

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}

¹Dep'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 is related to Information and Shannon's entropy in HTS datasets	2
2	Issues with DESeq2 and edgeR	5
3	Checking the scale assumptions of the RLE and TMM normalizations.	7
	References	7

1 GM is related to Information and Shannon's entropy in HTS datasets

Information content is a fundamental property of all systems. We can think about scale from an information theoretic point of view as a measure of how much information is encoded in a particular sample. Intuitively, systems with different scales will contain different amounts of information and so we would expect $W_n^\perp \sim H_n$. Below, we show that the scale of a system as defined by Nixon et al. (2023) is inversely related to the information content and entropy of the data.

Recall the underlying system is described by a $D \times N$ matrix of counts \mathbf{W} decomposed into the proportions for the n^{th} sample \mathbf{W}_n^\parallel (or the equivalent probability distribution $\mathbf{p}(w_n)$), and its scale \mathbf{W}_n^\perp , such that $\mathbf{W} = \mathbf{W}_n^\parallel \mathbf{W}_n^\perp$. Sequencing returns counts which are related to the underlying proportion; i.e., $\mathbf{Y}_n^\parallel \sim \mathbf{W}_n^\parallel$

For notational simplicity assume we will use a single discrete random variable to represent the probability distribution; i.e. $X = \mathbf{p}(w_n)$ over $1 \dots d$ elements. The geometric mean can be calculated as:

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

;

here every element in $\log_2 G(X)$ is the logarithm of the elemental probability $\log_2(p_i)$.

In information theory, the elemental amount of information or uncertainty for p_i is the inverse $I(X)_i = -\log_2(p_i)$ (Reza 1994).

The total entropy of the system $H(x)$ is the weighted sum of the elemental information

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

In this formulation, $\log_2(G(X))$ and $H(X)$ can be understood to be an unweighted and a weighted measure of 'surprisal' in the dataset. The total and the mean surprisal are related by the number of terms in $p(x)$ and by the weighting factor for each term. $H(X)$ corresponds to the amount of information needed in order to reproduce X . Conversely, G is an unweighted mean of the same distribution and G is the denominator to calculate the centred log-ratio normalization (CLR). Note that when calculating the CLR we are taking a ratio, and similarly for calculating entropy, but that the scale formula takes the product. Thus, the scale of a system is inversely related to the other two measures.

Empirically, we can see that $G(\mathbf{Y}_n^\parallel)$ is strongly correlated with Shannon's Entropy $H(\mathbf{Y}_n^\parallel)$ (Supplemental Figure 1 and Table 1). These two measures are expected to have different behaviours with different distributions of p_i . In the case of a uniform distribution both $H(\mathbf{Y}_n^\parallel)$ and $G(\mathbf{Y}_n^\parallel)$ 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 $H(\mathbf{Y}_n^\parallel)$ could ascribe

some uncertainty to the large number of low probability events, while $G(\mathbf{Y}_n^{\parallel})$ 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.

