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

1.1 Shannon's entropy has a volume or size

Information is a fundamental property of all measured systems. For discrete probability vectors, Shannon defined their information properties and launched the field of information theory for communications. Famously, for any probability vector, Shannon set the total entropy as the inverse of the sum of the probability weighted logarithm of the probabilities. Less well known is that an unweighted average entropy was also defined, and that this turns out to be the inverse of the logarithm of the geometric mean of the probability vector. Thus, information theory and compositional data analysis intersect through the geometric mean of a probability vector. Here I show that the information theoretic interpretation can help us understand the scale of a system as defined by Nixon and Silverman, and that this interpretation can help in the interpretation of highly asymmetric datasets. I will show the Asymptotic Equipartition Property of information can be defined as a volume for discrete distributions, and how that volume relates to the scale of a system. Finally, I will show how Shannon's entropy can substitute for the geometric mean in common CoDa operations and how that alters the interpretation of the results.

We can think about scale from an information theoretic point of view as a measure of how much information, or total uncertainty, is encoded in a particular sample (Shannon 1948; Jaynes and Bretthorst 2003). In the geometric interpretation of information theory used in quantum information theory, formally described as the Asymptotic Equipartition Property of information (Cover and Thomas 1991; Wilde 2017), entropy can be interpreted as the volume occupied by a probability distribution. See chapters 4 of the PhD thesis of Lecamwasam (2021) for a nice explanation of this.

For notational simplicity assume we have a single discrete random variable to represent a probability distribution with d elements; i.e. $X = \mathbf{p}_{i=(1\ldots d)}$. In information theory, the elemental amount of information or surprisal for p_i is the inverse of the logarithm of the elemental probability, $-log_2(p_i)$(Reza 1994). This measure is often called self-information.

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

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

$H(X)$ corresponds to the amount of information needed in order to reproduce X. We can also calculate the expected amount of information for each observation in the random variable, and this is the mean of the elemental probability. This measure is also called the sample average of the information (Wilde 2017);

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

The linkage between compositional analysis, scale inference and information theory comes when we realize that the logarithm of the geometric mean calculated in base(2) is:

$$l2G(X) = \log_2 G(X) = \frac{1}{d}\sum_{i=1}^{d} \log_2 p_i$$

;

We see that $h(X) = -l2G(x)$. Furthermore, $l2G$ is used as the basis for the centred log-ratio transform and is the starting point for scale-based inference:

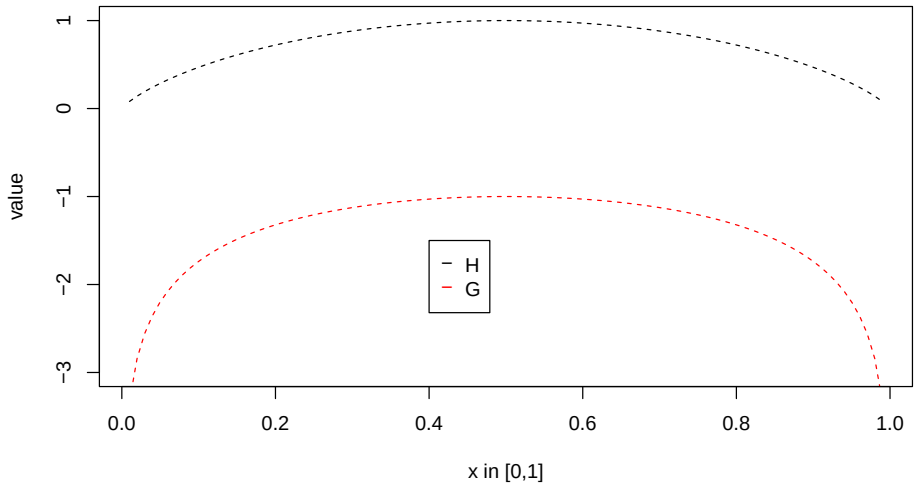
$$CLR(X) = log2(p_i) - l2G(X) = log2(p_i) + h(X)$$

