Estimating differences in scale for high throughput sequencing analysis with ALDEx2

Greg Gloor, Michelle Pistner, Justin Silverman [*1]

[1]Dep't of Biochemistry, University of Western Ontario, Penn State

[*]ggloor@uwo.ca

19 July 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

HTS is pervasive in the life sciences. It is used to read out many different types of experimental designs; bulk and single-cell transcriptomics, shotgun and amplicon metagenomics, in vitro selection experiments, and more. There are a large number of analysis tools that are purpose-built for each experimental design with limited cross-over. In part this is because different tools make different assumptions about the underlying data that may or may not be met.

The process of sequencing a DNA fragment with high throughput sequencing (HTS) includes wet-lab methods that introduce unacknowledged bias that affects the downstream statistical analysis. In many cases in many datasets these biases may be minor and manageable. However, often the biases are unknown and ignored by both the experimentalist and the analyst.

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 again the goal is to combine the libraries such that each library contributes the the same number of molecules to the pool. Finally, a fixed number of molecules is sampled from the pool 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. This process removes any relationship between actual numbers sampled from the environment and the number of molecules sequenced; that is, it removes the effect of scale on the output data.

All HTS data are normalized after sequencing to make the different libraries comparable. There are two main normalization methods. The first type of normalization uses an internal standard. This can be as simple as converting the counts to proportions; i.e., by dividing the count for each part in a sample by the total count of the sample. Derivatives of this method include the TPM and RPKM approaches where the proportion is multiplied by a constant (TPM by 10^6) or by a pair of constants (RPKM by 10^6 and the length of the part). The other internal-only normalization is the CLR, where the count of each part is divided by the geometric mean of the counts of the parts in each sample and a logarithm is taken of the resulting ratio. In this case, a small pseudo-count or prior is first applied. The second class of normalization uses an external standard. Normalizations such as the RLE, TMM and CSS methods assume that the majority of the parts are invariant and can be identified and then used to determine a relative scale between samples. Each sample is then normalized by multiplying or dividing by this relative scale to produce normalized counts.

It should be clear that all normalizations are ratios where the only difference is in how the denominator is chosen. A further complication arises during analysis when logarithmic transformations may or may not be applied, and the CLR (or other log-ratio transform) is the only transformation that explicitly uses as log-transformed data as the starting point. Several groups have pointed out that apparent differences between normalizations (2Paulson:2013aa, 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 the transforms exhibits behaviours that is similar to the logarithm of another transform (Gloor 2023).

These tranformations all have the effect of further reducing the variation between samples in the dataset such that the only variation remaining is relative variation. Nevertheless, there are many instances where the size of the measured system is an important 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 can have differnt gorwth 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). Finally, antibiotic treatment can change both the total abundance and the composition of microbiomes grown in culture and in the human gut [???].

Clearly we need some way to account for the variation in, and effect of, scale of the underlying systems (Nixon et al. 2023) when determining differential abundance using high throughput sequencing. A proper model for DA estimates both the difference in abundance (location) and the difference in size (scale), but the differences in scale are unavailable after sequencing. However, the tools currently used to estimate measurement error (shot noise) use either the NB (or modified derivative) or Dirichlet models post sequencing, and these are used to estimate only location. No tool accounts for differences in the intrinsic or extrinsic scale. While we cannot measure scale directly, we can estimate its effect post-hoc on results. By not accounting for scale we are usually committing Type I errors and presuming we have much more power than is actually available, but can also commit Type II errors (Nixon et al. 2023). As one example, recent guidance suggests *a very small number of samples* for RNA-sequencing because of overpowering (Schurch et al. 2016).

3 Results - examples:

We use two example datasets that show common types of problems.

The first dataset is a highly replicated yeast transcriptome where one condition is wild-type and the other has a snf-1 gene knockout. 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 because increasing sample size results in more false positive identifications of differential abundance (Schurch et al. 2016). As shown in Figure 1 A,B,D,E the root cause of the false positives is the very large number of genes with low or even negligible variance. We can see that almost all the transcripts that are differentially abundant with a FDR < 0.01 (red) have extremely low dispersion. In the most extreme cases transcripts with near 0 difference have a low FDR, leading to the common practice of choosing transcripts with a low FDR and at least a Difference of $2^{1.4}$, 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?".

