Methods for Outlier Detection Using Relative Abundance in Targeted RNASeq Applications

Dominic LaRoche, Bonnie LaFleur, Shripad Sinari, Kurt Michels, and Dean Billheimer

Monday, March 28, 2016

# Introduction

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 Consortium 2014, Van Keuren-Jensen et al. 2014). Recent studies evaluating RNA-seq have found generally high intra-platform and inter-platform congruence across multiple laboratories (Li et al. 2013, 't Hoen et al. 2013, SEQC/MAQC-III Consortium 2014). Despite these promising results, there is still a need to establish reliable diagnostics and quality control metrics for RNA-seq data. Accurately identifying batch effects, and differentiating these from true biological differences, will be necessary if smaller laboratories wish to utilize RNA-seq technology for clinical applications (cite). Moreover, the reliable identification and removal of poor quality data produced by RNA-seq pipelines has the potential to dramatically improve the analytical accuracy and reproducibility of RNA-seq data, thereby improving its clinical utility.

There are several software packages devoted to quality control of RNA-seq data (cite RNASeQC and RSeQC). However, none of these proposed models address a fundamental 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. The Illumina Mi-Seq, with shorters read lengths, is limited to 25 million reads per sequencing run. These reads are distributed across all of the samples included in a sequencing run and, therefore, impose a total sum constraint on the data. This constraint cascades down to each probe or tag within a 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 evidence 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 sample outliers and batch effects. Specifically, we utilize the total aligned reads for each sample to identify outlying samples and we utilize control samples to identify and control for batch effects.

# Methods

We begin with a brief introduction to compositional data, its properties, and some established analytical methods. Compositional data is defined as any data in which all elements are non-negative and sum to a constant. The positive sum constraint has some important consequences for many standard statistical methodologies including correlation and regression. Compositional data contain only relative information, i.e. the information about any individual component, or group of componenets, is relative to the other components and contain no information about the absolute value of the component. Potential problems associated with compositional data were identified as early as 1897 by Pearson. Despite the fact that compositional data naturally arises in a wide variety of scientific disciplines, a general method for analysis of compositional data was not developed until John Aitchison published his seminal book in 1986. Aitchison outlines some basic principles for compostional data analysis and provides some analysis tools for compositional data which conform to these principles.

For RNA-seq data, the total sum constraint is imposed by the sequencing technology. Since this total differs between platforms we will refer to the total number of available reads as . These reads are distributed among samples in a run such that: