Colloquium Biometricum 44
2014, 155−165

SELECTED STATISTICAL METHODS OF RNA−SEQ DATA ANALYSIS

Katarzyna Górczak, Idzi Siatkowski

Department of Mathematical and Statistical Methods
Poznań University of Life Sciences
Wojska Polskiego 28, 60-637 Poznań, Poland
e-mails: kgorczak@up.poznan.pl, idzi@up.poznan.pl

Summary

RNA−Seq technology has been widely used in the natural sciences, biology and medicine. It is a tool for investigating cellular processes in animals and plants. RNA−Seq experiments provide quantitative readouts in the form of count data. In this paper we would like to present some R packages used in RNA−Seq data analysis, especially for investigating differentially expressed genes.

Keywords and phrases: RNA−Seq, differentially expressed genes, edgeR, DESeq, EBSeq

Classification AMS 2010: 62-07, 62-09

1. Introduction

Next-generation sequencing (NGS) is nowadays the fastest-growing technology that can be used in genomic measurements, characterization and quantification of transcriptomes. One of the NGS−based applications is RNA−Seq used for analysis of gene expression. RNA−Seq eliminated several limitations inherent in the hybridization-based microarray technologies. In addition, this technology provides new knowledge of the range of gene expression levels and detection of alternative splicing events and gene fusion

transcripts. RNA−Seq is used in cancer research and disease diagnosis. It also helps to control cellular processes in plants or animals. An RNA−Seq experiment takes a sample of purified RNA, shears it and makes it possible to perform an RNA analysis through cDNA sequencing, and, in the effect, obtaining millions of short reads (Oshlack et al., 2010). Subsequently, this experiment covers a low – level analysis (such as base calling, read mapping, alignment), a high – level analysis (such as normalization, quantification expression, differential expression) and, finally, biological insight. In this paper we focus on the statistical testing of differential expression. We would like to compare the recently proposed statistical methods of detecting differentially expressed genes from edgeR, DESeq and EBSeq packages from the R environment. The purpose of the paper is to describe an RNA−Seq experiment and some of its most important aspects. Secondly, we compare edgeR, DESeq and EBSeq and review certain useful functions used by these methods. We would like to stress the fact that the R platform version 3.0.2 was used in all the computations.

2. RNA−Seq experiment

The sequencing process consists of three steps: library creation, amplification and establishing the precise order of nucleotides within an RNA molecule. An RNA−Seq experiment using an Illumina's Genome Analyzer approach takes an RNA sample, removes contaminants and divides it into small fragments in random positions. Subsequently, these fragments are converted to cDNA in the process of reverse transcription (Oshlack et al., 2010). A complementary strand is removed, special primers are attached and amplification using the polymerase chain reaction occurs. An RNA−Seq experiment requires special surface to which fragments with primers are attached. Nucleotides and enzyme are added to initiate bridge amplification. Several million of dense clusters of sequences are generated in each channel of the flow cell. In the last stage of the experiment all four labeled nucleotides are added (Bullard et al., 2010). After laser excitation, the image of emitted fluorescence from each cluster is captured. Markers are eluted and this cycle is repeated many times. The images make it possible to read the sequences and produce millions of short reads, which are typically mapped to a reference genome (Soneson and Delorenzi, 2013). Next generation sequencing requires a high-throughput platforms such as: Applied Biosystems' SOLiD (technology based on sequencing by ligation), Illumina's Genome Analyzer (technology based on sequencing by synthesis), Roche's 454 Life Sciences (technology based on pyrosequencing), Ion Torrent (technology based on Ion

semiconductor), Pacific Biosciences (technology based on single-molecule real-time sequencing) (Kvam et al., 2012). The main RNA−Seq data repository is SRA (Short Read Archive). This repository stores data in its own SRA format so that it can be searched for and downloaded in a convenient way. The basic and most commonly used extension of data from next-generation sequencing is FASTQ storing informations about read sequence and quality. The quality is calculated with the Phred Quality Score, follows $Q = -10\log_{10} p$, where p is the probability of a base call error. Smaller likelihood results in higher accuracy. A reference genome sequence is stored in FASTA format.

