INSERT_TITILE_HERE

Jia Rong Wu, Dr. Gregory Gloor

June 24, 2016

1 Abstract

High throughput sequencing is a technology that allows for the generation of millions of reads of genomic data regarding a study of interest. Subsequent analysis of this data can yield information ranging from transcription profiles, microbial diversity, or even cellular abundance in culture. Despite the compositional nature of sequencing data, many tools inappropriately treat the underlying features in sequencing data as counts and thus model them as such. This paper proposes a log-ratio transformation method that allows for feature comparisons between samples in a compositional context. The log-ratio transformation will be implemented as an extension to a general compositional data analysis tool known as ALDEx2 or ANOVA Like Differential Expression. It will be demonstrated using both simulated and real data that the (insert name here) log-ratio transformation provides results that are more robust and in line with expectations relative to ALDEx2.

2 Background

RNA-Seq is one fo the many tools that uses high throughput sequencing technology in order to generate information regarding the transcriptome of an organism. Traditionally, samples are comprised of a set of features whose identity depends on the experimental design. For example, features can considered genes in the case of RNA-seq and differential expression or operational taxonomic units (OTUs) when the objective is identifying microbial diversity. Count based data for RNA-Seq must fit the assumption that it is independent and identically distributed for statistical tests [2, 7]. However, in the context of RNA-Seq, counts per feature are not independent in the sense that they are dependent on the sequencing capacity of the machine. The probability of detecting rare transcripts in RNA-Seq is lower in the presence of highly expressed transcripts [8]. Absolute feature counts returned from a sequencing machine are uninformative due to the violation of the assumption of independence. Thus, RNA-Seq data is inherently compositional and must be treated as such. Traditional tools do not address the compositional nature of RNA-seq data [5].

Formally, a composition is defined as a vector whose components are proportions. These proportions sum to an arbitrary constrained constant c. This constant c is arbitrary in the sense that it can represent 1 if the components are proportions, 100 if the components are percentages or other constants such as 10^9 for parts-per-billion. Absolute values of components in a composition are uninformative. The only information provided in compositional data is the relative magnitudes of the components ratios between all pairs of components [1]. As compositional data carries relative information between components, analyzing the data as log-ratios is suggested by Aitchison [1]. The scale-invariant Centered-Log-Ratio (CLR) transformation proposed by Aitchison is defined as:

$$\begin{aligned} y &= \{\frac{x_i}{g(x)}\}_{i=1,\ldots,D} \\ \text{where } y &= \text{A composition transformed by CLR} \\ x &= \text{A component of the non-transformed composition (x)} \\ D &= \text{The number of components of x} \\ g(x) &= \text{Geometric mean of D components of x} \end{aligned} \tag{1}$$

For a given sample, ALDEx2 considers the sample to be a composition, and its features to be the components [5]. For each feature in a sample, ALDEx2 converts its count into a set of probabilities through Monte Carlo sampling from the Dirichlet distribution. Monte Carlo methods are defined as repeated random sampling from a probability distribution. Sampling features as Monte-Carlo instances from a Dirichlet distribution allows for sampling variance to be modeled [4]. Furthermore, the sum of each instance per sample is equivalent to one, making the features proportions. ALDEx2 transforms each proportional instance into log-ratios with Equation 1. This compositional approach is appropriate for any data derived from a high throughput sequencing machine.

This transformation normalizes the features of each sample to the geometric mean of each sample, and allows for comparisons of expression between and within RNA-Seq samples [3]. The CLR-transformed data accounts for differences in sequencing depth between samples [4, 6]. Much like how fluorescence in a qPCR

reaction provides an internal standard to determine the quantity of amplified DNA, the geometric mean provides an internal standard to determine the abundance of a feature as a ratio to all other features in the sample. Taking the CLR of each feature ensures a one-to-one correspondance between features in the composition. This allows for feature changes to be observed relative to the geometric mean [5].

