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

Dominic LaRoche[1,2*], Dean Billheimer[1], Kurt Michels[2], and Bonnie LaFleur[2]

[1] University of Arizona, Mel and Enid Zuckerman College of Public Health, Department of Biostatistics, [2]HTG Molecular Diagnostics, [*]Corresponding Author: dlaroche@email.arizona .edu

Introduction

We develop quality control diagnostics for extraction-free targeted RNA-Seq by recognizing the compositional nature of RNA-Seq and utilizing the existing body of work on compositional data.

Extraction-free Targeted (EFT) RNA-Seq

EFT RNA-Seq offers several benefits over traditional RNA-Seq for clinical use:

- Elimination of amplification bias
- Reduced sequencing cost
- Use of very small sample inputs.
- Simplified bioinformatics workflow (alignment)

EFT RNA-Seq creates the need for post-sequencing quality control metrics.

- No extraction step for discovering biological samples with little or degraded RNA
- No genome alignment for sequence quality assessment

Compositional Framework

RNA-Seq measures relative abundances of RNA transcripts..

- Finite number of transcript reads (counts) per run
- Hierarchical partitioning of sequence reads
  - Total reads in a sequencing run divided among samples
  - Total reads in a sample divided among probes

Statistical procedures must account for the compositional geometry.

- Work in ratios of components with centered log ratio (CLR) transformation.

  $CLR(x_i) = log\left(\frac{x_i}{g(\boldsymbol{x})}\right)$, where $g(\boldsymbol{x})$ is the geometric mean of $\boldsymbol{x}$.

- Account for geometry when measuring distance (Aitchison distance)

$$Dist(\boldsymbol{x},\boldsymbol{y}) = \left[\sum_{i=1}^{D}\left(log\left(\frac{x_i}{g(\boldsymbol{x})}\right) - log\left(\frac{y_i}{g(\boldsymbol{y})}\right)\right)^2\right]^{1/2}$$

Fractional Allocation of Aligned Reads to Samples

Problems with sample quality, library preparation, or sequencing may result in a low number of reads allocated to a sample.

- For most experiments we expect samples to get equal allocation of sequence reads

Method:

1. Use the CLR to transform the total number of reads allocated to each sample.
2. Apply outlier detection method to identify poor samples

Definition: x_i is a quality control sample failure if x_i < lower-quartile - 1.5 × IQR or x_i > upper-quartile + 1.5 × IQR, where IQR is the interquartile range of $\boldsymbol{x}$.

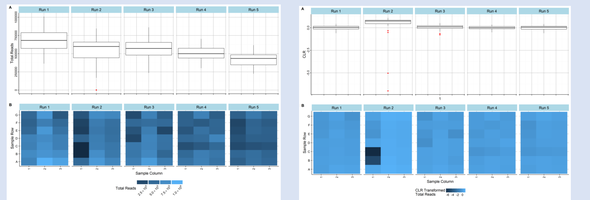

Figure 1. The raw (left) and CLR transformed (right) total counts for 120 mRNA samples sequenced in 5 runs. The CLR transformation substantially improves the detection of low total count outlying samples.

Testing for Compositional Invariance

RNA-Seq analyses assume the number of reads assigned to a sample is independent of the relative abundance of probes (Compositional Invariance (CI)).

- Insufficient availability of probes may lead to violations of CI.
- Modelling CI in high dimensional RNA-Seq can be problematic

Method:

1. Calculate all pairwise distances using Aitchison Distance
2. Create heatmap of distances with samples ordered by total reads
3. Plot multi-variate distance between each sample and the center of the top 25% of samples with respect to total reads

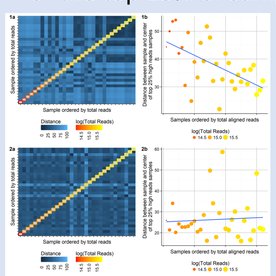

Figure 2. Distances between samples should be independent of the total reads assigned to each sample. Violations of compositional invariance can be visualized through the ordering samples by total aligned reads. Gradients along the diagonal in the heatmap (a), or non-zero slopes (b) indicate a violation of CI.

Batch Effects and Normalization

Identifying and controlling for batch effects is a critical step in the transition of RNA-Seq from the lab to the clinic.

- Identify sample groups related to batch with a hierarchical clustering (HC) or principal components analysis (PCA) on CLR transformed counts.
- CLR transformation has several benefits:
  - Double centering transformation improves PCA biplot interpretation
  - Works as single sample normalization without biological assumptions about differential expression among samples

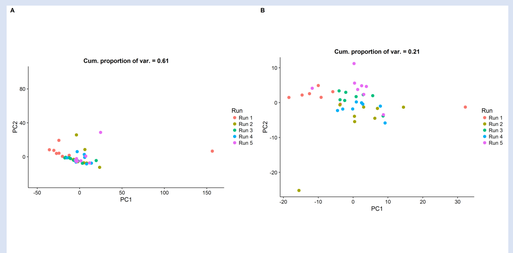

Figure 3. Plot of the first 2 components of a PCA of mi-RNA samples from 5 different sequencing runs. Batch effects are ore more clearly discernible in the CLR transformed data (B) than in the log-untransformed data (A).

Discussion

- Our fractional read allocation metric can identify problematic samples which arise from multiple failure modes, e.g. a low quality sample or a sequencing problem.
- The identification compositional invariance violations allows the investigator to account for the dependency between the total aligned reads and the composition when modelling.
- The CLR transformation improves PCA biplot interpretation and provides sample normalization.

Future Work

- Leverage compositional theory to improve analytical methods for targeted extraction-free RNA-Seq