NormExpression: an R package

to normalize gene expression data using evaluated methods

Zhenfeng Wu[12$], Weixiang Liu[34], Haishuo Ji[25], Deshui Yu[2],

Hua Wang[2], Liu Lin[2], Jishou Ruan[16*], Shan Gao[25*]

1. School of Mathematical Sciences, Nankai University, Tianjin 300071, P.R.China.
2. College of Life Sciences, Nankai University, Tianjin, Tianjin 300071, P.R.China.
3. School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, P.R.China.
4. Guangdong Provincial Key Laboratory of Biomedical Measurements and Ultrasound Imaging, Shenzhen 518060, P.R.China.
5. Institute of Statistics, Nankai University, Tianjin 300071, P.R.China.
6. State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin 300071, P.R.China.

[$] These authors contributed equally to this paper.

* The corresponding authors.

SG：gao_shan@mail.nankai.edu.cn

JR：jsruan@nankai.edu.cn

Abstract

Data normalization is a crucial step in the gene expression analysis as it determines the validity of its downstream analyses. Although many metrics has been designed to evaluate the relative success of these methods, the results by different metrics did not show consistency. Based on the previous work, we designed a new metric named Area Under normalized CV threshold Curve (AUCVC) to evaluate 13 commonly used normalization methods and achieved consistency in our evaluation results using both bulk RNA-seq and scRNA-seq data from the same library construction protocol. These gene expression data, normalization methods and evaluation metrics have been included in an R package named NormExpression. NormExpression provides a framework for researchers to select normalization methods with a fast and simple way to evaluate different methods, particularly some data-driven methods or their own methods.

Keyword: gene expression; normalization; evaluation; R package; scRNA-seq

Introduction

Global gene expression analysis provides quantitative information about the population of RNA species in cells and tissues [1]. High-throughput technologies to measure global gene expression levels started with Serial Analysis of Gene Expression method (SAGE) and are widely used with microarray and RNA-seq [2]. Recently, single-cell RNA sequencing (scRNA-seq) has been used to simultaneously measure the expression levels of genes from a single-cell and to provide a higher resolution of cellular differences than bulk RNA-seq, which can only produce an expression value for each gene by averaging its expression levels across a large population of cells [3]. Gene expression raw data from these high-throughput technologies must be normalized to remove technical variation so that meaningful biological comparisons can be made. Data normalization is a crucial step in the gene expression analysis as it determines the validity of its downstream analyses. Although the significance of gene expression data normalization has been demonstrated [4], how to successfully select a normalization method is still a controversial problem, particularly for scRNA-seq data.

Basically, two classes of methods are available to normalize gene expression data. They are the control-based normalization and the average-bulk normalization. The former class of methods assumes the total expression level summed over a small group of genes is approximately the same across all the samples. The latter class of methods assumes most of genes are not Differentially Expressed (DE) genes across all the samples. The control-based normalization often uses RNA from a group of internal control genes (*e.g.* housekeeping genes) or external spike−in RNA (*e.g.* ERCC RNA [5]), while the average-bulk normalization is more commonly used for their universality. Five average-bulk normalization methods designed to normalize bulk RNA-seq data are library size, median of the ratios of observed counts that is also referred to as the DESeq method [6], Relative Log Expression (RLE), upperquartile (UQ) and Trimmed Mean of M values (TMM) [7]. Recently, three new methods have been introduced as Total Ubiquitous (TU), Network Centrality Scaling (NCS) and Evolution Strategy (ES) with best performance among 15 tested methods [8].

Although many metrics has been designed to evaluate the relative success of these methods, the results by different metrics did not show consistency. In 2013, Gustavo *et al.* designed two novel and mutually independent metrics to evaluate 15 normalization methods and achieved consistent results using bulk RNA-seq data [8]. Based on their work, we designed a new metric named Area Under normalized CV threshold Curve (AUCVC) and tested it using both bulk RNA-seq and scRNA-seq data from the same library construction protocol. As a result, the evaluation by both our metric AUCVC and their metrics achieved consistency. On the other hand, with many new normalization methods developed, researchers need a fast and simple way to evaluate different methods, particularly some data-driven methods or their own methods rather than obtain information from published evaluation results, which could have bias or mistakes, *e.g.* misunderstanding of RLE, UQ and TMM methods [9]. To satisfy this demand, we developed an R package NormExpression to include gene expression data, normalization methods and evaluation metrics used in this study and provide a framework for researchers to evaluate and select methods for the normalization of their gene expression data.

