Dissertation Proposal: Methods for the Analysis of Compositional RNA Sequence Data

Dominic D LaRoche

July 25, 2017

Contents

Chapter 1

Background and Introduction to the Problem

1.1 RNA Sequencing Data

Ribonucleic acid (RNA) has become a major target of investigation for a wide variety of research areas in biology due to its role in the function of cells, including gene transcription and regulation. Measurements of RNA in biological samples is increasingly performed using RNA-Seq [29] on high-thoughput next generation sequencing (NGS) platforms. RNA-Seq offers several advantages over a traditional micro-array experiment and has been rapidly adopted by scientists [12]. However, the analysis of RNA-Seq data is complicated by the relative frequency nature of data produced by NGS platforms. The current research is restricted to the analysis of extraction-free targeted RNA-Seq data and the unique analytical challenges it creates.

1.1.1 RNA Sequencing Technology

Several general sequence-based methods exist to quantify RNA from a eukaryotic sample. These include traditional Sanger sequencing of the cDNA [7, 9], tag-based methods such as Serial Analysis of Gene Expression (SAGE), and micro-array experiments. The focus of this research is on extraction-free, targeted RNA-Seq using NGS platforms (typically an Illumina sequencer) so we limit the description of RNA-Seq to this type of platform. However, the methods developed in this proposal are likely to be easily extensible to more general extraction-based, whole transcriptome,

RNA-Seq data.

1.1.2 Process of Collecting and Sequencing RNA

RNA is sequenced from finite volumes of tissue or plasma/serum. In a traditional RNA-Seq experiments RNA must be isolated, purified and converted into double stranded complementary DNA (cDNA) sample libraries prior to sequencing. There are many methods available for RNA library preparation which vary with respect to the resulting libraries complexity, evenness and continuity of coverage, and accuracy for expression profiling [15].

The focus of this research is on extraction-free targeted RNA-Seq. Extraction free sample preparation enables measurement of *a priori* selected transcripts with greater precision and smaller sample inputs [**Rimsza2011**]. This research does not attempt to address the impacts of the various sample library preparation methods other than to note that RNA is obtained from finite, and potentially very small, biological samples.

To sequence the cDNA libraries, platform specific adapters are added to the ends of the cDNA fragments. Several platforms are available and the specific chemistry differs among them. Several articles have been published comparing these platforms [16, 10] for sequencing RNA. The differences in platforms, and the resulting differences in data produced by them, is beyond the scope of this research. Instead we focus on the common features of these platforms and the resulting impact on RNA-Seq data generally.

For all NGS-based sequencing platforms, adapted cDNA fragments are immobilized on a proprietary surface (flow cells or beads) with a finite amount of area. The resulting cDNA templates are then amplified to create clusters of copied cDNA (Illumina) or beads are placed into wells (Roche 454). Sequences for clonal clusters (wells) are then determined through Sequencing by Synthesis (SBS). The chemistry for the synthesis also varies by platform but the process involves repeatedly adding a single nucleotide to each template and recording the value of the nucleotide through either fluorescence or light emission. Each cluster then represents the sequence of a single cDNA transcript captured on the surface. Different platforms allow a different total number of transcripts to be

sequenced in a single sequencing run, ranging from 70,000 to 5 billion. This research is limited to evaluating data produced on Illumina sequencers which are the most widely adopted NGS technology.

The number of reads in a sequencing run allocated to a sample is referred to as the read depth. Read depth has been associated with data quality [27, 11, 17, 18] with greater read-depth associated with higher precision. There is a trade-off between cost, replication, and read-depth for any given experiment [18] because a given sequencing run has a limited number of reads to allocate to individual samples.

Once sequences have been measured by the sequencer they are aligned to a genome and various quality metrics are calculated. Alignment typically proceeds through a greedy algorithm implemented in programs such as Bowtie [**Langmead2009**]. Several different sequences might align with varying accuracy to the same gene and these must be counted in a way which accounts for the uncertainty of the alignment. In targeted sequencing the alignment is greatly simplified because the targets are known and are aligned with no ambiguity. However, cross-reactivity (when a probe binds to multiple targets) can be a problem in extraction-free targeted sequencing.

1.1.3 Compositional Properties of RNA Sequence Data

Previous authors have identified the relative abundance nature of RNA sequencing data [22, 5, 23, 14, 19]. For example, Robinson and Smyth (2007) [22] 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) [14] 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 relative frequency nature of RNA-seq data arises from two principle constraints: 1) the number of RNA transcripts that can fit into the finite sample collected, and 2) the number of

available reads of those transcripts available in a sequencing run. It is generally assumed that the sample will contain many more transcripts than can be measured by the sequencer so we focus this research on the second constraint.

