ISSN : 0974 - 7435 Volume 8 Issue 4

BioTechnology
An Indian Journal

FULL PAPER

BTAIJ, 8(4), 2013 [518-522]

A normalization method based on variance and median adjustment for massive mRNA polyadenylation data

Guoli Ji, Ying Wang, Mingchen Wu, Yangzi Zhang, Xiaohui Wu*
Department of Automation, Xiamen University, Xiamen 361005, Fujian, (CHINA)
E-mail : xhuister@xmu.edu.cn.

ABSTRACT

This paper proposed a normalization method based on minimum variance and median adjustment (MVM), and then made a comprehensive comparison of three normalization methods including DESeq, TMM and MVM. In this study, the MVM method was evaluated using polyadenylation [poly(A)] data and gene expression data from Arabidopsis by ways of empirical statistical criterias of mean square error (MSE) and Kolmogorov-Smirnov (K-S) statistic. Experimental results demonstrated the high performance of MVM method in that it could accurately remove the systematic bias and make the distributions of normalized data stable.

KEYWORDS

Normalization;
Minimum variance and median adjustment method;
Assessment;
MSE;
K-S test.

INTRODUCTION

With the improving of the next high-throughput generation sequencing (NGS), the RNA sequencing technology (RNA-seq) has developed[1, 2]. The systematic variation of RNA-seq data can be eliminated by effective normalization methods[3-5], which are divided into within-library method and between-library method. The within-library normalization method is able to make accurate comparisons of the gene-level expression within sample, but this method can not be used in differential analysis. The between-library normalization method uses total numbers of reads to balance the sample expression, so it is commonly used in the RNA-seq analysis.

The existing normalization methods for RNA-seq studies include TC (Total number of reads Count)[1], Q (Quaritile)[6], UQ (Upper Quaritile)[7], Med (Median), DESeq[8], TMM (Trimmed Mean of M values)[9] and RPKM (Reads Per Kilobase per Million mapped reads)[10] normalization. Both of the TC and RPKM methods are widely used, however, they are sensitive to the presence of majority genes and ineffective in gene differential analysis. The DESeq and TMM methods are implementation of statistical tests using NB distribution, producing similar results when the library compositions are robust[4]. In this study, the between-library normalization method was used to remove between-library variation of polydenylation [poly(A)] data and gene expression data. A minimum variance and median adjustment method (MVM) that based on DESeq and TMM methods was proposed to normalize poly (A) data and gene expression data. The performance of this MVM method was assessed by ways of data distributions and empirical statistical criterias of mean square

error (MSE) and Kolmogorov-Smirnov (K-S)[11].

DATA AND METHODS

Data description

The poly (A) data and gene expression data from Arabidopsis under different conditions were used (unpublished data) in this study. Each dataset includes four sequencing samples with different conditions, each condition contains three biological replicates, denoted it as wt1~3, oxt1~3, g1~3, gm1~3. Each row of poly (A) data represents a poly (A) site, each column represents the expression of the corresponding poly (A) site. Each row of gene expression data represents a gene, each column represents the expression of relevant gene.

Normalization methods

The process of the MVM method is shown in Figure 1.

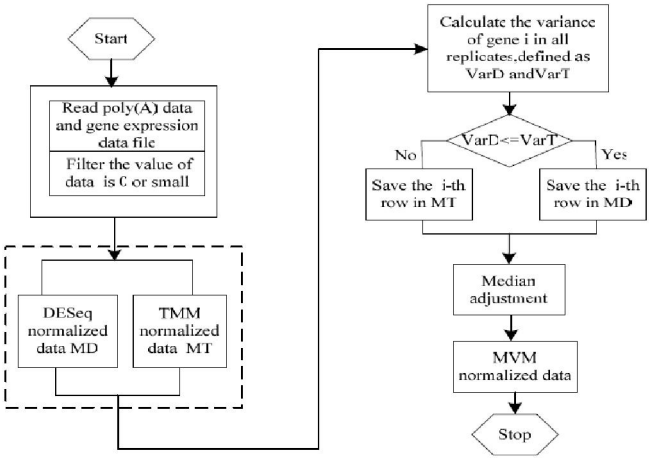

Figure 1 : The framework of the minimum variance and median adjustment method.