A.

$$\begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{bmatrix} \times \begin{bmatrix} f_1 \\ \vdots \\ f_j \\ \vdots \\ f_n \end{bmatrix}$$

B.

$$For\ all\ methods: f_j = f_j \Big/ exp\left[\frac{1}{n}\sum_{j=1}^{n} log\left(f_j\right)\right]$$

$$Library\ size: \quad f_j = 10^6 / N_j^*$$

$$DESeq(RLE): \quad f_j = 1/s_j$$

$$s_j = exp\left[\underset{i}{median}\left(log\left(x_{ij}\right) - \frac{1}{n}\sum_{k=1}^{n} log\left(x_{ik}\right)\right)\right]$$

$$Upperquartile: \quad f_j = 10^6 / N_j s_j \qquad s_j = \underset{i}{Q_3}\left(x_{ij}/N_j\right)$$

Figure 1

Results

Normalization factor and library size

In total, 13 normalization methods (**Materials and Methods)** have been included in the R package NormExpression. They are Housekeeping Genes (HG7), External RNA Control Consortium (ERCC), Total Read Number (TN), Total Read Count (TC), Cellular RNA (CR), Nuclear RNA (NR), the ratios of observed counts (DESeq), Relative Log Expression (RLE), upperquartile (UQ), Trimmed Mean of M values (TMM), Total Ubiquitous (TU), Network Centrality Scaling (NCS) and Evolution Strategy (ES). Currently, all the commonly used methods are used to normalize a raw gene expression matrix (n samples by m genes) by the multiplication of a factor to each column of it and produce a normalized gene expression matrix (**Figure 1A**). This factor is named as normalization factor in the package NormExpression or scaling factor in TU, NCS and ES methods. In NormExpression, the reciprocal of normalization factor is named as library size (**Figure 1B)**, which is also named as size factor in the Bioconductor package DESeq [6]. Definitions of normalization factor and size factor in the Bioconductor package edgeR [7] are different from the definition of normalization factor in NormExpression and the definition of size factor in DESeq. RLE, UQ and TMM in edgeR produce normalization factors to adjust library sizes, which should be used to calculate the Counts Per Million (CPM) for the normalization of gene expression data and CPM should be calculated by one million multiplying reciprocals of adjusted library sizes (**Figure 1B**). However, edgeR provides a function named calcNormFactors to produce normalization factors for library-size adjustment, which have been wrongly used for the normalization of gene expression data in many studies [9]. Since ES. HG7, ERCC, TN, TC, CR, NR and TU produce normalization factors by the estimation of library sizes as CPM, their normalization factors are amplified by one million for a uniform representation (**Figure 1B**) in NormExpression. DESeq, RLE, UQ and TMM have been modified to ignore zero values for both scRNA-seq and bulk RNA-seq data and the resulting normalization factors need be further normalized by their geometric mean values (**Figure 1B**). UQ and TMM use library sizes estimated by NR. After modification, RLE is identical to DESeq. We verified that these modifications did not change the evaluation or normalization results.

Evaluation of 13 normalization methods

In the previous study [8], Gustavo *et al.* had quantified success of normalization methods by the number of uniform genes (**Materials and Methods)** and used the Coefficient of Variation (CV) cutoff 0.25 to determine the number of uniform genes for each method. This metric was designed based on the theory that the relative values among different normalization methods were quite stable, although the absolute number of uniform genes depended on the cutoff value. However, it is almost impossible to determine a CV cutoff for scRNA-seq data since the CV in scRNA-seq data has a much more large dynamic range than that in bulk RNA-seq data.

