A Double-weighted Normalization Method for Identifying Differential Expression of RNA-seq Data

Xiaohui Wu[1], Chuang Zhao[1], Yaru Su[2*]

[1]*Department of Automation*
Xiamen University
Xiamen, 361005 China
E-mails: xhuister@xmu.edu.cn, zhaochuang200808@163.com

[2]*Forensic Science Division*
Department of Fujian Provincial Public Security
Fuzhou, 361003 China
E-mail: yarusu@gmail.com

[*]*Corresponding author*

Received: August 01, 2014 **Accepted: November 11, 2014**

Published: December 19, 2014

Abstract: *The normalization of high-throughput sequencing data from different sequencing conditions is a critical step of the entire high-throughput data analysis and processing. Normalization is important for the identification of gene structures and differentially expressed genes, which has great impact on the accuracy and reliability of downstream analysis procedures. Here, we propose a double-weighted normalization method for high-throughput sequencing data generated by RNA-seq, and present a p-value weighted method to detect differential expression from normalized data. This normalization method not only considers the overall expression level of all genes in a library, but also considers the impact of each individual gene. Experimental results show that our method can effectively normalize high-throughput data under different conditions to provide highly confident data for the downstream analysis of differential expression.*

Keywords: *RNA-seq, Normalization, Double-weighted, Differential expression.*

Introduction

The gene expression in higher organisms not only has the tissue specificity and developmental stage specificity, is also impacted by environmental factors. The differential expression analysis of digital gene expression data cannot only help conduct in-depth study of gene expression regulation and understand the nature of life processes, but also provide an important theoretical basis for gene diagnosis and treatment. Recently, RNA sequencing (RNA-seq) has become an important experimental protocol for the study of gene expression and transcriptome [12]. Due to that different samples are sequenced from different sequencing lanes (libraries), normalization is required to adjust the sequencing depth of different lanes and other potential technical errors for the objective differential expression inference [4, 7]. Experiments with microarray data have shown that normalization is an essential step in the processing pipeline for detecting differential expression [5]. However, although there are many approaches available for the normalization of microarray data [5, 10], they cannot be directly applied for RNA-seq data due to that the procedure for generating RNA-seq data is fundamentally different from that for microarray data [16]. The assumption of normalization is that most genes in the two samples have no global difference on gene expression [16], that is to say differential expression only occurs in a few genes. Normalization is a critical step of the entire high-throughput data analysis and processing, which has great impact on the accuracy and reliability of the subsequent analysis such as identification of gene structure and differentially expressed genes.

Normalization has great impact on the detection of differentially expressed genes. Due to that different lanes have different total number of sequences (i.e. sequencing depth), several widely used methods adopt the total number of sequences of all lanes to normalize the number of genes in each lane [4, 13, 14], such as RPKM (Reads Per Kilo bases per Million reads) method [14] or super geometry model [13]. However, the global normalization method tends to be influenced by a very small number of highly expressed genes, leading to errors in the detection of differentially expressed genes. Some other methods consider the normalization factor, such as geometric average method [19] and Trimmed Mean of M Values (TMM) [16], which can reduce the impact of a few extremely highly expressed genes on the global library to eliminate errors in a more robust way. Although the global influence of highly expressed genes could be eliminated using median metric, the geometric average method is too coarse to take consideration into account of the influence of normally expressed genes [16]. The TMM method weights each gene, but it fails to consider the overall influence of remaining genes in the whole library.

Here, a double-weighted normalization method for high-throughput sequencing data generated by RNA-seq is proposed. This method not only considers the overall expression level of all genes in the library, but also considers the impact of each individual gene. It can provide the dataset with high confidence for the analysis of differential expression and promote the research related to gene expression regulation.

Materials and methods

Double-weighted normalization method

Given a $n \times m$ matrix, each row represents a gene, each column represents a sequencing library, each element k_{gj} represents the number of sequences of gene *i* in sequencing library *j*, *g* = 1, …, *n*, *j* = 1, …, *m*. The global normalization method yields a global factor for each sequencing library (each column), which can adjust the expression levels of all genes in the sequencing library. To normalize sequencing library *j*, the sum of the number of sequences of all genes $s_j = \sum_g k_{gj}$ is normally considered as an estimate of the capacity of the sequencing library. However, as a sequencing library may be dominated by a small number of highly expressed genes, this kind of normalization cannot be applied on such data.

