Colloquium Biometricum 44
2014, 167−181

GRAPHICAL METHODS FOR DIFFERENTIAL ANALYSIS OF RNA−SEQ DATA

Katarzyna Górczak, Katarzyna Klamecka, Alicja Szabelska,
Joanna Zyprych-Walczak, 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 particularly used for investigating differential expression. There are several statistical methods and packages used to find differentially expressed genes (DEG) between two or more conditions. In order to compare the results, we need proper data visualizations. In this paper we focus on graphical methods of differential analysis.

Keywords and phrases: RNA−Seq, differential expression, statistical methods of differential analysis, graphical representation

Classification AMS 2010: 62-07, 62-09

1. Introduction

High-throughput sequencing technology is commonly used in genomic studies. One of the high-throughput sequencing applications is RNA−Seq used for measurement of gene expression levels. RNA−Seq allows one to study human diseases and biological systems in plant and animal (Kvam et al., 2012). Statistical analysis of RNA−Seq data has several steps. Finding genes with significant expression levels is an essential step in a differential analysis. In this paper we evaluated eight methods of graphical representation of DEG. Graphical presentation of the results facilitates the evaluation of the results. Four R packages were used to find DEGs: DESeq (Anders and Huber, 2010), edgeR (Robinson et al., 2010), EBSeq (Leng et al., 2013) and SAMSeq (Li and Tibshirani, 2011). They are freely available from the Bioconductor repository (www.bioconductor.org). All computations and graphs were performed in the R environment (R Core Team, 2013).

2. Methods

The analysis of an RNA−Seq experiment may be focused on the identification of differentially expressed isoforms, exons, transcripts and genes. In this paper we applied four methods to find differentially expressed genes. These methods are implemented in R packages: DESeq, edgeR, EBSeq and SAMSeq. An RNA−Seq experiment provides count data. In the R platform, the data is available in the form of matrix or a data frame, where each gene corresponds to the row and each sample corresponds to the column. The number of reads for gene g and a sample in class k may be denoted y_{gk}. DESeq and edgeR are based on the assumption that the number of reads for each gene can be modeled by a negative binomial distribution:

$$y_{gk} \sim NB(\mu_{gk}, \sigma_{gk}^2),$$

where μ_{gk} is the mean and σ_{gk}^2 is the variance of gene g in class k. The null hypothesis that the level of gene expression is equal between two classes is tested for each gene. EBSeq is particularly developed for isoform analysis, but it can be used to identify DEGs and also requires the gene counts. The false discovery rate (FDR) was calculated as the adjusted p-values. SAMSeq is a non-parametric method that uses permutation to determine the FDR rate. Calculated q-values were used in the further analysis. EBSeq uses a Bayesian approach and estimates the posterior probability of being differentially expressed for each gene separately. The genes that we considered as differentially expressed were the ones with the adjusted p-values at significance level 0.05 (for DESeq and edgeR methods), q-values lower that 0.05 (for the SAMSeq method) and probabilities posteriori greater than 0.95 (for the EBSeq method). In this paper we concentrate on the graphical presentation of DEG. We have taken into consideration eight methods that are described below.

ROC CURVES

Receiver operating characteristic (ROC) curves are useful for assessing the accuracy of predictions in evaluating and comparing models, algorithms or technologies. The most common way of showing the accuracy of prediction is using the true positives (TP), the false positive (FP), true negatives (TN) and false negatives (FN). Definitions of these are: $FPR = FP/(FP + TN)$, $FNR = FN/(FN + TP)$, where $TPR = 1 - FPR$ and $TNR = 1 - FNR$. The true positive rate (TRP) is called sensitivity and the true negative rate (TNR) is called specificity. The ROC curve is a graph of sensitivity (y - axis) vs. 1 – specificity (x - axis).

In a medical field, sensitivity is the fraction of people with the disease that the test correctly identifies as positive, whereas specificity is the fraction of people without the disease that the test correctly identifies as negative. ROC curves presented in this paper were based on a machine learning method called stacked regression.

Kohavi and Wolpert (1996) presented an idea called stacked generalization, for combining estimators. The idea was later used by Breiman, in 1984, who introduced the stacked regression principle.

In our study we design an average estimator by a linear combination of three machine learning methods: support vector machine, neural network and k-nearest neighbors algorithm. This approach is expected to be more accurate than each of the estimators taken individually.

