Briefings in Bioinformatics Advance Access published September 17, 2012

doi:10.1093/bib/bbs046

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium*

Submitted: 12th April 2012; Received (in revised form): 29th June 2012

Abstract

During the last 3 years, a number of approaches for the normalization of RNA sequencing data have emerged in the literature, differing both in the type of bias adjustment and in the statistical strategy adopted. However, as data continue to accumulate, there has been no clear consensus on the appropriate normalization method to be used or the impact of a chosen method on the downstream analysis. In this work, we focus on a comprehensive comparison of seven recently proposed normalization methods for the differential analysis of RNA-seq data, with an emphasis on the use of varied real and simulated datasets involving different species and experimental designs to represent data characteristics commonly observed in practice. Based on this comparison study, we propose practical recommendations on the appropriate normalization method to be used and its impact on the differential analysis of RNA-seq data.

Keywords: *high-throughput sequencing; RNA-seq; normalization; differential analysis*

INTRODUCTION

During the last decade, advances in Molecular Biology and substantial improvements in microarray technology have enabled biologists to make use of high-throughput genomic studies. In particular, the simultaneous measurement of the expression levels of tens of thousands of genes has become a mainstay of biological and biomedical research. For example, microarrays have been used to discover genes differentially expressed between two or more groups of interest in a variety of applications. These include the identification of disease biomarkers that may be important in the diagnosis of the different types and subtypes of diseases, with several implications in terms of prognosis and therapy [1, 2].

In recent years, the continuing technical improvements and decreasing cost of next-generation sequencing technology have made RNA sequencing (RNA-seq) a popular choice for gene expression studies. Such sequence-based methods have revolutionized studies of the transcriptome by enabling a wide range of novel applications, including detection of alternative splicing isoforms [3, 4], genome-guided [5, 6] or *de novo* assembly of transcripts [7–9], transcript fusion detection [10] or strand-specific expression [11]. In addition, RNA-seq has become an attractive alternative to microarrays for the identification of differentially expressed genes between several conditions or tissues, as it allows for high coverage of the genome and enables detection of weakly expressed genes [12].

Corresponding author. Marie-Agnès Dillies, Institut Pasteur, PF2 Plate-forme Transcriptome et Epigénome, 28 rue du Dr Roux, Paris CEDEX 15, F-75724 France. Tel.: +33 (0) 145688651; Fax: +33 (0) 145688406; E-mail: marie-agnes.dillies@pasteur.fr
*These authors have contributed equally to this work.

The French StatOmique Consortium gathers more than 40 statisticians and bioinformaticians involved in high-throughput transcriptome data analysis in France. The objective of this group, created in 2008, is to share among researchers and practitioners knowledge of the statistical analysis of high-throughput data.

In many ways, the progression of methodological development for RNA-seq data mirrors that of microarray data, although the bioinformatic and analytical pipelines differ considerably [13]. In particular, fragmented transcripts (short reads) are sequenced, rather than hybridized onto a chip, and must be assembled or aligned to a reference genome. Different sequencing technologies and protocols are currently available and share the same general pre-processing and analytical steps as follows [13]: (i) short reads are pre-processed (e.g. in order to remove adapters and low-quality sequences) and either mapped onto a genome reference sequence or assembled, (ii) the expression level is estimated for each biological entity (e.g. a gene) [5], (iii) the data are normalized and (iv) a statistical analysis is used to identify differentially expressed biological features [14]. Questions regarding all of these steps are still open and can have a strong impact on the analysis. In this work, we focus specifically on the third step, namely the issue of normalization for RNA-seq data in the context of differential analysis.

With both microarray and sequencing data, it has been shown that normalization is an essential step in the analysis of gene expression [15–17]. In microarray data analysis, normalization enables accurate comparisons of expression levels between and within samples by adjusting for systematic biases such as dye effect and hybridization artifacts [15, 18]. Although the technical biases inherent to microarray technology are not present in RNA-seq experiments, other sources of systematic variation have been reported, including between-sample differences such as library size (i.e. sequencing depth) [19] as well as within-sample gene-specific effects related to gene length [20] and GC-content [21]. In particular, larger library sizes result in higher counts for the entire sample. Although differences in library composition between samples may not be considered to be a source of systematic variation, they may contribute to a high level of biological variability.

During the last 3 years, a number of normalization approaches to treat RNA-seq data have emerged in the literature differing both in the type of bias adjustment and in the statistical strategy adopted. However, as data accumulate, there is still no clear indication of how the choice of normalization method impacts the downstream analysis. In addition, although effective and relevant methods have been derived and implemented to normalize RNA-seq data, they are not always properly used in practice. A small number of publications have compared normalization methods [16], providing useful yet preliminary results that must be confirmed with additional data to yield clear and robust guidelines to the community. To this end, we propose a systematic comparison of seven representative normalization methods for the differential analysis of RNA-seq data: Total Count (TC), Upper Quartile (UQ) [16], Median (Med), the DESeq normalization implemented in the `DESeq` Bioconductor package [14], Trimmed Mean of M values (TMM) implemented in the `edgeR` Bioconductor package [17], Quantile (Q) [22, 23] and the Reads Per Kilobase per Million mapped reads (RPKM) normalization [19].

In the past, comparisons among normalization methods for gene expression analysis have either made use of simulation studies or real calibration data [24–29]. Our comparison process is based on four real datasets sequenced using an Illumina sequencing machine, involving different species [*Homo sapiens* [30], *Mus musculus* (D. Castel, unpublished data), *Aspergillus fumigatus* (G. Janbon, unpublished data) and *Entamoeba histolytica* (C. C. Hon et al, submitted for publication)] and experimental designs, and dealing with both messenger RNAs and microRNAs. These four datasets were chosen to represent a broad range of characteristics and diversity typical of RNA-seq data analyses. Our comparison relies on both the qualitative characteristics of normalized data and the impact of the normalization method on the results from a differential expression (DE) analysis. In addition, a simulation study allows a further investigation of the impact of the normalization method on the false-positive rate and power of a DE analysis. Based on this study, we propose practical recommendations on the appropriate normalization method to be used and its impact on the differential analysis of RNA-seq data.