Here, we propose a double-weighted normalization method to estimate the normalization factor in a more robust way, which eliminates the impact of a few extremely highly expressed genes to provide a better estimation of the capacity of the sequencing library. For library *j*, two normalization factors are calculated:

$$f_j = \beta_j \cdot f_j^{(1)}. \tag{1}$$

Our method contains two weighting steps, the first step is calculating the normalization factor $f_j^{(1)}$ to weight for each gene. The second step is β weighting, which is a second weight for $f_j^{(1)}$ to adjust the overall expression level of the entire library.

(1) Calculation of the normalization factor $f^{(1)}$

M/A values of gene expression data from next-generation sequencing are used to calculate the adjusted factor based on the principle of MA plot that widely used in microarray analysis [6]. For microarray normalization, MA plot is normally used to characterize the distribution of the intensity ratio (*M*) of the Cy5 (red) to the Cy3 (green) and the average intensity value (*A*) of a two-color fluorescent hybrid gene chip [6], here:

$$M = \log_2 R - \log_2 G,\ A = (\log_2 R + \log_2 G)/2. \tag{2}$$

The general assumption for many microarray experiments on gene expression is that most genes are not differentially expressed, so *M* values of most genes should be in the vicinity of zero. Otherwise further normalization for the subsequent statistical analysis is required. For sequencing data, the logarithmic ratio *M* of gene *g* in library *i* to *j* and the absolute value of expression level A are defined by:

$$M_g = \log_2(k_{gi}/s_i) - \log 2(k_{gj}/s_j),\ A_g = \left(\log_2(k_{gi}/s_i) + \log 2(k_{gj}/s_j)\right)/2. \tag{3}$$

As logarithmic computation is required to calculate *M* and *A* values, genes that are not expressed in library *i* or *j* are removed by first. Next, genes with too large or too small *M* or *A* values are also discarded to prevent their effects on the adjustment of the whole library. Set A_0 as the minimum of *A* and M_0 as the minimum of *M*, lowly expressed genes that meet $A_g < A_0$ and $M_g < M_0$ are removed. Here, A_0 and M_0 can be set as the 5% or 1% quantile of *A* and *M* values, and

$$N_{A0} = \underset{g}{count}(A_g < A_0),\ N_{M0} = \underset{g}{count}(M_g < M_0). \tag{4}$$

Set the number of highly expressed genes to be removed equal with the number of lowly expressed genes to be removed, then M_g and A_g values are sorted to filter genes.

The next step is to weight each gene using the delta method [15] to estimate the approximate asymptotic variance and set the corresponding weight for each gene. Assuming that random variables X_n obey the binomial distribution with parameters *p* and *n*, then the approximate estimation of $\log(X_n/n)$ is $(1-p)/pn$. If the ratio of $\hat{p}$ to $\hat{q}$ is the estimation of the ratio of two independent samples of size *m* and *n*, then the estimated relative risk, $\hat{p}/\hat{q}$, approximately obeys the normal distribution of the variance of $(1-\hat{p})/\hat{p}n + (1-\hat{q})/\hat{q}m$. In our model, considering that the number of gene sequences in sequencing library *i* obeys the binomial distribution of parameters k_{gi}/s_i and s_i, then the variance of ratio of the sequence number of gene *g* in library *i* to *j* is estimated as:

$$\begin{aligned} v_g &= (1-\hat{p})/\hat{p}n + (1-\hat{q})/\hat{q}m = \\ &= (1-k_{gi}/s_i)/[(k_{gi}/s_i)s_i] + (1-k_{gj}/s_j)/[(k_{gj}/s_j)s_j] = \\ &= \frac{s_i - k_{gi}}{s_i k_{gi}} + \frac{s_j - k_{gj}}{s_j k_{gj}} \end{aligned} \tag{5}$$

Then the weight is set as the reciprocal of variance: $w_g = 1/v_g$.

