Supplementary figures and code

Greg Gloor

16 April, 2021

Contents

About this document

This document is an .Rmd document and can be found at:

github.com/ggloor/effect/effect_supplement.Rmd

The document is requires Rmarkdown and an installation of LaTeXto work properly. It contains interspersed markdown and R code that may be compiled into a pdf document and supports the figures and assertions in the main article. R code is not exposed in the pdf document but is either referred to by `R-code.r`, or by `R-block-n`. Both are are annotated R code chunks. The former are in the supplementary `code` directory, and the latter are interspersed in the document. Both are provided so that the interested reader can work through as need or interest arises.

R packages required

We will need the following R packages and add-ons (`R-block-1`). The code chunk `code/setup.r` processes the input and output files for analysis, and contains pointers to the files used to create the analysis data that follows.

1. knitr (CRAN)
2. Rmarkdown (CRAN)
3. ALDEx2 (Bioconductor)
4. CoDaSeq (github.com/ggloor/CoDaSeq)
5. distEffect (github.com/ggloor/distEffect)

6. source('code/setup.r')

Datasets

The transcriptome dataset in `data/barton_agg.tsv` is a 48-replicate, two-condition RNA-seq that was done to compare the transcriptome of the BY4741 strain (wild-type) of *Saccharomyces cerevisiae* to that of a SNF2 knockout mutant strain from the same genetic background (Gierliński *et al.*, 2015). The 96 samples were prepared in four batches of 24 samples. RNA was extracted from each of the biological replicates and enriched for polyadenylated RNA. ERCC external RNA spike-in mix was added to each sample (Jiang *et al.*, 2011). The RNA in each sample was fragmented and reverse transcribed to cDNA. The cDNA corresponding to each biological replicate was sequenced on seven lanes of an Illumina HiSeq 2000, thus giving seven technical replicates for each biological replicate. The sequencing data was downloaded from the European Nucleotide Archive under the project ID PRJEB5348. Each of the 672 fastq files, corresponding to one of the technical replicates was aligned to the complete and annotated yeast genome (NCBI project accession: PRJNA128), that had been modified to include the sequences for the 96 ERCC external RNA spike-ins (NIST SRM number: 2374), using bowtie2 v2.1.0 (Langmead and Salzberg, 2012)] with the default options. For each of the 672 replicates, the counts of sequencing reads mapped to each gene and to each of the ERCC external RNA spike-in sequences was determined using HTSeq v0.6.1 (Anders *et al.*, 2015). All reads with an alignment quality less than 0 were omitted. For each biological replicate, the counts for its technical replicates were aggregated by summing. Since the experiment done by (Gierliński *et al.*, 2015) enriched for polyadenylated RNA, counts for sequences that had been mapped to genes annotated as rRNA were removed as they were assumed to only contribute noise to the data.

The 16S rRNA gene sequencing dataset were obtained from the supplementary material of (Bian *et al.*, 2017) located at: https://doi.org/10.6084/m9.figshare.4535660. The two groups were extracted from the entire dataset using the extraction code in `data/effect_reproducibility_16S.R`. Samples were filtered to remove OTUs that were not observed in any sample and the count table saved in `data/tiyaini_pup_vs_ys.Rdata` for use.

For each dataset, a reference set of significant features for ALDEx2 was generated by performing and collecting 100 replicates of the entire dataset comparison using the code in `code\effect_reproducibility.R` and `code\effect_reproducibility_16S.R`, a parallel analysis of the transcriptome dataset with `edgeR` (Anders *et al.*, 2013) was conducted using `code\effect_reproducibility_edgeR.r`. Once the reference set was collected, we conducted 100 replicates of balanced sample size comparisons for each sample size between 2 and 40 samples for the transcriptome data and between 2 and 100 samples for the 16S rRNA gene sequencing data. We collected all features that were identified as passing the significance cutoffs chosen and saved them for use.

Reproducing the analysis

From an R command prompt you can compile this document into PDF if you have LaTeXand pandoc installed:

`rmarkdown::render('effect_supplement.Rmd')`, or you can do the same in bash (R -e "rmarkdown::render('effect_supplement.Rmd')") or you can open the file in RStudio and compile in that environment.

