Communicative & Integrative Biology

ISSN: (Print) 1942-0889 (Online) Journal homepage: https://www.tandfonline.com/loi/kcib20

Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments

A matter of relative size of studied transcriptomes

Elie Maza, Pierre Frasse, Pavel Senin, Mondher Bouzayen & Mohamed Zouine

To cite this article: Elie Maza, Pierre Frasse, Pavel Senin, Mondher Bouzayen & Mohamed Zouine (2013) Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments, Communicative & Integrative Biology, 6:6, e25849, DOI: 10.4161/cib.25849

To link to this article: https://doi.org/10.4161/cib.25849

Communicative & Integrative Biology 6:6, e25849; November/December 2013; © 2013 Landes Bioscience

Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments

A matter of relative size of studied transcriptomes

Elie Maza[1,2], Pierre Frasse[1,2], Pavel Senin[1,2], Mondher Bouzayen[1,2] and Mohamed Zouine[1,2,*]

[1]Université de Toulouse; INP-ENSA Toulouse; Laboratoire Génomique et Biotechnologie des Fruits; Castanet-Tolosan, France;
[2]INRA; Laboratoire Génomique et Biotechnologie des Fruits; Auzeville, Castanet-Tolosan, France

Keywords: Computational methods, Differential gene expression analysis, High throughput sequencing, Normalization methods, Relative transcriptome size, RNA-Seq data

In recent years, RNA-Seq technologies became a powerful tool for transcriptome studies. However, computational methods dedicated to the analysis of high-throughput sequencing data are yet to be standardized. In particular, it is known that the choice of a normalization procedure leads to a great variability in results of differential gene expression analysis. The present study compares the most widespread normalization procedures and proposes a novel one aiming at removing an inherent bias of studied transcriptomes related to their relative size. Comparisons of the normalization procedures are performed on real and simulated data sets. Real RNA-Seq data sets analyses, performed with all the different normalization methods, show that only 50% of significantly differentially expressed genes are common. This result highlights the influence of the normalization step on the differential expression analysis. Real and simulated data sets analyses give similar results showing 3 different groups of procedures having the same behavior. The group including the novel method named "Median Ratio Normalization" (MRN) gives the lower number of false discoveries. Within this group the MRN method is less sensitive to the modification of parameters related to the relative size of transcriptomes such as the number of down- and upregulated genes and the gene expression levels. The newly proposed MRN method efficiently deals with intrinsic bias resulting from relative size of studied transcriptomes. Validation with real and simulated data sets confirmed that MRN is more consistent and robust than existing methods.

Introduction

RNA-Seq approaches, based on high-throughput sequencing technologies, are becoming an essential tool in transcriptomics studies.[1] In the beginning, its ability to capture transcriptome dynamics (across different tissues and conditions) without sophisticated normalization of data sets was considered as a particular advantage over other techniques. Nevertheless, it was found that a normalization preprocessing step can significantly improve the quality of the analysis. Therefore, the effect of normalization was rigorously studied and many works related to this issue have been published in recent years. Notably, an evaluation of statistical methods for normalization in mRNA-Seq experiments stressed (ref. 2) the requirement for further research in the development of statistical and computational methods dedicated to the processing of mRNA-Seq data. In particular, Bullard, Purdom, Hansen and Dudoit (ref. 2) demonstrated the impact of the choice of the normalization procedure on the Differential Expression (DE) analysis.

Many of the normalization methods proposed in the literature are based on the correction of biases and "artifacts" directly related with RNA-Seq technology. For instance, the read (or fragment) number per lane and the gene (or RNA) length are trivial biases that have been pointed out very early.[3–5] Sequencing depth is another trivial and important bias highlighted in reference 6. Other technical biases, such as non-uniformity of the distributed reads within the transcripts, and the strong sample-sp.ecific GC-content effects on read counts, were later studied in references 7, 8, and 9.

The present paper focuses on another significant bias in RNA-Seq experiments that is not introduced by the technology, but is rather intrinsic to the studied transcriptomes. It is shown here that the relative size of studied transcriptomes represents an important bias and that a particular normalization procedure is needed to address this issue. The aim of the present study is therefore to propose an improved version of an existing normalization method and to evaluate its performance on simulated and

*Correspondence to: Mohamed Zouine; Email: Mohamed.Zouine@ensat.fr
Submitted: 05/13/13; Revised: 07/22/13; Accepted: 07/22/13
Citation: Maza E, Frasse P, Senin P, Bouzayen M, Zouine M. Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of relative size of studied transcriptomes. Communicative & Integrative Biology 2013; 6:e25849; http://dx.doi.org/10.4161/cib.25849

real data sets. The newly established method named "Median Ratio Normalization" (MRN) is then compared with normalization methods that have been described in the literature.[2,10,11]

Bias Related to the Relative Size of Transcriptomes

Thereafter, we will refer to different biological states of a given studied transcriptome as *conditions*, and to independent biological samples of a given condition as *replicates*.

