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

High throughput sequencing (HTS) is a universal tool to explore many biological phenomenon such as gene expression (single-cell sequencing, RNA-sequencing, meta-transcriptomics), microbial community composition (16S rRNA gene sequencing, shotgun metagenomics) and differential enzyme activity (selex, CRISPR killing). HTS proceeds by taking a fixed-number sample from the environment, making a library, multiplexing (merging) multiple libraries and applying a fixed number of molecules to the flow cell. In essence it is a poll of the environment that is mixed with other polls and then a poll of the mixture is taken. It should be clear that the total number of molecules sequenced is driven by the capacity of the instrument and not by the number of molecules in the sampled environment.

More formally, if the true information in the environment being sampled can be summarized by counts of features (genes, taxa, etc) in a matrix, then the elements of the matrix can be decomposed into its composition (relative information) and its scale (total sum). We can formally state this as $W = W^{\parallel} \times W^{\perp}$. If the matrix has N samples and D parts, then we can uniquely identify W_{dn} as the $d\hat{}\{\text{th}\}$ feature in sample n. Any data that we collect by sequencing is thought to contains only proportional information and we can represent the corresponding sequenced value as Y. Compositional data analysis makes the strong assumption that $Y = Y^{\parallel}_{dn} = W^{\parallel}_{dn}$; in other words that relative information is the only information available post-sequencing. However, the values after sequencing are only estimates of the true underlying values that we want to estimate and different technical and biological replicates vary. Note that under this strong assumption no corresponding estimate is obtained for the scale $Y^{\perp}_{dn}$.

One issue that was realized very early was that if the goal was to compare one sample to another then the output from HTS needed to be normalized in order to make the samples commensurate and to correct any minor asymmetries in the data from different samples. A large number of normalizations were developed that depended on the data source. These include proportions and derivatives (reads per kilobase per million, and transcripts per kilobase per million), the relative log expression (RLE) and trimmed mean of M values (TMM), and the centered log ratio (CLR). All of these are ratios with the major differences between approaches being how the denominator is chosen and whether the ratio is always log transformed or not.

Recently, Pistner Nixon et al showed that the denominator used to normalize $Y^{\parallel}_{dn}$ was an estimate of the scale of the system. In the context of compositional data analysis, this group proved that the geometric mean of $Y^{\parallel}_{n}$ was a biased estimate of $W^{\perp}_{n}$. Thus, $Y^{\perp}_{n}$ could be made a less biased estimate by adding uncertainty when geometric mean of $Y^{\parallel}_{n}$ was calculated. In terms of information theory, the geometric mean of $Y^{\parallel}_{n}$ is strongly correlated with Shannon's Information I, suggesting that some knowledge of W is contained in the post-sequencing data. This makes intuitive sense because underlying systems with different scales will have different amounts of information and so we would expect $W^{\perp}_{n} \sim I$.

The realization that a strong assumption about scale is built into each denominator explains why HTS data are notoriously fickle as analysis of the same dataset with different workflows, tools and assumptions giving widely different results when benchmarked. This can be understood as the result of unacknowledged bias about the scale estimate which inevitably leads to large Type 1 error rates as sample sizes increase (**Gustafson2015?**). Thus, the realization that the denominator is a scale estimate opens up the possibility that we can supply alternate scale estimates and use this to determine which features are likely to be robust to differential abundance analysis even when taking scale into account.

Data that are generated by sequencing come from systems where scale is usually important and may be a confounding variable (Lovell et al. 2015). For example, cells transformed by the cMyc oncogene have about 3 times the amount of mRNA and about twice the rRNA content than do non-transformed cells (Nie et al. 2012), and this dramatically skews transcriptome

analysis (Lovén et al. 2012). In addition, wild-type and mutant strains of cell lines, yeast or bacteria often have different growth rates, which would affect our ability to identify truly differentially abundant genes (Yoshikawa et al. 2011). As another example, the total bacterial load of the vaginal microbiome differs by 1-2 orders of magnitude in absolute abundance (Zozaya-Hinchliffe et al. 2010), and the composition is dramatically different as well (Ravel et al. 2011; Hummelen et al. 2010). Thus, the full description of these systems includes both relative change (composition) and absolute abundance (scale) but we can access only the composition directly.