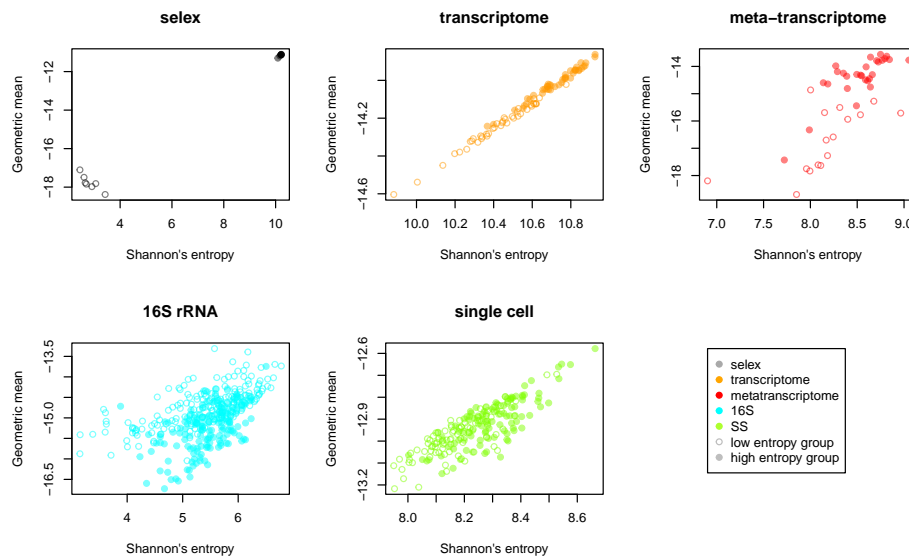


Figure 1: Plot of Shannon's entropy (H) vs geometric mean (G) for each sample in different datasets
The groups that each sample belong to are highlighted as filled or open circles. Each group in each dataset has different entropy with the groups in the selex and metatranscriptome datasets being highly 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	\bar{G}	\bar{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 \bar{G} in each group is relatively small. Most significantly, the selex dataset has a very large difference of about 100-fold, and both the 16S and the metatranscriptome dataset have about a 1.5 fold difference. These three datasets are candidates for a full scale model correction.

Nixon et al. (2023) showed that many operations on HTS datasets relied on both the proportion and scale components of the data. Moreover, all normalizations impose a scale model on the data, but the appropriateness of these models has never been explicitly acknowledged or

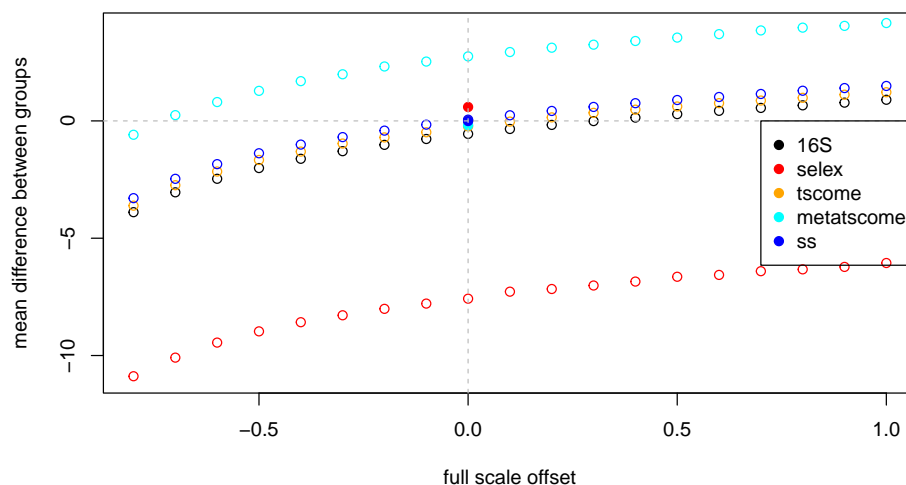


Figure 2: Plot of the offset of the mean difference between groups as a function of scale ratio
 For this, the default scale of 1:1 was altered in increments of 0.1 keeping the gamma parameter (dispersion) at 0.5. The filled circle shows the outcome when the calculation is done using the geometric mean and the same gamma parameter.
 (#fig:full.scale)

tested. Thus, the full scale model option allows the investigator to set both the offset between the geometric means of the features and their dispersion and observe how this affects the analysis outcome.

In the offset plot above we can see the cause and the effect of the full model with a fixed gamma of 0.5 and a base scale of 1 in each group. We see that the mean location of well centred datasets (yeast transcriptome, single-cell transcriptome) are close to 0, but could be centred better with small changes in scale ranging from 0 (single cell) to 1:1.1 for the yeast transcriptome dataset. In contrast, centring the 16S dataset requires about a 1:1.3 fold change. The metatranscriptome dataset would require about a 0.5:1 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 be used to illustrate the power of the full scale model and the relationship between G_n and scale. 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 so the non-selected group (NS) appears to have a similar location as the S group. The difference in entropy between the two groups and the differences in geometric mean are very large, with the difference in $\log_2 \bar{G}(S)$ and $\log_2 \bar{G}(NS)$ being about $2^{6.6}$. Setting the scale of both the S and NS groups to 1 we find that the difference in location is approximately $2^{7.7}$ in close agreement with the difference the geometric means. For this dataset to be centered we need to have a scale ratio $\approx 1:50$ or more. Note that the scale ratio is inverse to the ratio of geometric means as described above. In fact, in this dataset the relative abundances of the majority of features are nearly invariant, but this is masked by the large absolute changes in a small number of features (McMurrough et al. 2014), thus changing the scale of the data. Neither DESeq nor edgeR are able to provide a reasonable analysis of this dataset because the normalizations used assume equivalent scales (G. Gloor et al. 2016).