3. Mapping

Mapping and alignment constitute the first step in the RNA−Seq data analysis after checking the sequence quality and removing low quality reads in the image analysis. Millions of short reads obtained from the sequencing process must be turned into a quantification of expression. Read mapping makes it possible to find a region where a short read is identical to the reference genome. However, such matching may not be accurate. Short reads may be matched to several locations or can be derived from spliced regions, what may result in errors. Therefore, it is necessary to find the best location in the reference (Oshlack et al., 2010). Local mapping is possible using the Smith-Waterman algorithm, which compares small segments instead of looking at the total sequence. Similar to the Smith-Waterman algorithm is the BLAST search algorithm, which is, however, better in terms of speed. Mapping is possible using the Burrows-Wheeler transform or hash-tables. Mappers which have been recently available are: MAQ, SOAP, Bowtie, SHRiMP, BWA, TopHat, MIRA (Li et al., 2008; Flicek and Birney, 2009; Langmead et al., 2009). The first five mappers are general aligners. TopHat is a *de novo* annotator and MIRA is a *de novo* transcript assembler. This option can be used to identify novel transcripts when reference genome is not available. Bowtie is based on the Burrows-Wheeler transform. Summarizing mapped reads constitutes the next step in further analysis which relies on reads counting.

4. Statistical methods

Let K_{ij} denote the observed count for gene $i = 1, \dots, n$ and sample $j = 1, \dots, m$. We assume that the read counts K_{ij} are derived from a Negative Binomial (NB) distribution, as follows:

$$K_{ij} \sim NB(\mu_{ij}, \phi),$$

where μ_{ij} is a mean, and ϕ is the dispersion. Mean and variance are related by $\sigma_{ij}^2 = \mu_{ij} + \mu_{ij}^2 \phi$. Furthermore, let m_j be the library size for sample j. We assume that $\mu_{ij} = \lambda_{ij} m_j$, where λ_{ij} is the level of gene expression (of gene i from sample j). To assess differences in differential expression levels between gene i from sample A and gene i from sample B (where, for instance, sample A may represent disease, and sample B may represent control), the null hypothesis $H_0: \lambda_{iA} = \lambda_{iB}$ is tested against a two-sided alternative hypothesis and it is made for each gene (Anders and Huber, 2010). The hypothesis concerning the differential expression is tested using exact test – edgeR and a similar approach in DESeq (Robinson et al., 2010; Anders and Huber, 2010; respectively) or empirical Bayes approach – EBSeq (Leng et al., 2013).

5. Normalization

Normalization is an essential step in the analysis of differentially expressed genes. It allows us to compare the expression between samples with regard to some technical effects from the sequencing. There are several normalization methods used for a count-based differential analysis: Reads per Kilobase per Million reads (RPKM), TotalCount, trimmed mean of M-values (TMM), Median, Quantile, Upper Quartile or relative log expression (RLE) (Dillies et al., 2012). The simplest normalization procedure is RPKM, which divides the gene count by the total number of reads in each library. In this study we use two methods of normalization: TMM and Median. DESeq estimates a scaling factor by the median of the ratio of the observed counts given a geometric mean across all the samples and interpreted as a pseudo-reference sample, as follows:

$$\hat{s}_j = median_i \frac{K_{ij}}{(\prod_{v=1}^{m} K_{iv})^{1/m}}.$$

A similar approach is used in EBSeq. The method included in DESeq package is based on the hypothesis that most genes are not differentially expressed. The method implemented in the edgeR package based on a similar assumption uses a weighted TMM (Robinson and Oshlack, 2010). This procedure establishes gene-wise log-fold-changes, as follows:

$$M_{ij} = \log_2\left(\frac{K_{ij}}{N_j}\Big/\frac{K_{ij'}}{N_{j'}}\right),$$

and the absolute expression level

$$A_{ij} = \frac{1}{2}\log_2\left(\frac{K_{ij}}{N_j} * \frac{K_{ij'}}{N_{j'}}\right),$$

where N_j is a total number of reads from sample j ($N_j = \sum_{i=1}^{n} K_{ij}$). First, one sample ($j'$) is selected as the reference and values M_{ij} and A_{ij} are calculated. Subsequently, some values M_{ij} and values A_{ij} are trimmed (Robinson and Oshlack, 2010; propose 30% values M_{ij} and 5% values A_{ij}) and normalization factor for sample j using the reference sample j' is calculated as:

$$\log_2(TMM_j^{j'}) = \frac{\sum_{i\in G^*} w_{ij}M_{ij}}{\sum_{i\in G^*} w_{ij}},$$

where G^* is the set of genes with not trimmed values of M_{ij} and A_{ij} and $w_{ij} = \frac{N_j - K_{ij}}{N_j K_{ij}} + \frac{N_{j'} - K_{ij}}{N_{j'} K_{ij}}$.

