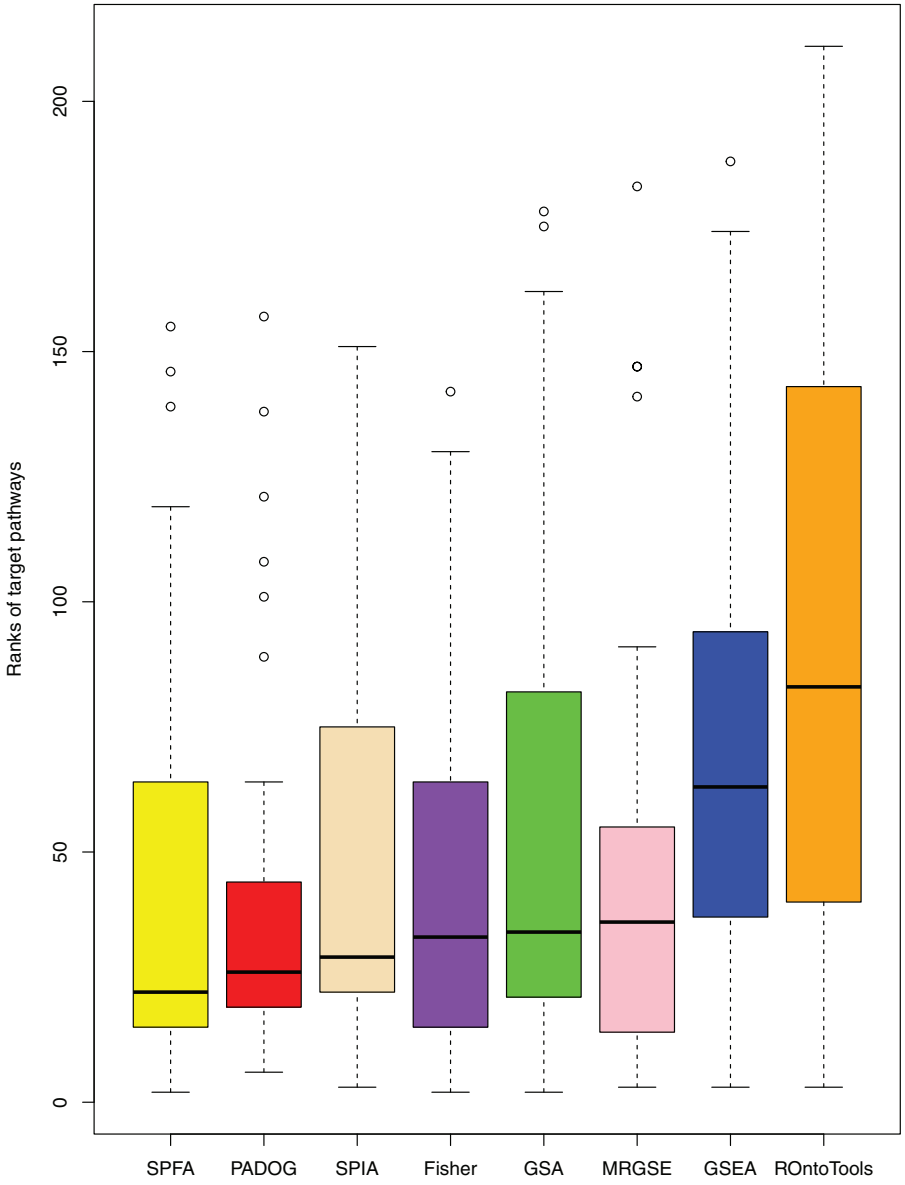

Figure 3 The distribution of the target pathways ranks of the eight methods using 33 datasets. SPFA performs the 1st among eight methods in terms of the median ranks of the 33 target pathways.
Full-size DOI: 10.7717/peerj.9695/fig-3

the median *p*-value and median rank values from each method together. We found that the combined value of SPFA and PADOG was the smallest (Table 3).

To further assess the performance of the eight methods, we collected the results from other general pathways typically associated with cancer using the 18 out of 33 datasets with a form of cancer in Table 4: Apoptosis and Pathways in cancer. When using the Apoptosis pathway and Pathway in cancer pathway instead of target pathways, SPFA's median ranks were both first, and the median *p*-values of MRGSE were also both ranked first. These results were in alignment with the target pathway results. However, when using