Figure 3 shows an effect plot of various scale models with this dataset. The full scale model, where the strong assumption that the mean G is assumed to be 1:1 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 the NS:S groups to be $\approx 1:50$ different moves the centre of the large number of relatively invariant features to the centreline of no difference, and recapitulates the default result obtained using G_n as the scale estimate where the ratio is $\sim 100 : 1$. Note that we get exactly the same answer (within random sampling error) with a scale of .02 for group NS and a scale of 1 for group S, or using a scale of 1 for group NS and a scale of 50 for group S. This shows that it is the relative difference between scales that is important in this dataset, 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 100-fold difference in \bar{G} between groups; the discrepancy being explained because the default scale model is applied uniformly to all samples, whereas the different values of the within-group geometric means ranging over a $\{ > 2.5 \}$ fold range. Thus, an advantage of a full scale model is that we have gained both information and understanding about the drivers of asymmetry underlying system.

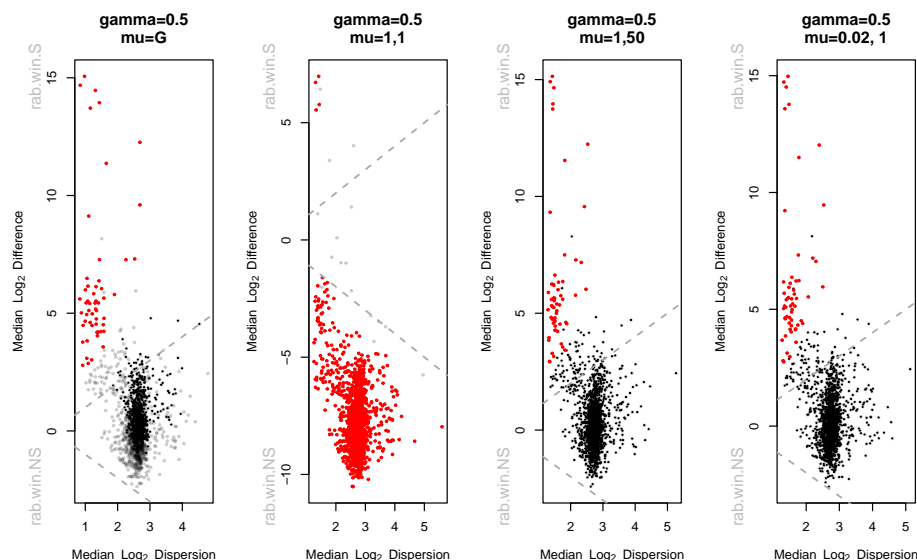


Figure 3: Effect plots of the selex dataset with various gamma and scale parameters
All scales are calculated with a logNormal distribution to ensure symmetry for the user.

2 Issues with DESeq2 and edgeR

DESeq2 and edgeR are two of the most commonly used tools for differential abundance analysis of bulk RNA sequencing datasets. They both operate by finding a scaling factor that makes all the samples commensurate. 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 commensurate. 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 (G. B. Gloor 2023)

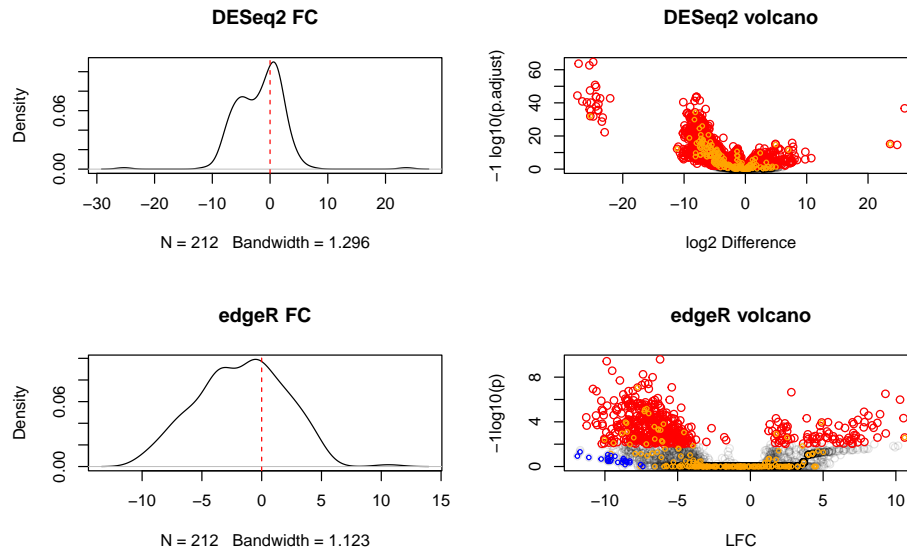


Figure 4: Shown here are the mean log2 fold change as a density plot, and a Volcano plot showing the location and adjusted p-value for each feature in the metatranscriptomic dataset

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 that DESeq2 identified as having a very large difference are shown in the edgeR volcano plot as blue circles.