1.2 Principles of Compositional Data Analysis

Compositional data are non-negative data which are subject to a sum constraint, i.e. all the elements must sum to unity. This simple constraint has some important consequences for many standard statistical methodologies including correlation and regression. Compositional data contain only relative information, i.e. the information about any individual component, or group of components, is relative to the other components and no absolute information about the absolute value of the component. For example, if we know that 20% of the food in a refrigerator is composed of fruit we do not know how much total fruit there is. If the refrigerator is full then there will be substantially more fruit than if the refrigerator is nearly empty. It is, therefore, important to recognize the types of inferences that can be made from compositional data, e.g. no inference can be made on the actual abundances.

Potential problems associated with compositional data were identified as early as 1897 by Pearson who noted that spurious correlations can be induced through ratios of independent variables, e.g. if X, Y, and Z are uncorrelated then X/Z and Y/Z will be correlated. Despite the fact that compositional data naturally arises in a wide variety of scientific disciplines, a general method for analysis of compositional data was not developed until John Aitchison published his seminal book in 1986. Aitchison outlines some basic principles for compositional data analysis (section 1.2.1) and provides some analysis tools for compositional data which conform to these principles (section 1.2.2). Additional methodology has been developed by a number of authors in the 29 years since the publication of Aitchison's book, although a number of problems remain.

1.2.1 Fundamental Principles

Aitchison outlined a set of fundamental principles to which all methods for compositional data should adhere [2]. These principles are outlined below.

Scale Invariance

Scale invariance requires that the results of a statistical procedure should not depend on the scale used. Any meaningful function x of a composition w must satisfy:

$$f(pw) = f(w), \text{ for every } p > 0.$$

Aitchison notes that any meaningful (scale-invariant) function of a composition can be expressed in terms of ratios of the components of the composition. For example, any method used for compositional data should not give different results whether the composition is given as proportions, percentages, parts per million, or any other scale.

Sub-compositional Coherence

Sub-compositional coherence requires that the results of a statistical procedure on a subset of components from a composition should depend only on the data contained in that subset. A sub-composition is defined as the (1, 2, ..., C) parts of a D-part composition $[x_1, ..., x_D]$:

$$[s_1, ..., s_c] = \frac{[x_1, ..., x_c]}{(x_1 + ... + x_c)}$$

Any changes in components $[x_{c+1}, ..., x_D]$ should not have an impact on the inference from the sub-composition. For example, if we measure the number of reads of 14 mRNA sequences from a sample that contains 4,000 unique mRNA sequences the inference we obtain from those 14 sequences should not be affected by the expression level of any of the other 3,986 sequences.

Permutation Invariance

Permutation invariance requires that the results of a statistical procedure should not depend on the ordering of the components.

1.2.2 Statistical Methods for Compositional Data

The Simplex

Traditional statistics is concerned with making inferences from points in R^D. However, the sample space for compositions is restricted to the *Simplex*, S^D because of the sum constraint. This fundamental difference in sample space necessitates an alternative methodology and renders inference from traditional regression and correlation meaningless. Aitchison (1986) developed much of the current methodology for compositional data through careful examination of the algebraic-geometric structure of the simplex.

It is typical to transform compositional data to the *unit simplex* by dividing each component by the sum of components such that the sum of transformed components is equal to 1:

$$\mathcal{C}\left[x_1 ... x_d\right] = \left[x_1 ... x_d\right] / \sum_{i=1}^{D} x_i \tag{1.1}$$

The notation $\mathcal{C}[\cdot]$ is termed the *closure operation* and scales a composition x to the unit simplex. There are two fundamental operations frequently used in R^D for which Aitchison defined equivalent operations in the simplex: 1) translation, and 2) scalar multiplication.

Perturbations

Aitchison identified the need for an operation in the simplex equivalent to $X = x + t$, the translation t. This operation, defined as a *perturbation* and denoted by $\oplus$, takes the form:

$$X = p \oplus x = \frac{\left[p_1 x_1 ... p_D x_D\right]}{\left(p_1 x_1 + ... + p_D x_D\right)} = \mathcal{C}\left[p_1 x_1 ... p_D x_D\right]$$

where, p is the perturbation which translates x to X and $\mathcal{C}[\cdot]$ is the *closure* operation which scales the composition to the unit simplex. The perturbation operator leads to several other useful metrics in the simplex such as the distance between two compositions (see 1.2.2). As well as for characterizing the imprecision or error around a measurement:

$$x_n = \Upsilon \oplus p_n \quad (n = 1, ..., N),$$

where the x_n are the observed measurements, the p_n are independent error perturbations characterizing the imprecision, and Υ is the true underlying composition.

