BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **17**, No 5
Special issue with selected papers from the workshop
"Two Years Avitohol: Advanced High Performance Computing Applications 2017"

Sofia • 2017 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2017-0055

Comparison of RNA-Seq Differential Expression Methods

Dean Palejev

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria
E-mail: palejev@math.bas.bg

***Abstract*:** *There are many methods designed to find differentially expressed genes using RNA-seq data. Their outputs differ a lot, some genes are determined to be differentially expressed by most or all methods, and others – by very few or even by just one method. Here we derive a systematic approach to quantifying the proximity of such methods, allowing us also to discover patterns and to determine whether some of them are significantly different than others.*

***Keywords*:** *partially ranked lists, RNA-seq, high-throughput sequencing, NGS, differential expression.*

1. Introduction

The high-throughput sequencing technologies (also known as Next Generation Sequencing or NGS) that became available during the last decade allowed researchers to produce vast amounts of DNA-seq and RNA-seq data. One of the basic questions that could be answered by utilizing RNA-seq data is finding differentially expressed genes or transcripts, namely those who have different overall expression levels between two groups of interest (e.g., patients and healthy controls, or two different strains of species). There are many methods designed to find differentially expressed genes, with some of them producing more similar results than others. An example of this is shown in Fig. 1, showing a Venn diagram of the differentially expressed genes found by four methods: DESeq [1], DESeq2 [2], edgeR [3] and the limma implementation in the R package voom [4], using the same dataset [5] consisting of two groups of mice of different strains. The dataset is available from the ReCount online resource [6]. We can see that 570 genes were detected as significant by all four methods. At the same time one of them (DESeq) detected only nine extra genes, whereas another one (limma) detected extra 641 differentially expressed genes, including the extra nine detected by DESeq. Overall for this dataset, limma finds more than twice as many differentially expressed genes as DESeq, which raises the question whether there are significant differences between such methods.

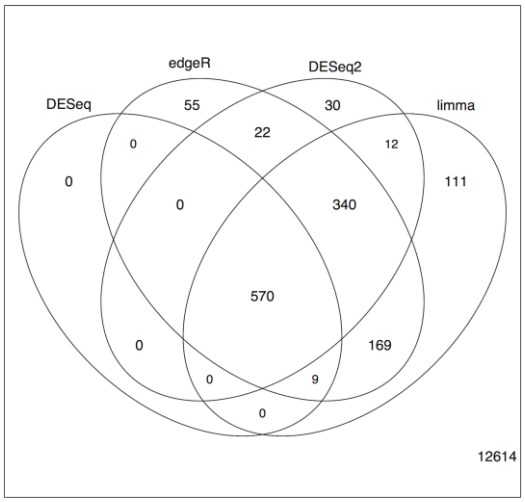

Fig. 1. Venn diagram showing the differentially expressed genes found by four different methods

There are many studies comparing RNA-seq differential expression methods, e.g., [7-9]. Often the comparisons are based on empirical considerations: FDR, empirical power, ROC curves (true positive rate vs. false positive rate), sensitivity and specificity, counts of genes detected as significant by different methods as a function of the sample size or sequencing depth, comparison of normalization methods, etc. In many cases, the comparison is often based on splitting the outputs into two groups – differentially expressed (or significant) genes and the rest. This is not optimal, as we would prefer to take into account the level of significance, with the most significant genes having the strongest candidates for subsequent biological verification and further investigation. At the same time, we are interested in the genes that are slightly above the desired significant level, because truly significant genes may have somehow larger *p*-values due to small sample size.

Because of that there is an added value in splitting the differentially expressed genes into several groups with similar *p*-values within each group, with each group considered distinct from the others, effectively creating partially ranked lists. Using such classification incorporates most of the useful information from the methods *p*-values, effectively mimicking and enhancing the way these methods are used by bioinformaticians.

2. RNA-seq data simulation

The R package compcodeR [7] was used in order to generate the in-silico data necessary for the study. The simulations were done on the High Performance Computing (HPC) complex Avitohol [10].

The typical RNA-seq dataset is a table of read counts, where each row represents a gene or transcript and each column represents a sample. Each entry shows how many RNA-seq reads from that particular sample were mapped to that particular gene or transcript.

Using compcodeR we generate read counts data for two populations, with 13932 genes, varying the sample sizes (values of 2, 5, 10, 20 for each sample) and the proportion of differentially expressed genes (0.01, 0.05, 0.1, 0.2), while keeping the other settings to their default values (e.g., the proportion of upregulated genes out of all differentially expressed genes is 0.5). The number of genes used here (13,932) is equal to the number of non-all-zero rows in the dataset from [5].

