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

Dominic LaRoche Dean Billheimer Shripad Sinari Kurt Michels
Bonnie LaFleur

May 17, 2017

1 Introduction

We develop quality control diagnostics for targeted RNA-Seq using the theory of compositional data. 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-transcriptome RNA-Seq for clinical use including the elimination of amplification bias, reduced sequencing cost, and a simplified bioinformatics workflow. Moreover, extraction-free targeted sequencing technologies, such as HTG EdgeSeq, permit the use of very 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. The post-sequencing methods described here should be easily extensible to traditional extraction-based RNA-Seq because targeted and traditional RNA-Seq data share many of the same properties.

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. Targeted and whole transcriptome RNA-Seq measurements from NGS-based instruments provide only relative frequencies of the measured transcripts. The measurement technology, along with sample preparation, preclude the measurement of absolute abundance. The total number of reads in a sequencing run for high-throughput RNA-Seq instruments is determined by the maximum number of avail-

able reads and not the absolute number of reads in a sample. For example, the Illumina Mi-Seq is limited to 25 million reads in a sequencing run while the Roche 454 GS Junior $^{(TM)}$, with longer read lengths, claims approximately 100,000 reads per run for shotgun sequencing. 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 abundance nature of RNA-Seq data [20, 6, 21, 12, 14]. 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. Finally, the commonly used $\log_2$ Counts per Million (CPM) re-scaling transformation proposed by Law et al. (2014) [12] divides each sequence count by the total number of reads allocated to the sample thereby transforming the data for each sample into a vector of proportions.

The positivity and summation constraint complicate the analysis of relative frequency data. As early as 1896 Karl Pearson [19] identified the spurious correlation problem associated with compositions. John Aitchison observed that relative frequency data is compositional and developed a methodology based on the geometric constraints of compositions [1]. Recent authors have argued that ignoring the sum constraint can lead to unexpected results and erroneous inference [15]. 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 develop two quality control metrics and improve batch effect detection for RNA-Seq data.

2 Methods

2.1 Compositional Data

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 number of reads allocated to target j in sample i. We highlight the hierarchical structure of RNA-Seq data as it leads to useful properties when developing 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 problems when using methods developed for standard Euclidean sample spaces such as interpreting the traditional $D \times D$ covariance structure or measuring the distance between vectors. 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 structurally induced negative values create problems for the interpretation of the covariance

matrix. Similarly, the use of naive distance metrics in the simplex may not be interpretable as in Euclidean space. Because of these difficulties, standard statistical methodology is not always appropriate [1] and can produce misleading results [14].

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:

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

The CLR transformation is similar to 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 geometric mean in the denominator of the CLR transformation. The use of the geometric mean results in subtracting the mean of the log transformed values from each log-transformed element thereby centering the vector of log-ratio transformed read counts. The difference appears minor but has important implications for the application of several common statistical methods.

Although the CLR transformation preserves the original dimension 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. Aitchison [1] proposed an alternative transformation, the additive log-ratio (ALR), which does not treat the components symmetrically but results

in a non-singular covariance matrix. The ALR transformation is defined as,

$$y_i = \text{ALR}(x_i) = log\left(\frac{x_i}{x_D}\right), \tag{6}$$

where x_D, the D^{th} component of x, can be any component.

As noted above, the compositional geometry must be accounted for when measuring the distance between two compositional vectors 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, 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 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:

$$d_A(x_i, x_j) = \left[\sum_{k=1}^{D}\left(log\left(\frac{x_{ik}}{g(x_i)}\right) - log\left(\frac{x_{jk}}{g(x_j)}\right)\right)^2\right]^{\frac{1}{2}} \tag{7}$$

or

$$d_A(x_i, x_j) = \left[\sum_{k=1}^{D}\left(clr(x_{ik}) - clr(x_{jk})\right)^2\right]^{\frac{1}{2}}. \tag{8}$$

To avoid numerical difficulties arising from sequence targets with 0 reads, Martin-Fernandez et al.

(2000) [17] suggest an additive-multiplicative hybrid transformation. If zeros are present in the data We recommend using the Martin-Fernandez transformation with a threshold value of $\delta = \frac{0.55}{\text{Total Reads}}$ to account for differences in sequencing depth. The CLR transformation is then applied to the Martin-Fernandez transformed data which contains no zeros.

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 outlined above for $\mathbb{T}$ and will refer exclusively to T hereafter.

3 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 or problems with sample pre-processing since these processes may affect cluster generation, and not just cluster quality. This is particularly important for extraction-free RNA-Seq technologies, such as the HTG EdgeSeq$^{(tm)}$, which allow for the use of smaller input amounts but lack the intermediate steps for checking sample quality. 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, namely the chemistry of the NGS-based measurement system, regardless of experimental condition. The objective is then to determine which samples arise from a different mechanism. Outlier detection is well suited for discovering observations that deviate so much from other observations that they are likely to have arisen from a different mechanism [11]. We base our method off Tukey's box-plots [24], which is a commonly 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 4. 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 sample 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 $\mathbf{x}$.

