Finding the centre:
correctioning asymmetry in high-throughput sequencing data

Jia R. Wu, Jean Macklaim, Briana L. Genge, Gregory B. Gloor
The University of Western Ontario, Department of Biochemistry
ggloor@uwo.ca : ggloor.github.io : github.com/ggloor/ALDEx2

Abstract

High throughput sequencing (HTS) generates millions of reads of genomic data regarding a study of interest, and data from high throughput sequencing platforms are count compositions. Subsequent analysis of such data yields information on transcription profiles, microbial diversity, or even relative cellular abundance in culture. Because of the high cost of acquisition, the data are usually sparse, and always contain far fewer observations than variables. However, an under-appreciated pathology of these data are their often unbalanced nature: i.e, there is often systematic variation between groups simply due to presence or absence of features. This variation is important to the biological interpretation of the data and causes samples in the comparison groups to exhibit widely varying centres. This work extends a previously described log-ratio transformation method that allows for variable comparisons between samples in a Bayesian compositional context.

Introduction to the problem

Aitchison [1] defined both the additive log-ratio and centred log-ratio normalizations. The CLR is at least in theory, scale-invariant because if the parts of $\boldsymbol{x}$ are counts with $\alpha = N$ reads, then:

$$\mathbf{x}_{clr} = log\big(\frac{Nx_i}{g(N\mathbf{x})}\big) = log\big(\frac{x_i}{g(\mathbf{x})}\big). \quad (1)$$

The caveat that is that the total read count, α, for each observation must be roughly similar.

$$\mathbf{x}_{alr} = log\big(\frac{x_i}{x_D}\big)_{i=1\ldots D-1} \quad (2)$$

For the ALR, the denominator is the D^{th} (constant) feature of $\boldsymbol{x}$.

The ALR is surprisingly similar to the qPCR approach in common use in molecular biology that measures relative abundance of molecules in a mixture where the feature of unknown abundance is determined relative to the abundance of a feature of (presumed) known abundance. It is well known that the relative abundance measure will change when a different species is used as the denominator; this also occurs if the denominator of the clr is not equivalent between samples or groups. The ALR and CLR can be viewed as the limits of a continuum of incomplete knowledge about the proper internal standard, or basis, by which relative abundance should be judged. We can however, choose to use combinations of other features as the basis.

The goal is to choose a basis that is invariant in a sparse, asymmetric dataset

Differential (relative) abundance with ALDEx2 [3, 4]

If we have a sample set composed of n samples with D parts. We assume that the dispersion of the j^{th} part across the samples contains both biological variation and random variation: i.e. $\tau_j = \nu_j + \epsilon_j$. We assume that the biological variation, ν_j, is encapsulated in the samples, but that the random variation, ϵ_j, derives from a multivariate random Poisson process with a defined total; i.e. a Dirichlet. Then if $S_i = [s_1, s_2, \ldots s_D]; \sum S_i \in 0 \rightarrow \infty$, the probability, $\mathbb{P}_{ij}$, of observing the counts for feature S_{ij} is similar to the fractional f_{ij} that feature j represents in S_i conditioned on the total read depth for the sample. $\mathbb{P}_{ij}(f_{ij}|\alpha_i)$ can be calculated as;

$$\mathrm{P}_{i(1\ldots k)} = \begin{pmatrix} \boldsymbol{p}_1 \\ \boldsymbol{p}_2 \\ \vdots \\ \boldsymbol{p}_k \end{pmatrix} = \begin{pmatrix} p_{i,11} & p_{i,21} & p_{i,31} & \cdots & p_{i,D1} \\ p_{i,12} & p_{i,22} & p_{i,32} & \cdots & p_{i,D2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{i,1k} & p_{i,2k} & p_{i,3k} & \cdots & p_{i,Dk} \end{pmatrix} \sim Dirichlet_{(1\ldots k)}(\boldsymbol{s}_i + 0.5) \quad (3)$$

Note that $\alpha_i \neq \sum S_i$ and in fact, the former contains no information about the latter.

This approach has consistent sampling properties and removes the problem of taking a logarithm of 0 when calculating the CLR (row-wise) since the count 0 values are replaced by positive non-zero values that are consistent with the observed count data. Each of the Monte-Carlo instances, by definition, conserves proportionality and accounts for the fact that there is more information when α_i, S_{ij} are large than when they are small.

