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

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.

Normalization and standardization methods for RNA-Seq generally assume that the total number of reads assigned to a sample does not affect the observed relative frequencies of probes within an assay. This assumpion, known as Compositional Invariance, is an important property for RNA-Seq data which enables the comparison of samples with differing read depths. Violations of the invariance property can lead to spurious differential expression results, even after normalization. We develop a diagnostic method to identify violations of the Compositional Invariance property.

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.