Let X_{gkr} be the observed number of reads (or count) of gene $g \in \{1,\ldots G\}$ in condition $k \in \{2,\ldots,K\}$ for replicate $r \in \{1,\ldots,R\}$; μ_{gk} the expectation of the true and unknown number of transcripts of gene g in condition k; L_g the length of gene g; and N_{kr} the total number of reads in condition k for replicate r. As described by Robinson and Oshlack (ref. 11), among others, we can model the expected value of X_{gkr} as:

$$\mathrm{E}\left(X_{gkr}\right) = \frac{\mu_{gk} L_g}{S_k} N_{kr} \qquad (1)$$

where S_k is the size of studied transcriptome in condition k, that is

$$S_k = \sum_{g=1}^{G} \mu_{gk} L_g. \qquad (2)$$

For each gene g, an approximation of the expected value of the ratio between 2 conditions, say 1 and 2, is given by the delta method (see, for example, ref. 12) as

$$\mathrm{E}\left(\frac{X_{g2r}}{X_{g1r}}\right) \approx \frac{\mu_{g2}}{\mu_{g1}} \frac{S_1}{S_2} \frac{N_{2r}}{N_{1r}}. \qquad (3)$$

As our interest is gene expression fold change, it is obviously given in the above equation (3) by μ_{g2}/μ_{g1}. Then, ratios N_{2r}/N_{1r} and S_1/S_2 are the biases introduced, respectively, by RNA-Seq technology and by the relative size of studied transcriptomes. The former bias can be easily corrected by division of the observed counts (in condition k for replicate r) by the total number of reads N_{kr} before the ratio calculation. The latter bias is related to the relative size of studied transcriptomes and cannot be removed directly because the values referring to μ_{g1} and μ_{g2} are obviously unknown. It is shown here that, for a given experiment, this bias is not directly related to the technology, but rather to the studied transcriptomes. Robinson and Oshlack (ref. 11) also clearly show this bias and they give a solution which seems nevertheless less robust than the one we propose. However, it is important to mention that the implementation of the MRN method requires taking into account a biological assumption based on (but less restrictive than) that made previously in reference 11. This assumption states the following: in transcriptomes under study, less than 50% of the genes are upregulated, whereas less than 50% are downregulated. We detail in the Methods section (subsection Median Ratio Normalization Method) the interest of such an assumption.

Evaluated Normalization Methods

The present paper evaluates a number of normalization methods designed to resolve (directly or indirectly) the bias induced by the relative size of studied transcriptomes.

First, methods directly addressing this issue are taken into consideration. The Trimmed Mean of M-values (TMM) method (developed in ref. 11) deals with estimating the "relative RNA production of 2 samples" under the assumption that the majority of genes are not differentially expressed. This method is implemented in the *edgeR* package available from the *Bioconductor* website at http://www.bioconductor.org (see refs. 13 and 14). Similarly, Anders and Huber (ref. 10), proposed a normalization method based on "size factors" that renders counts from different samples comparable. This method is called "Relative Log Expression" (RLE) and is implemented in *edgeR* and *DESeq* packages (available from the *Bioconductor* website).

Second, we evaluate 4 normalization methods that do not directly aim at removing this bias. Instead, they are dealing with normalization of the read counts in order "to adjust for varying lane sequencing depths and potentially other technical effects" (see ref. 2 and references therein). These are the Upper Quartile normalization (UpQu), the Median normalization (Medi), the Total Counts normalization (ToCo) and the FPKM normalization.[2,15] While in the UpQu normalization method, counts are normalized by division (in a given replicate) by the upper quartile of these counts, the Medi simply computes the median, and the ToCo uses the sum of all counts. Finally, the FPKM normalization method normalizes raw counts by both the length of the gene and the total counts. In addition, for this normalization method, the obtained quantities are multiplied by 10^9 in order to obtain "fragments per kilobase of transcript per million fragments mapped" abbreviated as FPKM (see refs. 3 and 15).

We compare the performance of all methods mentioned above against "No Normalization" (NoNo) and against our proposed method called "Median Ratio Normalization" (MRN). The MRN method follows the idea behind TMM and RLE normalization methods aiming at removing the bias due to the relative size of studied transcriptomes.

In total, 9 normalization methods, used for the DE analysis, were compared: TMM, TMM50 (TMM with 50% of trimmed M-values), RLE, UpQu, Medi, ToCo, FPKM, MRN and NoNo.

Materials and Methods

This section is divided into 4 independent parts. The first part is devoted to the description of the MRN procedure. We explicitly show, in this part, how the MRN procedure overcomes the bias due to the relative size of studied transcriptomes. The second part presents the pipeline used in this paper for the DE analysis. The third part describes the algorithm used to simulate RNA-Seq data sets that will be used to compare the normalization methods described above. Finally, the last part describes the tomato's RNA-Seq data set studied in this paper.

Median ratio normalization method

The computational steps of the MRN method aiming at removing the bias due to the relative size of studied transcriptomes, are given below. An R program of the MRN method is provided within a supplementary file (see Section *Supplementary Materials*). For sake of simplicity, the method is here described for 2 conditions (k = 1 and k = 2) with the same number of replicates R. Obviously, the MRN method can be generalized to an arbitrary number of conditions $K > 2$ with different numbers of replicates.

1. Calculation of weighted means of gene expressions for both conditions k $\in$ {1,2} and all genes $g \in \{1,...,G\}$:

$$\overline{X}_{gk} = \frac{1}{R}\sum_{r=1}^{R}\frac{X_{gkr}}{N_{kr}}.$$

2. Calculation of gene expression ratios with condition 1 as a reference, for all genes $g \in \{1,...,G\}$: $\tau_g = \text{median}_g\left(\overline{X}_{g2}/\overline{X}_{g1}\right)$.