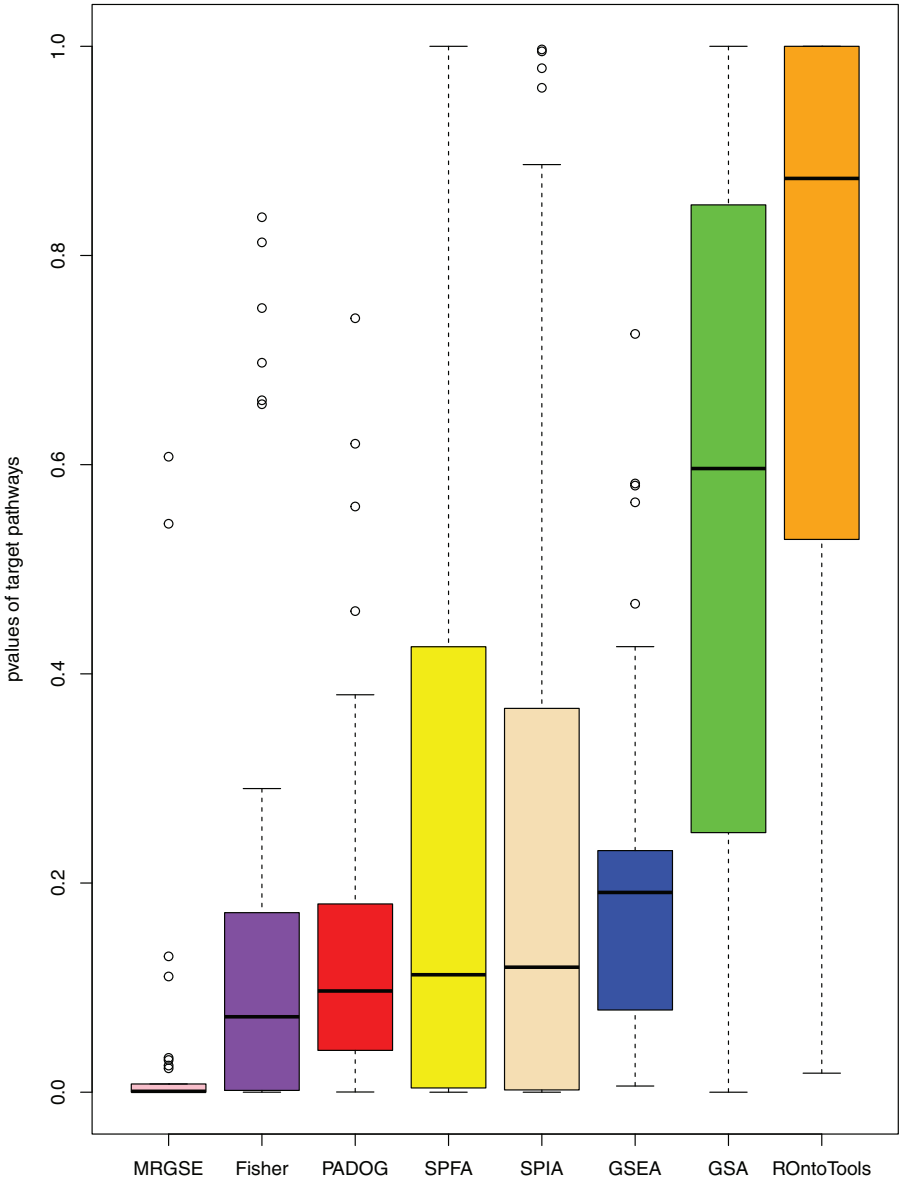

Figure 4 The distribution of the target pathways *p*-values of the eight methods using 33 datasets. SPFA performs the 4th among eight methods in terms of the median *p*-values of detecting the 33 target pathways. Full-size DOI: 10.7717/peerj.9695/fig-4

the Apoptosis pathway and Pathway in cancer pathway instead of the target pathways, PADOG's median *p*-values were both ranked fifth. When using the Apoptosis pathway, SPFA's median *p*-value ranked third. When using the Pathway in cancer pathway, SPFA's median *p*-value ranked fourth. All these results suggest that SPFA had the best accuracy and a good sensitivity when compared with the other seven methods.

Additionally, our results showed that SPFA's ratio of significant pathways was moderate, 0.16 (Fig. 5), compared to the others. MRGSE's ratio of significant pathways was almost 0.5, and it could be questioned whether a such number of pathways was realistic.

Table 3 **The combined rank values of the ranks in terms of the median *p*-values and the median ranks of target pathways of eight methods.**

	Methods	Ranks of the median *p*-values	Ranks of the median ranks	Sum
1	SPFA	4	1	5
2	PADOG	3	2	5
3	Fisher	2	4	6
4	MRGSE	1	6	7
5	SPIA	5	3	8
6	GSA	7	5	12
7	GSEA	6	7	13
8	ROnoTools	8	8	16

Table 4 **The results of other general pathways: apoptosis and pathway in cancer typically associated with cancer using the 18 out of 33 datasets with a form of cancer.** For each pathway, the values for the type of methods with the smallest median *p*-values and ranks (strongest association with the phenotype) are shown in bold.

Pathway statistic	Apoptosis		Pathway in cancer	
	p-Values median	Ranks median	*p*-Values median	Ranks median
SPFA	0.0658	**39.5**	7.94E−05	**3**
Fisher	0.0235	46	2.25E−05	4
SPIA	0.0661	53	1.62E−05	5
GSA	0.779	125	0.539	44.5
GSEA	0.393	116.5	0.291	102
MRGSE	**0.00213**	46	**2.7E−08**	**3**
RontoTools	0.647	70.5	1	210
PADOG	0.26	71	0.09	24

GSA's ratio of significant pathways was lower than 0.05, and it reflected that the GSA method had a high false negative rate. The methods had a modest ratio of significant pathways indicated that the method had a modest false positive rate and a modest false negative rate. Thus, the discriminative ability of SPFA was good when compared with the other seven methods. In conclusion, our results strongly supported that SPFA was well-suited for signaling pathway analysis and confirmed previously reported results in *Dong et al. (2016)*.

Sources of improvement for SPFA

The main source of improvement in SPFA is that it uses signal variations that effector genes received under normal and disease conditions. SPFA is compared to the simpler ORA-based method used to calculate the probability P_{de} without accounting for signal variations (Fig. 6). As shown in Fig. 6, the ORA-based method has a higher (worse) rank and *p*-value than SPFA for the target pathways.