Thus we see that the geometric mean used in the centred log ratio (CLR), often used for Compositional Data Analysis (CoDa) (Aitchison 1982) is directly related to entropy or H. Indeed, we can rearrange the CLR formula to show that it can be interpreted as computing the difference between the elemental information and the mean information content:

$$CLR(X) = -log2(p_i) - (-l2G(X)) = -(-log2(p_i) - h(X))$$

As defined in (Nixon et al. 2023), the scale is the inverse of $l2G$, which is $h(X)$. Thus, one way we can understand scale is that it is measuring the total complexity of the system, and this is expected to increase with absolute size in most cases. Moreover, $H(X)$ and $G(X)$ share similar shapes in the continuum between 0-1 for a bivariate distribution as shown below:

```
par(mfrow=c(1,1))
curve(mf.G, from=1e-2, to =.99,  col='red', lty=2, ylim=c(-3,1),
   xlab="x in [0,1]", ylab="value")
curve(mf.H, from=1e-2, to =.99,  col='black', lty=2, add=T )
legend(.4,-1.5, legend=c('H','G'), col=c('black','red'), pch="-")
```



The difference being that entropy is constructed to have a value of 0 at the margins because Shannon defined $0log(0) = 0$ while the geometric mean approaches negative infinity.

Now let's think about the idea of entropy as a volume which allows us to identify what amount of the available entropy space a given observation fills. The following is taken and modified from (Lecamwasam 2021), to which you should refer for a more fulsome discussion, and it is covered in Chapter 2 of Wilde(2017) and Chapter 3 of Cover and Thomas (1991). If we start with a four part system $X1 = [A, C, G, T]$ where the frequencies are equally and identically

distributed, then $p_A = p_C = p_G = p_T = \frac{1}{4}$. $H(X1) = -1 * 4 * (\frac{1}{4} * log2(\frac{1}{4})) = 2$. This is the maximum entropy possible. We can obtain the "volume" of $X1$ by exponentiating $H1$ using the same base as was used to calculate the entropy; $V1 = 2^{H1} = 4$. This is the same as the number of letters in the system; so the volume needed to explain the system is 4 units (in this case bits). But what happens in another system, $X2$ where A occurs with a much higher probability, say 0.7, and the other three are distributed equiprobably amongst the remainder with a probability of 0.1; i.e. $p_C = p_G = p_T = 0.1$. In this case $H2 = -1 * ((\frac{7}{10} * log2(\frac{7}{10})) + (3 * (\frac{1}{10} * log2(\frac{1}{10})))) = 1.358$. Here the volume of $X2 = 2^{H2} = 2.56$; meaning that less than the maximum volume is taken up by the information. Here $X2$ consumes about 64% of the volume of system $X1$. Thus, the volume is a measure of the total complexity or the scales of the two systems. In this way we can understand that scale is related to the information volume of a system.

Empirically, we can see that $G(\mathbf{Y}_n^{\|})$ is strongly correlated with Shannon's Entropy $H(\mathbf{Y}_n^{\|})$ as expected from the discussion above, and that this difference converges to a constant as the number of entries in the probability vector increases regardless of the distribution, although different distributions converge at different rates. For example, if we plot the relationship between H and G as a function of the length of the probability vector we can see a direct inverse relationship.