The ALDEx2 R package represents a general purpose toolbox to Bayesian estimation of HTS datasets. ALDEx2 was designed originally to convert point estimates of $Y_{dn}^{\parallel}$ into posterior distributions through Monte-Carlo sampling from the Dirichlet distribution, where the posterior more accurately represented variation on a per-sample basis. Each Monte-Carlo replicate is normalize using the CLR and used to calculate summary and test statistics, and finally expected values and confidence intervals are reported across the Monte-Carlo replicates. The CLR calculation uses as the denominator the geometric mean of each Monte-Carlo replicate of $Y_{dn}^{\parallel}$ with the geometric mean representing a point estimate of scale, i.e. of $Y_{n}^{\perp}$. We include a posterior distribution of the scale into this process by sampling $Y_{n}^{\perp}$ from a log-Normal distribution and using these as the denominator. The scale estimates can be calculated on a per-experiment, per-group or even a per-sample basis.

An advantage of incorporating scale is that the analysis can be made much more robust and that differences in scale can be accounted for explicitly. The examples below show how incorporating scale provides robust and interpretable differential abundance estimates in several different datasets.

1 Results

The first dataset is a highly replicated yeast transcriptome where one condition is wild-type and the other has a snf-1 gene knockout(Gierliński et al. 2015). Yeast deficient for snf-1 grow more slowly and are sensitive to a variety of common agents that cause cell stress (Yoshikawa et al. 2011). This dataset has been used to argue that only a small number of replicates need to be used to identity differentially abundant genes and that different tools should be used for datasets with different sample sizes because the tools have different intrinsic statistical power (Schurch et al. 2016). This guidance runs counter to standard statistical practice where power is intrinsically linked to sample size (Halsey et al. 2015), yet the concept of sample-size independent power is entrenched in all fields that use HTS as an experimental readout. Increasing false precision with increasing sample size can be understood as the result of unacknowledged bias brought about because of false certainty as to the scale of the data. Thus, by adding scale uncertainty we can be more confident that the analysis is not converging on a precise, but biased estimate of scale.

We start by examining the dispersion or variance of the data as counts and as the logarithm of the normalized counts as calculated by DESeq2 (log(RLE)) and by ALDEx2 (clr). The actual counts of data derived from sequencing are overdispersed with the mean value being less than the variance (Robinson, McCarthy, and Smyth 2010) as seen in Panel A of Figure 1. This relationship is why most tools model the counts with the Negative Binomial distribution(Frazee et al. 2023), and use it as the basis for batch correction(Zhang, Parmigiani, and Johnson 2020). However, the actual analysis of differential abundance is performed on the logarithm of the normalized counts [Robinson:2010, Love:2014aa], or on the clr values(Fernandes et al.

2013) both of which are log-ratios. The mean-dispersion distribution of these logarithmic-transformed data is quite different as shown in panels B and C of Figure 1 and as noted elsewhere Love, Huber, and Anders (2014).

```
FALSE [1] "Intercept"    "conds_W_vs_S"
```