We demonstrate the utility of our sample quality control measure using two sets of targeted RNA-Seq data: 1) 120 mRNA technical replicate universal-RNA samples prepared with the HTG EdgeSeq Immuno-Oncology assay and sequenced in 5 different equally sized runs, and 2) 105 miRNA technical replicate samples of human plasma, FFPE tissue, and Brain RNA prepared with the HTG EdgeSeq Whole Transcriptome miRNA assay. These two data sets differ in the both the type of RNA (mRNA versus miRNA) and the number of sequence targets in each assay (558 versus 2,280 targets, for the mRNA and miRNA assays respectively). All samples were prepared for sequencing using the HTG EdgeSeq Processor and sequenced with an Illumina Mi-Seq sequencer.

We compare the utility of our method to evaluation of the un-transformed total counts. Figure 1 shows a boxplot and heat-map of the total number of reads allocated to each sample for each of 5 sequencing runs. Figure 2 shows the same data after CLR transformation. After transformation the poor samples become much more visually evident in the heat maps. Additionally, the ability to detect outlying values increases and the number of poor samples detected increases from 1 to 6.

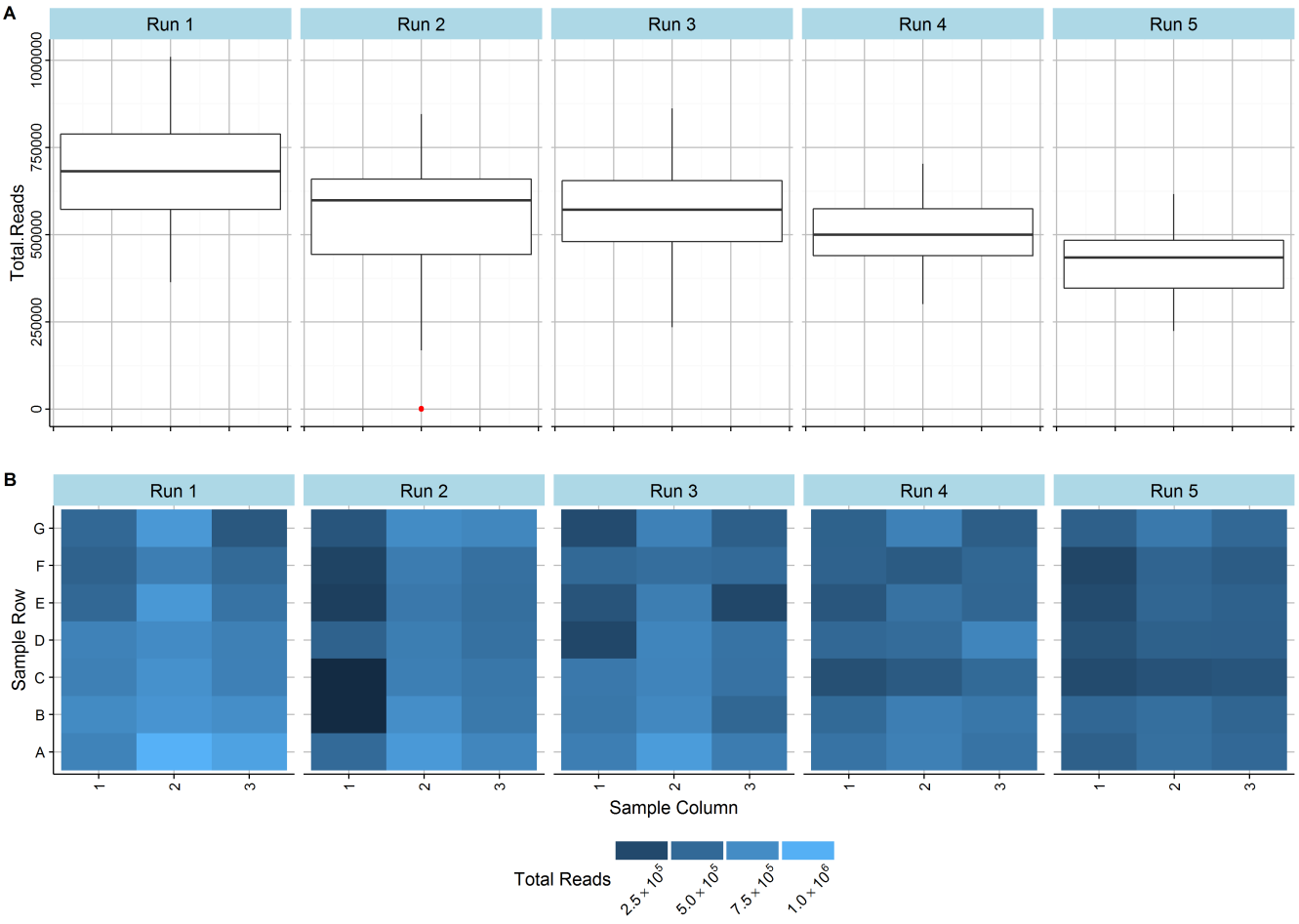

Figure 1: A) Distributions of total reads allocated to each sample in 5 runs on an Illumina Mi-Seq sequencer. Only 1 sample is identified as a problematic sample. B) Heat-maps showing the relative totals for each sample within each run. The darker heat-maps for runs 4 and 5 reflect the generally lower number of total reads in those sequencing runs as compared to runs 1 and 2. This is caused by normal variation in the number of reads available in a sequencing run.

