






# A Flexible and Comprehensive Platform for Analyzing Gene Expression Data

Bolin Chen<sup>1,2</sup> , Chenfei Wang<sup>1</sup>, Li Gao<sup>3</sup>, and Xuequn Shang<sup>1,2</sup>  

<sup>1</sup> School of Computer Science, Northwestern Polytechnical University, Xi'an, China  
npu\_bioinf@hotmail.com

<sup>2</sup> Key Laboratory of Big Data Storage and Management, Ministry of Industry and Information Technology, Northwestern Polytechnical University, Xi'an, China

<sup>3</sup> School of Software, Northwestern Polytechnical University, Xi'an, China

**Abstract.** Studying the original gene expression dataset is one of the essential methods for analyzing biological processes. Many platforms were developed to conduct this kind of study, such as GSEA, and the online gene list analysis portal Metascape. However, these well-known platforms sometimes are not friendly enough for inexperienced users due to the following reasons. Firstly, many biological experiments only have three duplicates, which make classical statistical methods lack of efficient and accuracy. Secondly, different experiments could result in different gene expression profiles, where standard differential expressed gene identification methods still have room to be further improved. Thirdly, many platforms work only for specific experimental conditions based on their default parameters, where users are not easily setup parameters for their own studies. In this study, we designed a comprehensive and flexible gene expression data analysis tool, where six novel differential expressed gene identification methods and three functional enrichment analysis methods were proposed. Majority parameters can be friendly setting by users and a variety of algorithms can be 9 according to the user's own study designing. Experiments show that our platform provides an effective way for gene set series analysis, and has great performance in both practicality and convenience.

**Keywords:** Gene expression · Differentially expressed genes · Microarray data · Functional enrichment analysis

## 1 Introduction

Detecting differentially expressed genes aims to find the classes of genes that are significantly expressed or depressed. In many biological experiments, taking drug discovery as an example, researchers need to find genes that have significant changes in expression levels due to drug action. Furthermore, biological changes typically involve a gene set in which multiple genes are associated with individual biological pathways or GO terms.

---

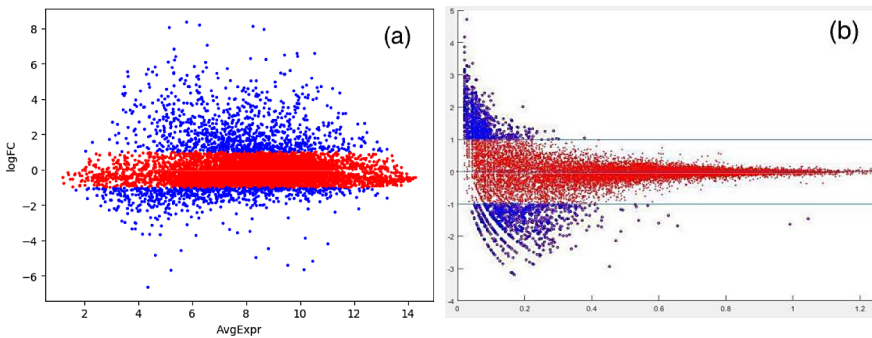
B. Chen and C. Wang—Equal contributors.

Therefore, researchers normally follow the following two steps to analyze gene expression datasets: i) identify differentially expressed genes, ii) perform functional enrichment analysis based on differentially expressed genes, such as pathway enrichment analysis or GO analysis.

Although many methods and tools have been widely developed for analyzing gene expression data sets, there are still some shortcomings. Considering the flexibility and practicality of the tool, we think that the current analysis tools can be further improved in the following three aspects.

Firstly, due to the limited budgets or low sample availability, many biological experiments have only three duplications for each condition. These limited number of samples makes many standard statistical methods inefficient, and the results based on these methods contain many false positives, which is not reliable.

Secondly, Fig. 1(a) shows gene expression dataset that the majority of genes were normally expressed, and only a small number of genes were expressed with very low or very high levels. The standard fold-change method [1] works very well for this kind of dataset. However, when it comes to the dataset like Fig. 1(b), this method may not work well. The standard fold-change method could only detect genes with small expression levels, because a slight change of the gene expression could result in at least two-fold changes between the test and control groups, but a larger expression change is needed for a gene if its expression level is high [2]. A scatter plot of this kind of gene expression with 3 test and 3 control samples was shown in Fig. 1(b). The simple fold change method is not sensitive to expression noises [3], which not only results in lots of false positives, but reduces the possibility of detecting highly expressed genes.