It should be obvious that the root cause is the very low dispersion parameter which arises because scale variation in the underlying data has been removed through sequencing. While we cannot determine the underlying scale of the data, we can estimate the effect on the

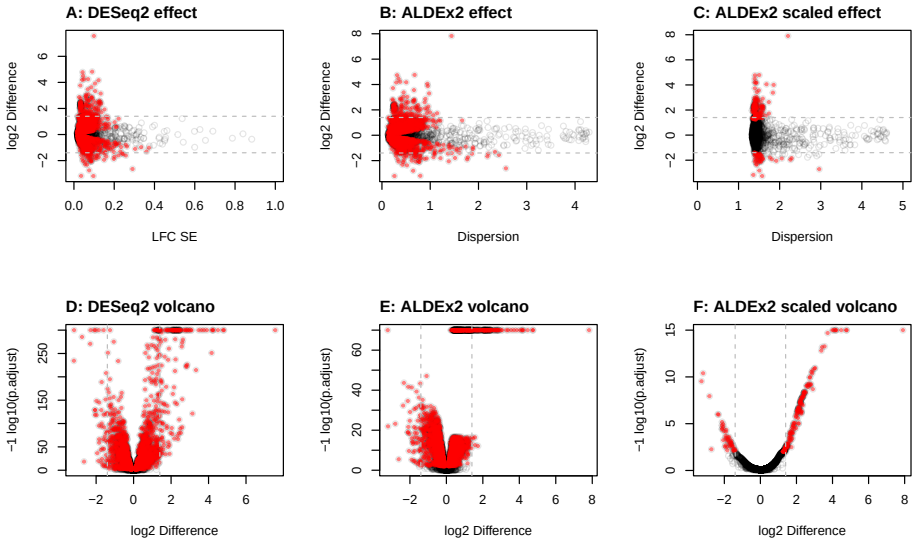
output if scale could be included. In the case of simple underestimation of variance as in this dataset, we can use the `gamma=1` parameter when we call either `aldex()`, or `aldex.clr()` to add uncertainty around the differential scale of the data.

Applying this parameter we can see that the large number of transcripts with near 0 dispersion now have substantial dispersion (Figure 1C), and this results in many fewer transcripts being called significantly different as shown in the volcano plot in Figure 1F. Indeed, adding scale reduces the significant transcripts to approximately the number observed with the somewhat arbitrary difference cutoff.

```
library(DESeq2)
devtools::load_all('~/Documents/0_git/ALDEx_bioc')
#library(ALDEx2)

yst <- read.table('~/Documents/0_git/datasets/transcriptome.txt', header=T,
  row.names=1)
# Gierlinski:2015aa
yst[,c('SNF2.6', 'SNF2.13','SNF2.25','SNF2.35')] <- NULL
yst[,c('WT.21','WT.22','WT.25','WT.28','WT.34','WT.36')] <- NULL
conds <- c(rep('S', 44), rep('W', 42))
coldata <- data.frame(conds)

# DESeq2
dds <- DESeqDataSetFromMatrix(countData = yst,
          colData = coldata, design= ~ conds)
dds <- DESeq(dds)
resultsNames(dds) # lists the coefficients
FALSE [1] "Intercept"    "conds_W_vs_S"
res <- results(dds, name="conds_W_vs_S")
set.seed(2023)
#ALDEx2
x <- aldex.clr(yst, conds, verbose=F)
x.e <- aldex.effect(x, verbose=F)
x.t <- aldex.ttest(x, verbose=F)
set.seed(2023)
x.s <- aldex.clr(yst, conds, gamma=1, verbose=F)
x.s.e <- aldex.effect(x.s, verbose=F)
x.s.t <- aldex.ttest(x.s, verbose=F)

sig.des <- which(res@listData$padj < 0.01)
sig.ald <- which(x.t$we.eBH < 0.01)
sig.s.ald <- which(x.s.t$we.eBH < 0.01)
```

In the unscaled situaiton, DESeq2 identifies 4264 and ALDEx2 identifies 3791 of 5892 transcripts. Applying the rule of at least a $2^{1.4}$ fold change reduces these outputs to 193 for DESeq2 and to 186 for ALDEx2. Applying only the `gamma=1` parameter results in 217 without resorting to the fold change cutoff. We can also see from the volcano plot that there is a substantially better relationship between fold change and the adjusted p-value.

