RNA-Seq as a Relative Abundance Measure: opportunities afforded by a compositional analysis framework.

D.D. LaRoche[1,2], D. Billheimer[1], S. Sinari[1], K. Michels[2], and B.J. LaFleur[2]
[1]University of Arizona Mel and Enid Zuckerman College of Public Health, Tucson, Arizona, USA;
dlaroche@email.arizona.edu
[2]HTG Molecular Diagnostics, Tucson, Arizona, USA

Abstract

The rapid rise in the use of RNA sequencing technology (RNA-seq) for scientific discovery has led to its consideration as a clinical diagnostic tool. However, as a new technology the analytical accuracy and reproducibility of RNA-seq must be established before it can realize its full clinical utility (SEQC/MAQC-III Consortium, 2014; VanKeuren-Jensen et al. 2014). We respond to the need for reliable diagnostics, quality control metrics and improved reproducibility of RNA-seq data by recognizing and capitalizing on the relative frequency nature of RNA-Seq data.

Problems with sample quality, library preparation, or sequencing may result in a low number of reads allocated to a given sample within a sequencing run. We propose a method, based on outlier detection of Centered Log-Ratio (CLR) transformed counts, for objectively identifying problematic samples based on the total number of reads allocated to the sample.

Batch effects arising from differing laboratory conditions or operator differences have been identified as a problem in high-throughput measurement systems (Leek et al. 2010; Chen et al. 2011). Batch effects are typically identified with a hierarchical clustering (HC) method or principal components analysis (PCA). For both methods, the multivariate distance between the samples is visualized, either in a biplot for PCA or a dendrogram for HC, to check for the existence of clusters of samples related to batch. We show that CLR transformed RNA-Seq data is appropriate for evaluation in a PCA biplot and improves batch effect detection over current methods.

As RNA-Seq makes the transition from the research laboratory to the clinic there is a need for robust quality control metrics. The realization that RNA-Seq data are compositional opens the door to the existing body of theory and methods developed by John Aitchison (1986) and others. We show that the properties of compositional data can be leveraged to develop new metrics and improve existing methods.

References

Aitchison, J. (1986). *The statistical analysis of compositional data.* Chapman & Hall, Ltd. isbn: 0-412-28060-4.

Chen, Chao, Kay Grennan, Judith Badner, Dandan Zhang, Elliot Gershon, Li Jin, and Chunyu Liu. (2011). Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS ONE 6* (2). issn: 19326203. doi:10.1371/journal.pone.0017238.

Leek, Jeffrey T, Robert B Scharpf, HÂ´ ector Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael a Irizarry. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews Genetics 11* (10): 733â“739. issn: 1471-0056. doi:10.1038/nrg2825. arXiv: NIHMS150003.

SEQC/MAQC-III Consortium. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature biotechnology 32* (9): 903-14. issn: 1546-1696. doi:10.1038/nbt.2957

Van Keuren-Jensen, Kendall, Jonathan J Keats, and David W Craig. (2014). Bringing RNA-seq closer to the clinic. *Nature biotechnology 32* (9): 884â“5. issn: 1546-1696. doi:10.1038/nbt.3017.