**Fig. 1.** The scatter plot of gene expression dataset. (a) A typical gene expression scatter plot where majority of genes were expressed normally, with only small number of genes were expressed with very low or very high levels. (b) Another gene expression scatter plot where the expression level of majority of genes were very low.

Thirdly, many gene expression analysis tools are designed only for general experiments and are not effective enough for other kinds of datasets. Many other tools only use default values to encapsulate the parameters of the differential expressed gene detection algorithms, and do not allow users to set parameters based on their own experimental conditions. Moreover, some platforms only focus on either differentially expressed genes

or functional enrichment analysis, which makes users have to combine at least two different platforms to complete the whole analysis processes. Since one platform generally is designed only for specific kind of biological condition [4], the mixed combination of different platform may introduce new biases. Besides, some online analysis tools only allow limited gene expression data, for example, 3000 genes [5], which results in large volumes of data that cannot be processed.

In this study, we designed a comprehensive and flexible platform for analyzing gene expression data that allows for custom parameter settings for almost all-important parameters. In addition, we proposed six new methods for identifying differential expressed genes and two functional enrichment analysis methods. The platform is applicable to different types of data, making the design platform more versatile and effective.

The structure of the article is arranged as follows. Section 2 introduces methods for data cleaning and differentially expressed identification. Section 3 compares the results of the proposed methods. Section 4 draws conclusions and discussions.

## 2 Methods

### 2.1 The Overall Framework of the Gene Expression Analysis Platform

The overall framework of the proposed platform is shown in Fig. 2.

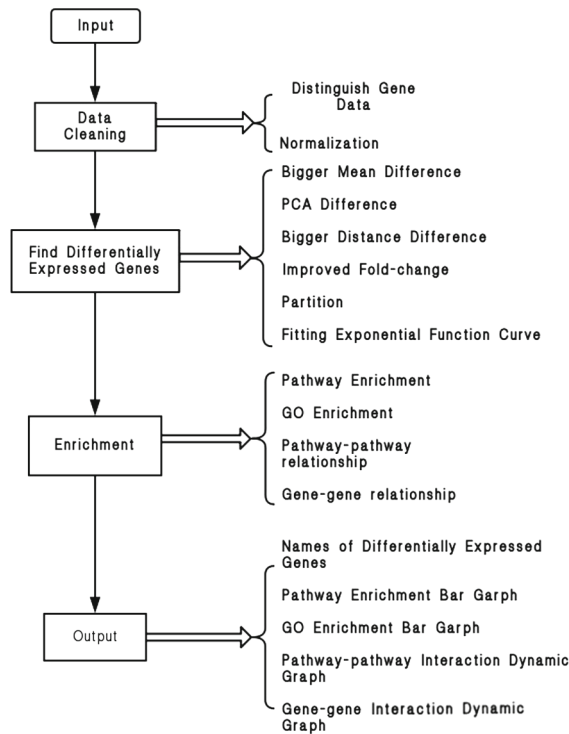
**A: Input file types.** The designed platform allows for the entire process of gene expression data analysis, from data cleansing to data normalization, from identification of expressed genes to functional enrichment analysis. Thus, it can support users either to upload the raw gene expression data sets or provide a list of differentially expressed genes for further enrichment analysis.

Two kinds of raw gene expression datasets are allowed as inputs, either in the form of one file that contains both the control and the test dataset together, or in the form of two files which contain the test group of data and the control group of data separately. This makes the platform has high compatibility and the scalability.

The proposed platform also allows users to define differentially expressed gene list as input, since in many cases, one only interest in the functional related analysis of a group of genes, without re-selecting genes from a given file. Several promising functional enrichment algorithms and network layout processes are employed to display the analysis results interactively.

**B: Data cleaning.** Data cleaning is one of the most overlooked but also the most critical steps in gene expression data analysis. In general, external experimental conditions affect gene expression data. For example, the drug injection dose of a sample may be proportional to the gene expression levels.

More importantly, gene expression data is affected by biological and technological variations, and a certain gene may correspond to multiple expression level. Therefore, the necessary data cleaning process must be performed prior to genetic analysis, which is an essential step in any data analysis.