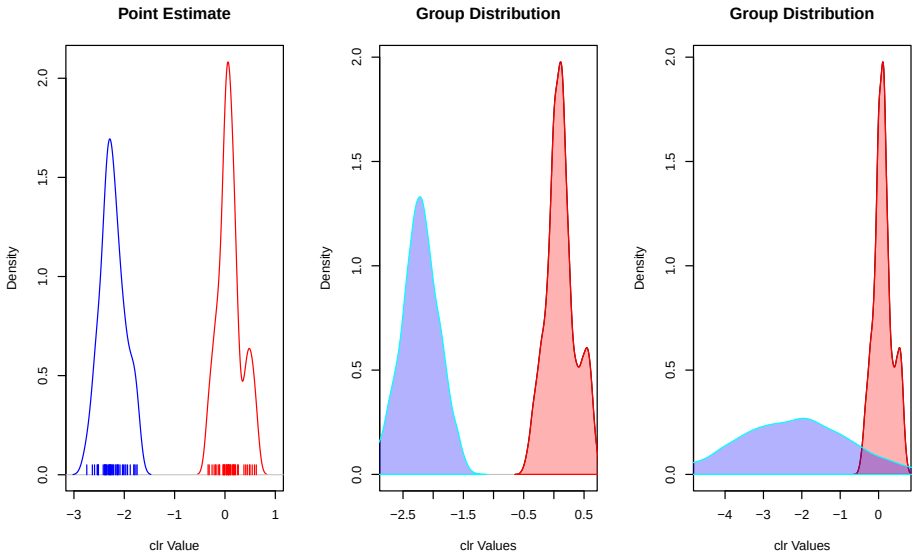
Scale ALDEx2



How is scale altering the dispersion? Figure 2 shows that the density of the clr values for the initial count data, for the unscaled posterior distribution and for the scaled posterior distribution for the gene 'NTS1-2'. In this example the distribution of expression values for the wild-type group is in blue and for the snf-2 knockout in red.

We can see that the major effect of generating a posterior distribution is to broaden the tails of what is possible to observe for both categories. Adding in scale uncertainty produces an even broader distribution for the posterior for both distributions with the greatest effect, in this gene, found in the wild-type distribution. Broadening the distribution increases the pooled dispersion and so reduces the likelihood of the false-positive identifications seen in the unscaled analysis.

```
yst.clr <- apply(yst+0.5, 2, function(x) log2(x) - mean(log2(x)))

par(mfrow=c(1,3))
plot(density(yst.clr['NTS1-2',1:44]), col='red', main='Point Estimate',
  xlab='clr Value', xlim=c(-3,1))
points(density(yst.clr['NTS1-2',45:86]), col='blue', type='l')
segments(yst.clr['NTS1-2',1:44], 0.05, yst.clr['NTS1-2',1:44], 0, col='red')
segments(yst.clr['NTS1-2',45:86], 0.05, yst.clr['NTS1-2',45:86], 0, col='blue')


  aldex.plotFeature(x, 'NTS1-2', pooledOnly=T, densityOnly=T)

    aldex.plotFeature(x.s, 'NTS1-2', pooledOnly=T, densityOnly=T)
```



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 expect that the total bacterial load 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 (Macklaim et al. 2013; Deng et al. 2018; Fettweis et al. 2019) because a pooled analysis falsely identifies many housekeeping genes as being differentially abundant between groups.

Figure 3A shows an effect plot of the data where reads are grouped into parts 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, and also because there is a greater diversity of organisms in BV than in H samples.