(Describe the effect plot here)

3 Problem

The assumption being made during the CLR transformation is that most features are either invariant or randomly varying. If there is systematic variation between groups simply due to presence or absence of features, then this assumption is broken. For example, when comparing microbial diversity between sampling sites, organisms present in one subsite may be absent from another. In the case of transcriptomics, organisms that reside in one condition may have a different expression profile than those that reside in a second condition. These differences are represented by either zeroes or low count features between samples. It will be demonstrated that exaggerated differences can be achieved with seemingly little variation within simulated data. Additionally, zeroes are problematic as they cannot be represented in logarithmic space. Therefore, prior to the log ratio transformation, zero count features are discarded if present in exclusive or conditions, or adjusted with an uninformative prior. Ideally, the goal is to identify a non-deterministic geometric mean that best represents each sample so features can be accurately compared.

In its current implementation, ALDEx2 computes a per-sample geometric mean for a samples' set of features and declares this as the baseline for feature comparisons. However as demonstrated by Figure 1, systematic variation between groups can influence the baseline such that the comparison between features is biased. In this case the majority of features cluster below the location 0. These figures represent a RNA-Seq experiment simulated such that 119 samples from one condition contain genes that are completely absent in the second condition. The assumption that most features are invariant is broken as the histogram of the mean expression per sample is not centered about the location 0 as depicted in Figure 1b. This shift of mean expression is problematic as it may lead to exaggerated differences and in turn, false positives. If there are a large proportion of features in a sample that do not follow the central tendency of the data, the geometric mean can be unreliable as a baseline. Therefore, identifying the features with the most consistent variance across all samples and generating the geometric mean from this subset will account for this systematic variation.

4 Methods

The geometric mean is the nth root of the product of a set of n numbers, and it represents the central tendency of the set of numbers. ALDEx2 computes the geometric mean on a per-sample basis using the entire set of features that constitute that sample. This method has been demonstrated to be effective on datasets in which there is no systematic variation across the entire dataset. The proposed (insert name here) log-ratio transformation subsets the features such that only the sets of features with the most typical variance are considered in the computation of the geometric mean. For each sample, the centered-log ratio transformation is applied to its set of features. The per-feature variance is computed across the set of log-ratio transformed data, and the resulting features whose variance falls into the inter-quartile range is declared as the invariant set. These are the features that are used to compute the (robust?) geometric mean.

4.1 Simulated Data

A series of RNA-Seq simulation data was generated for benchmarking purposes. Assemblies from *Saccharomyces cerevisiae* uid 128 and a complete reference genome of *S. cerevisiae* were drawn from NCBI. Using R package `polyester`, an RNA-Seq experiment with 2 groups of 10 replicates was generated, with a 20x sequencing coverage across the simulation experiment. Using bowtie2 these simulated reads were aligned to the *S. cerevisiae* reference genome. Labeling each group as A and B is arbitrary and hence samples s_1, s_2, ..., s_10 belong to condition A, and samples s_11, s_12, ..., s_20 belong to condition B. There are a total of

6349 features in these simulated data. On average, samples in the unmodified base dataset have a simulated read depth of approximately 870000 reads.

For the basic simulated dataset, all features in condition A were simulated to have no fold differences, where features 1-10, 11-20, 21-30, 31-40 in condition B were simulated with a fold difference of 2, 3, 4, 5 respectively. These 40 features serve as an internal control of true positives for each dataset as their fold changes are explicit and should always be displayed as differentially expressed. Figure 4a is an effect plot demonstrating how the 40 internal control features are considered statistically significant and differentially expressed. Systematic variation is simulated in an additional