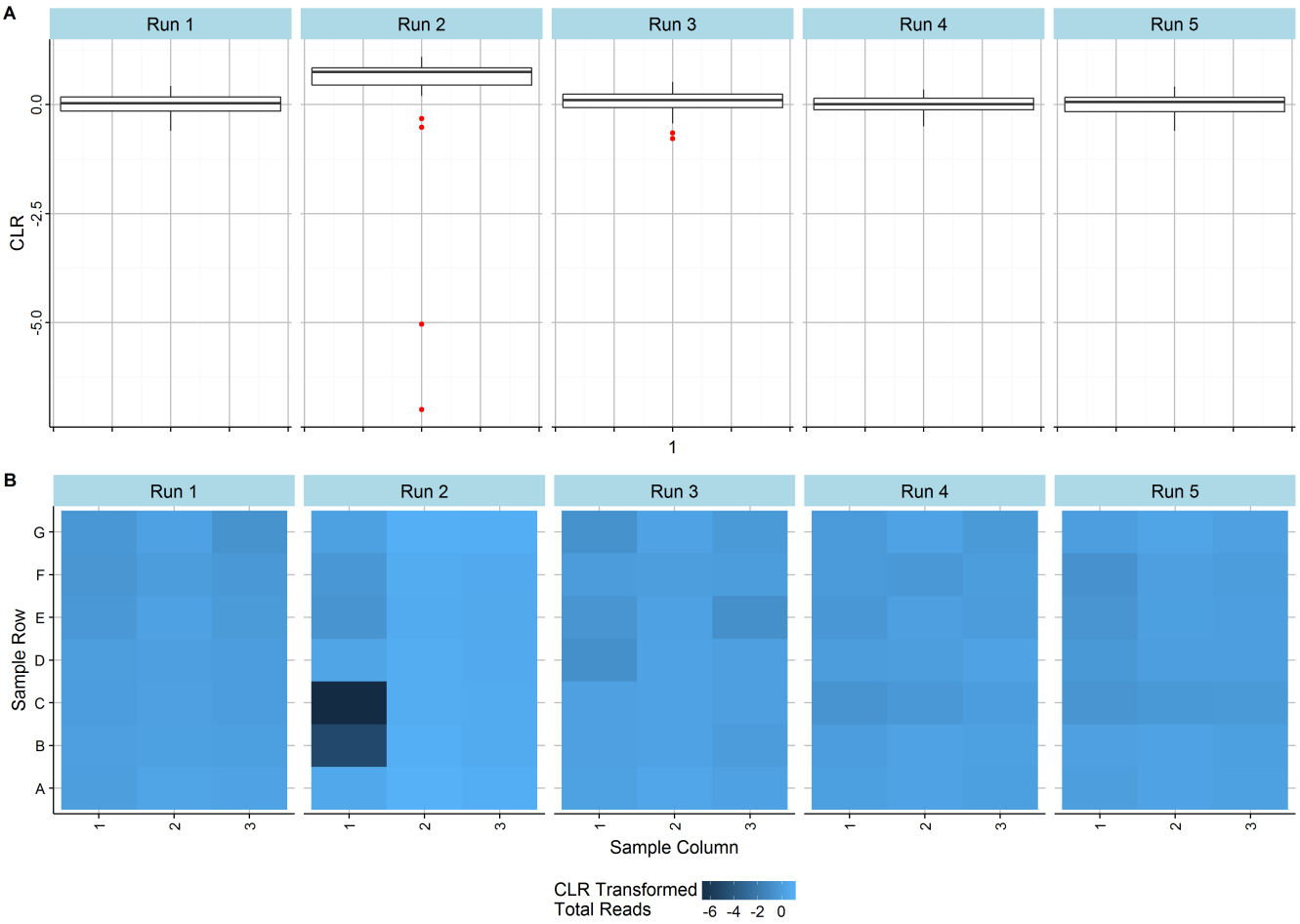

Figure 2: A) Distributions of CLR transformed total reads allocated to each sample in 5 runs on an Illumina Mi-Seq sequencer. After CLR transformation, 6 samples are identified as a problematic. B) Heat-maps showing the relative CLR transformed totals for each sample within each run.

4 Testing for Compositional Invariance

Normalization and standardization methods for RNA-Seq generally assume that the total number of reads assigned to a sample does not affect the observed relative frequencies of probes within an assay. For example, implicit in the CPM transformation is the idea that if you re-scale the counts (by dividing by the total for each sample) then the resulting counts are comparable and any differences are due to underlying differences in expression. Other methods which apply a scaling factor to each sample, such as Trimmed-mean of M values (TMM) or Quantile normalization, also rely on this assumption. In the parlance of compositional data these methods assume *Compositional Invariance*, i.e. the underlying composition is statistically independent of the total size of the composition (the total counts for a sample, t).