The DESeq normalization method[8] is based on the hypothesis that most genes in different samples are not differentially expressed, the underlying meaning of which is distributions of data are steady across samples. The DESeq normalization data can be received by dividing raw data by a sample-specific normalization factor, this factor is calculated as Eq. (1):

$$s_j = median_i\left(m_{ij} \Big/ \left(\prod_{k=1}^{n} m_{ik}\right)^{1/n}\right) \quad (1)$$

Where the scaling factor s_j is the DESeq normalization factor, m_{ij} is the expression of gene *i* in sample *j*, *n* is the number of samples in the experiment.

The TMM normalization method[12] is also based on the hypothesis that most genes are similar expressed. The procedure of TMM method is doubly trimmed, by default, trimming off M-value by 30% and A-value by 5%[9]. The TMM factor is calculated as Eqs. (2):

$$s_j = \frac{\sum_{i\in G'}\left(\left(\frac{N_k - m_{ik}}{N_k m_{ik}} + \frac{N'_k - m'_{ik}}{N'_k m'_{ik}}\right)\bullet\left(\frac{\log_2(m_{ik}/N_k)}{\log_2(m'_{ik}/N'_k)}\right)\right)}{\sum_{i\in G'}\left(\frac{N_k - m_{ik}}{N_k m_{ik}} + \frac{N'_k - m'_{ik}}{N'_k m'_{ik}}\right)} \quad (2)$$

Where s_j represents the scaling factor, G' represents genes set after removing the data whose value is 0. These TMM normalization factors should be re-scaled by the mean of the effective library sizes[4]. The normalized data set are obtained by scaling raw data by these re-scaled factors.

The MVM normalized value can be calculated as Eqs. (3):

$$m'_{ij} = \frac{m_{ij}}{median_j \Big/ \left(\frac{1}{n}\sum_i median_j\right)} \quad (3)$$

Where $median_j$ represents the median of genes expression in *j* sample, *n* represents number of experiment samples, m'_{ij} represents the expression level of *i* gene under *j* sample with MVM method.

This article implemented the MVM method using a series of R scripts with empirical statistical metrics. Both of the DESeq and TMM methods are implemented in appropriate R Bioconductor libraries. In package DESeq, calling *estimateSizeFactors()* and *sizeFactors()* functions can estimate the sample-specific normalization factors, then calling *counts()* function and setting the *normalized* parameter to TRUE, this study receives DESeq normalized data. The TMM normalization method is included in TMM package. The *calcNormFactors()* function provides TMM scale factors. The TMM normalized data are obtained by scaling raw data by re-scaled factors.

Appraisal procedures of the normalization methods

By examining the boxplots of data across samples, both before and after normalization, this study preliminary evaluates data discretization. If the method is an effective normalization scheme, the data distributions across samples should be stable. Moreover, the scatter plots are also used to describe data distributions quan-

FULL PAPER

titatively, as an effective normalization method should make bulk of M-value to lay on the horizontal line indicating equal value in two comparable samples.

Also, the MSE can be used to estimate the variation of the normalization data. In empirical statistics analysis, MSE is defined as the summation of variance and the square of bias, where the variance is a metric for precision and the bias is a criterion for accuracy. A small MSE shows that the difference of data between samples is small, the normalization method is better in overall[5]. The K-S statistic is another comparison criterion. By calculating the largest deviation statistic value D between two accumulative distributions, the K-S statistic measures similarities between these samples. An effective normalization method should produce smaller D value.

RESULTS

Profiles of the data before normalization

The M-values distribution between two technical replicates of the raw gene expression data is showed in Figure 2A, where the M value is calculated as log2 (read data). The M-values distribution is mostly centralized around zero, indicating that there is no significant difference expression in these technical replicates. However, Figure 2B shows that the log ratios between the two comparison biological replicates have significantly partial to the condition with higher expression genes. The M-value distributions between the two comparison biological replicates of gene expression (Figure 2C) and poly (A) (Figure 2D) data show the centers of M-values before normalization are deviated from zero. The median values are 1.15 for gene expression data and 1.66 for poly (A) data. Therefore, an effective normalization procedure for these data is needed in our analysis.