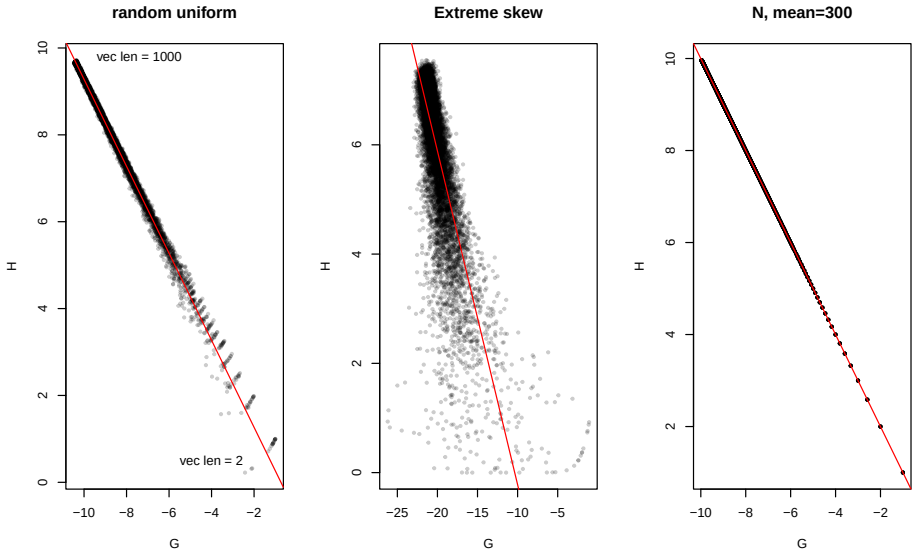


Figure 1: Association between entropy (H) and geometric mean (G) as a function of vector length
Twenty random vectors were constructed for each length between 2 and 1000 in increments of 2 for each of the random distributions in the legend; N = Normal, U = Uniform, B = Beta. The bottom right of each plot represents vectors of length 2, and the top left represents the vector of lenght 500. The maximum value of H increases as the vector length increases, and the maximum value of G decreases in lock-step. Each random distribution has an obviously distinct relationship between the two measures. For the purposes of high throughput sequencing the Beta distribution is most similar to that seen in the majority of instances.

```
## (Intercept)
##    -0.73613
## (Intercept)
##     -6.3303
##  (Intercept)
## -0.001424674
```

When we plot the relationship for any individual probability vector, we see that there is an direct relationship between the entropy and the log of the geometric mean, but that this relationship strongly depends on the underlying distribution of the probability distribution X.

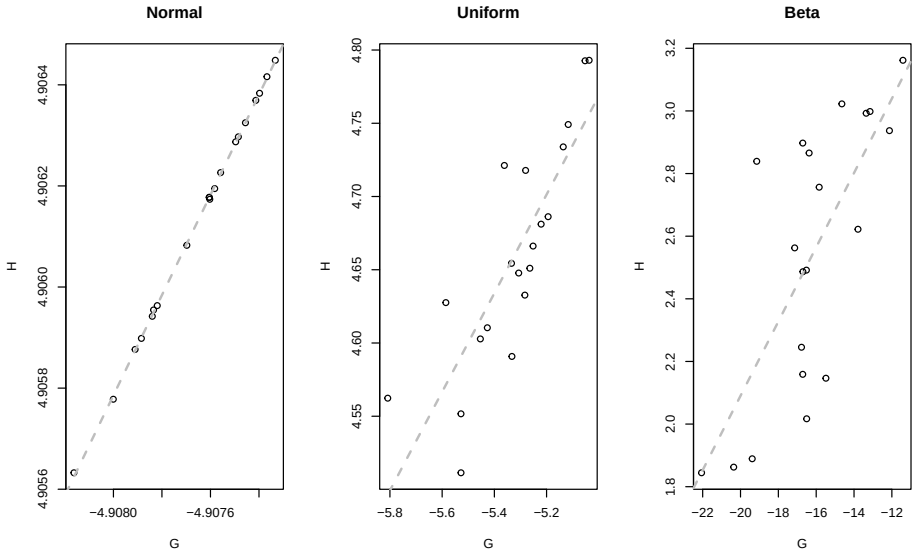


Figure 2: Plot of the association between H and G at a vector length of 30
The relationship between H and G is inverse, and the strength of that association depends on the distribution. The N distribution shows a very strong assocation, while the Beta distribution is less well defined. Associations are shown for a vector length of 30.

In real data, shown in Supplemental Figure 3 and Table 1 the correspondence is not as predictable, likely because the real data is a more complex distribution than any of the idealized distributions. Thus, these two measures have different behaviours with different distributions of p_i. In the case of a uniform distribution both $H(\mathbf{Y}_n^{\|})$ and $G(\mathbf{Y}_n^{\|})$ are maximal when $p(x)$ is equally and identically distributed. Thus, we expect that they are positively correlated here. In a Normal or a skewed distribution, we also observe 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 $H(\mathbf{Y}_n^{\|})$ could ascribe some uncertainty to the large number of low probability events, while $G(\mathbf{Y}_n^{\|})$ 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.

