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

true

One root cause of the large number of significant parts is the very low dispersion when the data are normalized and log-transformed. The raw data counts derived from sequencing are overdispersed with the mean value being less than the variance (1, 2) as seen in Panel A of Figure 1. However, the actual analysis of differential abundance is performed on the logarithm of the normalized counts (2–4) which has a very different abundance-dispersion relationship. Similar relationships are found in both DESeq2 (Panel B) and ALDEx2 (Panel C), and have been noted before (3, 5).

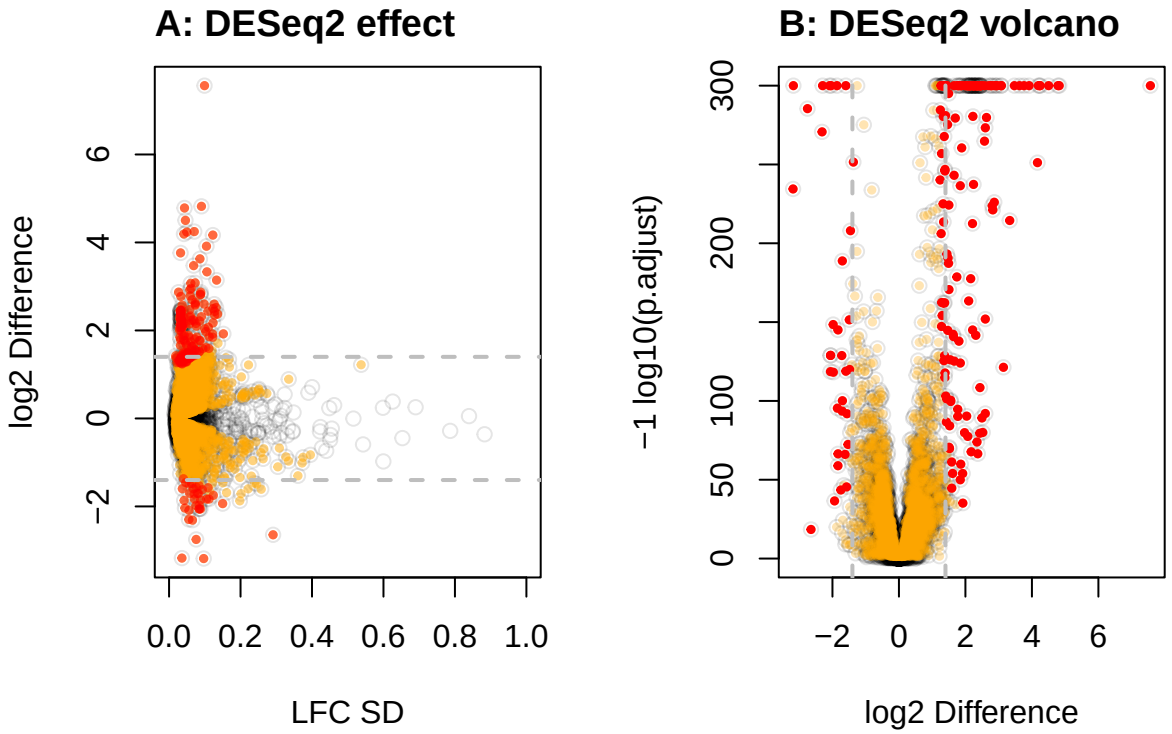

Figure 1: Effect and volcano plots for unscaled and scaled transcriptome analysis. DESeq2 was used to conduct a differential abundance (DA) analysis on the yeast transcriptome dataset. The results were plotted to show the relationship between difference and dispersion using effect plots or difference and the Benjamini-Hochberg corrected p-values (volcano plot). Each point represents the values for one transcript, with the color indicating if that transcript was significant in the ALDEx2 scaled analysis and the DESeq2 analysis (red) or in the DESeq2 unscaled analysis only (orange). Points in grey are not statistically signficantly different with any analysis. The horizontal dashed lines represent a log2(difference) of ±1.4

The variance or dispersion always increases with increasing raw (or normalized) read count but decreases when measured on the log-ratio transformed data (3, 5), reaching a minimum at some mid-point of the distribution. This makes the counter-intuitive suggestion that genes with moderate expression have more predictable expression than genes with very high expression such as housekeeping genes. This is at odds with