Effect size

Cohen's d, and similar statistics of a standardized mean difference (here-after effect size), use summary statistics to estimate the effect size (Hedges and Olkin, 1985; Cohen, 2013). The standard equation is in equation 1 and broadly speaking an effect size of this type is the ratio of the difference $diff$ and the dispersion $disp$ of two datasets.

$$z = \frac{\mu_a - \mu_b}{disp} = \frac{diff}{disp} \tag{1}$$

In equation 1 all values are summary statistics from distributions a and b, and the utility of z thus depends upon how well the assumptions of each summary statistic fits the underlying distributions. If z is calculated from the mean and standard deviation of a Normal distribution, then z is the number of z scores that separate the two mean values.

Effect size vs Difference

It is worth pointing out that a p-value is not a measure of magnitude of change (or difference), nor is a p-value a standard measure of effect size. Nevertheless, both p-values and difference are widely used in the high throughput sequencing literature as proxies for effect size. The most obvious example of this is when investigators use a Volcano plot which shows the relationship between p-values and the difference between groups (Cui and Churchill, 2003).

Recall that one common way a p-value is calculated is from the t-statistic which has the general form shown in equation 2. Comparing equation 1 and equation 2 we see that the denominator is different; $disp$ is the denominator when calculating z, and the square root of $disp$ divided by the sample size N is the denominator when calculating t. Thus, p-values are not stable estimates of effect size, but rather are strongly dependent on sample size.

$$t = \frac{\mu_a - \mu_b}{\sqrt{disp/\mathrm{N}}} = \frac{diff}{\sqrt{disp/\mathrm{N}}} \tag{2}$$

The distribution effect size ($\mathbb{E}$)

We now introduce and demonstrate the properties of the $\mathbb{E}$. The metric was first developed and used as a convenience for meta-RNAseq (Fernandes *et al.*, 2013; Macklaim *et al.*, 2013) and later for microbiome analyses (eg. Fernandes *et al.*, 2014; Bian *et al.*, 2017), however, these publications did not investigate its properties. The purpose of $\mathbb{E}$ is to determine the standardized difference between two distributions, rather than the standardized difference between the means (or midpoints) of the distributions. Let us briefly explain the difference.

The approach taken by $\mathbb{E}$ is to calculate the median of the standardized difference of the distributions, $\vec{eff}$. We will use vector format and start with our two distributions represented as $\vec{a}$ and $\vec{b}$. The $\mathbb{E}$ metric is calculated as in equation 6 from the outputs of the prior equations 3 and 4. Note that both the difference and dispersion estimates are vectors and not point estimates.

$$\vec{diff} = \vec{a} - \vec{b} \tag{3}$$

$$\vec{disp} = max\{|\vec{a} - \boldsymbol{\rho}\vec{a}|, |\vec{b} - \boldsymbol{\rho}\vec{b}|\} \tag{4}$$

$$\vec{eff} = \frac{\vec{diff}}{\vec{disp}} \tag{5}$$

$$\mathbb{E} = \mathrm{med}(\vec{eff}) \tag{6}$$

In equations 3-5, we use max to refer to the maximum value at each position in the vector, med to refer to the median of the vector, | | to indicated the absolute value of the elements in the enclosed vector, and $\boldsymbol{\rho}$ to denote one or more random permutations of the associated vector.

Note that both the numerator and the denominator in equation 5 are vectors as is the result since we are calculating the ratio of the vector values element-wise. The vectors recycle if one vector is different in length than the other. The numerator, $\vec{diff}$, is simply the signed difference between the distributions in $\vec{a}$ and $\vec{b}$, and the denominator, $\vec{disp}$, is the maximum absolute difference, a novel estimate of the pooled dispersion of the distributions. The $\vec{disp}$ metric is necessary since there is no vector-wise dispersion estimate in common use.

Properties of the median of $\vec{diff}$, MSD

The midpoints of both $\vec{diff}$ and $\vec{disp}$ have meaning, and are calculated by ALDEx2. The median of $\vec{diff}$ is the median signed difference (MSD, `diff.btw` in ALDEX2) and the median of $\vec{disp}$, or the median of the maximum absolute difference (MMAD), is the dispersion statistic `diff.win` returned by ALDEx2. The code contained in `R-block-2`, and Figure 1 shows the behaviour of the MSD relative to the difference of means for three distributions as a measure of location. We can see that the two methods are essentially equivalent except in the case of a Cauchy distribution, where the difference in means clearly fails to provide an reliable estimate of location. Thus, the MSD is an efficient and safe choice to determine location regardless of the underlying distribution.