**Fig. 2.** The framework of the gene expression dataset analysis.

**C: Identifying differentially expressed genes.** Although many traditional methods to find differentially expressed genes in a given dataset, such as fold-change method [2] and t-test method [6, 7], these traditional methods are not proper for certain situations. For example, the fold-change based methods are not appropriate to this situation where there are lots of low expression values (or zero values) in the raw data, and the t-test will not perform well when the number of samples is small. Our platform improves existing traditional methods with new five algorithms to identify the differentially expressed genes in a given expression data set. The platform also allows users to set parameters according to their own experimental conditions.

**D: Functional enrichment analysis.** Once obtaining a list of significantly differentially expressed genes, the functional enrichment analysis will be conducted by integrating pathways, Gene Ontology (GO) and some biological networks. The platform implements different algorithms to functional enrichment analysis, as well as find the relationships between pathways and genes, between GOs and genes, between pathways and pathways, or between genes and genes.

**E: Output information and results.** The platform will finally prepare the outcomes drawn from the above results. Users can get the following information: (1) significantly differentially expressed genes list, (2) pathways these genes enriched, (3) GOs that these

genes enriched, (4) the relationships between significantly enriched pathways and (5) the relationships between these differentially expressed genes.

For better observation of the relationships between pathways and pathways as well as genes and genes, the platform will generate functional networks to show these relationships by using d3.js. Also, the platform will give bar graph to show the degree of how significantly the pathways or GOs enriched.

## 2.2 Methods for Data Cleaning

**A: Distinguishing gene data.** In the raw gene expression data set, sometimes we notice that a certain gene may appear more than once, thus this gene may have multiple different expression level. In order to choose the exact data for further analysis, we propose three methods to distinguish gene data from the raw gene expression profiles.

*Method 1.* selecting the first occurrence of the data

For a certain gene that appears more than once, we use the expression data that appears earliest and ignore all following profiles of the same gene.

*Method 2.* using the sum value

For a certain gene that appears more than once, we will find all expression levels of this gene, and calculate the sum of the values in corresponding samples. Then, the new data will be used for the further analysis.

*Method 3.* using the mean value

The average of all expression values is calculated for a particular gene as its new profile.

**B: Data normalization.** To eliminate the effect of different concentration degrees of the raw samples brought to data, the platform also provides normalization based on the sum of columns.

## 2.3 Algorithms for Identifying Differentially Expressed Genes

Giving two sets of  $m$  control samples and  $n$  test samples, respectively. Let  $c_1, c_2, \dots, c_m$  be the expression value of control samples and  $t_1, t_2, \dots, t_n$  be expression value in test samples. Let

$$\bar{c} = \frac{\sum_{i=1}^m c_i}{m}$$

represents the mean value of the expression data in control samples, and

$$\bar{t} = \frac{\sum_{i=1}^n t_i}{n}$$

represents the mean value of the expression data in test samples, respectively.

The traditional fold-change based method can be describe briefly as the following formula:

$$s = \log_2 \frac{\bar{i}}{\bar{c}} \quad (1)$$

If  $s$  is larger than 1 or smaller than  $-1$  [8, 9], the gene will be regarded as significantly highly expressed or depressed. However, there are some potential problems when using this method. Firstly, a lower expressed gene is easier to obtain a two-fold change compared to a higher expressed gene in term of their change requirement. Hence, this method is biased to low expressed genes, and may misclassify differentially expressed genes with small ratios but large differences, leading to poor identification of genes at high expression levels [2]. Secondly, when the expression values in either the denominator or numerator are close or equal to zero, the ratio is not stable and even cannot be calculated, and the fold change value can be disproportionately affected by measurement noise. To solve these specific problems, we provide the following six methods to find differentially expressed genes.

**A: Improved fold-change method.** This method is designed for the situation where there are zero values in the raw expression data, because zero cannot be denominator. Traditional methods normally filtered those genes, but we think those genes still have valuable information to interpret gene expressions.

*Case 1.* For normal gene expression values

This is the normal case where the gene expression data can be used to do fold-change. For all the genes to be analyzed, we take the gene as differentially expressed gene when the calculated  $s$  in (1) is larger than a fold threshold. Users can set the fold parameter or choose the genes with top  $n\%$   $s$  as differentially expressed genes, where  $n$  can be set by themselves according to requirements.