Let sample *i* be the reference sample, the normalization adjusted factor of sample *j* is:

$$f_j^{(1)} = 2 \wedge \left(\frac{\sum_{g \in G} w_{gj} M_{gj}}{\sum_{g \in G} w_{gj}} \right), \tag{6}$$

where $M_{gj} = \log_2(k_{gj}/s_j) - \log 2(k_{gi}/s_i)$ and $w_{gj} = 1/\left(\frac{s_j - k_{gj}}{s_j k_{gj}} + \frac{s_i - k_{gi}}{s_i k_{gi}} \right)$, G denotes the remaining set of genes after removing genes according to the *M* value and *A* value.

(2) Estimation of the second weight β

For the set of remaining genes *G*, first the geometric average of each gene in all libraries ($r = 1, ..., m$) is set as a benchmark. Then the expression level of a gene is divided by this benchmarking level to get a scaling factor. Finally the weight of each column (each library) is the median of all scaling factors of this library:

$$\beta_j = \underset{g \in G}{median} \left\{ k_{gi} / \sqrt[m]{\prod_{r=1}^{m} k_{gr}} \right\}. \tag{7}$$

(3) Normalization of gene expression levels

The final normalization factor of library *j* is:

$$f_j = \underset{g \in G}{median} \left\{ k_{gi} / \sqrt[m]{\prod_{r=1}^{m} k_{gr}} \right\} \cdot \left[2 \wedge \left(\frac{\sum_{g \in G} w_{gj} M_{gj}}{\sum_{g \in G} w_{gj}} \right) \right]. \tag{8}$$

The adjusted capacity of each sequencing library *j* is estimated as $\hat{s}_j = f_j \sum_g k_{gj}$. The expression level of gene *g* after the global normalization using the adjusted library is $k_{gj}^{'} = f_j \cdot k_{gj}$. Only one adjusted factor is required to compare two samples. To normalize multiple sequencing libraries, a sample is selected as the reference by first, and then adjusted factors of other sequencing libraries are calculated on the reference library. This factor can be used to adjust the sample size of each sequence library for further analysis of differential expression.

Detection of differentially expressed genes based on a p-value weighted method

Currently, there are several methods for the detection of differential expression [1, 4, 6, 8, 11, 18, 19]. Both edgeR [17, 18] and DESeq [1] are based on the negative binomial distribution model and use precise testing method to infer differential expression. A *p*-value is obtained for each gene and genes are filtered according to their p-values. Let p_{DE} and p_{ed} be *p*-values of each gene that are obtained by edgeR [17] and DESeq [1], respectively. Here, we present a *p*-value weighted method which weights p_{DE} and p_{ed} in different ways to get an adjusted *p*-value p_f:

$$p_f = \alpha p_{DE} + \beta p_{ed} \tag{9}$$

where α and β are weights of p_{DE} and p_{ed}, respectively. The sum of them is 1. We use the following methods to determine α and β.

(1) Considering the simplest case $\alpha = \beta = 0.5$, i.e., p_f is the average value of p_{DE} and p_{ed}. Similarly, p_f is the geometric average, $p_f = \sqrt{p_{DE}\,p_{ed}}$.

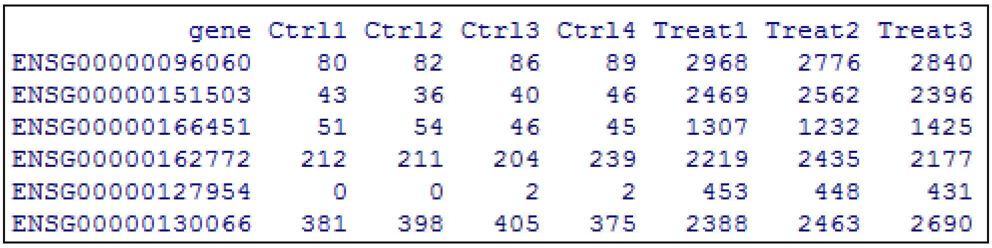
(2) Taking into account that the smaller the *p*-value is, the greater the possibility of differential expression of a gene and the corresponding weight is, we set $\alpha = p_{ed}/(p_{DE} + p_{ed})$ and $\beta = p_{DE}/(p_{DE} + p_{ed})$, then get $p_f = 2p_{DE}\,p_{ed}/(p_{DE} + p_{ed})$. After obtaining the adjusted *p*-value, the Bonferroni correction [2] is used to control the false discovery rate (FDR) for multiple comparisons. We then obtain a new FDR value for detecting differentially expressed genes. The selection of different weighting ways will affect the performance of the *p*-value weighted method.