6. Differential expression analysis

6.1. Data

In the analysis we take into consideration the datasets known from the literature. The data is presented in the form of a rectangular table of integer values, where genes correspond to rows and samples correspond to columns. Each cell of this table tells us how many reads have been mapped to some gene in some sample. The first dataset – 'fly' (Anders and Huber, 2010) includes counts from 17 605 genes in 4 samples. The second dataset – 'pnas' (Li et al.,

2008) concerns RNA−Seq data from a treatment vs control experiment with relatively low biological variability (37 435 genes in 7 samples).

6.2. DESeq, edgeR and EBSeq

Three R packages: edgeR, DESeq and EBSeq have implemented methods based on the negative binomial model. EdgeR has been used primarily for serial analysis of gene expression (SAGE). This package requires the estimation of the dispersion parameter. Firstly, we use a common dispersion for all genes (Robinson and Smyth, 2008), and further this method estimates the tagwise dispersion for each gene. Both edgeR and DESeq estimate the variance assuming the linear relationship between variance and mean expression levels (Kvam et al., 2012). The first step in any analysis is usually reading the table of counts into an R session. Then we can work with each package and create its objects. In the DESeq central data structure is a `CountDataSet()`, and edgeR stores data in a list-based object called `DGEList()`. EBSeq requires that data should be loaded as a matrix. It is necessary to estimate the effective library size. It can be obtained in R by writing the following DESeq code:

```
>cds1 <- estimateSizeFactors(cds)
>sizeFactors(cds1),
```

where `cds` is a DESeq `CountDataSet` object. EBSeq, simillary to DESeq, uses the median procedure to obtain the library size factor for each sample and it may be done via `MedianNorm()` function. In edgeR it may be obtained by `calcNormFactors()`.

After the normalization process we estimate the dispersions. In DESeq we can use functions:

```
>cds2 <- estimateDispersions(cds)
>str(fitInfo(cds2))
>plotDispEsts(cds2)
```

The first line of the code estimates the dispersion value for each gene and fits a curve through the estimates. The second line stores information about the per-gene estimate, fitted curve and estimated values. The last line generates a plot of empirical and fitted dispersion values per-gene against the mean of normalized counts. The plot for ‘pnas’ data is shown in Figure 1.

In this paper we focus on finding genes that are differentially expressed between two samples. In edgeR the testing can be done using the function `exactTest()`, in DESeq – using the function `nbinomTest()`, and in EBSeq – by the function `EBTest()`. The second function returns a data frame with useful information, e.g. the adjusted p-values, which are computed in such a way that the false discovery rate (FDR) is controlled at some level.

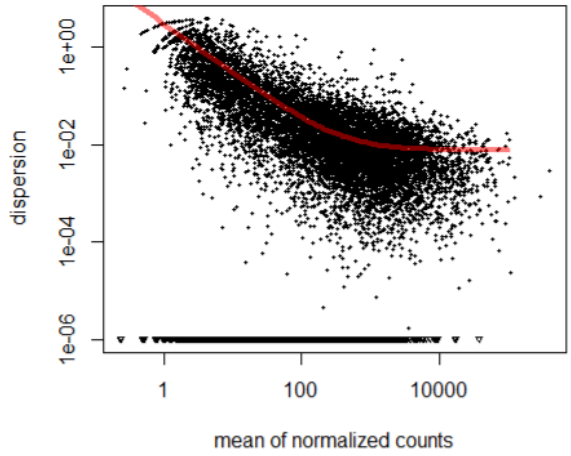


Fig. 1. Empirical and fitted dispersion values per-gene against the mean of normalized counts (DESeq; 'pnas' data)

Table 1. Data frame with some information obtained with a differential expression test (DESeq (a), edgeR (b), 'fly' data)

(a) DESeq

gene id	baseMean	baseMeanA	baseMeanB	Fold Change	log2Fold Change	pval	padj
Gene_14350	1656.4268	0.0000	3312.8536	Inf	Inf	5.69E-185	5.50E-181
Gene_2090	675.2460	0.0000	1350.4920	Inf	Inf	1.14E-139	5.50E-136
Gene_16627	263.0968	524.7669	1.4267	0.0027	-8.5228	4.42E-86	1.42E-82
Gene_11463	805.6701	1488.5990	122.7412	0.0825	-3.6003	5.25E-84	1.27E-80
Gene_14906	1106.6140	1989.3712	223.8567	0.1125	-3.1517	9.21E-79	1.78E-75
Gene_12596	154.4248	10.6986	298.1509	27.8681	4.8005	4.35E-36	7.01E-33

