NOISeq: An R package for differential expression in RNA-Seq using biological replicates

Sonia Tarazona, Pedro Furió, Alberto Ferrer, Ana Conesa

March 27, 2013

Contents

Abstract

NOISeq method for evaluating differential expression in RNA-Seq data was presented in Tarazona *et al.* [1]. In that paper, NOISeq was compared to other methods on technical replicates. Now that the RNA-Seq technique has become more affordable, many experiments including biological replicates have been conducted. Hence, we improved the NOISeq method so it could handle the biological variability in a better way. In addition, we created an R package which is publicly available in Bioconductor. This package not only includes the differential expression method itself but also some useful functions to explore graphically the count data and the differential expression results, and also an algorithm to simulate RNA-seq data.

1 Introduction

One of the most common analyses in transcriptomics is the identification of genes whose expression changes between two or more experimental conditions (*differential expression*). Microarrays is a widespread technique for measuring the gene expression level of thousand of genes simultaneously. However, the recently emerged high throughput sequencing technologies such as RNA-Seq are taking the place of microarrays. With RNA-seq, no previous knowledge about the genome is required. Furthermore, besides the estimation of expression levels, it also allows for studying the transcriptome in more depth: alternative splicing, 3' UTR regions, novel splice junctions, antisense regulation, intragenic expression, etc. [2]. It has been reported as well that technical variability tends to be lower for RNA-Seq than for microarrays [3, 4, 5].

Moreover, the nature of the expression data is different for both platforms. While microarrays return a continuous measurement, RNA-seq produces counts of reads mapping to each gene. Therefore, the methods traditionally used to identify differentially expressed genes between conditions when working with microarray data are not appropriate for RNA-seq data and many *ad hoc* differential expression methods have emerged in the last years, such as: DESeq [6], edgeR [7], baySeq [8], DEGseq [9], SAMseq [10] or our method NOISeq [1]. In DESeq, edgeR and baySeq, the read counts across biological replicates are assumed to follow a negative binomial distribution. Both DESeq and edgeR apply a generalized linear model but they estimate the variability in a different way, while a bayesian approach is used by baySeq. DEGseq . On the contrary, SAMseq and NOISeq are non-parametric methods, so they make no assumptions on the data distribution. SAMseq is based on Wilcoxon's rank statistic and uses a permutation approach to correct the sequencing deph bias. NOISeq compares the change in expression between conditions to a noise distribution generated by comparing pairs of replicates within the same condition (see 2.1 for more details).

NOISeq has been successfully applied in several studies [11, 12, 13] and benchmarked against the most popular differential expression methods with good results [14, 15].

The NOISeq method was conceived in the RNA-seq infancy, when sequencing was still expensive and it was very difficult to find experiments with biological replicates. Thus, we developed a method to efficiently select the genes changing between two experimental conditions when having technical replicates or no replicates at all. Hence, there was still room for improvement in the way NOISeq handled the biogical variability specific to each gene. In this work, we introduce the new NOISeq Bioconductor R package. This package includes the NOISeq method, the new version of NOISeq (NOISeqBIO), some useful exploratory plots for expression data and a simulation algorithm to generate RNA-seq data with biological replicates.

In section 2, we first briefly describe the former NOISeq method in order to facilitate the understanding of the modifications introduced in NOISeqBIO and because it is included in the Bioconductor package. In this section, the new NOISeqBIO is also presented along with the other functionalities in the R package. We compared NOISeqBIO to NOISeq and also to some of the most widely used differential expression methods, on both synthetic and experimental data sets. The compared methods were parametric methods such as edgeR [7] or DESeq [6], and SAMseq [10], which is non-parametric. The results are shown in section 3.

2 Methods

2.1 NOISeq method

The basic idea underlying NOISeq is that a given feature may be considered differentially expressed if their change in expression between two experimental conditions is greater (or has a higher probability of being greater) than the change observed among replicates within the same condition.

Let x_{ij}^g be the expression of gene g in condition i ($i = 1, 2$) and replicate j. To measure the expression level change between two conditions, NOIseq takes into consideration two statistics: the log fold change (M) and the difference (D), that are calculated as $M_s^g = log_2(\bar{x}_1^g/\bar{x}_2^g)$ and $D_s^g = \bar{x}_1^g - \bar{x}_2^g$. The reason of using these two statistics is to get more reliable measurements of the change, since the fold change for features with low read counts can be missleading and the same occurs for the difference in expression between two conditions in the range of high counts.

