Estimating differences in scale for high throughput sequencing 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

16 August 2023

Abstract

Introduction to scale simulation and FDR correction with ALDEx2.

Package

ALDEx2 1.33.1

Contents

1 scale simulation for the no-math crowd

Beyond compositionally in high throughput sequencing; estimating the importance of scale in data analysis

2 Introduction

High throughput sequencing (HTS) analysis is commonly used to read out many different types of experimental designs; bulk and single-cell transcriptomics, amplicon and shotgun metagenomics, in vitro selection experiments, and more. There are many analysis tools that are purpose-built for each experimental design and that work well when all assumptions are fulfilled. However, different tools make different assumptions about the underlying data and when the tools fail, they usually fail without error.

Despite the variety, all tools and analyses suffer from two common issues. First, that many features are statistically significantly different between groups even though the measured difference is very small. Second, that the results can be skewed if the underlying dataset has an asymmetric distribution between the groups. Here we show that both of these issues can be resolved by modelling the size, or scale, of the data in the underlying environment.

It is instructive to understand how the data are generated. HTS starts by making a 'library' which is a fixed-size sample from the environment. The library is usually combined with other libraries through 'multiplexing' where the goal is to combine the libraries such that each library contributes equivalent numbers of molecules to the pool. Finally, a fixed number of molecules is sampled from the pooled samples and loaded onto the instrument. Thus, the process of sequencing is akin to taking a poll (making a library), combining polls (multiplexing into a pool) and taking a poll of the pool. It should be clear that this process removes any relationship between actual numbers in the environment and the actual number of molecules actually sequenced. Thus, the total number of molecules that are sequenced for a sample is simply a nuisance parameter.

All HTS data must be standardized and normalized after sequencing to make the different samples comparable because the total count per sample is non-informative. All of these normalizations are ratios but there are two main main approaches to identifying the denominator. The first class of normalization uses a denominator that is internal to each sample. This can be as simple as converting the counts to proportions; i.e., by dividing the count for each feature in a sample by the total count of the sample. Derivatives of this method include the TPM (transcripts per million), FPKM(fragments per million), and RPKM (reads per kilobase per million) (Mortazavi et al. 2008). In these approaches the proportion is multiplied by several constants that depend on the instrument, gene characteristics in the sample and the sequencing depth. These normalizations track closely the counts of genes determined using fluorescent hybridization methods, with typical correlations above 0.8 Taniguchi et al. (2010). The other internal-only normalization is the CLR (centred log-ratio) transform(Aitchison 1982). Here, the count of each feature in each sample is divided by the geometric mean of the counts in the sample with a logarithm being taken of the resulting ratio. In this case, a small pseudo-count or prior is first applied since a log(0) is undefined. The second class of normalization uses a denominator that is determined from data external to the sample being normalized, but that is still derived only from the sequencing data itself. Normalizations such as the RLE (relative log expression)(Anders and Huber 2010), TMM (trimmed mean of M values)(Robinson and Oshlack 2010) and CSS (cumulative sum scaling)(Paulson et al.

2013) methods assume that a majority of the features are invariant and that a relative scale can be determined for each sample by dividing by a correction factor. These normalizations specifically normalize the read counts for each feature to a common total so that the counts can be compared between samples directly. In particular the TMM normalization was intended to account for asymmetry in occurrance or abundance in one group relative to the other, but assumes that the majority of features are invariant (Robinson and Oshlack 2010). Note that the CLR normalization also assumes that a majority of the parts are invariant (Wu et al. 2021).

A final complication arises during analysis when logarithmic transformations may or may not be applied, and the CLR is the only transformation that explicitly uses log-transformed data as the starting point. Several groups have pointed out that apparent differences between the behaviour of individual normalizations Skinnider, Squair, and Foster (2019) can be largely explained by differential application of logarithms Erb (2020). Indeed in many datasets a logarithm of one of these transforms exhibits behaviours that is similar to the logarithm of one or more other transforms (Gloor 2023). Finally, each normalization is a point estimate of the appropriate denominator and so an additional assumption is that the denominator is a good estimator of the measure.

These tranformations along with the process of sequencing itself remove all but the relative variation between samples in the dataset. Nevertheless, 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 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.

It is enlightening to examine the dispersion or variance of the data as counts and as the logarithm of counts. The 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 and this is why most tools model the counts with this distribution. However, the actual analysis is usually performed on the logarithm of the counts [Robinson:2010, Love:2014aa], or on the logarithm of the clr values(Fernandes et al. 2013) and the mean-dispersion distribution of the logarithm of the counts is quite different as shown in panel B of Figure 1 and as noted elsewhere Love, Huber, and Anders (2014).