For each pair of parameters (sample size and proportion differentially expressed genes), using compcodeR, we generate 10,000 sets of 13,932 genes with these parameters of the whole dataset. After that for each of the 10,000 sets we apply each of the following methods or combinations of methods (for simplicity, further we will call them methods):

1. DESeq.GLM (DESeq with option GLM)
2. DESeq.nbinom (DESeq with option nbinom)
3. DESeq2
4. DSS [11]
5. NBPSeq [12]
6. TCC [13]
7. edgeR.exact (edgeR with option exact)
8. edgeR.GLM (edgeR with option GLM)
9. logcpm.limma (limma after performing a log transformation on the counts per million)
10. sqrtcpm.limma (limma after performing a square root transformation on the counts per million)
11. (ordinary) *t*-test
12. voom.limma (applying the voom transformation and then a differential expression test with limma)
13. voom.ttest (applying the voom transformation followed by *t*-test)
14. vst.limma (applying the variance-stabilizing transformation from DESeq and the differential expression test with limma)
15. vst.ttest (applying the variance-stabilizing transformation from DESeq and *t*-test to determine the differentially expressed genes).

The typical workflow of applying each of these methods results in a set of *p*-values, one for each gene, which are ultimately adjusted for multiple comparisons using the Benjamini-Hochberg procedure [14]. Genes with adjusted *p*-values not greater than a particular threshold (typically 0.05) are determined to be differentially expressed by the method. Therefore in a typical usage setup, all the information that such method yields is included in the resulting adjusted *p*-values.

We are interested in determining a measure of the proximity of the 15 differentially expression methods listed above. For a given read count dataset, and for each pair of methods, we can calculate a Chebyshev-type distance (such as the one defined in [15]) between the respectively adjusted *p*-values that were produced by each of the two methods. The cutoff points used to split the interval [0, 1] into subintervals were: 0.0001, 0.001, 0.01, 0.05 and 0.1. Given the standard significance cutoff of 0.05, the interval from it to 0.1 represents the tests (genes) for which there is some, but not enough, evidence of being significant and we would

prefer to distinguish them from the clearly non-significant tests with Benjamini-Hochberg adjusted *p*-values. The practice shows that in some cases the significant genes determined by one method end up having adjusted *p*-values in that interval. For smaller sample sizes, it is also common for the truly significant genes to have adjusted *p*-values in that interval. As mentioned above, there is an added value in splitting the adjusted *p*-values into several intervals depending on their level of significance.

The set of adjusted *p*-values produced by each method does not have a uniform distribution. In fact, as shown in [16] the distribution of the unadjusted *p*-values has a distribution that is a mix between a distribution skewed towards 0 (corresponding to the genes determined significant by the method) and uniform distribution (corresponding to the non-significant genes). The distribution of the adjusted *p*-values has a small peak close to 0, another one at 1 and very few values in between. Thankfully the right-invariant property described in [15] allows us to use the distances between the *p*-values.

3. Results

For each pair of sample size and proportion of differentially expressed genes we perform hierarchical clustering using the distance mentioned above. For smaller values of both parameters, all methods are relatively close as shown in Fig. 2.

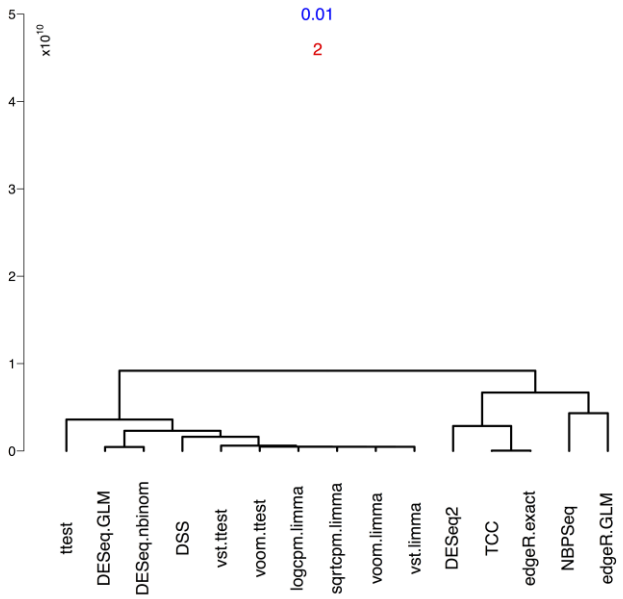

Fig. 2. Hierarchical clustering for sample size = 2 and proportion of differentially expressed genes = 0.01

This could be because for small sample size all reasonable estimates are close, even after different corrections or normalization. When either parameter increases we start seeing larger differences between some of the methods, e.g., as shown on Fig. 3 (sample size = 20, proportion of differentially expressed genes = 0.2).

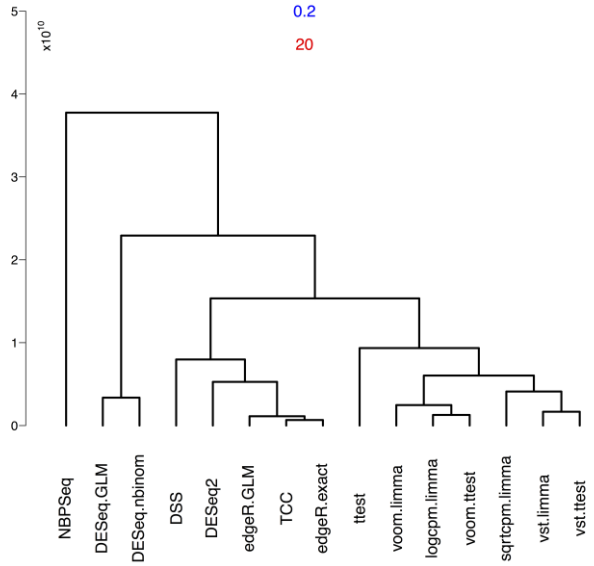

Fig. 3. Hierarchical clustering for sample size = 20 and proportion of differentially expressed genes = 0.2

In this case NBPseq differs a lot from the other methods. All methods that incorporate limma and *t*-test cluster are closely together. We can also see that the two submethods of DESeq (.nbinom and .GLM) are relatively close to each other, however for the smaller values of the parameters they are close to all other methods, and for the larger values of the parameters they are somehow further from the majority of the other methods. Due to space constrains we cannot show the dendrograms for all pairs of parameters. These dendrograms together with other plots are shown at the article webpage **http://www.math.bas.bg/~palejev/RNA-seq_comparison**. From the extra plots available on that webpage we can see that the limma and *t*-test based methods are generally close to each other.

Further we investigate whether any of these methods are significantly different. We approximate the null distribution described in [15] not by a chi-square distribution, but by a Gamma distribution in order to allow some more flexibility by having two parameters (although in this particular case both distributions are close and the approximation choice does not affect the results). Then for each pair of methods and a given set of parameters we compare the empirical distribution of the distance between the Benjamini-Hochberg-adjusted *p*-values produced by these methods with the Gamma distribution. One example is shown in Fig. 4. In it we first visualise boxplot of the null Gamma distribution and then for each pair of

parameters, a boxplot of the distribution of the distances between DESeq.nbinom and DESeq2 for that pair of parameters. The *x*-axis labels for the distribution boxplots, except for the Gamma, show the proportion of differentially expressed genes and then the sample size.

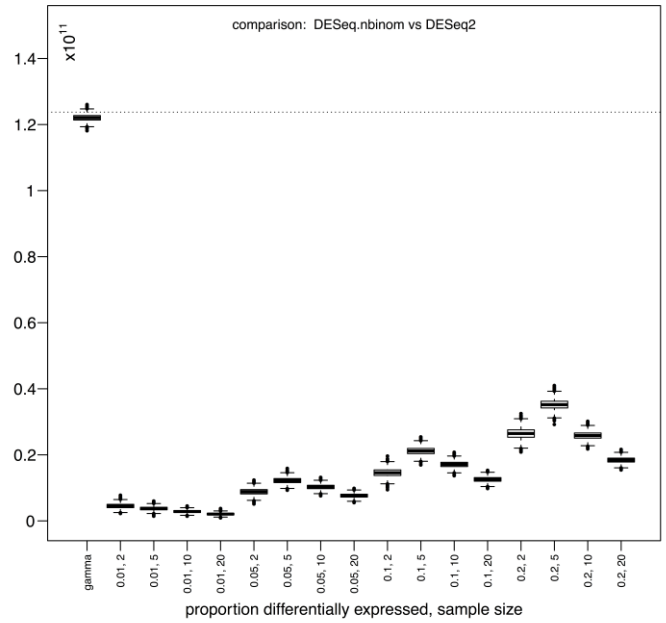

Fig. 4. Comparison of DESeq.nbinom and DESeq2 for all values of the parameters together with the null Gamma distribution

In this example all of the distances between these two methods are well below the null distribution, therefore there is no evidence that these two methods are significantly different. A pattern in this case, that appears in many other comparisons is that for a given proportion of differentially expressed genes, the methods differ the most for sample size of 5. In this case this is true for proportions of 0.05, 0.1 and 0.2.