Thus, in order to determine the probability of differential expression, the algorithm creates a so called "noise" distribution by pooling the (M_n, D_n) values computed among replicates within the same condition. When no replicates are available, NOISeq generates technical replicates from a multinomial distribution according to some parameters that can be specified by the user. However, we would like to remind that the conclusions obtained from technical replicates (simulated or not) cannot be extended to the whole population but only to the current experiment.

It can be derived an empirical cumulative distribution function $F(M, D)$ from the absolute value of these "noise" measurements (M_n, D_n). Given a certain gene with the corresponding (M_s, D_s) values ("signal") computed from the comparison of both experimental conditions, the probability of differential expression can be obtained as $F(|M_s|, |D_s|)$. The higher this probability, the higher the change in expression between conditions with regard to "noise" and the more we expect that change between conditions is due to the group effect and not to chance.

When using NOISeq, we recommended to use a threshold of 0.8 for this probability which is equivalent to an odds of 4:1, that means that the gene is four times more likely to be differentially expressed than nondifferentially expressed. However, the users often claimed for an equivalence between the differential expression probability returned by NOISeq and p-values. In this work, we addressed this issue and as it will be seen in section 2.2.

Regarding the normalization of counts, NOISeq had several possibilities that have been preserved in the new NOISeqBIO: taking the number of counts per million reads. the upper quartile method [16], the TMM method [17], or the RPKM values [3], if the length of the features is provided. No matter the normalization procedure selected, NOISeq permits applying a feature length correction that consists of dividing the expression level by a factor equal to any power of the feature length. NOISeq has also the advantage of accepting already normalized expression values instead of gene counts, so it can be used with any normalization procedure.

2.2 NOISeqBIO

NOISeqBIO was developed by joining the philosophy of our previous work together with the ideas from Efron *et al.* in [18]. These authors used an Empirical Bayes approach on microarray data in which they defined a statistic Z to evaluate the change in expression between two conditions. The probability distribution of this statistic 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 for a gene to have the same expression in both conditions and $p_1 = 1 - p_0$ is

the probability for a gene to have different expression between conditions. f_0 and f_1 are, respectively, the densities of Z for genes with no change in expression between conditions and for differentially expressed genes. If one of both distributions can be estimated, the probability of a gene to belong to one of the two groups can be calculated. We adapted this strategy to RNA-seq data by using some ideas already introduced in [1]. The algorithm consists of the following steps, which will be explained in more detail in this section:

1. Choosing an appropiate differential expression statistic Z.
2. Estimating the values of the Z statistic when there is no change in expression, i.e. the null statistic Z_0.
3. Estimating the probability density functions f and f_0.
4. Computing the probability of differential expression given the ratio f_0/f and an estimation $\hat{p}_0$ for p_0. If $Z = z$ for a given gene, this probability of differential expression can be computed as $p_1(z) = 1 - \hat{p}_0 f_0(z)/f(z)$.

1. **Differential expression statistic** Z

   The statistics used in NOISeq to compute differential expression were both the log-ratio of average expression values for the two conditions ($M_s = log_2(\bar{x}_1/\bar{x}_2)$) and the difference ($D_s = \bar{x}_1 - \bar{x}_2$). For this version of NOISeq, we corrected them by the biological variability. We defined, in this case, M_s^* and D_s^* as $M_s^* = \dfrac{M_s}{a_0 + \hat{\sigma}_M}$ and $D_s^* = \dfrac{D_s}{a_0 + \hat{\sigma}_D}$, where $\hat{\sigma}_M^2$ and $\hat{\sigma}_D^2$ are the standard errors of M_s and D_s statistics, respectively, and are computed as follows:

   - $\hat{\sigma}_M^2 = Var(log_2(\bar{x}_1/\bar{x}_2)) = Var(log_2(\bar{x}_1) - log_2(\bar{x}_2)) = Var(log_2(\bar{x}_1)) + Var(log_2(\bar{x}_2))$, if we assume that $\bar{x}_1$ and $\bar{x}_2$ are independent. We used second-order Taylor expansions (δ-method) to approximate the variance: $Var(log_2(X)) = \left(\dfrac{1}{E(X)log(2)}\right)^2 Var(X)$. For each condition i, we estimated $E(\bar{x}_i) = \bar{x}_i$ and $Var(\bar{x}_i) = S_i^2/n_i$. Hence, $\hat{\sigma}_M^2 \approx \dfrac{1}{\bar{x}_1^2 log(2)^2}\dfrac{S_1^2}{n_1} + \dfrac{1}{\bar{x}_2^2 log(2)^2}\dfrac{S_2^2}{n_2}$.
   - $\hat{\sigma}_D^2 = Var(\bar{x}_1 - \bar{x}_2) \approx \dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}$

   Thus, the corrected M and D values for signal are $M_s^* = M_s/\hat{\sigma}_M$ and $D_s^* = D_s/\hat{\sigma}_D$.

   a_0 is computed as a given percentile of all the values in $\hat{\sigma}_M$ or $\hat{\sigma}_D$, respectively, as in [18] (the authors suggest the percentile 90th as the best option). To compute the Z statistic, the M and D statistics are combined: $Z = \dfrac{M^* + D^*}{2}$.

2. **Null scores** Z_0

   Let $\mathbf{X}_i$ be the gene expression matrix for each experimental condition i ($i = 1, 2$). $\mathbf{X}_i$ has G rows (genes) and N_i columns (biological replicates for condition i). We assume that matrices $\mathbf{X}_i$ are normalized (e.g. by converting them to TMM [17] or RPKM [3] values) and that genes with no expression across all the replicates for both conditions are removed.

   In order to afterwards compute the null density f_0, we first need to estimate the values of the Z-scores for genes with no change between conditions. To do that, we permuted labels of samples between $\mathbf{X}_1$ and $\mathbf{X}_2$ and repeated the above procedure to obtain the matrix Z_0 with so many columns as the number of permutations. We pooled all the values in Z_0 to apply the next step.

3. **Estimation of densities**

   The next step is estimating the ratio f_0/f from the Z_0 and Z scores. In [18], they estimate f_0/f by nonparametric logistic regression but, in NOISeqBIO, we estimate separately f and f_0 using a kernel density estimator (KDE) with gaussian kernel because we observed a better performance of the method with this option.

4. **Probability of differential expression**

   Given a gene with a value z for the Z score, let $p_1(z)$ be the probability that that gene is differentially expressed between both experimental conditions. This probability can be derived from Bayes Rule as follows:

$$p_1(z) = \frac{p_1 f_1(z)}{f(z)} = 1 - p_0 \frac{f_0(z)}{f(z)} \tag{1}$$

Moreover, as Efron *et al.* show in their work [18], the FDR defined by Benjamini and Hochberg is closely connected to the *a posteriori* probability $p_0(z) = 1 - p_1(z)$ we are calculating.

Since we already estimated the ratio f_0/f in the previous section, we only need to estimate p_0. We took an upper bound of p_0 as it is suggested in [18]. Taking into account that $p_1(z)$ must be nonnegative leads to the following restriction $p_0 \leq \min_Z\{f(Z)/f_0(Z)\}$, which can be used as the estimate for p_0.

2.3 Functionalities in the R package

The Bioconductor R package includes the two differential expression methods: NOISeq and NOISeqBIO. The reason why we did not exclude the NOISeq method from the package is because it can be observed in publications and databases (see for instance [19]) that, despite the decreasing of RNA-seq costs, there are still many experiments with no replication or even experiments that are used as preliminary tests. Hence, we think that NOISeq could be still useful to analyze this kind of experiments.

Some exploratory tools can also be found in the package, for both the count data and the differential expression results. We describe all these features in the following sections.

More information about the use of all the functions in the package can be found in the NOISeq User's Guide at http://www.bioconductor.org/packages/2.12/bioc/html/NOISeq.html.

2.3.1 Exploration of expression data

The package contains a set of plots to explore the characteristics of the features detected according to the counts and other biologial information provided such as the position of the feature in the chromosome or the biological group where it belongs (biotypes), and also to increasing sequencing depths (simulated from the given sample). Most of these plots were used for our previous work [1] and most of them were implemented in the sofware Qualimap [20], so we will not describe them again in this paper.