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

By not accounting for scale we are usually identifying statistically significant changes that are biologically meaningless but are found because of the artificially low dispersion. As one example, recent guidance suggests *using only a small number of samples* for bulk RNA-sequencing because only a few samples is needed to identify the majority of statistically significant features found with a large sample size (Schurch et al. 2016). This popularized the useage of both p-values and log2 fold change between groups as dual cutoffs when determining significance. However, it is also possible to fail to identify true positive differences because of scale (Nixon et al. 2023).

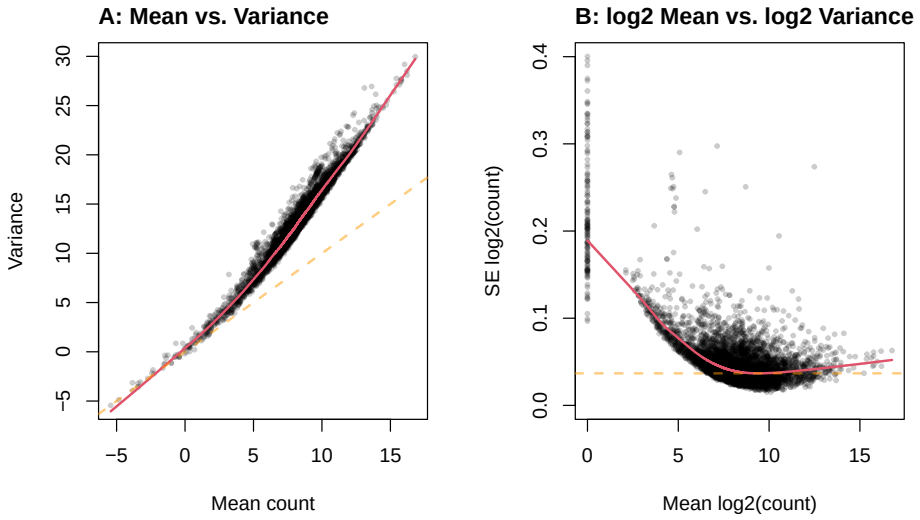

Figure 1: Plot of abundance v dispersion for a typical transcriptome dataset as both counts and as logarithms of counts
Panel A shows that the data are over-dispersed relative to a Poisson distribution which is represented by the dashed line when plotted on a log-log scale. Panel B shows that the relationship between the mean and the dispersion, here the standard error (SE) of the mean, is very different when the data are log-transformed first. In addition, the amount of dispersion reaches a minimum at moderate count 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 is the minimum y value.

Clearly we need some way to account for the variation in, and effect of, the scale of the underlying systems (Nixon et al. 2023) when determining differential abundance (DA) using high throughput sequencing. While normalizaitons attempt to correct for differences in relative counts, no tool accounts for differences in the intrinsic or extrinsic scale of the underlying system. Moreover, while we cannot measure scale directly, we can estimate its effect post-hoc on results.

3 Results - examples:

We can model the conundrum of scale by assuming that the true values from the system that we want to measure are counts composed of both compositional and total values, and that their product is the full scaled system that fits the equation: $W = W^{\parallel}W^{\perp}$. If the true data have D features (genes, taxa, etc) and N samples, then W is a $D \times N$ matrix of values and we can denote the true count of a feature in a sample as W_{dn}. However, sequencing only does not allow us to access $W^{\parallel}$ directly, but instead returns only a single estimate of the composition for each feature $Y^{\parallel}_{dn}$. This single value is not definitive since different technical replicates return different numbers for $Y^{\parallel}_{dn}$ because of sampling variation Fernandes et al. (2013).

We can use Bayesian techniques to estimate both the sampling variation of $W^{\parallel}$ and the scale uncertainty of $W^{\perp}$ by Monte Carlo sampling with the ALDEx2 R package. ALDEx2 generates pseudo technical replicates of $Y^{\parallel}$ by Monte Carlo sampling from a Dirichlet distribution to produce a posterior distribution for $Y^{\parallel}$ that is an estimate of $W^{\parallel}$Gloor et al. (2016). Then a CLR transformation is applied to each Monte Carlo sample to produce a distribution of log-ratio transformed values. Crucially, the denominator used for the CLR can be seen as a

point estimate of the scale of the system (Nixon et al. 2023). Thus, modifying ALDEx2 by replacing the point estimate for the CLR denominator with Monte Carlo sampled values from a log-Normal distribution gives a distribution of credible denominators. Thus, the CLR values now are credible estimates for both $W^{\parallel}$ and $W^{\perp}$. This allows the scaled CLR distribution to be used to estimate the effect of scale on the output and so make $Y \sim W$ under the assumption of the scale uncertainty. Below we use two example datasets to show how including scale uncertainty can be used to make inferences about differentially abundant features more reliable.