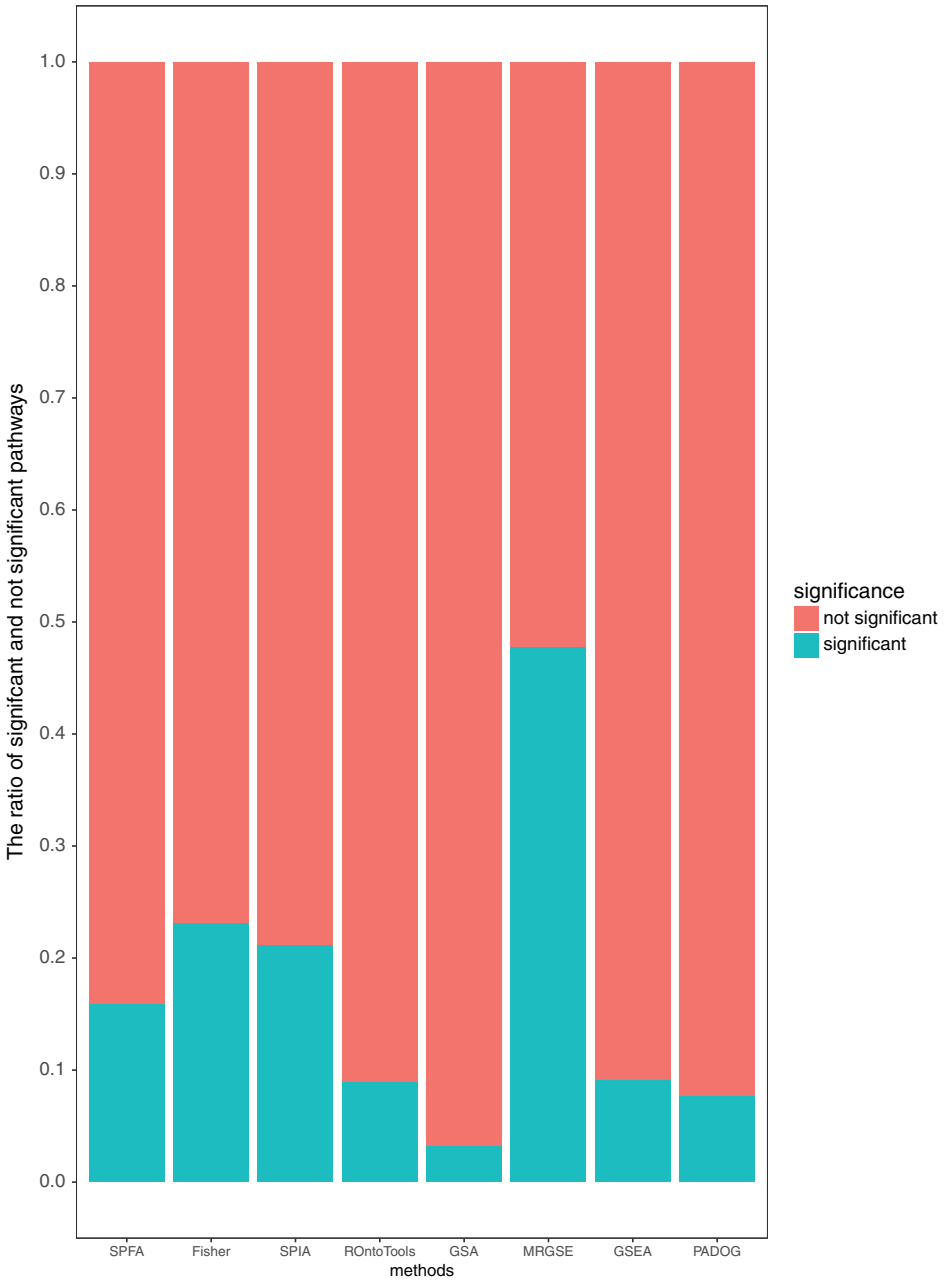

Figure 5 **Average percentage of the pathways detected as significant and not significant by each method using the threshold of *p*-values ≤ 0.05.** Full-size DOI: 10.7717/peerj.9695/fig-5

Validating the correlation between diseases and the signal variations that effector genes received under two different conditions

To validate the correlation between diseases and the signal variations that effector genes received under two different conditions (normal vs. disease), we analyzed a colorectal cancer dataset (GSE4183) and an Alzheimer's disease dataset (GSE16759). The colorectal cancer microarray GSE4183 (Affymetrix array HG-U133 Plus2.0) included 15 colorectal cancer samples and 8 normal samples (*Galamb et al., 2008*; *Gyorffy et al., 2009*).

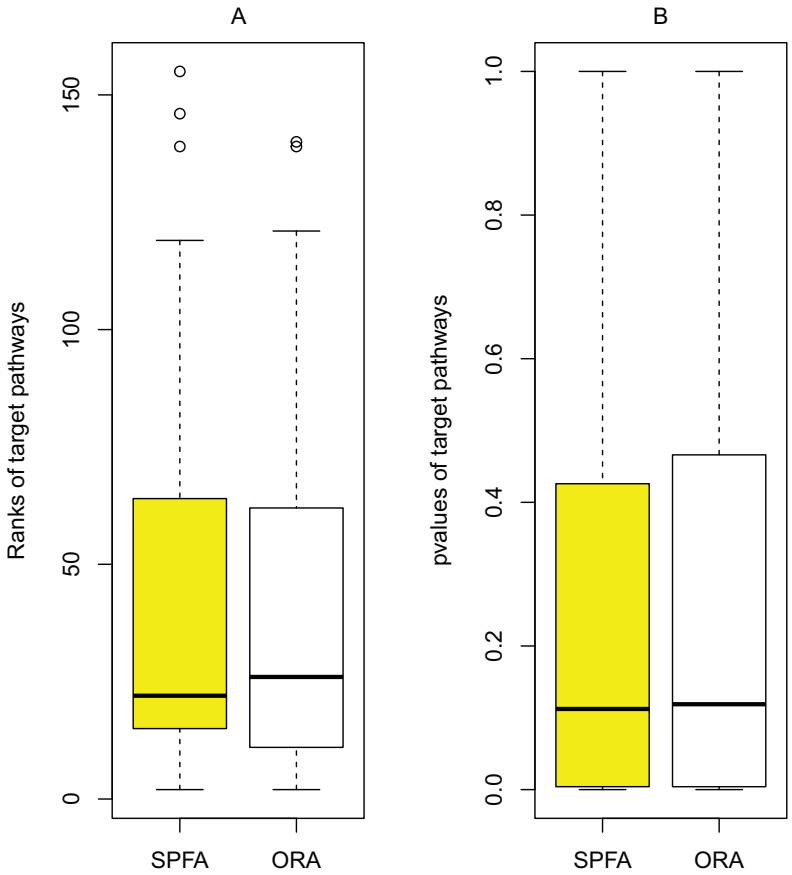

Figure 6 Determining the contribution of signal variations received by effector genes between two different conditions (normal vs. disease) in SPFA performance. The boxplots show the distribution of the target pathways ranks (A) and *p*-values (B). Full-size DOI: 10.7717/peerj.9695/fig-6

The Alzheimer's disease dataset GSE16759 included four disease samples and four normal samples (*Juan et al., 2010*).

The Wnt signaling pathway was altered in 90% of the colorectal cancer samples (*Galamb et al., 2008*). We assessed the signal variations that effector genes received in the Wnt signaling pathway using the GSE4183 dataset (Fig. 7). The results of (*Galamb et al., 2008*) coincided with our signal variation results (*Galamb et al., 2008*) reported that overexpression of TNS1 could induce the activation of JNK (ENTREZID: 5599, 5601, and 5602). The signal variation that the effector gene ENTREZID: 5602 received ranked first in our results. *Galamb et al. (2008)* detected that RBMS1 is another overexpressed gene and modulator of c-myc (ENTREZID: 4609). c-myc can regulate cell cycles and cause cells to transform pathways. The signal variation that the effector gene ENTREZID: 4609 received ranked second in our results. *Galamb et al. (2008)* also identified that TCF4 is an overexpressed gene that can participate in the transcriptional regulation of genes associated with colon carcinogenesis. These colon carcinogenesis associated genes include c-myc (ENTREZID: 4609), cy-clin D1 (ENTREZID: 595), PPARδ (ENTREZID: 5467), and MMP7 (ENTREZID: 4316). The signal variations that these effector genes received ranked second, fourth, fifth, and sixth, respectively.

