doi: 10.1093/nar/gkv711

Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package

Sonia Tarazona[1,2], Pedro Furió-Tarí[1], David Turrà[3], Antonio Di Pietro[3], María José Nueda[4], Alberto Ferrer[2] and Ana Conesa[1,5,*]

[1]Genomics of Gene Expression Lab, Centro de Investigación Príncipe Felipe, Eduardo Primo Yúfera 3, 46012, Valencia, Spain, [2]Department of Applied Statistics, Operations Research and Quality, Universidad Politécnica de Valencia, Camí de Vera, 46022, Valencia, Spain, [3]Department of Genetics, Universidad de Córdoba, Campus de Rabanales Edificio Gregor Mendel, 14071, Córdoba, Spain, [4]Statistics and Operational Research Department, Universidad de Alicante, Carretera San Vicente del Raspeig s/n, 03690, Alicante, Spain and [5]Microbiology and Cell Science Department, Institute for Food and Agricultural Sciences, University of Florida, Gainesville, FL 32603, USA

Received April 21, 2015; Revised June 12, 2015; Accepted July 01, 2015

ABSTRACT

As the use of RNA-seq has popularized, there is an increasing consciousness of the importance of experimental design, bias removal, accurate quantification and control of false positives for proper data analysis. We introduce the NOISeq R-package for quality control and analysis of count data. We show how the available diagnostic tools can be used to monitor quality issues, make pre-processing decisions and improve analysis. We demonstrate that the nonparametric NOISeqBIO efficiently controls false discoveries in experiments with biological replication and outperforms state-of-the-art methods. NOISeq is a comprehensive resource that meets current needs for robust data-aware analysis of RNA-seq differential expression.

INTRODUCTION

RNA-seq has rapidly become the technology of choice for high-throughput gene expression analysis. RNA-seq relies on cDNA sequencing as a way to determine the sequence characteristics of transcripts and to estimate the gene expression level. High-throughput sequencing makes the study of transcriptomes readily approachable, including alternative splicing, the discovery of novel splice junctions, the delimitation of UTR boundaries or the identification of antisense or extra-exonic expression (1,2). Additionally, RNA-seq technology can be applied either with the support of genome annotation to facilitate transcript identification or without a reference genome, making it a powerful tool for *de novo* transcriptome characterization. This versatility makes RNA-seq a potent and increasingly used technology for the global study of transcriptomes.

One of the most wide-spread applications of RNA-seq is the transcript quantification and the differential gene expression analysis (3,4). It has been claimed that RNA-seq has a number of advantages over its predecessors (arrays), such as a wider dynamic range of measurements (5), the capacity to detect transcripts with low expression level (3) and the ability to identify differences in isoform or allele expression (6,7). RNA-seq was initially described as highly reproducible, and it was claimed to provide more 'direct' and reliable gene expression measurements (3), but it is now generally accepted that it also has limitations which make it far from perfect. Although the high reproducibility of the technology reduces the need of technical replication (3,4), the precision at the low expression level is still limited (4,8) and, nonetheless, sufficient biological replicates are needed to adequately infer properties about the population (9,10). Therefore, the number of replicates and the sequencing depth at which one should sample remain important considerations when designing an RNA-seq experiment (11). Neither planning the RNA-seq experiment nor processing the data is straightforward. RNA-seq data might be biased because of the inaccuracies introduced at different stages of the protocol, from RNA isolation to library construction, or in the actual sequencing process (2). Technology biases, such as the transcript length (12), GC content (13), PCR artifacts, uneven transcript read coverage, contamination by off-target transcripts, or large differences in transcript distributions (14), are factors that interfere in the linear relationship between transcript abundance and the number of mapped reads at a gene locus. Normalization is therefore a substantial step in RNA-seq data processing and so different methods are available for addressing RNA-seq normalization based on different initial assumptions (13,15–17).

Finally, most existing methods for differential expression (DE) analysis make assumptions about the probability dis-

[*]To whom correspondence should be addressed. Tel: +34 963 289 680; Fax: +34 963 289 701; Email: aconesa@cipf.es

tribution of the data and only accept raw counts as input (18,19), but these assumptions might not be fulfilled or count data could have been transformed (e.g. to correct batch effects) or normalized. Moreover, it has been shown that control of the False Discovery Rate (FDR) is inadequate in most cases (20).

All these factors impact DE calls and the biological conclusions extracted from RNA-seq experiments (21). It is therefore absolutely necessary that RNA-seq data analysis follows a thorough procedure to evaluate data quality, detect biases and correct them when possible. Several approaches have been presented that address these issues (2,9) and attractive tools have been designed that deal with some of them (10,22,23). However, none of the existing solutions provide a comprehensive resource that supports RNA-seq procedures through the whole process of sequencing planning, quality control (QC) and DE analysis. With this purpose in mind, we developed the NOISeq R package, which is publicly available at the Bioconductor repository (24). The NOISeq R package integrates very useful tools for guiding users when planning sequencing experiments to quantify gene expression, assessing the quality of the expression data obtained, choosing appropriate normalization or filtering methods according to the biases detected, performing DE analysis and visualizing the results, among other functionalities. The package also includes two robust non-parametric approaches for DE analysis: NOISeq and NOISeqBIO. NOISeq (25) was published as a methodology to handle RNA-seq data with technical replicates or no replications. The method has been used in several studies (26–34) and benchmarked against the most popular DE methods with good results (20,35–37). In this work, we introduce NOISeqBIO method for biological replicates, which implements an empirical Bayes approach that improves the handling of biological variability specific to each gene, and is very successful in controlling the high FDR in experiments with biological replicates. Although parametric methods are said to have more power than non-parametric approaches, they may lead to unreliable results if the distributional assumptions do not hold. Consequently, the development of non-parametric DE tools for RNA-seq has increased in the last years and some examples are SAMseq (38), NPEBseq (36) or LFCseq (39), among others. The strategies followed by these methods are quite different. While SAMseq uses the rank of the expression values in a Wilcoxon statistic, NPEBseq applies an empirical Bayesian approach where the test statistic is based on expression fold-change, and LFCseq is inspired in the NOISeq method but uses only the fold-change as the test statistic instead of considering both the fold-change and the difference in expression.

This paper describes the QC and DE analysis pipeline implemented in the NOISeq package. We also present the statistical formulation of the new NOISeqBIO and compare this approach to popular RNA-seq DE methods (edgeR (18), DESeq2 (19) and SAMseq (38)) using both real and simulated datasets to demonstrate the efficiency of the method to control false calls in a wide variety of analysis scenarios.

MATERIALS AND METHODS

The NOISeq package

The NOISeq R package is a comprehensive resource for the analysis of RNA-seq data, which can be divided into three blocks: (i) count data QC, (ii) filtering of low-count features, normalization and batch effect correction and (iii) DE analysis (Figure 1). Within each block, the package offers both visualization plots and processing functions that help to perform a comprehensive diagnosis and analysis of count datasets. The package includes an option for easily generating a QC report pdf file containing all the plots described in this section (see NOISeq package in Bioconductor for an example).

Diagnostic and visualization plots. In the NOISeq package there are a total of 14 different analytical plots available to evaluate the quality of the data and the results of the DE analysis. Table 1 summarizes these graphical resources and indicates the type of plot and its main application. More details are given in the Results section and the Supplementary Material.

Low-count filtering. The estimation of gene expression is less reliable for low-count genes, which therefore represent a source of noise that negatively affects sensitivity and specificity in most DE analysis methods (20), hence the removal of low-count genes before further analysis is undertaken is recommended (40). We propose three different methods to filter out the low-count features which are implemented in the NOISeq package: counts per million (CPM), proportion test and Wilcoxon test. In contrast to other commonly used methods for low-count filtering, all these NOISeq methods take the experimental design into account and apply the filtering criterion to every experimental condition in the dataset, removing features that are below the threshold in all conditions. In the CPM method, given a low expression threshold *cpm*, in CPM, features with an average CPM per condition below this threshold in all experimental conditions are removed. It is also possible to set a cutoff for the coefficient of variation per condition to eliminate features with inconsistent expression values. For the proportion test, H_0: $p = p_0$ is tested against H_1: $p > p_0$ for each feature and condition, where p is the feature-relative expression, and $p_0 = cpm/10^6$. Features with a P-value $>\alpha$ in all conditions are filtered out. The Wilcoxon test is a similar procedure but it tests H_0: $m = 0$ versus H_1: $m > 0$ (m being the median expression per condition). No CPM threshold needs to be set in this case, but it should be applied when at least five replicates per condition are available. A more detailed description of these filtering methods, together with a comparative evaluation of their performance, can be found in the Supplementary Material.

Normalization. The NOISeq package includes three commonly used normalization methods: RPKM (4), Upper Quartile (15) and TMM (14). However, since NOISeq accepts previously normalized data, any other normalization procedure that the user requires can be applied.