The first dataset is a highly replicated yeast transcriptome where one condition is wild-type and the other has a snf-1 gene knocK0ut(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 is entrenched in all fields that use HTS as an experimental readout.

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.

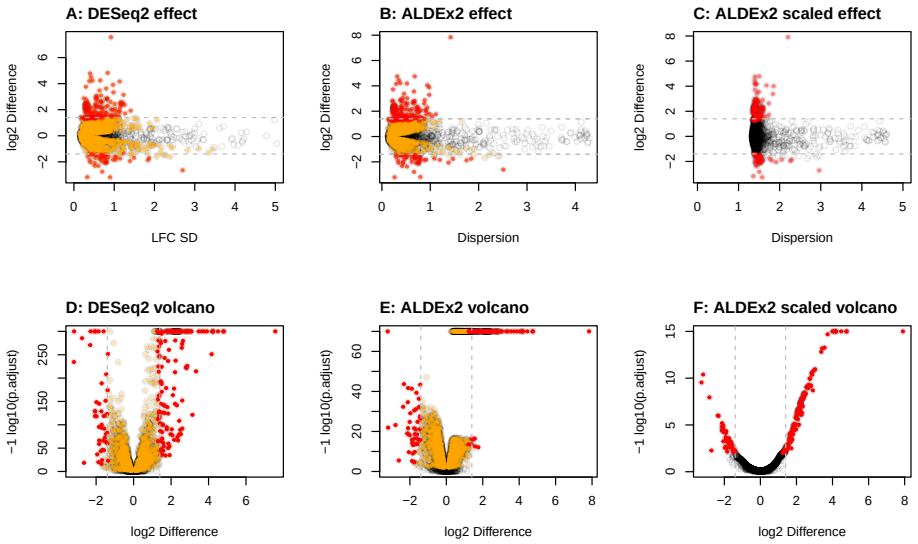

Figure 2: Effect and volcano plots for unscaled and scaled transcriptome analysis
DESeq2 or ALDEx2 were used to conduct a differential abundance (DA) analysis on the yeast transcriptome dataset. The results were plotted to show the relationship between difference and dispersion (effect plot) or difference 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.