MA-PLOT

MA plot describes the relationship between the base mean expression and the $\log_2$ fold change of the genes (log2FC). The base mean expression is calculated using normalized values of counts for each sample. The log2FC is calculated as:

$$\log_2 FC = \frac{\bar{y}_{g1}}{\bar{y}_{g2}},$$

where $\bar{y}_{g1}$, $\bar{y}_{g1}$ are the normalized mean values of counts in class 1 and 2, respectively. On the x axis in the plot there is a normalized base mean for each gene, on the y axis there is the log2FC of the genes (Anders and Huber, 2010).

BOXPLOT

This figure shows the distribution of some statistical features. In the current study it will be a normalized base mean of counts for each gene and log2FC. Boxplot consists of a box, where a thick line inside represents the median, the top line is the third quartile (Q_3), and the bottom line is the first quartile (Q_1).

Outside of the box are extended sections called whiskers. The lower end of whiskers cannot be less than $Q_1 - 1.5 \cdot (Q_3 - Q_1)$, and the upper end of whiskers cannot be greater than $Q_3 + 1.5 \cdot (Q_3 - Q_1)$. The observations outside this range are draped on a graph individually and we call them outlier values. Variations of box plots can be found in McGill et al. (1978).

BARPLOTS

The barplot is a chart, which uses vertical or horizontal bars to represent categorical data. The length of each bar corresponds to the value of each category (Kelley and Donnelly, 2009). Barplots summarize differentially expressed genes for considered methods. It present not only the numbers of DEGs for each method, but also the structure of these sets of genes with respect to levels of counts abundances. The colors in the plot represent genes with different levels of counts abundance. Four levels of abundance were chosen: genes with the mean number of counts between samples<100, between 100-1000, between 1000-5000 and >5000.

VENN DIAGRAM

Venn diagrams introduced by John Venn show any possible relationships between several sets (Baron, 1969). These diagrams are commonly presented by overlapping circles. The overlapping surfaces of the wheels are the part of a common collection, so there are elements that belong simultaneously to both sets. This scheme shows the number of common differentially expressed genes between each combination of methods.

DENDROGRAM

In this method, we calculate the similarity between two differential expression methods based on the differentially expressed gene ranks. For four different methods we obtained some differentially expressed genes. From these sets we chose genes common to all the sets. Then, for all four methods, we ranked these genes, thus obtaining four ranking lists of genes. Based on these lists we computed the distance matrix by using Euclidian distance. Finally, the comparison of the gene rankings was used in dendrograms. The dendrogram was constructed using complete-linkage clustering (Jain et al., 1999).

OVERLAPS

Overlap is an asymmetric matrix whose cells contain percentages of the number of commonly detected differentially expressed genes between the i-th and j-th methods (Seyednasrollah et.al., 2013). In (i, j)-th cell we have a proportion of common detections with respect to the i-th method

$$P_{ij} = \frac{D_{ij}}{D_i} \cdot 100,$$

where $i \neq j$, D_{ij} is the number of differentially expressed genes commonly detected by the i-th method and the j-th method, D_i is the number of differentially expressed genes detected by the i-th method.

HEATMAPS

The heatmap shows the correlation between the samples based on the differentially expressed genes found by each method. Based on this plot, the researcher can realize if probes fulfill the requirements of experimental design and the proper distinction of considered classes. The correlation is calculated with Spearman's correlation coefficient. In the current study the heatmaps were created for each considered method (Sneath, 1957).

3. Data

We considered four datasets: 'bodymap', 'montgomery', 'bottomly' and 'wang' obtained using Illumina's Genome Analyzer high-throughput sequencing system. The first dataset contains 52580 genes and 15 samples and is derived from the Body Map project (Asmann et al., 2012). The second dataset consists of 52580 genes and 129 human samples from two conditions: 60 Caucasian individuals from the HapMap3 project and 69 Nigerian individuals as a part of the International HapMap project (Montgomery et al. 2013 and Pickrell et al., 2010, respectively). The third dataset, related to research on mice (Bottomly et al., 2011), includes 36536 genes and 21 samples. Ten of them were of the C57BL/6J (B6) strain and 11 of the DBA/2J (D2) strain. The 'wang' dataset concerns the comparison of human tissues. Tissue samples were derived from single anonymous unrelated individuals of both sexes. These data include 52580 genes and 15 samples from two conditions: male and female (Wang et al., 2008). After filtering (throwing out all genes which had the mean value of counts across the samples equal to 0) the datasets had the following number of genes: 12953 (bodymap), 12984 (montpick), 13932 (bottomly) and 12627 (wang).

4. Results

In this paper we applied eight graphical methods described in the previous section for each result from the DESeq, edgeR, EBSeq and SAMSeq packages. Depending on the graphical method the results for chosen datasets were shown to present interesting features. To obtain the overview of the datasets we present them using Venn diagrams. The plots were created in ellipse shape with four subsets for the list containing DEGs detected using four statistical methods. The results are presented in Figure 1.