*Case 2.* data For gene expression values where  $\bar{c} = 0$  or  $\bar{i} = 0$

For all the genes with  $\bar{c} = 0$ , we look at the corresponding  $\bar{i}$ . If a certain gene with  $\bar{c} = 0$  while its  $\bar{i}$  is in the highest  $k\%$  of this kind of genes, we think this gene as significantly expressed, where  $k$  can be set by users according to their requirements. Following the same procedure, we process genes with  $\bar{i} = 0$ . Then, we put the two parts of differentially expressed genes together as significantly expressed genes.

*Case 3.* For gene expression values where  $\bar{c}$  or  $\bar{i}$  closes to zero

Either in test group or control group, if there are more than two thirds of the gene expression data is 0, we may consider that the non-zero values may be caused by noise. So, in this case, we need to do some pretreatments, transferring the gene dataset to above two cases. For a certain gene, if there are more than two thirds of the expression data is 0 in test group, then we calculate its  $\bar{i}$  After processing all this kind of genes, if the  $\bar{i}$  is smaller than the  $k\%$  percentile of all  $\bar{i}$   $s$  of this kind of genes, we will view that the gene does not express in test group and the small value is drawn from noise. The  $k\%$  can be set by users according to their requirements. Following the same procedure, we process the genes with more than two thirds of the expression data is 0 in control group. We will set its expression data in test group all to 0s. For the rest of the genes, when selecting differentially expressed genes, do as the Case 1 and Case 2 do.

The expected results of the differentially expressed genes should scatter uniformly in different expression levels, rather than just concentrate on the areas where the gene expression level is pretty low. To solve this problem, we provide the following methods.

**B: A divided section-based method.** To identify differentially expressed genes scattered not uniformly in different expression levels, a set of divided sections was given according to the mean expression values. Taking the highest  $n\%$  logarithm of the ratio in every divided section as differentially expressed genes, where  $n$  can be set by users. Still, we will focus on the gene dataset in normal cases, the procedures are same in previous improved fold-change method. The pretreatment procedure is discussed before.

*Step 1.* Pretreatments should be done to transfer the gene dataset to normal cases. Identify differentially expressed genes in special situations firstly, and then rank the expression level of genes according to  $\bar{t}$ .

*Step 2.* Divide the genes into sections, ensuring that the number of the genes in every section is same (except the last section). It will lead to under-fitting or overfitting if the number of genes is too small or too big. In expected situation, every section includes 0.02%–0.03% of the total genes.

*Step 3.* In every section, calculate the fold-change score  $s$  in (1), taking the genes with highest  $n\%$   $s$  score as differentially expressed ones.

This partition method has met the requirement of selecting differentially expressed genes from low to high levels. However, there is still a little problem associating with partition method. We divide the genes into lots of groups, but the boundaries between groups will be very obvious if the number of groups is too small. In turn, it will lead to overfitting, greatly lower the quality of the outcome if the number of groups is too big. Thus, we design the following method to find a smooth curve, which shows that genes above or below are selected to be differentially expressed genes.

**C: Fitting Exponential Function Curve.** We notice that the exponential function can be a monotonically decreasing function, meeting the expectation that differentially expressed genes are not uniformly distributed in low and high expression level. Here, we just focus on the gene dataset in normal cases, and the pretreatment procedure is discussed in section A: *Improved fold-change method*.

*Step 1.* Rank the data of the genes according to the mean expression value in test samples  $\bar{t}$ .

*Step 2.* Finding out the gene with its mean expression value in test group, if its expression value is in  $k^{\text{th}}$  percentile of all mean expression value in test group, where  $k$  can be set by users according to their requirements, denote this value as  $x$ . Similarly, calculate its corresponding mean expression value in control samples, and denote this value as  $y$ .

*Step 3.* According to  $x$  and  $y$ , we can get a specific exponential function (2).

$$y = a^x \tag{2}$$

However, the exponential function always intersects with y-axis on (0,1). So, the exponential function should be moved rightward. The distance of the right-move is the

$k^{th}$  percentile of all mean expression value in test group of genes, and denote this value as  $x_0$ . Then, we will get a new exponential function as follows.

$$y = a^{x-x_0} \tag{3}$$

*Step 4.* We use the mean expression value in test group of the genes as abscissa, and absolute value of the logarithm of 2-fold change between mean expression value in two treatments. We choose the corresponding gene as a differentially expressed gene if the point is beyond the exponential function curve.