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 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)) ))
}
```

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,

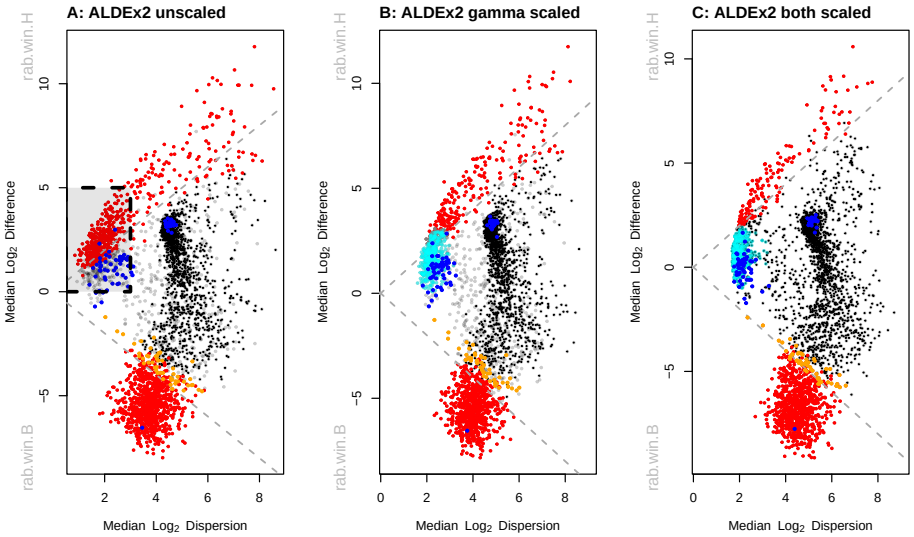

Figure 3: Analysis of vaginal transcriptome data aggregated at the Kegg Orthology (KO) functional level
Panel A shows the default analysis with samples from healthy individuals at the top and from BV individuals at the bottom. Highlighed in the box are highly abundant KOs that are almost exlusively housekeeping 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 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.

4 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.

5 Methods or supplement

Count data can in the underlying system be decomposed into two parts; the relative amount (composition) and the total amount (scale). We can describe the full scaled system made up of the composition and the scale with the equation: $W = W^{\parallel}\dot{W}^{\perp}$. What we measure by sequencing is only the composition part $Y^{parallel}$ which is a single point estimate of the composition $W^{parallel}$ of the underlying true system. We can denote the sequence count dataset as an $D \times N$ matrix, with elements Y_{dn} indicating the number of sequenced DNA molecules mapping to the d^{th} entity (e.g., taxa or gene) in the n^{th} sample. We will use hat notation to denote an estimate of a quantity (e.g., $\hat{W}$ is an estimate of W).

It is possible that the relative and size information can vary independently. For example, the total number of RNA molecules in a cell may vary by type, but the relative amounts of many of them may be constant; alternatively the total number of molecules may be the same, but the relative proportions may change. Of course, there could be a mixture of both possibilities as well.

Making a library from the molecules in W, sequencing and generating the count table from the output provides a single point estimate of $W^{\parallel}$. The accuracy with which $Y^{\parallel}$ is a good estimate of $W^{\parallel}$ is determined by the number of samples, the number of DNA molecules sequenced relative to the size of the system and is affected by the bioinformatic pipelines. Nevertheless, it is assumed that $Y^{\parallel} \sim W^{\parallel}$ under these constraints.

- $Y^{\parallel}$ is a point estimate of the data and under-samples the true underlying distribution because of sparse random sampling and because the number of samples sequenced is usually very small.
- ALDEx2 uses $Y^{\parallel}_{d.}$ as the basis to generate a posterior distribution by Monte-Carlo sampling from a Dirichlet distribution, where the posterior distribution contains a large number of credible values for $W^{\parallel}$.
- The Monte-Carlo instances are then CLR transformed. Here the denominator of the CLR is a single point estimate of the underlying scale.
- ALDEx2 has been modified include Monte-Carlo sampling from a log-Normal distribution to generate a posterior distribution of credible scale values, given the estimated scale in W. The amount of scale is controlled either by providing a single value γ which is the standard deviation (SD) of the scale, or by providing a matrix of SDs that vary between groups.
- This has the result of further smoothing and broadening the distribution to account for the uncertainty in the scale of the system from which the data was drawn.
- Applying uniform scale uncertainty increases the dispersion of the

So in formal terms, the system we want to measure is a set of N samples with D parts (genes, species, etc) contained in a matrix or data table W. The observed data is in a matrix Y, which contains the same We incorporate measurement uncertainty as uncertainty around the observed value We incorporate scale uncertainty as uncertainty around the correction used (Gx) - contant uncertainty increases dispersion - different amounts of uncertainty increases dispersion and moves the center

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.

Costea, Paul I, Georg Zeller, Shinichi Sunagawa, and Peer Bork. 2014. "A Fair Comparison." *Nat Methods* 11 (4): 359. https://doi.org/10.1038/nmeth.2897.

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.

Erb, I. 2020. "Partial Correlations in Compositional Data Analysis." *Applied Computing and Geosciences* 6 (6): 100026.

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.

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. 2023. "amIcompositional: Simple Tests for Compositional Behaviour of High Throughput Data with Common Transformations." *Austrian Journal of Statistics* 52 (4): 180–97.

Gloor, Gregory B, Jean M. Macklaim, 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.

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

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.

Marioni, John C, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. 2008. "RNA-seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays." *Genome Res* 18 (9): 1509–17. https://doi.org/10.1101/gr.079558.108.

Moffitt, Jeffrey R, Junjie Hao, Guiping Wang, Kok Hao Chen, Hazen P Babcock, and Xiaowei Zhuang. 2016. "High-Throughput Single-Cell Gene-Expression Profiling with Multiplexed Error-Robust Fluorescence in Situ Hybridization." *Proc Natl Acad Sci U S A* 113 (39): 11046–51. https://doi.org/10.1073/pnas.1612826113.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-seq." *Nat Methods* 5 (7): 621–28. https://doi.org/10.1038/nmeth.1226.

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.

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. 2013. "Differential Abundance Analysis for Microbial Marker-Gene Surveys." *Nat Methods* 10 (12): 1200–1202. https://doi.org/10.1038/nmeth.2658.

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.

Skinnider, Michael A, Jordan W Squair, and Leonard J Foster. 2019. "Evaluating Measures of Association for Single-Cell Transcriptomics." *Nat Methods* 16 (5): 381–86. https://doi.org/10.1038/s41592-019-0372-4.

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.

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.