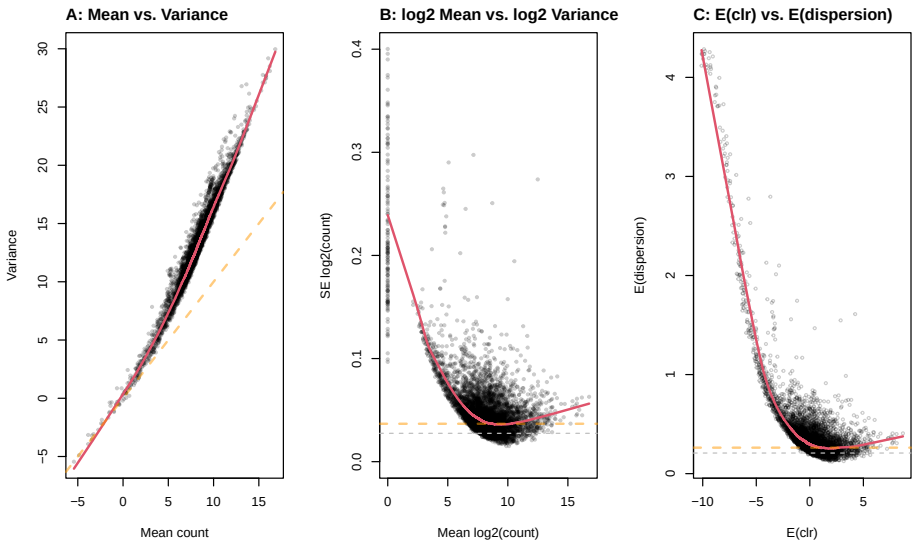

Figure 1: Plot of abundance v dispersion for a typical transcriptome dataset as counts, as logarithms of counts, and as clr values
Panel A shows that the data are over-dispersed relative to a Poisson distribution which is represented by the dashed line when plotted on a log-log scale. Panel B shows that 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. The expected clr value for each transcipt are plotted vs. the expected dispersion. In panels B and C the amount of dispersion reaches a minimum at moderate values. The red line in each panel shows the loess line of fit to the mid-point of the distributions. The dashed orange line in panel A is the line of equivalence, and in panel B and C is the minimum y value.

Here we can see that while the variance or dispersion always increases with increasing raw read count, it actually decreases when measured on the log-ratio transformed dataLove, Huber, and Anders (2014) and reaches 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 the known biology of cells where single cell counting of housekeeping transcripts shows that they are both highly expressed and have little intrinsic variation(Taniguchi et al. 2010). Furthermore, the actual amount of dispersion is very small and for many transcripts is almost negligible. To show this point more clearly, the majority of the transcripts in the lowest decile of dispersion are statistically significantly different (75% with DESeq2, 69% with ALDEx2), suggesting that low dispersion estimates lead to many false positives, and we suggest that this is because of the biased estimate of $Y^{\perp}$ that is being calculated. Thus, the actual variation of highly expressed genes is not captured accurately by current approaches to analysis and supports the idea that HTS have unacknowledged bias because scale is not taken into account.

Using either DESeq2 or ALDEx2, we observe that a majority of transcripts are statistically significantly different between groups even with a FDR of 0.01 ;4264 or 3791 of 5891 transcripts. Applying the rule of thumb of at least a $2^{1.4}$ fold change reduces these outputs to 193 for DESeq2 and to 188 for ALDEx2. Clearly, the necessary assumption of most features being invariant is not justified.

As shown in Figure 2 A,B,D,E the root cause of the many statistically significant positive transcripts is the very large number of transcripts with negligible variance with both DESeq2 and ALDEx2. We can see that almost all the transcripts that are differentially abundant with an FDR < 0.01 (orange and red points) have extremely low dispersion and a very low difference between groups. In the most extreme cases transcripts with near 0 difference have a low FDR. This issue lead to the common practice of choosing transcripts with a low FDR and a fold-change threshold (commonly set at $\pm 2^{1.4}$), and these limits are shown by the dashed grey lines. A similar sitation arises when using the ALDEx2 package, and indeed the two methods identify substantially similar transcripts. This results begs the question, "why bother with the signficance test?".

The very low dispersion estimate for most of the features arises because scale variation in the underlying data has been removed through sequencing and normalization. The actual scale of the data is unaccessible post-sequencing but we can estimate the effect of scale on the output. To do this, we add uncertainty to the denominator that is used to calculate the log-ratio of the samples, and combine this with the posterior probabilities that ALDEx2 calculates on a per-feature basis (Fernandes et al. 2013). Scale uncertainty is incorporated using the `gamma` parameter that controls the amount uncertainty being included when we call either `aldex()`, or `aldex.clr()`. The `ALDEx2` package contains a sensitivity analysis function, `aldex.senAnalysis()`, that can be used to explore the effect of different amounts of scale uncertainty. In practice we suggest that a `gamma` parameter between 0.5 and 1 is realistic for most experimental designs.

Applying `gamma=1` as a parameter we can see that the large number of transcripts with near 0 dispersion have had their dispersion increased (Figure 2C), and this results in many fewer transcripts (217) being called significantly different as shown in the volcano plot in Figure 1F (red points). Furthermore, overplotting the significant transcripts identified after adding scale uncertainty on the un-scaled analysis shows that adding scale uncertainty removes the need for the dual cutoff. Indeed, adding scale uncertainty reduces the significant transcripts to approximately the number observed with the somewhat arbitrary difference cutoff. Thus, incorporating scale uncertainty through the default scale model allows us to determine which variables are likely to be significant due to sequencing and normalization, and which are significantly different even with scale uncertainty included.