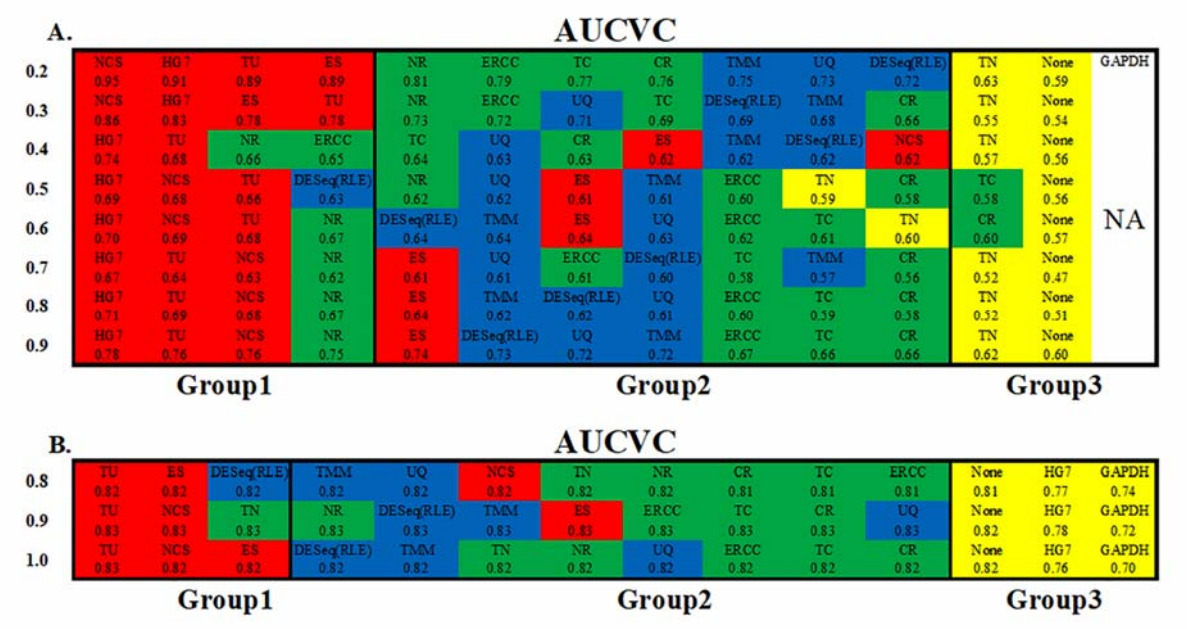

Figure 2.

Inspired by Area Under the receiver operating characteristic Curve (AUC) [10], we designed a new metric named Area Under normalized CV threshold Curve (AUCVC) to evaluate 13 normalization methods using one scRNA-seq dataset scRNA663 and one bulk RNA-seq dataset bkRNA18 (**Materials and Methods)**. A single housekeeping gene GAPDH was also used for comparision in the evaluation of normalization methods using bulk RNA-seq data, but it was not available for that using scRNA-seq data due to zero counts of GAPDH in many samples. Parameter grid of non-zero ratio (**Materials and Methods**) from 0.2 to 0.9 for scRNA663 and from 0.8 to 1 for bkRNA18 was used to

produce AUCVC values of all methods (**Figure 2**). For each non-zero ratio, TU used the maximum AUCVC, which had been determined by testing all possible combinations of presence rate, lower and upper cutoffs (**Materials and Methods)** at 5% resolution. The presence rate was tested from 0.2 to 0.6 for scRNA663 and set 1 for bkRNA18. The lower cutoff was tested from 5% to 40% and the upper cutoff was tested from 60% to 95%. In addition, the calculation only considered each combination of lower and upper cutoffs which produced ubiquitous genes (**Materials and Methods**) more than 1,00 for scRNA663 and more than 1,000 for bkRNA18. For each non-zero ratio, NCS and ES used the ubiquitous genes produced by the TU method, when it achieved the maximum AUCVC. The raw gene expression matrix (None) was also used to produce AUCVC values for comparison.