Batch effect correction. The Principal Component Analysis (PCA) function in the NOISeq package allows for ex-

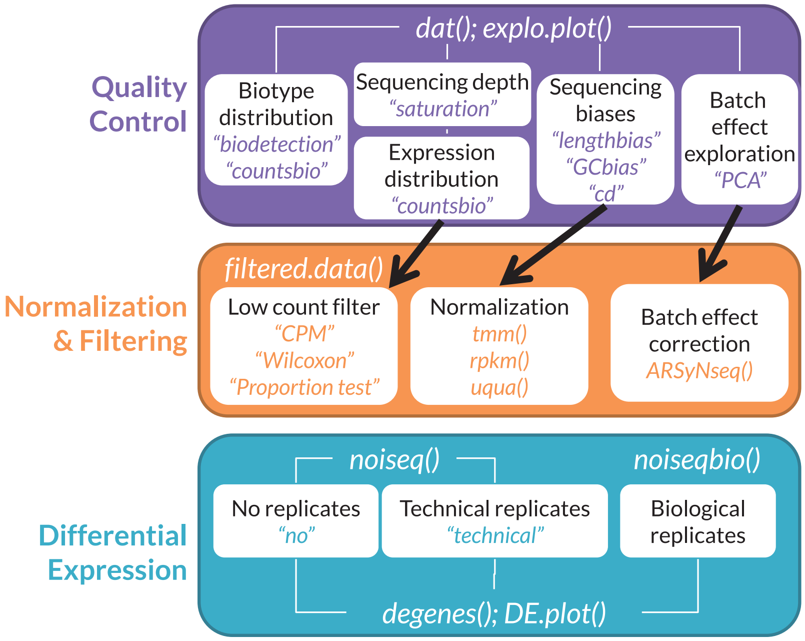

dat(); explo.plot()
Quality Control
Biotype distribution "biodetection" "countsbio"
Sequencing depth "saturation"
Expression distribution "countsbio"
Sequencing biases "lengthbias" "GCbias" "cd"
Batch effect exploration "PCA"
Normalization & Filtering
filtered.data()
Low count filter "CPM" "Wilcoxon" "Proportion test"
Normalization tmm() rpkm() uqua()
Batch effect correction ARSyNseq()
Differential Expression
noiseq()
noiseqbio()
No replicates "no"
Technical replicates "technical"
Biological replicates
degenes(); DE.plot()

Figure 1. Outline of the NOISeq package functionalities. Black arrows highlight that some QC plots are used to make data processing decisions. Terms in color and between quotation marks refer to arguments of the NOISeq package functions.

Table 1. Graphical tools included in the NOISeq package

Type of plot	R function	Application

Biotype detection	*dat(...,type=‘biodetection’);*
	explo.plot(...,plottype=‘persample’)
		Percentage of genes detected per biotype from their total representation in the genome in a given sample or condition

Biotype comparison	*dat(...,type=‘biodetection’);*
	*explo.plot(...,plottype=‘comparison’*)
		Biotype detection comparison for two samples or conditions

Biotype expression range	*dat(...,type=‘countsbio’);*
	explo.plot(...,plottype=‘boxplot’)
		Range of expression levels within each biotype in a given sample or condition

Saturation	*dat(...,type=‘saturation’); explo.plot(...)*	Number of detected genes at the given sequencing depth and at simulated higher and lower depths, and number of newly detected genes per million additional reads

Dynamic range of expression	*dat(..., type=‘countsbio’);*
	explo.plot(...,plottype=‘boxplot’)
		CPM distribution for all the samples in the experiment

Sensitivity	*dat(..., type=‘countsbio’);*
	explo.plot(...,plottype=‘barplot’)
		Percentage of genes with more than 0, 1, 2, 5 or 10 CPM per sample, and in any of the samples

Feature length	*dat(..., type=‘lengthbias’); explo.plot(...)*	Gene expression as a function of length
GC content	*dat(..., type=‘GCbias’); explo.plot(...)*	Gene expression as a function of GC content
RNA composition	*dat(..., type=‘cd’); explo.plot(...)*	Comparison of RNA composition (count distribution) in all samples
PCA plot	*dat(..., type=‘PCA’); explo.plot(...)*	Principal Component Analysis plot for either samples or genes
Expression	*DE.plot(...,graphic=‘expr’)*	Mean expression values for both conditions where DEGs are highlighted
(M,D)	*DE.plot(...,graphic=‘MD’)*	(M,D) values from the comparison of both conditions where DEGs are highlighted
Manhattan	*DE.plot(...,graphic=‘chrom’)*	Expression across chromosomal positions where up and down DEGs are highlighted
DEG distribution	*DE.plot(...,graphic=‘distr’)*	Distribution of DEGs per biotype and chromosome

ploring datasets and detecting possible batch effects. When a batch effect is present in the data, the package offers the possibility of removing it by applying an adaptation of ARSyN method (41) to RNA-seq data. Furthermore, this function can even remove systematic noise from unknown sources from the data when the batch information is not available. See NOISeq Bioconductor user's guide for a more detailed description.

DE. Two distribution-free DE methods are implemented in the package: NOISeq and NOISeqBIO. The NOISeq method (25) was developed to deal with datasets with only technical, or even no replicates. The next section introduces

NOISeqBIO, a novel approach which adapts the NOISeq method to handle biological variability.

NOISeqBIO

NOISeqBIO combines the non-parametric framework of NOISeq with an empirical Bayes approach inspired by the work of Efron *et al.* (42). This method assumes that genes can be classified into two different populations: genes with invariant expression between two experimental conditions and genes whose expression changes between conditions. In NOISeqBIO, a statistic Z is defined to evaluate this change in expression and the probability distribution of Z can be described as a mixture of two distributions: one for genes changing between conditions and the other for invariant genes. Thus, the mixture distribution f can be written as: $f(z) = p_0 f_0(z) + p_1 f_1(z)$, where p_0 is the probability of a gene having the same expression level in both conditions, i.e. the ratio of non-differentially expressed genes (non-DEGs), and $p_1 = 1 - p_0$ is the probability of a gene being differentially expressed between conditions, i.e. the DEG ratio. f_0 and f_1 are, respectively, the densities of Z for non-DEGs and DEGs. If one of either distribution can be estimated, the probability of a gene belonging to one of the two groups can be calculated. The algorithm consists of the following three steps:

(i) Computing DE statistic Z

The DE statistics used in NOISeq, the log-ratio of average expression values for the two conditions ($M_s = log_2(\bar{x}_1/\bar{x}_2)$) and the difference ($D_s = \bar{x}_1 - \bar{x}_2$) are also used in NOISeqBIO. As in NOISeq, the 0's in expression data are replaced by a small value higher than 0 to avoid indeterminations when computing the statistics. In NOISeqBIO, M and D are corrected by the biological variability:

$M_s^* = \dfrac{M_s}{a_0 + \hat{\sigma}_M}$ and $D_s^* = \dfrac{D_s}{a_0 + \hat{\sigma}_D}$, where $\hat{\sigma}_M$ and $\hat{\sigma}_D$ are the standard errors of M_s and D_s, respectively, and are computed as follows:

$\hat{\sigma}_M^2 = Var(log_2(\bar{x}_1/\bar{x}_2)) = Var(log_2(\bar{x}_1) - log_2(\bar{x}_2)) = Var(log_2(\bar{x}_1)) + Var(log_2(\bar{x}_2))$, assuming that $\bar{x}_1$ and $\bar{x}_2$ are independent. We used the delta method (i.e. a Taylor series approximation) to estimate $Var(log_2(X)) \approx \left(\dfrac{1}{E(X)log(2)}\right)^2 Var(X)$. For each condition i, we estimated $E(\bar{x}_i) = \bar{x}_i$ and $Var(\bar{x}_i) = S_i^2/n_i$. Hence, $\hat{\sigma}_M^2 \approx \dfrac{1}{\bar{x}_1^2 log(2)^2}\dfrac{S_1^2}{n_1} + \dfrac{1}{\bar{x}_2^2 log(2)^2}\dfrac{S_2^2}{n_2}$.

$$\hat{\sigma}_D^2 = Var(\bar{x}_1 - \bar{x}_2) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

a_0 is computed as a given percentile of all the values in $\hat{\sigma}_M$ or, respectively, as in (42), where the authors suggest taking the 90th percentile. Finally, several combinations of M^* and D^* statistics were tested to define the Z statistic (results not shown), and $Z = \dfrac{M^* + D^*}{2}$ was selected as the best one.

(ii) Estimating null scores Z_0

Let $\mathbf{X}_i$ be the gene expression matrix for each experimental condition i ($i = 1, 2$) of dimensions $G \times N_i$, where G is the number of genes and N_i is the number of biological replicates in condition i. We assume that matrices $\mathbf{X}_i$ have been previously normalized and that non-expressed genes have been removed according to the filtering criteria defined by the user (see low-count filtering section for some proposals on low-count filtering).

To estimate the Z-scores of genes with no change between conditions (Z_0), we permute the labels of samples between $\mathbf{X}_1$ and $\mathbf{X}_2$ r times, and compute Z statistic as above. A matrix with r columns and G rows is obtained and Z_0 is generated by pooling all its values.

When fewer than five replicates are available per condition, this null distribution is poor as the number of possible permutations is low. In these cases we borrow information from across similar genes. Genes are grouped according to their expression values across replicates by k-means clustering. For each cluster k, we consider the expression values of all the g_k genes in the cluster as observations within the corresponding condition and then shuffle this submatrix $r \times g_k$ times. For each permutation, we calculate Z_0. When $g_k \geq 1000$, the cluster is subdivided again into subclusters.

(iii) Obtaining the probability of DE

Given a gene with a score z, the posterior probability of DE $p_1(z)$ can be derived from the Bayes rule as: $p_1(z) = \frac{p_1 f_1(z)}{f(z)} = 1 - p_0 \frac{f_0(z)}{f(z)}$

A Kernel Density Estimator with a Gaussian kernel is used to approximate $f(z)$ and $f_0(z)$. For p_0, we take an upper bound, as previously suggested (42) to avoid negative p_1 values: $p_0 \leq \min_Z\{f(Z)/f_0(Z)\}$.

According to (42), the FDR defined by Benjamini and Hochberg is closely connected to the *a posteriori* probability $p_0(z) = 1 - p_1(z)$ that we are calculating. Hence, $p_1(z) = 1 - FDR$ and so $1 - p_1(z)$ can be used as an adjusted P-value. Note that this is an important difference with regard to NOISeq method, which was designed for technical replication and therefore the DE probability returned by NOISeq could not be considered to be equivalent to a P-value.