Results and discussion

Data preprocessing

The raw RNA-seq data is provided by Li et al. [9], which consists of 7 lanes of 35 bp reads. A data set containing approximately 10 million sequence tags was generated from both control and hormone-treated cells (Treat), which is sufficient for quantitative analysis of gene expression. The raw data should be filtered and normalized before downstream analysis. Since the analysis of genes with very low expression level is normally not statistically significant [16], these genes should be discarded from further analysis. We filtered out very lowly expressed tags, keeping genes that were expressed at a reasonable level in at least one treatment condition. Since the smallest group size is three, we kept genes that achieved at least one tag per million (TPM) in at least three groups.

The raw data is comprised of 37435 genes (rows), the number of valid genes after filtering is 16494. As shown in Fig. 1, the first column denotes the gene ID, each row represents the number of reads of a gene under different conditions. Ctrl1~4 denote the four control groups and Treat1~3 represent the three experimental groups.



```
           gene Ctrl1 Ctrl2 Ctrl3 Ctrl4 Treat1 Treat2 Treat3
ENSG00000096060    80    82    86    89   2968   2776   2840
ENSG00000151503    43    36    40    46   2469   2562   2396
ENSG00000166451    51    54    46    45   1307   1232   1425
ENSG00000162772   212   211   204   239   2219   2435   2177
ENSG00000127954     0     0     2     2    453    448    431
ENSG00000130066   381   398   405   375   2388   2463   2690
```

Fig. 1 Format of the RNA-seq data after pre-processing

Data quality assessment is an essential step in any data analysis, which should typically be performed early in the analysis. To explore the count table, we applied the dist function [3] to the count matrix to get sample-to-sample distances and plotted a heat map for an overview over similarities and dissimilarities between samples. As can be seen from Fig. 2A, the clustering correctly reflects the experimental design, where samples are more similar when they have the same treatment. We also used the principal component plot [1] to visualize the overall effect of experimental covariates and batch effects (Fig. 2B), where no batch effects besides the known effects are observed. Therefore, subsequent normalization and differential expression testing was applied on this data set.

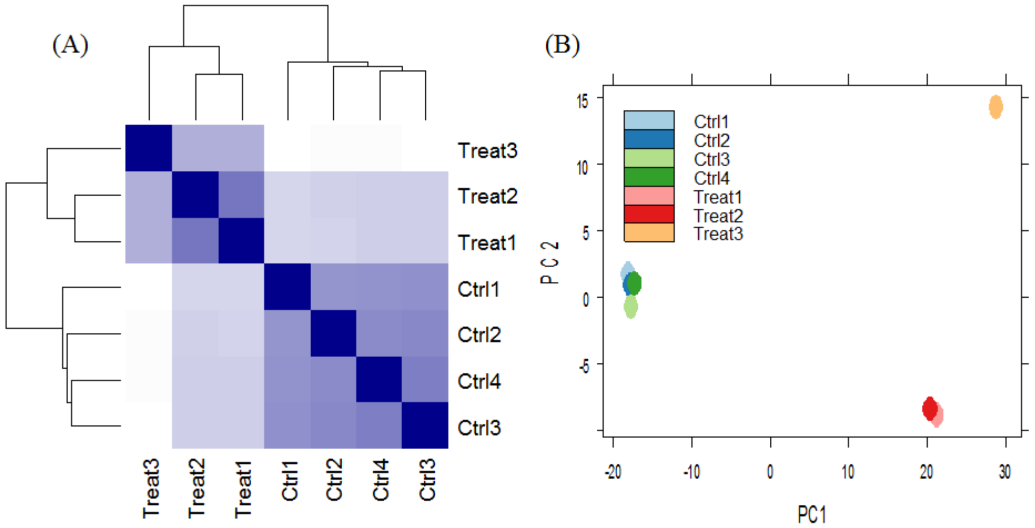


Fig. 2 Data quality assessment
(A) Heat map of the Euclidean distances between samples
(B) Principal component plot for the first two principals