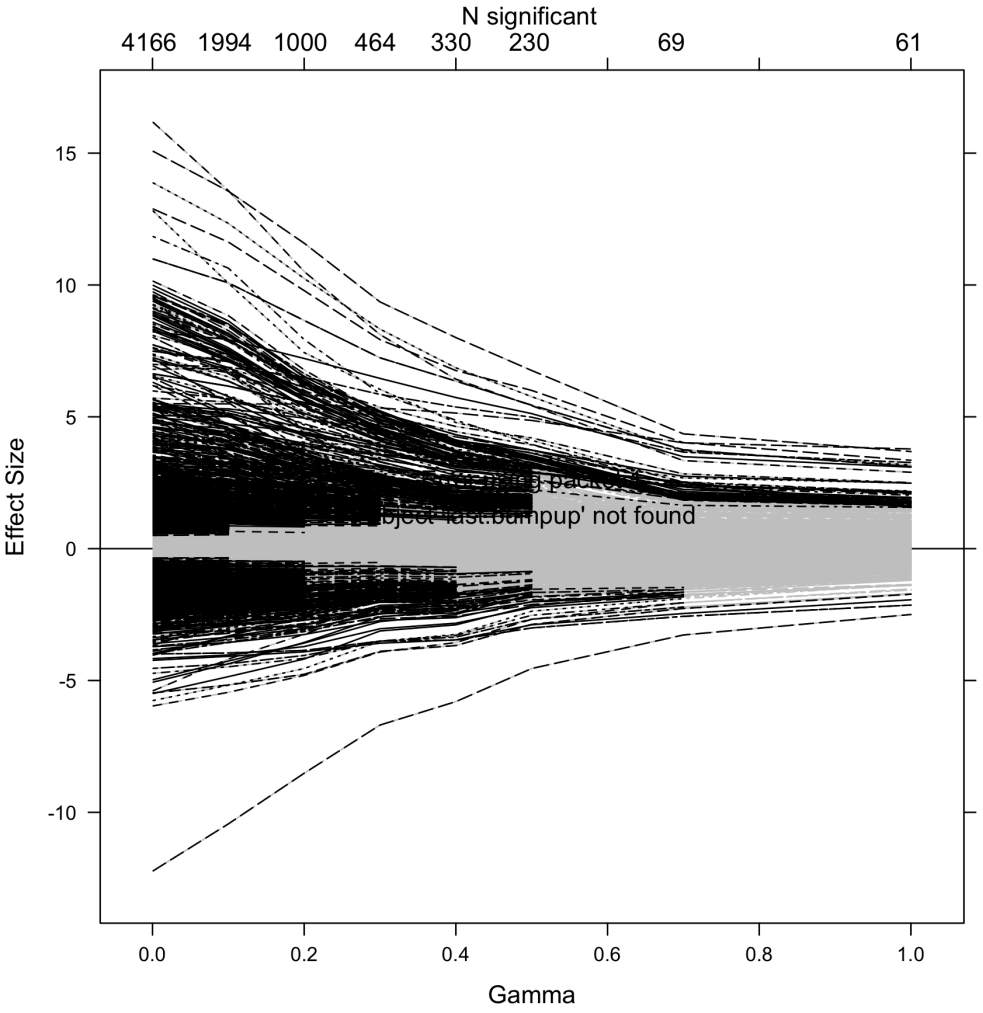

Figure 2: the aldex.senAnalysis() function was used to generate a dataset with γ values of 1e-3, 0.1, 0.2, 0.3, 0.4,0.5, 0.7 and 1. The plotGamma() function was used to plot the result. Transcripts that are statistically significant are shown in black, and if not significant are in grey. The γ values, standardized effect size and number of significant transcripts at each value are given on the axes.