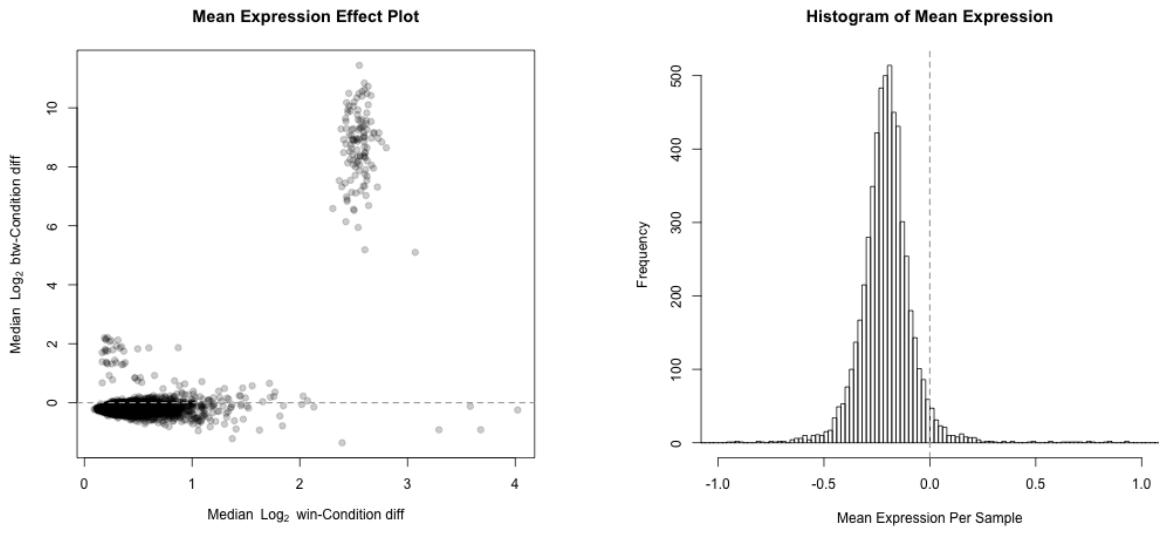
An additional dataset with systematic variation

To rationalize computing on a subset (Add this in supplementary or supporting figs) For example, a set of 135 random variates is drawn with a population mean of 20 and a standard deviation of 2 is drawn. An additional 15 variables with a value of 0.5 are added to simulate a 10% sparsity across the 150 variables. Computing the geometric mean across this set of 150 variables yields a value that fails to estimate the centrality of the data. However, taking only the subset of features whose variance falls within the interquartile range of the variance allows for a geometric mean that is more representative of the sample as demonstrated in Figure S1

5 Discussion

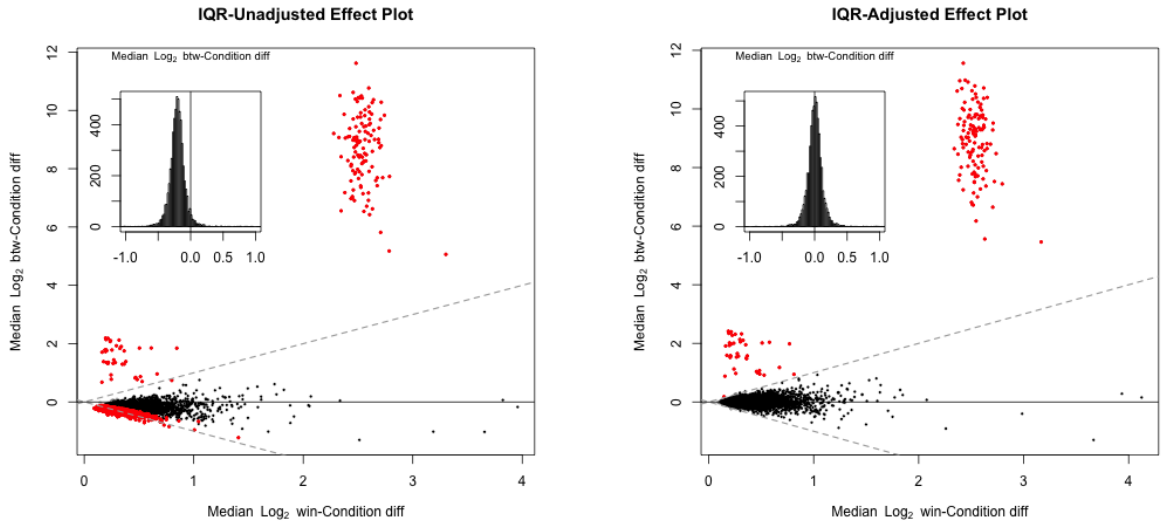
Given a dataset with no systematic variation, (insert name here) log-ratio transformation will provide a result that is consistent with the centred log-ratio transformation. Using the set of invariant features which can be considered a trimmed estimator over the set of all features yields a geometric mean that is for all intents and purposes equivalent to the one generated by the set of all features.