The Power Operation

Like perturbation is the simplical equivalent of translation, Aitchison defined the power operation as the simplical equivalent of multiplication. For any real number $a \in R^1$ and any composition $x \in S^D$ the power transform of x is defined as:

$$X = a \otimes x = \mathcal{C}\left[x_1^a ... x_D^a\right].$$

The power transformation enables a compositional form for regression between a fixed variable v and a composition x:

$$x = \Upsilon \oplus \{logv \otimes \beta\} \oplus p.$$

In this formulation β is composition analogous to regression coefficients and p is a perturbation analogous to an error term in regression.

Log-ratio Analysis

In order to satisfy the scale invariance of compositional data it is typical to work on ratios of the components. Furthermore, since the sample space of ratios of positive numbers is not in the whole of real numbers, it is typical to work in the logarithms of ratios. The log-ratio transformation maps the composition to the whole of real numbers Three popular transformation exist for a D-part composition: 1) additive log-ratio, 2) centered log-ratio, and 3) ilr. The alr transformation is defined as:

$$alr(X) = [log(\frac{x_1}{x_D})\ log(\frac{x_2}{x_D})\ ...\ log(\frac{x_{D-1}}{x_D})] \tag{1.2}$$

and reduces the dimension of the compositional vector from $D \to D-1$. The choice of divisor does not impact the inference made from the data [2], although the divisor must be reliably greater than 0 in all measurements. Moreover, since the transformation is 1:1, inferences on the ratios can be made back to the parts of the composition. Parts of the composition with 0 values will return $log(0/x_D) = log(0) = -\infty$. While this preserves the rank order of the magnitudes of the components

it is not useful in application.

Since the ALR transformation reduces the dimension of the compositional vector and does not treat all elements of the composition equally, Aitchison proposed the *centered log-ratio* transformation (clr). The clr is defined as:

$$clr(X) = [log\left(\frac{x_1}{g(x)}\right), \ ... \ , log\left(\frac{x_D}{g(x)}\right)], \tag{1.3}$$

where $g(x)$ is the geometric mean of X. The clr transformation takes $S^D \rightarrow U^D$, where U^D is a hyper-plane of R^D, thereby preserving the dimension of X and providing symmetric treatment of all elements of X. The CLR transformation will fail if any x_i is equal to 0. This is because the geometric mean, defined as $g(x) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}$, will equal 0. The CLR will then become,

$$clr(x) = log\left(\frac{x_1}{0}\right) = log(x_1) - log(0) = log(x_1) + \infty.$$

For any $x_i > 0$,

$$clr(x_i) = log(x_i) + \infty = \infty$$

whereas for any $x_i = 0$,

$$clr(x_i) = log(0) + \infty = -\infty + \infty,$$

which is indeterminate. Both log-ratio transformations proposed by Aitchison fail to accommodate zeros so various mechanisms have been proposed for handling zeros while maintaining the essential properties of compositional data analysis (see section 1.2.2).

Distance Between Compositions

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, 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 [21]. However, the perturbation invariance and subcompositional dominance are not generally satisfied.

Aitchison [2, 4] suggests using the sum of squares of all log-ratio differences. Billheimer, Guttorp, and Fagan [6] use the geometry of compositions to define a norm which, along with the perturbation operator defined by Aitchison [2], allow the interpretation of differences in compositions. Martin-Fernandez et al. [21] 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

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

Compositional Invariance

Up to this point I have assumed that the size of a composition, i.e. the sum of the components, contains no information. Aitchison (1986) [2] refers to a raw composition as a basis and points out that for any given composition, $\mathbf{x} : \sum \mathbf{x} = 1$ there are an uncountable number of bases which can be acheived by scaling the composition by a constant;

$$\mathbf{w} = t\mathbf{x}, \tag{1.4}$$

where $t = \sum \mathbf{w}$.

Aitchison defines compositonal invariance as the statistical independence of $\mathbf{x}$ and t. He then defines a test for compositional invariance based on the covariance between the composition formed from the basis through the closure operation and the basis size, t. This is accomplished through the

linear model,

$$\mathbf{Y} = \mathbf{A}\mathbf{\Theta} + \mathbf{E} \tag{1.5}$$

where $\mathbf{Y}$ is the ALR transformed composition array, $\mathbf{A}$ is an $N \times 2$ matrix with rth row $[1 \log t_r]$, and

$$\mathbf{\Theta} = \begin{bmatrix} \alpha_1 \dots \alpha_d \\ \beta_1 \dots \beta_d \end{bmatrix}. \tag{1.6}$$

The statistical hypothesis of compositional invariance is that $\beta_1 = \cdots = \beta_d = 0$

If the data satisfy the assumption of logistic normality then hypothesis testing can be acheived

Zeros and Missing Components

Many compositional methods do not work with missing values, or zeros, such as the ALR or CLR transformations as noted above. Because of this, several strategies for handling missing values and zeros through imputation have been proposed. Missing values are classified into missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). These classifications have the same definitions as their non-compositional counterparts. However, imputation methods for non-compositional data are not immediately applicable due to the lack of independence between the components. Missing values, in the traditional sense, are not typical for RNA-seq data so I do not elaborate on these methods here instead focusing on the handling of zeros.