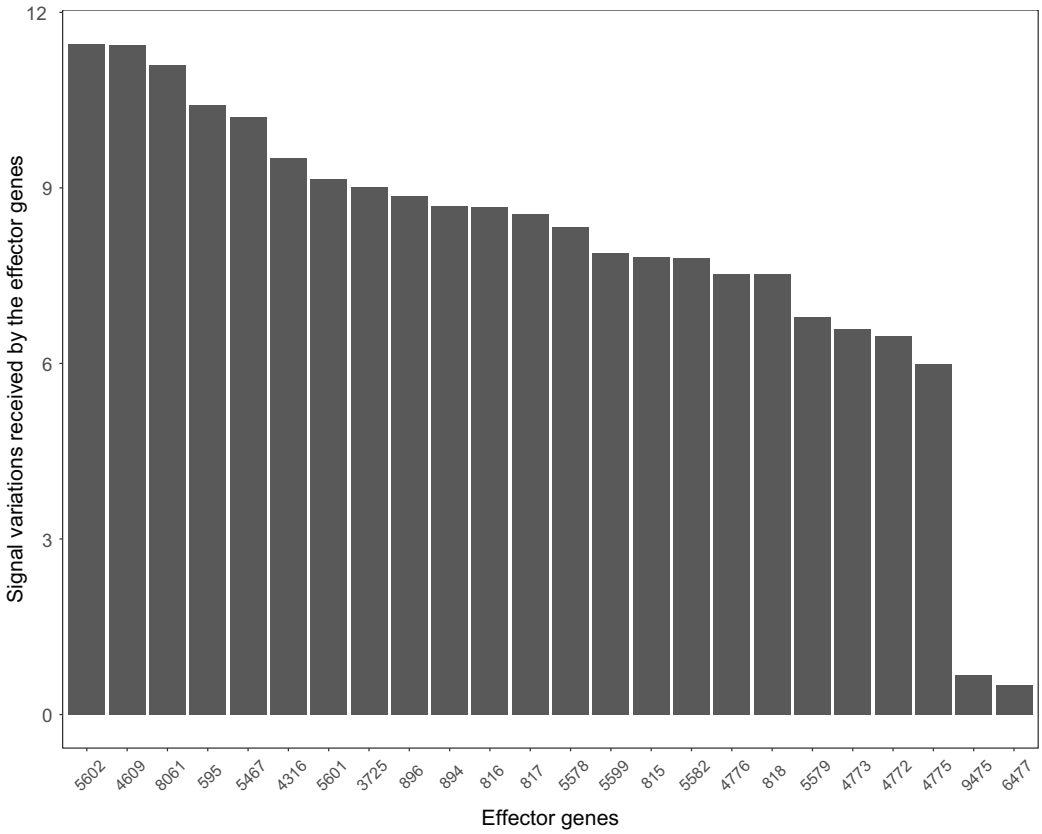

Figure 7 The signal variations received by effector genes from the upstream genes in the Wnt signaling pathway using colorectal cancer datasets (GSE4183).

Full-size DOI: 10.7717/peerj.9695/fig-7

Many pathways can be studied in colorectal cancer datasets. For example, the PI3K-Akt signaling pathway plays a critical role in the growth and progression of colorectal cancer (*Johnson et al., 2010*). The effector genes ENTREZID:596, ENTREZID:842, and ENTREZID:1027 have the highest signal variations and are linked to cell cycle progression and cell survival (Fig. 8). The GSE4183 dataset results further confirmed the role of this pathway in colorectal cancer development.

The Wnt signaling pathway is also closely related to the occurrence and development of Alzheimer's disease (*Inestrosa et al., 2007*). The signal variations that different effector genes received calculating based on the Alzheimer's disease dataset GSE16759 in the Wnt signaling pathway were shown in Fig. 9. The signal variations that the effector genes: ENTREZID: 595 and 896 received were considerably higher than the other effector genes in the Wnt signaling pathway. This result validated evidence of crosstalk between the Alzheimer's disease signaling pathway and the two effector genes' upstream genes in the Wnt signaling pathway.

All these results indicated the high correlation between diseases and the signal variations calculated using the SPFA method.

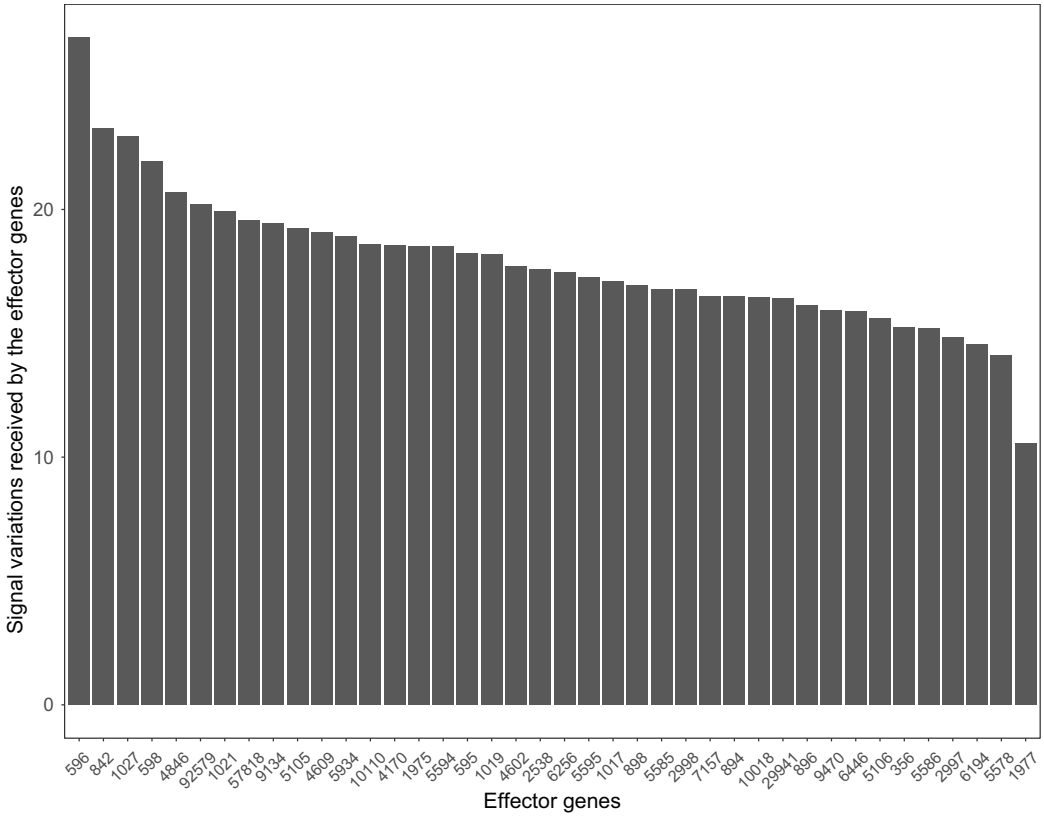

Figure 8 **The signal variations received by effector genes from the upstream genes in the PI3K-Akt signaling pathway using colorectal cancer datasets (GSE4183).**

Full-size DOI: 10.7717/peerj.9695/fig-8

The other usages of the signal variations that effector genes received under two different conditions

The signal variations that effector genes received under two different conditions can show the different contributions of different functional attributes contributed to their corresponding diseases. We can also identify which parts of the pathway contribute to their corresponding diseases through the signal variations that effector genes received.