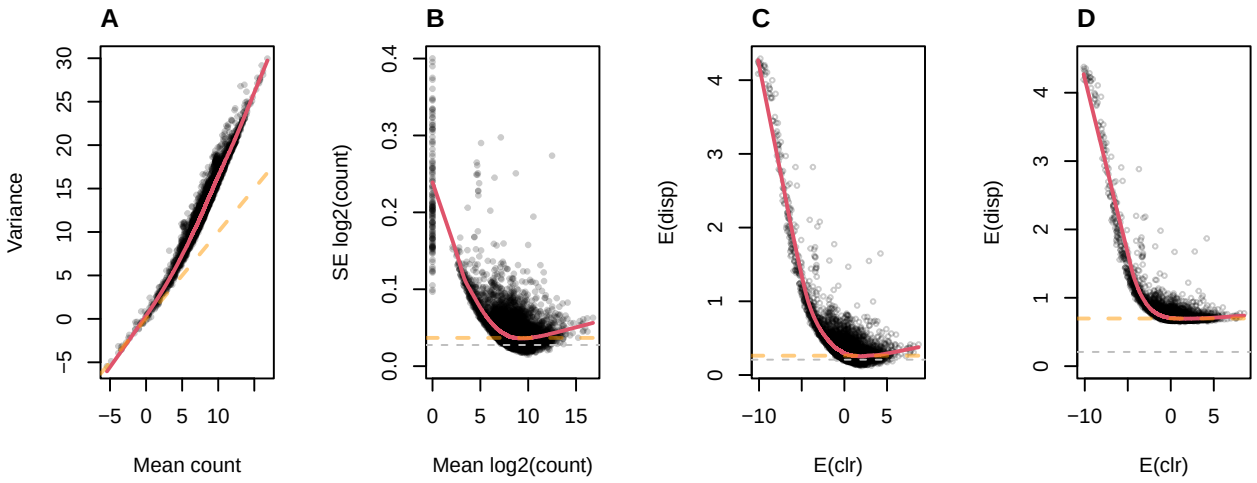

Figure 3: Plot of abundance v dispersion for a typical transcriptome dataset as counts, as logarithms of counts, and as CLR values. Panel A shows that the data are over-dispersed relative to a Poisson distribution which is represented by the dashed line when plotted on a log-log scale. Panel B shows that the relationship between the mean and the dispersion calculated in DESeq2, here the standard error (SE) of the mean, is very different when the data are log-transformed first. Panel C shows the equivalent values calculated by ALDEx2 in which the expected CLR value for each transcipt are plotted vs. the expected dispersion. Panel D shows the output for ALDEx2 with $\gamma = 0.5$. The red line in each panel shows the LOESS line of fit to the mid-point of the distributions. In panels B and C the amount of dispersion reaches a minimum at moderate values. The dashed orange line in panel A is the line of equivalence, and in panel B and C is the minimum y value. The values below the dashed grey line in panels B and C represent those below the first decile of dispersion.

the known biology of cells where single cell counting of housekeeping transcripts shows that they are both highly expressed and have little intrinsic variation (6). Furthermore, the dispersion is exceeding small being, for many transcripts, almost negligible. To show this point more clearly, the majority of the transcripts in the lowest decile of dispersion indicated below the dashed grey line are statistically significantly different (75% with DESeq2, 69% with ALDEx2), suggesting that low dispersion estimates lead to many false positives. Indeed, benchmarking and comparison studies repeatedly show that the choice of normalization plays a role in which results are returned as significant (7–9). From the perspective of Nixon et al. (10), it is reasonable to conclude that some of these results may be due to the choice of normalization.

While not necessary in all datasets (e.g., the transcriptome example discussed in the previous section and see the Supplementary material), it can greatly improve modeling in certain cases, especially when the scale is highly asymmetric between conditions. In order to specify a scale model from scratch, we need to revisit the concept that all normalizations in widespread use are actually ratios with the denominator implied by the normalization. Therefore, we can easily deviate from a certain normalization (e.g., the geometric mean assumption implied by the CLR) by specifying a total model based on the mean difference between conditions. While knowing the mean difference between conditions may seem cumbersome in practice, it is the *relationship* between the group scale values that is important, not their raw values (Nixon et al. 2023). This can be illustrated quite simply by starting with the mean ratio in G between groups for the yeast transcriptome dataset which are 6.05×10^{-5} for snf2 and 5.08×10^{-5} for WT; their ratio being 1.17, or a 0.17-fold difference. Using this information, we can recapitulate the differential abundance analysis in Figure 2B and 2C exactly by using setting the mean denominator of group 1 to 1, and group 2 to 1.17 with a gamma of 0.5 as shown in Supplemental Figure 2. This ratio can be adjusted to alter the mean assumption placed on the group scale values.