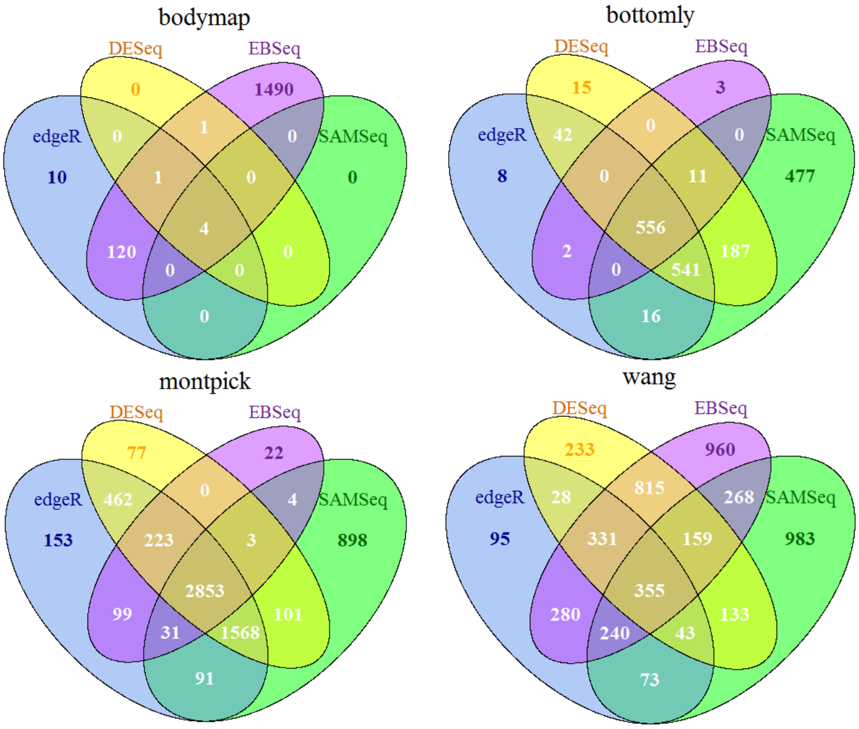
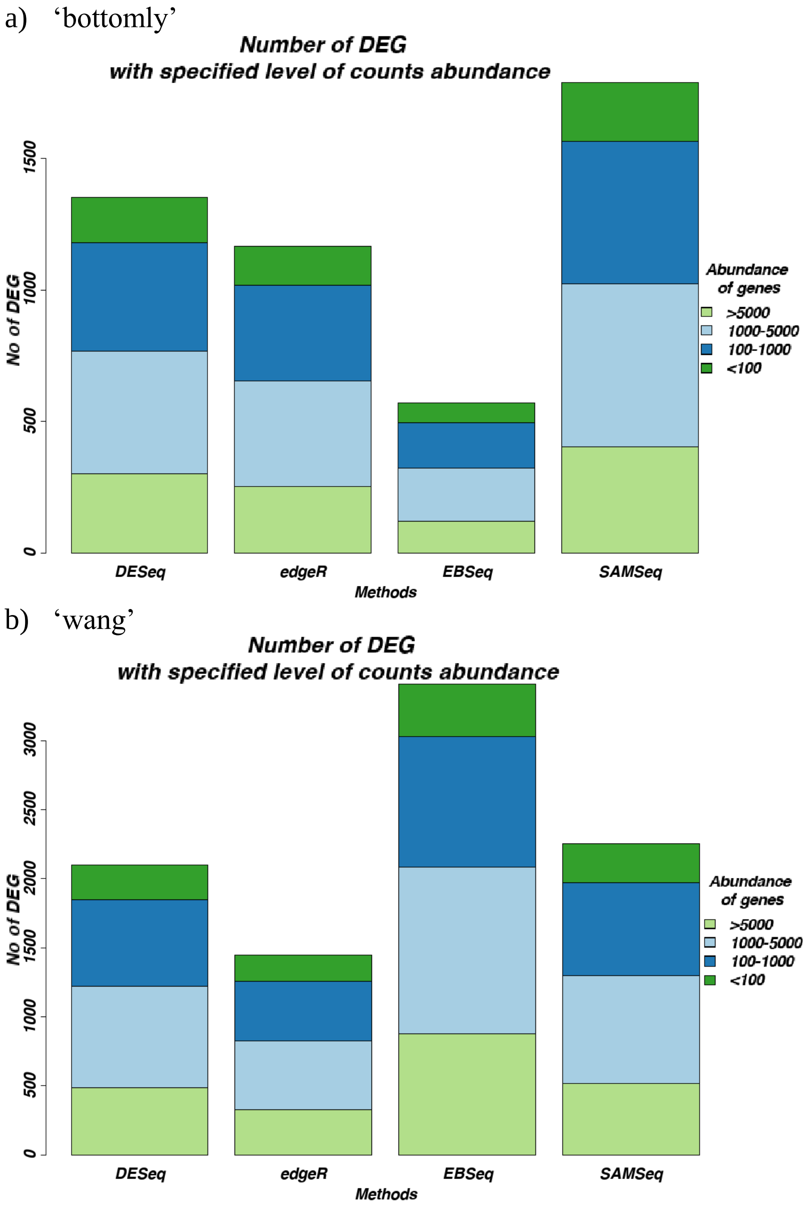