Normalization of RNA-seq data

To compare the two libraries, our double-weighted method was first used to normalize the data. Fig. 3 shows the MA plot before and after normalization. There is a downward shift of the log ratio, probably due to that most reads are dominated by some highly expressed genes. MA values after normalization in Fig. 3B show that the majority of M values are near zero, which is consistent with the normalization assumption that differential expression only occurs in a few genes.

It can be seen from Fig. 4 that median values of experimental data after normalization are all in the same level, which suggests that our double-weighted method could help to normalize the sequencing depth under different conditions to make the libraries under different conditions comparable.

Detection of differential expression

We then conducted the *p*-value weighted test to infer differential expression between the control and treatment. The smear plot in Fig. 5A shows the log-fold changes with differentially expressed genes highlighted. It can be seen that the lower the level of gene expression is, the greater difference the sample size required to call the differential expression is. This is because that the error from the sequencing will be large if the number of genes in a sample is too small. Therefore, the sample of small size is insufficient to determine the differential expression, while a relatively small difference is allowed in the detection of differentially expressed genes with high expression level. Lines in Fig. 5A denoting 2-fold changes highlight differential expressed genes, which are in accord with results of the multiplicity method [13, 14] for differential expression detection. Multiplicity method is a relatively simple and straightforward method for detecting differential expressed genes [16], which considers genes with fold-change greater than 2 or less than 0.5 as differential expressed genes. However, the result of this method is too simple to find clues for higher level functions. Moreover, except for differentially expressed genes with significant changes, the reliability of the other differentially expressed genes with a small change is relatively low.

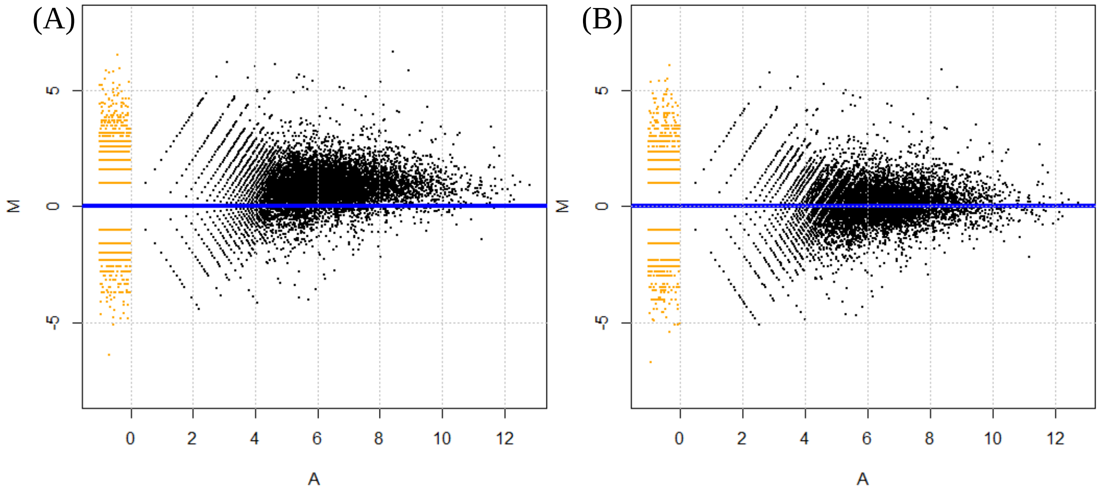


Fig. 3 MA plots before (A) and after (B) normalization. Blue line marks M = 0.