In RNA-seq data, zeros can naturally arise for multiple reasons. If reads for a probe cannot exist because of known reasons, e.g. an exogenous negative control, then a 0 for this probe would be considered a *structural zero*. If a probe is assigned 0 reads for a sample because that gene is not expressed in the sample then this would be a *true zero*. Finally, if a probe is assigned 0 reads for a sample because it has very low expression, then this would be considered *Below Detection Limit* (BDL). For many RNA-seq technologies there is no way to differentiate between true zeros and BDL zeros in practice as it is impossible to know if increasing the read-depth would eventually result in detection of reads for a gene.

Aitchison (1986) proposes several approaches to dealing with zeros in the absence of a one-to-one monotonic transformation which accommodates zeros: amalgamation, imputation with a constant, addition of a constant to every observation, replacing the log-ratio transformation with a modified Box-Cox transformation, and a conditional model which explicitly models the probability of a component having a 0 proportion. Amalgamation, model-based imputation, and conditional models may all be good solutions for particular analyses goals and data sets but are difficult or impossible to apply universally to individual samples, a necessary requirement for clinical utility.

The addition of a constant to every observation (as in the case of the Law et al. [14] log2(Counts per Million) transformation) is a tempting option for its simplicity and ease of implementation for individual samples. However, this additive transformation alters the proportionality within a composition i.e. the ratio $\frac{a}{b} \neq \frac{a+c}{b+c}\ \forall a, b, c \neq 0$. For the CLR transformation the addition of a constant, c to every observation will reduce the variation in the clr transformed data as can be seen in the limit as $c \to \infty$, through repeated applications of l'Hopital's rule, it can be shown that:

$$lim_{c\to\infty} \frac{x_i + c}{\left(\prod_{i=1}^{D}(x_i + c)\right)^{1/D}} = 0 \ \forall \ x_i \in X$$

Since each element of the clr transformed data will converge to a constant the variance of the composition will also converge to 0. Also, as $c \to 0$, the zero elements of the composition will approach $-\infty$ and these components may have undue leverage on downstream analyses. The log-ratio transformation is clearly sensitive to the choice of constant.

Martin-Fernandez et al. [20] address this issue through an additive-multiplicative hybrid transformation :

$$r_j = \begin{cases} c_j, & \text{if } x_j = 0, \\ x_j(1 - \sum_{k|x_k=0} c_k), & \text{if } x_j > 0 \end{cases} \tag{1.7}$$

This transformation is additive on the zero components but multiplicative on the non-zero components. It has several advantages over the simple additive transformation including: 1) perturbation invariance, 2) power transformation invariance, and 3) sub-composition invariance. Importantly, if all $c_j = c$, the Aitchison distance between two transformed data sets does not depend on c.

The choice of c can be different for each zero component but determining the appropriate value

for each 0 in a single sample would be challenging and would likely provide a limited benefit for samples with a relatively small proportion of 0 components. Martin-Fernandez et al. recommend using 0.55 the threshold value as originally suggested by Sanford et al. [13, 24]. 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, I define the threshold value for a sample as $\delta = \frac{1}{\text{Total Reads}}$, and $c = 0.55 \times \delta$.

Sub-Compositions

It is often desirable to evaluate only a portion of the initial set of genes measured, e.g. when comparing two assays with different but overlapping probe sets or when interrogating a subset of oncogenes from a larger assay. It is then necessary to ensure that the resulting sub-composition does not depend on the other components in the initial composition such that variations in a component outside the sub-composition do not affect inferences made from the sub-composition. I.e. the resulting sub-composition should adhere to the principle of sub-compositional coherence (see 1.2.1). In order to ensure the sub-composition is independent of the omitted values from the full composition.