Fig. 1. Venn diagrams for DEG ('bodymap', 'bottomly', 'montpick' and 'wang' datasets)

We may notice that for the 'bodymap' dataset there are only 4 statistically significant genes common for all considered methods, whereas for 'montpick' almost 3000 common differential genes were found.

The structure of significant genes can be found in Figure 2. The results are shown in barplots for two datasets: 'bottomly' and 'wang'.



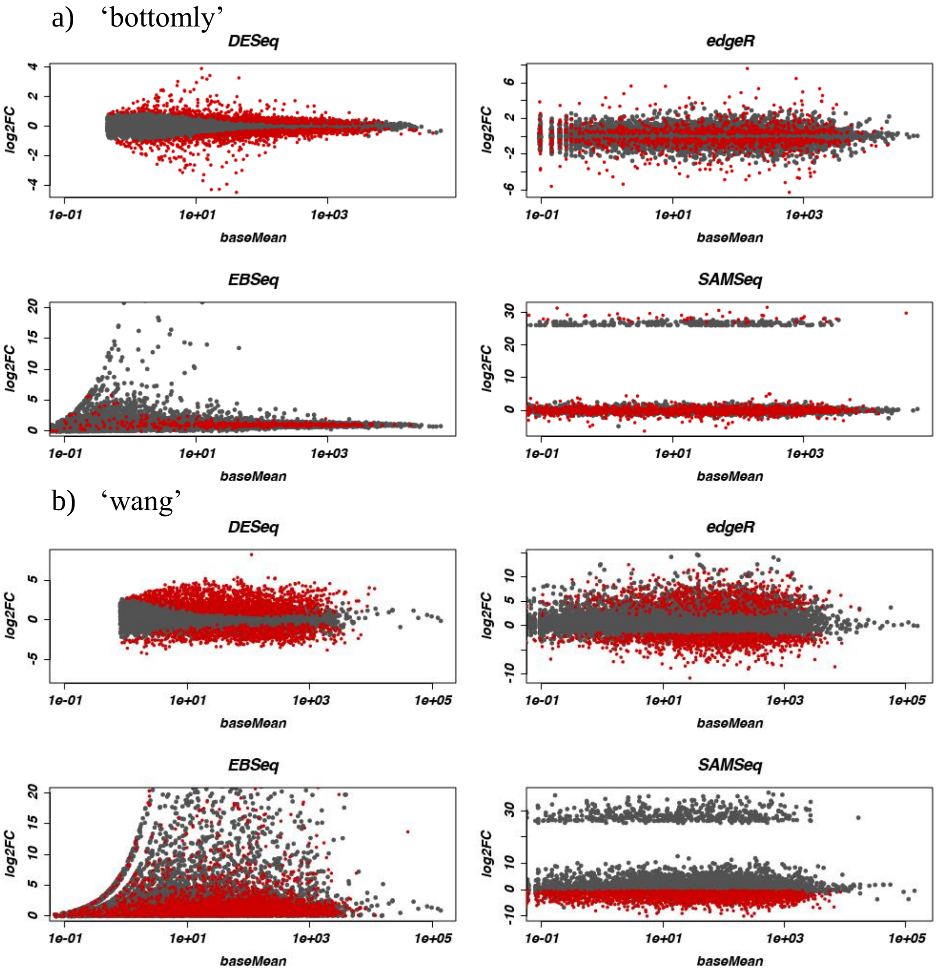
a) ‘bottomly’

b) ‘wang’

Fig. 2. Barplots for the number of DEG with a specified level of counts abundance (‘bottomly’ (a), ‘wang’ (b) datasets)

As we can see in Figure 2, the number of genes in each considered group of the abundance. The number of genes for each method is similar for the abundance of genes >5000. It is worth noticing that the SAMSeq method gives the lowest number of significant genes for the ‘bottomly’ dataset, whereas for the ‘wang’ dataset it is the opposite case. The numbers of genes with abundances between 10-100 and 1000-5000 are similar between one another for each statistical method.