6 Figures



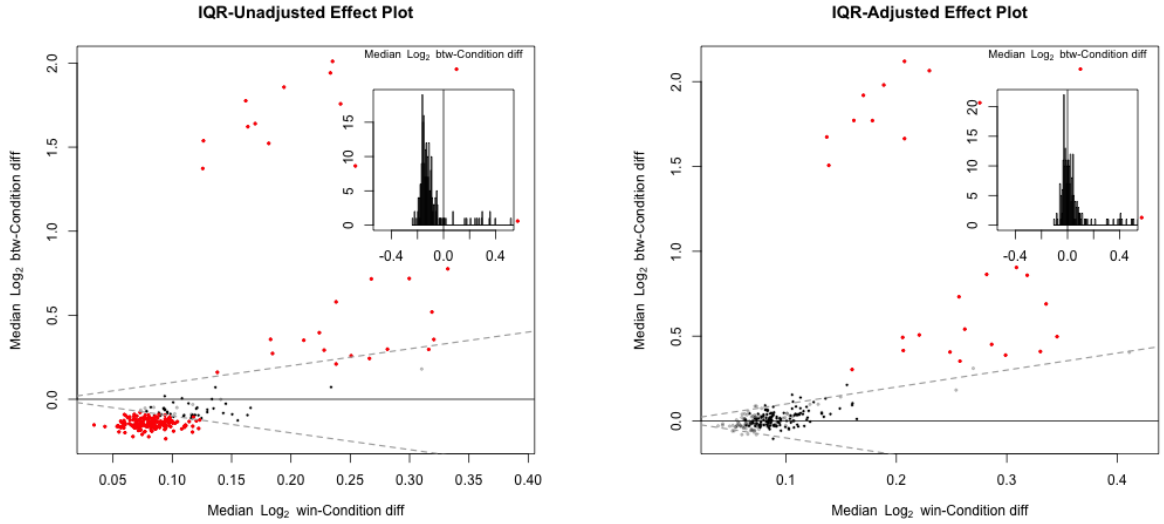
(a) (b)

Figure 1: 1a: Effect plot depicting how the majority of features are shifted below the location y=0. This is an artifact that reflects how dissimilar the samples are, and must be accounted for in order to make accurate comparisons. 1b: Histogram of the mean expression per sample. In ideal datasets, the mean expression per sample should be centred about the location x=0 however due to systematic variation, the centering is off.