Due to space constrains we cannot display all of the results here. All of the graphs are shown on the article webpage, and they show that that for all pairs of methods, and all of the considered parameter values, the distances distributions are well below the null distribution. Therefore for all pairs of methods, the differences do not appear statistically significant.

4. Conclusion

Here different methods are shown for finding differentially expressed genes might result in very different outputs for the same dataset. We also show the usefulness of defining a distance between such methods on a particular dataset by using the Spearman's distance on partially ranked lists of Benjamini-Hochberg-adjusted

p-values. Finally, we demonstrate that for reasonable values of the number of genes, the sample sizes and the proportion of differentially expressed genes, although there are large differences in outputs of these methods, they do not appear to be statistically significant.

***Acknowledgments*:** The author acknowledges the support of the Bulgarian National Science Fund grant I02/19.

References

1. Anders, S., W. Huber. Differential Expression Analysis for Sequence Count Data. – Genome Biology, Vol. **11**, 2010, R106.
https://doi.org/10.1186/gb-2010-11-10-r106
2. Love, M. I., W. Huber, S. Anders. Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2. – Genome Biology, Vol. **12**, 2014, No 12, 550.
https://doi.org/10.1186/s13059-014-0550-8
3. Robinson, M. D., D. J. Mccarthy, G. K. Smyth. EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. – Bioinformatics, Vol. **26**, 2010, No 1, 149-140.
https://doi.org/10.1093/bioinformatics/btp616
4. Law, C. W., Y. Chen, W. Shi, G. Smyth. Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-seq Read Counts. – Genome Biology, Vol. **15**, 2014, R29.
https://doi.org/10.1186/gb-2014-15-2-r29
5. Bottomly, D., N. A. R. Walter, J. E. Hunter et al. Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays. – PLoS ONE, Vol. **6**, 2011, No 3, e17820.
https://doi.org/10.1371/journal.pone.0017820
6. Frazee, A. C., B. Langmead, J. T. Leak. ReCount: A Multi-Experiment Resource of Analysis-Ready RNA-Seq Gene Count Datasets. – BMC Bioinformatics, Vol. **12**, 2011, 449.
https://doi.org/10.1186/1471-2105-12-449
7. Soneson, C., M. Delorenzi. A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data. – BMC Bioinformatics, Vol. **14**, 2013, 91.
https://doi.org/10.1186/1471-2105-14-91
8. Conesa, A., P. Madrigal, S. Tarazona et al. A Survey of Best Practices for RNA-Seq Data Analysis. – Genome Biology, Vol. **17**, 2016, 13.
https://doi.org/10.1186/s13059-016-0881-8
9. Rapaport, F., R. Khanin, Y. Liang et al. Comprehensive Evaluation of Differential Gene Expression Analysis Methods for RNA-Seq Data – Genome Biology, Vol. **19**, 2013, No 9, R95.
https://doi.org/10.1186/gb-2013-14-9-r95
10. Atanassov, A., T. Gurov, A. Karaivanova et al. On the Parallelization Approaches for Intel MIC Architecture. – AIP Conference Proceedings, Vol. **1773**, 2016, No 070001.
https://doi.org/10.1063/1.4964983
11. Wu, H., C. Wang, Z. Wu. A New Shrinkage Estimator for Dispersion Improves Differential Expression Detection in RNA-Seq Data. – Biostatistics, Vol. **14**, 2013, No 2, 232-43.
https://doi.org/10.1093/biostatistics/kxs033
12. Di, Y., D. Schafer, J. S. Cumbie, J. H. Chang. The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. – Statistical Applications in Genetics and Molecular Biology, Vol. **10**, 2011, No 1.
https://doi.org/10.2202/1544-6115.1637
13. Sun, J., T. Nishiyama, K. Shimizu, K. Kadota. TCC: An R Package for Comparing Tag Count Data with Robust Normalization Strategies. – BMC Bioinformatics, Vol. **14**, 2013, 219.

https://doi.org/10.1186/1471-2105-14-219

14. B e n j a m i n i, Y., Y. H o h b e r g. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. – Journal of the Royal Statistical Society. Series B (Methodological), Vol. **59**, 1995, No 1, pp. 289-300.
15. S t o i m e n o v a, E. Comparison of Partially Ranked Lists. – Austrian Journal of Statistics, Vol. **46**, 2014, No 3-4, pp. 107-115.
https://doi.org/10.17713/ajs.v46i3-4.676
16. F e r g u s o n, J. P., D. P a l e j e v. p-Value Calibration for Multiple Testing Problems in Genomics – Statistical Applications in Genetics and Molecular Biology, Vol. **13**, 2014, No 6, pp. 659-673.
https://doi.org/10.1515/sagmb-2013-0074