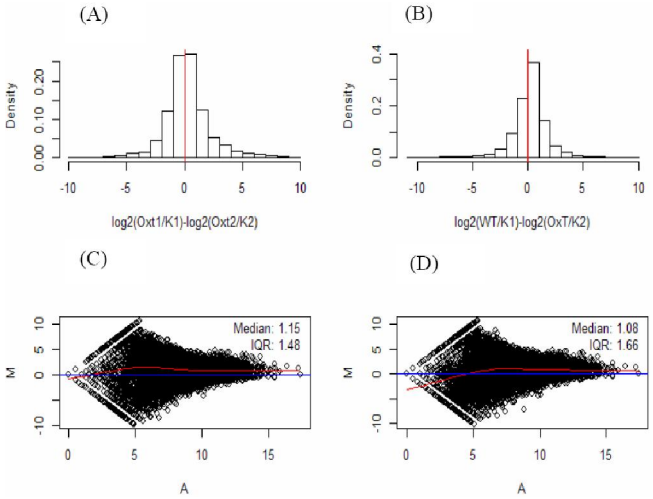

Figure 2 : Profiles of the data before normalization. Comparison on the log ratios of (A) technical replicates and (B) biological replicates expression levels. M-A plot of biological comparison replicates in (C) gene expression and (D) poly(A) data.

Assessment of the MVM normalization method

Comparison of methods with data distributions

As to the poly (A) data, Figure 3A shows the distributions of data across replicates, both before and after normalization. The DESeq and TMM methods appear to perform similarly. The distributions of MVM normalized data from different samples are more stable than DESeq and TMM methods, illustrating that the difference among these samples is minimum. In the M-A plot (Figure 3B), the M-values of the raw data are not centered on M = 0, the M-values of normalization data are all centered close to zero. Especially, in the M-A plot after MVM normalization, the bulk of data lie on the line of zero. All the results consistently exhib-

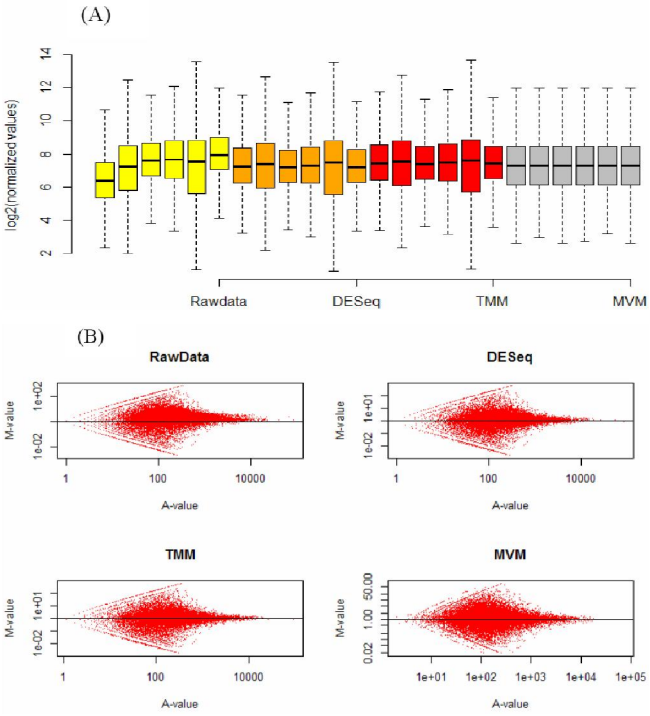

Figure 3 : Comparison of methods with data distributions. (A) Boxplots of data both before and after normalization for all replicates to assess. (B) M-A plots of the same data.

ited that the MVM is an effective normalization method.

