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

Dominic LaRoche Dean Billheimer Shripad Sinari Kurt Michels
Bonnie LaFleur

November 26, 2016

1 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 [23, 26]. Recent studies evaluating RNA-seq have found generally high intra-platform and inter-platform congruence across multiple laboratories [**Li2013**, 24, 23]. Despite these promising results, there is still a need to establish reliable diagnostics, quality control metrics and improve the reproducibility of RNA-seq data. Understanding, and capatilizing on, the relative frequency nature of RNA-Seq data provides tools for identifying batch effects, creating quality control metrics, and improving reproducibility.

This research is focused on developing diagnostics for targeted RNA-Seq. Targeted sequencing allows researchers to efficiently measure transcripts of interest for a particular disease by focusing sequencing efforts on a select subset of transcript targets. Targeted sequencing offers several benefits over traditional whole-transciptome RNA-Seq for clinical use including the elminiation of amplification bias, reduced sequencing cost, and a simplified bioinformatics workflow. However, traditional RNA-Seq and targeted RNA-Seq data share many of the same properties so the methods described here should be easily extensible to traditional RNA-Seq.

Targeted and whole transcriptome RNA-Seq data provide relative frequencies of the measured transcripts.

Relative frequency measures are characterized as a vector of proportions of some whole. These proportions are necessarily positive and sum to a constant which is determined by the measurement system and not the measurand. For example, as an experiment, we take a large bag of marbles of different colors and pour them over a shallow bowl. The bag holds many more marbles than the bowl so most of the marbles spill out and remain unmeasured. Since we measure an unknown portion of the marbles in the bag we cannot know the total number of each marble color that was contained in the bag. However, we can estimate the relative frequencies of each marble color in the bag. The total number of marbles we observe is a function of the measurement, i.e. the size of the bowl.

Suppose we repeat this experiment a second time with a new bag of marbles to which we add a bunch of red marbles. Since the bowl is fixed in size, if capture more rd marbles in the bowl (as we expect) then the other colors int he bowl must then decrease in frequency proportionally even if the absolute number of the other colors remains constant. This proportionality is strictly enforced when the bowl is small and the bag is large, but one can imagine a scenario where the bowl is large and the bag is small in which the propotionality is somewhat relaxed.

Similarly, NGS-based RNA-Seq methods are also relative frequencies. 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 shorter 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 thereby creating a natural hierarchical structure to RNA-Seq data.

Previous authors have identified the relative nature of RNA-Seq data [20, 6, 21, 12, 15]. For example, Robinson and Smyth (2007) [20] 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. 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.

The positivity and summation constraint complicate the analysis of relative frequency data. John Aitchison observed that relative frequency data is compositional and developed a methodology based on the geometric constraints of composiitons [1]. As early as 1896 Karl Pearson [19] identified the spurious correlation problem associated with compositions. Aitchison (1986) [1] identified the difficulty of interpreting the covariance matrix of a composition. Recent authors have argued that ignoring the sum constraint can lead to unexpected results and erroneous inference [16]. Despite the evidence that RNA-Seq data are compositional in nature, few researchers have extended the broad set of compositional data analysis theory and operations for use in RNA-Seq analysis problems. We provide a brief background on compositional methods. We then extend existing compositional data methodology to include 3 (?) quality control metrics for RNA-Seq data. Finally, we show how compositional properties can be exploited to facilitate exploratortion of high-dimensional RNA-Seq data.

Illumina incorporates several sequencing specific quality control metrics including percentage of clusters passing filters and cluster density analysis. Other quality control metrics are also available, such as HTQC [**Yang2013**]. However, most of the quality control metrics, while informative, are subjective...

Extraction-free sequencing technologies, such as HTG EdgeSeq, permit the use of very small sample volumes but create the need for additional quality control metrics since poor quality samples, which would likely be removed after unsucessful RNA extraction in extraction-based technologies, can be sequenced.