Sub-compositional coherence can be accomplished through the formation of the sub-composition itself. In forming a c dimensional sub-composition we are transforming the composition from $\mathcal{S}^D \rightarrow \mathcal{S}^c$ (with $c < D$). Aitchison [2] formalizes the formation of a c-dimensional sub-composition x_s from a D-dimensional composition x as

$$x_s = \mathcal{C}(Sx),$$

where S is a $c \times D$ selecting matrix and $\mathcal{C}(\cdot)$ is the closure operation. Under the assumption that all of the original components of the composition were independent, the resulting sub-composition will not depend on the values of the original composition. Furthermore, the formation of sub-compositions in this way leads to 3 useful properties: 1) the ratio of any two components from the sub-composition is identical to the ratio of the two components in the full composition, 2) the resulting sub-composition can be interpreted as resulting from a linear projection and 3) the resulting sub-compositions will be on the same scale.

Aitchison [2] defines *complete subcompositional independence* for a composition if "sub-compositions

formed from any partition of the composition form an independent set."

1.3 Current RNA-Seq Methodology with Respect to Compositional Data Methods and Theory

1.3.1 Counts per Million (CPM) Transformation

The log Counts per Million (CPM) transformation is a simple as commonly used standardization method and is implemented in the R package limma [14]. The CPM transformation standardizes each read count with the total number of reads allocated to the sample (library size):

$$log_2\left(\frac{r_{gi}+0.5}{t_i+1}\times 10^6\right), \tag{1.8}$$

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 size) for each sample t_i (scaled by a constant 1 to ensure the proportional read to library size ratio is greater than zero).

The counts per million transformation is a compositional closure operation, similar to 1.1, which has been modifed to accomodate zeros (assuming zeros occur as a result of insufficient sensitivity, see section 1.2.2) and rescaled by 10^6. At the surface, the primary difference between the CPM transformation proposed by Law et al. (2014) and the CLR transformation proposed by Aitchison (1986) is that the CPM transformation uses the sum of the components as the denominator whereas the CLR transformation uses the geometric mean of the components as the denominator. This difference, however minor, has important implication for the interpretation of the resulting transformation.

The CLR transformation of a D-dimensional composition $\mathbf{X}$ can be acheived through the matrix operation:

$$CLR(\mathbf{X}) = G_D \log(\mathbf{X}), \tag{1.9}$$

where $G_D = I_D - D^{-1}J_D$ with I_D the $D \times D$ identity matrix and J_D a $D \times D$ matrix of 1's. The matrix G_D is an idempotent linear transformation and, therefore, a projection which takes a

D-dimensional composition from $\mathcal{S}^d$ to $\mathcal{R}^d$. Recognizing the CLR as a projection from the simplex space to Euclidean space has important implications for the application of traditional statistical methods to the transformed data.

The CPM transformation can be simplified to the log of the compositional closure operation by removing the multiplication by 10^6 and the addition of constants to both the numerator and denominator of the ratio. The resulting operation, CPM^* can then be achieved through the matrix operation:

$$CPM^*(\mathbf{X}) = log(C_N\mathbf{X}), \tag{1.10}$$

where C_N is a $N \times N$ matrix with the inverse of the compostition (sample) totals along the diagonal. From equations 1.9 and 1.10 it becomes obvious that in the CLR transformation the log transformation is applied directly to X, whereas in the CPM transformation the log transformation is applied only after applying the closure operation. Less obvious is that the matrix C_N does not satisfy the properties of a projection matrix, namely C_N is not idempotent. This means the CPM operation cannot be interpreted as a projection. In fact, without the scaling multiplier (10^6), the CPM operation simply replaces the positivity constraint on the data with a negativity constraint.

The geometric mean as the denominator in the CLR transformation arises naturally from the log transformation of X. Dividing by the geometric mean is equivalent to subtracting the mean of the log-transformed components from each log-transformed component; thereby centering the data (equation 1.14).

$$CLR(x_i) = log\left(\frac{x_i}{g(\mathbf{x})}\right) \tag{1.11}$$

$$= log(x_i) - log(g(\mathbf{x})) \tag{1.12}$$

$$= log(x_i) - \frac{1}{D}log(\prod \mathbf{x}) \tag{1.13}$$

$$= log(x_i) - \frac{\sum_{j=1}^{D} log(x_j)}{D} \tag{1.14}$$

The simplified CPM^* does not have a similar interpretation. Rather, the CPM subtracts the

log of the sum of un-transformed components which has no clear interpretation.

$$CPM^*(x_i) = log\left(\frac{x_i}{\sum \mathbf{x}}\right) \tag{1.15}$$

$$= log(x_i) - log(\sum \mathbf{x}) \tag{1.16}$$

This difference plays an important role in principal components analysis (PCA) and the measurement of distance. With respect to PCA, Aitchison (1983) [1] identifies 3 important properties of a transformation which make it suitable for principle components analysis: 1) it must have an interpretable covariance structure, 2) it must be symmetric in its treatment of the components (invariant to permutations), and 3) it must deal with the inherent curvature of compositions.