If the investigators are interested in finding more information about down-regulated and up-regulated genes, they can refer to the MA-plots presented in Figure 3. These MA-plots, drawn for the ‘bottomly’ and ‘wang’ datasets, show how the effectiveness of detection of significant genes depends on the expression level.

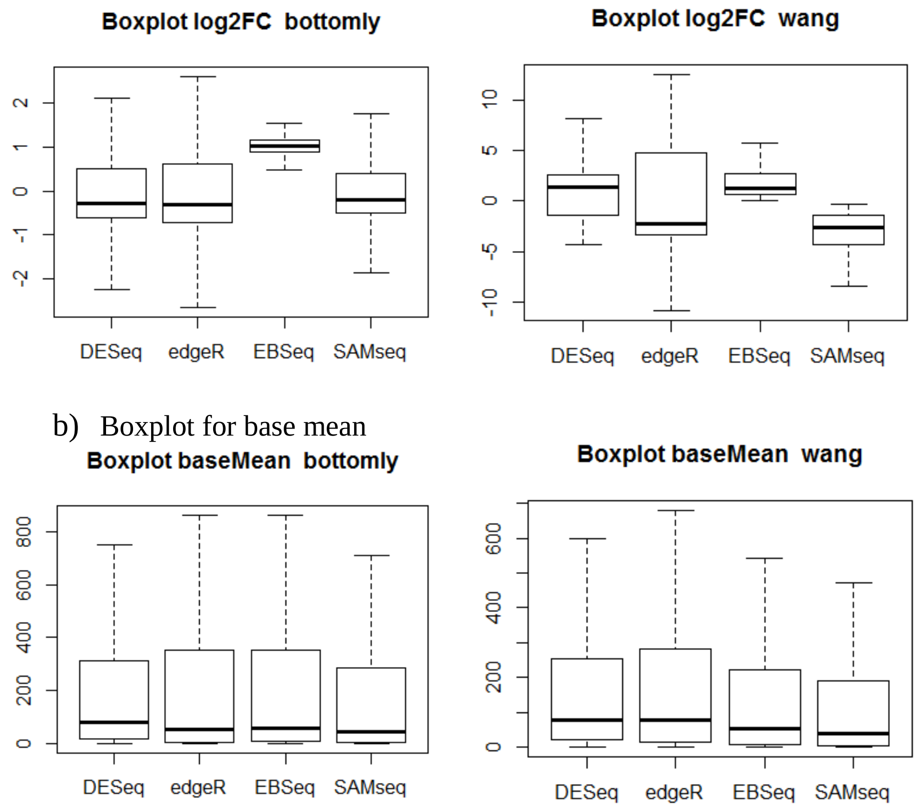

a) ‘bottomly’

b) ‘wang’

Fig. 3. MA-plots of DEG for each differential method (‘bottomly’ (a), ‘wang’ (b) datasets). The DEG found by the method are colored in red (in black when white-black printed)

From Figure 3, we can observe that normalization used for the DESeq method cuts the genes with a low value of the base mean of counts. In turn, the EBSeq method cuts down-regulated genes and finds only up-regulated DEGs.

Furthermore, the normalization used in SAMSeq divides the genes into two groups: the one around zero and the second one with highly up-regulated genes. The level of log2FC is much higher for the EBSeq and SAMSeq methods compared to DESeq and edgeR.


a) Boxplot for log2FC

b) Boxplot for base mean

Fig. 4. Boxplot of log2FC (a) and base mean (b) (‘bottomly’, ‘wang’ datasets)

The summarization of base means and log2FC for the ‘bottomly’ and ‘wang’ datasets is shown on Figure 4 as boxplots. There were a lot of observations strikingly far from the median values independently marked on the plot. Since individual points littered the graph, the outlier values were removed for the sake of clarity. Deleting them from the charts did not affect the characteristic values (the upper and lower quartiles, and the median), which remained the same. The values of the median of log2FC for the ‘wang’ dataset are different. However, for the ‘bottomly’ dataset the results are similar apart from the EBSeq. On boxplot for log2FC we can see in which range DEGs should be indicated in red in the previous Figure 3 (for instance, log2FC for the ‘wang’ dataset the EBSeq most frequently adopts the values between 0 and 6). In Figure 3 we may notice that for this dataset and this method indeed most DEGs are found in this range. The remaining genes, which are shown above, are outliers that were not marked on the boxplot. There were 573 such outliers in this case.

