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

October 12, 2016

1 Abstract

2 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/MAQC-III Consortium 2014, Van Keuren-Jensen, Keats, and Craig 2014). Recent studies evaluating RNA-seq have found generally high intra-platform and inter-platform congruence across multiple laboratories (**Li2013**; 't Hoen et al. 2013; SEQC/MAQC-III Consortium 2014). 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.

Relative frequency measures (hereafter reffered to as *compositional data* for consistency with other disciplines) are characterzied as a vector of proportions of some whole. These proportions are necessarily positive and sum to a constant. The positivity and summation constraint complicate the analysis of compositions. For example, John Aitchison (Aitchison 1986) identified the difficulty of interpreting the covariance matrix of a composition that results from the depency in the data induced by the sum constraint. As early as 1896 Karl Pearson (Pearson 1896) identified the spurious correlation problem associated with compositions.

NGS-based RNA-Seq methods are inherently compositional because 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. Previous authors have identified the relative nature of RNA-Seq data (Robinson and Smyth 2007; Anders and Huber 2010; Robinson and Oshlack 2010; Law et al. 2014; Lovell et al. 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. 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 1896; 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 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. We also show how compositional properties can be exploited to improve exploratory analyses and improve reproducibility.

3 Methods

3.1 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 (Aitchison 1986). 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.

For RNA-seq data, the total sum constraint is imposed by the limited number of available reads in each sequencing platform. Since this total differs between 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 $\mathbb{T}$:

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

In 2, any of the elements can be chosen for t_D with the remaining elements labeled $1, ..., d$ in any order (Aitchison 1986).

From equations 1 and 2 it is clear that the D samples represent a $D-1 = d$ dimensional simplex (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,

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

which means at least one element of each row of the covariance matrix mustbe negative. Aitchison refers to this as the "negative bias difficulty" (although 'bias' is not used in the traditional sense; Aitchison 1986, p. 53). The existence of these negative values creates problems for the interpretation of the covariance matrix since values are no longer free to take values between 0 and 1.

Similarly, the compositional geometry must be accounted for when measuring the distance between two compositions or finding the center of a group of compositions (Aitchison et al. 2000). Aitchison (Aitchison 1992) outlined several properties for any compositional difference metric which must be met: scale invariance, permutation invariance, perturbation 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). The permutation invariance is generally satisfied by existing methods (Martín-Fernández et al. 1998). However, the perturbation invariance and subcompositional dominance are not generally satisfied.

Becuase of the difficulties outlined above, standard statistical methodology is not always appropriate (Aitchison 1986) and can produce misleading results (Lovell et al. 2015). To overcome these obstacles, Aitchison (Aitchison and Shen 1980) 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 $\mathbf{t}$ as:

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

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

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

Although the CLR transformation gives equal treatment to every element of $\mathbf{t}$, the resulting covariance matrix, Γ, is singular. Therefore, care should be taken when using general multivariate methods on CLR transformed data.

Aitchison (Aitchison 1986; Aitchison 1992) suggests using the sum of squares of all log-ratio differnces. Billheimer, Guttorp, and Fagan (2001) use the geometry of compositions to define a norm which, along with the perturbation operator defined by Aitchison (Aitchison 1986), allow the interpretation of differences in compositions (Billheimer, Guttorp, and Fagan 2001). Briefly, denote the elementwise multiplication of two positive D-vectors $\mathbf{u}$ and $\mathbf{v}$ by

$$\mathbf{u} \cdot \mathbf{v} \equiv (u_1 v_1, u2_v 2, ..., u_D v_D)'.$$

Further define the perturbation operator for composition $\mathbf{x}$ and perturbation $\alpha \in S^d$ as

$$\mathbf{z} = \mathbf{x} \oplus \alpha = C\left(\mathbf{x}\alpha\right)$$

for compositional addition. Compositional multiplication is acheived via the power transformation,

$$\mathbf{x}^{\alpha} \equiv C\left(x_1^{\alpha}, x_2^{\alpha}, ..., x_k^{\alpha}\right).$$

Billheimer et al. (Billheimer, Guttorp, and Fagan 2001) show that S^d, with a defined perturbation orperator and scalar multiplication, constitutes a complete inner product space an inner product defined as

$$\langle \mathbf{u}, \mathbf{z} \rangle = \left(\theta' \mathcal{N}^{-1} \eta\right)^{1/2},$$

where, θ and η are the CLR transformations of $\mathbf{u}$, and $\mathbf{z}$ respectively and $\mathcal{N} = I_D + j_D j_D'$ (I_D is a D-dimensional identity matrix and j_D is a D-length vector of 1's).

