Quality control metrics for extraction-free targeted RNA-Seq: methods afforded by a compositional framework.

D.D. LaRoche[1,2,*], D.D. Billheimer[1], K.. Michels[2], and B.L. LaFleur[2]
[1]University of Arizona, Tucson, Arizona, USA;
[2]HTG Molecular Diagnostics, Tucson, Arizona, USA
[*] *dlaroche@email.arizona.edu*

Abstract

We develop quality control diagnostics for targeted RNA-Seq using the theory of compositional data. Targeted sequencing using extraction-free sample preparatrion allows researchers to efficiently measure transcripts of interest for a particular disease by focusing sequencing efforts on a select subset of transcript targets from small sample volumes. However, extraction free technologies create the need for post-sequencing quality control metrics since poor quality samples, which would likely be removed after unsuccessful RNA extraction in extraction-based technologies, can still be sequenced. We capatilize on the relative frequency poperty of RNS-Seq data to identify poor qulaity samples, samples that violate the relative frequency expectation, and batch effects using only post-sequencing 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. Similarly, most RNA-Seq analyses assume that the relative frequencies of target read counts are not affected by the total number of reads allocated to the sample, a property known as Compositional Invariance. We develop a method for evaluating sequencing runs or experiements for violations of compositional invariance. Finally, batch effects arising from differing laboratory conditions or operator differences have been identified as a problem in high-throughput measurement systems. We show that CLR transformed RNA-Seq data is appropriate for evaluation in a PCA biplot and improves batch effect detection over current methods.

Key words: RNA-Seq, Transcriptome, Quality Control, Sequencing.