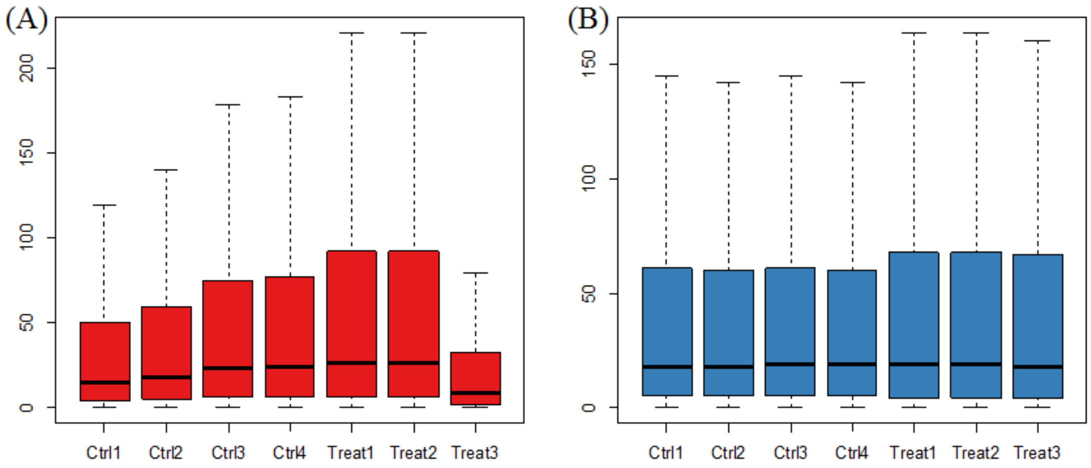


Fig. 4 Box plots of four samples of control group and three samples of treat group before (A) and after (B) normalization

As can be seen from Fig. 5B that frequencies are significantly higher when *p*-values are close to 0 or 1, while frequencies are more uniformly distributed when *p*-values are between 0 to 1. In fact, the *p*-value is supposed to obey the uniform distribution when there is no differential expression. Therefore, the high frequency at the low *p*-value is caused by differentially expressed genes, whereas the high frequency at *p*-value = 1 is caused by a small number of genes.

We also compared differential expression results from our *p*-value weighted method with DESeq and EdgeR. Top 1000 (top 1500 or 2000 genes were also tested, but the results showed no difference) differentially expressed genes were selected according to their FDR values from each method to count the overlapping**.** From these 1000 genes, 634 genes are found by all the three methods (Fig. 6A). The overlapping between EdgeR and other two methods is relatively low, where 254 genes are not overlapped. Up to 967 and 926 genes detected by our *p*-value weighted method can also be found by DESeq and EdgeR, respectively. The extent of overlapping of our method is the highest among these three methods (Fig. 6B), indicating that our method is the most robust in the detection of differential expression.

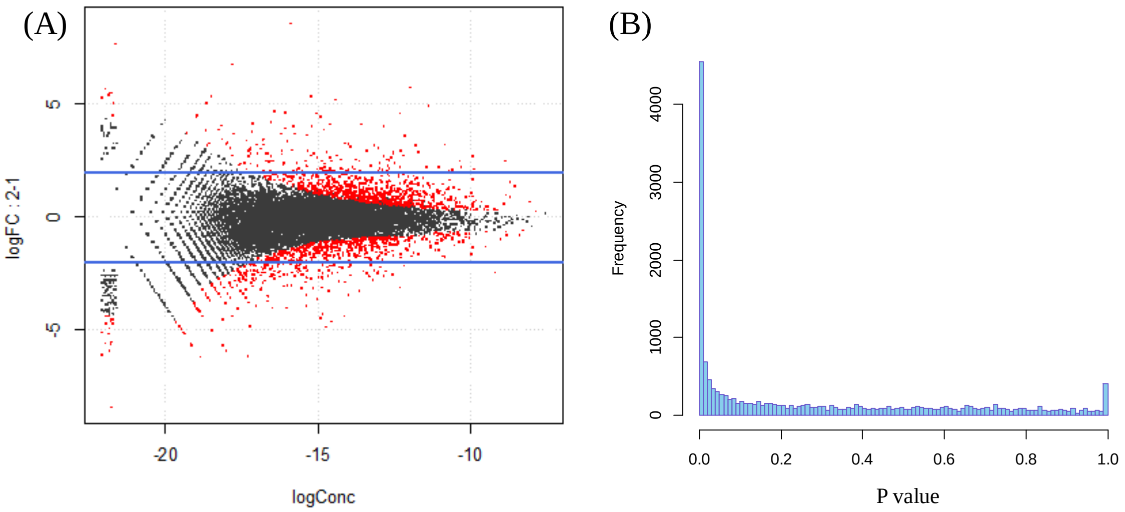


Fig. 5 (A) Smear plot showing the log-fold changes with differentially expressed genes highlighted. Blue lines indicate 2-fold changes;
(B) Histogram of *p*-values from the *p*-value weighted test

