Supplement: Beyond compositionally in high throughput sequencing; estimating the importance of scale in data analysis with ALDEx2

Greg Gloor, Michelle Pistner Nixon, Justin Silverman *[1]

[1] Dep't of Biochemistry, University of Western Ontario, Penn State

*`ggloor@uwo.ca`

21 August 2023

Abstract

Introduction to scale simulation and FDR correction with ALDEx2.

Package

ALDEx2 1.33.1

Contents

1 GM correlates with Shannon's entropy in HTS datasets

We can think about the scale relationship from an information theoretic point of view. Empirically in most datasets the geometric mean of $Y_n^{\parallel}$ is strongly correlated with Shannon's Entropy H (supplemental table and supplemental Figure 1) suggesting that some knowledge of W is contained in the post-sequencing data that is not strictly compositional. Intuitively, underlying systems with different scales will contain different amounts of information and so we would expect $W_n^{\perp} \sim H_n$.

The logarithm of the geometric mean $\log(G)$ and H can be understood from an information theoretic point of view to be an unweighted G or a weighted measure H of 'surprisal' in the dataset. Entropy H and the geometric mean G (or its logarithm $\log_2 G$) are not simple to relate algebraically, but they can be understood in terms of what they are measuring if their description is rephrased in a common language. Recall have a $D \times N$ matrix of counts W decomposed into the proportions for the n^{th} sample $W_n^{\parallel}$ (or the equivalent probability distribution $p(w_n)$), and its scale $W_n^{\perp}$.

For notational simplicity assume a single discrete random variable X with a probability distribution $p(x)$ over $1 \dots d$ features. The entropy $H(X)$ in bits is:

$$H(X) = -\sum_{i=1}^{d} p_i \log_2 p_i$$

and for the same distribution log2 of the geometric mean G is:

$$\log_2 G = \frac{1}{d} \sum_{i=1}^{d} \log_2 p_i$$

As defined here $H(X)$ is a weighted total of the uncertainty or 'surprisal' contained in $p(x)$, and relates to the amount of information we would need to have in order to reproduce $p(x)$. Conversely, G is an unweighted mean of the same distribution. G is used as the denominator to calculate the centred log-ratio normalization (clr):

$$clr = \log_2(p_i) - \log_2 G$$

over $i = 1 \dots d$. This form shows explicitly that each p_i is compared with the geometric mean G. Thus G can be interpreted as a measure of difference for how far each p_i is from G, or in information theoretic terms as an unweighted measure of the mean surprisal for the distribution.

Thus, we can thus understand $H(X)$ as a measure of the total weighed surprisal and G as a measure of the average unweighted surprisal for $p(x)$. The total and the mean surprisal are related by the number of terms in $p(x)$ and by the weighting factor for each term.

These two measures are expected to have different behaviours in different distributions of p_i. In the case of a uniform distribution both $H(X)$ and G are maximal since $p(x)$ is equally and identically distributed. Thus, we expect that they are positively correlated here. In a Normal or a skewed distribution, we also anticipate a positive correlation because both are affected in the same direction by outlier values. In very sparse datasets, the two measures could become uncoupled because $H(X)$ could ascribe some uncertainty to the large number of low probability events, while G would tend to be very small. Here these two measures could be either uncorrelated or exhibit negative correlation. We can see this distributional behaviour in different datasets.