3. Calculation of the median of obtained ratios: $\tau = \text{median}_g(\tau_g)$.

4. Calculation of normalization factors for both conditions k $\in$ {1,2} (taking into account factor τ) and for each replicate r $\in$ {1,…,R} (taking into account the sequencing depth N_{kr}):

$$\begin{cases} e_{1r} = 1 \times N_{1r}, \\ e_{2r} = \tau \times N_{2r}. \end{cases}$$

5. Calculation of adjusted normalization factors to multiply, for symmetry, to 1: $f_{kr} = e_{kr}/\widetilde{f}$ for k $\in$ {1,K = 2} and r $\in$ {1,…,R} with

$$\widetilde{f} = \exp\left(\frac{1}{KR}\sum_{k=1}^{K}\sum_{r=1}^{R}\log\left(e_{kr}\right)\right).$$

6. Normalization of counts: X_{gkr}/f_{kr}.

The theoretical bases of the above computational workflow are the followings. Step 1 and Equation (1) imply that the expected values of means $E(\overline{X}_{gk})$ are equal to $\mu_{gk}L_g/S_k$. Hence, in Step 2, Equation (3) implies that the expected values of ratios $E(\tau_g)$ are approximately equal to $(\mu_{g2}/\mu_{g1})(S_1/S_2)$ where the bias of the relative size of transcriptomes is given by S_1/S_2. By assuming that less than half the genes are upregulated, and less than half are downregulated, the median τ of all ratios (calculated in Step 3) is then an approximation of S_1/S_2. In Step 4, in order to remove this bias due to the relative size of transcriptomes, normalization factors are fixed to e_{1r} for the reference condition (condition 1) and to e_{2r} for condition 2. Finally, Step 5 only aims at having adjusted factors to multiply to 1 for symmetry (as in ref. 14).

Pipeline for the differential expression analysis

Since the objective of the present work is to evaluate the impact of normalization methods on the DE analysis, each of studied data sets was analyzed by the application of the same universal pipeline for DE analysis. However, the normalization step was obviously different for each of the studied methods.

All computations were done with R environment.[16] TMM, RLE, UpQu and Medi normalization methods were performed with function *calcNormFactors()* of the *edgeR* package. Computations of MRN, FPKM and ToCo normalization methods were implemented by us and provided within a supplementary file (see Supplementary Materials).

For the DE analysis, an error model based on the Negative Binomial distribution was used, with variance and mean linked by local regression. This model is well suited for modeling count data dispersion (refs. 17, 18, and 19) and is implemented in *edgeR* and *DESeq* packages. In this study, we used the 2 functions *estimateDispersions()* and *nbinomTest()* of the *DESeq* package. The R program for the DE analyses is also provided in a supplementary file (see Supplementary Materials).

Simulation of RNA-Seq data

Simulated RNA-Seq data sets for K = 2 different conditions and R = 3 biological replicates were generated by the following 2-steps procedure.

Step I. This step deals with the simulation of gene expression values μ_{gk} for both conditions 1 and 2. The input consists of various parameters, some of which are not fixed (underlined parameters) and allow the simulation of different data types by changing their values. The aim of this step is to start from simulated M and A-values (from Log-Normal and Normal distributions respectively) and then to resolve the expression values. (Simulations of Gaussian laws are performed with the *rnorm()* function of R.)

1. Number of simulated genes: G = 30000.

2. Percentages of differentially expressed genes: p_u% of upregulated genes and p_d% of downregulated genes.

3. Simulation of M-values from Log-Normal distributions for upregulated genes M_g^u and downregulated ones M_g^d:

$$\begin{cases} M_g^u \sim 1 + \text{LN}\left(\underline{m_u}, s_u\right) \quad \text{and} \quad s_u = 0.7, \\ M_g^d \sim -1 - \text{LN}\left(\underline{m_d}, s_d\right) \quad \text{and} \quad s_d = 0.7, \end{cases}$$

where -1 (resp. 1) implies a 2-fold downregulation (resp. upregulation).

4. Simulation of A-values from a Normal distribution: $A_g \sim N(m_A, s_A)$ with $m_A = 7$ and $s_A = 3$.

5. Calculation of gene expression values for conditions 1 and 2: $\mu_{g1} = \left\lceil 2^{\left(2A_g - M_g\right)/2} \right\rceil$ and $\mu_{g2} = \left\lceil 2^{\left(2A_g + M_g\right)/2} \right\rceil$ where $\lceil x \rceil$ is the integer immediately higher than x.

6. Calculation of simulated M and A-values:

$$\begin{cases} M_g = \log_2\left(\mu_{g2}\right) - \log_2\left(\mu_{g1}\right), \\ A_g = \left(\log_2\left(\mu_{g1}\right) + \log_2\left(\mu_{g2}\right)\right)/2. \end{cases}$$

Step II. This step is dedicated to the simulation of observed counts X_{gkr} from previously computed gene expression values and Negative Binomial distributions. Simulated Negative Binomial random counts are computed with the *rnbinom()* function of R which depends upon 2 parameters: *mu* for the mean and *size* for