The inherent curavature of compositions leads to problems with measuring distance. Specifically, the distance metric between any two compositions must be invariant to changes of location. In the simplex, changes in location are acheived through the perturbation operator which is the compositional equivalent of translation in euclidean space (see 1.2.2). Therefore, any distance metric must be invariant to equal changes in location of the two compositions.

Definition 1. *Perturbation invariance for a distance metric, D, is defined as;*

$$D(q \odot x, q \odot y) = D(x, y), \tag{1.17}$$

where $D()$ is any distance metric applied to two compositions x and y.

This property is analogous to translation invariance for Euclidean distance metrics. The CLR transformation results in perturbation invariance when used with standard Euclidean distance. Let $\mathbf{q}$ be a D-dimensional perturbation. Additionally, let $\mathbf{x}$ and $\mathbf{y}$ represent two D-dimensional compositions. Applying the Euclidean distance metric to CLR transformed x and y we have,

$$D(x,y) = \left[\sum_i^D \left(log\left(\frac{x_i}{g(\mathbf{x})}\right) - log\left(\frac{y_i}{g(\mathbf{y})}\right)\right)^2\right]^{1/2} \tag{1.18}$$

$$= \left[\sum_i^D \left(log\left(\frac{x_i}{y_i}\right) - log(g(\mathbf{x})) + log(g(\mathbf{y}))\right)^2\right]^{1/2} \tag{1.19}$$

$$= \left[\sum_i^D \left(log\left(\frac{x_i}{y_i}\right) - \frac{\sum_i^D log\mathbf{x}}{D} + \frac{\sum_i^D log\mathbf{y}}{D}\right)^2\right]^{1/2} \tag{1.20}$$

Set $\mathbf{u} = \mathbf{q} \odot \mathbf{x}$ and $\mathbf{v} = \mathbf{q} \odot \mathbf{y}$. Equation 1.20 then becomes,

$$D(u,v) = \left[\sum_i^D \left(log\left(\frac{u_i}{v_i}\right) - \frac{\sum log\mathbf{u}}{D} + \frac{\sum log\mathbf{v}}{D}\right)^2\right]^{1/2} \tag{1.21}$$

$$= \left[\sum_i^D \left(log\left(\frac{q_i x_i}{q_i y_i}\right) - \frac{\sum_i^D logq_i x_i}{D} + \frac{\sum_i^D logq_i y_i}{D}\right)^2\right]^{1/2} \tag{1.22}$$

$$= \left[\sum_i^D \left(log\left(\frac{q_i x_i}{q_i y_i}\right) - \frac{\sum_i^D logq_i + \sum_i^D logx_i}{D} + \frac{\sum_i^D logq_i + \sum_i^D y_i}{D}\right)^2\right]^{1/2} \tag{1.23}$$

$$= \left[\sum_i^D \left(log\left(\frac{x_i}{y_i}\right) - \frac{\sum_i^D logq_i}{D} - \frac{\sum_i^D logx_i}{D} + \frac{\sum_i^D logq_i}{D} + \frac{\sum_i^D y_i}{D}\right)^2\right]^{1/2} \tag{1.24}$$

$$= \left[\sum_i^D \left(log\left(\frac{x_i}{y_i}\right) - \frac{\sum_i^D log\mathbf{x}}{D} + \frac{\sum_i^D log\mathbf{y}}{D}\right)^2\right]^{1/2} \tag{1.25}$$

$$= D(x,y) \tag{1.26}$$

The same distance metric is applied to CPM^* transformed data results in distances which depend on the perturbation $\mathbf{q}$.

$$D(x,y) = \left[\sum_{i=1}^D \left(log\left(\frac{x_i}{\sum \mathbf{x}}\right) - log\left(\frac{y_i}{\sum \mathbf{y}}\right)\right)^2\right]^{1/2} \tag{1.27}$$

$$= \left[\sum_{i=1}^D \left(log(x_i) - log(y_i) - log(\sum \mathbf{x}) + log(\sum \mathbf{y})\right)^2\right]^{1/2} \tag{1.28}$$

$$= \left[\sum_{i=1}^D \left(log\left(\frac{x_i}{y_i}\right) + log\left(\frac{\sum \mathbf{x}}{\sum \mathbf{y}}\right)\right)^2\right]^{1/2} \tag{1.29}$$