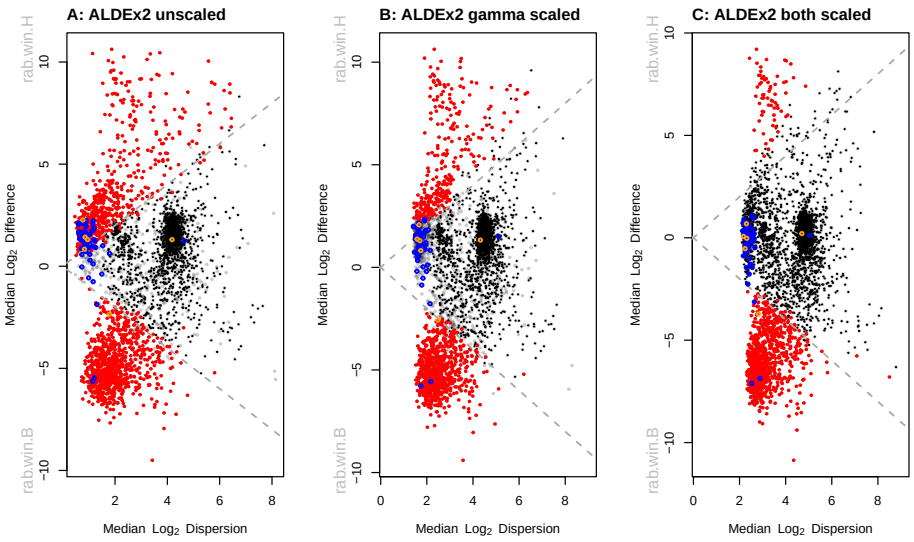
We can also see that there are a large number of functions that are shared between the two groups, this largely corresponds to core metabolic functions that would not be expected not contribute to differences in pathogenicity. Overplotting the core translation (blue) or glycolysis functions (orange) shows anomalies in both their their location and scale in Figure 3A. The major group of these 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 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.

```
e.min <- read.table('~/Documents/0_git/Log-Ratio-Publication/data/twntyfr.txt',
  header=T, row.names=1, check.names=F, sep="\t", comment.char="", quote="")
```

Scale ALDEx2

```
ribo <- c(grep("LSU", rownames(e.min)), grep("SSU",
   rownames(e.min)))
glycol <- c(2418,1392,1305,1306,2421,1049)
sparse.set <- names(which(apply(e.min, 1, min) == 0))

conds <-c("H","H","H","H","B","H","B","B","H","B","H","B","B","B","B",
  "B","B","B","H","B","H","H")

xt <- aldex.clr(e.min, conds)
xt.e <- aldex.effect(xt)
xt.t <- aldex.ttest(xt)
xt.all <- cbind(xt.e, xt.t)

## single gamma model
xg <- aldex.clr(e.min, conds, gamma=1)
xg.e <- aldex.effect(xg)
xg.t <- aldex.ttest(xg)
xg.all <- cbind(xg.e, xg.t)
## new scale model

mu.vec <- gsub('H', log2(1), conds)
mu.vec <- as.numeric(gsub('B', log2(1.05), mu.vec))
mu.mod <- sapply(mu.vec, FUN = function(mu) rlnorm(128, mu, 1))
xt.m <- aldex.clr(e.min, conds, gamma=t(mu.mod))
xt.m.e <- aldex.effect(xt.m)
#hist(xt.m.e$diff.btw, breaks=99)
#plot(density(xt.m.e$diff.btw))
#abline(v=0, lty=2, lwd=2, col='red')
xt.m.t <- aldex.ttest(xt.m)
xt.m.all <- cbind(xt.m.e, xt.m.t)

par(mfrow=c(1,3))
aldex.plot(xt.all)
title('A: ALDEx2 unscaled', adj=0, line= 0.8)
points(xt.all$diff.win[ribo], xt.all$diff.btw[ribo], col='blue',
  cex=0.6, lwd=1.5)
points(xt.all$diff.win[glycol], xt.all$diff.btw[glycol], col='orange',
  cex=0.6, lwd=1.5)

aldex.plot(xg.all, xlim=c(0.3,9))
title('B: ALDEx2 gamma scaled', adj=0, line= 0.8)
points(xg.all$diff.win[ribo], xg.all$diff.btw[ribo], col='blue', cex=0.6,
  lwd=1.5)
points(xg.all$diff.win[glycol], xg.all$diff.btw[glycol], col='orange',
  cex=0.6, lwd=1.5)


aldex.plot(xt.m.all, xlim=c(0.3,9))
title('C: ALDEx2 both scaled', adj=0, line= 0.8)
points(xt.m.all$diff.win[ribo], xt.m.all$diff.btw[ribo], col='blue',
```

```
  cex=0.6, lwd=1.5)
points(xt.m.all$diff.win[glycol], xt.m.all$diff.btw[glycol], col='orange',
  cex=0.6, lwd=1.5)
```