Another method of presenting results compares the numbers of common differentially expressed genes. For overlapping forms in Figure 5 we show the results for the ‘bottomly’ and ‘wang’ datasets. In this figure we may notice that for the ‘bottomly’ dataset, the methods reveal a high level of similarity. The highest percentages are between 96 – 99% of the number of commonly detected DEGs. The lowest percentages are obtained for the EBSeq – DESeq, EBSeq – edgeR, EBSeq – SAMSeq combinations. On the other hand, the combinations EBSeq – DESeq, EBSeq – edgeR give the highest percentages for the ‘wang’ dataset.

We may notice that using the above mentioned methods for examining the “bottomly” dataset we have found a relatively large proportion of common genes. None of the methods has many genes in common with EBSeq, which is due to the fact that this method has identified the smallest number of DEGs (see Figure 1), whereas at least 98-99% of the genes detected by EBSeq were usually detected by the other methods. For the ‘wang’ dataset percentages have lover values, which means that the methods found more various DEGs (see Figure 1). The highest percentage of common genes in this dataset is found in DESeq – EBSeq and edgeR – EBSeq because DESeq and edgeR have not found many more common differentially expressed genes with EBSeq (see Figure 1).

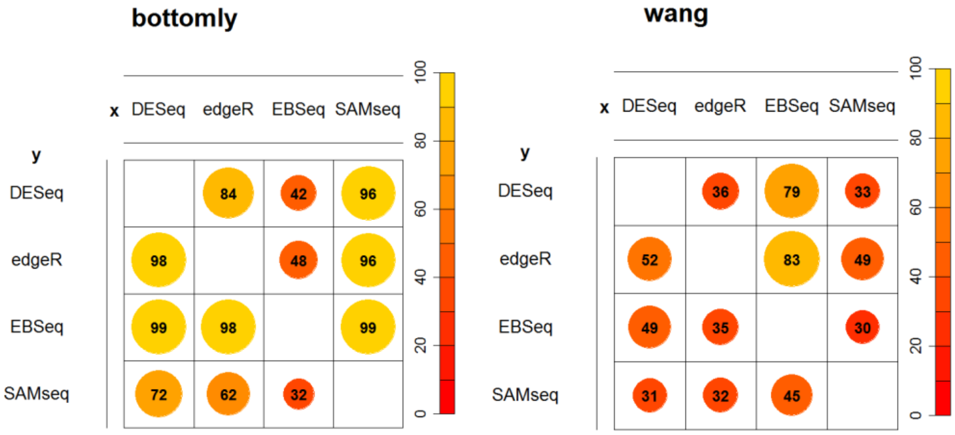
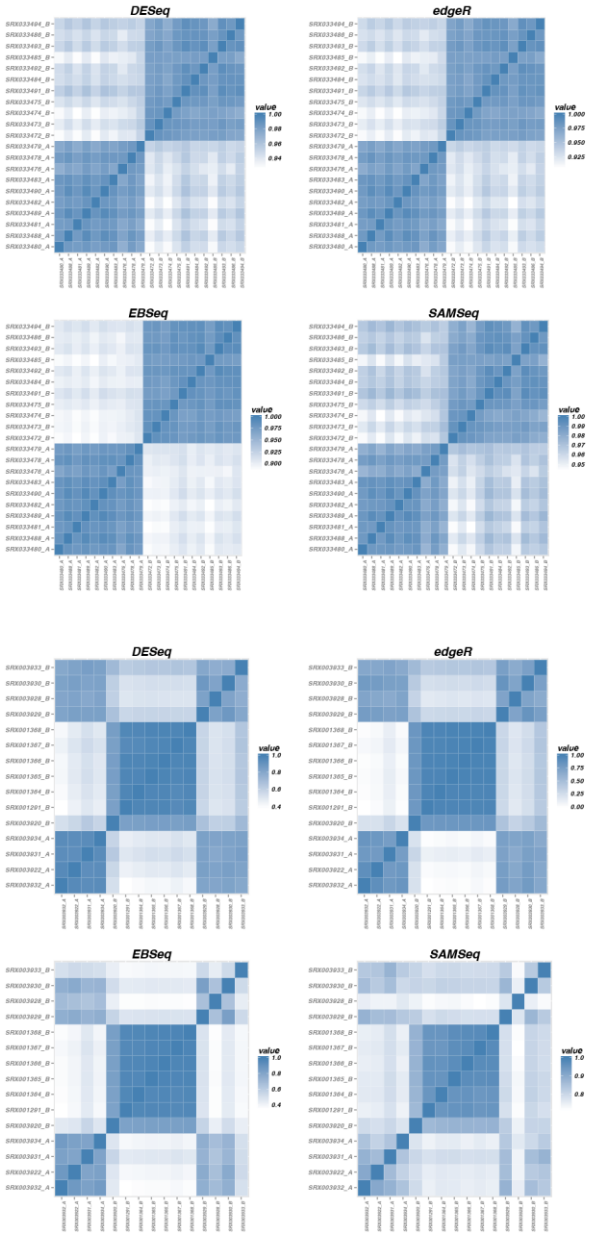