(b) edgeR

gene id	logFC	logCPM	p-value	FDR
Gene_14350	14.70667751	6.372796573	1.02E-280	9.81E-277
Gene_2090	13.41294393	5.080570537	2.56E-180	1.24E-176
Gene_9780	-7.019562141	4.925005904	1.03E-121	3.33E-118
Gene_16627	-8.306116176	3.719361236	1.57E-99	3.79E-96
Gene_16573	5.94773123	4.079433516	3.07E-79	5.93E-76
Gene_9774	-8.790872362	3.673782042	5.74E-78	9.25E-75

The data frame for ‘fly’ data in DESeq is presented in Table 1a. The ‘baseMeanA’ and the ‘baseMeanB’ are the means of normalized counts calculated for each gene within sample A and sample B, the ‘baseMean’ is the mean of the ‘baseMeanA’ and ‘baseMeanB’ values. The ‘foldChange’ is the ratio of ‘baseMeanB’ to ‘baseMeanA’. A similar table can be obtained from edgeR (Table 1b). The ‘logFC’ corresponds to $\log_2$ fold change of the genes, the ‘logCPM’ value is calculated as the counts divided by the library sizes and multiplied by one million. In EBSeq we may also get a list of differentially expressed genes.

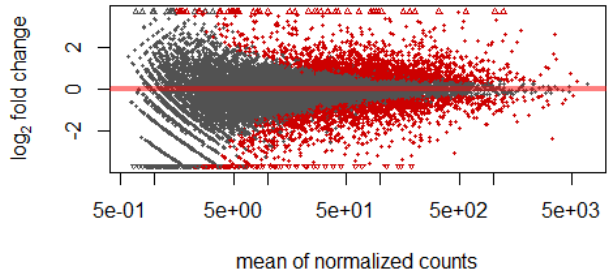


Fig. 2. Plot for genes identified as being differentially expressed (DESeq; ‘pnas’ data)

By the function `GetPPMat()` we obtain a matrix containing the posterior probabilities of being equivalently expressed genes or differentially expressed genes. The graphical representation of the sixth column (‘log2FoldChange’) against the second column (‘baseMean’) from Table 1a, for example for ‘pnas’ data, is shown in Figure 2. The red points are genes which are significant at a 5% false discovery rate.

A similar plot can be obtained in edgeR, which is shown in Figure 3.

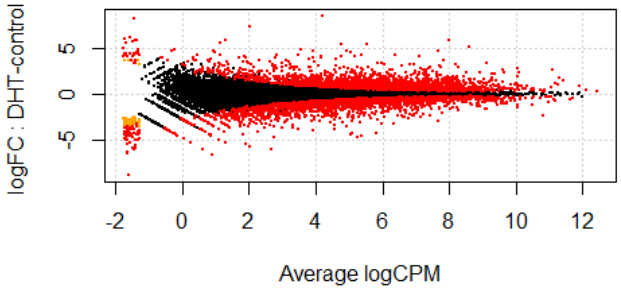


Fig. 3. Plot for genes identified as being differentially expressed (edgeR; ‘pnas’ data)

In R it is possible to obtain information about the most significantly differentially expressed genes, but also about the most strongly down-regulated or up-regulated genes (Table 2). Numbers -1, 0 and 1 are for down-regulated, non-differentially expressed and up-regulated genes, respectively. The following line of code displays the number of genes significantly differentially expressed at a false discovery rate of 5%:

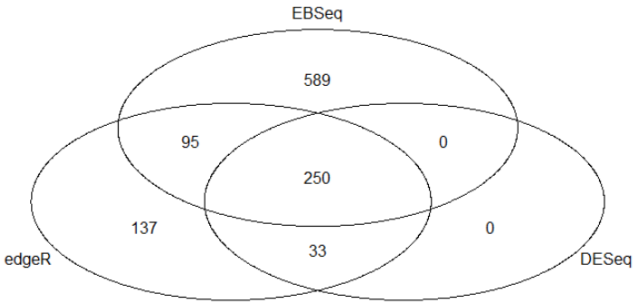
```
>length(which(res$padj<0.05))
```

Table 2. Down-regulated, non-differentially expressed genes, up-regulated (DESeq, edgeR, ‘pnas’ data)

edgeR			DESeq		
-1	0	1	-1	0	1
2094	12060	2340	1270	13698	1526

In this paper we have taken into consideration two data sets. For edgeR and DESeq, genes were recognized as differentially expressed with adjusted p-value lower than 0,05. For EBSeq, differentially expressed genes were selected with posterior probabilities greater than 0,95. The first set (‘fly’) consisted of 17 605 genes, and the second one (‘pnas’) – of 37 435 genes. We compared packages in finding differentially expressed genes by the Venn diagram (Figure 4). We can see the number of differentially expressed genes detected by each method, and the number of commonly detected genes. For ‘fly’ data set, edgeR found 515 differentially expressed genes, DESeq – nearly half as many, and EBSeq almost twice more than edgeR.