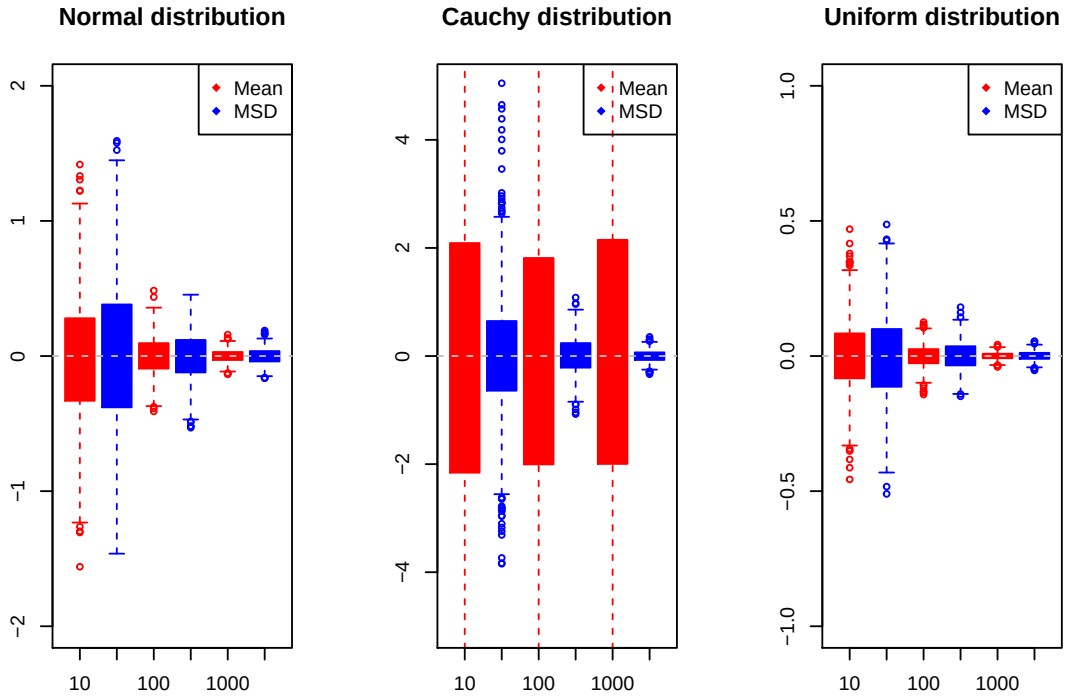

Figure 1: Boxplots of residual values of the difference in means, or the MSD for 1000 trials between two random distributions with a known difference of 1. The number of samples in the distributions was 10, 100 or 1000. A perfect estimator of location would have a residual of 0 without any variation. The difference in means (Mean, red) and the median of the difference vector (MSD) return very similar results except in the case of a Cauchy distribution, where the MSD is clearly preferable. The y-axis of the Cauchy distribution plot has been truncated since the limit of the difference in mean values is very large.

Properties of the median of $\vec{disp}$ (MMAD)

The MMAD is a somewhat robust, and efficient estimator of scale. `R-block-3` shows the code used to support this, but is not run for efficiency since the number of random values needed to estimate with precision is extreme.

We estimate that MMAD is 1.418 ± 0.0001 that of the standard deviation for a normal distribution. The efficiency of the MMAD for estimating dispersion in a Normal distribution is 52%, which compares favourably with that for the median absolute deviation (MAD) of 37%. Thus, the MMAD requires, at most, double the sample size to determine dispersion at the same precision as the standard deviation. Finally, Figure 2 and `R-block-4` show the breakdown point of the MMAD for a Normal distribution. We can see that the expected value of the difference between two distributions is unchanged until 20% of the observations in one sample are changed. This compares poorly with the MAD which has the maximum breakdown point of 50%, but favourably with 0% breakdown point of the standard deviation. We conclude that the MMAD and the MSD, while not perfect, are reasonable measures of dispersion and difference for a broad array of distributions.

