Published online 16 July 2015

Nucleic Acids Research, 2015, Vol. 43, No. 21 ***e140***
doi: 10.1093/nar/gkv711

Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package

Sonia Tarazona[1,2], Pedro Furió-Tarí[1], David Turrà[3], Antonio Di Pietro[3], María José Nueda[4], Alberto Ferrer[2] and Ana Conesa[1,5,*]

[1]Genomics of Gene Expression Lab, Centro de Investigación Príncipe Felipe, Eduardo Primo Yúfera 3, 46012, Valencia, Spain, [2]Department of Applied Statistics, Operations Research and Quality, Universidad Politécnica de Valencia, Camí de Vera, 46022, Valencia, Spain, [3]Department of Genetics, Universidad de Córdoba, Campus de Rabanales Edificio Gregor Mendel, 14071, Córdoba, Spain, [4]Statistics and Operational Research Department, Universidad de Alicante, Carretera San Vicente del Raspeig s/n, 03690, Alicante, Spain and [5]Microbiology and Cell Science Department, Institute for Food and Agricultural Sciences, University of Florida, Gainesville, FL 32603, USA

Received April 21, 2015; Revised June 12, 2015; Accepted July 01, 2015

ABSTRACT

As the use of RNA-seq has popularized, there is an increasing consciousness of the importance of experimental design, bias removal, accurate quantification and control of false positives for proper data analysis. We introduce the NOISeq R-package for quality control and analysis of count data. We show how the available diagnostic tools can be used to monitor quality issues, make pre-processing decisions and improve analysis. We demonstrate that the nonparametric NOISeqBIO efficiently controls false discoveries in experiments with biological replication and outperforms state-of-the-art methods. NOISeq is a comprehensive resource that meets current needs for robust data-aware analysis of RNA-seq differential expression.

One of the most wide-spread applications of RNA-seq is the transcript quantification and the differential gene expression analysis (3,4). It has been claimed that RNA-seq has a number of advantages over its predecessors (arrays), such as a wider dynamic range of measurements (5), the capacity to detect transcripts with low expression level (3) and the ability to identify differences in isoform or allele expression (6,7). RNA-seq was initially described as highly reproducible, and it was claimed to provide more 'direct' and reliable gene expression measurements (3), but it is now generally accepted that it also has limitations which make it far from perfect. Although the high reproducibility of the technology reduces the need of technical replication (3,4), the precision at the low expression level is still limited (4,8) and, nonetheless, sufficient biological replicates are needed to adequately infer properties about the population (9,10). Therefore, the number of replicates and the sequencing depth at which one should sample remain important considerations when designing an RNA-seq exper-