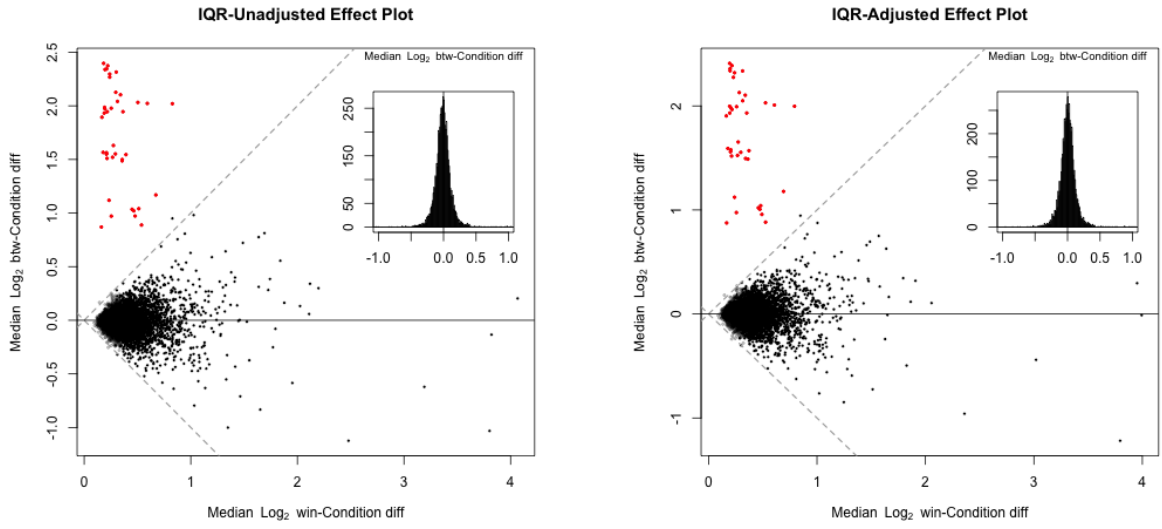
(a) (b)

Figure 2: 2a: Effect plot and mean expression per sample histogram that depict how computing the geometric mean can be inappropriate if computed across the set of all features. 2b: Effect plot and mean expression per sample histogram of a dataset using the invariant set features.



(a) (b)

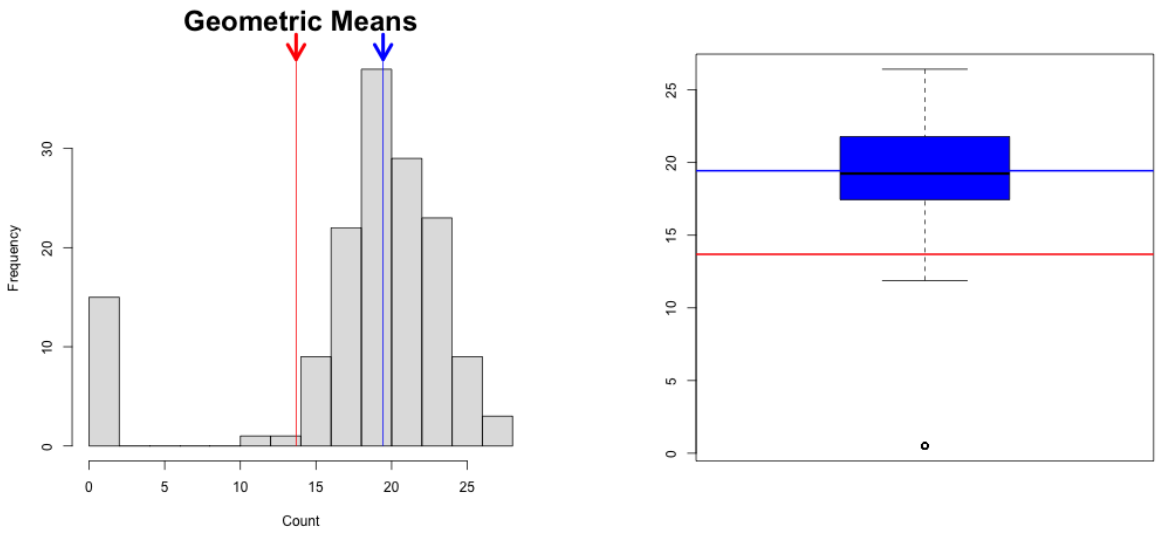
Figure 3: Caption for Fig 3 a_b



(a) (b)