(a) ‘fly’ data



(b) ‘pnas’ data

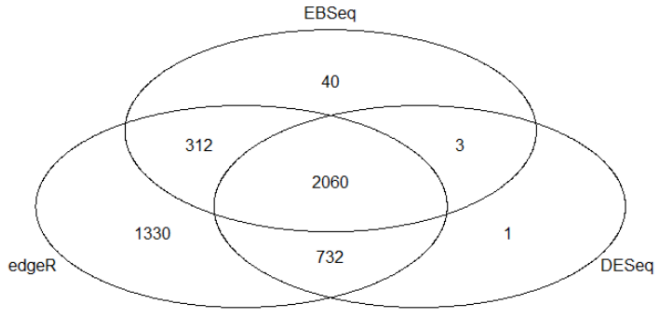


Fig. 4. Venn diagrams for differentially expressed genes for ‘fly’ data set (a) and ‘pnas’ data set (b)

7. Conclusion

The paper has tried to make a brief review of the three statistical methods of detecting differentially expressed genes. All methods based on the negative binomial distribution are comparable. Both edgeR and DESeq produce p-values and adjusted p-values. EBSeq provides the posterior probability of being differential expression and each of them allows us to control the FDR rate. During the normalization process, edgeR uses a trimmed mean of M-values between each pair of samples, whereas DESeq and EBSeq use median normalization. The analyses were performed in DESeq, edgeR and EBSeq, which are a packages for the statistical environment R and are available from a Bioconductor repository. These packages have implemented functions useful in assessing the results of the RNA−Seq experiment. Gene expression measured by the number of reads mapped to a reference genome helps to understand the impact of these genes on certain diseases and cellular processes. We have tested differential expression within a pairwise comparison. EdgeR and EBSeq are the methods for analyzing the data involving two or more samples with replicates. DESeq can be applied in analyzing data from an experiment without replicates. A comprehensive description of the above mentioned methods as well as other methods for detecting differentially expressed genes may be found in Soneson and Delorenzi (2013).

Acknowledgements

The authors would like to thank the reviewers for a thorough evaluation of the paper and valuable comments.

References

Anders S., Huber W. (2010). Differential expression analysis for sequence count data. *Genome Biology* 11, R106.

Bullard J.H., Purdom E., Hansen K.D., Dudoit S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11, 94.

Dillies M., 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., Jaffrézic F.: French StatOmique Consortium. (2012). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*. doi:10.1093/bib/bbs046.

Flicek P., Birney E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature Methods* 6, S6 – S12.

Kvam V.M., Liu P., Si Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from RNA-Seq data. *American Journal of Botany* 99(2), 248–256.

Langmead B., Trapnell C., Pop M., Salzberg S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10, R25. doi: 10.1186/gb-2009-10-3-r25

Leng N., Dawson J.A., Thomson J.A., Ruotti V., Rissman R.A., Smits B.M.G., Hagg J.D., Gould M.N., Stwart R.M., Kendziorski C. (2013). Ebseq: An empirical Bayes hierarchical model for inference in RNA-Seq experiments. *Bioinformatics* 29 (8), 1035–1043. doi: 10.1093/bioinformatics/btt087

Li H., Lovci M.T., Kwon Y.-S., Rosenfeld M.G., Fu X.-D., Yeo G.W. (2008). Determination of tag density required for digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model. *Proceedings of the National Academy of Sciences of the USA* 105, 20179–20184.

Li H., Ruan J., Durbin R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18(11), 1851–1858. doi: 10.1101/gr.078212.108

Oshlack A., Robinson M.D., Young M.D. (2010). From RNA-seq reads to differential expression results. *Genome Biology* 11, 220.

Robinson M.D., McCarthy D.J., Smyth G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Robinson M.D., Oshlack A. (2010). A scaling normalization method for differential expression analysis RNA-Seq data. *Genome Biology* 11, R25.

Robinson M.D., Smyth G.K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9, 2, 321–332.

Soneson Ch., Delorenzi M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14, 91.