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.

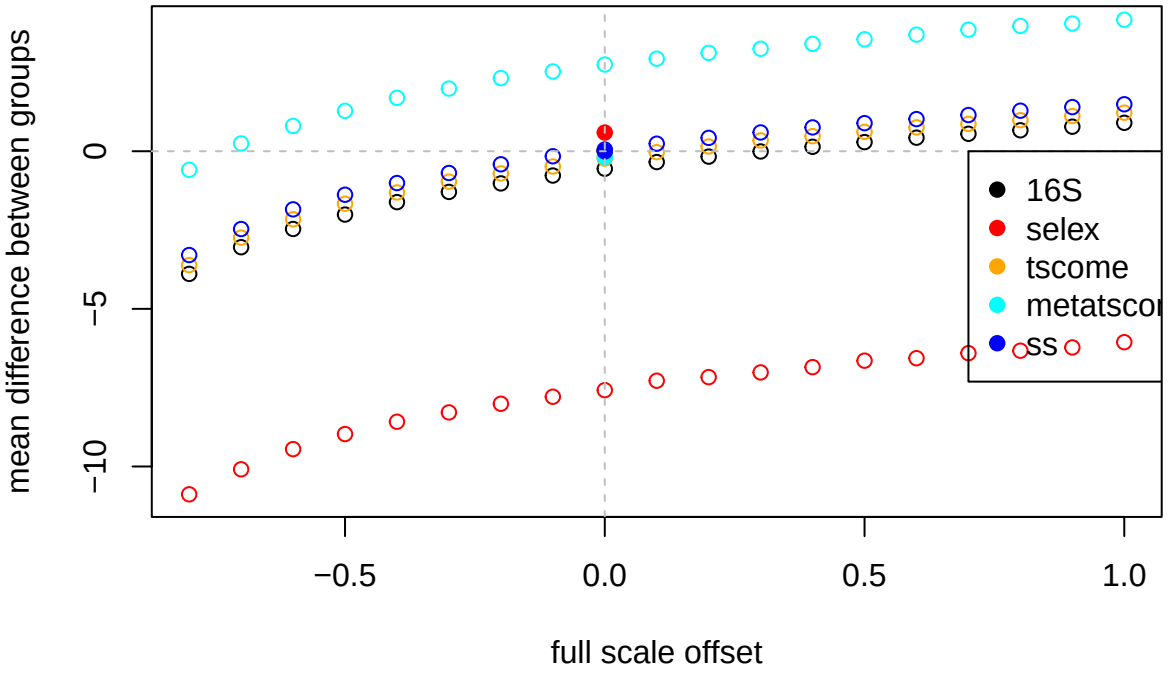

Figure 4: Plot of the offset of the mean differnece between groups as a function of scale ratio. For this, the default scale of 1:1 was altered in increments of 0.1 keeping the gamma parameter (dispersion) at 0.5. The filled circle shows the outcome when the calculation is done using the geometric mean and the same gamma parameter.

Nixon et al. (10) 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 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 $\log_2 \overline{G}(S)$ 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 (11), 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 (12).

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 $\approx$ 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.