When looking at the Wnt signaling pathway results of GSE4183 (Fig. 7), first, we know the functional attributes participating in the cell cycle have abnormal signal variations because most effector genes with high signal variations participate in the pathway cell cycle (including c-myc (ENTREZID: 4609), cy-clin D1 (ENTREZID: 595, 894, and 896), PPARδ (ENTREZID: 5467), and MMP7 (ENTREZID: 4316)). Second, we can know that the abnormal state of the first and second parts of the Wnt signaling pathway may contribute more to colorectal cancer because that the effector genes with high signal variations are all in the two parts. If we were only to observe DEG distribution in the Wnt signaling pathway using the GSE4183 dataset, we would not know which abnormal part contributed to the disease (Fig. 10). Through the result of the Wnt signaling pathway in GSE16759 (Fig. 9), on one hand, according to this result, we can know that the functional attributes linked with the effector genes: ENTREZID: 595 and 896 which had the highest signal

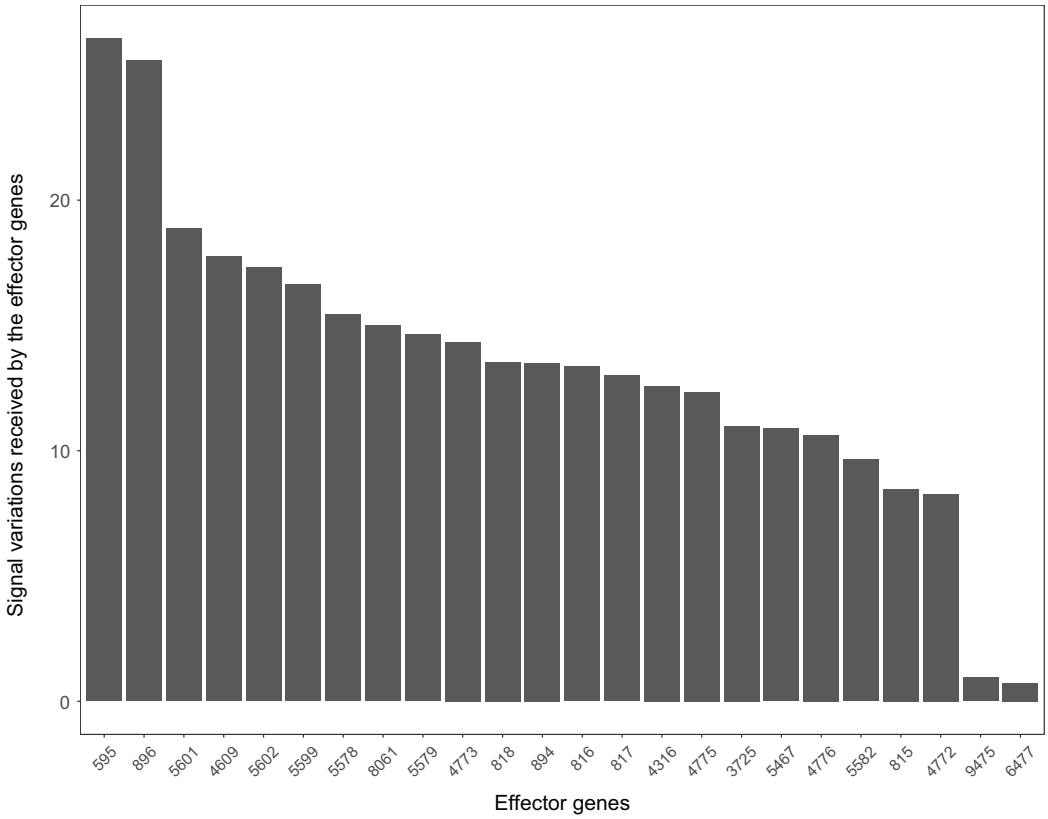

Figure 9 The signal variations received by effector genes from the upstream genes in the Wnt signaling pathway using Alzheimer's disease datasets (GSE16759).

Full-size DOI: 10.7717/peerj.9695/fig-9

variations were abnormal in Alzheimer's disease. On the other hand, this may dominate that the first part of the Wnt signaling pathway may be more related to the occurrence and development of Alzheimer's disease because of crosstalk between the Alzheimer's disease pathway and the first part of the Wnt signaling pathway contained the two effector genes: ENTREZID: 595 and 896.

DISCUSSION

Functional attributes (associated with biological behaviors of disease cells) are the responses that effector genes respond to the signal they received. Disease cells always have abnormal functional attributes. Thus, the signal that the effector genes received can be very different. However, no current pathway analysis method takes this factor into consideration. Most pathway analysis methods only include the activation and significance of pathways. Their results give us inadequate information on functional attributes that can help explain the biological behaviors of disease cells. Here, we proposed SPFA, a novel signaling pathway analysis method that takes into account signal variations that effector genes receive under disease and normal conditions. Our results showed that SPFA was comparable to seven other signaling pathway analysis methods. We also found that the signal variations that effector genes receive can reflect the contribution of different