2 Methods

2.1 Compositional Data

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 fixed constant [1]. For RNA-seq data, the total sum constraint is imposed by the limited number of available reads in each sequencing run. Since this total differs between sequencing platforms we will refer to the total number of available reads as $\mathbb{T}$. These reads are distributed among the D samples in a sequencing run such that:

$$\sum_{i=1}^{D} t_i = \mathbb{T} \tag{1}$$

where t_i represents the total reads for sample i. Because of the total sum constraint, the vector $\mathbf{t}$ is completely determined by $D-1$ elements since the D^{th} element of $\mathbf{t}$ can be determined from the other $d = D-1$ elements and the total $\mathbb{T}$:

$$t_D = \mathbb{T} - \sum_{i=1}^{d} \mathbf{t_i} \tag{2}$$

In 2, any of the elements can be chosen for t_D with the remaining elements labeled $1, ..., d$ in any order [1]. Similarly, the total reads for each sample (t_i) are distributed among the P transcript targets in the assay such that $\sum_{j=1}^{P} p_{ij} = t_i$, where p_{ij} is the total reads allocated to target j in sample i. We highlight the hierarchichal structure of RNA-Seq data as it leads to useful properties when devloping quality control metrics.

From equations 1 and 2 it is clear that the total reads allocated to each of the D samples represent a $D - 1 = d$ dimensional simplex ($\mathcal{S}^d$). This leads to a diffculty in interpreting the traditional $D \times D$ covariance structure. In particular, it is clear that for a D-part composition $\mathbf{x}$, $\text{cov}(x_1, x_1 + \cdots + x_D) = 0$ since $x_1 + \cdots + x_D$ is a constant. Moreover, the sum constraint induces negativity in the covariance matrix,

$$\text{cov}(x_1, x_2) + \cdots + \text{cov}(x_1, x_D) = -\text{var}(x_1). \tag{3}$$

Equation 3 shows that at least one element of each row of the covariance matrix must be negative. Aitchison refers to this as the "negative bias difficulty" (although 'bias' is not used in the traditional sense; [1], p. 53). The existence of these negative values creates problems for the interpretation of the covariance matrix since values are no longer free to range between 0 and 1.

Because of the difficulties outlined above, standard statistical methodology is not always appropriate [1] and can produce misleading results [15]. To overcome these obstacles, Aitchison [2] proposed working in ratios of components. We focus on the Centered Log-Ratio (CLR) which treats the parts of the composition symmetrically and provides an informative covariance structure. The CLR transformation is defined for a D-part composition $\mathbf{x}$ as:

$$y_i = \text{CLR}(x_i) = log\left(\frac{x_i}{g(\mathbf{x})}\right), \tag{4}$$

where $g(\mathbf{x})$ is the geometric mean of $\mathbf{x}$. The $D \times D$ covariance matrix is then defined as:

$$\Gamma = [\text{cov}\,(y_i, y_j) : i,\ j = 1, ..., D] \tag{5}$$

To avoid numerical difficulties arising from sequence targets with 0 reads Martin-Fernandez et al. [17] an additive-multiplicative hybrid transformation. This transformation is additive on the zero components but multiplicative on the non-zero components. It has several advantages over the simple additive transformation since it preserves several important compositional properties. Martin-Fernandez et al. [17] recommend using 0.55 × the smallest detectable value as originally suggested by Sandford et al. [11, 22]. The threshold value for RNA-seq data must account for read depth since a 0 in a sample with a library size of 1 thousand reads would potentially not be 0 if the total number of reads was increased to 1 million. Therefore, we define the threshold value for a sample as $\delta = \frac{0.55}{\text{Total Reads}}$. The Martin-Fernandez transformation then becomes,

$$v_i = \frac{x_i}{\sum_{i=1}^{D} x_i} \tag{6}$$

$$u_i = \begin{cases} \delta & \text{if } v_i = 0 \\ v_i \times \left[1 - \left(\sum_{i=1}^{D} \mathcal{I}_{(v_i=0)}\right) \times \delta\right] & \text{if } x_i \neq 0. \end{cases} \tag{7}$$

The CLR transformation is then applied to the vector $\mathbf{u}$.

The CLR transformation can be viewed as an extension of the familiar Counts per Million (CPM) transformation [12] defined as, $log_2\left(\frac{r_{gi}+0.5}{t_i+1} \times 10^6\right)$, where r_{gi} is the number of sequence reads for each probe (g) and sample (i), (scaled to avoid zero counts), adjusted for the number of mapped reads (library count) for each sample t_i (scaled by a constant 1 to ensure the proportional read to library size ratio is greater than zero). The primary difference between the CLR and log(CPM) transformations is in the use of the geomtric mean in the denominator of the CLR transformation. Although the CLR transformation preserves the original dimmension of the data, and gives equal treatment to every element of $\mathbf{x}$, the resulting covariance matrix, Γ, is singular. Therefore, care should be taken when using general multivariate methods on CLR transformed data.