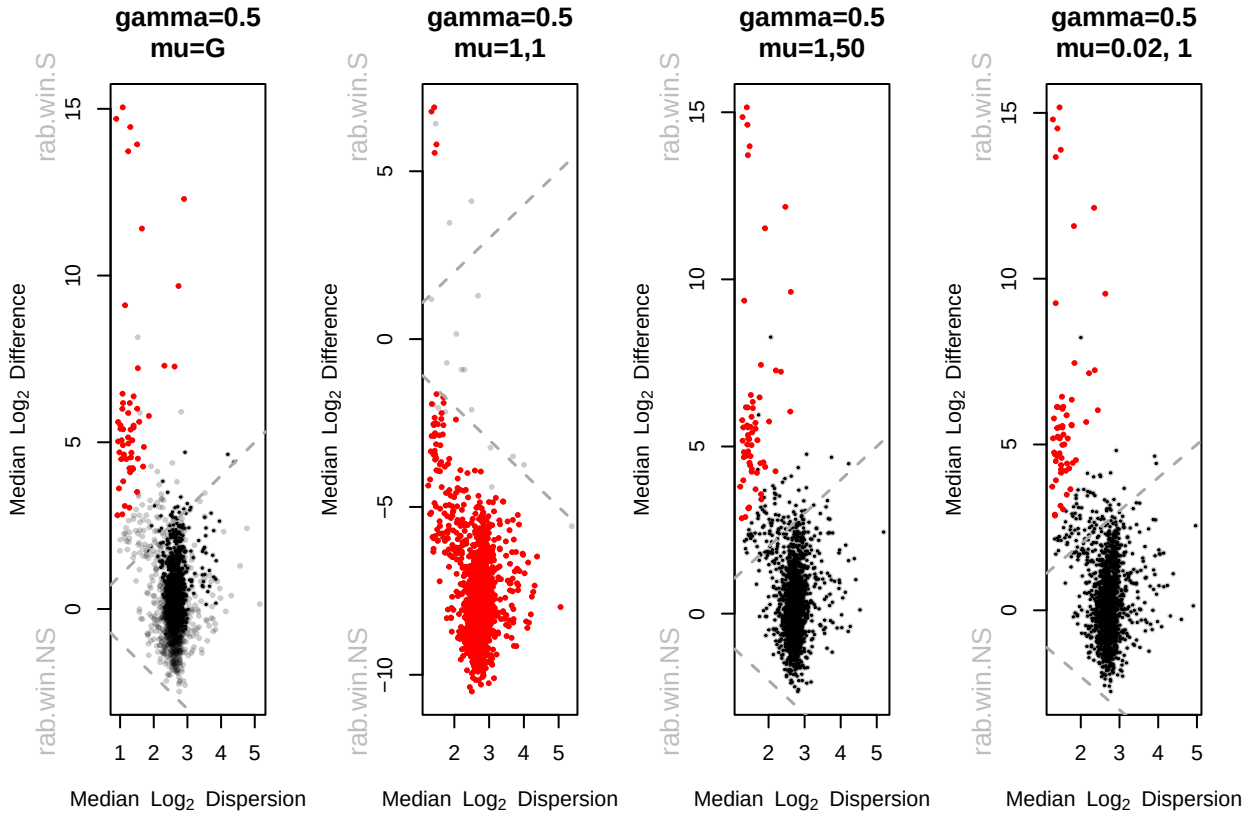

Figure 5: Effect plots of the selex dataset with various gamma and scale parameters. All scales are calculated with a logNormal distribution to ensure symmetry for the user.

How scaling affects dispersion

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 (13)