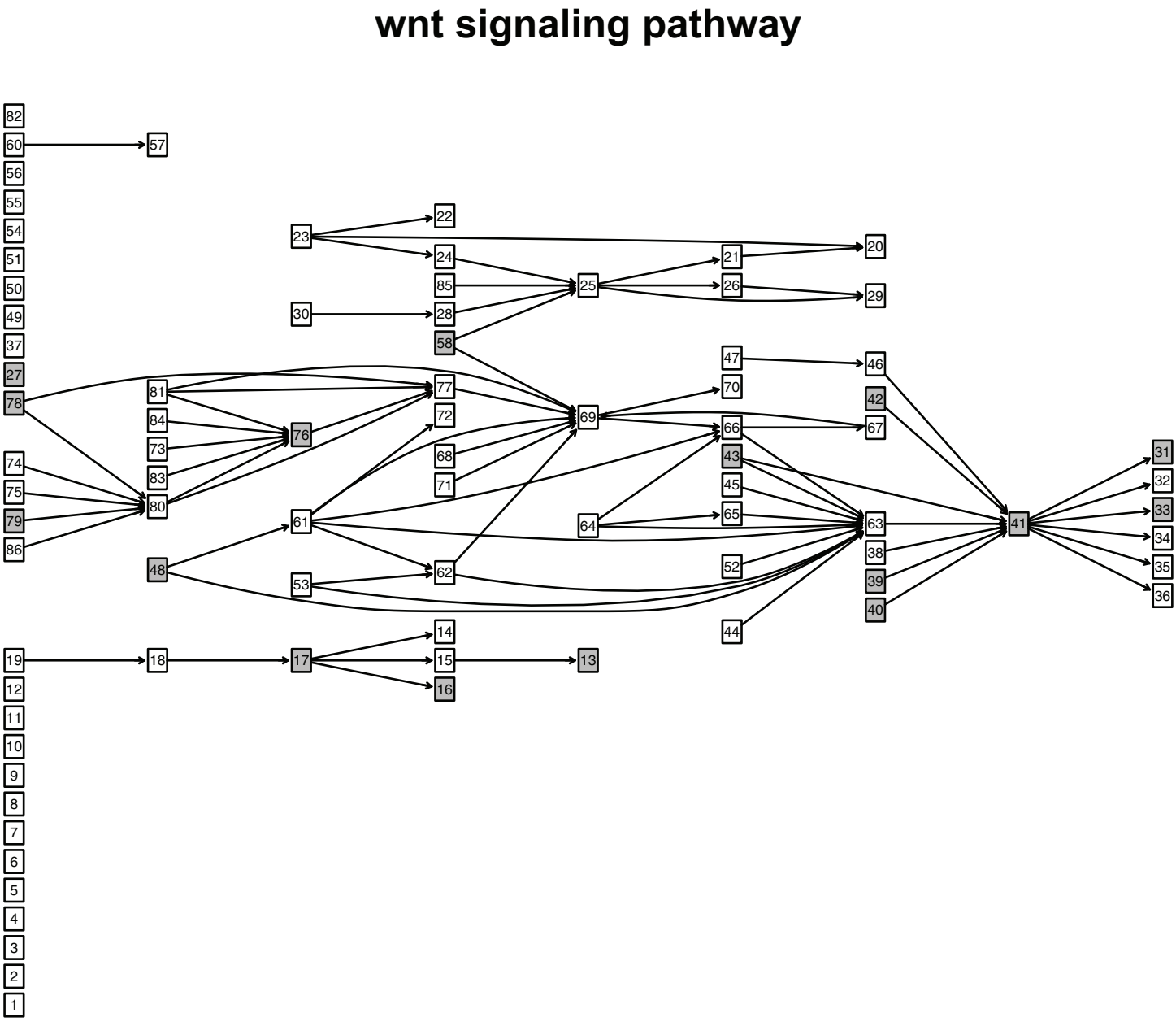

Figure 10 The distribution of DEGs in Wnt signaling pathway using colorectal cancer datasets (GSE4183). The nodes with grey color mean that these nodes contain DEGs; the nodes with white color mean that these nodes do not contain DEGs. Full-size DOI: 10.7717/peerj.9695/fig-10

functional attributes in the signaling pathway, deepening our understanding of disease cells' biological behaviors. Additionally, SPFA used the effector genes with high signal variations to find the abnormal part of the disease-related pathway.

However, SPFA was weaker than MRGSE, Fisher, and PADOG when comparing the median p-values of target pathways. We assume this is due to the statistical models used. The probability P_{sd} is evaluated by gene IDs permutation. Correlation differences are sometimes used to establish differential co-expression networks. This indicates that high correlation differences may exist in randomly-selected paired genes. The p-values may

increase when paired genes with high correlation differences are randomly selected. Future studies should use a better statistical model to resolve this problem. Additionally, the 33 gene expression datasets used in this work were still limited. More experiments need to be conducted to further validate SPFA's performance. A large number of normal and disease samples are also needed to locate the effector genes with high signal variations in disease-related pathways. These genes could then serve as effective module biomarkers for accurately detecting or diagnosing complex diseases, or as drug discovery targets. SPFA depends on manually-curated signaling pathways which play a small role in complex cellular progression. More signaling pathways need to be discovered for SPFA's optimal performance.

CONCLUSIONS

In this study, we developed a new signaling pathway analysis method called SPFA. We compared this method's ability to identify altered signaling pathways against the other seven methods. SPFA showed better results than the seven other methods. Our results also showed that the SPFA method could help identify abnormal functional attributes under normal and disease conditions and the abnormal parts of a pathway during the disease biological process.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by grants from the National Natural Science Foundation of China (61871121, 61271055, and 61471112). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China: 61871121, 61271055, and 61471112.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Zhenshen Bao conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Bing Zhang performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Li Li analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Qinyu Ge analyzed the data, prepared figures and/or tables, and approved the final draft.
- Wanjun Gu analyzed the data, prepared figures and/or tables, and approved the final draft.
- Yunfei Bai conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The R code of SPFA can be accessed at GitHub: https://github.com/ZhenshenBao/SPFA. The accession numbers of the 33 datasets used are available in Table 1.

REFERENCES

Affer M, Dao S, Liu C, Olshen AB, Mo Q, Viale A, Lambek CL, Marr TG, Clarkson BD. 2011. Gene expression differences between enriched normal and chronic myelogenous leukemia quiescent stem/progenitor cells and correlations with biological abnormalities. *Journal of Oncology* **2011(5894)**:798592 DOI 10.1155/2011/798592.

Badea L, Herlea V, Dima SO, Dumitrascu T, Popescu I. 2008. Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepatogastroenterology* **55(88)**:2016–2027.

Bao Z, Li X, Zan X, Shen L, Ma R, Liu W. 2016. Signalling pathway impact analysis based on the strength of interaction between genes. *IET Systems Biology* **10(4)**:147–152 DOI 10.1049/iet-syb.2015.0089.

Barth AS, Kuner R, Buness A, Ruschhaupt M, Merk S, Zwermann L, Kaab S, Kreuzer E, Steinbeck G, Mansmann U, Poustka A, Nabauer M, Sultmann H. 2006. Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies. *Journal of the American College of Cardiology* **48(8)**:1610–1617.

Blake JA, Dolan M, Drabkin H, Hill DP, Li N, Sitnikov D, Bridges S, Burgess S, Buza T, McCarthy F, Peddinti D, Pillai L, Carbon S, Dietze H, Ireland A, Lewis SE, Mungall CJ, Gaudet P, Chrisholm RL, Fey P, Kibbe WA, Basu S, Siegele DA, McIntosh BK, Renfro DP, Zweifel AE, Hu JC, Brown NH, Tweedie S, Alam-Faruque Y, Apweiler R, Auchinchloss A, Axelsen K, Bely B, Blatter M, Bonilla C, Bouguerleret L, Boutet E, Breuza L, Bridge A, Chan WM, Chavali G, Coudert E, Dimmer E, Estreicher A, Famiglietti L, Feuermann M, Gos A, Gruaz-Gumowski N, Hieta R, Hinz C, Hulo C, Huntley R, James J, Jungo F, Keller G, Laiho K, Legge D, Lemercier P, Lieberherr D, Magrane M, Martin MJ, Masson P, Mutowo-Muellenet P, O'Donovan C, Pedruzzi I, Pichler K, Poggioli D, Porras Millan P, Poux S, Rivoire C, Roechert B, Sawford T, Schneider M, Stutz A, Sundaram S, Tognolli M, Xenarios I, Foulgar R, Lomax J, Roncaglia P, Khodiyar VK, Lovering RC, Talmud PJ, Chibucos M, Giglio MG, Chang H, Hunter S, McAnulla C, Mitchell A, Sangrador A, Stephan R, Harris MA, Oliver SG, Rutherford K, Wood V, Bahler J, Lock A, Kersey PJ, McDowall DM, Staines DM, Dwinell M, Shimoyama M, Lauledkind S, Hayman T, Wang S, Petri V, Lowry T, D'Eustachio P, Matthews L, Balakrishnan R, Binkley G, Cherry JM, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hitz BC, Hong EL, Karra K, Miyasato SR, Nash RS, Park J, Skrzypek MS, Weng S, Wong ED, Berardini TZ, Huala E, Mi H, Thomas PD, Chan J, Kishore R, Sternberg P, Van Auken K, Howe D, Westerfield M. 2013. Gene ontology annotations and resources. *Park J* **41**:D530–D535 DOI 10.1093/nar/gks1050.

Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW. 2004. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences of the United States of America* **101(7)**:2173–2178 DOI 10.1073/pnas.0308512100.

Donahue TR, Tran LM, Hill R, Li Y, Kovochich A, Calvopina JH, Patel SG, Wu N, Hindoyan A, Farrell JJ, Li X, Dawson DW, Wu H. 2012. Integrative survival-based molecular profiling of human pancreatic cancer. *Clinical Cancer Research* **18(5)**:1352–1363 DOI 10.1158/1078-0432.CCR-11-1539.

Dong X, Yun H, Xiao W, Tian W. 2016. LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. *Scientific Reports* **6(1)**:18871 DOI 10.1038/srep18871.

Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. 2003. Global functional profiling of gene expression. *Genomics* **81(2)**:98–104.

Efron B, Tibshirani R. 2006. On testing the significance of sets of genes. *Annals of Applied Statistics* **1(1)**:107–129 DOI 10.1214/07-aoas101.

Fang H, Li X, Zan X, Shen L, Ma R, Liu W. 2017. Signaling pathway impact analysis by incorporating the importance and specificity of genes (SPIA-IS). *Computational Biology and Chemistry* **71**:236–244 DOI 10.1016/j.compbiolchem.2017.09.009.

Galamb O, Gyorffy B, Sipos F, Spisak S, Nemeth AM, Miheller P, Tulassay Z, Dinya E, Molnar B. 2008. Inflammation, adenoma and cancer: objective classification of colon biopsy specimens with gene expression signature. *Disease Markers* **25(1)**:1–16 DOI 10.1155/2008/586721.

Geistlinger L, Csaba G, Kuffner R, Mulder N, Zimmer R. 2011. From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics* **27(13)**:i366–i373 DOI 10.1093/bioinformatics/btr228.

Gyorffy B, Molnar B, Lage H, Szallasi Z, Eklund AC. 2009. Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples. *PLOS ONE* **4(5)**:e5645 DOI 10.1371/journal.pone.0005645.

Gyorffy B, Molnar B, Lage H, Szallasi Z, Eklund AC. 2009. Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples. *PLOS ONE* **4(5)**:e5645 DOI 10.1371/journal.pone.0005645.

He H, Jazdzewski K, Li W, Liyanarachchi S, Nagy R, Volinia S, Calin GA, Liu CG, Franssila K, Suster S, Kloos RT, Croce CM, De la Chapelle A. 2005. The role of microRNA genes in papillary thyroid carcinoma. *Proceedings of the National Academy of Sciences of the United States of America* **102(52)**:19075–19080 DOI 10.1073/pnas.0509603102.

Hever A, Roth RB, Hevezi P, Marin ME, Acosta JA, Acosta H, Rojas J, Herrera R, Grigoriadis D, White E, Conlon PJ, Maki RA, Zlotnik A. 2007. Human endometriosis is associated with plasma cells and overexpression of B lymphocyte stimulator. *Proceedings of the National Academy of Sciences of the United States of America* **104(30)**:12451–12456 DOI 10.1073/pnas.0703451104.

Hong Y, Downey T, Eu KW, Koh PK, Cheah PY. 2010. A 'metastasis-prone' signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics. *Clinical & Experimental Metastasis* **27(2)**:83–90 DOI 10.1007/s10585-010-9305-4.

Hong Y, Ho KS, Eu KW, Cheah PY. 2007. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clinical Cancer Research* Vol. 13. **(4)**:1107–1114 DOI 10.1158/1078-0432.CCR-06-1633.

Inestrosa NC, Varela-Nallar L, Grabowski CP, Colombres M. 2007. Synaptotoxicity in Alzheimer's disease: the Wnt signaling pathway as a molecular target. *International Union of Biochemistry and Molecular Biology Life* **59(4–5)**:316–321.

Johnson SM, Gulhati P, Rampy BA, Han Y, Rychahou PG, Doan HQ, Weiss HL, Evers BM. 2010. Novel expression patterns of PI3K/Akt/mTOR signaling pathway components in colorectal cancer. *Journal of the American College of Surgeons* **210(5)**:767–778 DOI 10.1016/j.jamcollsurg.2009.12.008.

Joshitope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, Bono BD, Jassal B, Gopinath GR, Wu GR, Matthews L. 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research* **33(Database issue)**:428–432 DOI 10.1093/nar/gki072.

Juan NI, Chun-Chi L, Morgan TE, Finch CE, Xianghong Jasmine Z. 2010. Joint genome-wide profiling of miRNA and mRNA expression in Alzheimer's disease cortex reveals altered miRNA regulation. *PLOS ONE* **5(2)**:e8898 DOI 10.1371/journal.pone.0008898.

Jung S. 2018. KEDDY: a knowledge-based statistical gene set test method to detect differential functional protein-protein interactions. *Bioinformatics* **35(4)**:619–627 DOI 10.1093/bioinformatics/bty686.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44(D1)**:D457–D462 DOI 10.1093/nar/gkv1070.

Khatri P, Drăghici S, Ostermeier GC, Krawetz SA. 2002. Profiling gene expression using onto-express. *Genomics* **79(2)**:266–270 DOI 10.1006/geno.2002.6698.

Khatri P, Sirota M, Butte AJ. 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Computational Biology* **8(2)**:e1002375 DOI 10.1371/journal.pcbi.1002375.

Le Dieu R, Taussig DC, Ramsay AG, Mitter R, Miraki-Moud F, Fatah R, Lee AM, Lister TA, Gribben JG. 2009. 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* **114(18)**:3909–3916 DOI 10.1182/blood-2009-02-206946.

Lenburg ME, Liou LS, Gerry NP, Frampton GM, Cohen HT, Christman MF. 2003. Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data. *BMC Cancer* **3(1)**:5 DOI 10.1186/1471-2407-3-31.

Li X, Shen L, Shang X, Liu W. 2015. Subpathway analysis based on signaling-pathway impact analysis of signaling pathway. *PLOS ONE* **10(7)**:e0132813 DOI 10.1371/journal.pone.0132813.