Again, let $\mathbf{u} = \mathbf{q} \odot \mathbf{x}$ and $\mathbf{v} = \mathbf{q} \odot \mathbf{y}$. Equation 1.29 then becomes,

$$D(u,v) = \left[\sum_{i=1}^{D}\left(log\left(\frac{u_i}{v_i}\right) + log\left(\frac{\sum \mathbf{u}}{\sum \mathbf{v}}\right)\right)^2\right]^{1/2} \tag{1.30}$$

$$= \left[\sum_{i=1}^{D}\left(log\left(\frac{q_i x_i}{q_i y_i}\right) + log\left(\frac{\sum_{i=1}^{D} q_i y_i}{\sum_{i=1}^{D} q_i x_i}\right)\right)^2\right]^{1/2} \tag{1.31}$$

$$= \left[\sum_{i=1}^{D}\left(log\left(\frac{x_i}{y_i}\right) + log\left(\frac{\sum_{i=1}^{D} q_i y_i}{\sum_{i=1}^{D} q_i x_i}\right)\right)^2\right]^{1/2} \tag{1.32}$$

$$= D(x,y) \iff q_i = k \;\forall\; i \in 1 \ldots D, \tag{1.33}$$

where k is a constant. The distance between two CPM^* transformed compositions will change when both compositions are shifted by the same perturbation unless the perturbation is a constant.

Both the CPM and the CLR transformations suffer from the lack of subcompositional coherance.

1.3.2 Median Normalization

1.3.3 Trimmed Mean of M-values Normalization Method

Robinson and Oshlack (2010) primarily focused on the mapped read constraint when developing their Trimmed-Mean of M-values (TMM) normalization method for RNA sequence data. Like many others [5], they also assume that the majority of genes in an assay are not differentially expressed. For sequencing data Robinson and Oshlack first define the gene-wise log-fold-changes as:

$$M_g = log_2 \frac{Y_{gk}/N_k}{Y_{gk\prime}/N_{k\prime}},$$

where Y_{gk} is the read count for gene g in sample k (here we assume 1 library for each sample) and N_k as the total number of reads for sample k. They then define the absolute expression levels as:

$$A_g = \frac{1}{2} log_2 \left(Y_{gk}/N_k \cdot Y_{gk\prime}/N_{k\prime}\right) \text{ for } Y_{g\bullet} \neq 0,$$

assuming the absolute expression level is the same for the two samples. Both the M-values and A-values are trimmed removing the upper and lower 30% of the M-values and the upper and lower

5% of the A-values (although these values are defaults and can be tailored for a given experiment). The resulting normalization factor for sample k is then a weighted average of the M-values:

$$log2\left(TMM_k^r\right) = \frac{\sum_{g \in G} w_{gk}^r M_{gk}^r}{\sum_{g \in G} w_{gk}^r}$$

where

$$M_{gk}^r = log_2 \frac{Y_{gk}/N_k}{Y_{gr}/N_r},\ Y_{gk}, Y_{gr} > 0, \text{ for reference sample } r,$$

and the weights, w_{gk}^r, are defined as:

$$w_{gk}^r = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}},\ \text{for } Y_{gk}, Y_{gr} > 0.$$

The weights are a result of taking the inverse of the approximate asymptotic variance using the delta method [8]. Note that these calculations omit any genes for which either observed read count is 0; these observations are generally removed by the trimming procedure.

The TMM normalization method attempts to derive a single value, f, such that $S_k = f \cdot S_{k\prime}$, where S_k is the (unobserved) total count for sample k. Unfortunately, this does not adequately account for the compositional nature of the data. In particular, the authors assume that the expression levels will be the same among samples for the majority of genes (probes). However, if we view the data as a composition, even if only a small number of genes differ with respect to expression level then all of the probes will differ with respect to what we actually observe, their *proportion* of the sample. The authors correctly identify that some probes will be under-represented, even if their absolute expression remains unchanged, in samples with a large total RNA output as compared to samples with a smaller total RNA output. This is because these probes will constitute a smaller proportion of the total number of reads, even if the absolute number will not have changed.

The authors do not explicitly reference Aitchison, or any other compositional data publications, but their method relies heavily on these methods. The TMM operation first converts all samples to the unit simplex by dividing each gene specific read count by the total counts for that sample (Y_{gk}/N_k). The full set of M-values calculated from these compositions represent the perturbation, p from a reference sample to another. However, the M-values are trimmed so p is incomplete and we

are now in a subcompositional sample space. The assumption that most genes are not differentially expressed leads to the authors to use a weighted average of the M-values (p_i's) to determine a single correction factor which can translate one composition to another. This method, therefore, then assumes $p_i = p_j$ for *most* of the genes i and j in the trimmed set.

The authors fail to formulate the problem in terms of the components of a composition, instead, believing that inferences can still be made on absolute abundances. This is clearly evidenced in their definition of the A_g's, which they term the 'absolute abundances' for each gene in the sample. As stated in section 1.1.3, RNA sequence data is compositional in nature and, as such, inferences on the absolute abundances are not possible.