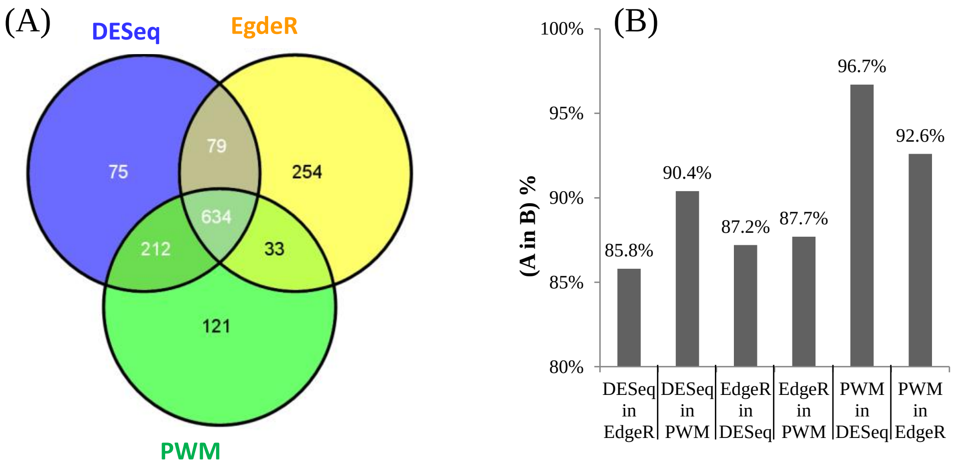


Fig. 6 Comparison of different methods for calling differential expression
(A) Venn diagram showing the overlapping among DESeq, EdgeR and *p*-value weighted method (PWM);
(B) Number of overlapping differential expressed genes between each two methods

Conclusion

In this paper, a double-weighted normalization method for high-throughput sequencing data was proposed. This method considers both the overall expression level of all genes in the library as well as the impact of each individual gene, which provides data with high confidence for the subsequent bioinformatics analysis. In contrast, two other widely used methods, DESeq and EdgeR, were both implemented by statistical tests and were sensitive to the number of replicates of samples. Experimental results show that our method can effectively normalize sequencing data under different conditions and provide highly confident data for the detection of differential expression. We also proposed the *p*-value weighted method to improve the existing algorithm for detecting differentially expressed genes with statistical significance. Experimental results show the robustness of the proposed *p*-value weighted method in that the overlapping rate of differential expressed genes of our method is higher than that of DESeq and EdgeR. In summary, this work provides a pipeline for the analysis of sequencing data, which will promote the research related to gene expression regulation to some extent.

Acknowledgements

This work was funded by the National Natural Science Foundation of China (No. 61201358), the Natural Science Foundation of Fujian Province of China (No. 2012J01154), the specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20120121120038), and the Fundamental Research Funds for the Central Universities in China (Xiamen University, No. 2013121025).