Data

Experimental data. Three experimental RNA-seq datasets were used to illustrate the use of the NOISeq diagnostic plots while the performance of the methodology was evaluated in the last two. The count data matrices for the three of them are available from NOISeq website (http://bioinfo.cipf.es/noiseq/doku.php).

ENCODE dataset. RNA-seq data from human B-cells ($CD20+$ cell line) and monocytes ($CD14+$ cell line) were obtained by Cold Spring Harbor Laboratory for the ENCODE project (43). Two different RNA extracting protocols were applied: the PolyA+ extraction method (Pap) and PolyA- selection procedure (Pam). Sequencing was performed with an Illumina GAIIx platform. The read files were downloaded from ENCODE website (see Section 1 in Supplementary Material) and mapped to the reference genome downloaded from UCSC (hg19 GRCh37) (44) using TopHat v2.0.8 (45). Gene expression was quantified using the HTSeq Python package version 0.5.3p3 (46) and an in-house script to take multihits into account by equitably dividing each read mapping to different genes among all of them.

Fusarium oxysporum ***dataset.*** *Fusarium oxysporum* f.sp.*lycopersici* race 2 isolate 4287 was used in this experiment. Freshly obtained microconidia were germinated for 16 h at 28°C in minimal medium (MM) (47) with 25 mM sodium glutamate and 20 mM HEPES buffer, at pH 7.4. The mixture was then moved to 37°C for 4 h, and then transferred to fresh MM or to heparinized human whole blood (Dunn Labortechnik GmbH) for 30 min at 37°C. Mycelia were recovered, flash frozen and used for RNA extraction as previously described (48). The poly(A) RNA fraction was enriched using the MicroPoly(A)Purist kit (Ambion, Darmstadt, Germany) and fragment libraries were prepared using the SOLiD Total RNA-Seq Kit (Ambion). Approximately 700 million library beads were loaded onto one full slide and sequenced to a level of 50 bases using the Applied Biosystems SOLiD 4 system with SOLiD MM50 chemistry. We obtained two biological replicates for both blood (*wt_B_30_37*) and MM (*wt_M_30_37*) conditions and mapped them onto the reference genome from the Ensembl Fungi database (49) (release 14) using Lifescope software. CLC Bio tools were used to quantify the gene expression.

Human prostate cancer dataset. This RNA-seq dataset was directly obtained from the Sequence Read Archive (SRA) repository (ERP000550). In this study Ren *et al.* sequenced samples of tumoral and healthy prostate which came from Chinese patients (50). There were 11 biological replicates for tumoral prostate (*T*) and 12 replicates for healthy prostate (*N*). The sequencing was done with an Illumina HiSeqTM 2000 and the reads were mapped to the reference human genome downloaded from Ensembl (49) (release 68) using TopHat 1.4.1 (51). Gene expression was quantified using the HTSeq Python package, version 0.5.3p3 (46).

Simulated data. To better evaluate the performance of NOISeqBIO we designed a simulation algorithm for synthetic datasets that mimics the sample structure and values of real data, while controlling the level of noise and the magnitude of gene expression changes (see Supplementary Material for a detailed description). Using this algorithm we generated 10 different datasets for each of the following parameter combinations:

- **Dataset size**: The data were simulated from the *F. oxysporum* and *Prostate Cancer* experimental samples after removing those genes with 0 counts in all samples. Hence, the resulting simulated data contained either 16 235 genes if simulated from *F. oxysporum* samples or 41 365 if simulated from *Prostate Cancer* samples.
- **Noise**: We considered no noise (0) and 20% noise (0.2), as described in the simulation algorithm.
- **Replicates**: We simulated data with a low number of replications (two or three replicates), and data with a larger number of replicates: 5 or 10.
- **DEG**: The proportion of DEGs was set to 5% or 10%.

In addition, we estimated the biological variability from several experimental datasets and defined two scenarios: high and low biological variability. We obtained a total of 320 simulated datasets for each one of the scenarios. The resulting simulated fold-change between the average expression of both conditions after normalizing data varied from 1.3 to 1400 with a median value of 40.

Data analysis. To assess the performance of NOISeqBIO (2.6.0) on the datasets described above, we compared it to the most widely used DE methods for RNA-seq: edgeR 3.4.2 (18) and DESeq2 1.2.10 (19). These two parametric methods assume the data follow a negative binomial distribution. While edgeR performs an exact test in the case of pair-wise comparisons, DESeq2 applies a Wald test on the estimated coefficients of a generalized linear model. Both methods also differ in the way they estimate variability. Since NOISeqBIO is a non-parametric method, we also included a non-parametric method in the comparison: SAMseq (38), which has been reported to perform well (20). SAMseq (from R package samr 2.0) uses a Wilcoxon rank statistic with resampling to account for the different sequencing depths and to estimate the FDR. Data were filtered using the CPM method included in the NOISeq package in order to reduce noise. The CPM value per sample was set to one in order to be conservative and not exclude too many genes. Normalization factors were computed within each DE method. The TMM method (14) was used for NOISeqBIO and edgeR. For DESeq and SAMseq, their own normalization procedure was applied.

The R code used in this work to simulate RNA-seq data, compute DE and analyze DE results for all the methods is available from NOISeq website (http://bioinfo.cipf.es/noiseq/doku.php).

RESULTS

QC of gene expression data

RNA extraction, library preparation and sequencing protocols affect the characteristics of the RNA-seq sample and can introduce different types of errors. The diagnostic plots in NOISeq package help to detect these biases and are focused on three different aspects of the data: the composition of RNA biotypes, the detection of transcripts and quantification of gene expression, and the sequencing biases.

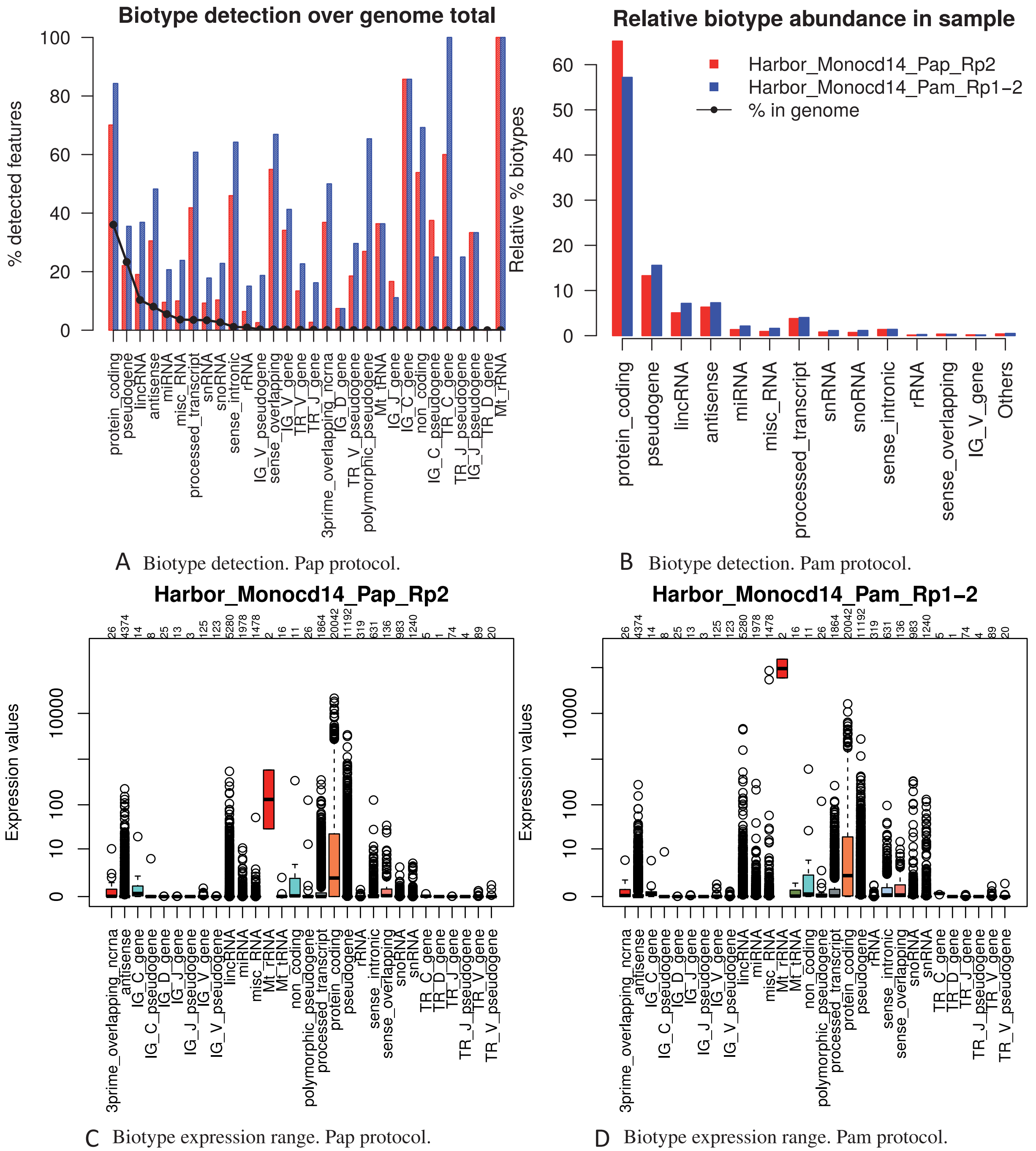
Biotype distribution. RNA-seq experiments may follow different RNA purification protocols to select specific target RNA species (i.e. long mRNAs or microRNAs) or different library preparations (e.g. stranded or non-stranded). These experimental procedures may result in RNA-seq data having a non-uniform RNA composition that may not be directly comparable (25). This can be relevant, for example, when trying to combine data from different sources or when the technical variability of specific protocols is high. The NOISeq package contains diagnostic plots that analyze the count distribution across RNA biotypes. The ‘Biotype detection’ and ‘Biotype comparison’ plots show the distribution of mapped reads among different RNA biotypes, indicating the proportion of genes detected for each biotype from their total annotated representation in the genome (Figures 2A, B, Supplementary Figures S1, S2, S3A and B, and Supplementary Section 2.1.3). The ‘Biotype expression range’ boxplots show the range of expression levels in CPM reads within each biotype (Figures 2C, D, Supplementary Figure S3C and D). Different biotype distributions across samples may indicate contamination, a technical problem or expected differences among samples.