Fig. 5. Overlaps for DEG (‘bottomly’ – left, ‘wang’ – right)

The similarity of the samples in and between the classes may be presented using heatmaps. Color intensity corresponds to the value of the Spearman’s correlation coefficient. The results are presented in Figure 6.

a) ‘bottomly’



b) ‘wang’

Fig. 6. Heatmaps for DEGs between two conditions (‘bottomly’ (a), ‘wang’ (b) datasets). Each sample has additional label “A” and “B” corresponding to the group it belongs to

We can see from this Figure 6 that in the case of the 'bottomly' dataset, genes found by each method separate samples of different classes. On the other hand, in the 'wang' dataset there are four samples revealing a higher correlation to the samples from the other class than to the samples from the same one.

The effectiveness of classification based on differential methods is a way of comparing the results. ROC curves allow us to graphically present the similarity of the methods. Figure 7 shows ROC curves for the 'bottomly' and 'wang' datasets.

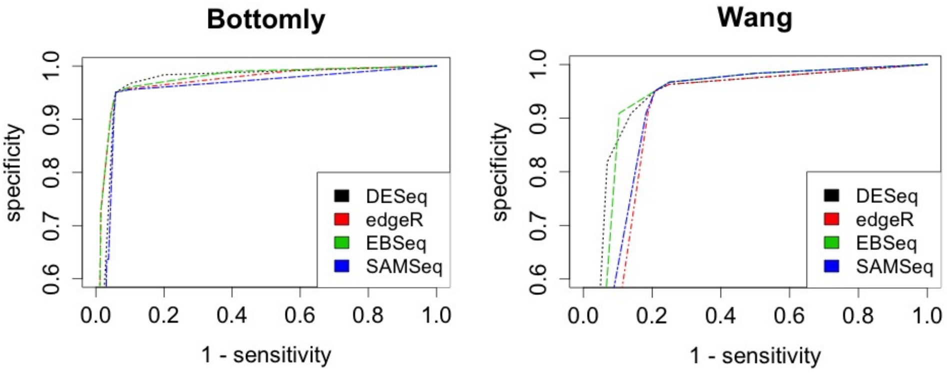


Fig. 7. Roc curves for 200 the most significant genes ('bottomly' – left, 'wang' – right datasets)

In the case of the 'bottomly' dataset the methods show small differences, whereas in the 'wang' dataset they reveal poor effectiveness of the edgeR method.

The last graphical presentation of the results is the dendrogram, based on common DEG ranks. In Figure 8 we show the dendrograms for the 'bodymap' and 'wang' datasets.

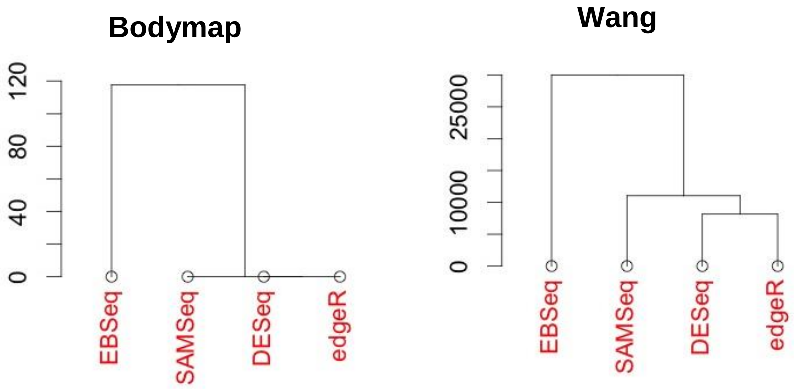


Fig. 8. Dendrograms based on common DEG ranks ('bodymap' – left, 'wang' – right datasets)

In both datasets we observe that EBSeq is the most distinct method, whereas the other methods are in the same branch. It is also worth pointing out that the scale in the 'bodymap' dataset is much lower than in the 'wang' dataset. It means that similarity of the methods in the 'bodymap' dataset is higher.