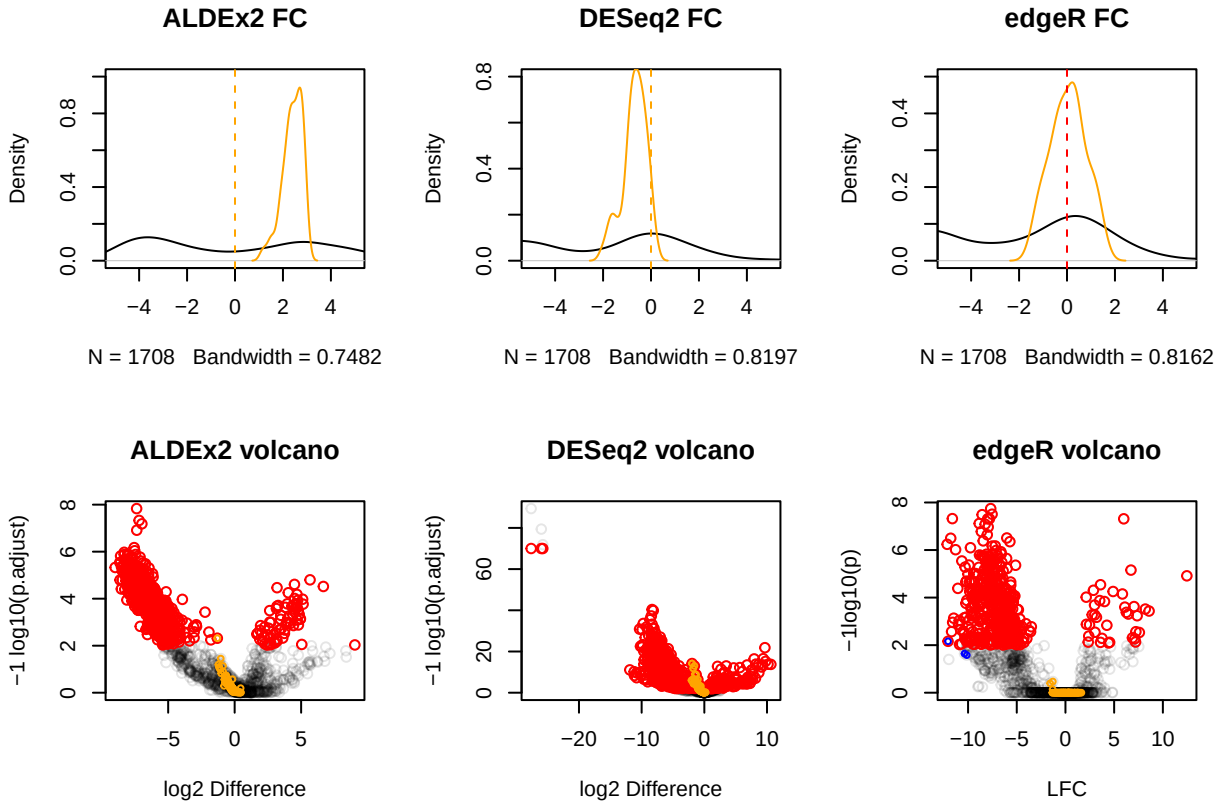

Figure 6: Shown here are the mean log2 fold change as a density plot, and a Volcano plot showing the location and adjusted p-value for each feature in the metatranscriptomic dataset. The DESeq2 approach does a good job of centring this data, while edgeR is less suitable. The volcano plots show dramatically 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 0.0515571 and minimum FDR value is 0.3382316. 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 -0.6359311 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.