Compositional invariance (CI) is an important property for RNA-Seq data which enables the comparison of samples with differing read depths. However, it is well documented that the quality of RNA-Seq depends on the read depth of the sequencing run with higher read-depths associated with higher quality data [23, 22]. Read depth may affect the measurement of relative abundances for the target RNA sequences as some targets may receive proportionally more reads as the read depth increases. This would be a direct violation of CI and could lead seemingly differential expression between samples with different read depths, even after normalization. Another form of CI violation, that is perhaps more likely in RNA-Seq experiments, is the dependence between the variance of read counts and the read depth.

Aitchison [1] outlined a simple model for testing compositional covariance using the ALR transformation,

$$[y_1 \dots y_d] = \begin{bmatrix} 1 & t \end{bmatrix} \begin{bmatrix} \alpha_1 & \cdots & \alpha_d \\ \beta_1 & \cdots & \beta_d \end{bmatrix} + [e_1 \dots e_d], \tag{9}$$

where $y_1 \dots y_d$ are the d ALR transformed components, t is the vector of sample total aligned reads, $\alpha_1 \dots \alpha_d$ are the probe specific log-ratio intercepts, and $\beta_1 \dots \beta_d$ are the coefficients relating the the total aligned reads to the relative expression of the probe. A test for compositional invariance for the experiment then becomes a test of the null hypothesis, $H_o : \beta_1 = \cdots = \beta_d = 0$. This test can be re-parameterized to test for dependence between the variance and total aligned reads as well.

Unfortunately, the small sample sizes and large number of probes typically associated with RNA-Seq

experiments complicates the application of Aitchison's model. We propose an alternative visualization for simultaneously detecting both violations of compositional invariance described above. We use the multivariate Aitchison distance (8) between all pairs of samples in a heat-map with the samples ordered by total aligned reads. If CI is violated we expect pairs samples with similar total aligned reads will have smaller scalar distances than those with large differences in total aligned reads. This will result in visual clustering around the 45 degree axis. If the variance depends on the total aligned reads, we expect the scalar distance between sample pairs to decrease with increasing read depth resulting in a visual gradient in the distance heat map. To reduce the visual noise associated with outlier samples in the heat-map we also provide a dot-plot of the distance between each CLR transformed sample and the compositional center of the samples in the top quartile of total reads.

We demonstrate this visualization with two sets of miRNA samples (Fig. 4) and two sets of mRNA samples(Fig. 4). The miRNA samples are composed of 40 technical replicates each of (1) plasma samples and (2) brain samples. In the miRNA data there is a clear gradient along the 45 degree axis for the plasma samples (Fig. 4.A). This indicates a dependence between the total aligned reads and the variance of the samples (as indicated by the increasing multivariate distance between replicates as the total aligned reads decreases). In contrast, there is no clear gradient in the brain samples (Fig. 4.B). The mRNA samples are composed of (A) 16 technical replicates of diseased pancreas tissue and (B) 16 technical replicates of normal pancreas tissue. In the diseased pancreas samples there is a clear gradient with low total aligned read samples more distant from samples with greater total aligned reads (Fig. 4.1). This indicates that the composition is dependent on the total aligned reads, a violation of compositional invariance for these samples. In contrast, the normal pancreas samples show no such pattern related to total aligned reads (Fig. 4.2).