Chapter 2

Proposed Research

2.0.1 Composition based quality control for targeted RNA-Seq

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 [25, 28]. Recent studies evaluating RNA-seq have found generally high intra-platform and inter-platform congruence across multiple laboratories [**Li2013**, 26, 25]. 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.

2.0.2 Paper 2- Compositional data for biologists

2.0.3 Paper 3- Telescoping amalgamations

Bibliography

[1] J. AITCHISON. "Principal component analysis of compositional data". In: *Biometrika* 70.1 (1983), pp. 57–65. ISSN: 0006-3444. DOI: 10.1093/biomet/70.1.57. URL: http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/70.1.57 (cit. on p. 19).

[2] 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. 8, 11, 13, 16).

[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. 12).

[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. 12, 13).

[5] 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. 7, 21).

[6] 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. 13).

[7] M S Boguski, C M Tolstoshev, and D E Bassett. "Gene discovery in dbEST." In: *Science (New York, N.Y.)* 265.5181 (1994), pp. 1993–4. ISSN: 0036-8075. URL: http://www.ncbi.nlm.nih.gov/pubmed/8091218 (cit. on p. 5).

[8] George. Casella and Roger L. Berger. *Statistical Inference, 2nd Edition*. Australia ; Pacific Grove, CA : Thomson Learning c2002, 2002. ISBN: 0534243126 (cit. on p. 22).

[9] Daniela S Gerhard et al. "The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC)." In: *Genome research* 14.10B (2004), pp. 2121–7. ISSN: 1088-9051. DOI: 10.1101/gr.2596504. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=528928{\&}tool=pmcentrez{\&}rendertype=abstract (cit. on p. 5).

[10] Travis C Glenn. "Field guide to next-generation DNA sequencers." In: *Molecular ecology resources* 11.5 (2011), pp. 759–69. ISSN: 1755-0998. DOI: 10.1111/j.1755-0998.2011.03024.x. URL: http://www.ncbi.nlm.nih.gov/pubmed/21592312 (cit. on p. 6).

[11] Brian J Haas et al. "How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes?" In: *BMC Genomics* 13.1 (2012), p. 734. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-734. URL: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-734 (cit. on p. 7).

[12] Illumina. "Buyer's Guide: Simple, Customized RNA-Sequencing Workflows Overcome Limitations with RNA-Seq". In: (). URL: http://www.illumina.com/content/dam/illumina-marketing/documents/products/other/rna-sequencing-workflow-buyers-guide-476-2015-003.pdf (cit. on p. 5).

[13] 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. 16).

[14] 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. 7, 15, 17).

[15] Joshua Z Levin et al. "Comprehensive comparative analysis of strand-specific RNA sequencing methods." In: *Nature methods* 7.9 (2010), pp. 709–15. ISSN: 1548-7105. DOI: 10.1038/nmeth.1491. URL: http://www.ncbi.nlm.nih.gov/pubmed/20711195http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3005310 (cit. on p. 6).

[16] Lin Liu et al. "Comparison of Next-Generation Sequencing Systems". In: *Journal of Biomedicine and Biotechnology* 2012 (2012), pp. 1–11. ISSN: 1110-7243. DOI: 10.1155/2012/251364. URL: http://www.hindawi.com/journals/bmri/2012/251364/ (cit. on p. 6).

[17] Yichuan Liu et al. "Evaluating the Impact of Sequencing Depth on Transcriptome Profiling in Human Adipose". In: *PLoS ONE* 8.6 (2013). Ed. by Zhanjiang Liu, e66883. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0066883. URL: http://dx.plos.org/10.1371/journal.pone.0066883 (cit. on p. 7).

[18] Yuwen Liu, Jie Zhou, and Kevin P White. "RNA-seq differential expression studies: more sequence or more replication?" In: *Bioinformatics (Oxford, England)* 30.3 (2014), pp. 301–4. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt688. URL: http://www.ncbi.nlm.nih.gov/pubmed/24319002http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3904521 (cit. on p. 7).

[19] 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 p. 7).

[20] 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. 15).

[21] 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 p. 13).

[22] 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. 7).

[23] 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. 7).

[24] 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. 16).

[25] 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. 25).

[26] 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. 25).

[27] 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. 7).

[28] 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. 25).

[29] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics." In: *Nature reviews. Genetics* 10.1 (2009), pp. 57–63. ISSN: 1471-0064. DOI: 10.1038/nrg2484. arXiv: NIHMS150003. URL: http://www.ncbi.nlm.nih.gov/pubmed/19015660 (cit. on p. 5).