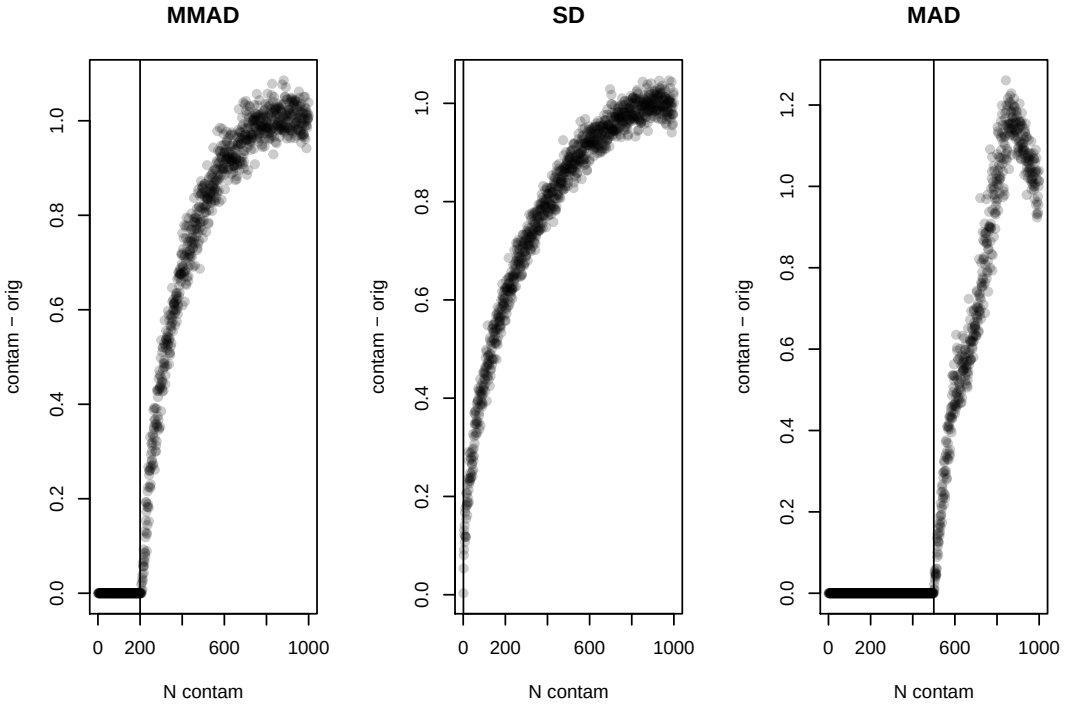

Figure 2: The breakdown point is the percent contamination that can be introduced into a distribution without changing the measure of scale. The breakdown point of the MMAD is about 0.2, while the breakdown point of the standard deviation is 0. The vertical lines indicate the breakdown points.

Comparing $\mathbb{E}$ and Cohen's d

Finally, we examine how $\mathbb{E}$ and Cohen's d compare when determining the standardized difference between two distributions. The code is shown in `R-block-5` and Figure 3 shows how the two statistics compare for Normal and Cauchy distributions, with an expected difference of 2 and a scale parameter of 1.

Figure 3 shows that as expected Cohen's d had a standardized effect size that was distributed around the mean value of 2.0 when two Normal distributions were compared. The $\mathbb{E}$ standardized effect size was distributed around a mean value of 1.4, which is expected given

that the denominator of $\mathbb{E}$ is 1.418 times larger than the pooled standard deviation of 1. Applying this correction, the standardized effect size for $\mathbb{E}$ would be 2.0, the same as Cohen's d for a Normal distribution. Thus as instantiated, the $\mathbb{E}$ is less than Cohen's d, but can be easily scaled if needed. Scaling is not performed by default, since there is no guarantee that we are comparing Normal distributions in high throughput sequencing datasets.