The evaluation results using both scRNA663 and bkRNA18 achieved consistency that all the normalization methods were classified into three groups (**Figure 2**) based on their AUCVC values sorted in descending order. The first group including TU, NCS and ES achieved the best performances using both scRNA663 and bkRNA18. The second group including ERCC, TC, CR, NR, DESeq, RLE, UQ and TMM achieved medial performances using both scRNA663 and bkRNA18. In the second group, ERCC, TC, CR and NR outperformed DESeq, RLE, UQ and TMM using scRNA663, while DESeq, RLE, UQ and TMM outperformed ERCC, TC, CR and NR using bkRNA18. The third group achieved the poorest performances, including TN and None for scRNA663 (**Figure 2A**) and HG7, GAPDH and None for bkRNA18 (**Figure 2B**). HG7 and GAPDH achieved the poorest performances using bkRNA18, which suggested that a predefined set of housekeeping genes could not be appropriate guides for data normalization of bulk RNA-seq data. However, it could be coincidental that HG7 was classified into the first group using scRNA663. TN outperformed the second group of methods using bkRNA18 but was outperformed by the second group of methods using scRNA663.

The evaluation results by the medians of Spearman Correlation Coefficients (SCCs) (**Materials and Methods**) and the cluster analysis results were also consistent with the evaluation results by AUCVC. Generally speaking, a normalization method with a higher AUCVC value produced a lower median of Spearman Correlation Coefficients (SCCs)

between normalized expression profiles of ubiquitous gene pairs using both scRNA-seq and bulk RNA-seq data. The hierarchical clustering result showed that 13 methods had been classified into the same groups (**Figure 3CF**) by SCCs between normalization factor pairs as those (**Figure 2AB**) by AUCVC. By our new designed metric AUCVC, TU, NCS and ES were evaluated as the best normalization methods using both scRNA-seq and bulk RNA-seq data, which enhanced the discovery using only bulk RNA-seq data in the previous study [8]. Since the non-zero ratio 0.2 allowed the maximum number of uniform genes for calculation, we presented this snapshot of evaluation results to show the consistency of the evaluation results using both scRNA663 (**Figure 3ABC**) and bkRNA18 (**Figure 3DEF**).

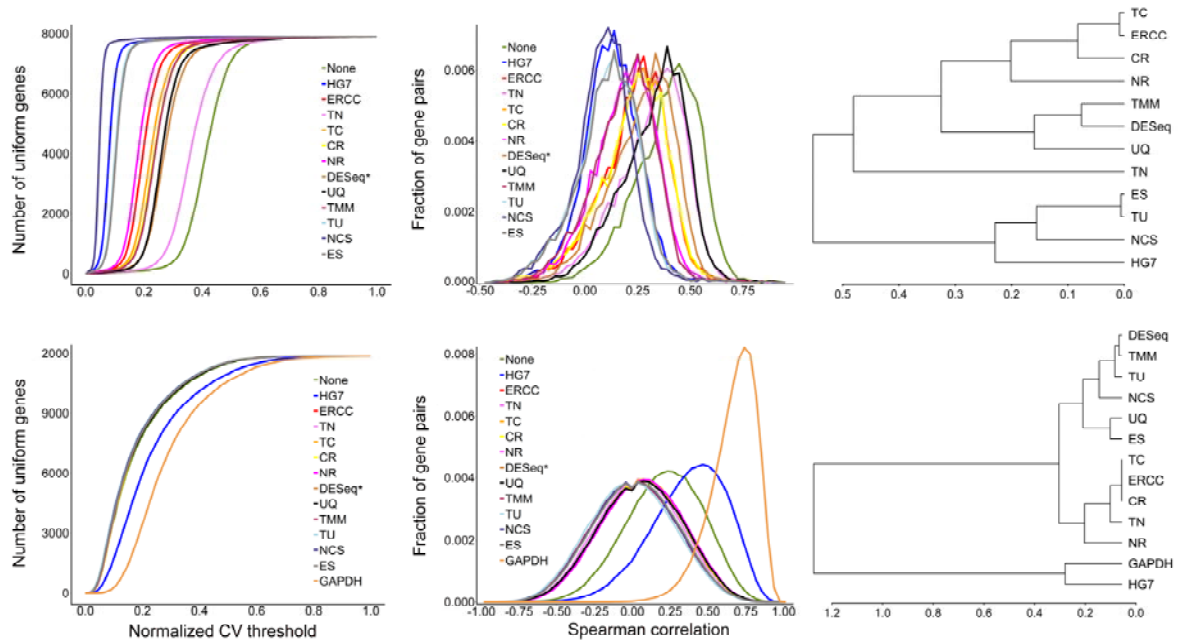

Figure 3.