5. Discussion

In the paper we focused on eight graphical presentations of the results of a differential analysis. All of them allow for a quick interpretation and understanding of the data. By means of these visualizations we can verify if a dataset makes the results comparable between different methods. As we deduce from the presented visualizations, the 'bottomly' dataset reveals a close resemblance of the results for the four statistical methods used in the analysis. The proposed graphical presentations enables us to compare the results of the four R packages. If we only want to match the number of DEGs identified by each method and the overlap between the sets of DEGs, we use the Venn diagram. Unfortunately, this approach has limited applicability as the diagram would be illegible if we used it to compare several methods. We can also visualize these sets by MA-plots with clearly marked unique genes. MA-plots show the relationship between the base mean of the counts for each gene and the log fold change between the considered classes and can reveal artefacts between these values. In a perfect situation, the intensity of the base mean values should be evenly distributed around zero across all intensity of the log fold change values. MA-plot is a very good techniques to provide information about the up-regulated and down-regulated genes. Boxplots are the most convenient way for summarizing data and presenting a large amount of information including the median, upper and lower quartile, minimum and maximum data value. These charts are very efficient in handing large datasets. They show outliers values which may be detected below and above the whiskers but do not provide any other information about distribution, such as a histogram that more resembles the probability density function. Considering the advantages and disadvantages of the graphical methods presented above, it can be concluded that they are complex and quick tools to determine distinctions between the results of a differential analysis. Additionally, the proposed methods can help to determine which method in a differential analysis produces the results lending support to the assumptions of the experiment.

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.

Asmann Y.W., Necela B.M., Kalari K.R., Hossain A., Baker T.R., Carr J.M., Davis C., Getz J.E., Hostetter G., Li X., McLaughlin S.A., Radisky D.C., Schroth G.P., Cunliffe H.E., Perez E.A., Thompson E.A. (2012). Detection of redundant fusion transcripts as biomarkers or disease – specific therapeutic targets in breast cancer. *Cancer Res*. 72 (8), 1921–8, doi: 10.1158/0008–5472.

Baron M.E. (1969). A note on the historical development of logic diagrams: Leibniz, Euler and Venn. *The Mathematical Gazette* 53(384), 113–125.

Bottomly D., Walter N.A.R., Hunter J.E., Darakjian P., Kawane S., Buck K.J., Searles R.P., Mooney M., McWeeney S.K., Hitzemann R. (2011). Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS One* 6(3), e17820, doi: 10.1371/journal.pone.0017820.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Jain A.K., Murty M.N., Flynn P.J. (1999). Data clustering: a review. *ACM Computing Surveys* 31(3), 264–323, doi: 10.1145/331499.331504.

Kelley W.M., Donnelly R.A. (2009). *The Humongous Book of Statistics Problems*. New York, NY, Alpha Books.

Kohavi R., Wolpert D. H. (1996). Bias plus variance decomposition for zero–one loss functions. In *Proceedings of the 13th International Conference on Machine Learning*, 275–283.

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.

Leng N., Dawson J., Thomson J.A., Ruotti V., Rissman A.I., Smits B.M.G., Haag J.D., Gould M.N., Stewart R.M., Kendziorski Ch. (2013). EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics,* doi: 10.1093/ bioinformatics/btt087.

Li J., Tibshirani R. (2011). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research* 22(5), 519–36, doi: 10.1177/0962280211428386.

McGill R., Tukey J.W., Larsen W.A. (1978). Variations of box plots. *The American Statistician* 32(1), 12–16.

Montgomery S.B., Sammeth M., Gutierrez-Arcelus M., Lach R.P., Ingle C., Nisbett J., Guigo R., Dermitzakis E.T. (2013). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464 (7289), doi:10.1038/nature08903.

Pickrell J.K., Marioni J.B., Pai A.A., Degner J.F., Engelhardt B.E., Nkadori E., Veyrieras J.B., Stephens M., Gilad Y., Pritchard J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464 (7289), 768–772, doi:10.1038/nature08872.

R Core Team (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. URL http://www.R-project.org/.

Robinson M., McCarthy D., 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.

Seyednasrollah F., Laiho A., Elo L.L. (2013). Comparison of software packages for detecting differential expression in RNA-Seq studies. *Brief. Bioinform.*, doi: 10.1093/bib/bbt086.

Sneath P.H.A. (1957). The application of computers to taxonomy. *Journal of General Microbiology* 17, 201–226, doi: 10.1099/00221287–17–1–201.

Wang E.T., Sandberg R., Luo S., Khrebtukova I., Zhang L., Mayr C., Kingsmore S.F., Schroth G.P., Burge C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456(7221), 470–476, doi:10.1038/nature07509.