Main features and evaluation of different workflows for differential gene expression studies through RNA-Seq samples.

A. Pérez Sala[1,2], I.M. Larrayoz Roldán[1]

[1]Biomarkers and Molecular Signaling, Center for biomedical research of La Rioja
(aperez.ext,ilarrayoz)@riojasalud.es
[2]Masters on Biomedical Engineering, Master and Doctoral School, University of the Basque Country

1. Introduction and objectives

Recently the number of researchers using high throughput sequencing techniques in their studies has grown significantly, and high throughput RNA-Seq has validated its utility across a wide set of experiments. Moreover, differential gene expression analysis has been one of the most important procedures to analyse mRNA expression level between different conditions.

RNA-Seq data analysis require specific bioinformatic processes depending on the experimental paradigm. This study focuses on the bioinformatic protocols for differential gene expression quantification between different conditions. Differential expression analysis workflows have evolved including new features, that aim to improve the specificity and sensibility of these analyses. Also, these workflows not only improved by giving more accurate results, but also have incorporated new algorithms reducing computational requirements. The analysis of strengths and weakness between different protocols could be interesting for the scientific community.

2. Materials

Already described protocols like "tuxedo" [1](TopHat, Cufflinks and CummeRbund) and "new tuxedo" [2](HISAT,StringTie and Ballgown) have been chosen in addition to protocols composed by general purpose programs [3] (STAR [4], FeatureCounts [5], and EdgeR [6] / DeSeq [7]) for this study. FeatureCounts, DeSeq and EdgeR have been combined with data generated by all aligners to study variations on the results through the different tools.

3. Methods

Many features of these three different analysis workflows has been evaluated across the three main protocol steps looking for differences, advantages and disadvantages of each one. Meanwhile we have cross combined the three main protocol steps between them to compare the combined performance versus the original workflows.

4. Results and discussion

After aligners evaluation, STAR got the best unique mapped read ratio compared with TopHat and HiSat. FeatureCounts was the quickest in the second step, and provides a simple and easy to understand dataset. Finally, EdgeR, using FeatureCounts, presents advantages in flexibility in the experimental design, reliability of statistical models using empiric bayes methods, and normalization methods (TMM, Trimmed mean of m-values).

To sum up, we propose STAR, FeatureCounts and EdgeR workflow as a low computational requirement and reliable protocol in differential gene expression analysis with high experimental design flexibility.

5. Thanks

This work was supported by Instituto de Salud Carlos III CP15/00198, Fondo Social Europeo (Periodo de Programación 2014 – 2020) and Fondo Europeo de Desarrollo Regional granted to IML.

6. References

[1] C. Trapnell *et al.*, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.," *Nat. Protoc.*, vol. 7, no. 3, pp. 562–78, 2012.

[2] M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, and S. L. Salzberg, "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown," *Nat Protoc.*, vol. 11, no. 9, pp. 1650–1667, 2016.

[3] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.," *Genome Biol.*, vol. 15, no. 12, p. 550, 2014.

[4] A. Dobin *et al.*, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013.

[5] Y. Liao, G. K. Smyth, and W. Shi, "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, vol. 30, no. 7, pp. 923–930, Apr. 2014.

[6] 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, no. 1, pp. 139–140, Jan. 2010.

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