Liang WS, Dunckley T, Beach TG, Grover A, Mastroeni D, Walker DG, Caselli RJ, Kukull WA, McKeel D, Morris JC, Hulette C, Schmechel D, Alexander GE, Reiman EM, Rogers J, Stephan DA. 2007. Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiological Genomics* **28(3)**:311–322 DOI 10.1152/physiolgenomics.00208.2006.

Liu M, Ping C, Frédéric S, Ritchie ME, Catherine C, Tim B, Karine BP, Robert E, Simpson KM, Joëlle M. 2008. Integrative analysis of RUNX1 downstream pathways and target genes. *BMC Genomics* **9(1)**:363 DOI 10.1186/1471-2164-9-363.

Liu W, Xu P, Bao Z. 2019. Understanding the mechanisms of cancers based on function sub-pathways. *Computational Biology and Chemistry* **78**:491–496 DOI 10.1016/j.compbiolchem.2018.11.011.

Liu Z, Yao Z, Li C, Lu Y, Gao C. 2011. Gene expression profiling in human high-grade astrocytomas. *Comparative and Functional Genomics* **2011(3)**:245137 DOI 10.1155/2011/245137.

Nishimura D. 2001. BioCarta. *Biotech Software & Internet Report* **2(3)**:117–120.

Pei H, Li L, Fridley BL, Jenkins GD, Kalari KR, Lingle W, Petersen G, Lou Z, Wang L. 2009. FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt. *Cancer Cell* **16(3)**:259–266 DOI 10.1016/j.ccr.2009.07.016.

Runne H, Kuhn A, Wild EJ, Pratyaksha W, Kristiansen M, Isaacs JD, Regulier E, Delorenzi M, Tabrizi SJ, Luthi-Carter R. 2007. Analysis of potential transcriptomic biomarkers for

Huntington's disease in peripheral blood. *Proceedings of the National Academy of Sciences of the United States of America* **104(36)**:14424–14429 DOI 10.1073/pnas.0703652104.

Sabates-Bellver J, Van der Flier LG, De Palo M, Cattaneo E, Maake C, Rehrauer H, Laczko E, Kurowski MA, Bujnicki JM, Menigatti M, Luz J, Ranalli TV, Gomes V, Pastorelli A, Faggiani R, Anti M, Jiricny J, Clevers H, Marra G. 2007. Transcriptome profile of human colorectal adenomas. *Molecular Cancer Research* **5(12)**:1263–1275 DOI 10.1158/1541-7786.mcr-07-0267.

Sanchez-Palencia A, Gomez-Morales M, Gomez-Capilla JA, Pedraza V, Boyero L, Rosell R, Farez-Vidal ME. 2010. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *International Journal of Cancer* **129(2)**:355–364 DOI 10.1002/ijc.25704.

Sebastian-Leon P, Vidal E, Minguez P, Conesa A, Tarazona S, Amadoz A, Armero C, Salavert F, Vidal-Puig A, Montaner D, Dopazo J. 2014. Understanding disease mechanisms with models of signaling pathway activities. *BMC Systems Biology* **8(1)**:121 DOI 10.1186/s12918-014-0121-3.

Stirewalt DL, Meshinchi S, Kopecky KJ, Fan W, Pogosova-Agadjanyan EL, Engel JH, Cronk MR, Dorcy KS, McQuary AR, Hockenbery D, Wood B, Heimfeld S, Radich JP. 2008. Identification of genes with abnormal expression changes in acute myeloid leukemia. *Genes Chromosomes Cancer* **47(1)**:8–20 DOI 10.1002/gcc.20500.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102(43)**:15545–15550.

Tarca AL, Bhatti G, Romero R. 2013. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLOS ONE* **8(11)**:e79217 DOI 10.1371/journal.pone.0079217.

Tarca AL, Draghici S, Bhatti G, Romero R. 2012. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics* **13(1)**:136 DOI 10.1186/1471-2105-13-136.

Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R. 2009. A novel signaling pathway impact analysis. *Bioinformatics* **25(1)**:75–82 DOI 10.1093/bioinformatics/btn577.

Uddin S, Ahmed M, Hussain A, Abubaker J, Al-Sanea N, AbdulJabbar A, Ashari LH, Alhomoud S, Al-Dayel F, Jehan Z, Bavi P, Siraj AK, Al-Kuraya KS. 2011. Genome-wide expression analysis of Middle Eastern colorectal cancer reveals FOXM1 as a novel target for cancer therapy. *American Journal of Pathology* **178(2)**:537–547 DOI 10.1016/j.ajpath.2010.10.020.

Ullah MO. 2013. Improving the output of signaling pathway impact analysis. *Romanian Statistical Review* **61**:38–43.

Voichita C, Donato M, Draghici S. 2012. Incorporating gene significance in the impact analysis of signaling pathways. In: *2012 11th International Conference on Machine Learning and Applications, 12–15 December 2012 Boca Raton, FL, USA*. 126–131.

Wallace TA, Prueitt RL, Yi M, Howe TM, Gillespie JW, Yfantis HG, Stephens RM, Caporaso NE, Loffredo CA, Ambs S. 2008. Tumor immunobiological differences in prostate cancer between African–American and European-American men. *Cancer Research* **68(3)**:927–936 DOI 10.1158/0008-5472.CAN-07-2608.

Wang Y, Roche O, Yan MS, Finak G, Evans AJ, Metcalf JL, Hast BE, Hanna SC, Wondergem B, Furge KA, Irwin MS, Kim WY, Teh BT, Grinstein S, Park M, Marsden PA, Ohh M. 2009. Regulation of endocytosis via the oxygen-sensing pathway. *Nature Medicine* **15(3)**:319–324 DOI 10.1038/nm.1922.

Zhang Y, James M, Middleton FA, Davis RL. 2005. 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. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **137B(1)**:5–16 DOI 10.1002/ajmg.b.30195.

Zheng B, Liao Z, Locascio JJ, Lesniak KA, Roderick SS, Watt ML, Eklund AC, Zhang-James Y, Kim PD, Hauser MA, Grunblatt E, Moran LB, Mandel SA, Riederer P, Miller RM, Federoff HJ, Wullner U, Papapetropoulos S, Youdim MB, Cantuti-Castelvetri I, Young AB, Vance JM, Davis RL, Hedreen JC, Adler CH, Beach TG, Graeber MB, Middleton FA, Rochet JC, Scherzer CR. 2010. PGC-1alpha, a potential therapeutic target for early intervention in Parkinson's disease. *Science Translational Medicine* **2(52)**:52ra73 DOI 10.1126/scitranslmed.3001059.

Zheng Q, Wang XJ. 2008. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Research* **36**:W358–W363.