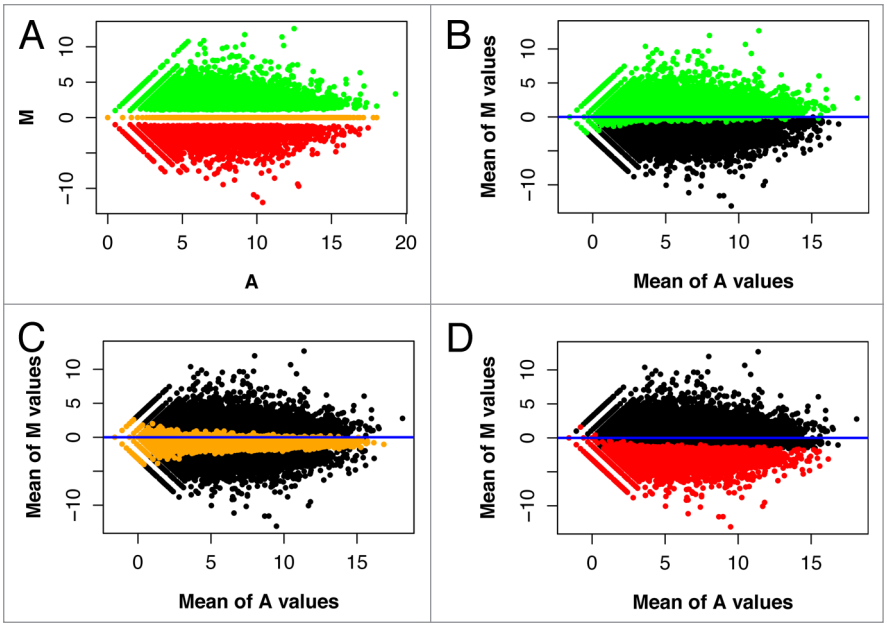

Figure 1. MA-plots of a simulated data set. This figure corresponds to a simulated data set from simulation type 1 of **Table 1**. Panel A shows MA-plot of gene expression values (simulated from Step I of the simulation algorithm) with green dots for upregulated genes, orange dots for non DE genes and red dots for downregulated genes. Panels B, C, and D show the same MA-plot of means of observed values (simulated from Step II of the simulation algorithm) with black dots for all genes and, respectively, green dots for upregulated genes in panel B, orange dots for non DE genes in panel C and red dots for downregulated genes in panel D.

the dispersion. (In this case, the variance is linked to the mean by the equation $mu+mu^2/size$.)

1. It is assumed that all genes have the same length: $L_g = 1000$ bp.

2. Calculation of transcriptome sizes for conditions 1 and 2 from Equation (2):

$$S_1 = \sum_{g=1}^{G} \mu_{g1} L_g \quad \text{and} \quad S_2 = \sum_{g=1}^{G} \mu_{g2} L_g.$$

3. Calculation of probabilities of gene observations:

$$p_{g1} = \frac{\mu_{g1} L_g}{S_1} \quad \text{and} \quad p_{g2} = \frac{\mu_{g2} L_g}{S_2}.$$

4. Library sizes of replicates are drawn independently from a uniform distribution: $N_{kr} \sim U[N_0,N_1]$with $N_0 = 15 \times 10^6$ and $N_1 = 25 \times 10^6$.

Table 1. Parameters of 4 simulation types

Parameters	*Type 1*	*Type 2*	*Type 3*	*Type 4*
p_u	40%	30%	15%	10%
p_d	20%	30%	5%	10%
m_u	0	0	0	0.5
m_d	0	0	0	0.1

This table contains parameters of the 4 simulation types studied in the paper. Parameters correspond to underlined parameters of Step I of the simulation algorithm described in the Methods section (subsection Simulation of RNA-Seq Data).

5. Calculation of gene expectations from Equation (1): $E_{gkr} = p_{gk} N_{kr}$.

6. Simulations of gene counts are drawn from Negative Binomial distributions: $X_{gkr} \sim NB(size,E_{gkr})$ with size = 10.

In order to test the performance of normalization methods on different data sets, we performed 4 simulation types by varying underlined parameters in Step I of the above algorithm. Some of these simulated data sets are performed with quite extreme situations leading to some challenging data sets in order to really discriminate the performance of studied normalization methods. The full parameter details of these 4 simulations are given in **Table 1**. Simulation types 1, 2 and 3 differ from each other only by the amounts of upregulated and downregulated genes: p_u and p_d. While having the same amount of upregulated and downregulated genes, simulation type 4 has a higher mean expression: $m_u > m_d$. Theoretically, simulation types 1, 3 and 4 should result on transcriptomes with different sizes. Indeed, for simulation types 1 and 3, the amount of upregulated genes (p_u) is greater than the amount of downregulated ones (p_d) and, as a consequence of independent and identical distributions of simulated expressions, S_2 should be greater than S_1. For the simulation type 4, while the percentages of upregulated and downregulated genes are the same, the mean expression of upregulated genes m_u is greater than the mean expression of downregulated genes m_d and, in turn, S_2 should also be greater than S_1. Only simulation type 2 should have equal transcriptome sizes.

An example of a data set from simulation type 1 is shown in **Figure 1**. Panel A shows the MA-plot of simulated gene expression values μ_{gk} (from Step I). As described in **Table 1**, 40% of genes are upregulated (green dots) with M-values greater than 1, 20% of genes are downregulated (red dots) with M-values less than –1, and then 40% of genes are non DE with M-values equal to 0. Panels B, C, and D of **Figure 1** display the MA-plot of means of observed values (from Step II) with green dots for upregulated genes, orange dots for non DE genes and red dots for downregulated genes. Panels B, C, and D show an asymmetry of M-values due to the greater size of transcriptome in condition 2 following Equation (3) and implying that the symmetry is not around 0 (blue line of **Figure 1**) but around $\log_2(S_1/S_2)$ which is smaller than 0. This is the reason why M-values of upregulated, non DE and downregulated genes are all down-shifted from 0.

The full R program of these simulations is provided on a supplementary file (see Supplementary Materials) with parameters from simulation type 1. These parameters can obviously be changed in order to simulate different biological conditions.