To illustrate the utility of these diagnostic plots we used them to compare RNA-seq samples generated with different purification protocols (ENCODE data). By looking at the ‘Biotype comparison’ plots for Pap and Pam protocols, some differences are readily evident: the Pap protocol identifies a significantly higher relative proportion of protein-coding genes (>60%) than the Pam protocol (≈55%), as shown in Supplementary Section 2.1.3. As a consequence, the second protocol provides a relatively higher level of other RNA species such as pseudogenes, lincRNA or antisense transcripts (Figure 2B, and Supplementary Figures S1, S2, S3B). Differences in the relative percentages of biotypes detected also impact the quantification of the different RNA species, as revealed by the ‘Biotype expression range’ plots (Figure 2C, D, Supplementary Figures S3C and D). The Pap protocol results in a wider dynamic range of expression for protein-coding genes than the Pam protocol. However, the two Mt_rRNAs detected accumulate a huge number of reads when using Pam protocol (around 267 000 and 64 000 CPM, respectively) when compared to the Pap protocol (around 1, 100 and 230 CPM, respectively). These differences in the transcript quantification may affect the DE analysis. To prove this, we used NOISeq to select DEGs between B-cell and monocyte cell lines for each experimental protocol. In total, 15 346 and 14 357 DEGs were called in data obtained with the Pap or Pam procedures, respectively. Only 8929 genes (around 60%) were common to both extraction methods (Supplementary Figure S4), and the differences mostly affected protein-coding genes, pseudogenes, antisense and lincRNAs, which were the most abundant and differentially enriched biotypes between the Pap and Pam protocols (Supplementary Figure S5). These results highlight the importance of the RNA sample composition in RNA-seq analysis and show how the biotype break-down plots included in the NOISeq package can be used to reveal these characteristics in the data. The actions required after detecting an abnormality in these QC plots depend on the magnitude of the detected problem and on the goal of the study. Options range from removal of outlier samples, restricting analysis to well-quantified biotypes, or choosing the most adequate library preparation procedure. Unless studying biotypes differences is a goal, we recommend using samples with homogeneous biotype distributions.

Sequencing depth and quantification of expression. A key issue when analyzing RNA-seq data is to determine whether the available sequencing depth provides sufficient coverage of expressed transcripts and accurate gene expression quantification. It is generally accepted that genes detected by only a few reads are not reliably quantified and should be removed before further statistical analysis. Some of the QC plots in the NOISeq package are specifically targeted at answering these questions. We use the Prostate Cancer dataset to illustrate their utility. The ‘Saturation’ plot (Figure 3A, B, and Supplementary Figure S6) indicates the number of detected genes (left axis) at the given sequencing depth (solid dot), and also at simulated higher and lower numbers of reads. The bars (right axis) show the new discovery rate (NDR), i.e., the number of newly detected genes per million additional reads (25). If more reads do not lead to a higher number of new detections, then saturation is considered to have been reached and any additional sequencing will mostly improve the quantification of the previously detected genes. In the Prostate Cancer data we observed that around 50% of the annotated genes are found at the nominal sequencing depth of between 20 and 25 million reads (Supplementary Figure S6). The ‘Saturation’ plot estimates that in this range of total reads, around 250 additional genes are detected per additional million reads. This implies that increasing the sequencing depth by 10 million reads will increase transcriptome coverage by 10%. However, analyzing this information for each biotype (Figure 3A and B) shows more relevant results. For protein-coding genes, the NDR was 40, so the improvement in feature detection for 10 million additional reads is estimated at around 2%. In contrast, the NDR for lncRNAs stayed at 35 which translates into an estimated 25% more lncRNAs at a 10 million sequencing depth increase. Depending on the goal and scientific questions of the study, decisions on the need for additional reads may change. If only protein-coding genes are to be analyzed in the study, sequencing depth might be sufficient, while this might not be the case if the target are also lncRNAs.

The ‘Dynamic range of expression’ plot (Figure 3C) compares the distribution of read CPM for all the samples in the experiment and is useful for identifying differences in count distributions within the dataset. In the Prostate Cancer data, we observed that the distribution of expression levels for detected genes varies considerably among samples, and suggests that a normalization approach that corrects for these differences would be needed to make the samples comparable. This plot could also be used to reject samples with odd expression level distributions. For example, one could consider removing sample *N_10* from the analysis for having too low median expression levels. The analysis of expression quantification is complemented by the ‘Sensitivity’ plot (Figure 3D), which displays the number of genes with more than 0, 1, 2, 5 or 10 CPMs for each sample (bars) or in any of the samples (horizontal lines). This plot reflects the total fraction low-expression genes represent within the total number of transcripts. In the Prostate Cancer example, less than 35% of the genes have more than one CPM in any of the samples. Therefore it provides a graphical representation that helps the user to make decisions on what CPM threshold should be used, as it shows the percentage of features that would be removed at different CPM values.

To illustrate how low-count diagnosis and filtering affects DE analysis, we used one of the filtering options provided by NOISeq package: the CPM method (see Materials and Methods and Supplementary Section 4). We chose a CPM threshold of one to remove low-count genes, and



A Biotype detection. Pap protocol.

B Biotype detection. Pam protocol.

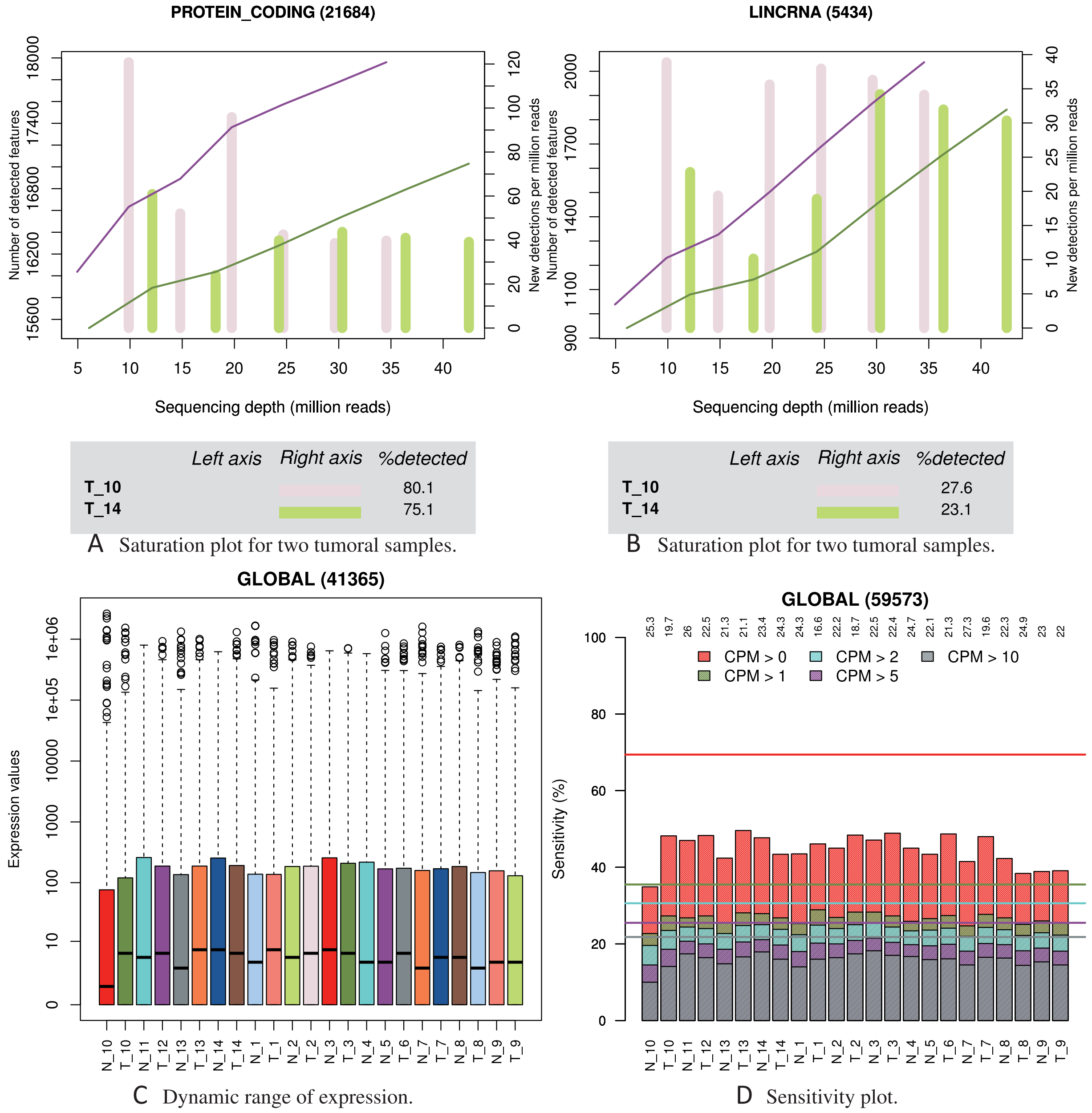
C Biotype expression range. Pap protocol.