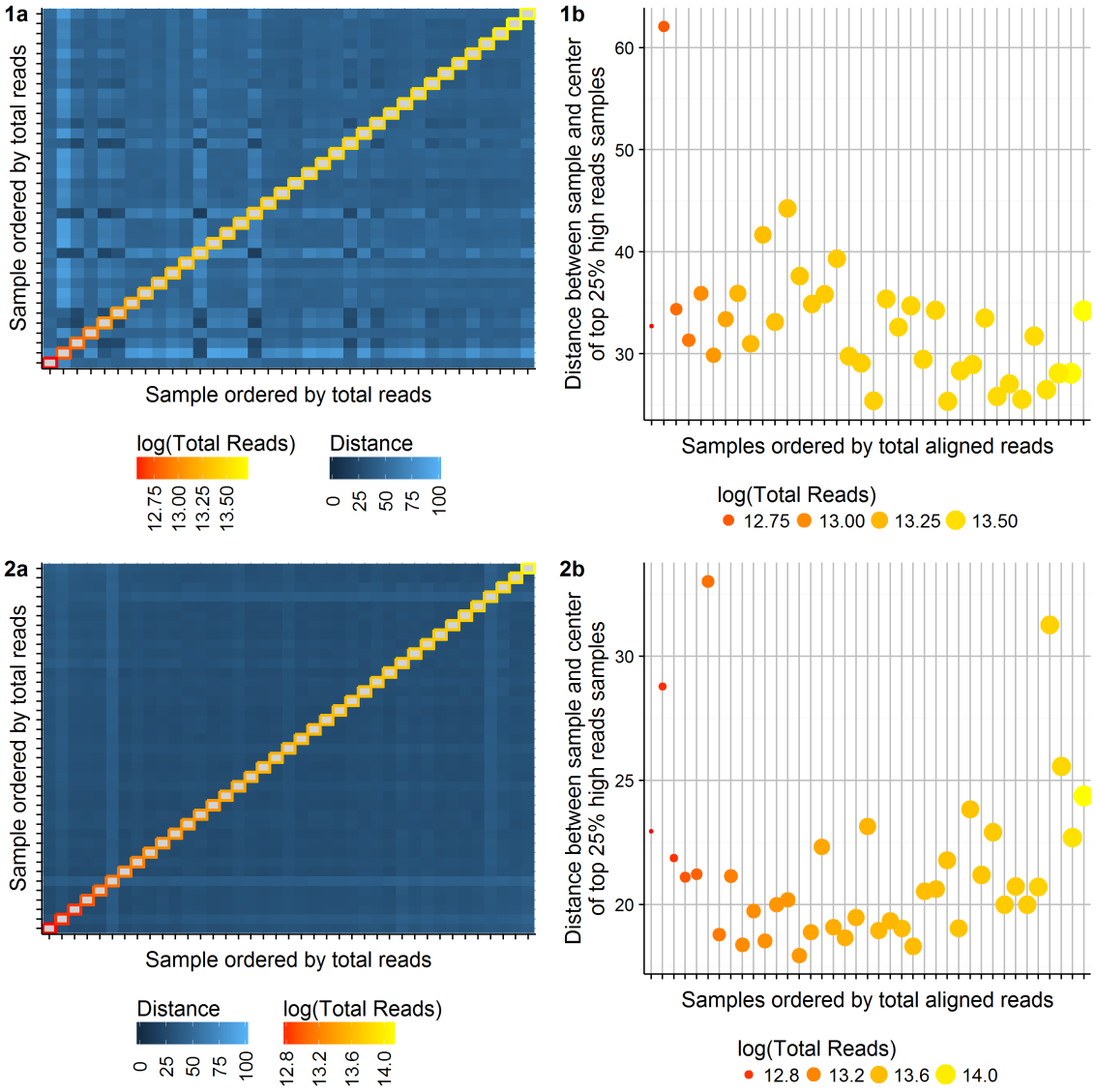

Figure 3: Two sets of miRNA samples with samples in (1.) showing a violation of compositional invariance and (2) showing compositional invariance.