Comparison of methods with empirical statistical criteria

As to the gene expression data, Figure 4 presents the results assessed by methods of MSE and K-S statistics. The bias and MSE of raw data both are large (Figure 4A); in contrast, the MVM method could remove the bias and minimize MSE effectively. As shown in Figure 4B, the K-S statistics value from all of these methods is much lower than the raw data. The D value from our MVM method is least among all these methods, showing the difference of data from these biological replicates is smaller.

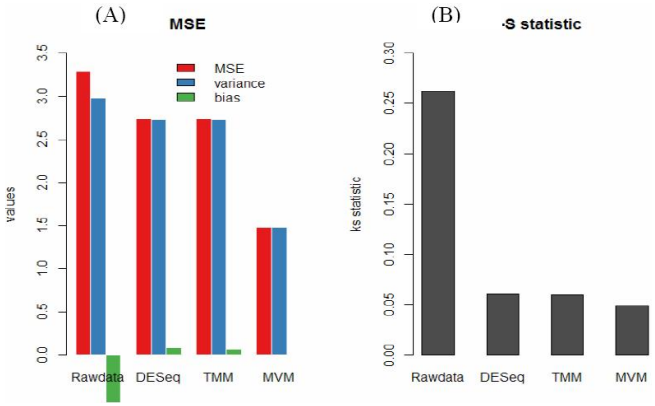

Figure 4 : Comparison of methods with statistical criterias. Bar graphs show the results of (A) MSEs and (B) K-S statistics.

CONCLUSION

There are many deviation and uncertainties factors in RNA-seq data, hence the raw data within and between samples can not be directly compared. The MVM method that based on DESeq and TMM methods was proposed to normalize the poly (A) data and gene expression data form Arabidopsis. Then, this MVM method was evaluated by different assessment criteria, including data distributions both before and after normalization, MSE and K-S statistics. These evaluation results showed that the normalization continues to be an essential step in RNA-seq data analysis. The DESeq and TMM normalization methods can produce similar results in differential analysis, but both of which are ineffective in poly (A) data and gene expression data analysis. The MVM method can effectively remove the system variation and minimize the difference of gene expression across replicates. In this study, the MVM method turned in a good performance on stabilization of data distributions across replicates.

ACKNOWLEDGEMENTS

This project was funded by the National Natural Science Foundation of China (Nos. 61174161, 61201358 and 61203176), the Natural Science Foundation of Fujian Province of China (No. 2012J01154), the specialized Research Fund for the Doctoral Program of Higher Education of China (Nos. 20100121120022 and 20120121120038), the Key Research Project of Xiamen City of China (No. 3502Z20123014), and the Fundamental Research Funds for the Central Universities in China (Xiamen University: Nos. 2011121047, 2013121025 and CBX2013015).

REFERENCES

[1] J.C.Marioni, et al.; RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res, **18(9)**, p. 1509-17 **(2008)**.

[2] S.Marguerat, J.Bahler; RNA-seq: from technology to biology. Cell Mol Life Sci, **67(4)**, 569-79 **(2010)**.

[3] S.W.Chua, et al.; A novel normalization method for effective removal of systematic variation in microarray data. Nucleic Acids Res, **34(5)**, e38 **(2006)**.

[4] M.A.Dillies, et al.; A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform, **(2012)**.

[5] L.X.Garmire, S.Subramaniam; Evaluation of normalization methods in mammalian microRNA-Seq data. RNA, **18(6)**, 1279-88 **(2012)**.

[6] B.M.Bolstad, et al.; A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, **19(2)**, 185-193 **(2003)**.

[7] J.H.Bullard, et al.; Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics, **11**, 94 **(2010)**.

[8] S.Anders, W.Huber; Differential expression analysis for sequence count data. Genome Biol, **11(10)**,

R106 **(2010)**.

[9] M.D.Robinson, A.Oshlack; A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol, **11(3)**, R25 **(2010)**.

[10] M.A, W. Ba, M.K; Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods, **(2008)**.

[11] Z.Sun, Y.Zhu; Systematic comparison of RNA-Seq normalization methods using measurement error models. Bioinformatics, **28(20)**, 2584-91 **(2012)**.

[12] M.D.Robinson, , D.J.McCarthy, G.K.Smyth; edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, **26(1)**, 139-40 **(2010)**.