METHODS

In this section, we describe the normalization methods and real datasets used in our study, as well as the specific criteria used in our comparison.

Definitions

The datasets included in this study were obtained from two different Illumina sequencing machines, differing in their read length and overall throughput but sharing the same sequencing technology that

takes place on a glass slide called a 'flow cell'. A flow cell is made up of eight independent sequencing areas, or 'lanes'. Libraries are deposited on these lanes in order to be sequenced. A library contains cDNAs representative of the RNA molecules that are extracted from a given culture or tissue and are pre-processed in order to be adapted to the sequencing procedure. Similarly to microarrays, the library composition reflects the RNA repertoire expressed in the corresponding culture or tissue. The 'library size' refers to the number of mapped short reads obtained from the sequencing process of the library. In this study, a single library was sequenced in each lane.

Normalization methods

Because the most obvious source of variation between lanes is the differences in library size (i.e. sequencing depth), the simplest form of inter-sample normalization is achieved by scaling raw read counts in each lane by a single lane-specific factor reflecting its library size. We consider five different methods for calculating these scaling factors, described as follows:

Total count (TC): Gene counts are divided by the total number of mapped reads (or library size) associated with their lane and multiplied by the mean total count across all the samples of the dataset.

Upper Quartile (UQ): Very similar in principle to TC, the total counts are replaced by the upper quartile of counts different from 0 in the computation of the normalization factors [16].

Median (Med): Also similar to TC, the total counts are replaced by the median counts different from 0 in the computation of the normalization factors.

DESeq: This normalization method [14] is included in the `DESeq` Bioconductor package (version 1.6.0) [14] and is based on the hypothesis that most genes are not DE. A DESeq scaling factor for a given lane is computed as the median of the ratio, for each gene, of its read count over its geometric mean across all lanes. The underlying idea is that non-DE genes should have similar read counts across samples, leading to a ratio of 1. Assuming most genes are not DE, the median of this ratio for the lane provides an estimate of the correction factor that should be applied to all read counts of this lane to fulfill the hypothesis. By calling the `estimateSizeFactors()` and `sizeFactors()` functions in the `DESeq` Bioconductor package, this factor is computed for each lane, and raw read counts are divided by the factor associated with their sequencing lane.

Trimmed Mean of *M*-values (TMM): This normalization method [17] is implemented in the `edgeR` Bioconductor package (version 2.4.0). It is also based on the hypothesis that most genes are not DE. The TMM factor is computed for each lane, with one lane being considered as a reference sample and the others as test samples. For each test sample, TMM is computed as the weighted mean of log ratios between this test and the reference, after exclusion of the most expressed genes and the genes with the largest log ratios. According to the hypothesis of low DE, this TMM should be close to 1. If it is not, its value provides an estimate of the correction factor that must be applied to the library sizes (and not the raw counts) in order to fulfill the hypothesis. The `calcNormFactors()` function in the `edgeR` Bioconductor package provides these scaling factors. To obtain normalized read counts, these normalization factors are re-scaled by the mean of the normalized library sizes. Normalized read counts are obtained by dividing raw read counts by these re-scaled normalization factors.

In addition to these scaling methods, we consider two alternative strategies:

Quantile (Q): First proposed in the context of microarray data, this normalization method consists in matching distributions of gene counts across lanes [22, 23]. It is implemented in the Bioconductor package `limma` [31] by calling the `normalizeQuantiles()` function.

Reads Per Kilobase per Million mapped reads (RPKM): This approach was initially introduced to facilitate comparisons between genes within a sample and combines between- and within-sample normalization, as it re-scales gene counts to correct for differences in both library sizes and gene length [19]. However, it has been shown that attempting to correct for differences in gene length in a differential analysis actually has the effect of introducing a bias in the per-gene variances, in particular for lowly expressed genes [20]. Despite these findings, the RPKM method continues to be a popular choice in many practical applications.

All of these methods can be divided into two subgroups referring to the library size concept

(TMM and DESeq) or distribution adjustment of read counts (TC, UQ, Med, Q, RPKM). Both TMM and DESeq rely on the hypothesis that most of the genes are not DE. They both propose a scaling factor based on a mean, or median, ratio. However, for TMM this ratio is computed between each test lane and the reference one, while for DESeq all samples are taken into account. Finally, DESeq scaling factors apply to read counts, while those calculated using TMM apply to library sizes. The second group is composed of methods that assume similarities between read count distributions, either on a single quantile (TC, Med, UQ, RPKM) or on all quantiles (Q). RPKM includes both a TC and gene length normalization.

Finally, in addition to the main methods described above, some proposed strategies for RNA-seq data normalization focus on the use of housekeeping genes [16] or on the putative bias associated to GC-content [30, 32]. We did not include such a normalization strategy in our study because a close inspection of our datasets did not confirm the presence of such a bias (Supplementary Figure S13). As such, we assume that the GC bias associated with each gene is constant across conditions and does not need to be corrected in the context of a differential analysis. However, these normalization methods are further discussed in Supplementary Data.

The seven normalization methods are also compared to the raw unnormalized data, denoted by Raw Counts (RC). All the analyses are performed with `R` 2.14; the scripts used to implement each method are available in Supplementary Data. It is worth noting that all of the scaling normalization approaches described above can easily be modified to produce an offset parameter to be incorporated within a statistical model for DE.

Real data

The seven normalization methods previously described are compared based on four real RNA-seq datasets involving different species and experimental designs as well as very different characteristics in terms of reproducibility between replicates, the presence of high-count sequences, the library sizes, differences in library composition between biological conditions and the importance of gene length in estimates of gene expression (Table 1). The four datasets as well as additional details about each experiment, data pre-processing and bioinformatics steps are included in Supplementary Data.

Dataset descriptions

Homo sapiens melanoma cell lines (Hs): These human data correspond to a comparison between a melanoma cell line expressing the Microphtalmia Transcription Factor (MiTF) and a melanoma cell line in which small interfering RNAs (siRNAs) are used against MiTF in order to lower its expression [33].