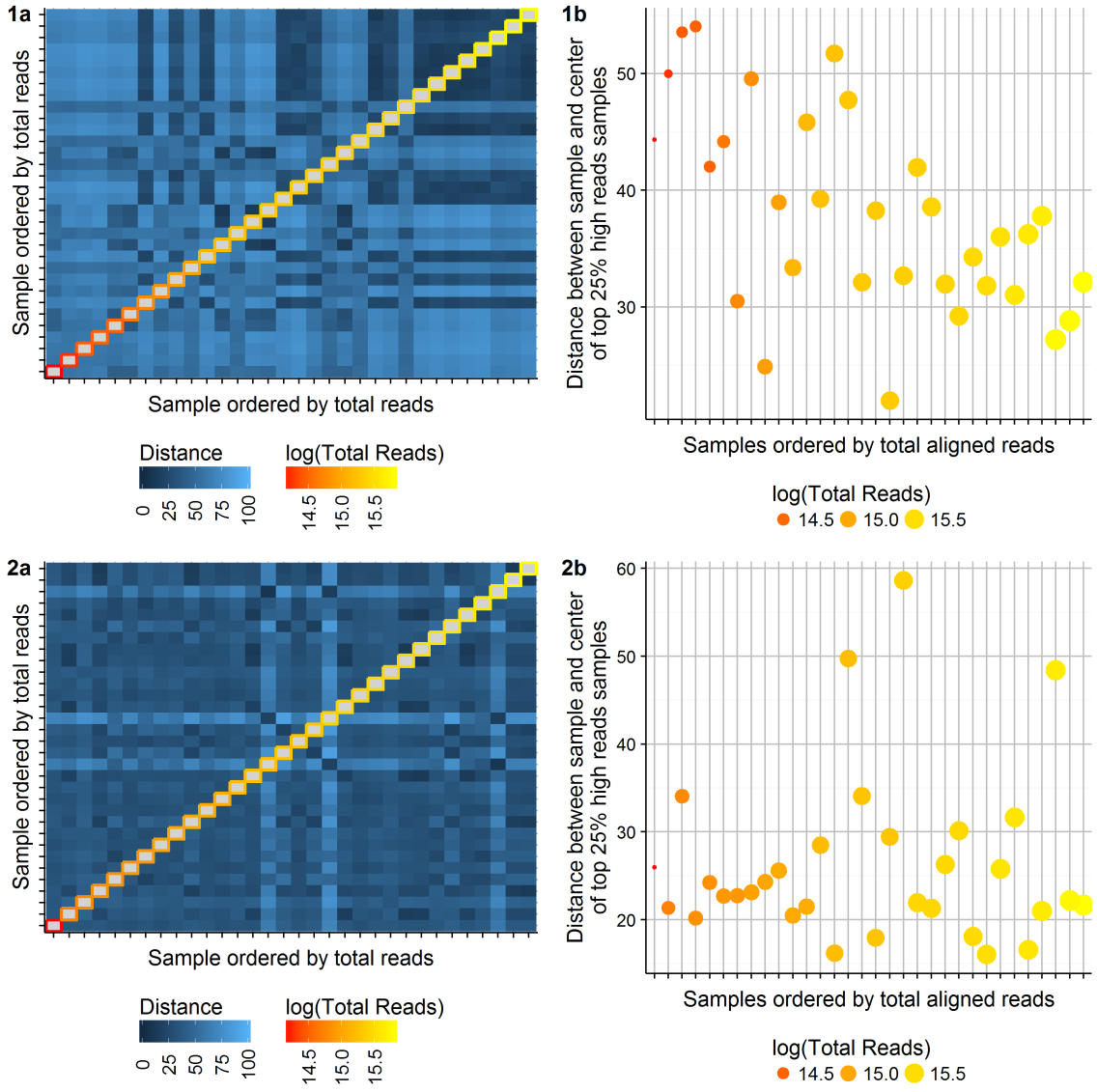

Figure 4: Two sets of mRNA samples with samples in (1.) showing a violation of compositional invariance and (2) showing compositional invariance.

5 Batch Effects and Normalization

Batch effects arising from differing laboratory conditions or operator differences have been identified as a problem in high-throughput measurement systems [13, 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 (HC) method or principal components analysis (PCA). For both methods, the multivariate distance between the samples is visualized. either in a biplot for PCA or a dendrogram for HC, to check for the existence of clusters of samples related to batch. The compositional nature of RNA-Seq data has important implications for the detection of batch effects due to the incompatibility with standard measures of distance between compositions as noted above [1, 18].

The next generation sequencing process results in arbitrary differences in scale among samples as some samples will receive more total reads than others. Principle components analysis is sensitive to differences in scale among the variables, failure to remove these difference can mask potential batch effects and leave unwanted technical variation in the data. Most normalization methods use a scaling factor calculated for each sample to re-scale the read count for each gene within the sample [10]. The CLR transformation can similarly be viewed as a scaling normalization (with the scale factor chosen as the inverse of the geometric mean $1/g(x)$). Unlike other normalization methods, the CLR transformation has the added benefit of being applied at the individual sample level, not experiment wise, and requires no assumptions about differential expression among samples, unlike other popular normalizations. This makes it particularly well suited for the clinic where there are generally no reference samples for normalization.

Aitchison demonstrated that the CLR transformation has several other useful properties in addition to re-scaling the data [1], particularly with respect to PCA biplots [5]. Most notably for the detection of batch effects, the distance between any two points representing samples in the form-biplot approximates the Euclidean distance between the two samples or the Mahalanobis distance for covariance biplots. The CLR transformation retains the property that this distance is at least as great as the distance between any corresponding subset of these two compositions (subspace dominance). Additionally, the euclidean distance between two CLR transformed samples is location invariant. Other scaling and normalization methods do not necessarily satisfy these properties and, therefore, batch effects may be masked or artificial.

We demonstrate the use of the compositional biplot to detect batch effects using technical replicates

of three sample types: brain, plasma, and fresh frozen paraffin embedded (FFPE). Each sample is replicated 8 times in each of 5 sequencing runs for a total of 120 samples. Samples were prepared using the EdgeSeq Whole Transcriptome miRNA assay which measures 2,280 targets including including 11 control probes and 2,269 unique miRNA probes. All sequencing was performed on an Illumina Mi-seq$^{(tm)}$ sequencer.

We create a second data set, by re-scaling the original data, to better illustrate the effects of changes in read depth on batch effect detection. To re-scale the samples from the original we multiply every read count in a given sample by a factor, ranging from 0.5 to 1.5, randomly generated from the uniform distribution. We then obtain a new data set in which the proportions between the read counts remains unchanged but the variance in the total number of reads among the samples is increased.

We perform a PCA on log-transformed and CLR transformed data. We then construct form-biplots of the first two principle components for each transformed data set (Fig. 5). The differences between the 3 samples types (brain, plasma, and FFPE) dominate the first two principle components for both data sets. However, the CLR transformed data provides tighter clusters, relative to the distance between the clusters, than the log-transformed raw data. There is also a single FFPE sample which is closer to the brain samples than the other samples. It is worth noting that this sample would have been removed using our proposed quality control metric.

Since the sample type differences overwhelm the potential batch effects we performed a second PCA on only the brain samples for both transformed data sets (Fig. 5). Both biplots exhibit clustering by batch but the CLR transformed data shows better separation between the batches, although batches are still overlapping.

Some of the batch effects detected in the log-transformed data may be attributable to the differences in total reads between batches. By randomly re-scaling each sample by a constant we are able to break the relationship between batch and the total reads in a sample. Figure **??** gives the biplots for log-transformed and CLR-transformed randomly re-scaled data. The sample type clusters in the log-transformed data become more diffuse while the CLR-transformed biplot remains unchanged. Most notably, the batch effects previously visible in the log-transformed brain samples become completely obscured in the randomly re-scaled data but remain unchanged in the CLR-transformed data (Fig. 5).