Tomato's RNA-Seq data set

To investigate the tomato transcriptome dynamics of fruit set, RNA were isolated from flower buds and flowers at Anthesis

and Post-Anthesis stages. For each stage, cDNA libraries were generated from 3 biological replicates and subjected to Illumina mRNA-Seq technology sequencing. We generated 30–40 million high-quality sequence reads for each replicate. A total of 287.5 millions of 101 bp paired-end high quality reads: 94 M reads from Bud, 91M reads from Anthesis and 102M reads from Post-Anthesis. Mapping of these reads to the Tomato genome sequence Sly2.40 was performed by the TopHat software with default parameters, at most 2 mismatches and no indels allowed.[20,21] More than 90% of the reads were aligned to the genomic sequence. Expressed loci were assessed with Cufflinks software with default parameters and using the tomato gene annotation file iTAG2.30.[20,22]

Results

This section is devoted to the comparison of results of DE analyses obtained with the various normalization methods described before on both simulated data sets and the tomato's RNA-Seq data set.

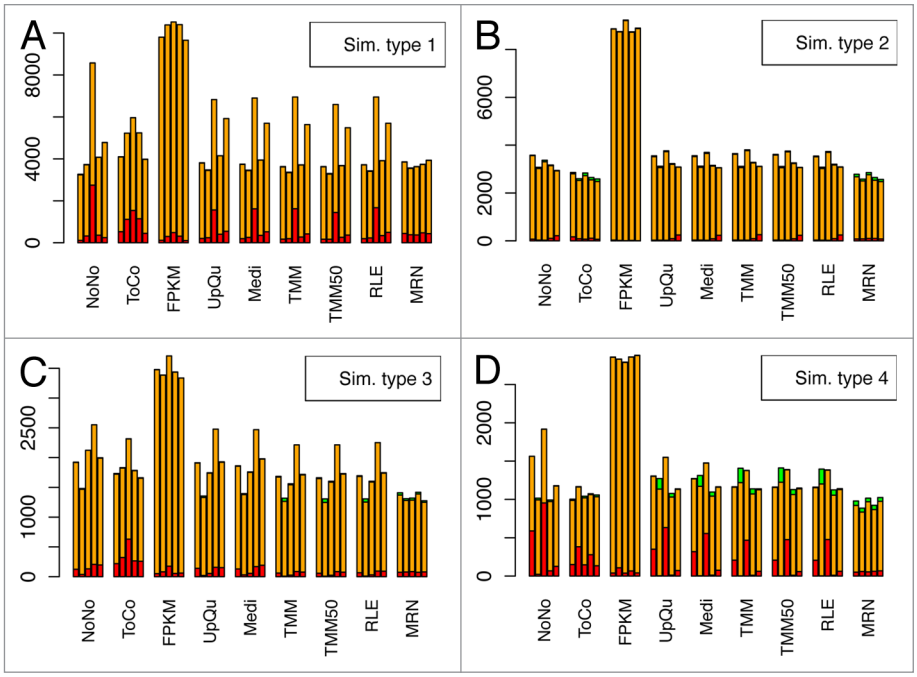

Figure 2. Bar-plots of false discoveries. For each of the 4 simulation types of **Table 1**, bar-plots of above panels show the number of false discoveries of each normalization method for 5 independent simulation repetitions (one bar per repetition). With green and red colors, respectively, the number of false upregulated and downregulated genes. With orange color, the number of false DE genes. False upregulated and false downregulated genes correspond to non-DE genes. In the same way, false DE genes correspond obviously to upregulated or downregulated genes.

Results on simulated data sets

The impact of normalization methods on the DE analysis of the 4 different simulation types from **Table 1** is presented in **Figures 2 and 3**, and in **Table 2**. In order to compare the variability of the studied methods, 5 data sets have been simulated and compared for each simulation type. As described below, obtained results are remarkably stable for the MRN method from one repetition to another and, consequently, no more than 5 repetitions seem to be needed. **Figure 2** shows the number of false discoveries produced by each of the normalization methods for each of the simulation types. **Figure 3** shows the Mean Squared Error (MSE) of each normalization method for each of the simulation types, that is, the mean of squared differences between ratios of gene expressions obtained by a normalization method and the true ratios of simulated data. Results presented in these figures can be obtained with the R program of a given supplementary file (see Supplementary Materials). **Table 2** contains, for each method, some interesting indicators as power and false positive and negative rates.

Number of false discoveries

Globally, as we can see on **Figure 2**, the total number of false discoveries (maximum ordinate of each graph) is decreasing starting from the simulation type 1 toward the simulation type 4. That is obviously due to the decreasing amount of DE genes (see **Table 1**). For any of the simulation types, the greatest amount of errors is always associated to the FPKM normalization method. Another global trend is that for all 4 simulation types, the MRN has the smallest number of errors closely followed by UpQu, Medi, TMM, TMM50, and RLE methods. Moreover, the MRN gives also more robust results with almost the same number of false discoveries for each of the 5 repetitions for each of the simulation types.

For simulation type 2, the differences between normalization methods (excepted FPKM) are less important. That is mostly due to the fact that, as described previously, the sizes of the 2 transcriptomes are nearly the same, and consequently, no normalization is needed. For simulation types 1, 3, and 4, only MRN seems to produce symmetric errors (although less significant for simulation type 1). The other methods, at least for some of the 5 repetitions, produce more false downregulated genes. This drawback is due to the non-symmetry of ratios around value 1 for simulation types 1, 3, and 4 as quoted in the Methods section (subsection Simulation of RNA-Seq Data). Also, as expected, this drawback is more important for NoNo and ToCo methods where no normalization for the relative size of transcriptomes is applied.

Mean squared errors

Mean Squared Errors (MSE) of estimated differential expressions ($\log_2$-ratios) are shown on **Figure 3**. Similarly to the performance on false discoveries, the FPKM normalization method is producing the worst MSE. Also, in the same way, the MRN gives the best results in both consistency and robustness: the MRN gives globally the smallest MSE in any of the 5 independent simulations.

