Differential expression in RNA-seq: A matter of depth

Sonia Tarazona¹,², Fernando García-Alcalde¹, Joaquín Dopazo¹, Alberto Ferrer², Ana Conesa¹

¹ Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia
² Department of Applied Statistics, Operations Research and Quality, Univ. Politécnica Valencia

Aim

RNA-seq technology is increasingly being used for gene expression profiling. However, the properties of RNA-seq data have not been yet fully established, and additional research is needed for understanding how these data respond to differential expression analysis. In this work, we study how the sequencing depth affects the detection of transcripts and their identification as differentially expressed. We evaluate some differential expression algorithms and propose a novel approach, NOISeq. Our results reveal that most existing methodologies suffer from a strong dependency on sequencing depth for their differential expression calls, that results in an increasing number of false positives as the number of reads grows. NOISeq models the noise distribution from the actual data, so it can better adapt to the size of the data set, and is more effective in controlling the rate of false discoveries.

Fig. 1 SEQUENCING DEPTH. Number of genes with more than 5 reads for several sequencing depths in three different public datasets [1,2,3]. The more sequenced the more detected. No plateau reached, even with 200 million reads.

Fig. 2 BIOTYPE DETECTION Percentage of gene biotype in the brain sample (dataset [2]) at different sequencing depths. The distribution of biotypes observed among detected features evolve with increasing sequencing depth, with the relative abundance of protein-coding transcripts steadily decreasing, whereas noncoding genes gain a proportional presence.

Fig. 3 M-D values in noise (black dots) and for differentially expressed genes (red dots).

Fig. 4 Precision-recall curves (A) and false discovery rates (B) for the differential expression methods compared on data set [2] using RT-PCR results as a gold-standard. NOISeq outperforms the other methods (A) while keeping a low false discovery rate (B).

FET method is Fisher’s Exact test. “nlc” means “no length correction” and “RPKM” is Reads Per Kilobase and per Million Reads [4].

Conclusions

The sequencing depth affects the detection and expression quantification of the transcripts.

As more sequencing output is considered, the diversity and quantity of detected off-target RNA species increase.

NOISeq method shows a good performance when comparing it to other differential expression methods: Fisher’s Exact Test (FET), edger [5], DESeq [6] and baySeq [7].

NOISeq and FET are more robust to sequencing depth than the other methods: the number of differentially expressed genes keeps similar at increasing sequencing depths.

NOISeq maintains good True Positive and False Positive rates when increasing sequencing depth, while FET shows a poor detection rate and the other methods generate an increasing number of False Positives.

References