Similarly, the compositional geometry must be accounted for when measuring the distance between two compositions or finding the center of a group of compositions [3]. Aitchison [4] outlined several properties for any compositional difference metric which must be met: scale invariance, permutation invariance, perturba-

tion invariance (similar to translation invariance for Euclidean distance), and subcompositional dominance (similar to subspace dominance of Euclidean distance). The scale invariance requirement is ignorable if the difference metric is applied to data on the same scale (which is generally not satisfied in raw RNA-seq data due to differences in read depth). The permutation invariance is generally satisfied by existing methods such as Euclidean distance [18]. However, the perturbation invariance and subcompositional dominance are not generally satisfied [18].

Aitchison [1, 4] suggests using the sum of squares of all log-ratio differences. Billheimer, Guttorp, and Fagan [8] use the geometry of compositions to define a norm which, along with the perturbation operator defined by Aitchison [1], allow the interpretation of differences in compositions. Martin-Fernandez et al. [18] showed that applying either Euclidean distance or Mahalanobis distance metric to CLR transformed data satisfies all the requirements of a compositional distance metric. Euclidean distance on CLR transformed compositions is referred to as Aitchison distance:

$$d_A(x_i, x_j) = \left[\sum_{k=1}^{D} \left(log\left(\frac{x_{ik}}{g(x_i)}\right) - log\left(\frac{x_{jk}}{g(x_j)}\right) \right)^2 \right]^{\frac{1}{2}}$$

or

$$d_A(x_i, x_j) = \left[\sum_{k=1}^{D} \left(clr(x_{ik}) - clr(x_{jk}) \right)^2 \right]^{\frac{1}{2}}.$$

Up to this point we have referred to the total reads available per sequencing run, $\mathbb{T}$. However, it is more typical to work with the aligned reads in practice. The total aligned reads, T, is always a fraction of the total reads available for a sequencing run, $\mathbb{T}$. The fraction of the total reads aligned can be a affected by multiple factors, including the choice of alignment algorithm, which we do not address here. We assume that T imposes the same constraints on the data as oultined above for $\mathbb{T}$ and will refer exclusively to T hereafter.

2.2 Sample Quality Control

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. The Percent Pass Filter (% PF) metric provided on

Illumina sequencers provides a subjective measure that can identify problems with sequencing that result in a low number of reads allocated to a sample. However, % PF will not necessarily catch problems associated with poor sample quality of or problems with sample pre-processing since these processes may affect cluster generation, and not just cluster quality. Morevoer, there is currently no objective way to evaluate sample quality based on the total number of reads attributed to a sample. We propose a method for objectively identifying problematic samples based on the total number of reads allocated to the sample.

For most experimental designs we expect the number of reads allocated to each sample in a sequencing run to arise from the same general data generating mechanism, regardless of experimental condition. The objective is then to determine which samples arise from a different mechanism. Outlier detection is well suited for discovering obervations that deviate so much from other observations that they are likely to have arisen from a different mechanism [10]. We base our method off Tukey's box-plots[25] which is a well used and robust method for detecting outliers [7].

We expect the total number of reads allocated to each sample, t_i, to be equivalent notwithstanding random variation. For a given sequencing run with D samples we define the vector of total reads allocated to each sample as $\mathbf{t}$. Since the D dimensional vector $\mathbf{t}$ is a composition we have $\mathbf{t} \in \mathcal{S}^{D-1}$, the $D-1$-dimensional simplex. As noted above, traditional statistical methods may not be appropriate for data in the simplex. Therefore, we map $\mathbf{t} \in \mathcal{S}^{D-1} \rightarrow \mathbf{x} = CLR(\mathbf{t}) \in \mathcal{R}^D$ using the Centered Log Ratio transformation. We then apply Tukey's method for detecting outliers to $\mathbf{x}$, which simply identifies those observations which lie outside 1.5 times the inter-quartile range.

Definition 1. x_i is a quality control failure if $x_i <$ lower-quartile$-1.5\times$ IQR *or* $x_i >$ upper-quartile$+1.5\times$ IQR, where IQR is the interquartile range of $\mathbf{x}$.