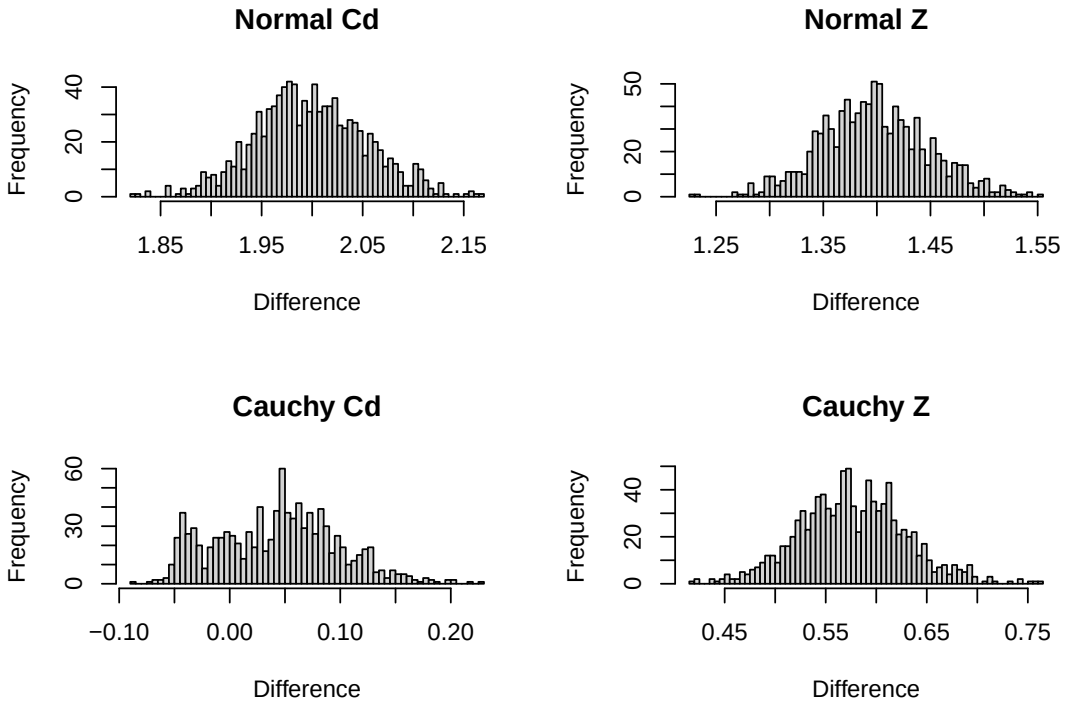

Figure 3: The standardized effect size of $\mathbb{E}$ and Cohen's d (Cd) are compared for Normal and Cauchy distributions. The histograms plot one thousand random tests of the effect size for distributions of size 1000. The difference between the distributions was set at 2 with a scale parameter of 1 in each case.

The situation is very different when using the standardized effect size to compare two Cauchy distributions. Here the mean effect size for Cohen's d is 0.044, or almost no effect. Figure 3 shows that this distribution is bimodal, and the median standardized effect is 0.046. However, the $\mathbb{E}$ metric gives a mean standardized effect size of 0.58 and is symmetrically distributed over a comparatively narrow range. Recall that Cauchy distributions have very broad tails, and so a scale of 1 with a Cauchy distribution will result in a considerably broader distribution than will a Normal distribution with the same scale, and consequently the standardized effect size should be smaller than observed with a Normal distribution with the same location and scale parameters.

The yeast transcriptome dataset

The main manuscript shows the results for a well-described transcriptome dataset from a SNF1 *Saccharomyces cerevisae* gene knockout with a cutoff of $\mathcal{Z} > 1$ or 2. Later we expand on this analysis and show a similar analyses from a 16S rRNA gene sequencing experiment in a Chinese population.

p-value based approaches

We used the *Saccharomyces cerevisia* 48-replicate SNF2 gene knockout dataset described in (Schurch *et al.*, 2016). We first used the `CoDaSeq R` package to remove outlier samples using a robust compositionally appropriate process (Peter *et al.*, 2009), which selects those samples that are further from all other samples than expected. The actual approach is to generate the compositionally appropriate Aitchison distance matrix after centered log-ratio transformation (Aitchison, 1986) and to identity the total distance between all samples in each group. Those samples that contribute more than twice the interquartile range to the total distance of the matrix are removed. Using this approach we removed samples 6, *10*, 13, *15*, 25, *31*, and 35 from the SNF2 group, leaving 41 "clean" samples, and we removed samples 21,25,28,34 and 36 from the WT group leaving 43 samples (samples that are removed by this approach but were not removed in the (Schurch *et al.*, 2016) dataset are in italics). Sample 22 was not removed by this approach from the SNF2 group, but was using the approach in (Gierliński *et al.*, 2015). In general, this approach is slightly more aggressive in removing outliers that was used in outliers than the one used in (Gierliński *et al.*, 2015; Schurch *et al.*, 2016). Discrepencies between the methods will largely occur because of the compositionally-appropriate method adopted here (Aitchison, 1986). In the dataset used here there were 6236 genes rather than the larger number in the original report; the number of features is not expected to alter the conclusions. The code is for this step is in `code/setup.r`.

```
## Warning in kable_pipe(x = structure(c("Tool", "edgeR", "edgeR", "edgeR", : The
## table should have a header (column names)
```

Table 1: TP set by method.