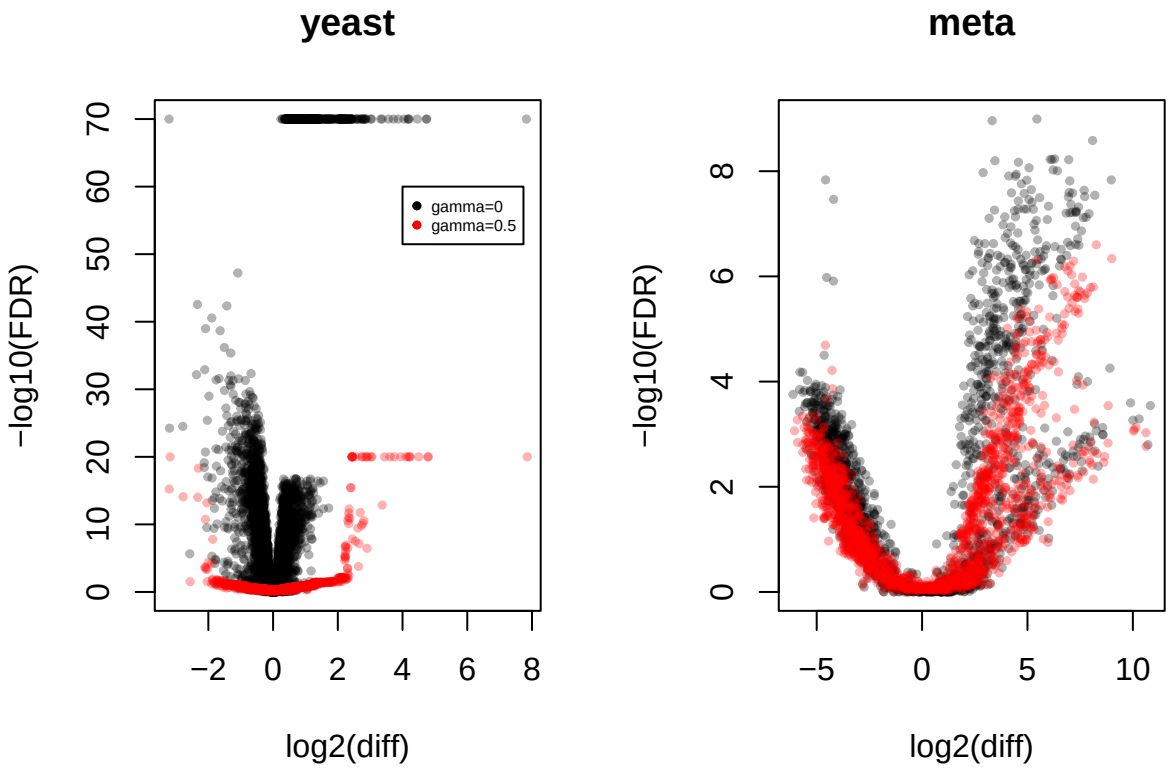

Figure 7: Shown here are two volcano plots that plot the mean log2 fold change plotted vs the log of the FDR for the yeast transcriptome dataset and for the metatranscriptome dataset. Data were generated in ALDEx2 with $\gamma = [0, 0.5]$ in black or in red.

Adding a small amount of scale, $\gamma = 0.5$, has a major effect on the volcano plot for the yeast transcriptome dataset, but mimimal effect on the vaginal metatranscriptome dataset. In both datasets the FDR values are raised, but no effect is observed on the difference between values. However, the concordance between the FDR values and the difference between become very tight for the yeast transcriptome dataset, but the concordance is not visibly unchanged for the other dataset. This is likely because both the difference between and the dispersion were both very small in the former and substantially larger in the latter.

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 (1, 14), but have become widely used in the analysis of other data modalities; e.g. (15).

Table 1: 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 2: 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

Thus far we have observed that adding scale uncertainty has a negligible effect on either the differences between groups or the relative abundance measure with a correlation of 0.999 and 0.998. The effect is entirely restricted to the dispersion estimats as shown below. Panel A 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 B shows that this increase is non-linear, being more pronounced among those features that had minimal dispersion when gamma=0. Panel C shows that increasing gamma rotates the dispersion estimates at high relative abundances, such that housekeeping genes no longer have larger mean dispersions (rAB > 5) than do regulated genes (rAB >0, < 5). Overplotted are the lowess lines of fit through the densities of transcripts that are judged as significantly different using either scaled or unscaled data, and with or without thresholding. Statistical significance was determined with a FDR < 0.05.

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, and rotate them such that the parts with the lowest dispersions have the greatest increase. This increase in dispersion comes about because the underlying distributions are widened with the inclusion of scale uncertainty.

References

1. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, **11**, R25.1–R25.9.

2. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–40.

3. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 550.1–550.21.

