Methods for Outlier Detection Using Relative Abundance in Targeted RNASeq Applications

Dominic LaRoche, Bonnie Lafluer, Shripad Sinari, Kurt Michels, and Dean Billhiemer

Monday, March 28, 2016

High-throughput RNA sequence (RNA-Seq) data generated from next-generation sequencers, such as those from Illumina and Roche, is generally modeled as count data arising from a Poisson or Negative-binomial distribution with various methods of estimating the parameters (Robinson and Smyth 2008, Anders and Huber 2010, Zhou et al. 2011, Law et al. 2014). However, none of these proposed models address an additional feature of RNA-Seq data. Specifically, high-throughput RNA-Seq instruments have a maximum number of reads available per run. For example, the Roche 454 GS Junior (tm) claims approximately 100,000 reads per run for shotgun sequencing and 70,000 reads per run for amplicon sequencing. These reads are distributed across all of the samples included in the run and imposes a total sum constraint on the data. This constraint cascades down to each sample which is, in turn, constrained by the total number of reads allocated to the sample.

The total sum constraint is common in biological sampling. For example, if a 1 ml sample of blood is taken this sample could be divided into several components such as plasma, red blood cells, white blood cells, and platelets. If the amount of any 1 component were to increase some other component (or all the other components) must decrease due to the fixed volume of the sample. Previous authors have identified the compositional nature of RNA-Seq data (Robinson and Smyth 2007, Anders and Huber 2010, Robinson and Oshlack 2010, Law et al. 2014, Lovell 2015). For example, Robinson and Smyth (2007) consider counts of RNA tags as relative abundances in their development of a model for estimating differential gene expression implemented in the Bioconductor package edgeR (Robinson et al. 2009). Similarly, Robinson and Oshlack (2010) explicitly acknowledge the mapped read constraint when developing their widely used Trimmed-Mean of M-values (TMM) normalization method for RNA-Seq data.

Ignoring the sum constraint can lead to unexpected results and erroneous inference (Pearson 1897, Aitchison 1986, Lovell et. al 2011). Despite the general acknowledgement that RNA-Seq data are compositional in nature, few researchers have extended the broad set of compositional data analysis theory and tools created by Aitchison (1986) for use in RNA-Seq analysis problems. We extend existing compositional data methodology to include statistical diagnostic tests for the identification of outliers and batch effects.