2.3.2 Exploration of differential expression results

Once the differential expression results have been obtained, it may be interesting to visualize them graphically. To illustrate the tools implemented in NOISeq package for this purpose (function *DE.plot*), we will use the *F. oxysporum* data set (see section **??**) and will compare the expression in blood (B) and in minimal medium (MM) at 30 minutes for wild type strain.

Figure 1 shows in each axis the average expression for each experimental condition (also in log-expression if indicated). When a threshold q to select differentially expressed genes (DEG) is given, the DEG are highlighted in red.

In Figure 2, the (M, D) signal values are represented. Again, if a threshold q to select differentially expressed genes (DEG) is provided, the DEG are highlighted in red.

When position in chromosomes has been provided for each feature, it may be useful to plot the expression across all the positions and chromosomes (in grey) and also see where the DEG are located (in red the up-regulated genes and in green the down-regulated genes) as for example in Figure 3, in which two chromosomes were depicted.

Finally, the distribution of DEG across chromosomes or biotypes (if provided) can be visualized, as in Figure 4.

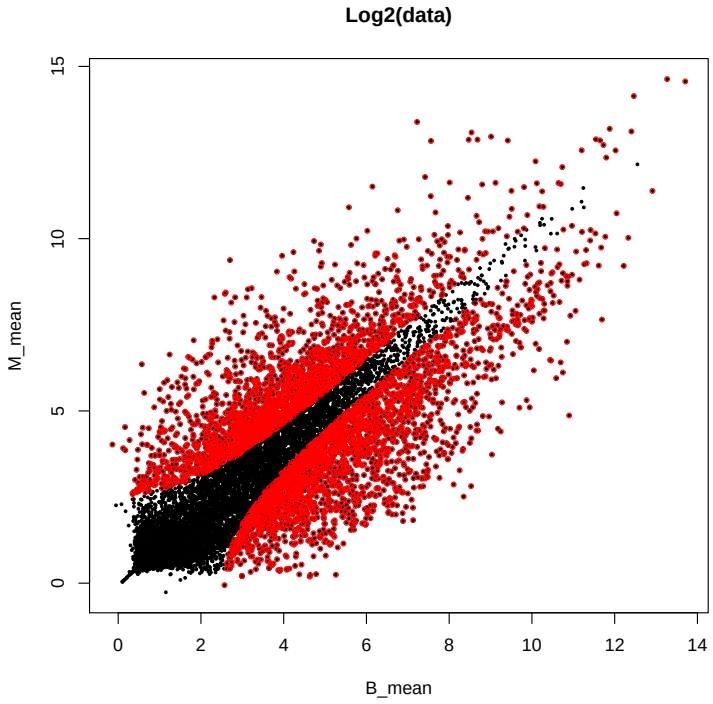

Figure 1: Example of Expression plot in NOISeq package from *F. oxysporum* data. X-axis: blood condition. Y-axis: MM condition. In both axis the log_2 of the average expression of the corresponding condition is shown. In red, the differentially expressed genes.

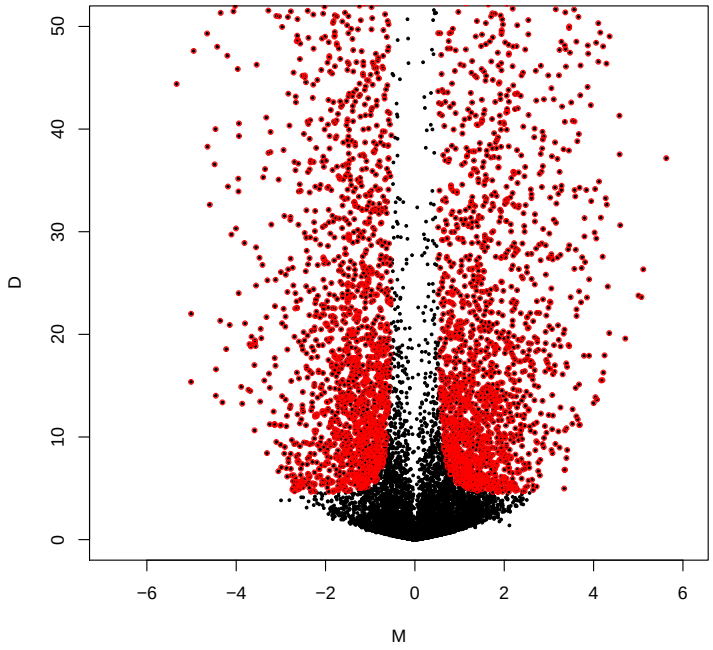

Figure 2: Example of (M, D) plot in NOISeq package from *F. oxysporum* data. In red, the differentially expressed genes.