2.3 Batch Effects and Normalization

Batch effects arising from differing labratory conditions or operator differences have been identified as a problem in high-throughput measurement systems ([14, 9]). Identifying and controlling for batch effects is a critical step in the transition of RNA-Seq from the lab to the clinic. Batch effects are typically identified with a hierarchical clustering method or principal components analysis (PCA) and removed through various normalization methods ([20, 6, 21, 12, 13]).

The compositional nature of RNA-Seq data has important implications for the detection of batch effects becuase of the difficulty of interpreting the covariace matrix ([1]) and the incompatibility with standard measures of distance ([18]). The CLR transformation facilitates both batch effect detection and normalization. The CLR transformed covariance matrix is suitable for exploration through PCA ([5]) or hierarchical clustering using standard Euclidean distance ([18]).

References

[1] J Aitchison. *The statistical analysis of compositional data*. Chapman & Hall, Ltd., 1986. ISBN: 0-412-28060-4. URL: http://dl.acm.org/citation.cfm?id=17272 (cit. on pp. 3, 4, 6, 8).

[2] J. Aitchison and S.M. Shen. "Logistic-normal distributions:Some properties and uses". In: *Biometrika* 67.2 (1980), pp. 261–272. ISSN: 0006-3444. DOI: 10.1093/biomet/67.2.261. URL: https://www.researchgate.net/publication/229099731{_}Logistic-Normal{_}Distributions{_}Some{_}Properties{_}and{_}Uses (cit. on p. 4).

[3] J. Aitchison et al. "Logratio analysis and compositional distance". In: *Mathematical Geology* 32.3 (2000), pp. 271–275. ISSN: 08828121. DOI: 10.1023/A:1007529726302 (cit. on p. 5).

[4] John Aitchison. "On criteria for measures of compositional difference". In: *Mathematical Geology* 24.4 (1992), pp. 365–379. ISSN: 0882-8121. DOI: 10.1007/BF00891269. URL: http://link.springer.com/10.1007/BF00891269 (cit. on pp. 5, 6).

[5] John Aitchison and Michael Greenacre. "Biplots of compositional data". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 51.4 (2002), pp. 375–392. ISSN: 0035-9254. DOI: 10.1111/1467-9876.00275. URL: http://doi.wiley.com/10.1111/1467-9876.00275 (cit. on p. 8).

[6] Simon Anders and W Huber. "Differential expression analysis for sequence count data". In: *Genome Biol* 11.10 (2010), R106. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-10-r106. URL: http://www.biomedcentral.com/content/pdf/gb-2010-11-10-r106.pdf (cit. on pp. 2, 7).

[7] Irad Ben-Gal. "Outlier Detection". In: *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, 2009, pp. 117–130. DOI: 10.1007/978-0-387-09823-4_7. URL: http://link.springer.com/10.1007/978-0-387-09823-4{_}7 (cit. on p. 7).

[8] Dean Billheimer, Peter Guttorp, and William F Fagan. "Statistical Interpretation of Species Composition". In: *Journal of the American Statistical Association* 96.456 (2001), pp. 1205–1214. ISSN: 0162-1459. DOI: 10.1198/016214501753381850. URL: http://www.jstor.org/stable/3085883 (cit. on p. 6).

[9] Chao Chen et al. "Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods". In: *PLoS ONE* 6.2 (2011). ISSN: 19326203. DOI: 10.1371/journal.pone.0017238 (cit. on p. 7).

[10] D. M. Hawkins. *Identification of Outliers*. Dordrecht: Springer Netherlands, 1980. ISBN: 978-94-015-3996-8. DOI: 10.1007/978-94-015-3994-4. URL: http://link.springer.com/10.1007/978-94-015-3994-4 (cit. on p. 7).

[11] Henk A. L. Kiers et al., eds. *Data Analysis, Classification, and Related Methods*. Studies in Classification, Data Analysis, and Knowledge Organization. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000. ISBN: 978-3-540-67521-1. DOI: 10.1007/978-3-642-59789-3. URL: http://www.springerlink.com/index/10.1007/978-3-642-59789-3 (cit. on p. 5).

[12] Charity W Law et al. "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts." En. In: *Genome biology* 15.2 (2014), R29. ISSN: 1465-6914. DOI: 10.1186/gb-2014-15-2-r29. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-2-r29 (cit. on pp. 2, 5, 7).