Implementation and availability

The gene expression data (scRNA663 and bkRNA18), normalization methods and evaluation metrics (AUCVC and SCCs) have been included in the R package NormExpression. All the functions except the NCS and ES methods have been implemented in R programs [2] for their running on R platforms of any version. The NCS and ES methods had been implemented in Perl programs on the Linux system by Gustavo *et al.* [8] but they need be installed with many Perl modules. We have modified them into a stand-alone program (**Supplementary file 2**).

A quick evaluation is usually started with 10 normalization methods, which are HG7, ERCC (if available), TC, CR, NR, DESeq, RLE, UQ, TMM and TU. The quick evaluation produce AUCVC values of 10 methods and the raw gene expression matrix for users to evaluate and select methods. NCS and ES are not included in 10 methods, since they have a similar performance of TU but are much more time consuming. The non-zero ratio and presence rate can be set to 1 to calculate AUCVC for bulk RNA-seq data, while they need be set to appropriate values (default 0.2) for scRNA-seq data to avoid parameter grid. Based on our experiences, both non-zero ratio and presence rate need be set to the values to ensure that both the product of the sample number multiplying non-zero ratio and that of the sample number multiplying presence rate are larger than 100 for scRNA-seq data.

Materials and Methods

Datasets

In the previous study by Lin Liu *et al*. (SRA: SRP113436), 663 single-cell samples and 18 bulk samples had been sequenced using the Smart-seq2 scRNA-seq protocol. In this study, we built a scRNA-seq dataset including 653 single cells from colon tumor tissues and 10 single cells from distal tissues (>10 cm) as control. We also built a bulk RNA-seq dataset including nine samples from colon tumor tissues and nine samples from distal tissues. Samples with total read number less than 288,289 were removed in the data filtering step. The cleaning and quality control of both scRNA-seq and bulk RNA-seq data were performed using the pipeline Fastq_clean [15] that was optimized to clean the raw reads from Illumina platforms. Using the software STAR [11] v2.5.2b, we aligned all the cleaned scRNA-seq and bulk RNA-seq reads to the human genome GRCh38/hg38 and the expression levels of 57,992 annotated genes (57,955 nuclear genes and 37 mitochondrial genes) were quantified. Non-polyA RNAs were not discarded to test the robustness of normalization methods, although the Smart-seq2 protocol theoretically had only captured polyA RNAs. In addition, the expression levels of 92 ERCC RNA and the long non-coding RNA (lncRNA) MDL1 in human mitochondrial [12] were also quantified. ERCC RNA had been spiked into 208 single-cell samples before library construction, the expression levels

of ERCC RNA in other 455 single-cell samples and 18 bulk samples were simulated by linear regression. Finally, two datasets were named scRNA663 (58085 × 663) and bkRNA18 (58085 × 18) and included into the R package NormExpression.

Normalization methods

All 13 methods in the package NormExpression are HG7, ERCC, TN, TC, CR, NR, DESeq, RLE, UQ, TMM, TU, NCS and ES. HG7, ERCC, TN, TC, CR, NR and TU are based on a set of pre-selected genes and each of these methods uses the gene expression level summed over these pre-selected genes in a sample as the library size (**Figure 1B**) to calculate the normalization factor. HG7 includes seven genes (UBC, HMBS, TBP, GAPDH, HPRT1, RPL13A and ACTB), which had been used to achieve the best evaluation result among those using all possible combinations of tested housekeeping genes in the previous study by Gustavo *et al.* [8]. ERCC is a set of commonly used spike-in RNA consisting of 92 polyadenylated transcripts with short 3' polyA tails but without 5' caps [5]. The pre-selected genes used by HG7, ERCC, and TU are seven housekeeping genes, 92 ERCC RNA and ubiquitous genes (described below), respectively. NR only counts reads which have been aligned to nuclear genomes, while CR counts reads which have been aligned to both nuclear and mitochondrial genomes. The library size estimated by TC is equal to that estimated by CR plus that estimated by ERCC. TN uses the number of all reads which can be aligned to ERCC RNA, nuclear and mitochondrial genomes.