For simulation type 2, where transcriptomes have the same sizes, all methods (except FPKM) give similar results. For simulations types 1, 3, and 4, TMM, TMM50 and RLE seem to have

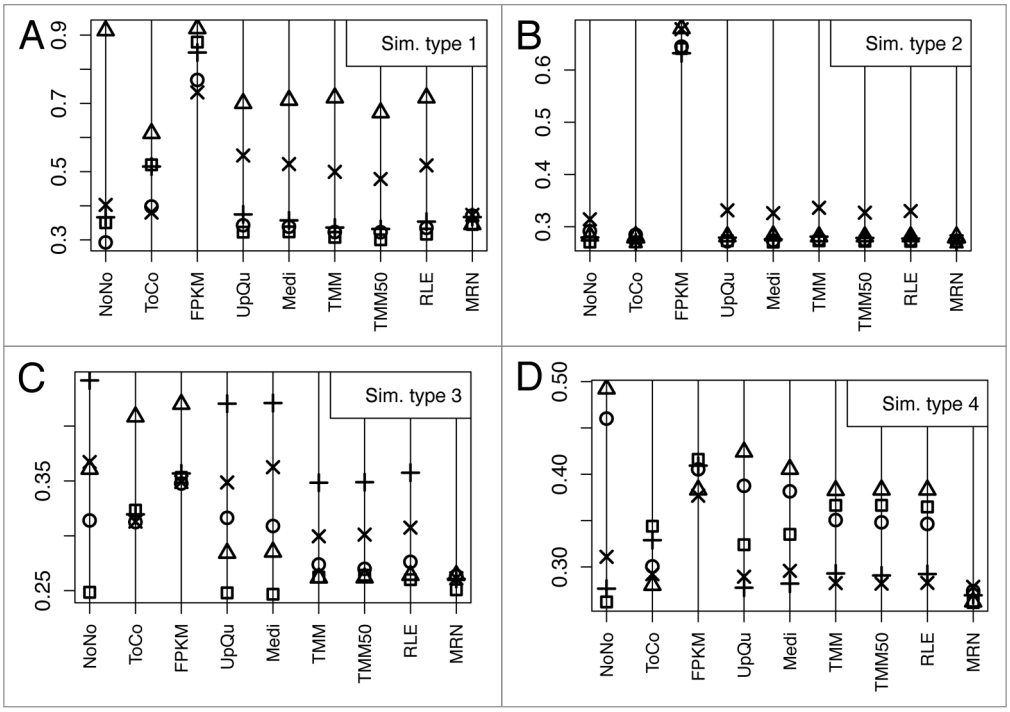

Figure 3. Scatter plots of MSE. For each of the 4 simulation types of **Table 1**, scatter plots of above panels show the MSE of each normalization method for 5 independent simulation repetitions. For each simulation type, each repetition has the same symbol for each normalization method.

Table 2. Quantitative results

Results	*NoNo*	*ToCo*	*FPKM*	*UpQu*	*Medi*	*TMM*	*TMM50*	*RLE*	*MRN*
Signif.	0.3279	**0.3411**	0.1956	0.3237	0.3243	0.3241	0.3243	0.3247	**0.3412**
Power	0.7898	**0.8159**	0.4794	0.7858	0.7878	0.7899	0.7917	0.7902	**0.8348**
F. Pos.	0.0195	0.0241	**0.0062**	0.0151	0.0149	0.0131	0.0123	0.0139	**0.0117**
F. Neg.	0.2102	**0.1841**	0.5206	0.2142	0.2122	0.2101	0.2083	0.2098	**0.1652**

This table contains, for each method, the rate of significantly DE genes (Signif. = Number of signif. DE genes / Total number of simulated genes), the power to detect a DE gene (Power = Number of rightly signif. DE genes / Total number of DE genes), the false positive rate (F. Pos. = Number of false positive genes / Total number of non DE genes) and the false negative rate (F. Neg. = Number of false negative genes / Total number of DE genes). We notice that F. Neg. = 1 - Power. For Signif. and Power indicators (resp. F. Pos. and F. Neg. indicators), the maximum (resp. minimum) value is bolded in a grayed cell, the second one is only bolded, and the minimum (resp. the maximum) value is underlined.

the same behavior. Also, NoNo, UpQu, and Medi normalization methods are found to be very sensitive to the variability in all simulations.

Quantitative results

Table 2 summarizes, for each method, the results obtained from the 20 simulations (from the 4 simulation types with 5 independent repetitions for each one). The summary was done by adding all genes of all simulations (working then with a sample size of 20 × 30000 = 600000 genes). Excepting the FPKM method, all methods give between 32% and 34% of significantly DE genes. The MRN method maximizes this indicator with 34%. The FPKM method gives about 20% of significant DE genes. These values are consistent with the power to detect a DE gene: the MRN method holds the best power value (about 83%) and the FPKM method holds the worst one (about 48%). In terms of false positive rate, the FPKM has the best result (about 0.62%) followed by the MRN method (about 1.17%). Globally, the MRN method seems to give the best results.

Results on the tomato fruit set data set

Results of DE analyses performed on the tomato RNA-Seq data with the various normalization methods described above are shown on **Figure 4**. Panels A and B are related, respectively, to flower Buds to Anthesis transition and to Anthesis to Post-Anthesis transition. Obviously, these results do not intend to validate one or another of the studied normalization methods. Nevertheless, this study shows interesting links between the different normalization methods leading to different groups of methods having the same behavior. Moreover, these groups can be compared with those obtained in the previous simulation study.