The second example dataset is a vaginal metatranscriptome dataset used in Wu et al. (2021), where we are comparing gene expression in bacteria collected from healthy (H) and BV-affected women. In this environment, both the relative abundance of species between groups is different as is the gene expression levels within a species (Macklaim et al. 2013). We further expect that the total number of bacteria is about 10X more in BV than in H (Zozaya-Hinchliffe et al. 2010). Thus, this is an extremely challenging environment to determine differential abundance. Indeed, the accepted method to analyze vaginal metatranscriptomes is to conduct a taxon by taxon analysis rather than a system-wide analysis (Macklaim et al. 2013; Deng et al. 2018; Fettweis et al. 2019) because a pooled analysis unexpectedly identifies many housekeeping genes as being differentially abundant between groups.

In this example we show how to specify the scale model explicitly and show that applying different scale models to each group can control for the very large difference in scale in the underlying data. When specifying the whole scale model we can pass a matrix of scale values instead of a single to `aldex.clr()`. This matrix should have the same number of rows as the of Monte-Carlo Dirichlet samples, and the same number of columns as the number of samples. This matrix encapsulates the additional uncertainty of the scale model on a per-sample basis.

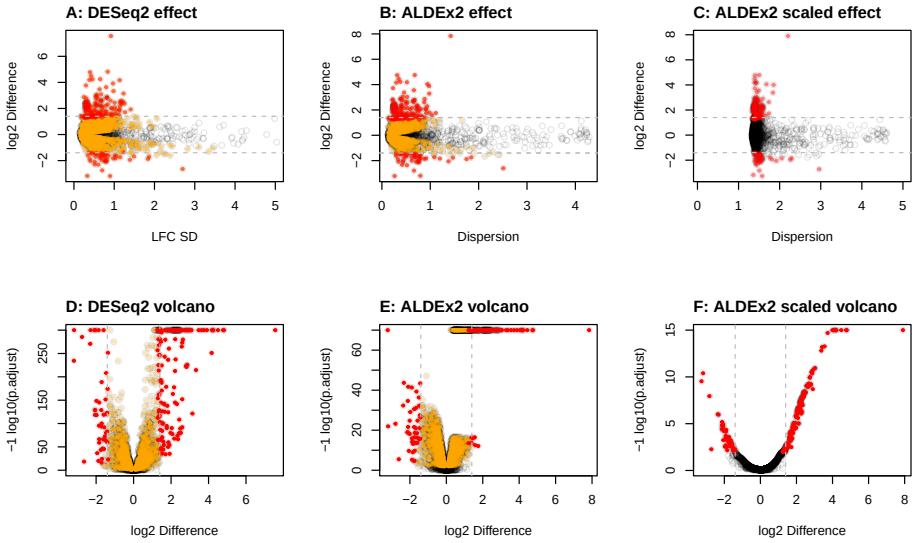

Figure 2: Effect and volcano plots for unscaled and scaled transcriptome analysis
DESeq2 or ALDEx2 were used to conduct a differential abundance (DA) analysis on the yeast transcriptome dataset. The results were plotted to show the relationship between difference and dispersion (effect plot) or difference and the Benjamini-Hochberg corrected p-values (volcano plot). Panels A,B,D,E are for the unscaled analysis, and Panels C,F are for the scaled analysis. Each point represents the values for one transcript, with the color indicating if that transcript was significant in the scaled analysis and unscaled analysis (red) or in the 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, which is a commonly applied cutoff when the majority of features are statistically significant.

Figure 3A shows an effect plot (Gloor, Macklaim, and Fernandes 2016) of the data where reads are grouped by function, corresponding approximately to orthologous proteins regardless of the organism of origin. Each point represents one of the 3728 functions, and we can see that there are many more functions represented in the BV group (bottom) than in the healthy group (top). This is because the *Lactobacilli* that dominate a healthy vaginal microbiome have reduced genome content relative to the anaerobic organisms that dominate in BV, because there is a greater diversity of organisms in BV than in H samples and because the BV condition has at least an order of magnitude more bacteria than does the H condition.