Entamoeba histolytica strains (Eh): *Entamoeba histolytica* is a unicellular protozoa that can be ingested through soiled water. This human parasite is the causative agent of amebiasis, one of the three most common causes of death worldwide. The data included in this study compare gene expression between two strains of *E. histolytica* (Eh), one being virulent (HM1:IMSS) and the other being attenuated (Rahman) (C. C. Hon et al, submitted for publication).

Aspergillus fumigatus (Af): *Aspergillus fumigatus* is a fungus whose spores are present not only in the air we breathe but also in soils and decaying organic matter. It does not normally cause illness but can induce fatal pulmonary infections to individuals with a weakened immune status. These RNA-seq data compare the transcriptome of *A. fumigatus* strain 1163 in two different growth media.

Mus musculus muscle stem cells (Mm): These data are related to a transcriptome study where the expression of miRNAs was measured in three different cellular stages of the skeletal muscle lineage in adult mouse.

Comparison procedures

Qualitative characteristics of normalized data. For each dataset, the seven normalization methods are compared based on qualitative characteristics of normalized data, including the count distributions and variability between biological replicates. Boxplots of raw and normalized read counts are calculated as log2(read count + 1) in order to avoid problems associated with zero values. The within-condition variability measure is based on the coefficient of variation per gene. Boxplots represent the distribution of this coefficient across samples.

We also investigated the average variation of a set of 30 housekeeping genes in the human data, assuming that these genes are similarly expressed across samples (lanes). The housekeeping genes were selected from a previously described list [34] and presented the least variation across the 84 human cell types of the GeneAtlas data [35] available on `GEO`

Table 1: Summary of datasets used for comparison of normalization methods, including the organism, type of sequencing data, number of genes, number of replicates per condition, minimum and maximum library sizes, Pearson correlation between replicates and between samples of different conditions (minimum, maximum), percentage of reads associated with the most expressed RNA (minimum, maximum), library type (SR = single-read or PE = paired-end read, read length, D = directional or ND = non-directional) and Illumina sequencing machine

Organism	Type	Number of genes	Replicates per condition	Minimum library size	Maximum library size	Correlation between replicates	Correlation between conditions	% Most expressed gene	Library type	Sequencing machine
H. sapiens	RNA	26 437	{3, 3}	2.0×10^7	2.8×10^7	(0.98, 0.99)	(0.93, 0.96)	≈ 1%	SR 54, ND	GaIIx
A. fumigatus	RNA	9248	{2, 2}	8.6×10^6	2.9×10^7	(0.92, 0.94)	(0.88, 0.94)	≈ 1%	SR 50, D	HiSeq2000
E. histolytica	RNA	5277	{3, 3}	2.1×10^7	3.3×10^7	(0.85, 0.92)	(0.81, 0.98)	6.4–16.2%	PE 100, ND	HiSeq2000
M. musculus	miRNA	669	{3, 2, 2}	2.0×10^6	5.9×10^6	(0.95, 0.99)	(0.09, 0.75)	17.4–51.1%	SR 36, D	GaIIx

(http://www.ncbi.nlm.nih.gov/geo) with the accession number GSE1133.

Differential expression analysis. The seven normalization methods are compared based on results from a DE analysis performed with the Bioconductor package `DESeq` and the Two-Stage Poisson Model (TSPM) [36], both described below. In addition to comparing the number of DE genes and the number of common DE genes found among the methods, we generate, for each real dataset, a dendrogram representing the similarity between the DE gene lists obtained with each normalization method, based on the binary distance and the Ward linkage algorithm (`dist()` and `hclust()` functions in `R`). The four dendrograms (Supplementary Figure S4) are subsequently merged into a consensus dendrogram resulting from the mean of the distance matrices obtained from each real dataset.

Simulations

Simulation model

The simulation model is similar to one previously used [29] and adapted to counts. Let N be the number of genes and M the number of samples divided into two conditions, and let x_{ij} be the expression value of a given gene i in sample j. We assume x_{ij} follows a Poisson distribution of parameter λ_{jk} according to the condition k to which sample j belongs. Under this model, the null hypothesis H_0 of no difference between the two conditions is equivalent to $\lambda_{i2} = \lambda_{i1}$; the alternative hypothesis H_1 of DE between the two conditions is equivalent to $\lambda_{i2} \neq \lambda_{i1}$. Finally, let π_0 (resp. π_1) be the proportion of genes generated under H_0 (resp. H_1) among the N genes.

Data were simulated with $N = 15\,000$, $M = 20$ (10 samples per condition) and π_1 increasing from 0% to 30%. In order to generate realistic data, the parameter λ_{i1} used to sample the gene i from a Poisson distribution for the first condition corresponds to the observed mean expression for each gene estimated from the *M. musculus* data; the parameter λ_{i2} used to sample the gene i from a Poisson distribution for the second condition is equal to λ_{i1} under H_0 and to $(1 + \tau)\lambda_{i1}$ under H_1, with $\tau = \pm 0.2$. To assess the impact of non-equivalent library sizes, we added the possibility of multiplying all gene expression values x_{ij} for a given sample j by a constant K_j taken to be equal to $|\varepsilon_j|$, where ε_j is drawn from a $N(1, 1)$ distribution. In addition, the *M. musculus* data contain a set of highly expressed genes contributing to the majority of total counts, which enables an assessment of the impact of such high-count genes in the simulated data.

False-positive rate and power

For each simulated dataset, the false-positive rate (power) can be estimated based on the genes simulated under H_0 (H_1). We consider three settings: (i) equivalent library sizes across lanes and no high-count genes, (ii) non-equivalent library sizes across lanes and no high-count genes and (iii) equivalent library sizes across lanes and presence of high-count genes. For each scenario and proportion of H_1 tested, the false-positive rate and the power were averaged over 10 simulated datasets to ensure a reasonable precision.

