RNA sequencing data: biases and normalization

F. Finotello[1] ✉, E. Lavezzo[2], L. Barzon[2], P. Fontana[3], A. Si-Ammour[3], S. Toppo[2], B. Di Camillo[1]

[1]Department of Information Engineering, University of Padova, Italy
[2]Department of Molecular Medicine, University of Padova, Italy
[3]Edmund Mach Foundation, San Michele all'Adige, Trento, Italy

Motivations

In recent years, RNA sequencing (RNA-seq) has rapidly become the method of choice for measuring and comparing gene transcription levels. Despite its wide application, it is now clear that this methodology is not free from biases and that a careful normalization procedure is the basis for a correct data interpretation. The most common normalization techniques account for: library size, gene or transcript length and sequence-specific biases such as GC-content effects. The aim of the present work is to investigate biases affecting RNA seq data and their effect on differential expression analysis. In order to reduce biases due to over-simplification of gene transcription models, we consider exon-based counts.

Methods

We two used publicly available RNA-seq data sets from two-group comparison studies which are characterized by multiple technical replicates. We summarized read counts at exon level and investigated their dependence on sequence-specific covariates: GC-content and exon length. In addition, we considered the effect of library size correction on between-groups comparison and the impact of the above mentioned biases on the detection of differentially expressed exons. The assessment is performed on raw data, as well as on data normalized with different approaches: RPKM [1], library size scaling, based on Trimmed Mean of M-values (TMM) [2] and on Poisson goodness-of-fit statistic applied to non differentially expressed genes [3], and within-lane normalization based on loess regression of log-counts on GC-content and exon length [4]. We selected differentially expressed exons using the GLM-based version of edgeR [5] as it can consider an "offset" matrix which codifies counts normalization, that can be computed with the desired approach, and library size scaling factors specified by the user.

Results

In our study, read counts show a significant dependence on exon length and a moderate dependence on GC-content. Exon length bias also affects differential expression analysis: longer exons tend to have lower P-values and to be selected as differentially expressed more frequently than shorter exons. The tested normalization techniques do not completely remove biases and, in particular, RPKM approach over-corrects for exon length bias. Moreover, the choice of the strategy for library size adjustment has a great impact on the direction of the detected differential expression. The results obtained on these data sets demonstrate that RNA-seq data normalization is still an open issue. Further efforts should be directed towards the clarification of the relationship between read counts and sequence-specific biases, which are, in turn, correlated to each other, and the definition of new models for their correction.

References

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature methods. 2008;5(7):621-8.
2. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.
3. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics. 2011.
4. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-seq data. BMC Bioinformatics. 2011;12(1):480.
5. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. Nucleic Acids Res. 2012.