Differential Analysis for RNA-seq using Intensive Randomization

Dorota Desaulle, StatOmique 2022, the 25th October 2022

Names and affiliations

Dorota Desaulle, MCF, UR 7537 BioSTM - Biostatistique, Traitement et Modélisation des données biologiques, Faculté of Pharmacie, Université de Paris Cité, France (principal investigator - speaker)

email : dorota.desaulle@u-paris.fr

Yves Rozenholc, PU, UR 7537 BioSTM - Biostatistique, Traitement et Modélisation des données biologiques, Faculté de Pharmacie, Université de Paris Cité, France (BioSTM team leader)

Céline Hoffmann, PhD, CNRS Researcher at CNRS UMR 8258-INSERM U1267, Faculté de Pharmacie, Université de Paris Cité, France (collaboration partner)

Abstract

The RNA-seq data aims to quantify the transcriptome of biological samples. The sequencing pipeline provides a quantification of RNA fragments in terms of read counts. Among other applications, they enable comparisons of genes expressions between different experimental conditions within a differential analysis or an identification of biological patterns by means of classification tools. These data are subject to sample specific systematic biases resumed as so-called scaling factors. Most existing methods consider them as nuisance model parameters and estimate them together with gene specific effect within the model. In other approaches, counts data are normalized first and the downstream analysis is conducted on the normalized measures next.

Despite methodological advances, the optimal approach to normalize RNA-seq data has not reached a consensus to date (Abrams et al. 2019). Recently, we have proposed a novel statistical framework for differential analysis in transcriptomics (Desaulle et al. 2021) supported by a strong theoretical basis.

Our concept is related to the housekeeping genes normalization. However, since the latter are not always unknown, we propose an new iterative procedure. In each iteration, few randomly selected genes are used as a reference in the normalization step and a differential analysis is conducted on the remaining genes. Finally, the detection of differential expressions (DE) is obtained by combining the results from all iterations. This intensive procedure is proven to provide good control of the statistical errors and has been implemented in the R package DArand (Desaulle and Rozenholc 2021) and is publicly available from the Comprehensive R Archive Network.

We compare our results with DESeq2 (Love, Huber, and Anders 2014) and edgeR (Robinson, McCarthy, and Smyth 2010) methods widely chosen in practice. Under the common assumption that most genes are not DE, both are able to control the false positive rates while maintaning the power (Dillies et al. 2013) however in simulations our new method shows promising results when the number of DE increases.

Presentation time

20-30 minutes

References

Abrams, Zachary B., Travis S. Johnson, Kun Huang, Philip R. O. Payne, and Kevin Coombes. 2019. "A Protocol to Evaluate RNA Sequencing Normalization Methods." *BMC Bioinformatics* 20 (24): 679.

https://doi.org/10.1186/s12859-019-3247-x.

Desaulle, Dorota, Céline Hoffmann, Bernard Hainque, and Yves Rozenholc. 2021. "Differential Analysis in Transcriptomic: The Strength of Randomly Picking 'Reference' Genes." https://arxiv.org/abs/2103.09872.

Desaulle, Dorota, and Yves Rozenholc. 2021. *DArand: Differential Analysis with Random Reference Genes*. https://CRAN.R-project.org/package=DArand.

Dillies, Marie-Agnès, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, et al. 2013. "A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis." *Brief Bioinform* 14 (6): 671–83. https://doi.org/10.1093/bib/bbs046.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15: 550. https://doi.org/10.1186/s13059-014-0550-8.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.