We can see that there are a large number of functions that are shared between the two groups (Figure 3A), and inspection shows that these largely correspond to core metabolic functions that would not be expected to contribute to differences in ecosystem behaviour. As a proxy for housekeeping functions the core ribosome functions (blue) shows that their mean location is not centred on 0.The major group of these housekeeping functions is located off the line of no difference (being approximately located at $+1.5$) and not surprisingly have among the lowest dispersion in the dataset. Nevertheless, they are identified as differentially abundant (red) along with many others. While changes in the abundance of housekeeping functions is a useful proxy for relative abundance of species in the environment, they tell us nothing about the functional capacity of the two groups because these are functions in common to every organism. Of more interest is determining the functions that are different between groups that are unique or over-expressed in one group relative to the other.

```
aldex.makeScaleMatrix <- function(gamma, diff, conditions){
  ## new scale model
  # mu1 is the base relative variance
  # mu2 is the other relative variance
```

```
  # gamma is the dispersion paramenter
  gamma=gamma
  mu1 = 1 # set base to 1
  mu2 = 1 + diff # set other to adjust this until the housekeeping is centred

  # note: it is the log2 difference between mu1 and mu2 that is key here
  # eg; mu1=1, mu2=1.15 is equivalent to mu1=4, mu2=4.6
  # log2(1)=0, log2(1.15)~0.2; log2(4)=2, log2(4.6)~2.2

  mu.vec <- gsub(levels(factor(conditions))[1], log2(mu1), conditions)
  mu.vec <- as.numeric(gsub(levels(factor(conditions))[2], log2(mu2), mu.vec))
  return(t( sapply(mu.vec, FUN = function(mu) rlnorm(128, mu, gamma)) ))
}
```

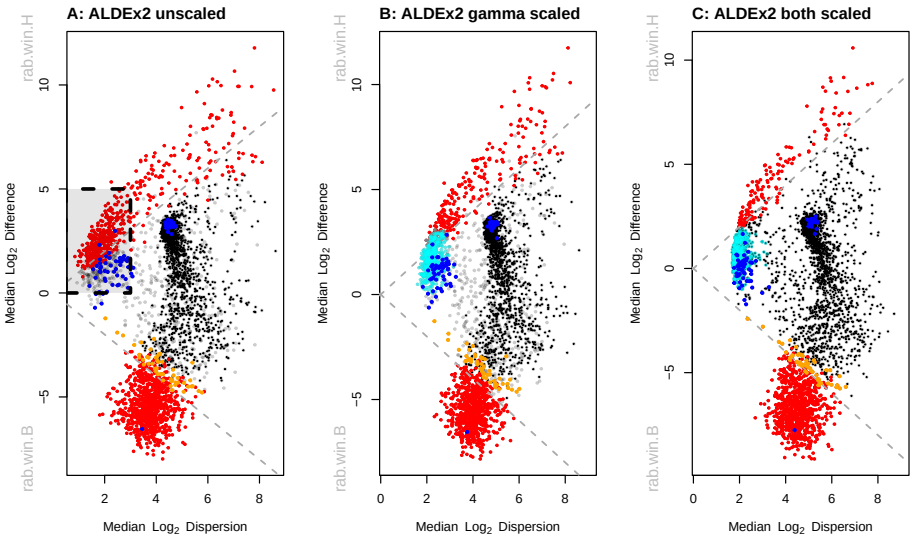

Figure 3: Analysis of vaginal transcriptome data aggregated at the Kegg Orthology (KO) functional level
Panel A shows the default analysis with samples from healthy individuals at the top and from BV individuals at the bottom. Highlighted in the box are highly abundant KOs that are almost exlusively housekeeping functions, with ribosomal KOs highlighted in blue, statistically significant (FDR < 0.01) functions in red, and non-significant functions in black or orange. These housekeeping functions are off the midline of no difference. Panel B shows the same data scaled with 'gamma $= 1$', which increase the minimum dispersion approximatly by one unit. Here the housekeeping functions from Panel A are colored cyan or blue for reference. Panel C shows the same data scaled with 'gamma $= 0.75$' and a 0.15 fold difference in dispersion applied to the BV samples relative to the H samples. The orange functions are now statistically significant. Note that this shifts the midpoint of the housekeeping functions towards the midline.