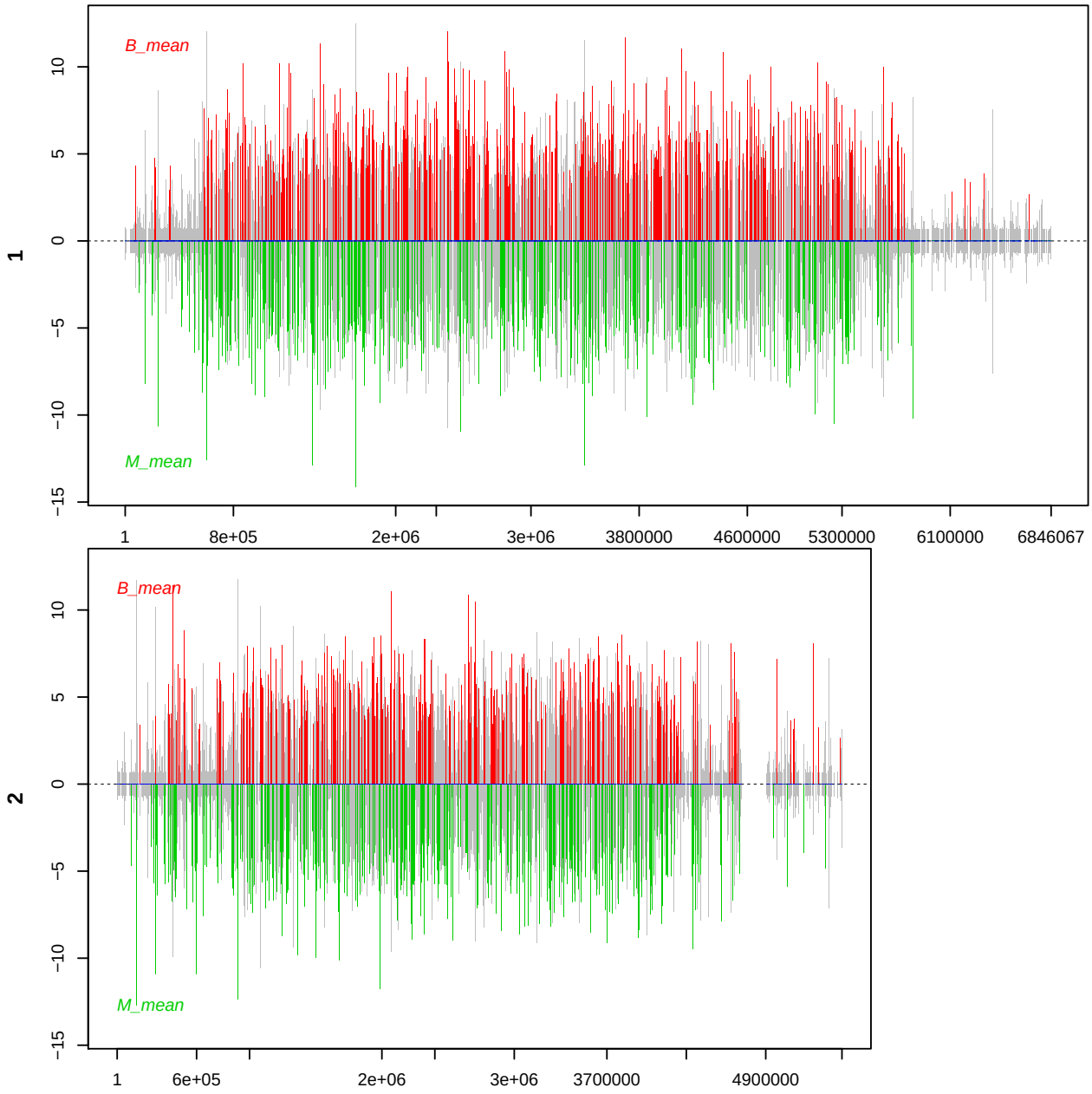

Figure 3: Example of Manhattan plot in NOISeq package from *F. oxysporum* data. X-axis: position in chromosome. Y-axis: log_2 of average expression levels (positive and negative values for up and down regulated genes, respectively). In red and green, the differentially expressed genes (up and down regulated, respectively).