D Biotype expression range. Pam protocol.

Figure 2. Biotype distribution. Data: *Monocytes (CD14-positive cells from human leukapheresis production) from ENCODE project*. (**A**) shows the percentage of genes in the genome detected (with at least 1 read) in our sample per biotype. Red and blue bars correspond to the two samples compared as indicate in legend of (B). Black line indicates the abundance of each biotype within the genome. (**B**) displays the abundance of each biotype within the genes detected in each of the two samples. (**C, D**) Expression values (Y axis) are given in CPM of sequencing reads. Numbers in the upper part of the plot are the number of genes per biotype that are represented in each boxplot.

computed the DEGs using NOISeqBIO and edgeR (18). Figure 4 shows the results of this filtering approach. A total of 42 366 low-count genes were removed after applying the CPM threshold, of which 292 and 887 had been detected as differentially expressed by NOISeqBIO and edgeR, respectively. In turn, removing these low-count genes resulted in 683 (NOISeqBIO) and 1195 (edgeR) newly detected genes which belonged to a higher expression range. These results highlight the impact of low-count filtering in RNA-seq DE analysis and how the NOISeq package resources can be used to easily address this task.

Sequencing biases. Finally, when sequencing artifacts are present in the data, the quantification of gene expression could be biased and lead to the wrong biological conclu-



A Saturation plot for two tumoral samples.

B Saturation plot for two tumoral samples.

C Dynamic range of expression.

D Sensitivity plot.

Figure 3. Sequencing depth and expression quantification. Data: Prostate cancer. (**A,B**) show the number of detected genes (Y left axis) and the new detections per each million of additional reads sequenced (Y right axis) at increasing sequencing depths for two different samples and for 'protein-coding' genes and 'lincRNAs', respectively. (**C**) shows the distribution of expression values in CPM for each sample considering genes with more than 0 counts in any of the samples. (**D**) summarizes the proportion of genes with more than 0, 1, 2, 5 or 10 CPM in each sample (bars) or in any of the samples (horizontal lines).

sions. Proper and timely detection of these biases is needed to choose an appropriate normalization procedure that corrects data errors and improves downstream statistical analyses. NOISeq implements diagnostic plots for three of the most frequently cited sequencing biases in RNA-seq data, namely 'feature length' (12), 'GC content' (13) and 'RNA composition' (14). To describe them we used the *F. oxysporum* dataset.

The 'feature length' and 'GC content' plots (Figure 5A–D) display gene expression as a function of length or GC content, taking bins of 200 genes. To assess the relationship between the length or GC content and the average gene expression, a cubic spline regression model was fitted (red and blue line for length and GC content plots, respectively). A model P-value lower than the significance level (e.g. 0.05) and R^2 higher than 70% is considered to indicate a significant length or GC content effect on the expression level. Both types of bias were evident in the *F. oxysporum* data (Figure 5A and C). The 'RNA composition' plot (Figure 5E and F) indicates if significant differences in the RNA sample composition are present. In these plots, M values are computed between each sample s and a reference sam-

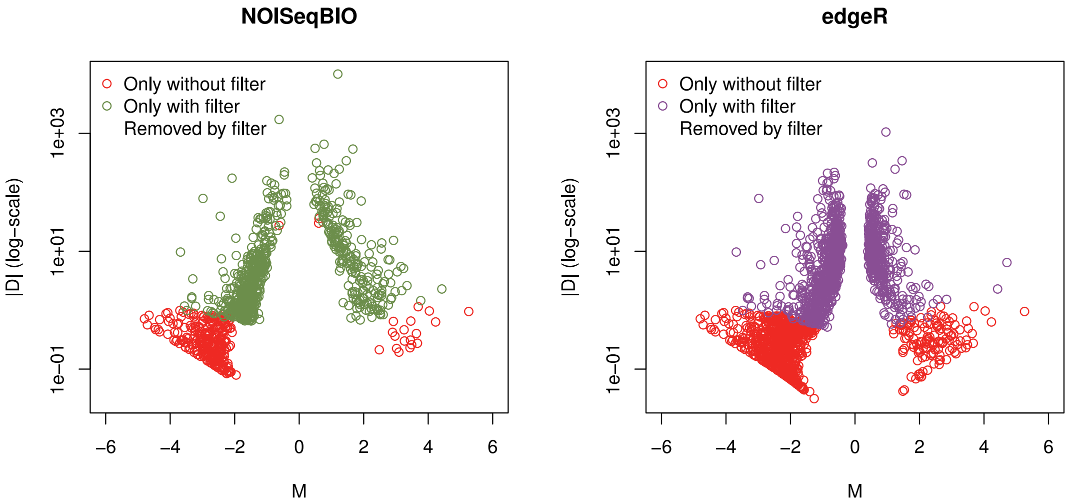


Figure 4. (M, D) plots for DEGs from NOISeqBIO (left plot) and edgeR (right plot) on Prostate Cancer data. X axis represents $M = log_2(\bar{x}_{\text{healty}}/\bar{x}_{\text{tumoral}})$ values, while the absolute value of $D = \bar{x}_{\text{healty}} - \bar{x}_{\text{tumoral}}$ is depicted on the Y axis. Red dots correspond to DEGs which were only obtained when no filter was applied (CPM method). Black dots correspond to genes removed by the filtering method. Green and purple dots correspond to the DEGs obtained with NOISeqBIO and edge R methods, respectively, only when the low-count filter was applied.

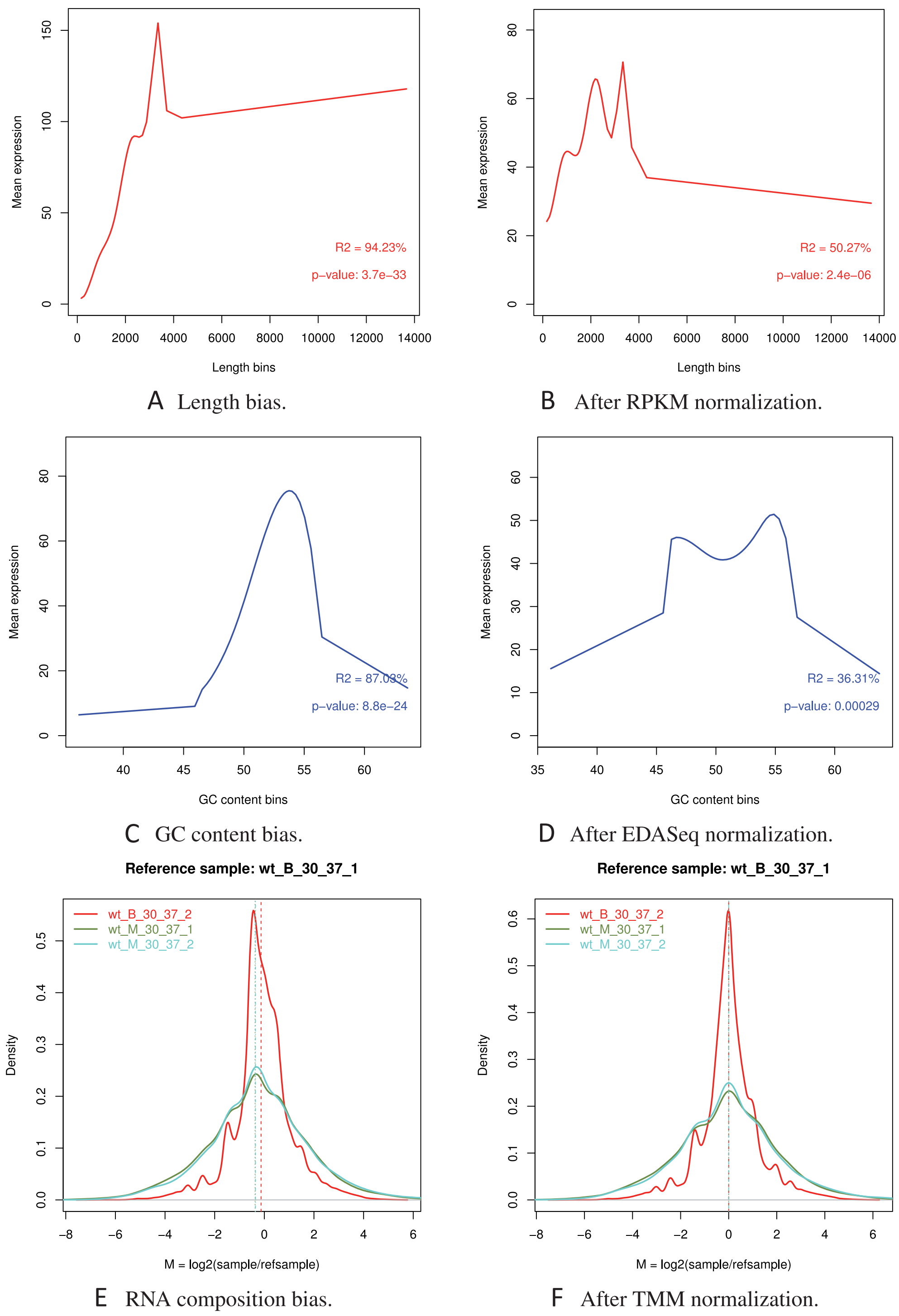
ple r (which can be arbitrarily chosen) as $M = log_2(x_s/x_r)$, where x_s are the counts in sample s. If no bias is present, the median of M values for each comparison is expected to be 0 (14). Deviations from this value indicate that the expression levels for a fraction of genes in one sample tend to be higher than in the others, and therefore that the data violate the assumption of uniform global RNA distributions which is frequently made in genome-wide gene expression experiments. Figure 5E shows a deviation from 0 in the M medians for the *F. oxysporum* data. Confidence intervals for the M median (Supplementary Section 2.3.1) are also computed and showed that this deviation was statistically significant in this case. These diagnostic plots give clues about the specific normalization procedures that are required for removing the observed biases. Table in Section 3 of Supplementary Material displays some of the normalization methods or R packages available for correcting each bias. Figure 5B, D and F show the three diagnostic plots after applying different types of normalization procedures designed to target each specific bias: RPKM (4) (included in the NOISeq package), ‘full’ within-sample normalization in the EDASeq package (13) and TMM (14) which is also included in the NOISeq package. We can see that, in all three cases, the biased pattern was significantly decreased after normalization.