Differential expression analysis

In both the real and simulated data, the impact of the normalization methods is assessed using the results from a DE analysis. For this test, we choose to use two methods based on different models: the `DESeq`

Bioconductor package [14] and the TSPM [36], which may be implemented using an `R` script found at the corresponding author's website. The DESeq method, which was specifically developed to find differentially expressed genes between two conditions for RNA-seq data with small sample size and overdispersion, uses a model based on a negative binomial distribution and local regression to estimate the relationship between the mean and variance of each gene. `DESeq` was chosen because it is widely used in practice. In addition, it allows scaling factors to be easily included in the statistical test, and in contrast to `edgeR`, the statistical test does not assume comparable distribution of read counts. The `DESeq` Bioconductor package (version 1.6.0) with default setting was employed. The package accommodates each normalization method via the specification of size factors in the following function:

```
AnnotatedDataSet(pData(cds)$sizeFactor <-...)
```

In order to confirm these results using an alternative method, we have also applied the TSPM [36], which makes use of a model based on the Poisson, rather than negative binomial, distribution. The TSPM evaluates the presence of overdispersion on a gene-by-gene basis in a first stage, and subsequently tests for DE using a standard likelihood approach for genes displaying evidence of overdispersion, or a likelihood ratio test statistic for those without evidence of overdispersion.

For both methods, raw *P*-values were adjusted for multiple comparisons by the Benjamini–Hochberg procedure [37], which controls the false discovery rate. Genes with an adjusted *P*-value < 0.05 were considered to be differentially expressed.

RESULTS

In comparing the seven normalization methods, we aim to identify methods that appear to have both robust and stable performance across real or simulated datasets exhibiting a variety of characteristics. We note that RNA-seq technology provides the opportunity to explore the expression of transcripts rather than genes for organisms exhibiting complex transcription patterns. Although some of the normalization strategies included in this study apply to both read counts and estimated expression levels, others are adapted only to read counts. As such, all data included in this study contain gene-level read counts rather than estimated transcript expression levels.

Real data

We consider two criteria for the comparisons made on four real datasets, described in detail in Table 1: (i) the qualitative characteristics of normalized data and (ii) results from DE analyses. For the former, we focus on boxplots of the distribution of counts as well as a study of intra-group variability. We remark that drawing definitive conclusions from these qualitative comparisons concerning the performance of each normalization method is typically not possible; however, such exploratory analyses are often undertaken in the early stages of an analysis and help shed light on characteristics of the data and the impact of the normalization process on the data distribution prior to further analysis. For the latter, we study the lists of differentially expressed genes between conditions identified following the use of each normalization method in each dataset.

Qualitative characteristics of normalized data

Like microarray data, in typical DE analyses the majority of genes under consideration are often assumed to be non-differentially expressed between conditions. For this reason, it is useful to examine boxplots of counts across samples in each dataset, both before and after normalization; an effective normalization scheme should result in a stabilization of read count distributions across replicates. For data with small differences in library size and little inter-sample variability (e.g. the *H. sapiens* data), it is perhaps unsurprising that all methods, including the unnormalized raw counts, yield comparable results (Supplementary Figure S1). However, when large differences in library size exist (e.g. the *A. fumigatus* and *M. musculus* data), these sample-to-sample differences are evident in the boxplots for the unnormalized raw counts.

In the case of the *M. musculus* miRNA-seq data, we note that although most of the other methods appear to perform similarly in stabilizing these differences, TC and RPKM do not improve over the raw counts (Figure 1a). A similar pattern may be seen in the results for the *A. fumigatus* data (Supplementary Figure S1). In addition to large differences in library size, the *M. musculus* data also exhibit the presence of high-count genes (i.e. a few genes whose read counts contribute to a large proportion of the total count for a given sample) associated with different expressed RNA repertoires. The TC normalization method thus corrects for differences in sequencing depth, but it is unable to handle differences in RNA