Applying `gamma=1` as before increases the dispersion as expected, but does little to move the large number of housekeeping functions toward the midline of no difference. Nevertheless, about 50% of the housekeeping functions are no longer statistically significantly different.

Up to this point, scale uncertainty has been applied uniformly to both conditions, but the scale adjustment can be applied to each condition, or even each sample independently through a custom scale matrix. This can be done quite simply with the `aldex.makeScaleMatrix()` function which produces a matrix of scale uncertainties that are distinct for each group. Applying a differential scale of 0.15, or 15% of the base scale now moves the housekeeping functions to the midline of no difference. Differential scale has the property that if the

differences in underlying scale of the system is large, then the sign of the differential scale is irrelevant because . . . [HELP]. Note that this identifies a significant number of functions that are differentially up in BV that were formerly classed as not different without scale, or when only a uniform scale was applied. These former false negatives are noted in orange in each panel. Inspection of the functions shows that these are largely missing from the Lactobacillus species and so should actually be captured as differentially abundant. Thus, applying differential scale allows us to distinguish between both false positives (housekeeping functions in cyan) and false negatives (orange functions) even in a very difficult to analyze dataset.

Differential scale has the property that if the differences in underlying scale of the system is large, then the sign of the differential scale is irrelevant.

2 Discussion

Biological count data can be decomposed into two parts the relative (compositional) and the absolute (scale), and the product of these generates a fully scaled biological system (Nixon et al. 2023). Biological systems are inherently variable and stochastic and current measurement methods that rely on high throughput sequencing fail to capture all of that variation. In the absence of information external to the sequencing run itself, no normalisation method can recapture any of the scale information, including scale variation (Lovén et al. 2012).

While the underlying scale of the system cannot be measured easily, the effect on analysis can by including scale uncertainty in the analysis (Nixon et al. 2023). Nixon et al, showed that this can be done by including uncertainty in the denominator used for the normalization. The ALDEx2 R package is ideally suited for this since this tool builds a Bayesian posterior of the compositional component of the dataset at the outset and then conducts the analysis on that posterior. Adding scale uncertainty can be done at the same time thus producing a posterior model that incorporates both compositional and scale uncertainty. For this, the compositional uncertainty is sampled from a Dirichlet distribution, and the scale uncertainty is sampled from a log-Normal distribution.

All normalizations attempt to make the samples in a dataset commensurate but cannot explicitly address the scale of the underlying system. However, the general lack of scale information has important consequences for the analysis of HTS datasets. One issue is that analysis tools seem over-powered with even moderate sample sizes (Schurch et al. 2016). Using small sample sizes in analysis leads to less reliability and reproducibility in analyses since surprisingly large sample sizes are needed to determine reliable p-values (eg. (Halsey et al. 2015)). Thus, recommendations to use small sample sizes in multivariate datasets such as RNA-seq datasets are not supported by simple modelling in the univariate case. Another issue is that datasets are difficult to analyze when there they contain systematic asymmetry, with different tools exhibiting differing pathologies with these datasets Wu et al. (2021).

In the case of overpowering, HTS analyses seem to be more robust when applying a dual cutoff of both p-value and difference between group means (Schurch et al. 2016). Figure 2 shows one reason for this robustness could be that the dual cutoff is mimicking the effect of including scale uncertainty, since substantially similar transcripts are identified by the two approaches. However, while using the post-hoc difference cutoff is useful for differential abundance analysis it is not clear how this can be incorporated into other kinds of downstream analyses. Conversely data that include scale uncertainty are fully compatible with existing downstream analyses.