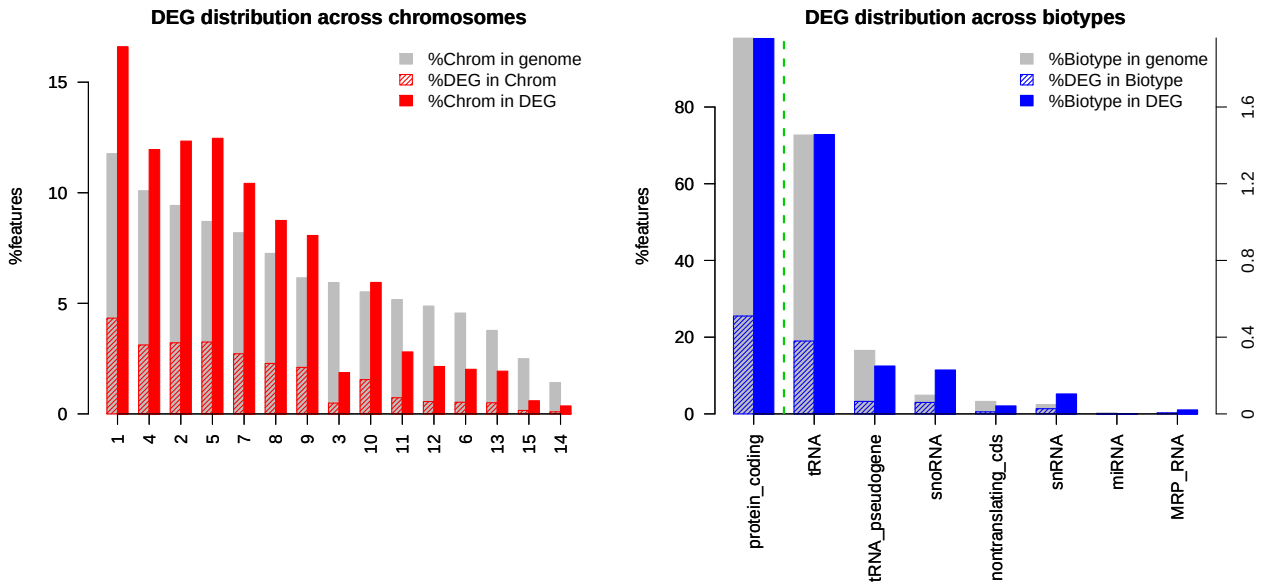

Figure 4: Example of plot showing distribution of DEG across chromosomes and biotypes in NOISeq package from *F. oxysporum* data. Left plot (red) shows the chromosomes ordered according to the number of genes they contain. Grey bars represent the percentage of total genes in each chromosome. Color stripped bars are the proportion of genes in each chromosome that are differentially expressed. Solid color bars are the percentage of each chromosome within the differentially expressed genes. Right plot (blue) shows the biological groups (biotypes) ordered according to the number of genes in each one. To the left of the green vertical line, the most abundant biotypes. Grey bars represent the percentage of total genes in each biotype. Color stripped bars are the proportion of genes in each biotype that are differentially expressed. Solid color bars are the percentage of each biotype within the differentially expressed genes.