**D: Mean difference-based method.** Another alternative method is based on the mean difference. We use the following formula to evaluate the changes of expression level (4) for a certain gene  $g_i$ ,

$$x_i = |\bar{t}_i - \bar{c}_i| \tag{4}$$

The genes with the difference in highest  $n\%$  of all the differences are selected as differentially expressed genes, where  $n$  can be set by users according to their requirements.

**E: PCA difference-based method.** Similar to the previous method, but the only difference is doing pretreatment before calculating the needed differences.

*Step 1.* For all the genes, drawing PCA to one dimension on their expression values in test and control samples [10], respectively.

*Step 2.* For a specific gene, calculate the absolute value of the difference between the principle component in test and control samples.

*Step 3.* After processing all the genes, choose the genes with the difference in highest  $n\%$  as differentially expressed genes.

**F: Geometrical distance-based method.** In this method, we use the geometrical distance of the raw expression data between two conditions. We view gene values in test and control samples as two points respectively, if the “distance” between the two points is very large, we may say that the expression level of the gene changed greatly in test group and control group. Here, we use 2-norm to describe the “distance” between the two points. Notice that the method is just for the situation where the numbers of samples in test group and in control group are same.

*Step 1.* For the genes, rank the expression values in test group and control group respectively.

*Step 2.* Two vectors are constructed by using the sorted values for the test and control groups, respectively.

$$T : [t'_1 t'_2 \dots t'_n] \tag{5}$$



$$C : [c'_1 c'_2 \dots c'_m] \quad (6)$$

*Step 3.* Calculate the 2-norm of the two vectors, just like calculate the distance between two points (7).

$$d = |T - C| \quad (7)$$

*Step 4.* After processing all the genes, choose the genes with the  $d$  in highest  $n\%$  of all the  $d$ s as differentially expressed genes, where  $n$  can be set by users according to their requirements.

## 2.4 Algorithms for Functional Enrichment Analysis

The comparison of individual gene expression values is not sensitive enough to detect the subtle functional changes of biological systems, because cellular changes typically involve in many groups of genes. Multiple genes are linked to a single biological pathway [11, 12], and it is the additive change in expression within gene sets that leads to the difference in phenotypic expression. Thus, researchers will not just focus on the isolated genes, but turn to the pathways and GOs that these differentially expressed genes enriched. The enrichment analysis helps researchers to understand and interpret omics data from the point of view of canonical prior knowledge structured in the forms of pathways and GO diagrams. This allows us finding distinct cellular processes, biological changes or diseases that are statistically associated with selection of differentially expressed genes between two samples [7].

The proposed platform supports four kinds of enrichment analyses, where two of which are proposed to meet the basic requirements for the pathway enrichment and GO enrichment analysis, respectively.

**A: Pathway enrichment.** Traditional enrichment analysis algorithm is based on Fisher's exact test [13]. In addition, the platform also supports additional two more methods for enrichment analysis for different experimental conditions.

*Method 1.* Fisher's exact test

The detail methods of the fisher's exact test can be found in [14].

*Method 2.* Using the relationship between gene sets

We know that the membership overlap of two gene sets can be used to describe the relationship about them. The bigger the intersection of the two sets, the closer of their relationships. In this method, we apply this concept to enrichment analysis as follows. For a certain pathway, we can easily know what genes it includes according the information in the database at the backend of the platform. We use a new set  $U$  to store the intersection part of the genes included in pathway and the differentially expressed genes found before, and use another set  $V$  to store the union part of the genes included in pathway and the differentially expressed genes found before. If the  $U$  is not null, we calculate the quotient of the number of genes in  $U$  and  $V$ . The bigger the value, the more the two sets overlap, which means the more significantly the pathway is enriched.