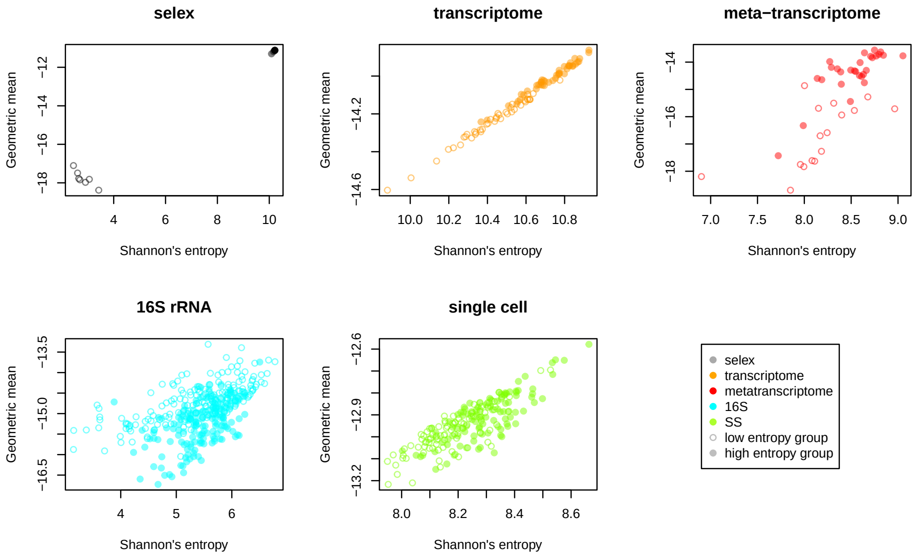
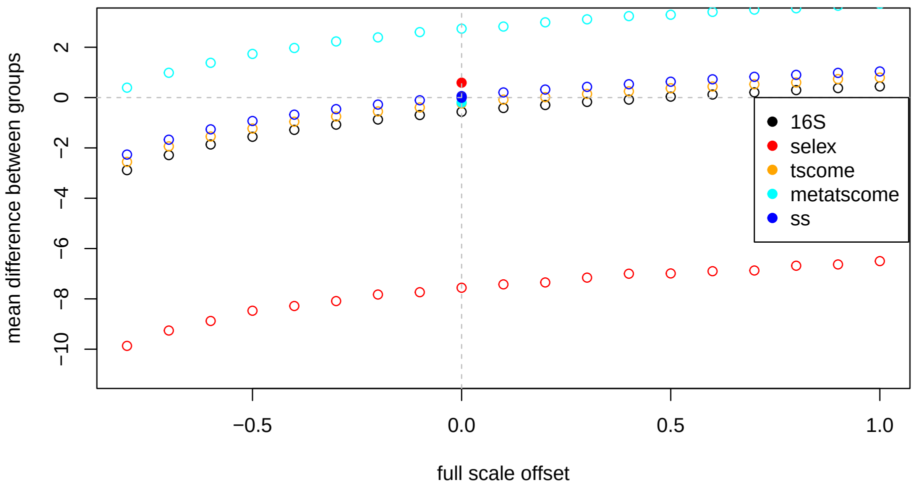


Figure 1: **Plot of Shannon's entropy (H) vs geometric mean (G) for each sample in different datasets** The groups that each sample belong to are highlighted as filled or open circles. Each group in each dataset has different entropy with the groups in the selex and metatranscriptome datasets being highly distinct.

The table below summarizes the mean values for, and the correlation between, G and H (cor) and the sparsity defined as the proportion of features with less than 1 count per sample (spar) for each association in each group of samples:

Dataset	group	$\overline{G}$	$\overline{H}$	cor	spar
Selex	control	-11.2	10.2	0.99	0
"	selected	-17.8	2.8	-0.88	0.802
yeast	snf2 ko	-14.0	10.7	0.99	0.004
"	WT	-14.2	10.4	0.99	0.007
Meta	H	-18.8	8.6	0.78	0.451
"	BV	-18.2	8.9	0.79	0.238
16S	Pup	-14.7	5.4	0.68	0.079
"	Cent	-15.2	5.4	0.53	0.251
SS	A	-13.0	8.2	0.83	0.978
"	B	-12.9	8.3	0.80	0.977

We can see that for most datasets the difference between the $\overline{G}$ in each group is relatively small. Most significantly, the selex dataset has a very large difference of about 64 fold, and both the 16S and the metatranscriptome dataset have about a 1.5 fold difference. These datasets are candidates for a full scale model correction.



The full scale model option allows the investigator to set both the offset between the geometric means of the features and their dispersion. In the offset plot above we can see the cause and the effect of the full model with a fixed gamma of 0.5. We see that the mean location of well centred datasets (yeast transcriptome, single-cell transcriptome and 16S rRNA dataset) are relatively well centred, but could be centred better with small changes in scale ranging from 0 (single cell) to ~0.1 for the 16S and yeast datasets. In contrast, centring the metatranscriptome dataset would require about a 50% change in relative scale between groups, but as shown in the main text, centring the housekeeping genes is more apt.