3 Preliminary results on simulated data sets

To evaluate the performance of NOISeqBIO and compare it to the traditional NOISeq and to other methods (edgeR [7], DESeq [6] and SAMseq [10]), we simulated data that mimic RNA-seq counts. We took a sample from an experimental data set of the fungal pathogen *A. fumigatus* and used the negative binomial distribution to model the biologial variability among replicates within the same condition. The differentially expressed genes (DEG) had different means for each condition. For this preliminary evaluation we simulated 5 replicates per experimental condition, different levels of technical noise (0%, 20% and 40%) and different proportion of DEG (1%, 5% and 10%). We generated 10 data sets for each combination. We chose a threshold of 0.05 for the adjusted p-value for all methods or, equivalently, a threshold of 0.95 for the probability of differential expression for NOISeqBIO. In the case of NOISeq, we used the recommended threshold of 0.8. The results of this preliminary study are shown in Figure 5. We measured the performance of the methods by computing the sensitivity (SE), the False Discovery Rate (FDR) and the Matthews Correlation Coefficient (MCC), that is a kind of summary of the two first indicators. It can be observed the excellent performance of NOISeqBIO (Bio4 in the plot) when compared to the rest of methods. NOISeqBIO presents a good sensitivity while keeping a low FDR, which results in a high MCC.