Intuitively, systems with different scales will contain different amounts of information and so we would expect $W_n^{\perp} \sim H_n$. As the scale of a system as defined by Nixon et al. (2023) is inversely related to G, this means that scale is directly proportional to the information content and entropy of the data.

Below I show that we can replace G with H in the calculations performed by ALDEx2 without loss of utility.

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 $\mathbf{W}_n^{\|}$ (or the equivalent probability distribution $\mathbf{p}(w_n)$), and its scale $\mathbf{W}_n^{\perp}$, such that $\mathbf{W} = \mathbf{W}^{\|}\mathbf{W}^{\perp}$. Sequencing returns counts which are related to the underlying proportion; i.e., $\mathbf{Y}_n^{\|} \sim \mathbf{W}_n^{\|}$

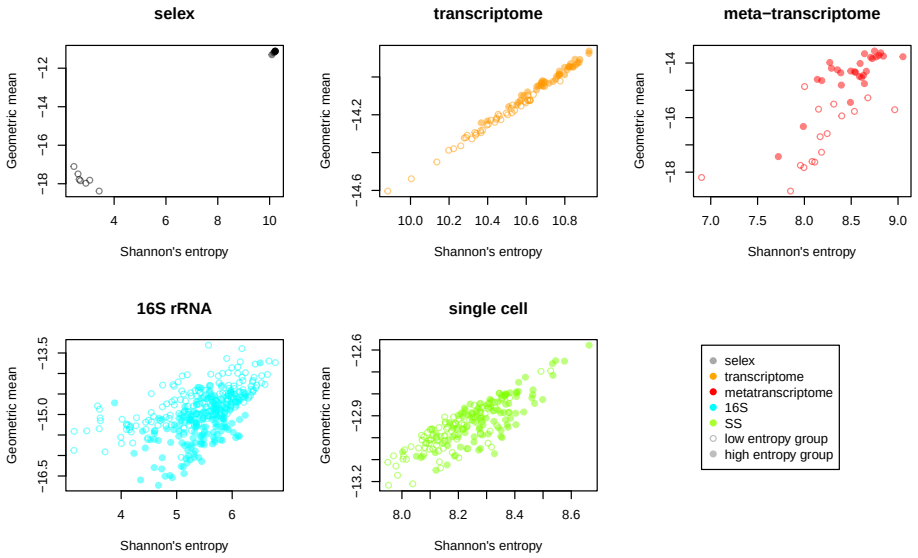
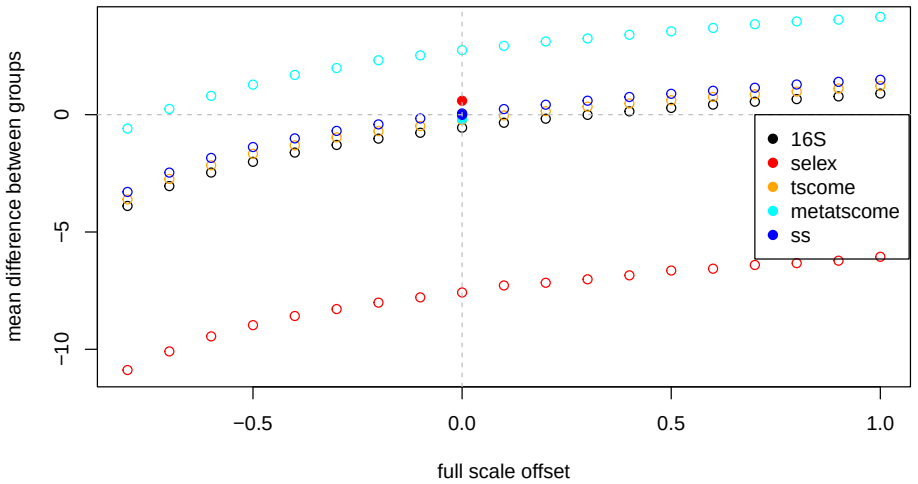


Figure 3: Plot of Shannon's entropy (H) vs geometric mean (G) for each sample in different datasets The groups that each sample belong to are highlighted as filled or open circles. Each group in each dataset has different entropy with the groups in the selex and metatranscriptome datasets being highly 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 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.