In the case of asymmetry, the use of a user-specified scale model can be very useful for otherwise difficult to analyze datasets such as meta-transcriptomes and in-vitro selection datasets where the majority of features can change. We showed one such example in Figure 3 where the dataset was highly asymmetrical, and the TMM and RLE normalizations cannot be used. Incorporating differential scale on a per-group basis moves the mass of the data towards the midline of no difference and so affects both Type I and Type II error rates. Differential scale has the property that if the differences in underlying scale of the system is large, then the sign of the differential scale is irrelevant. In this analysis, transcripts that were previously not classed as differentially abundant are now called as significantly different, and the housekeeping transcripts move from being significantly different to not being identified as such. While we acknowledge that some prior information on which housekeeping transcripts should not be classed as DA is needed, we suggest that this information is widely available and used when performing the gold-standard quantitative PCR test of differential abundance SEQC/MAQC-III Consortium (2014). Thus, the use of this prior knowledge is not unique to our approach.

In summary, while the underlying scale of the system is generally inaccessible, the effect of scale on the analysis outcomes can be modelled. Adding scale information to the analysis allows for more robust inference because the features that are sensitive to scale can be identified and their impact on the analysis weighted accordingly. Additionally, the use of differential scale models permits difficult to analyze datasets to be examined in a robust and principled manner even when the majority of features are asymmetrically distributed or expressed (or both) in the groups. Thus, reportin scale uncertainty should become a standard practice in the analysis of HTS datasets as a way to identify which features are most robust to differences in the underlying system. Furhtermore, we supply a set of toosl that make incorporating scale simple even for datasets that come from highly asymmetrical environments.

Deng, Zhi-Luo, Cornelia Gottschick, Sabin Bhuju, Clarissa Masur, Christoph Abels, and Irene Wagner-Döbler. 2018. "Metatranscriptome Analysis of the Vaginal Microbiota Reveals Potential Mechanisms for Protection Against Metronidazole in Bacterial Vaginosis." Edited by Craig D. Ellermeier, Janet Hill, and Andrew Onderdonk. *mSphere* 3 (3). https://doi.org/10.1128/mSphereDirect.00262-18.

Fernandes, Andrew D, Jean M Macklaim, Thomas G Linn, Gregor Reid, and Gregory B Gloor. 2013. "ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq." *PLoS One* 8 (7): e67019. https://doi.org/10.1371/journal.pone.0067019.

Fettweis, Jennifer M, Myrna G Serrano, J Paul Brooks, David J Edwards, Philippe H Girerd, Hardik I Parikh, Bernice Huang, et al. 2019. "The Vaginal Microbiome and Preterm Birth." *Nat Med* 25 (6): 1012–21. https://doi.org/10.1038/s41591-019-0450-2.

Frazee, Alyssa C., Andrew E. Jaffe, Rory Kirchner, and Jeffrey T. Leek. 2023. "polyester: Simulate RNA-seq reads." https://doi.org/10.18129/B9.bioc.polyester.

Gierliński, Marek, Christian Cole, Pietà Schofield, Nicholas J Schurch, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, et al. 2015. "Statistical Models for RNA-Seq Data Derived from a Two-Condition 48-Replicate Experiment." *Bioinformatics* 31 (22): 3625–30. https://doi.org/10.1093/bioinformatics/btv425.

Gloor, Gregory B, Jean M. Macklaim, and Andrew D. Fernandes. 2016. "Displaying Variation in Large Datasets: Plotting a Visual Summary of Effect Sizes." *Journal of Computational and Graphical Statistics* 25 (3C): 971–79. https://doi.org/10.1080/10618600.2015.1131161.

Halsey, Lewis G, Douglas Curran-Everett, Sarah L Vowler, and Gordon B Drummond. 2015. "The Fickle p Value Generates Irreproducible Results." *Nat Methods* 12 (3): 179–85. https://doi.org/10.1038/nmeth.3288.

Hummelen, Ruben, Andrew D Fernandes, Jean M Macklaim, Russell J Dickson, John Changalucha, Gregory B Gloor, and Gregor Reid. 2010. "Deep Sequencing of the Vaginal Microbiota of Women with HIV." *PLoS One* 5 (8): e12078. https://doi.org/10.1371/journal.pone.0012078.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biol* 15 (12): 550.1–21. https://doi.org/10.1186/s13059-014-0550-8.

Lovell, David, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. 2015. "Proportionality: A Valid Alternative to Correlation for Relative Data." *PLoS Comput Biol* 11 (3): e1004075. https://doi.org/https://doi.org/10.1371/journal.pcbi.1004075.