Summary statistics from the distribution of CLR values for each feature are calculated and reported as either expected values or as medians of the distributions. If we have two groups, A and B, where the indices of the samples in the first group are $1 \ldots i_a$ and the indices of the samples in the second group are $i_{a+1} \ldots n$, then the distributions of CLR values for the j^{th} feature of the two groups can be contained in the vectors: $\boldsymbol{a}_j = \mathrm{C}_{(1\ldots i_a)j(1\ldots k)}$ and, $\boldsymbol{b}_j = \mathrm{C}_{(i_{a+1}\ldots n)j(1\ldots k)}$. Summary statistics use for plotting and analysis are:

- Log-ratio abundance of a feature is the median of the joint distribution of CLR values from groups A and B; i.e., it is the median of $\boldsymbol{a}_j \cup \boldsymbol{b}_j$.
- Dispersion is the median of the vector $\Delta_{\boldsymbol{a}_j \vee \boldsymbol{b}_j} = maximum(|\boldsymbol{a}_j - \boldsymbol{a}_{\langle j \rangle}|\ ,\ |\boldsymbol{b}_j - \boldsymbol{b}_{\langle j \rangle}|)$, where $\langle j \rangle$ indicates a random permutation of the vector. The reported dispersion for each feature is denoted as $\tilde{\Delta}_{\boldsymbol{a}_j \vee \boldsymbol{b}_j}$ and is a conservative surrogate for the median absolute deviation when $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$ contain many entries.
- Difference between groups is the median of the vector $\Delta_{\boldsymbol{a}_j - \boldsymbol{b}_j} = (\boldsymbol{a}_j - \boldsymbol{b}_{\langle j \rangle})$, i.e., $\tilde{\Delta}_{\boldsymbol{a}_j - \boldsymbol{b}_j}$.
- Effect size for a given feature is the median the vector derived from $\Delta_{\boldsymbol{a}_j - \boldsymbol{b}_j} / \Delta_{\boldsymbol{a}_j \vee \boldsymbol{b}_j}$, and is thus a standardized difference between the distributions in $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$.

Results

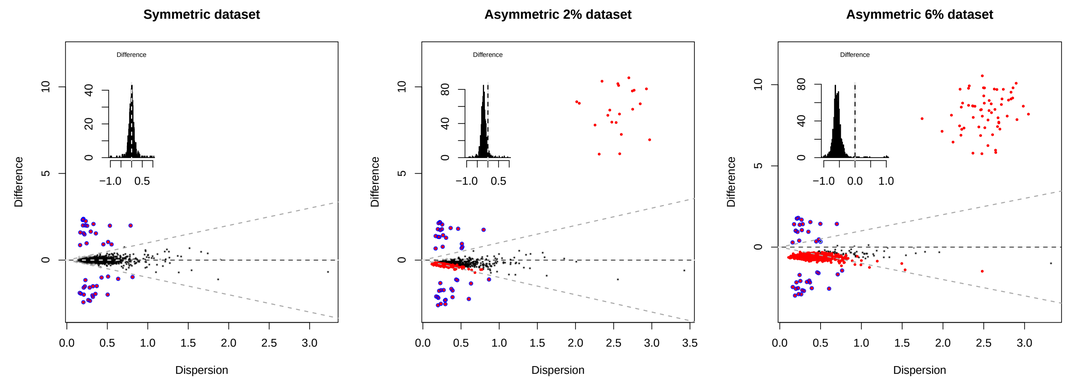

Figure 1: The effect plots [5] show the difference between two conditions in simulated RNA-seq datasets with 1000 genes where 40 genes are modelled to have true difference between groups. Each point is a feature (gene), they are coloured in black if not different between groups, red if significantly different between groups, and red with a blue circle if they are one of the 40 genes modelled to be true positives. The red points in the top right quadrant are the genes modelled to be asymmetrically variable between groups. The inset histograms show the distribution of the differences between groups as calculated by ALDEx2, and the vertical line shows a difference of 0. These x-axis of these plots are truncated to show only differences near the midpoint. The centre of the dataset must be on 0 to prevent the inclusion of false positive and the exclusion of false negative differential abundance identifications.

Proposed alternative denominators

In all cases the geometric mean of the set of features is used and substituted for $g(x)$ in equation 1.