Tool	Method	Total genes	Significant genes
edgeR	glm	6349	4705
edgeR	et	6349	4684
edgeR	et & T>2	6349	101
ALDEx2	Wilcox	6236	4318
ALDEx2	Welch's	6236	4352
ALDEx2	$\mathcal{Z} > 1$	6236	2020
ALDEx2	$\mathcal{Z} > 2$	6236	545

Following the original report, we used the set of genes identified by each tool with the full set of samples as the gold standard set. Table 1 shows the number of significant genes by each method. We can see that the edgeR methods and the ALDEx2 p-value based methods return similar sets of genes at an FDR cutoff of $q < 0.05$. In fact, the different toolsets return almost the same set of genes, with 4060 of the genes being found by the four the p-value based approaches. The two $\mathbb{E}$ cutoffs return a subset of the p-value based core set. Note that filtering by p-value and fold-change is extraordinarily restrictive: only 101 genes pass the p-value filter and a Threshold of > 2 as defined in Gierlinski et al (2015), and indeed, the Threshold of > 2 is all that is needed to identify that set of 101 genes in this dataset.

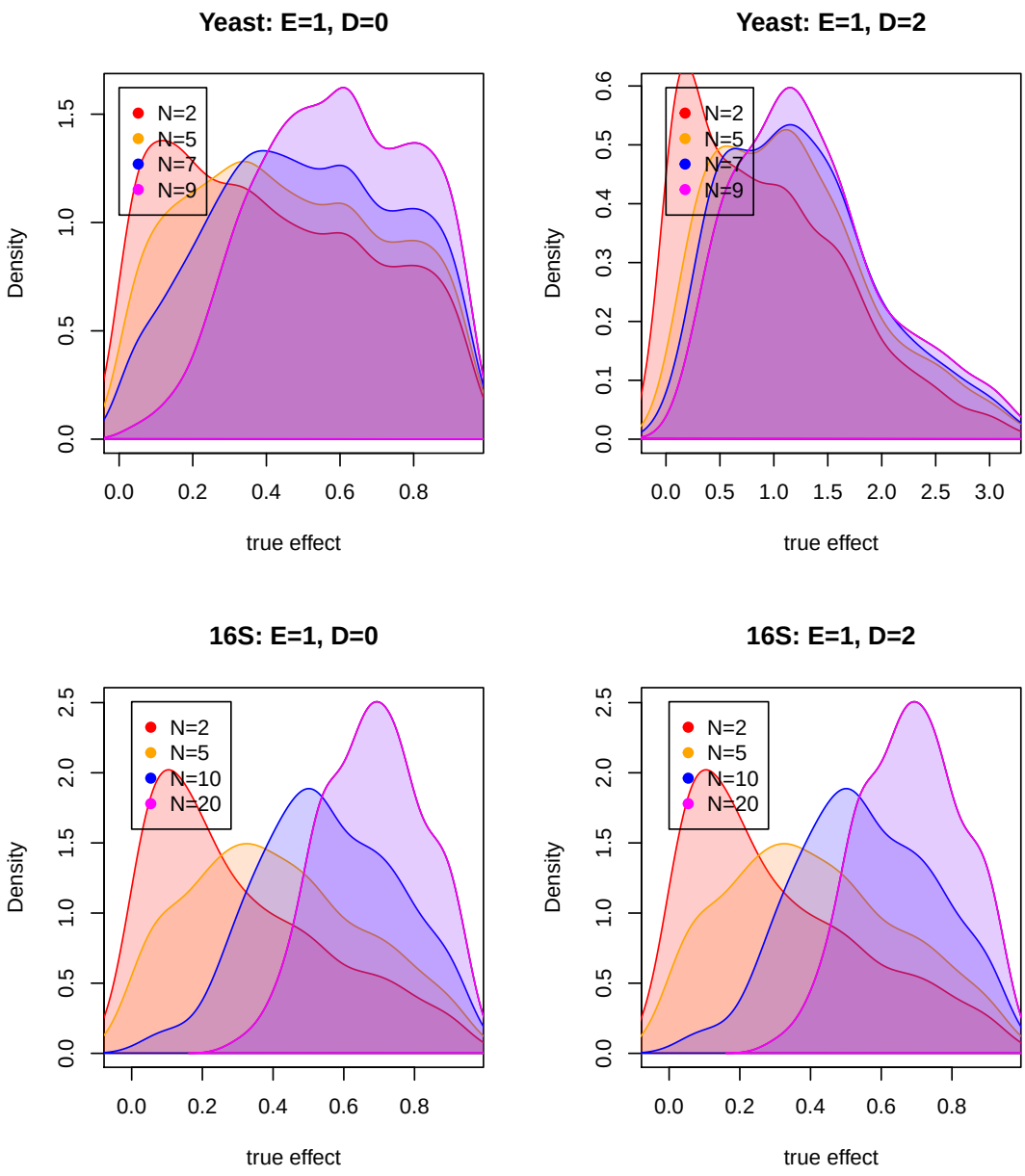

Figure 4: The panels show the effect size distribution of features identified as false positives by $\mathbb{E}$ at four different sample sizes in two different datasets.

We can ask what is the actual effect size in the whole dataset for FP features in the subsets. Figure 4 addresses this question for both datasets and we can see the effect of including the difference cutoff. We can see that if a FP feature is identified in a subset that the feature has a strong likelihood of being actually different if the sample size is large enough. At a per-group sample size of 2, the modal effect size is close