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, as shown in Figure 4 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 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-significant p-values.

3 Checking the scale assumptions of the RLE and TMM normalizations

We can show that the normalizations built into the edgeR Bioconductor package (RLE, TMM, TMMwsp, upperquartile) are scale assumptions by using the normalization factor as an input to `aldex.makeScaleMatrix()` and then measure the mean location of the data as above. The ideal behavior of a normalization is that the mean location of the data should be close to 0 or unchanged. This behavior will ensure that Type 1 and Type 2 errors due to scale assumptions are minimized.

The results, shown in the tables below compare these normalizations to the no scale assumption (iso) and the geometric mean normalization (GM) assuming that the normalizations are either on a log scale or a linear scale. That these affect scale can be observed by their effect on the mean location of the data. For the most part assuming no scale differences between groups centres the data less well than does the geometric mean. In most cases the other normalizations perform poorer than assuming no scale at all for the rRNA dataset and about as well as the no scale assumption in the metatranscriptome datasets and the single-cell transcriptome dataset. All normalizations perform poorly in the selex dataset. Overall, these results may not be surprising because the TMM and RLE (and related normalizations) were developed specifically to scale and normalize transcriptome datasets (Robinson and Oshlack 2010; Anders and Huber 2010), but have become widely used in the analysis of other data modalities; e.g. (McMurdie and Holmes 2014).

Table 2: Log scale models

	RLE	TMM	TMMwsp	upperquartile	iso	GM
rRNA	-2.1529252	-0.1763102	-0.3699510	0.0865392	-0.5599236	0.0093805
selex	-1.9343851	2.1047060	1.9617191	NaN	-7.5096353	-0.0062949
yst	0.0746243	0.0297339	0.0639645	-0.0694199	-0.2160747	-0.0129932
meta	2.7626908	2.5125909	2.8658800	2.5469506	2.7145893	-0.0381394
ss	0.1384218	0.0843315	0.0010290	0.1008664	0.0578294	-0.0276270

Table 3: Linear scale models

	RLE	TMM	TMMwsp	upperquartile	iso	GM
rRNA	-2.2112427	0.0072079	-0.1431558	0.2898137	-0.5577481	0.0093805
selex	-1.6912748	0.5256675	0.5113871	NaN	-7.5296170	-0.0062949
yst	0.1448714	0.1524261	0.1634843	0.0022118	-0.2288918	-0.0129932
meta	3.2141857	2.6937636	3.1516101	2.5240846	2.7581217	-0.0381394
ss	0.1735196	0.0961419	-0.0140678	0.1074003	0.0524419	-0.0276270

References

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Genome Biol* 11 (10): R106. <https://doi.org/10.1186/gb-2010-11-10-r106>.

- Gloor, GB, JM Macklaim, M Vu, and AD 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>.
- Gloor, Gregory B. 2023. "amlcompositional: Simple Tests for Compositional Behaviour of High Throughput Data with Common Transformations." *Austrian Journal of Statistics* 52 (4): 180–97.
- McMurdie, Paul J, and Susan Holmes. 2014. "Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible." *PLoS Comput Biol* 10 (4): e1003531. <https://doi.org/10.1371/journal.pcbi.1003531>.
- 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>.
- Nixon, Michelle Pistner, Jeffrey Letourneau, Lawrence A. David, Nicole A. Lazar, Sayan Mukherjee, and Justin D. Silverman. 2023. "Scale Reliant Inference." <https://arxiv.org/abs/2201.03616>.
- Reza, Fazlollah M. 1994. *An Introduction to Information Theory*. Courier Corporation.
- Robinson, Mark D, and Alicia Oshlack. 2010. "A Scaling Normalization Method for Differential Expression Analysis of RNA-seq Data." *Genome Biol* 11 (3): R25.1–9. <https://doi.org/10.1186/gb-2010-11-3-r25>.