NOIseq: a RNA-seq differential expression method robust for sequencing depth biases

Sonia Tarazona, Fernando García, Alberto Ferrer, Joaquín Dopazo, Ana Conesa
Genomics of Gene Expression Lab, Centro de Investigaciones Príncipe Felipe, Valencia, Spain
http://bioinfo.cipf.es/aconesa

Introduction

Next Generation Sequencing (NGS) technologies have brought a revolution to research in genome and genome regulation. One of the most breaking applications of NGS is in transcriptome analysis. RNA-seq has revealed exciting new data on gene models, alternative splicing and extra-genic expression. Also RNA-seq permits the quantification of gene expression across a large dynamic range and with more reproducibility than microarrays. Several methods for the assessment of differential expression from count data have been proposed but biases associated to transcript length and transcript frequency distributions have been reported. It is still not clear how much sequencing reads should be generated in a RNA-seq experiment to obtain reliable results and what's exactly being detected. In general we observed that many RNA-seq datasets have not reached saturation for detection of expressed genes and that the relative proportion of different transcript biotypes changes with increasing sequencing depth. In this work we investigate the effect that library size has on the assessment of differential expression on different aspects of the selected genes. We show that current statistical methods suffer from a strong dependency of their significant calls on the number of mapped reads considered and proposed a novel differential expression methodology – **NOISeq**[1]- that is robust to the amount of reads.

Results

NOISeq is a non-parametric approach for the differential expression analysis of RNseq-data. NOISeq creates a null or noise distribution of count changes by comparing the number of reads of each gene in samples within the same condition. This reference distribution is then used to assess whether the change in count number between two conditions for a given gene is likely to be part of the noise or represents a true differential expression. Two variants of the method are implemented: NOISeq-real uses replicates, when available, to compute the noise distribution and, NOISeq-sim simulates them in absence of replication. We compared our method with edgeR2, DESeq3, baySeq4 and Fisher Exact Test (FET) using three different experimental datasets. Results are presented for MAQC experiment where the transcriptome of brain and Universal Human Reference (HUR) samples were sequenced at about 45 million Solexa reads each.

We first determined that although protein-coding gene is the most abundant transcript type within differential expression calls for all methodologies, other RNA types, such as processed-transcript, pseudogenes and lincRNAs are readily detected. NOISeq dected comparatively more protein-coding genes than other methods that called significant a considerable number of non-coding and small RNA transcripts. Additionally, all comparing methods except FET greatly increased the number of detected (non-coding) genes as sequencing depth raised while NOISeq showed a constant pattern. Also these other methods tend to select shorter genes and smaller fold change differences with the increasing amounts of reads. In general, parametric approaches selected much more genes than NOISeq, specially at high sequencing depth rates. When analyzing the functional content of these genes by functional enrichment analysis, we observed that the pool of genes detected both by NOISeq and the parametric methods where highly enriched in functional categories, while genes selected only by parametric methods did not. To check whether this differences were indicative of different false calls between methods, we used the RT-PCR data available at the MAQC project that contains 330 true positive and 83 true negative differentially expressed genes. Performance plots indicate that edgeR, DESeq, baySeq strongly increased the number of false calls with sequencing depth, while NOISeq was constant and low. On the contrary true discoveries were slightly better for these methods, presumably consequence of their large number of selected genes. FET showed in low false and true discovery rates, due to its general lower detection power.

Conclusions

We showed that most current RNA-seq statistical analysis methods fail to control the number of false discoveries as the size of the sequenced library increases. These false positive are mainly short, non-coding genes and/or genes with small fold changes. NOISeq, but adopting an empirical approach to model the null distribution of differential expression captures better the shape of noise in RNA-seq data, resulting in a sequencing-depth robust method for differential expression analysis.

References

1. Tarazona S., Garcia-Alcalde F., Ferrer A., Dopazo J., Conesa, A. Differential expression in RNA-seq:a matter of depth. Genome Research, Sep 2011, doi:10.1101/gr.124321.11.
2. Robinson, MD, McCarthy, DJ, and Smyth, GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139,140.
3. Anders, S and Huber, W. 2010. Differential expression analysis for sequence count data. Genome Biology 11(10):R106.
4. Hardcastle, T and Kelly, K. 2010. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11(1):422+.

Relevant Web sites

5. http://bioinfo.cipf.es/noiseq