Martin-Fernandez et al. (1998) 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

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

3.2 Outlier Detection

Problems with RNA isolation, library preparation, or sequencing may result in a low number of reads for the sample. There is currently no objective way to evaluate sample quality based on the total number of reads attributed to a sample. We develop a method grounded in the compositional nature of RNA-Seq data for objectively identifying samples with potentially poor quality.

For most experimental designs we expect the number of reads in each sample, t_i, to be equivalent notwithstanding random variation. Since these reads are part of a composition it is natural to view them as arising from a Multinomial distribution with equal probabilities. Since each cell has the same probability we test

for outlying values using the Binomial distribution with probability $1/D$ and size $n =$ total available reads.

3.3 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 (Leek et al. 2010; Chen et al. 2011). 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 (Robinson and Smyth 2007; Anders and Huber 2010; Robinson and Oshlack 2010; Law et al. 2014; Leek 2014).

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 (Aitchison 1986) and the incompatibility with standard measures of distance (Martín-Fernández et al. 1998). The CLR transformation facilitates both batch effect detection and normalization. The CLR transformed covariance matrix is suitable for exploration through PCA (Aitchison and Greenacre 2002) or hierarchical clustering using standard Euclidean distance (Martín-Fernández et al. 1998).

References

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

Aitchison, J. and S.M. Shen (1980). "Logistic-normal distributions:Some properties and uses". In: *Biometrika* 67.2, 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).

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

Aitchison, John (1992). "On criteria for measures of compositional difference". In: *Mathematical Geology* 24.4, pp. 365–379. ISSN: 0882-8121. DOI: 10.1007/BF00891269. URL: http://link.springer.com/10.1007/BF00891269 (cit. on pp. 3, 4).

Aitchison, John and Michael Greenacre (2002). "Biplots of compositional data". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 51.4, 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. 6).

Anders, Simon and W Huber (2010). "Differential expression analysis for sequence count data". In: *Genome Biol* 11.10, 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, 6).

Billheimer, Dean, Peter Guttorp, and William F Fagan (2001). "Statistical Interpretation of Species Composition". In: *Journal of the American Statistical Association* 96.456, pp. 1205–1214. ISSN: 0162-1459. DOI: 10.1198/016214501753381850. URL: http://www.jstor.org/stable/3085883 (cit. on pp. 4, 5).

Chen, Chao et al. (2011). "Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods". In: *PLoS ONE* 6.2. ISSN: 19326203. DOI: 10.1371/journal.pone.0017238 (cit. on p. 6).

Law, Charity W et al. (2014). "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts." En. In: *Genome biology* 15.2, 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, 6).

Leek, Jeffrey T. (2014). "Svaseq: Removing Batch Effects and Other Unwanted Noise From Sequencing Data". In: *Nucleic acids research* 42.21, pp. 1–9. ISSN: 13624962. DOI: 10.1093/nar/gku864 (cit. on p. 6).

Leek, Jeffrey T et al. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data." In: *Nature reviews. Genetics* 11.10, pp. 733–739. ISSN: 1471-0056. DOI: 10.1038/nrg2825. arXiv: NIHMS150003. URL: http://dx.doi.org/10.1038/nrg2825 (cit. on p. 6).

Lovell, David et al. (2011). "Proportions, Percentages, PPM: Do The Molecular Biosciences Treat Compositional Data Right?" In: *Compositional Data Analysis: Theory and Applications*. October. John Wiley & Sons, Ltd, pp. 191–207. ISBN: 9780470711354. DOI: 10.1002/9781119976462.ch14. URL: http://dx.doi.org/10.1002/9781119976462.ch14 (cit. on p. 2).

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

Martín-Fernández, J A et al. (1998). "Measures of difference for compositional data and hierarchical clustering methods". In: *Proceedings of IAMG* 98.1, pp. 526–531 (cit. on pp. 4, 6).

Pearson, Karl (1896). "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, pp. 489–498. URL: https://archive.org/details/philtrans00847732 (cit. on pp. 1, 2).

Robinson, M. D. and G. K. Smyth (2007). "Small-sample estimation of negative binomial dispersion, with applications to SAGE data". In: *Biostatistics* 9.2, 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, 6).

Robinson, Mark D and Alicia Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." In: *Genome biology* 11.3, R25. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-3-r25 (cit. on pp. 2, 6).

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

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

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