The DESeq method was obtained from the Bioconductor package DESeq [6] and modified to process scRNA-seq data. RLE, UQ and TMM were obtained from the Bioconductor package edgeR [7] and modified to process scRNA-seq data. TU, NCS and ES were obtained from the previous study by Gustavo *et al.* [8]. Since TU sums counts of all ubiquitous genes as the library size to calculate the normalization factor, a process to select ubiquitous genes (describe below) has been integrated into the TU method. TU maximizes AUCVC instead of the number of resulting uniform genes to select ubiquitous genes in the R package NormExpression.

Uniform genes and ubiquitous genes

A uniform gene was defined if the Coefficient of Variation (CV, **Formula 1**) of its post-normalization expression levels across all samples was not more than a cutoff. Ubiquitous genes were defined as the intersection of a trimmed sets of all samples [8]. This trimmed set of genes were selected for each sample by 1) excluding genes with zero values, 2) sorting the non-zero genes by expression level in that sample, and 3) removing the upper and lower ends of the sample-specific expression distribution. Gustavo *et al.* determined the upper and lower cutoffs by testing all possible combinations of lower and upper cutoffs at 5% resolution to maximize the number of resulting uniform genes using one bulk RNA-seq dataset [8]. The size of a scRNA-seq dataset is usually very large, which could result in a very small or even empty set of ubiquitous genes, since the number of ubiquitous genes depends on the sizes of datasets. To select ubiquitous genes using scRNA-seq data, we defined a parameter named presence rate, which required that one selected ubiquitous gene must appear in at least a proportion of the trimmed sets.

Evaluation metrics

In the previous study [8], Gustavo *et al.* designed two novel and mutually independent metrics, which were the number of uniform genes and Spearman Correlation Coefficients (SCCs) between expression profiles of gene pairs. Two basic ideas to support these two evaluation metrics are sucessiful normalization methods increase the number of uniform genes and decrease the correlation between the expression profiles of gene pairs. In this study, we designed a new metric AUCVC instead of the number of uniform genes to evaluate normalization methods. We randomly selected 1,000,000 ubiquitous gene pairs to calculate the medians of SCCs. Then, we compared the evaluation results of all the normalizaiton methods by the medians of SCCs with those by AUCVC.

AUCVC is created by plotting the number of uniform genes (y-axis) at each normalized CV (**Formula 2**) threshold (x-axis). To determine the number of uniform genes using scRNA-seq data containing a high frequency of zeros, we only considered genes with non-zero expression values divided by the sample number not less than a thresthold, which was designed as a parameter non-zero ratio. Since a high or a low normalized CV threshold produces more false or less uniform genes, it is reasonable to consider the overall

288 performance of each method at various threshold settings instead of that at one specific
289 threshold setting. In formula 1 and 2, symbols have the same meanings as those in figure 1
290 and n* does not count zero elements in each sample.
291
292

293
$$CV_i = \left(\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(x_{ij} - \overline{x}_i\right)^2} \right) \Big/ \overline{x}_i \ , \ \overline{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \qquad (1)$$

294
$$Normalized\ CV_i = \left\{ CV_i - \min_i\left(CV_i\right) \right\} \Big/ \left\{ \max_i\left(CV_i\right) - \min_i\left(CV_i\right) \right\}$$
$$CV_i = \left(\sqrt{\frac{1}{n^*}\sum_{j=1}^{n^*}\left(log_2\left(x_{ij}\right) - \overline{x}_i\right)^2} \right) \Big/ \overline{x}_i \ , \ \overline{x}_i = \frac{1}{n^*}\sum_{j=1}^{n^*} log_2\left(x_{ij}\right), \ x_{ij} > 0 \qquad (2)$$

295 # Conclusion and Discussion

296

297 # Acknowledgments

298 We thank Professor Sui Huang and Dr. Joseph Zhou from Institute for Systems
299 Biology (ISB) their hosts on Shan Gao' visiting to ISB.

300

301 # Funding

302 This work was supported by grants from by National Key Research and
303 Development Program of China (2016YFC0502304-03) to Defu Chen and Fundamental
304 Research Funds for the Central Universities (for Nankai University) to Shan Gao.

305

306 # Competing interests

307 Non-financial competing interests are claimed in this study.
308
309

Authors' contributions

SG conceived this project. SG and JR supervised this project. ZW, WL and SL performed programing. ZW, DY and HJ analyzed the data. WD prepared all the figures, tables and additional files. SG drafted the main manuscript. XX and XX revised the manuscript. **All authors have read and approved the manuscript.**

REFERENCES

1. Lovén, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B., Levens, D.L., Lee, T.I., and Young, R.A., *Revisiting global gene expression analysis.* Cell, 2012. **151**(3): 476-482.
2. Gao, S., Ou, J., and Xiao, K., *R language and Bioconductor in bioinformatics applications(Chinese Edition)*. 2014, Tianjin: Tianjin Science and Technology Translation Publishing Ltd.
3. Tao, H. and Gao, S., *Computational Systems Biology: Methods and Protocols*. 2018, New York: Springer.
4. Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S., *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.* BMC Bioinformatics, 2010. **11**(1): 94.
5. Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B., *Synthetic spike-in standards for RNA-seq experiments.* Genome Research, 2011. **21**(9): 1543.
6. Anders, S. and Huber, W., *Differential expression analysis for sequence count data.* Genome biology, 2010. **11**(10): R106.
7. Robinson, M.D., McCarthy, D.J., and Smyth, G.K., *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): 139-140.
8. Glusman, G., Caballero, J., Robinson, M., Kutlu, B., and Hood, L., *Optimal scaling of digital transcriptomes.* Plos One, 2013. **8**(11): e77885.
9. Li, P., Piao, Y., Shon, H.S., and Ryu, K.H., *Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data.* Bmc Bioinformatics, 2015. **16**(1): 347.
10. Gao, S., Zhang, N., Duan, G.Y., Yang, Z., Ruan, J.S., and Zhang, T., *Prediction of function changes associated with single-point protein mutations using support vector machines (SVMs).* Human Mutation, 2009. **30**(8): 1161-6.
11. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): 15-21.

12. Gao, S., Tian, X., Chang, H., Sun, Y., Wu, Z., Cheng, Z., Dong, P., Zhao, Q., Ruan, J., and Bu, W., *Two novel lncRNAs discovered in human mitochondrial DNA using PacBio full-length transcriptome data.* Mitochondrion, 2017.

Figure legends

Figure 1. Normalization factor and library size

(**A**). A raw gene expression matrix can be transformed into a normalized gene expression matrix by the multiplication of a factor f_j to each column. Each column represents the expression levels of all genes from a sample and each row represents the expression levels of a gene across all samples. (B). HG7, ERCC, TN, TC, CR, NR and TU use library sizes Nj* to calculate normalization factors. Nj represents the library size estimated by TC. DESeq, RLE, UQ and TMM have been modified in NormExpression to ignore zero values and the resulting normalization factors need be further normalized by their geometric mean values. After modification, RLE is identical to DESeq. Q3 means that about 75% of genes in the jth sample have expression levels below Q3 and about 25% have those above Q3. For all methods, log represents the natural logarithm.

Figure 2. Parameter grid to evaluate normalization methods

Parameter grid of non-zero ratio from 0.2 to 0.9 for scRNA663 and from 0.8 to 1 for bkRNA18 was used to produce AUCVC values All the normalization methods were classified into three groups based on their AUCVC values sorted in descending order using one scRNA-seq dataset scRNA663 (**A**) and one bulk RNA-seq dataset bkRNA18 (**B**).

Figure 3. Consistency in the evaluation results by different metrics

A normalization method with a higher AUCVC value produced a lower median of Spearman Correlation Coefficients (SCCs) between normalized expression profiles of ubiquitous gene pairs using both scRNA-seq (**AB**) and bulk RNA-seq data (**DE**). The hierarchical clustering result showed that 13 methods had been classified into the same groups (**CF**) by SCCs between normalization factor pairs as those (**Figure 2AB**) by AUCVC.