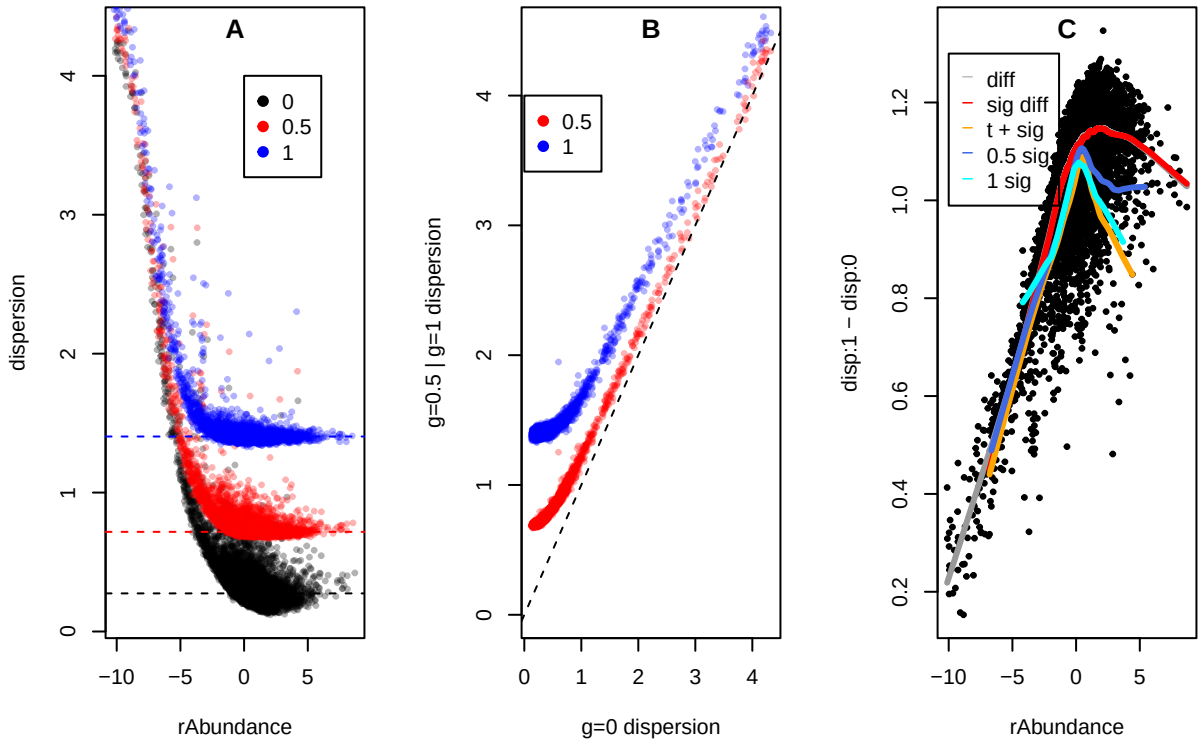

Figure 8: Adding scale uncertainty changes the dispersion distribution. Panel A shows a plot of the expected value for relative abundance vs the expected value for the pooled dispersion as output by aldex.effect. The dashed horizontal lines show the median value for the features with a rAbundance between -0.5 and 0.5. Panel B plots the unscaled vs scaled dispersion, note the non-linear relationship. Panel C plots relative abundance vs the difference between dispersion with gamma set to 0 and to 1 to highlight the rotation that is evident in Panel A. The colored lines indicate the lowess line of fit through the centre of mass of the plot. Line colors represent different populations of points. The grey line is the total population, the red line is the population of significant transcripts with no scale, the orange line is the population of significant transcripts with a difference threshold of about 2.5-fold change, the blue line is the population of significant transcripts with gamma =0.5, and the cyan line is the significant population with gamma = 1.

4. Andrews,T.S. and Hemberg,M. (2019) M3Drop: Dropout-based feature selection for scRNASeq. *Bioinformatics*, **35**, 2865–2867.

5. Fernandes,A.D., Macklaim,J.M., Linn,T.G., Reid,G. and Gloor,G.B. (2013) ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-seq. *PLoS One*, **8**, e67019.

6. Taniguchi,Y., Choi,P.J., Li,G.-W., Chen,H., Babu,M., Hearn,J., Emili,A. and Xie,X.S. (2010) Quantifying e. Coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, **329**, 533–8.

7. Maza,E., Frasse,P., Senin,P., Bouzayen,M. and Zouine,M. (2013) Comparison of normalization methods for differential gene expression analysis in RNA-seq experiments: A matter of relative size of studied transcriptomes. *Commun Integr Biol*, **6**, e25849.

8. Dillies,M.-A., Rau,A., Aubert,J., Hennequet-Antier,C., Jeanmougin,M., Servant,N., Keime,C., Marot,G., Castel,D., Estelle,J., *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*, **14**, 671–83.

9. Weiss,S., Xu,Z.Z., Peddada,S., Amir,A., Bittinger,K., Gonzalez,A., Lozupone,C., Zaneveld,J.R., Vázquez-Baeza,Y., Birmingham,A., *et al.* (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, **5**, 27.

10. Nixon,M.P., Letourneau,J., David,L.A., Lazar,N.A., Mukherjee,S. and Silverman,J.D. (2023) Scale reliant inference.

11. McMurrough,T.A., Dickson,R.J., Thibert,S.M.F., Gloor,G.B. and Edgell,D.R. (2014) Control of catalytic efficiency by a coevolving network of catalytic and noncatalytic residues. *Proc Natl Acad Sci U S A*, **111**, E2376–83.

12. Gloor,G., Macklaim,J., Vu,M. and Fernandes,A. (2016) Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Austrian Journal of Statistics*, **45**, 73–87.

13. Gloor,G.B. (2023) amIcompositional: Simple tests for compositional behaviour of high throughput data with common transformations. *Austrian Journal of Statistics*, **52**, 180–197.

14. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol*, **11**, R106.

15. McMurdie,P.J. and Holmes,S. (2014) Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Comput Biol*, **10**, e1003531.