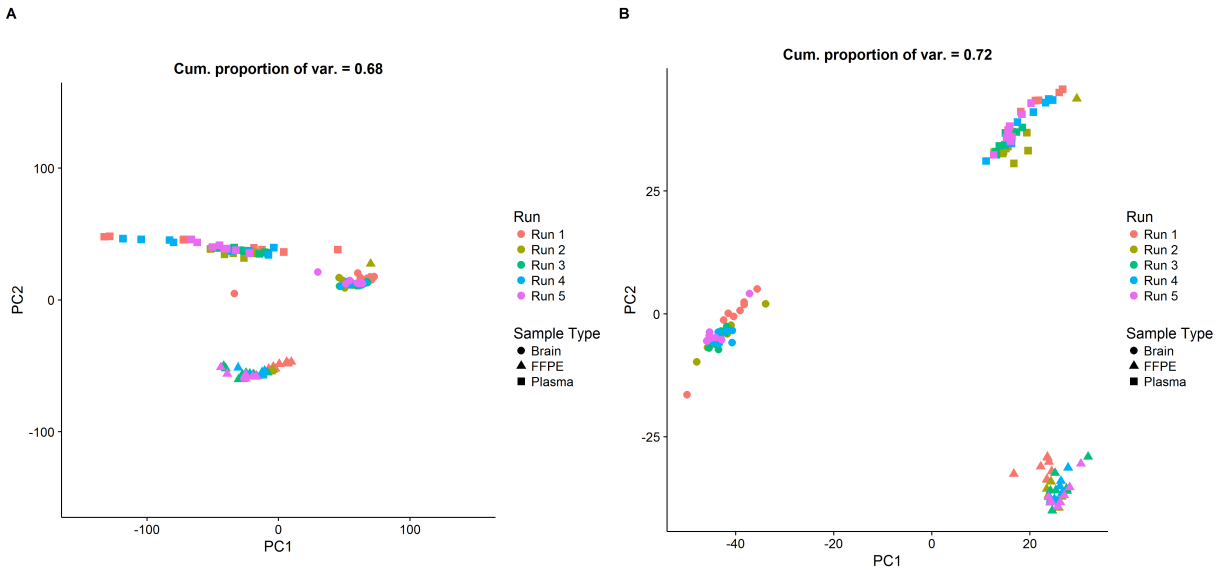

Figure 5: Principle component analysis of A) log-transformed and B) CLR-transformed read count data. The differences between sample types is much greater than the batch effects in both transformation. The CLR transformation results in tighter sample type clusters resulting from less variation along the first principle component.

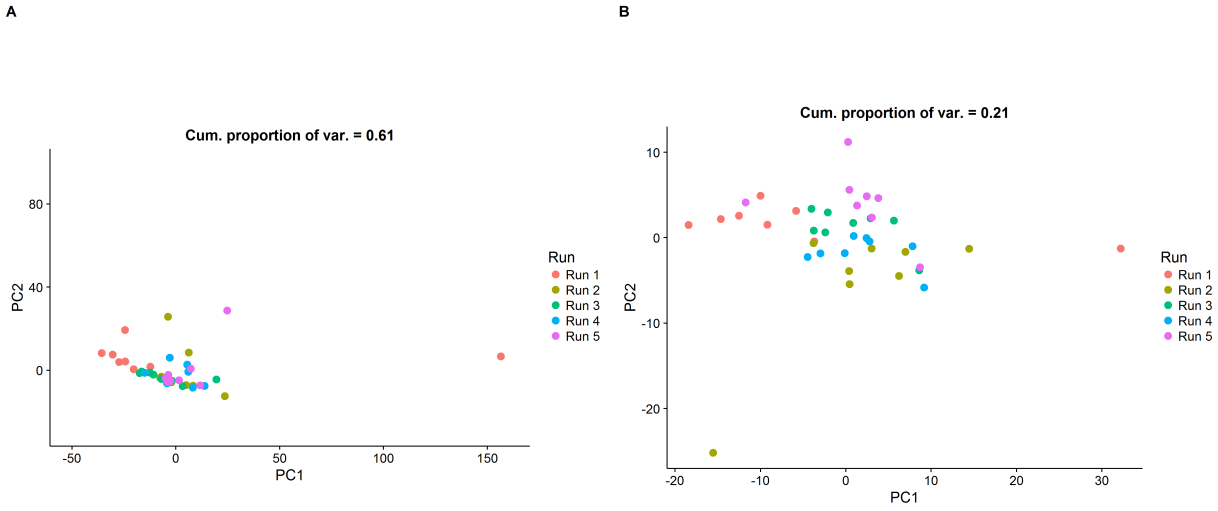

Figure 6: Principle component analysis of only brain samples from A) log-transformed and B) CLR-transformed read count data. The batch effects are more easily identified in the CLR transformed data.