The in vitro selection dataset is clearly an outlier in both the difference between the average group geometric mean, and in the offset plot. However, this dataset can also be used to illustrate the power of the full scale model. In Figure 3 we can see that the default output of ALDEx2 has a centered output. This occurs largely by chance, as the high sparsity of the selected (S) group is balanced almost exactly by the arbitrarily chosen sequencing depth. In this dataset the relative abundances of the majority of features are invariant, but this is masked by the large absolute changes in a small number of features (McMurrough et al. 2014). Neither DESeq nor edgeR are able to provide a reasonable analysis of this dataset (Gloor et al. 2016).

A full scale model, where the strong assumption that the mean G is assumed to be the same between the two groups, dramatically skews the output and the large number of relatively invariant features are now identified as significantly different. While not wrong as long as the assumption that the scales are identical is stated, this is not a useful analysis outcome. Modifying the mean scale difference between groups to be 50-fold different moves the centre of the large number of relatively invariant features to the centreline of no difference, and recapitulates the default result obtained without the full scale model. Note that we get exactly the same answer (within random sampling error) with a scale of 0.02 for group 1 and a scale of 1 for group 2, or using a scale of 1 for group 1 and a scale of 50 for group 2. This shows that it is the relative difference between scales that is important, not the absolute values. From this result we can conclude that, on average, the difference in underlying scale in the system is about 50-fold, and this is congruent with the circa 64-fold difference in mean G between groups. Thus, an advantage of a full scale model is that we have gained information and understanding about the underlying system.

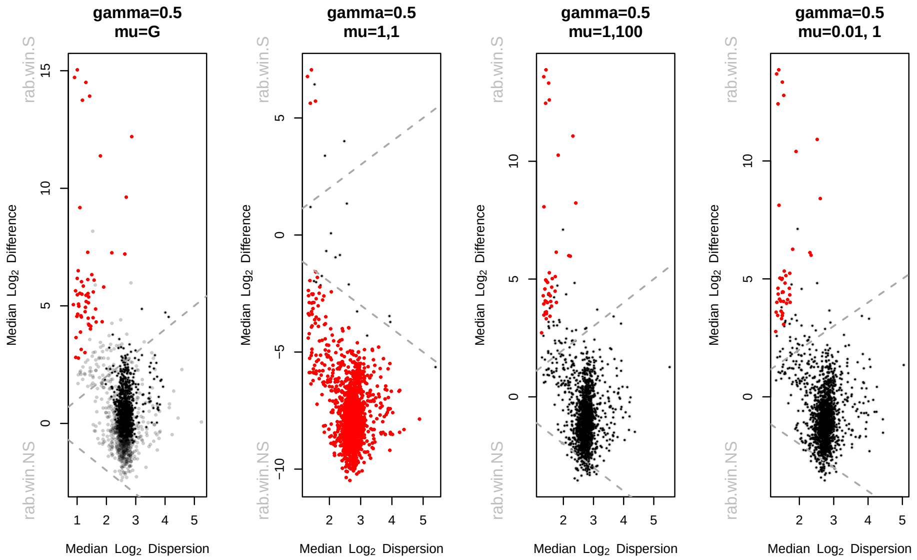


Figure 2: Effect plots of the selex dataset with various gamma and scale parameters

2 Issues with DESeq2 and edgeR

DESeq2 and edgeR are two of the most commonly used tools for differential abundance analysis of bule RNA sequencing datasets. They both operate by finding a scaling factor that makes all the samples commesurate. DESeq2 does this by finding a midpoint feature that can be used as a reference in each sample; this can be different for different samples. The edgeR reference finds the midpoint of the 'typical' sample instead. In both cases the data are then scaled by dividing by a small factor that makes the read counts commesurate. Differential abundance analysis is then performed on the scaled values after taking their logarithm to base 2. In some ways this is similar to the log-ratio approach used by ALDEx2, but is more prone to dataset and sample effects than is the log-ratio method (Gloor 2023)

Examining the plots, edge R clearly not centred with the median housekeeping functions offset by -1.4869633 and minimum FDR value is 8.6385521×10^{-8}. Additionally, there is little range in the p-values relative to those seen for DESeq2, and the values are more in line with the range seen with ALDEx2.