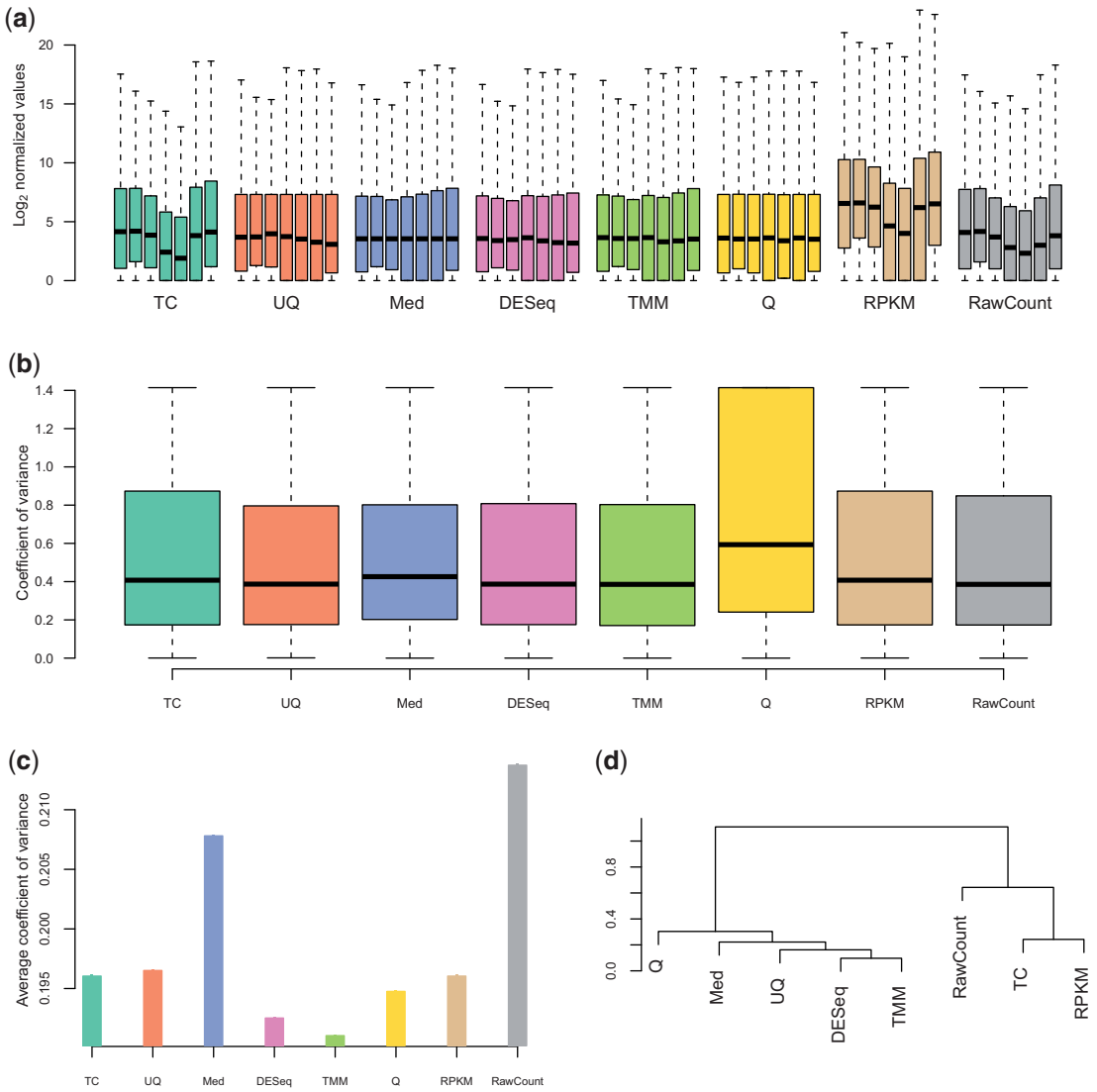

Figure 1: Comparison of normalization methods for real data. (**A**) Boxplots of log2(counts + 1) for all conditions and replicates in the *M. musculus* data, by normalization method. (**B**) Boxplots of intra-group variance for one of the conditions (labeled 'B' in the corresponding data found in Supplementary Data) in the *M. musculus* data, by normalization method. (**C**) Analysis of housekeeping genes for the *H. sapiens* data. (**D**) Consensus dendrogram of differential analysis results, using the **DESeq** Bioconductor package, for all normalization methods across the four datasets under consideration.

composition among samples [17]. In addition, we note that even following a normalization using the Q method, the first quantiles of the samples in the *M. musculus* are not aligned; this is due to a subset of samples that contain a much higher proportion of 0 counts (Supplementary Figure S12).

These boxplots of normalized values also indicate subtle discrepancies between normalization methods that are similar in nature. As an example, we remark upon the differences between the Med and UQ methods in the *M. musculus* data; the former aligns the median values for counts across all samples, while the latter aligns the upper quantile of counts across all samples. However, differences in library composition across samples, such as the aforementioned presence of high-count genes or a large numbers of 0 counts, affect the calculation of these scaling factors unequally (Supplementary Figures S9–S11 and S14).

It is also of interest to consider which normalizations are able to minimize intra-condition variance. In most of the datasets considered here, little difference is observed among the normalization methods (Supplementary Figure S2). One exception occurs in the *M. musculus* data, where Q normalization actually

appears to increase, rather than decrease, the intra-group variance for one of the conditions (Figure 1b). This can be explained by looking at read count distributions across the seven mouse samples. In particular, the read count distributions in one of the conditions (labeled 'B' in the corresponding data found in Supplementary Data) are quite different from those in the other two conditions, with more extreme counts (very low or very high) but fewer moderate counts (data not shown). As the Q normalization process corrects gene counts by matching distributions across all samples on the basis of the mean distribution, read counts of this condition are corrected more than read counts of the others. This over-correction in turn increases intra-condition variability, especially for genes with moderate counts.

Finally, we consider the effect of the various normalization methods on the variation in expression among a set of housekeeping genes in the human data, which may be assumed to be similarly expressed across samples. Figure 1c represents the average coefficient of variation of 30 known housekeeping genes in the human data (see Supplementary Data for further detail). Considering that these genes are assumed to have relatively constant expression, we note that the DESeq and TMM normalization methods lead to the smallest coefficient of variation. Although choosing an appropriate set of such housekeeping genes can be difficult, these results complement the previous qualitative observations concerning the behavior of the normalization methods under consideration.

Differential expression analysis

Because the aim of this comparative study is to determine the downstream effect of the choice of normalization method, we also consider results from a DE analysis based on the `DESeq` Bioconductor package and TPSM method. With real data, it is difficult to determine whether a particular normalization method is superior to the others (e.g. through the false-positive rate). However, the advantage of such a comparison is that it allows us to determine which methods perform similarly.

Table 2 indicates that there is a great deal of overlap among all of the normalization methods in data with little inter-sample variability (e.g. the *E. histolytica* data) using the `DESeq` package; the same general trend may be seen with results from the TSPM (Supplementary Table S10). However, across datasets Q and RPKM tend to uniquely identify weakly expressed genes as differentially expressed (Supplementary Figure S3). These same patterns were observed across all datasets for the `DESeq` method (Supplementary Figure S4, Supplementary Tables S5–S9) and are displayed in the consensus dendrogram tree in Figure 1d. This consensus dendrogram illustrates a trend, namely that in the results from a DE analysis, the TC normalization tends to group with RPKM and the unnormalized raw counts, while the remaining methods tend to group together. We note that although the number of genes identified as differentially expressed differs between the DESeq and TSPM methods (Supplementary Tables S5–S10), the same general relationships may be observed among the different normalization methods, and the consensus dendrogram tree constructed using results from the TSPM is nearly identical to that constructed from the DESeq results (Supplementary Figure S15). This suggests that the relationships identified among the normalization methods are not simply linked to the model used for the differential analysis.

Simulations

Although comparisons using real data are informative, simulations complement these results by allowing different factors, including differences in library size and RNA composition, to be controlled. With this in mind, the false-positive rate and power resulting from the DE analysis may be calculated in a variety of scenarios: equivalent or non-equivalent library sizes between lanes and presence or not of high-count genes contributing to a large proportion of the total count for a given sample. By varying these factors, differences among the normalization methods become more apparent.