1. **all** This is the entire set of features; i.e., the CLR
2. **IQLR** This is the set of features that have variance in the dataset between the first and last quartiles
3. **zero** This is the set of non-zero features in group A, or in group B. Groups A and B thus have different sets of features composing their denominators
4. **User** This is a set of user-defined features. In Figure 2, these features are the ribosomal protein functions. This is conceptually the ALR with a small number of features instead of a single one

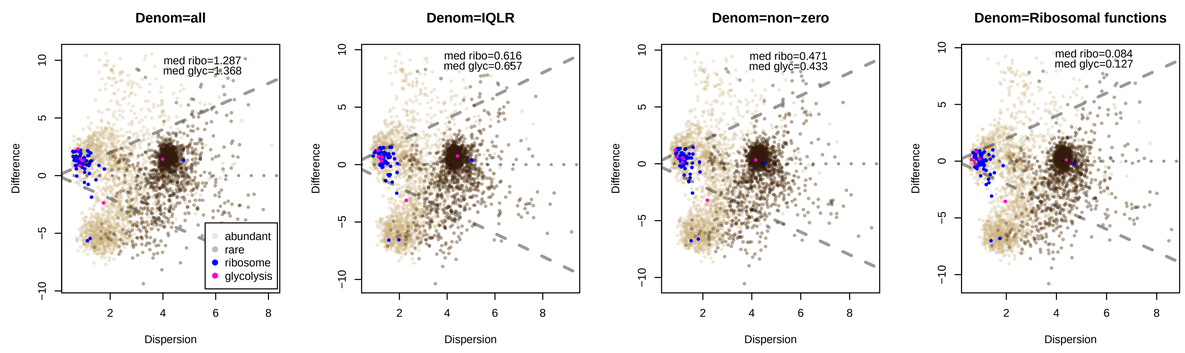

Figure 2: Effect plot and Bland-Altman plots summarizing gene expression in two different states from a meta-transcriptome of health and bacterial vaginosis. Each point is a function, coloured dark brown if rare, light brown if abundant, blue if part of the ribosome, magenta if part of glycolysis. The latter two sets of functions are presumed housekeeping and should be constant between conditions. The mean displacement of these two functional groups is noted.

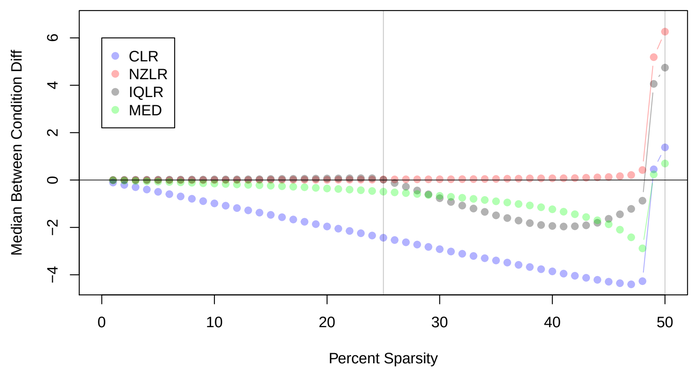

Figure 3: Each point represents the median between condition difference for a given transformation in a dataset with a specified sparsity. Points closer to the location y=0 are favourable. The CLR transformation fails as soon as asymmetric sparsity is introduced. The IQLR transformation is effective on datasets with up to 25% asymmetric sparsity from zeroes or extreme count features. The NZLR transformation is effective on datasets with up to 50% sparsity . Replacing the geometric mean with the median in Equation 1, is an improvement, but results in a generally small shift in midpoint.

Conclusions

- Asymmetric data is problematic and must be corrected
- The IQLR approach is the best all purpose starting approach: the zero method is not recommended since sparsity may not be the precipitating issue: in extreme cases such as Figure 2, only a user-defined denominator may be appropriate
- These approaches have be included in versions of ALDEx2 since 1.6, available on Bioconductor

References

[1] J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, 1986.

[2] John Aitchison and Michael Greenacre. Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(4):375–392, 2002.

[3] A. D. Fernandes, J. M. Macklaim, T.G Linn, G. Reid, and G. B. Gloor. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-seq. *PLoS ONE*, 8(7):e67019, July 2013.

[4] Andrew D Fernandes, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16s rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2:15.1–15.13, 2014.

[5] Gregory B. Gloor, Jean M. Macklaim, and Andrew D. Fernandes. Displaying variation in large datasets: Plotting a visual summary of effect sizes. *Journal of Computational and Graphical Statistics*, 25(3):971–979, 2016.

Acknowledgements

Funded by a grants to GG from the National Science and Engineering Research Council of Canada and Agriculture and Agrifoods Canada