Simply adding in scale uncertainty with `gamma=1` as before has the expected effect of increasing the dispersion as seen in Figure 3B. However, there is little effect on the location of the housekeeping functions and many of those in common between the two groups are still identified as differentially abundant; a Type I error. Furthermore, there are many functions that are in the BV group that are not identified because of this location shift; a Type II error.

Figure 3C shows the effect of adding in differential scale to the two groups to account for the fact that there is a large difference in underlying abundance in the environment. Here we keep the `gamma` parameter near 1, but shift the location of the parameter in the BV samples. This has the effect of moving the location of the common housekeeping genes to the midline of no difference between groups, and provides us with a fair comparison between these two highly disparate microbiomes.

Discussion Scale is important and overlooked - solves some issues with HTS that are ignored Scale robustness should be an important part of data analysis Sca

4 Methods or supplement

We can decompose the underlying environment data into two parts composed of the relative amount and the total amount. We will denote a sequence count dataset as an $D \times N$ matrix Y, with elements Y_{dn} denoting the number of sequenced DNA molecules mapping to the d^{th} entity (e.g., taxa or gene) in the n^{th} sample. Y_{dn} is an estimate of the relative part in the underlying environment.

A key element of scale reliant inference (SRI) is that the observed data is an imperfect measurement of the underlying biological system which we denote W and call a *scaled system*. Like Y, W is a $D \times N$ matrix. Unlike Y, the elements W_{dn} denote the true (as opposed to measured) amount of entity d in the biological system from which the n-th sample was taken.

The counts W have two parts; first scale (i.e., total amounts) and second composition (i.e., relative amounts). We denote the scale of the system as the N-vector $W^{\perp}$ with elements $W_n^{\perp} = \sum_{d=1}^{D} W_{dn}$. We denote the composition of the system as a $D \times N$ matrix $W^{\parallel}$ with elements $W_{dn}^{\parallel} = W_{dn}/W_n^{\perp}$. It follows that the composition and scale of the system uniquely determine the system via: $W_{dn} = W_{dn}^{\parallel} W_n^{\perp}$. Throughout this paper, we use hat notation to denote an estimate of a quantity (e.g., $\hat{W}$ is an estimate of W).

We can let W be the data matrix or table that describes the system we want to measure. This systems contains N samples and D parts where each part is a gene, species, etc and we denote the d-th part in the n-th sample as W_{dn}. The actual data is composed of both proportional (relative) and scale (size) information, such that $W = W^{\parallel}$ and $W^{\perp}$. 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.

We can directly estimate $W^{\parallel}$ by taking a random sample and determining the counts of the parts in the sample. The properties of this measure Y should be well known, and the accuracy with which Y is a good estimate of $W^{\parallel}$ is determined by the number of samples.

- Y is a point estimate of the data and under-samples the true underlying distribution because of sparse random sampling.
- $W^{\parallel}$ is a more accurate estimate of the sampling uncertainty derived from Dirichlet Monte-Carlo sampling to 'fill-in' and provide a posterior distribution of plausible values based on the point estimates
- W with a given amount of scale uncertainty can be estimated by adding scale uncertainty $W^{\perp}$ with $\gamma = 1$. 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.

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

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.

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.

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.

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.

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.

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.

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.

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.