In situations where library sizes are simulated to be equivalent and no high-count genes are present, all normalization methods considered perform nearly identically to the unnormalized raw counts in terms of the false-positive rate and power, using the `DESeq` Bioconductor package; this is unsurprising, as normalization is unneeded in such a case (Supplementary Figure S5a). In situations where library sizes are different (Supplementary Figure S5b), we note that the nominal false-positive rate is not maintained and the power is significantly decreased for the unnormalized data. All of the normalization methods are able to correct for these differences in library sizes, as all control the false-positive rate and

Table 2: Number of differentially expressed genes found in common for each of the normalization methods using the **DESeq** Bioconductor package, as well as the unnormalized raw counts (RC), in the *E. histolytica* data

	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	548	547	547	543	547	543	399	175
UQ		**1213**	**1195**	**1160**	**1172**	**1054**	416	184
Med			**1218**	**1147**	**1160**	**1043**	416	183
DESeq				**1249**	**1169**	**1058**	413	184
TMM					**1190**	**1051**	516	184
Q						**1092**	414	184
RPKM							417	149
RC								184

Counts along the diagonal indicate the number of DE genes per method (i.e. 548 DE genes for the TC method, etc.), while counts off the diagonal indicate the number of DE genes in common per pair of methods (i.e. 547 DE genes in common between TC and UQ). Numbers in bold correspond to pairs of methods with very similar lists of DE genes.

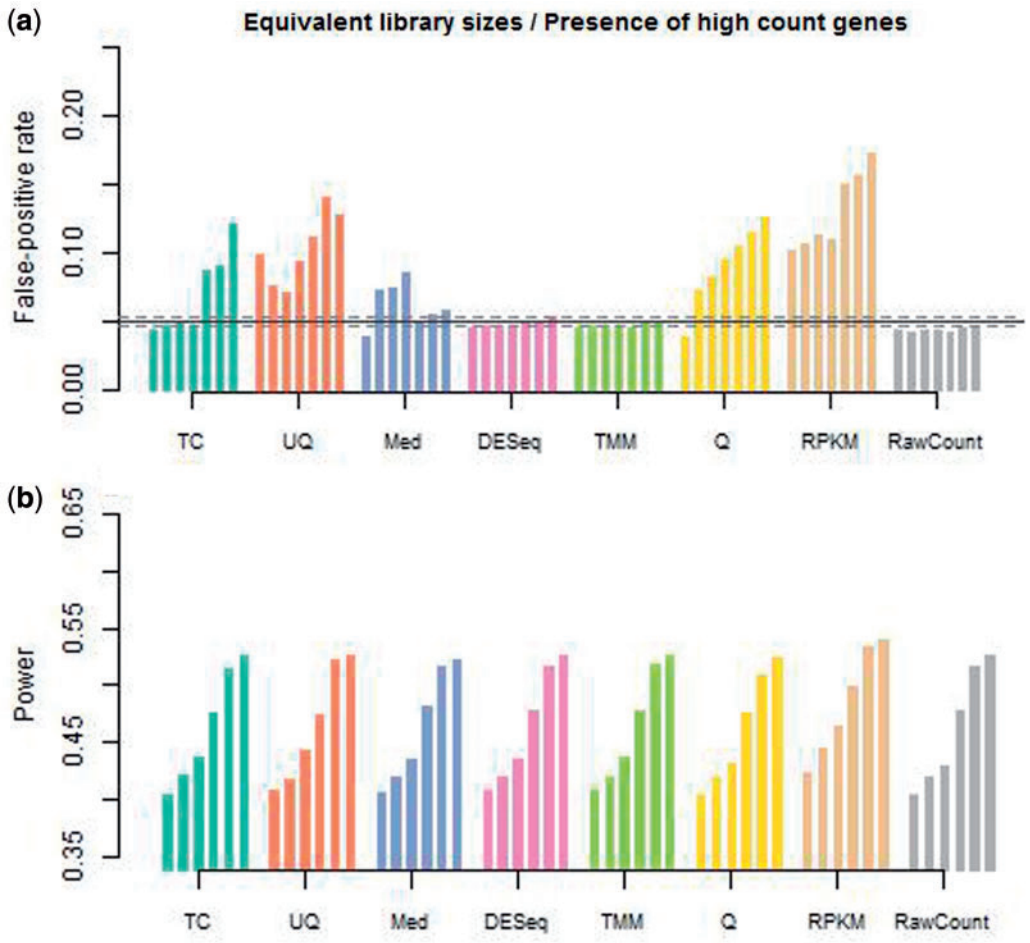

Figure 2: Comparison of normalization methods for simulated data with equal library sizes and the presence of high-count genes. (**A**) Average false-positive rate over 10 independent datasets simulated with varying proportions of differentially expressed genes (from 0% to 30% for each normalization method). (**B**) Power over 10 independent datasets simulated with varying proportions of differentially expressed genes (from 5% to 30% for each normalization method).

maintain a reasonable power. Figure 2 presents results from the most discriminant simulation setting, where the library sizes are simulated to be equivalent for all samples with the presence of a few high-count genes. This setting indicates that contrary to the situation with varying library sizes, the presence of high-count genes does not impact the performance of raw counts; this seemingly contradictory result is due to the fact that the data are simulated under the model used for the differential analysis. However, the presence of these high-count genes clearly results in an inflated false-positive rate for five of the normalization methods (TC, UQ, Med, Q and RPKM). Only DESeq and TMM are able to control the false-positive rate while also maintaining the power to detect differentially expressed genes.

DISCUSSION

Despite initial optimistic claims that RNA-seq data do not require sophisticated normalization [38], in practice normalization remains an important issue since raw counts are often not directly comparable within and between samples. While this subject has received some attention in the literature, the increasing number of RNA-seq normalization methods makes it challenging for scientists to decide which method to use for their data analysis. Given the fact that the choice of normalization has a great influence on the subsequent statistical analyses, the quality and credibility of these methods need to be assessed fairly [39]. To this end, our comparison study deals with seven representative normalization strategies compared on four real datasets involving different species and experimental designs, and on simulated datasets representing various scenarios.