DESeq2 is better centred with median housekeeping functions offset by -1.3531112 but the minimum FDR value is NA. In addition there are a large number of functions with 0 variance, these are very low count functions with very high sparsity in the dataset. These are not differential in the DESeq2 analysis. However, there are a number of functions in the DESeq2 analysis that are very differentially abundant, with a log2 fold change of $<$ -20. Examination of the raw counts shows that these are uniformly 0 or 1 in the H dataset, and present at high counts in some, but not all of the BV dataset. That these stand out is odd, since inspection of the count table shows that there are many other functions that are missing from the H dataset, and have higher and more uniform counts in the BV dataset. Thus, the importance

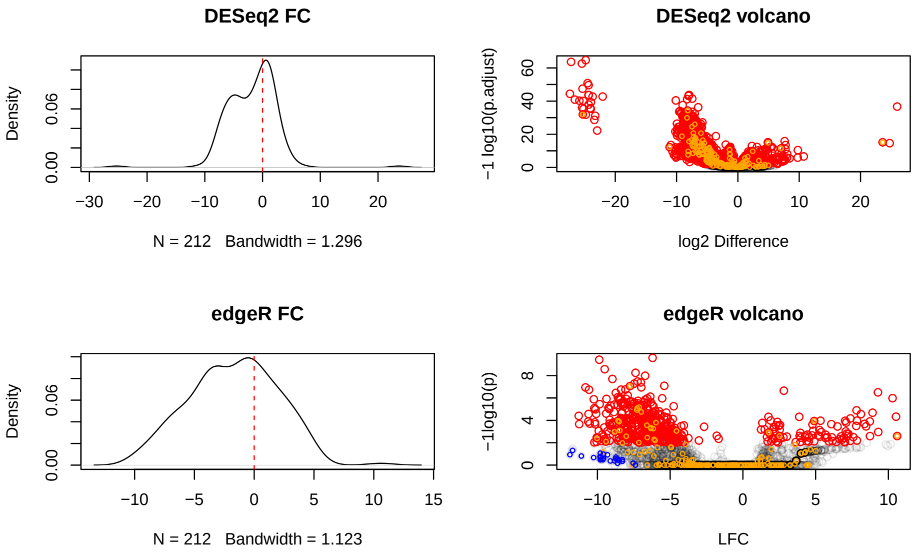


Figure 3: **Shown here are the mean log2 fold change as a density plot, and a Volcano plot showing the location and adjusted p-value for each feature in the metatranscriptomic dataset**
The DESeq2 approach does a good job of centring this data, while edgeR is less suitable. The volcano plots show dramatically different outcomes. The DESeq2 algorithm assigns very large fold changes to features that have only moderate change, and further identifies a very large proportion of features as significantly different. In contrast, edgeR exhibits a much smaller number of differentially abundant features. In both volcano plots, the housekeeping genes in the main Figure 3 are shown in orange. We can see that these are asymmetrically distributed in both plots. Additionally the location of the features taht DESeq2 identified as having a very large difference are shown in the edgeR volcano plot as blue circles.

of the outlier functions seen in the DESeq2 volcano plot should be viewed with suspicion and may be an artefact of the normalization used. When overplotted on the edgeR volcano plot (blue), it is clear that these functions have non-signficant p-values.

Gloor, Gregory B. 2023. “amIcompositional: Simple Tests for Compositional Behaviour of High Throughput Data with Common Transformations.” *Austrian Journal of Statistics* 52 (4): 180–97.

Gloor, Gregory B, Jean M Macklaim, Michael Vu, and Andrew D Fernandes. 2016. “Compositional Uncertainty Should Not Be Ignored in High-Throughput Sequencing Data Analysis.” *Austrian Journal of Statistics* 45: 73–87. https://doi.org/doi:10.17713/ajs.v45i4.122.

McMurrough, Thomas A, Russell J Dickson, Stephanie M F Thibert, Gregory B Gloor, and David R Edgell. 2014. “Control of Catalytic Efficiency by a Coevolving Network of Catalytic and Noncatalytic Residues.” *Proc Natl Acad Sci U S A* 111 (23): E2376–83. https://doi.org/10.1073/pnas.1322352111.