Hierarchical classifications

The 2 upper graphs of **Figure 4** are hierarchical classifications of all studied normalization methods for both studied transitions.

These hierarchical classifications are performed with results of each DE analysis (with a type I error equal to 5%) by setting to 1 the genes that are upregulated, to –1 those who are downregulated and to 0 those who are not significantly DE. (Hierarchical classifications are done with the R function *hclust()* with Euclidean distance matrices and argument *method* = *"complete"*.)

We can easily see here that 5 normalization methods seem to have the same behavior in both transition studies and are close one from another: TMM, TMM50, UpQu, Medi, and RLE. The MRN, ToCo and NoNo normalization methods are close to the 5 methods cited above but not in the same way for both transitions. It is also obvious here that the FPKM normalization method is always far from the others.

Bar-plots

The 2 bottom graphs of **Figure** 4 are bar-plots of numbers of significantly DE genes for each normalization method for both studied transitions. Globally, the amount of DE genes is greater between Anthesis and Post-Anthesis stages. Nevertheless, behaviors of normalization methods seem to be the same in both transitions.

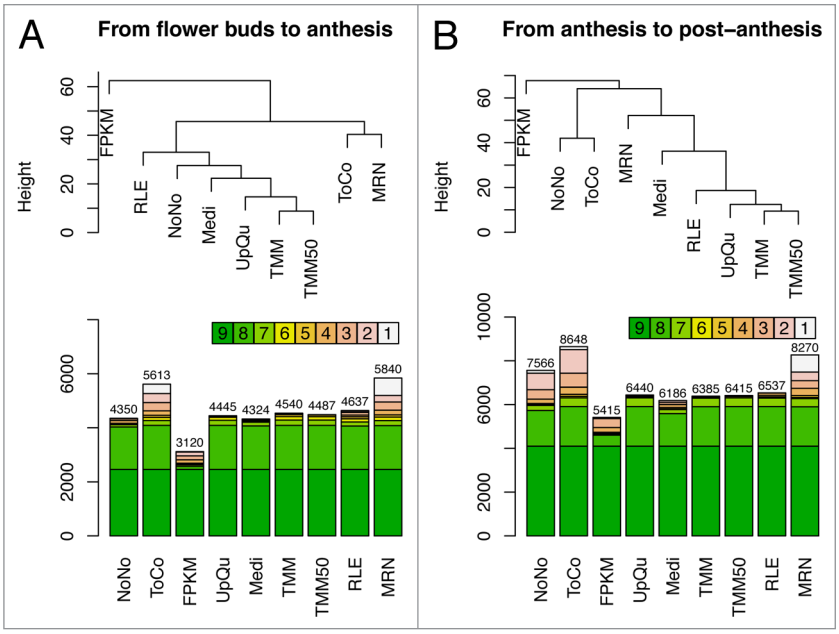

Figure 4. Results of the DE analyses on tomato RNA-Seq data. Results of the DE analyses on tomato RNA-Seq data for both studied transitions: from Bud to Anthesis (on the left) and from Anthesis to Post-Anthesis (on the right).

The FPKM normalization method gives a less amount of DE genes. On the contrary, MRN and ToCo give the greater amount of DE genes. Also, once again, TMM, TMM50, UpQu, Medi, and RLE normalization methods seem to have the same behavior.

We notice here that these global behaviors of our normalization methods (both on hierarchical classifications and bar-plots) were also encountered in the simulation study, which would tend to validate the simulation approach.

Moreover, for these bar-plots, for each normalization method, each color (**Fig.** 4) represents the amount of normalization methods that have the same common DE genes. For example, dark green color represents the number of DE genes that are common for all methods. We can then see that about 2500 and 4000 genes are common for all normalization methods for the first and the second transition respectively. That roughly represents up to 50% of common DE genes (for both transitions), which underlines the great importance of an accurate normalization method. In the same way, white color represents the number of sp.ecific DE genes. Here, the MRN gives the great amount of sp.ecific DE genes (650 and 800 for the 2 transitions). The other methods seem not to give sp.ecific DE genes or a few ones for the ToCo method in the first transition.

Discussion

In this paper, we identified and discussed an important issue related to RNA-Seq data normalization for differential gene expression analysis. Simulated data analysis revealed that the bias due to the relative size of transcriptomes leads to poor estimations of ratios of gene expressions, and consequently to biased DE analysis. To address this issue, we benchmarked the performance of the most widespread normalization methods together with a novel proposed method named "Median Ratio Normalization" (MRN).

The newly established method is a modified version of the normalization method already proposed by Robinson and Oshlack (ref. 11), where authors also stressed the need to remove the bias due to the relative size of transcriptomes, but proposed a different computational workflow. Moreover, the assumption we made is less restrictive than the one made in reference 11, which assumes that "the majority of genes are non DE." The same issue has been also addressed by the RLE normalization method (ref. 10), which uses also another computational workflow. The processing procedure of RNA-Seq data described in the present study clearly outperforms comparatively to previous methods.

Of particular interest, the notion of relative size is also considered for high throughput metabonomics and proteomics data analyses where count data are commonly used.[23,24]

Based on poor benchmark performance, it clearly emerges that the FPKM normalization method should be largely avoided in DE analysis. This conclusion is also in agreement with the evaluation made by Bullard, Purdom, Hansen and Dudoit (ref. 2). Two other normalization methods, UpQu and Medi, assessed in our study display poor and very variable performances. On the other hand, our study indicated that the use of TMM or RLE methods lead to good performance on simulated data sets, though giving rise to some variability from one repetition to another. Finally, the MRN proposed method is found to be consistent and robust, producing globally better results that are remarkably stable from one repetition to another.