[13] Jeffrey T. Leek. "Svaseq: Removing Batch Effects and Other Unwanted Noise From Sequencing Data". In: *Nucleic acids research* 42.21 (2014), pp. 1–9. ISSN: 13624962. DOI: 10.1093/nar/gku864 (cit. on p. 7).

[14] Jeffrey T Leek et al. "Tackling the widespread and critical impact of batch effects in high-throughput data." In: *Nature reviews. Genetics* 11.10 (2010), pp. 733–739. ISSN: 1471-0056. DOI: 10.1038/nrg2825. arXiv: NIHMS150003. URL: http://dx.doi.org/10.1038/nrg2825 (cit. on p. 7).

[15] David Lovell et al. "Proportionality: A Valid Alternative to Correlation for Relative Data." In: *PLoS computational biology* 11.3 (2015), e1004075. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004075. URL: http://www.ncbi.nlm.nih.gov/pubmed/25775355 (cit. on pp. 2, 4).

[16] David Lovell et al. "Proportions, Percentages, PPM: Do The Molecular Biosciences Treat Compositional Data Right?" In: *Compositional Data Analysis: Theory and Applications*. October. John Wiley & Sons, Ltd, 2011, pp. 191–207. ISBN: 9780470711354. DOI: 10.1002/9781119976462.ch14. URL: http://dx.doi.org/10.1002/9781119976462.ch14 (cit. on p. 3).

[17] J. A. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn. "Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation". en. In: *Mathematical Geology* 35.3 (2000), pp. 253–278. ISSN: 1573-8868. DOI: 10.1023/A:1023866030544. URL: http://link.springer.com/article/10.1023/A{\%}3A1023866030544 (cit. on p. 5).

[18] J A Martín-Fernández et al. "Measures of difference for compositional data and hierarchical clustering methods". In: *Proceedings of IAMG* 98.1 (1998), pp. 526–531 (cit. on pp. 6, 8).

[19] Karl Pearson. "Mathematical Contributions to the Theory of Evolution.–On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs : Pearson, K. : Free Download & Streaming : Internet Archive". In: *Proceedings of the Royal Society of London* 60 (1896), pp. 489–498. URL: https://archive.org/details/philtrans00847732 (cit. on p. 3).

[20] M. D. Robinson and G. K. Smyth. "Small-sample estimation of negative binomial dispersion, with applications to SAGE data". In: *Biostatistics* 9.2 (2007), pp. 321–332. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxm030. URL: http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxm030 (cit. on pp. 2, 7).

[21] Mark D Robinson and Alicia Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data." In: *Genome biology* 11.3 (2010), R25. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-3-r25 (cit. on pp. 2, 7).

[22] Richard F. Sanford, Charles T. Pierson, and Robert A. Crovelli. "An objective replacement method for censored geochemical data". In: *Mathematical Geology* 25.1 (1993), pp. 59–80. ISSN: 0882-8121. DOI: 10.1007/BF00890676. URL: http://link.springer.com/10.1007/BF00890676 (cit. on p. 5).

[23] SEQC/MAQC-III Consortium. "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium." In: *Nature biotechnology* 32.9 (2014), pp. 903–14. ISSN: 1546-1696. DOI: 10.1038/nbt.2957. URL: http://dx.doi.org/10.1038/nbt.2957 (cit. on p. 1).

[24] Peter A C 't Hoen et al. "Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories." In: *Nature biotechnology* 31.11 (2013), pp. 1015–22. ISSN: 1546-1696. DOI: 10.1038/nbt.2702. URL: http://dx.doi.org/10.1038/nbt.2702 (cit. on p. 1).

[25] John W. (John Wilder) Tukey. *Exploratory data analysis*. Addison-Wesley Pub. Co, 1977, p. 688. ISBN: 9780201076165 (cit. on p. 7).

[26] Kendall Van Keuren-Jensen, Jonathan J Keats, and David W Craig. "Bringing RNA-seq closer to the clinic." In: *Nature biotechnology* 32.9 (2014), pp. 884–5. ISSN: 1546-1696. DOI: 10.1038/nbt.3017. URL: http://dx.doi.org/10.1038/nbt.3017 (cit. on p. 1).