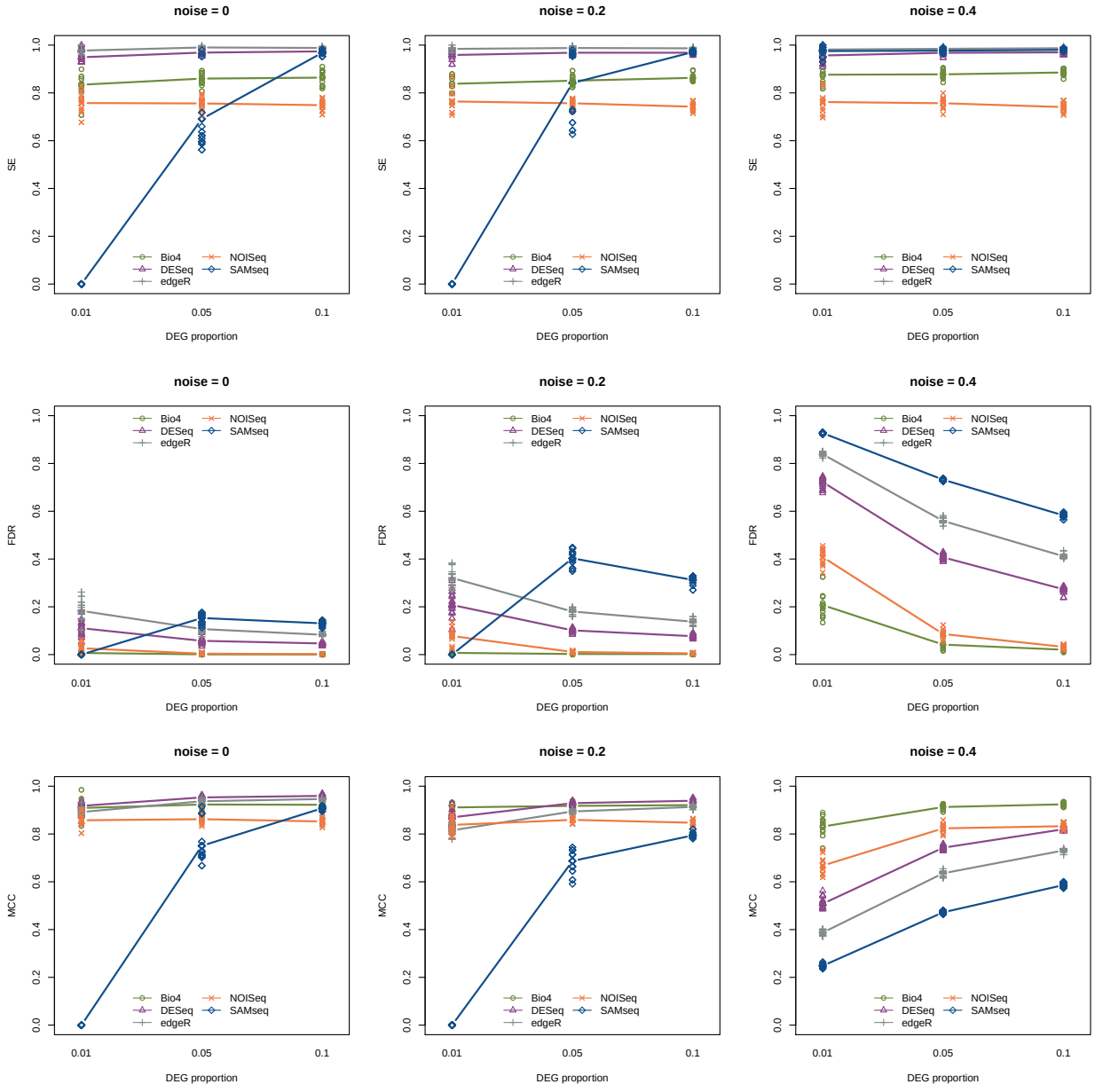

Figure 5: Comparison of methods performance on simulated data sets with 5 replicates.

References

[1] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa, "Differential expression in RNA-seq: A matter of depth," *Genome Research*, vol. 21, pp. 2213–2223, Dec. 2011.

[2] J. Malone and B. Oliver, "Microarrays, deep sequencing and the true measure of the transcriptome," *BMC biology*, vol. 9, no. 1, p. 34, 2011.

[3] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, 2008.

[4] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.," *Genome research*, vol. 18, pp. 1509–1517, Sept. 2008.

[5] A. Oshlack, M. Robinson, and M. Young, "From RNA-seq reads to differential expression results," *Genome Biology*, vol. 11, pp. 220+, Dec. 2010.

[6] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biol*, vol. 11, no. 10, p. R106, 2010.

[7] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, pp. 139–140, Jan. 2010.

[8] T. Hardcastle and K. Kelly, "baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics*, vol. 11, no. 1, pp. 422+, 2010.

[9] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang, "DEGseq: an R package for identifying differentially expressed genes from RNA-seq data," *Bioinformatics*, vol. 26, pp. 136–138, Jan. 2010.

[10] J. Li and R. Tibshirani, "Finding consistent patterns: A nonparametric approach for identifying differential expression in rna-seq data," *Statistical Methods in Medical Research*, 2011.

[11] P. G. Ferreira, S. Patalano, R. Chauhan, R. Ffrench-Constant, T. Gabaldon, R. Guigo, and S. Sumner, "Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes," *Genome Biology*, vol. 14, no. 2, p. R20, 2013.

[12] J. Carcel-Trullols, C. Aguilar-Gallardo, F. Garcia-Alcalde, M. A. Pardo-Cea, J. Dopazo, A. Conesa, and C. Simón, "Transdifferentiation of malme-3m and mcf-7 cells toward adipocyte-like cells is dependent on clathrin-mediated endocytosis," *SpringerPlus*, vol. 1, no. 1, pp. 1–12, 2012.

[13] Q.-H. Zhu, S. Stephen, K. Kazan, G. Jin, L. Fan, J. Taylor, E. S. Dennis, C. A. Helliwell, and M.-B. Wang, "Characterization of the defense transcriptome responsive to fusarium oxysporum-infection in arabidopsis using rna-seq," *Gene*, vol. 512, no. 2, pp. 259–66, 2013.

[14] C. Soneson and M. Delorenzi, "A comparison of methods for differential expression analysis of rna-seq data," *BMC Bioinformatics*, vol. 14, no. 1, p. 91, 2013.

[15] I. Nookaew, M. Papini, N. Pornputtpong, G. Scalcinati, L. Fagerberg, M. Uhlén, and J. Nielsen, "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 Research*, 2012.

[16] J. Bullard, E. Purdom, K. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinformatics*, vol. 11, pp. 94+, Feb. 2010.

[17] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biology*, vol. 11, pp. R25+, Mar. 2010.

[18] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher, "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 2001.

[19] J. Feng, C. A. Meyer, Q. Wang, J. S. Liu, X. S. Liu, and Y. Zhang, "Gfold: a generalized fold change for ranking differentially expressed genes from rna-seq data," *Bioinformatics*, vol. 28, no. 21, pp. 2782–8, 2012.

[20] F. García-Alcalde, K. Okonechnikov, J. Carbonell, L. M. Ruiz, S. Götz, S. Tarazona, T. F. Meyer, and A. Conesa, "Qualimap: evaluating next generation sequencing alignment data," *Bioinformatics*, vol. 28, no. 20, pp. 2678–9, 2012.