Lovén, Jakob, David A Orlando, Alla A Sigova, Charles Y Lin, Peter B Rahl, Christopher B Burge, David L Levens, Tong Ihn Lee, and Richard A Young. 2012. "Revisiting Global Gene Expression Analysis." *Cell* 151 (3): 476–82. https://doi.org/10.1016/j.cell.2012.10.012.

Macklaim, Jean M, Andrew D Fernandes, Julia M Di Bella, Jo-Anne Hammond, Gregor Reid, and Gregory B Gloor. 2013. "Comparative Meta-RNA-Seq of the Vaginal Microbiota and Differential Expression by Lactobacillus Iners in Health and Dysbiosis." *Microbiome* 1 (1): 12. https://doi.org/10.1186/2049-2618-1-12.

Macklaim, Jean M, and Gregory B Gloor. 2018. "From RNA-Seq to Biological Inference: Using Compositional Data Analysis in Meta-Transcriptomics." *Methods Mol Biol* 1849: 193–213. https://doi.org/10.1007/978-1-4939-8728-3/_13.

Nie, Zuqin, Gangqing Hu, Gang Wei, Kairong Cui, Arito Yamane, Wolfgang Resch, Ruoning Wang, et al. 2012. "C-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells." *Cell* 151 (1): 68–79. https://doi.org/10.1016/j.cell.2012.08.033.

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.

Ravel, Jacques, Pawel Gajer, Zaid Abdo, G Maria Schneider, Sara S K Koenig, Stacey L McCulle, Shara Karlebach, et al. 2011. "Vaginal Microbiome of Reproductive-Age Women." *Proc Natl Acad Sci U S A*, no. 108: 4680–87. https://doi.org/doi/10.1073/pnas.100611107.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

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.

Schurch, Nicholas J, Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, et al. 2016. "How Many Biological Replicates Are Needed in an RNA-Seq Experiment and Which Differential Expression Tool Should You Use?" *RNA* 22 (6): 839–51. https://doi.org/10.1261/rna.053959.115.

SEQC/MAQC-III Consortium. 2014. "A Comprehensive Assessment of RNA-Seq Accuracy, Reproducibility and Information Content by the Sequencing Quality Control Consortium." *Nat Biotechnol* 32 (9): 903–14. https://doi.org/10.1038/nbt.2957.

Taniguchi, Yuichi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. 2010. "Quantifying e. Coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells." *Science* 329 (5991): 533–38. https://doi.org/10.1126/science.1188308.

Thellin, O, W Zorzi, B Lakaye, B De Borman, B Coumans, G Hennen, T Grisar, A Igout, and E Heinen. 1999. "Housekeeping Genes as Internal Standards: Use and Limits." *J Biotechnol* 75 (2-3): 291–95.

Wu, Jia R., Jean M. Macklaim, Briana L. Genge, and Gregory B. Gloor. 2021. "Finding the Centre: Compositional Asymmetry in High-Throughput Sequencing Datasets." In *Advances in Compositional Data Analysis: Festschrift in Honour of Vera Pawlowsky-Glahn*, edited by Peter Filzmoser, Karel Hron, Josep Antoni Martìn-Fernàndez, and Javier Palarea-Albaladejo, 329–46. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-71175-7_17.

Yoshikawa, Katsunori, Tadamasa Tanaka, Yoshihiro Ida, Chikara Furusawa, Takashi Hirasawa, and Hiroshi Shimizu. 2011. "Comprehensive Phenotypic Analysis of Single-Gene Deletion and Overexpression Strains of Saccharomyces Cerevisiae." *Yeast* 28 (5): 349–61. https://doi.org/10.1002/yea.1843.

Zhang, Yuqing, Giovanni Parmigiani, and W Evan Johnson. 2020. "ComBat-Seq: Batch Effect Adjustment for RNA-Seq Count Data." *NAR Genom Bioinform* 2 (3): lqaa078. https://doi.org/10.1093/nargab/lqaa078.

Zozaya-Hinchliffe, Marcela, Rebecca Lillis, David H Martin, and Michael J Ferris. 2010. "Quantitative PCR Assessments of Bacterial Species in Women with and Without Bacterial Vaginosis." *J Clin Microbiol* 48 (5): 1812–19. https://doi.org/10.1128/JCM.00851-09.