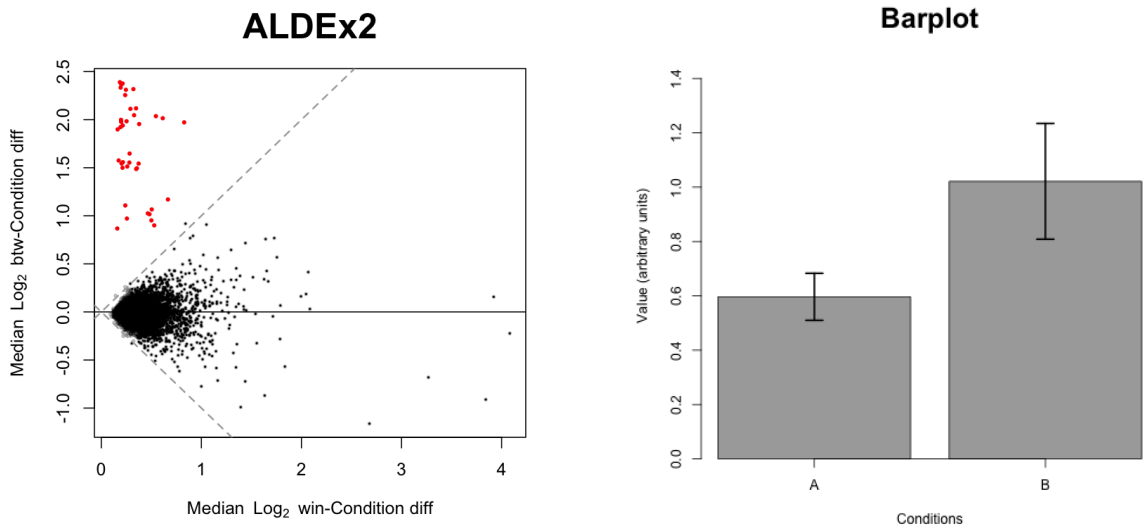
Figure 4: Effect plots and mean expression per sample histograms demonstrating how the IQR log-ratio transformation does NOT affect data that is unaffected by systematic variation.

7 Supplementary Information



(a) (b)

Figure S1: S1a: A histogram of 135 random variates with a population mean of 20 and a standard deviation of 2. 15 additional values of 0.5 were added to simulate a 10% sparsity across the 150 variables. Red represents the geometric mean of the entire set of features. Blue represents the geometric mean of the variates that fall within the IQR of the dataset. S1b: Blue is the geometric mean of the IQR features, Red is the geometric mean of the entire set of features.



(a) (b)

Figure S2: S2a: An ALDEx2 effect plot. Each individual point can be considered a feature whose location on the plot explains its effect size. The X-axis refers to dispersion, or the error bars in a barplot. The Y-Axis refers to difference, or the absolute difference between conditions A and B on the bar chart. Points that fall above the line of y=x have an effect size >1, where points that fall below the line of y=-x have an effect size of <-1. Red features are ones denoted as statistically significant by a Welch's T-test.

References

[1] J. Aitchison. *The statistical analysis of compositional data.* Chapman & Hall, 1986. ISBN: 0-412-28060-4.

[2] P. L. Auer and R.W. Doerge. "Statistical Design and Analysis of RNA Sequencing Data". In: *GEnetics* 185 (2 2010). DOI: doi:10.1534/genetics.110.114983..

[3] Marie-Agne's Dillies et al. "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis". In: *Briefings In Bioinformatics* 14 (6 2012).

[4] A.D. Fernandes et al. "ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq." In: *PLoS ONE* 8 (7 2013).

[5] A.D. Fernandes et al. "Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis." In: *Microbiome* 2 (15 2014).

[6] D. Lovell et al. "Proportions, percentages, ppm: do the molecular biosciences treat compositional data right?" In: *Compositional Data Anal: Theory Appl.* (2011).

[7] J.H. McDonald. *Handbook of Biological Statistics (3rd ed.)* Sparky House Publishing, pp. 77–85.

[8] S. Tarazona et al. "Differential expression in RNA-seq: A matter of depth." In: *Genome Research.* 22 (12 2011). DOI: doi:10.1101/gr.124321.111.