The study also shows that normalization methods globally behave similarly when trained with the simulated data sets and with the real data sets, which somehow tends to validate our simulation approach. In that regard, while, to our knowledge, a standard simulation procedure is still lacking, our current work provides a unique benchmark simulation procedure that could be useful for future researches on transcriptomics, metabolomics and proteomics data analyses.

We notice here that, independently to the normalization method used, the issue of isoform switch can influence the expression results, leading to erroneous DE analysis. Indeed, read counts calculation may change depending on the data treatment applied prior to the normalization step. This is obviously an important issue but not considered in our paper since it is prior to the normalization step.

Finally, we wanted to draw attention on the recent article written by Dillies, Rau, Aubert, Hennequet-Antier, Jeanmougin et al. (ref. 25), which appears when we were writing the present article. In this paper, the authors also give an interesting evaluation of normalization methods for RNA-Seq data analysis. We find that many of the results described in this article are consistent with our own results. Obviously, the MRN method that we propose here is not evaluated.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

This work was supported by the Laboratoire d'Excellence, LabEx, TULIP (ANR-10-LABX-41). This work benefited from the networking activities within the European funded COST ACTION FA1106 "QualityFruit."

Supplementary Materials

The reader is referred to the online Supplementary Materials for the 3 R programs (ref. 15) described below:

• **Simulation of RNA-Seq data.** Supplementary file 1 contains an R program for the simulation algorithm described and used in this paper. Parameters actually correspond to simulation type 1 (see **Table 1**), but can easily be changed in order to carry out other simulation types.

• **Normalizations and differential expression analyses.** Supplementary file 2 contains an R program for differential expression analyses with all the 9 normalization methods studied in this paper. This program can be directly executed with the output of Supplementary file 1.

• **Some graphical results.** Supplementary file 3 contains an R program providing some graphical results of differential expression analyses performed with the various normalization methods studied in this paper. This program can be directly executed with the output of Supplementary file 2.

Supplementary material may be found here:
http://www.landesbioscience.com/journals/cib/article/25849/

References

1. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010; 11:R106; PMID:20979621; http://dx.doi.org/10.1186/gb-2010-11-10-r106
2. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 2010; 11:94; PMID:20167110; http://dx.doi.org/10.1186/1471-2105-11-94
3. Cameron AC, Trivedi PK. Regression analysis of count data. Cambridge University Press, 1998.
4. Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. Anal Chem 2006; 78:4281-90; PMID:16808434; http://dx.doi.org/10.1021/ac051632c
5. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, et al.; on behalf of The French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. [First published online.]. Brief Bioinform 2012; •••:1-13; PMID:22988256
6. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 2011; 8:469-77; PMID:21623353; http://dx.doi.org/10.1038/nmeth.1613
7. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004; 5:R80; PMID:15461798; http://dx.doi.org/10.1186/gb-2004-5-10-r80
8. Gokce E, Shuford CM, Franck WL, Dean RA, Muddiman DC. Evaluation of normalization methods on GeLC-MS/MS label-free spectral counting data to correct for variation during proteomic workflows. J Am Soc Mass Spectrom 2011; 22:2199-208; PMID:21952779; http://dx.doi.org/10.1007/s13361-011-0237-2
9. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 2012; 13:204-16; PMID:22285995; http://dx.doi.org/10.1093/biostatistics/kxr054
10. Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am J Bot 2012; 99:248-56; PMID:22268221; http://dx.doi.org/10.3732/ajb.1100340
11. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008; 5:621-8; PMID:18516045; http://dx.doi.org/10.1038/nmeth.1226
12. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct 2009; 4:14; PMID:19371405; http://dx.doi.org/10.1186/1745-6150-4-14
13. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.
14. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. BMC Bioinformatics 2011; 12:480; PMID:22177264; http://dx.doi.org/10.1186/1471-2105-12-480
15. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol 2011; 12:R22; PMID:21410973; http://dx.doi.org/10.1186/gb-2011-12-3-r22
16. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010; 26:139-40; PMID:19910308; http://dx.doi.org/10.1093/bioinformatics/btp616
17. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010; 11:R25; PMID:20196867; http://dx.doi.org/10.1186/gb-2010-11-3-r25
18. Sato S, et al.; Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 2012; 485:635-41; PMID:22660326; http://dx.doi.org/10.1038/nature11119
19. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. Genome Res 2011; 21:2213-23; PMID:21903743; http://dx.doi.org/10.1101/gr.124321.111
20. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009; 25:1105-11; PMID:19289445; http://dx.doi.org/10.1093/bioinformatics/btp120
21. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012; 7:562-78; PMID:22383036; http://dx.doi.org/10.1038/nprot.2012.016
22. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010; 28:511-5; PMID:20436464; http://dx.doi.org/10.1038/nbt.1621
23. van der Vaart AW. Asymptotic statistics. Cambridge University Press, Cambridge, 1998. Cambridge Series in Statistical and Probabilistic Mathematics.

24. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009; 10:57-63; PMID:19015660; http://dx.doi.org/10.1038/nrg2484
25. Zenoni S, Ferrarini A, Giacomelli E, Xumerle L, Fasoli M, Malerba G, Bellin D, Pezzotti M, Delledonne M. Characterization of transcriptional complexity during berry development in Vitis vinifera using RNA-Seq. Plant Physiol 2010; 152:1787-95; PMID:20118272; http://dx.doi.org/10.1104/pp.109.149716