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 $\log_2 \overline{G}(S)$ and $log_2\overline{G}(NS)$ 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 $\overline{G}$ 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.

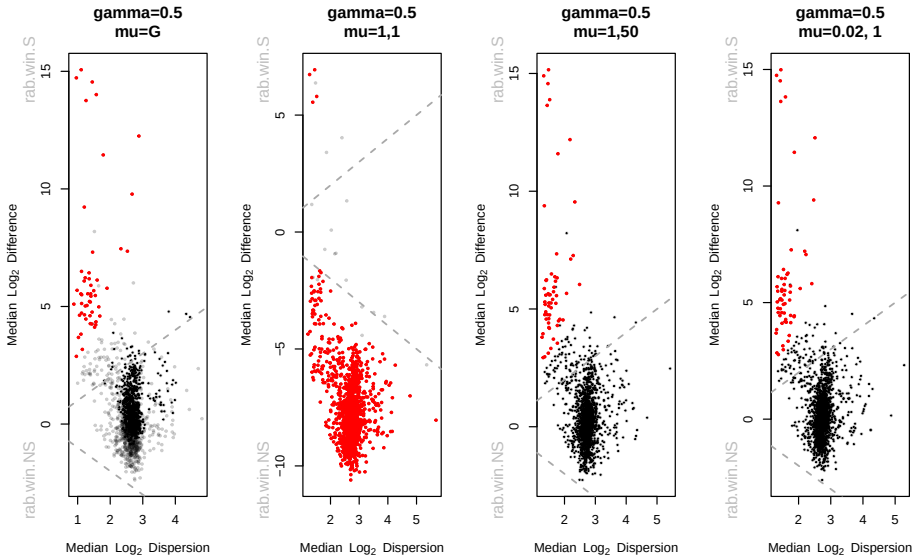


Figure 5: **Effect plots of the selex dataset with various gamma and scale parameters**
All scales are calculated with a logNormal distribution to ensure symmetry for the user.

2 How scaling affects dispersion

Thus far we have observed that adding scale uncertainty increases the minimum dispersion value with little effect on the difference between values as in panel A. Panel B shows the change in dispersion relative to the rAbundance of the features. Here we can see that the minimum dispersion is increased as gamma increases. The horizontal lines show the median value for the features with a rAbundance between -0.5 and 0.5. Panel C shows that this increase is non-linear, being more pronounced among those features that had minimal dispersion when gamma=0. Panel D shows that increasing gamma actually shrinks the dispersion estimates towards the mean dispersion, while increasing the total dispersion as shown in panel C.

Thus, adding scale uncertainty has two effects both on dispersion. The first is to increase the dispersion estimate for each part and this has its primary effect to reduce the p-values and standardized effect sizes returned by ALDEx2. The second is to reduce the range of dispersions such that the parts with the lowest dispersions have the greatest increase. This increase in dispersion comes about because the underlying distributions are widened because of the additional uncertainty added by the inclusion of scale uncertainty.