Based on three real mRNA and one miRNA-seq datasets, we confirm previous observations that RPKM and TC, both of which are still widely in use [40, 41], are ineffective and should be definitively abandoned in the context of differential analysis. The RPKM approach was initially proposed to account for differences in gene length [19]; however, the relationship between gene length and DE actually varies among the datasets considered here (Supplementary Figures S6–S8). Even in cases where a strong positive association between gene counts and length is observed, scaling counts by gene length with RPKM is not sufficient for removing this bias [16, 20]. Several alternative approaches to account for gene length at the steps of normalization, differential analysis or gene-set analysis have been proposed [19, 32, 42], but no standard strategy has yet been identified. The TC approach, on the other hand, ignores the fact that different biological samples may express different RNA repertoires. In addition, it may too often be biased by the behavior of a relatively small number of high-count genes that are not guaranteed to have similar levels of expression across different biological conditions [16]. Similarly, Q is based on the strong assumption that all samples must have identical read count distributions. As shown in our comparison, this may lead to increased within-condition variability and should be avoided. The other normalization methods (UQ, Med, DESeq and TMM) perform similarly on the varied datasets considered here, both in terms of the qualitative characteristics of the normalized data and the results of DE analyses.

Simulations allow a further discrimination of the seven methods, in particular in the presence of high-count genes, where it appears that only DESeq and TMM are able to maintain a reasonable false-positive rate without any loss of power. One should note that DESeq and TMM are also indicated through an investigation of the variation of housekeeping genes in the *H. sapiens* data, although this analysis should be interpreted with caution. Housekeeping genes are assumed to have similar expression levels across samples of different tissues, but there is no guarantee that this hypothesis holds in every condition tested. However, taken together with the previous conclusions, these results confirm the satisfactory behavior of the DESeq and TMM methods. We also remark that in terms of the scaling factors used, DESeq and TMM are the most similar normalization methods. Finally, these two methods do not explicitly include an adjustment of count distributions across samples, allowing samples to exhibit differences in library composition. It is not surprising, then, that these two methods performed much better than the others for data with differences in library composition. A summary of these conclusions is shown in Table 3.

It is important to keep in mind that most normalization strategies (including DESeq and TMM) rely on the rather strong assumptions that most genes are not differentially expressed, and that for those differentially expressed there is an approximately balanced proportion of over- and under-expression [22, 43]. Though these assumptions appear reasonable in many studies, including those considered here, there are experiments in which they are not met.

Table 3: Summary of comparison results for the seven normalization methods under consideration

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	−	+	+	−	−
UQ	++	++	+	++	−
Med	++	++	−	++	−
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
Q	++	−	+	++	−
RPKM	−	+	+	−	−

A '−' indicates that the method provided unsatisfactory results for the given criterion, while a '+' and '++' indicate satisfactory and very satisfactory results for the given criterion.

Unfortunately, these assumptions are rarely checked in practice; in fact, it would be extremely difficult to do so. In recent work, to address the observation that the proportion of DE genes can affect normalization quality, Kadota *et al*. [44] proposed an alternative multi-step normalization strategy in which genes that are determined to be potentially DE are removed prior to estimation of scaling factors using the TMM normalization method. This work suggests that in some cases, the appropriate choice of parameters can lead to slight improvements in performance in the TMM method.

On a practical note, DESeq and TMM are straightforward to apply through a command of the `DESeq` and `edgeR` Bioconductor packages, respectively. We note that unlike the other methods, TMM and DESeq use a normalization factor within the statistical model for differential analysis, rather than on the data themselves; one consequence of this approach is that the corresponding packages do not automatically provide normalized read counts to the end user, although this information is often appreciated and requested by biologists. However, normalized read counts for the DESeq and TMM methods can be obtained through a simple command in the `DESeq` package or a series of `R` commands, respectively, as shown in Supplementary Data. As the two packages implement normalization methods with comparable performance, a comparison of their respective statistical models dedicated to differential analysis may provide further arguments to favor one of the two methods.

The present study represents a major step toward a more comprehensive use of normalization methods for RNA-seq data and will be of great help to biologists that are confronted with RNA-seq data analyses. As sequencing technology continues to mature, the use of multiplexed experiments will likely become increasingly common, paving the way to a dramatic growth in the amount of data produced; additional work will be needed to determine how to include such multiplexed samples within a normalization scheme. In addition, this work is restricted to normalization methods for processing read counts, and as such its conclusions are limited to this context. In particular, it assumes that complex transcriptomes are studied at the gene, rather than transcript, level. Normalization and differential analysis at the transcript level require the use of sophisticated statistical models such as Cufflinks [5] or RSEM [45] in order to estimate, rather than count, expression levels of these transcripts. These estimates do not have the same statistical properties as read counts and may not be described by the same models or processed by the same normalization algorithms. An exception can be made for the `DEXseq` Bioconductor package [46], which proposes a detection of differential exon usage based on read counts *per exon* and applying the DESeq normalization. Another comparative study will be carried out in the future to address this more complex yet fruitful area.

SUPPLEMENTARY DATA

Supplementary Data are available online at http://bib.oxfordjournals.org/.

Key points

- Normalization of RNA-seq data in the context of differential analysis is essential in order to account for the presence of systematic variation between samples as well as differences in library composition.
- The Total Count and RPKM normalization methods, both of which are still widely in use, are ineffective and should be definitively abandoned in the context of differential analysis.
- Only the DESeq and TMM normalization methods are robust to the presence of different library sizes and widely different library compositions, both of which are typical of real RNA-seq data.