*Method 3. Using the weighted f-measure*

We borrow the ideas from the concepts of precision and recall [15] in this method. Let  $P_i$  denote the gene set of pathway  $i$ , and  $D$  the set of differentially expressed genes. The precision is defined as the quotient between the number of genes in the intersection of the two sets and the number of genes in difference, and the recall is defined as the quotient between the number of genes in the intersection of the two sets and the number of genes in interested. Then we calculate the f-measure to describe how the pathway is enriched in given differentially expressed gene set. However, we notice that the number of genes in the intersection of significantly differential is small when doing pathway enrichment, which leads to the fact that precision is small. Thus, the f-measure is largely depended on recall. So, we calculate the weighted f-measure to balance importance of precision and recall, for example, we will make precision more important when we are calculating the f-measure.

**B: Pathway Relationship.** After finding out the pathways the differentially expressed genes are included, we use the degree of overlap of the pathways to describe the relationship between them.

*Step 1.* Find out the differentially expressed genes that contained in the pathway.

*Step 2.* Calculating the number of genes in their intersections and unions for the given pathways.

*Step 3.* If the numbers of genes in the intersections and unions are not equal to zero, calculate the quotient of the numbers of genes in the intersections and in the unions. Then we use this quotient to describe the relationship between pathways. The bigger the quotient, the closer connection between two pathways.

**C: Gene Ontology Enrichment.** Gene Ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species. The way we take GO enrichment analysis is the same to the previous pathway enrichment. We support traditional Fisher's exact test, as well as use the relationship between precision and recall. The difference is that we use f-measure as a standard when we use precision and recall to describe whether the GO is significantly enriched.

**D: GO Relationship.** As we do in section *B: Pathway Relationship*, we use the same method to find the relationships between the selected significantly enriched GOs. However, GO is kind of a big term which contains many information. So, we don't pay much attention in this study. The updated version will contain this part.

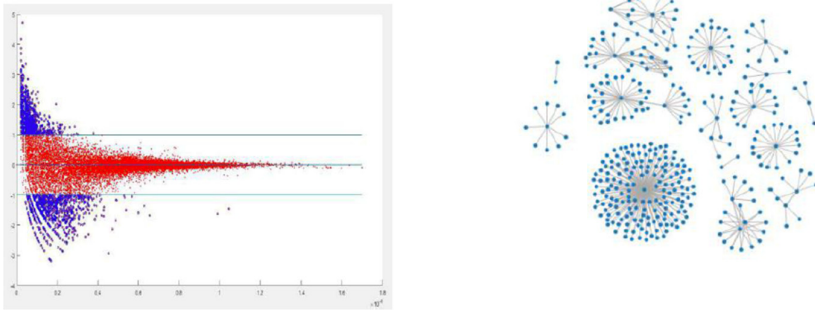
**E: Functional relationship network.** A functional relationship network between genes will then be generated by using PPI networks, where each node represents a gene and edge represents the interaction between two genes. This functional relationship network between differentially expressed genes should show potential gene communities related to different biological functions, which is an important supplementary for those canonical prior knowledges structured.

### 3 Results

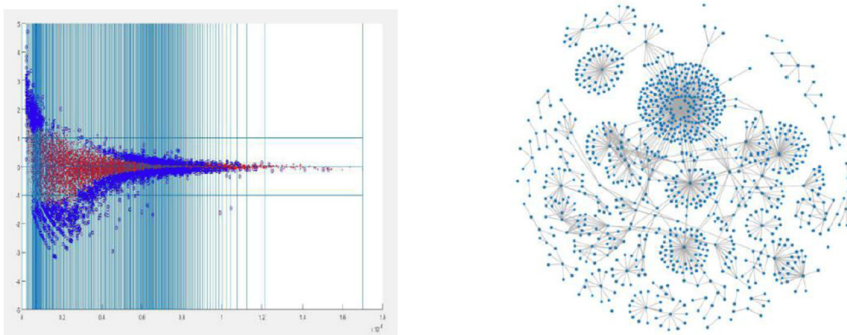
#### 3.1 Effects of the Methods for Detecting Differentially Expressed Genes

We compared the performance of the proposed methods in terms of detecting differentially expressed genes and their functional communities, as following figures shows.