DE

In this section we evaluate the performance of NOISeqBIO with the Prostate Cancer and *F. oxysporum* studies, that represent analysis scenarios with different replication levels, biological variability, and number of genes. We compare our method to the widely used edgeR (18) and DESeq2 (19), and also to the non-parametric SAMseq approach (38).

Results on experimental datasets. We computed DE in the experimental data, taking an adjusted P-value cutoff of 0.05, or equivalently, a probability cutoff of 0.95 for NOISeqBIO. Results are summarized in Figure 6 and Supplementary Figures S10–S13. In both datasets, we observed a significant difference in the percentage of DEGs called by NOISeqBIO in comparison to the edgeR and DESeq2 parametric approaches, and the non-parametric SAMseq method, which gave the highest rate of DEGs. NOISeqBIO found more DEGs (31.5%) than the parametric approaches (around 25%) in the *F. oxysporum* dataset, which has a comparatively low coefficient of variance and replicate number (Figure 6). The opposite result was obtained with the more variable Prostate Cancer dataset, where NOISeqBIO called 3% DEGs and edgeR and DESeq2 called around 7% DEGs. When analyzing the Prostate Cancer expression plots (Supplementary Figure S11) we noticed that the DEGs selected by NOISeqBIO had larger expression fold-changes than those selected by the parametric methods, while our method was comparably more sensitive to narrower expression changes in the *F. oxysporum* dataset (Supplementary Figure S10). In both cases SAMseq gave more permissive fold-change thresholds. Gene ranking was similar between edgeR, DESeq2 and NOISeqBIO, with Spearman’s correlation coefficient for FDR values between 0.95 and 0.98 (Supplementary Figures S12 and S13) indicating, that regardless of the significance thresholds, all methods similarly captured DE. A functional enrichment analysis was performed by comparing the DEG called by each method to the rest of the genome using GOseq (21). Interestingly, enrichment results were equivalent across methods, suggesting no major biological differences in the DEG sets detected by the tested algorithms (not shown).

We hypothesize that this different behavior in gene selection was due to the way the methods handle variability and replication: while the parametric methods tend to render more significant calls in highly replicated but variable data, NOISeqBIO more strongly penalizes values with a high coefficient of variation. On the contrary, when data variability is lower, NOISeqBIO might be more effective in calling DEGs. To verify this hypothesis and to further characterize the performance of NOISeqBIO, we compared the DE



A Length bias.

B After RPKM normalization.

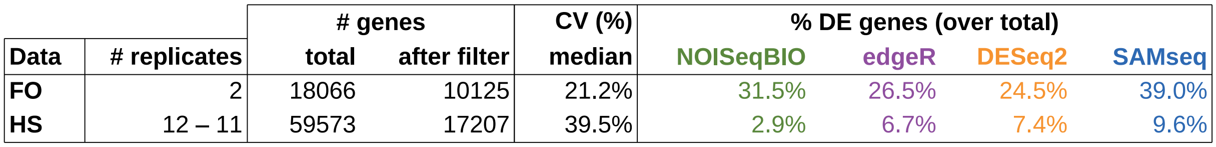
C GC content bias.

D After EDASeq normalization.

E RNA composition bias.

F After TMM normalization.

Figure 5. Sequencing biases. Data: *F. oxysporum*. (**A,B**) show the influence of the gene length on the gene expression before and after RPKM normalization. Each dot represents a bin of 200 genes and the red line was fitted with a spline regression model. (**C, D**) are analogous but for studying the influence of GC content on expression, before and after EDAseq normalization. (**E, F**) display the distribution of the log-ratio between each sample and the sample taken as reference (before and after TMM normalization), in order to check if the RNA composition differ among samples.



		# genes		CV (%)	% DE genes (over total)			
Data	**# replicates**	**total**	**after filter**	**median**	**NOISeqBIO**	**edgeR**	**DESeq2**	**SAMseq**
FO	2	18066	10125	21.2%	31.5%	26.5%	24.5%	39.0%
HS	12 – 11	59573	17207	39.5%	2.9%	6.7%	7.4%	9.6%

Figure 6. Characteristics of the *F. oxysporum* (FO) and Prostate Cancer (HS) datasets showing the number of replicates, number of genes, variability and percentage of DEGs called by each DE method.

methods on synthetic datasets where we mimicked the data structure of both experiments and introduced controlled levels of expression changes and noise.

Results on simulated datasets. Synthetic datasets were created using the simulation algorithm (described in the Supplementary Material) considering different numbers of genes, numbers of replicates per condition, levels of technical noise and proportions of DEGs. We also simulated two different biological variability scenarios: high (similar to the Prostate Cancer data) and low (similar to the *F. oxysporum* data) biological variability. A total of 320 datasets were obtained for each scenario according to different combinations of the simulation parameters. The method performance was evaluated on a fixed adjusted *P*-value cutoff (0.05), by looking at sensitivity (the proportion of true DE calls out of the total number of DEGs), the FDR (i.e. proportion of false DE calls out of the total number of DE calls) and the Matthews' Correlation Coefficient, a combined measure of all of the potential classification errors (also known as phi coefficient). Figures 7 and 8, Supplementary Figures S14 and S15 show the results of the simulation study for the high and low biological variability scenarios, respectively, as a function of the number of replicates, aggregating all technical noise, DEG proportions and gene-number scenarios. The significance of the differences between methods was estimated by an ANOVA model with repeated measures. Supplementary Figures S16 and S17 also show the percentage of DE calls for each method.

NOISeqBIO clearly outperformed the other nonparametric method SAMseq, which had serious problems both in obtaining good sensitivity at low replication numbers and in controlling the FDR when the number of replicates was high. These differences might be due to the very different strategies adopted by these two approaches: while SAMseq is based on permutations (which might break down with few replicates), NOISeqBIO uses the joint distribution of all genes in the dataset to estimate the null distribution and therefore may better capture the variability of the data needed to call significant changes.

Even more interesting results were obtained when comparing NOISeqBIO to the parametric methods. Sensitivity was shown to be high (90–100%) and relatively constant across all analysis scenarios and methods, although the two parametric approaches showed to have more statistical power, with the exception of the 10-replicate condition. Note here that the data were simulated using the Negative Binomial distribution, which is the probability distribution assumed by edgeR and DESeq2, so a better performance is expected for these two methods. In contrast, the FDR results significantly varied: at a low biological variability and replication level (two replicates), the median FDR was above the nominal 0.05 cutoff for all the methods, but the actual FDR values fluctuated considerably depending on the scenario. In these conditions DESeq2 had a significantly higher false positive rate while NOISeqBIO and edgeR showed comparable results (Figure 8 and Supplementary Figure S15). However, when biological variability was high (Figure 7 and Supplementary Figure S14) or when the number of replicates increased, the parametric approaches tended to call too many false positives, and only NOISeqBIO provided significantly better FDR control, and remained stable even through multiple simulated scenarios (Figures 7 and 8). These results are similar to those observed for the original NOISeq method when applied to technical replicates (25) and corroborate the results obtained with the real data: the high percentage DEGs called in the *F. oxysporum* dataset might reflect its low variability but also include false discoveries that might be more pronounced with the NOISeqBIO algorithm, while the lower DEG number provided by NOISeqBIO with the Prostate Cancer dataset might reflect the better FDR control of our method when used for high-variability data.

Similar performance patterns were observed when outliers were introduced in the simulations. NOISeqBIO maintained or improved the FDR respect to other methods in 3, 5 or 10 sample datasets, while the FDR was higher at two samples (Supplementary Figure S18). Finally, we tested the performance of the methods when no DEGs are present in the data (Supplementary Figure S19), and found that, in general, all the methods obtain a False Positive Rate (FPR) below the significance level of 5%.