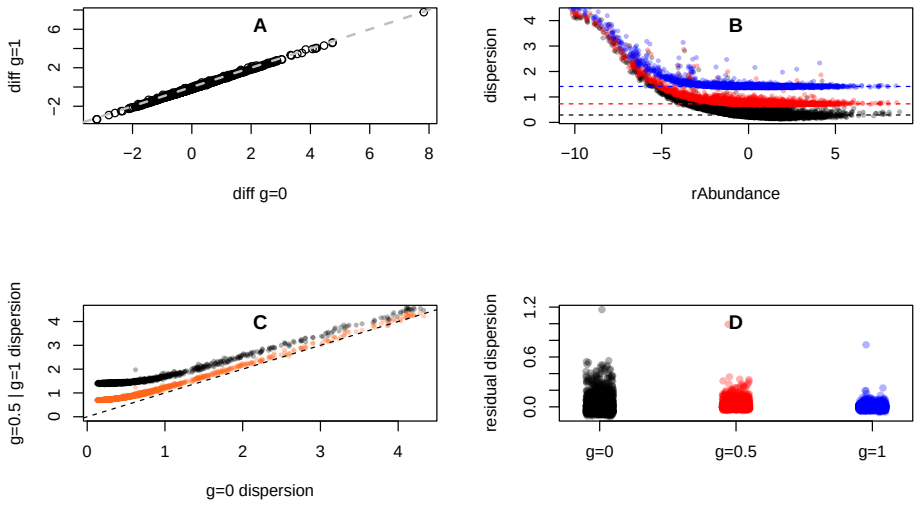


Figure 6: jnk

3 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 (G. B. Gloor 2023)

Examining the plots, edge R clearly not centred with the median housekeeping functions offset by -0.2888154 and minimum FDR value is 0.0104767. Additionally, as shown in Figure 4there 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 -0.6535995 but the minimum FDR value is $6.9497481 \times 10^{-15}$. 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-signficant p-values.

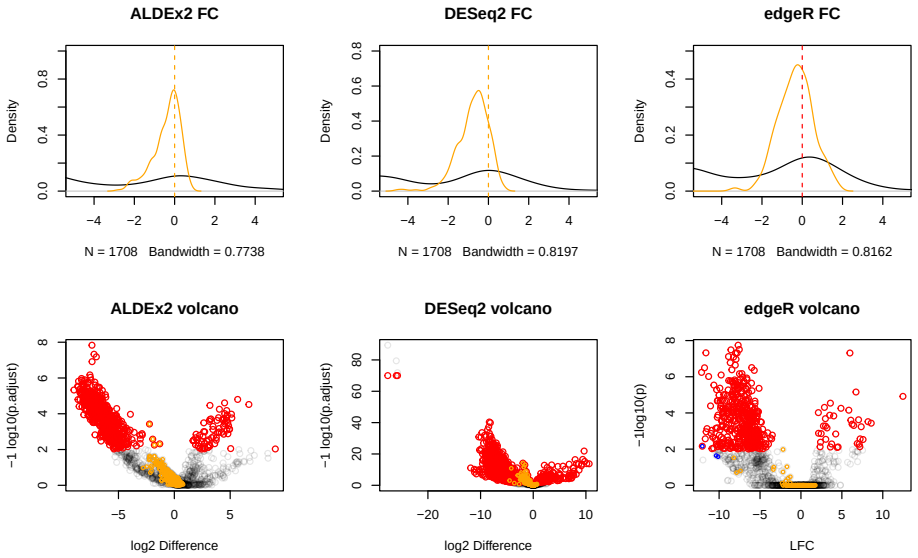


Figure 7: Shown here are the mean log2 fold change as a density plot, and a Volcano plot showing the location and adjusted p-value for each feature in the metatranscriptomic dataset
The DESeq2 approach does a good job of centring this data, while edgeR is less suitable. The volcano plots show dramatically 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.

4 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, upperquantile) 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 date 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

Aitchison, John. 1982. "The Statistical Analysis of Compositional Data." *Journal of the Royal Statistical Society: Series B (Methodological)* 44 (2): 139–60.

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.

Cover, T. M., and Joy A Thomas. 1991. *Elements of Information Theory*. New York: Wiley. http://www.loc.gov/catdir/bios/wiley043/90045484.html.

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.

Jaynes, E. T, and G. Larry Bretthorst. 2003. *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press. http://www.loc.gov/catdir/samples/cam033/2002071486.html.

Lecamwasam, Ruvindha. 2021. "Investigations of Metrology in Optomechanics and Quantum Information Theory." PhD thesis, Research School of Physics, ANU College of Science, The Australian National University.

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.

Shannon, C. E. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27 (3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Wilde, Mark M. 2017. *Quantum Information Theory*. 2nd ed. Cambridge University Press. https://doi.org/10.1017/9781316809976.