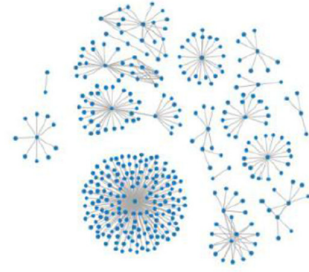
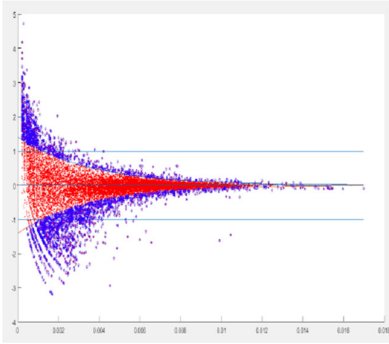
The x-axis represents the mean expression value in test group of the genes, and the y-axis is the absolute value of the logarithm of the ratio of the mean expression value between test group and control group (Figs. 3, 4, 5, 6, 7 and 8).



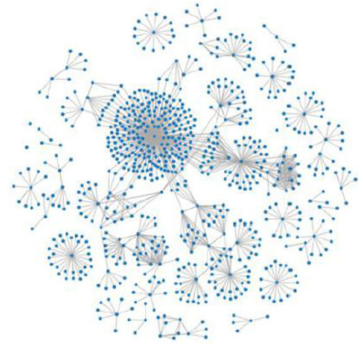
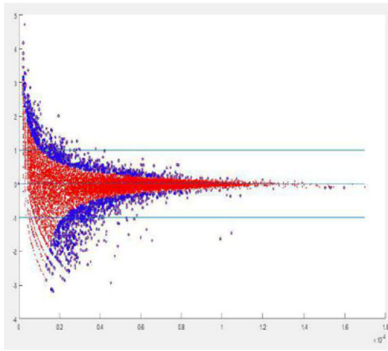
**Fig. 3.** The performance of *the improved fold-change method* for identifying differentially expressed genes. Red spots represent all raw gene data in the expression data set, the blue spots represent the gene that are differentially expressed. There are 291 nodes and 301 edges in the right functional relationship graph. (Colour figure online)



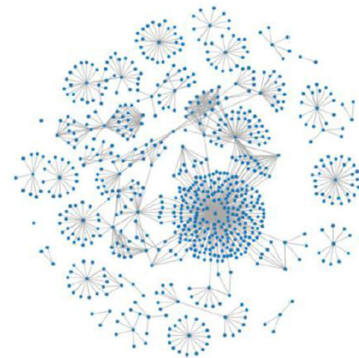
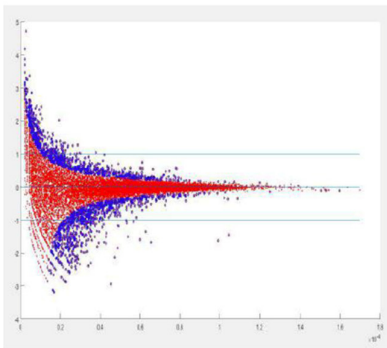
**Fig. 4.** The performance of *the divided section-based method* for identifying differentially expressed genes. The red spots represent all raw gene data in the expression data set, the blue spots represent the gene that are differentially expressed. There are 955 nodes and 1254 edges in the right functional relationship graph. (Colour figure online)



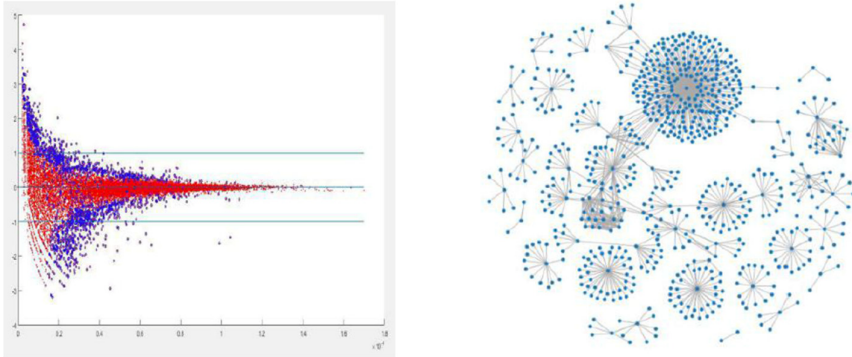
**Fig. 5.** The performance of the *fitting exponential function curve-based method* for identifying differentially expressed genes. The red spots represent all raw gene data in the expression data set, the blue spots represent the gene that are differentially expressed. There are 369 nodes and 404 edges in the right functional relationship graph. (Colour figure online)