DISCUSSION AND CONCLUSION

Although RNA-seq has become the technology of choice for genome-wide transcriptome profiling, there is growing awareness of the need to thoroughly examine the quality parameters imposed on both the raw and processed data to detect and eventually remove potential technology biases. Not surprisingly, recent results from the SEQC consortium on RNA-seq QC (52) highlighted quantification errors introduced by library preparation, GC content, gene coverage and read duplication as the key factors that affect the reproducibility of RNA-seq data across sites and technologies, and suggested that extensive QC should be a fundamental part of any RNA-seq analysis pipeline. Here we present the NOISeq package as an extensive resource for exploratory analysis, pre-processing and DE analysis of count data, which complements existing software tools for QC on raw (FastQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and mapped (RSeQC (53), RNA-SeQC

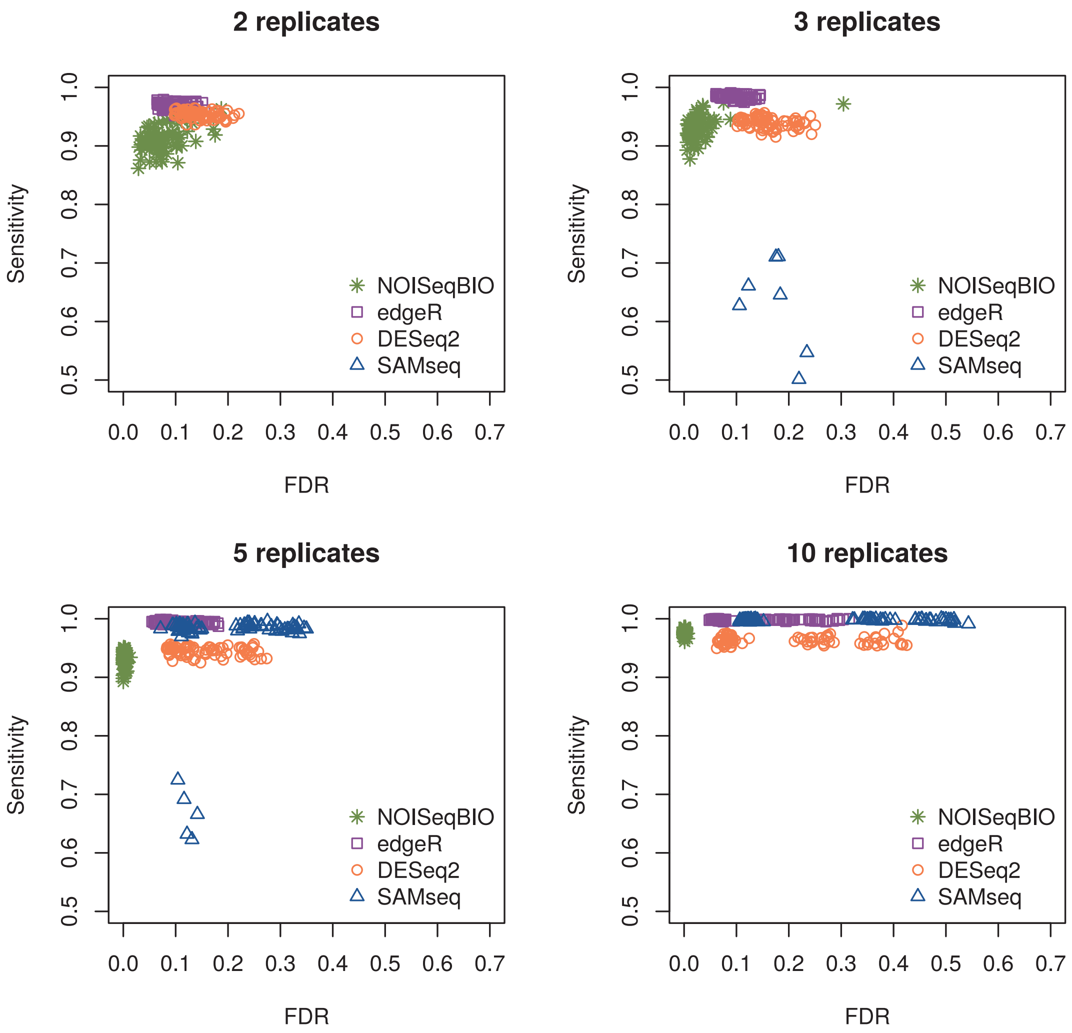


Figure 7. Trade-off between FDR and sensitivity for each DE method at a significance level of 5% in the high-variability scenario (320 simulations). Please note that for SAMseq there are results below the axis limits that are not displayed for the sake of clarity.

(54), Qualimap (55) or RNASeqGUI (23)) RNA-seq data. NOISeq provides a wide array of diagnostic and visualization plots which are relevant for understanding the characteristics of the data, biotype composition, technology biases and DE; moreover, it also contains tools to process the data accordingly, for example, by normalizing away biases or removing low-count genes. Some of these plots were advanced at the publication of the NOISeq method. In this work we extend QC diagnostic tools, show how to use these pre-processing resources and what their impact might be on downstream analysis.

In addition, the package also includes a statistical framework for RNA-seq DE analysis based on creating an empirical distribution, rather than relying on parametric assumptions, to assess differences in gene expression between conditions. We have shown that this approach works well in both NOISeq (for technical replicates) (25) and NOISeqBIO (for biological replicates, as described in this work) across different data scenarios and significantly reduces the false call problems which still remain present when using parametric methods. The main differences between NOISeqBIO and our previous method NOISeq are that (i) NOISeqBIO corrects the statistics for the biological variability specific of each gene, while NOISeq considered a global variability because it was conceived for technical replicates; and (ii) NOISeqBIO returns a DE probability that is equivalent to FDR adjusted *P*-values, but it is not comparable to the DE probability given by NOISeq.

Lastly, and importantly, one particularly interesting result from this study was obtained from the synthetic datasets that were simulated for different numbers of replicates. Our data suggest that duplicates might be insufficient to provide accurate RNA-seq DE results when using state-of-the art methodologies, including our own. The relevance of sufficient replication has also been brought to attention in recent work that evaluated experimental design considerations (20,56,57), a conclusion that was echoed by the SEQC project (58). We anticipate that, as awareness of RNA-seq experimental design issues increases and the technology becomes more affordable, experiments with higher replication levels will proliferate and therefore DE methods which efficiently deal with the FDR, while maintaining good sensitiv-

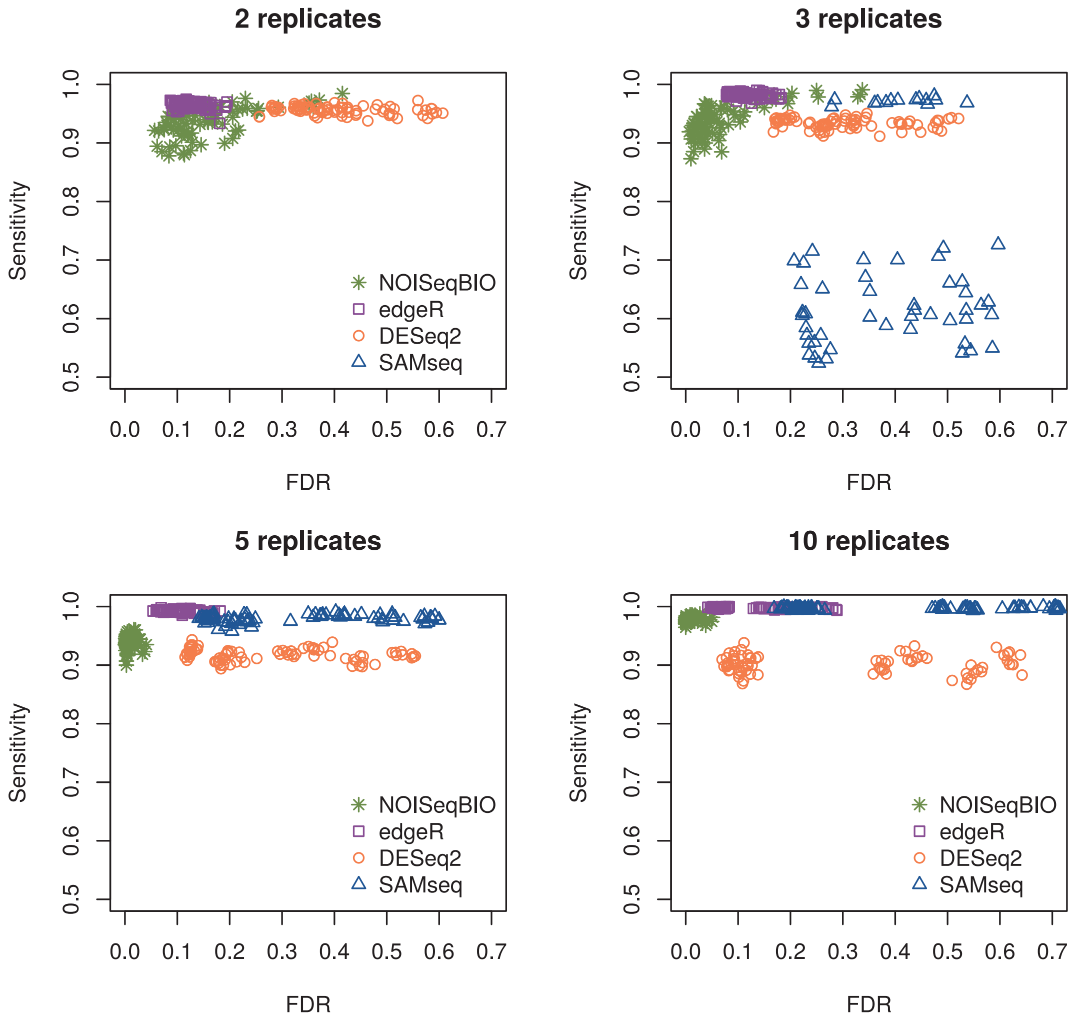


Figure 8. Trade-off between FDR and sensitivity for each DE method at a significance level of 5% in the LOW variability scenario (320 simulations). Please note that for SAMseq there are results below the axis limits that are not displayed for the sake of clarity.

ity, will increasingly be required. In this sense, NOISeqBIO perfectly fits these requirements. More sophisticated experimental designs than pair-wise comparisons are becoming available for RNA-seq data, and edgeR, DESeq2 or SAMseq can deal with this type of designs. However, it is still very common and necessary in these kind of studies the comparison of two groups, and NOISeqBIO provides an efficient solution. Moreover, our method has the advantage of being non-parametric, meaning that no distributional assumptions have to be made and no model validation is necessary. Therefore, it can be applied even when the data have been transformed in order to correct strong biases or batch effects, where parametric assumptions would not hold anymore.

Thus, taken together we believe that the NOISeq R package offers a suitable pipeline for RNA-seq robust analysis from expression quantification data to DE.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the University of Florida for supporting the publication charges of this manuscript.

FUNDING

European Union Seventh Framework Programme [FP7/2007-2013, 306000]; Spanish Ministry of Science and Innovation [MICINN, BIO2008-04638-E], in the framework of ERA-Net Pathogenomics; MICINN [DPI2008-06880-C03-03/DPI]. Funding for open access charge: University of Florida, Publication Funds.
Conflict of interest statement. None declared.