Acknowledgements

We thank Chung Chau Hon from Institut Pasteur (whose work was supported by the French National Research Agency (ANR-10-GENM-011)) for his very helpful support in providing up-to-date annotations of *E. histolytica*, as well as Thomas Strub, Irwin Davidson and the IGBMC sequencing platform for supplying the *H. sapiens* RNA-seq data. We also thank Guilhem Janbon from Institut Pasteur, who designed the *A. fumigatus* experiment and kindly accepted that his data be included in this study, and Delphine Charif, who participated in discussions concerning this work. On behalf of the French StatOmique Consortium and in alphabetical order: J.A., M.-A. D., M.G., C.H.-A., F.J., M.J., C.K., S.L.C., A.R. and N.S. wrote the manuscript. J.A., D.C., M.-A.D., J.E., M.G., G.G., C.H.-A., F.J., B.J., M.J., L.J., C.K., D.L., S.L.C., C.L.G., G.M., A.R., B.S. and N.S. designed and performed the analyses.

FUNDING

This work was supported by the Groupe de Rercherche Bioinformatique Moléculaire (GdR BiM, http://www.gdr-bim.u-psud.fr). D.C. was supported by a DIM STEM-Pole fellowship and Association Française contre les Myopathies.

References

1. van 't Veer LJ, Dai H, van de Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;**415**:530–6.
2. Sørlie T, Tibshirani R, Parker J, *et al*. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* 2003;**100**:8418–23.
3. Wang ET, Sandberg R, Luo S, *et al*. Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008; **456**(7221):470–6.
4. Pan Q, Shai O, Lee LJ, *et al*. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 2008;**40**(12):1413–5.
5. Trapnell C, Williams BA, Pertea G, *et al*. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;**28**:511–5.
6. Guttman M, Garber M, Levin JZ, *et al*. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* 2010;**28**:503–10.
7. Robertson G, Schein J, Chiu R, *et al*. De novo assembly and analysis of RNA-seq data. *Nat Methods* 2010;**11**:909–12.
8. Grabherr MG, Haas BJ, Yassour M, *et al*. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol* 2011;**29**(7):644–52.
9. Schulz MH, Zerbino DR, Vingron M, *et al*. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 2012;**28**(8):1086–92.
10. Maher CA, Kumar-Sinha C, Cao X, *et al*. Transcriptome sequencing to detect gene fusions in cancer. *Nature* 2009; **458(7234)**:97–101.
11. Levin JZ, Yassour M, Adiconis X, *et al*. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* 2010;**7(9)**:709–15.
12. Marguerat S, Bähler J. RNA-seq: from technology to biology. *Cell Mol Life Sci* 2010;**67(4)**:569–79.
13. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. *Genome Biol* 2010; **11**(220).
14. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11(R106)**:R106.
15. Park T, Yi SG, Kang SH, *et al*. Evaluation of normalization methods for microarray data. *BMC Bioinformatics* 2003; **4**(33).
16. Bullard JH, Purdom E, Hansen KD, *et al*. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics* 2010; **11**(94).
17. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010;**11**(R25).
18. Quackenbush J. Microarray data normalization and transformation. *Nat Genet* 2002;**32**:496–501.
19. Mortazavi A, Williams BA, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* 2008;**5**:621–8.
20. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* 2009;**4**(14).
21. Pickrell JK, Marioni JC, Pai AA, *et al*. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010;**464**(7289): 768–72.
22. Bolstad BM, Irizarry RA, Astrand M, *et al*. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003; **19**:185–93.
23. Yang YH, Thorne NP. Normalization for two-color cDNA microarray data. *Science and Statistics: A Festschrift for Terry Speed*, Vol. 40. IMS Lecture Notes – Monograph Series, 2003, 403–18.
24. Shedden K, Chen W, Kuick R, *et al*. Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data. *BMC Bioinformatics* 2005;**6**(26).
25. Qin LX, Beyer RP, Hudson FN, *et al*. Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. *BMC Bioinformatics* 2006;**7**(23).
26. Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* 2006;**7**(359).
27. Jaffrézic F, Marot G, Degrelle S, *et al*. A structural mixed model for variances in differential gene expression studies. *Genet Res* 2007;**89**:19–25.
28. McCall MN, Irizarry RA. Consolidated strategy for the analysis of microarray spike-in data. *Nucleic Acids Res* 2008; **36**(17):e108.
29. Jeanmougin M, de Reynies A, Marisa L, *et al*. Should we abandon the t-test in the analysis of gene expression microarray data: A comparison of variance modeling strategies. *PLoS ONE* 2010;**5**(9):e12336.

30. Strub T, Giuliano S, Ye T, *et al*. Essential role of microphthalmia transcription factor for DNA replication, mitosis and genomic stability in melanoma. *Oncogene* 2011;**30**:2319–32.
31. Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, (eds). *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. New York: Springer, 2005;397–420.
32. Risso D, Schwartz K, Sherlock G, *et al*. GC-content normalization for RNA-seq. *BMC Bioinformatics* 2011;**12**:480.
33. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 2012;**13**(2):204–216.
34. Eisenberg E, Levanon EY. Human housekeeping genes are compact. *Trends Genet* 2003;**19(7)**:362–5.
35. Su AI, Wiltshire T, Batalov S, *et al*. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 2004;**101(16)**:6062–7.
36. Auer PL, Doerge RW. A two-stage Poisson model for testing RNA-seq data. *Stat Appl Genet Mol Biol* 2011;**10**:1–28.
37. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Stat Methodol* 1995;**57**(1):289–300.
38. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
39. Hofmann R, Seidl T, Dugas M. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol* 2002;**3**.11:research0033–research0033.
40. Liu S, Lin L, Jiang P, *et al*. A comparison of RNA-seq and high-density exon array for detecting differential gene expression between related species. *Nucleic Acids Res* 2011; **39**(2):578–88.
41. Isabella VM, Clark VL. Deep sequencing-based analysis of the anaerobic stimulon in Neisseria gonorrhoeae. *BMC Genomics* 2011;**12**(51).
42. Young MD, Wakefield MJ, Smyth GK, *et al*. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010;**11(2)**:R14.
43. Calza S, Pawitan Y. Normalization of gene-expression microarray data. *Methods Mol Biol* 2010;**673**:37–52.
44. Kadota K, Nishiyama T, Shimizu K. A normalization strategy for comparing tag count data. *Algorithms Mol. Biol.* 2012;**7**:5.
45. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* 2011;**12**:323.
46. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Research* 2012. doi:10.1101/gr.133744.111.