**Fig. 6.** The performance of the *mean difference-based method* for identifying differentially expressed genes. The red spots represent all raw gene data in the expression data set, the blue spots represent the gene that are differentially expressed. There are 750 nodes and 979 edges in the right functional relationship graph. (Colour figure online)



**Fig. 7.** The performance of the *PCA difference based method* for identifying differentially expressed genes. The red spots represent all raw gene data in the expression data set, the blue spots represent the gene that are differentially expressed. There are 755 nodes and 982 edges in the right functional relationship graph. (Colour figure online)



**Fig. 8.** The performance of the *geometrical difference-based method* for identifying differentially expressed genes. The red spots represent all raw gene data in the expression data set, the blue spots represent the gene that are differentially expressed. There are 729 nodes and 888 edges in the right functional relationship graph. (Colour figure online)

## 4 Conclusion

We started with microarray data and explained data cleaning and data normalization. Considering that the distribution of data samples may be uneven, we analyzed the distribution of expression data. From the perspective of expressing data types, we proposed six methods for data analyzing, which is greatly reflected in our platform. We also integrated the enrichment analysis function in the platform.

This flexible genetic analysis platform eliminates the need for a separate platform to perform different analysis processes. More importantly, the platform allows users to input raw microarray data and can help filter data noise, users also can set parameters themselves, which greatly improving the shortcomings of traditional platforms. Such an integrated platform returns the final results in text and graphics, facilitating user analysis and visualization of gene expression data. Experiments proved that the proposed platform has good utilization value in both practicability and efficiency.

**Acknowledgement.** This work was supported by the National Natural Science Foundation of China under Grant Nos. 61972320, 61772426, 61702161, 61702420, 61702421, and 61602386, the education and teaching reform research project of Northwestern Polytechnical University (Grant No 2020JGY23), the Fundamental Research Funds for the Central Universities under Grant No. 3102019DX1003, the Key Research and Development and Promotion Program of Henan Province of China under Grant 182102210213, the Key Research Fund for Higher Education of Henan Province of China under Grant 18A520003, and the Top International University Visiting Program for Outstanding Young Scholars of Northwestern Polytechnical University.

## References

1. Tusher, V.G., Tibshirani, R., Chu, G.E.: Significance analysis of microarrays applied to the ionizing radiation response. In: Proceedings of the National Academy of Science of the United

- States of America (24 April 2001) 98 (5116–5121)). Proceedings of the National Academy of Sciences of the United States of America, p. 98 (2001)
2. Mutch, D.M., et al.: The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinform.* **3**(1), 17–20 (2002)
  3. Raser, J.M.: Noise in gene expression: origins, consequences, and control. *Science* (Washington DC), **309**(5743), 2010–2013 (2005)
  4. Zhou, Y., et al.: Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature communications* (2019)
  5. Metascape homepage. <http://metascape.org/gp/index.html#/main/step1>
  6. Dalman, M.R., Deeter, A., Nimishakavi, G., Duan, Z.H: Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinform.* **13**, 256–303 (2012)
  7. Witten, D.M., Tibshirani, R.A.: comparison of fold-change and the t-statistic for microarray data analysis. *Analysis* (2007)
  8. Robinson, M.D, Smyth, G.K.: Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* (2007)
  9. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-SEQ data with DESEQ2. *Genome Biol.* **15**(12), 550 (2014)
  10. Mika, S.: Kernel PCA and de-noising in feature spaces. *Adv. Neural Inf. Process. Syst.* **11**, 65–92 (1999)
  11. Hong, M.G., Pawitan, Y., Magnusson, P.K.E., Prince, J.A.: Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum. Genet.* **126**(2), 289–301 (2009)
  12. Gene Ontology Consortium. The Gene Ontology project in 2008. *Nucleic acids research*, 36(Database issue). D440–D444 (2007). <https://doi.org/10.1093/nar/gkm883>
  13. Huang, D.W., Sherman, B.T., Lempicki, R.A.: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**(1), 1–13 (2009)
  14. Fisher, R.A.: On the interpretation of  $X^2$  from contingency tables, and the calculation of P. *J. Royal Stat. Soc.* **85**(1), 87–94 (1922)
  15. Powers, D.: Evaluation: fom precision, recall and f-measure to roc, informedness, markedness and correlation. *J. Mach. Learn. Technol.* **2**, 37–63 (2007)