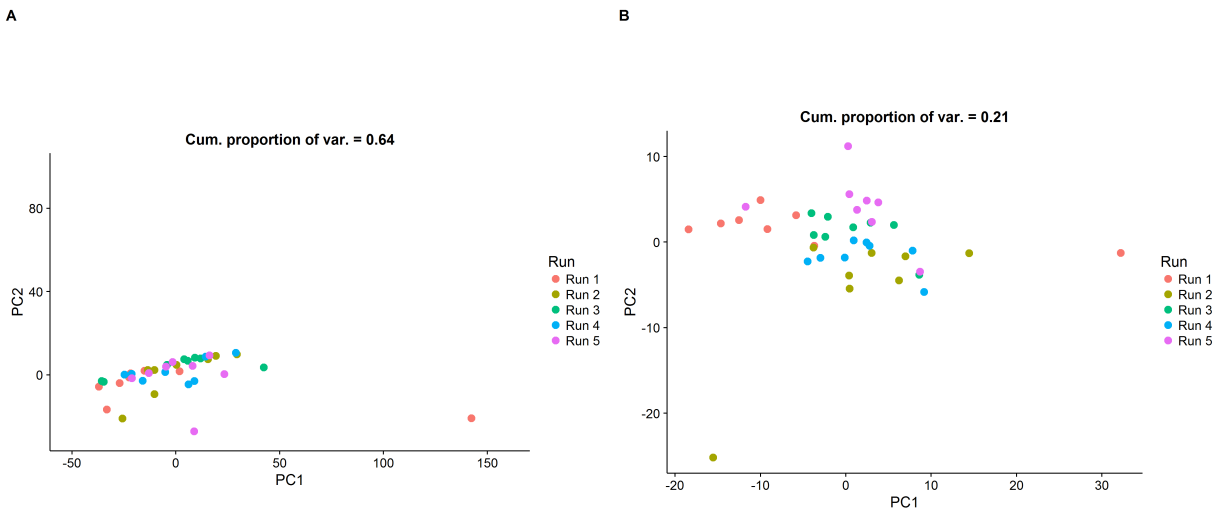

Figure 7: Principle components analysis of the randomly re-scaled brain samples for A) log-transformed and B) CLR-transformed read count data. The batch effects visual in the log-transformed raw data disappear after random re-scaling whereas the batch effects remain identifiable in the CLR transformed data.

6 Discussion

Our sample quality control metric can identify problematic samples which arise from multiple failure modes, e.g. a low quality sample or a sequencing problem. However, it is conceivable that a sample might have an unusually low (or high) number of reads and still provide quality information. In certain experimental designs one might be able to further evaluate these samples with a PCA biplot on the CLR transformed data. In our PCA analysis we identified a FFPE sample which would have failed our quality control and was clearly very different from the other technical replicates. However, if this sample had remained quite similar to the other FFPE replicates this would have provided information that the sample may still be valuable. In this way, the quality control metric and PCA biplot can be used in tandem to provide additional information about the quality of a sample.

The compositional invariance visualization is a logical extension of the sample quality control metric since the assumption of the sample quality control is that the total number of aligned reads is related to the proportional allocation of reads within the sample. As noted above samples which violate the compositional invariance property may still contain valuable information. The identification compositional invariance violations allows the investigator to account for the dependency between the total aligned reads and the

relative abundance of transcripts within the samples when modelling.

The principal components analysis biplot is a well know dimension reduction visualization. For the current data the dimension is reduced from 2,280 probes to 2 principle components. The utility of the data reduction, including the quality of the approximation of the multivariate distance between the samples, is proportional to the amount of variance explained by these two principle components. In our data the first two principle components explain between 72 and 21 percent of the variation in the data. The analysis with the lowest percent of variation explained by the first 2 components is of the CLR-transformed brain samples. Surprisingly, batch effects are still visible in this plot, in which case they can be removed [16].

As RNA-Seq makes the transition from the research laboratory to the clinic there is a need for robust quality control metrics. The realization that RNA-Seq data are compositional opens the door to the existing body of theory and methods developed by John Atichison and others. We show that the properties of compositional data can be leveraged to develop new metrics and enhance existing methods.

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. 2–5, 10, 14).

[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 p. 5).

[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. 14).

[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 p. 2).

[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. 6).

[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. 5).

[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. 14).

[10] Marie Agn??s Dillies et al. “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis”. In: *Briefings in Bioinformatics* 14.6 (2013), pp. 671–683. ISSN: 14675463. DOI: 10.1093/bib/bbs046 (cit. on p. 14).

[11] 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. 6).

[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, 4).

[13] 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. 14).

[14] 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).

[15] 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. 2).

[16] J Luo et al. “A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data.” In: *The pharmacogenomics journal* 10.4 (2010), pp. 278–91. ISSN: 1473-1150. DOI: 10.1038/tpj.2010.57. URL: http://www.ncbi.nlm.nih.gov/pubmed/20676067http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2920074 (cit. on p. 18).

[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. 6).

[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. 5, 14).

[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. 2).

[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 p. 2).

[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 p. 2).

[22] David Sims et al. "Sequencing depth and coverage: key considerations in genomic analyses". In: *Nature Reviews Genetics* 15.2 (2014), pp. 121–132. ISSN: 1471-0056. DOI: 10.1038/nrg3642. URL: http://www.nature.com/doifinder/10.1038/nrg3642 (cit. on p. 10).

[23] Sonia Tarazona et al. "Differential expression in RNA-seq: a matter of depth." In: *Genome research* 21.12 (2011), pp. 2213–23. ISSN: 1549-5469. DOI: 10.1101/gr.124321.111. URL: http://www.ncbi.nlm.nih.gov/pubmed/21903743http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3227109 (cit. on p. 10).

[24] John W. (John Wilder) Tukey. *Exploratory data analysis*. Addison-Wesley Pub. Co, 1977, p. 688. ISBN: 9780201076165 (cit. on p. 6).