REFERENCES

1. Malone,J. and Oliver,B. (2011) Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.*, **9**, 34.
2. Robles,J.A., Qureshi,S.E., Stephen,S.J., Wilson,S.R., Burden,C.J. and Taylor,J.M. (2012) Efficient experimental design and analysis strategies for the detection of differential expression using RNA-sequencing. *BMC Genomics*, **13**, 484.

3. Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
4. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) *Nat. Methods*, **5**, 621–628.
5. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
6. Wagner,J.R., Ge,B., Pokholok,D., Gunderson,K.L., Pastinen,T. and Blanchette,M. (2010) Computational analysis of whole-genome differential allelic expression data in human. *PLoS Comput. Biol.*, **6**, e1000849.
7. Bell,G.D., Kane,N.C., Rieseberg,L.H. and Adams,K.L. (2013) RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome Biol. Evol.*, **5**, 1309–1323.
8. Labaj,P.P., Leparc,G.G., Linggi,B.E., Markillie,L.M., Wiley,H.S. and Kreil,D.P. (2011) Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics*, **27**, i383–i391.
9. Auer,P.L. and Doerge,R.W. (2010) Statistical design and analysis of RNA sequencing data. *Genetics*, **185**, 405–416.
10. Busby,M.A., Stewart,C., Miller,C.A., Grzeda,K.R. and Marth,G.T. (2013) Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics*, **29**, 656–657.
11. Cai,G., Li,H., Lu,Y., Huang,X., Lee,J., Müller,P., Ji,Y. and Liang,S. (2012) Accuracy of RNA-Seq and its dependence on sequencing depth. *BMC Bioinformatics*, **13**, S5.
12. Oshlack,A. and Wakefield,M. (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct*, **4**, 14–10.
13. Risso,D., Schwartz,K., Sherlock,G. and Dudoit,S. (2011) GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, **12**, 480.
14. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
15. Bullard,J., Purdom,E., Hansen,K. and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.
16. Dillies,M.-A., Rau,A., Aubert,J., Hennequet-Antier,C., Jeanmougin,M., Servant,N., Keime,C., Marot,G., Castel,D., Estelle,J., Guernec,G., Jagla,B., Jouneau,L., Laloë,D., Le Gall,C., Schaëffer,B., Le Crom,S., Guedj,M. and Jaffrézic,F. (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.*, **14**, 671–683.
17. Zheng,W., Chung,L.M. and Zhao,H. (2011) Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics*, **12**, 290.
18. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
19. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol.*, **15**, 550.
20. Soneson,C. and Delorenzi,M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, 91.
21. Young,M.D., Wakefield,M.J., Smyth,G.K. and Oshlack,A. (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.*, **11**, R14.
22. Bashir,A., Bansal,V. and Bafna,V. (2010) Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance. *BMC Genomics*, **11**, 385.
23. Russo,F. and Angelini,C. (2014) RNASeqGUI: a GUI for analysing RNA-Seq data. *Bioinformatics*, **30**, 2514.
24. Gentleman,R.C., Carey,V.J. and Bates,D.M. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
25. Tarazona,S., García-Alcalde,F., Dopazo,J., Ferrer,A. and Conesa,A. (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res.*, **21**, 2213–2223.
26. Ferreira,P.G., Patalano,S., Chauhan,R., Ffrench-Constant,R., Gabaldon,T., Guigo,R. and Sumner,S. (2013) Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes. *Genome Biol.*, **14**, R20.
27. Carcel-Trullols,J., Aguilar-Gallardo,C., Garcia-Alcalde,F., Pardo-Cea,M.A., Dopazo,J., Conesa,A. and Simón,C. (2012) Transdifferentiation of MALME-3M and MCF-7 cells toward adipocyte-like cells is dependent on clathrin-mediated endocytosis. *SpringerPlus*, **1**, 1–12.
28. Zhu,Q.-H., Stephen,S., Kazan,K., Jin,G., Fan,L., Taylor,J., Dennis,E.S., Helliwell,C.A. and Wang,M.-B. (2013) Characterization of the defense transcriptome responsive to Fusarium oxysporum-infection in Arabidopsis using RNA-seq. *Gene*, **512**, 259–266.
29. Shearman,J.R., Jantasuriyarat,C., Sangsrakru,D., Yoocha,T., Vannavichit,A., Tragoonrung,S. and Tangphatsornruang,S. (2013) Transcriptome analysis of normal and mantled developing oil palm flower and fruit. *Genomics*, **101**, 306–312.
30. Chen,G., Chen,J., Shi,C., Shi,L., Tong,W. and Shi,T. (2013) Dissecting the characteristics and dynamics of human protein complexes at transcriptome cascade using RNA-seq data. *PLOS ONE*, **8**, e66521.
31. Durban,J., Pérez,A., Sanz,L., Gómez,A., Bonilla,F., Chacón,D., Sasa,M., Angulo,Y., Gutiérrez,J.M. and Calvete,J.J. (2013) Integrated 'omics' profiling indicates that miRNAs are modulators of the ontogenetic venom composition shift in the Central American rattlesnake, Crotalus simus simus. *BMC Genomics*, **14**, 234.
32. Liu,W.-Y., Chang,Y.-M., Chen,S. C.-C., Lu,C.-H., Wu,Y.-H., Lu,M.-Y.J., Chen,D.-R., Shih,A. C.-C., Sheue,C.-R., Huang,H.-C. *et al.* (2013) Anatomical and transcriptional dynamics of maize embryonic leaves during seed germination. *Proc. Natl Acad. Sci. U.S.A.*, **110**, 3979–3984.
33. Su,L., Zhou,L., Liu,J., Cen,Z., Wu,C., Wang,T., Zhou,T., Chang,D., Guo,Y., Fang,X. *et al.* (2014) Phenotypic, genomic, transcriptomic and proteomic changes in Bacillus cereus after a short-term space flight. *Adv. Space Res.*, **53**, 18–29.
34. Xia,J.H., Liu,P., Liu,F., Lin,G., Sun,F., Tu,R. and Yue,G.H. (2013) Analysis of stress-responsive transcriptome in the intestine of Asian seabass (Lates calcarifer) using RNA-Seq. *DNA Res.*, **20**, 449–460.
35. Nookaew,I., Papini,M., Pornputtpong,N., Scalcinati,G., Fagerberg,L., Uhlén,M. and Nielsen,J. (2012) A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **40**, 10084–10097.
36. Bi,Y. and Davuluri,R.V. (2013) NPEBseq: nonparametric empirical bayesian-based procedure for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, 262.
37. Klambauer,G., Unterthiner,T. and Hochreiter,S. (2013) DEXUS: identifying differential expression in RNA-Seq studies with unknown conditions. *Nucleic Acids Res.*, **41**, e198.
38. Li,J. and Tibshirani,R. (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.*, **22**, 519–536.
39. Lin,B., Zhang,L.-F. and Chen,X. (2014) LFCseq: a nonparametric approach for differential expression analysis of RNA-seq data. *BMC Genomics*, **15**, S7.
40. Anders,S., McCarthy,D.J., Chen,Y., Okoniewski,M., Smyth,G.K., Huber,W. and Robinson,M.D. (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.*, **8**, 1765–1786.
41. Nueda,M., Ferrer,A. and Conesa,A. (2011) ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments. *Biostatistics*, **13**, 553–566.
42. Efron,B., Tibshirani,R., Storey,J.D. and Tusher,V. (2001) Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.
43. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
44. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.*, **22**, 1760–1774.
45. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

46. Anders,S., Pyl,P.T. and Huber,W. (2015) HTSeq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166.
47. Puhalla,J.E. (1968) Compatibility reactions on solid medium and interstrain inhibition in Ustilago maydis. *Genetics*, **60**, 461–474.
48. López-Berges,M.S., Capilla,J., Turrà,D., Schafferer,L., Matthijs,S., Jöchl,C., Cornelis,P., Guarro,J., Haas,H. and Di Pietro,A. (2012) HapX-mediated iron homeostasis is essential for rhizosphere competence and virulence of the soilborne pathogen Fusarium oxysporum. *Plant Cell Online*, **24**, 3805–3822.
49. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
50. Ren,S., Peng,Z., Mao,J.-H., Yu,Y., Yin,C., Gao,X., Cui,Z., Zhang,J., Yi,K., Xu,W. *et al.* (2012) RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res.*, **22**, 806–821.
51. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
52. Li,S., Labaj,P.P., Zumbo,P., Sykacek,P., Shi,W., Shi,L., Phan,J., Wu,P.-Y., Wang,M., Wang,C. *et al.* (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.*, **32**, 888–895.
53. Wang,L., Wang,S. and Li,W. (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, **28**, 2184–2185.
54. DeLuca,D.S., Levin,J.Z., Sivachenko,A., Fennell,T., Nazaire,M.-D., Williams,C., Reich,M., Winckler,W. and Getz,G. (2012) RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, **28**, 1530–1532.
55. García-Alcalde,F., Okonechnikov,K., Carbonell,J., Ruiz,L.M., Götz,S., Tarazona,S., Meyer,T.F. and Conesa,A. (2012) Qualimap: evaluating next generation sequencing alignment data. *Bioinformatics*, **28**, 2678–2679.
56. Liu,Y., Zhou,J. and White,K.P. (2014) RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, **30**, 301–304.
57. Rapaport,F., Khanin,R., Liang,Y., Pirun,M., Krek,A., Zumbo,P., Mason,C.E., Socci,N.D. and Betel,D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, **14**, R95.
58. SEQC/MAQC-III Consortium. (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat. Biotechnol.*, **32**, 903–914.