References

1. Anders S., W. Huber (2010). Differential Expression Analysis for Sequence Count Data, Genome Biology, 11, R106, doi: 10.1186/gb-2010-11-10-r106.
2. Benjamini Y., Y. Hochberg (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society. Series B (Methodological), 289-300.
3. Borg I., P. J. Groenen (2005). Modern Multidimensional Scaling: Theory and Applications, Springer, Verlag.
4. Bullard J. H., E. Purdom, K. D. Hansen, S. Dudoit (2010). Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-seq Experiments, BMC Bioinformatics, 11, 94, doi: 10.1186/1471-2105-11-94.
5. Chua S.-W., P. Vijayakumar, P. M. Nissom, C.-Y. Yam, V. V. T. Wong, H. Yang (2006). A Novel Normalization Method for Effective Removal of Systematic Variation in Microarray Data, Nucleic Acids Research, 34(5), e38, doi: 10.1093/nar/gkl024.
6. Dudoit S., Y. H. Yang, M. J. Callow, T. P. Speed (2002). Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments, Statistica Sinica, 12, 111-139.
7. Filloux C., M. Cédric, P. Romain, F. Lionel, K. Christophe, R. Dominique, M. Abderrahman, P. Daniel (2014). An Integrative Method to Normalize RNA-seq Data, BMC Bioinformatics, 15, 188, doi: 10.1186/1471-2105-15-188.
8. Hardcastle T., K. Kelly (2010). baySeq: Empirical Bayesian Methods for Identifying Differential Expression in Sequence Count Data, BMC Bioinformatics, 11, 422, doi:10.1186/1471-2105-11-422.
9. Li H., M. T. Lovci, Y.-S. Kwon, M. G. Rosenfeld, X.-D. Fu, G. W. Yeo (2008). Determination of Tag Density Required for Digital Transcriptome Analysis: Application to an Androgen-sensitive Prostate Cancer Model, Proceedings of the National Academy of Sciences of the USA, 105(51), 20179-20184.
10. Li X., H. Lu, M. Wang (2013). A Hybrid Gene Selection Method for Multi-category Tumor Classification using Microarray Data, International Journal Bioautomation, 17, 249-258.
11. Lu J., J. Tomfohr, T. Kepler (2005). Identifying Differential Expression in Multiple SAGE Libraries: An Overdispersed Log-linear Model Approach, BMC Bioinformatics, 6, 165, doi: 10.1186/1471-2105-6-165.
12. Marguerat S., J. Bähler (2010). RNA-seq: From Technology to Biology, Cellular and Molecular Life Sciences, 67, 569-579.

13. Marioni J. C., C. E. Mason, S. M. Mane, M. Stephens, Y. Gilad (2008). RNA-seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays, Genome Research, 18, 1509-1517.
14. Mortazavi A., B. A. Williams, K. McCue, L. Schaeffer, B. Wold (2008). Mapping and Quantifying Mammalian Transcriptomes by RNA-seq, Nature Methods, 5, 621-628.
15. Oehlert G. W. (1992). A Note on the Delta Method, The American Statistician, 46, 27-29.
16. Robinson M., A. Oshlack (2010). A Scaling Normalization Method for Differential Expression Analysis of RNA-seq Data, Genome Biology, 11, R25, doi: 10.1186/gb-2010-11-3-r25.
17. Robinson M. D., D. J. McCarthy, G. K. Smyth (2010). EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data, Bioinformatics, 26, 139-140.
18. Robinson M. D., G. K. Smyth (2007). Moderated Statistical Tests for Assessing Differences in Tag Abundance, Bioinformatics, 23, 2881-2887.
19. Wang L., Z. Feng, X. Wang, X. Wang, X. Zhang (2010). DEGseq: An R Package for Identifying Differentially Expressed Genes from RNA-seq Data, Bioinformatics, 26, 136-138.

Assist. Prof. Xiaohui Wu, Ph.D.
E-mail: xhuister@xmu.edu.cn

Xiaohui Wu received her B.Sc. degree (2006) and Ph.D. degree (2011) from Xiamen University, China. She is currently an Assistant Professor with the Department of Automation in Xiamen University. Her research interests are bioinformatics and biocomputing.

Chuang Zhao
E-mail: zhaochuang200808@163.com

Chuang Zhao is a graduate student with the Department of Automation, Xiamen University, China. His research interests are bioinformatics and artificial intelligence.

Yaru Su, Ph.D.
E-mail: yarusu@gmail.com

Yaru Su received her B.Sc. degree (2006) in Automation, Xiamen University, and Ph.D. degree (2009) in Pattern Recognition and Intelligent System, University of Science and Technology of China. She currently works at Forensic Science Division, Department of Fujian Provincial Public Security. Her research interests include bioinformatics, data mining, pattern recognition, and machine learning.