to 0, indicating that there is little discriminatory power at this sample size. This is not surprising. However, at a sample size of 5 in each group, the modal effect size of FP features abbou 0.3 in both sample sets, and increases to 0.5 with 10 samples. Importantly, at 10 samples per group the FP features all have effect sizes that are larger than 0, indicating that there is a diffenrece in the full dataset, but the difference is smaller than expected. This is consistent with the findings of Halsey et al (Halsey *et al.*, 2015). Thus, any FP features identified at a sample size greater than 10 are likely substantially different between the two groups, although not as different as the cutoff would suggest.

Note however that in the small variance transcriptome dataset that a substantial fraction

of the FP features have effect sizes larger than the cutoff. This can be ascribed to the FP features having larger differences in the subsets than in the real datasets.

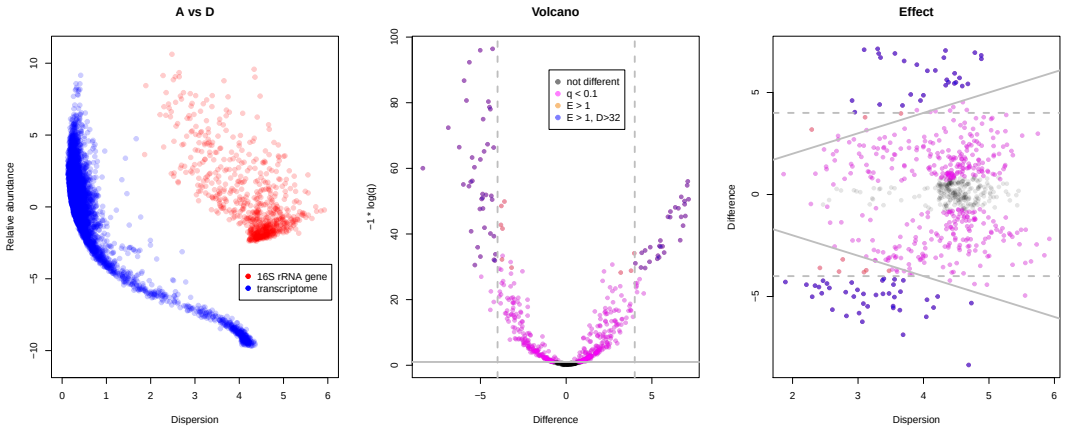

Figure 5: Abundance vs. Dispersion, Volcano and Effect plots. The A vs D plot shows the relationship between the dispersion in the dataset vs. the relative abundance for the yeast transcriptome dataset in the main text, and the 16S rRNA gene sequencing dataset. Dispersion is the MMAD for each feature, and the Relative abundance is the median of the clr-transformed Dir Monte-Carlo instances (rab.all and diff.win from ALDEx2). The Volcano and Effect plots show features identified by Ed , adjusted p-values and absolute differences when detecting differential features between two groups for the 16S rRNA gene sequencing dataset. These plots compare the features identified by Ed and by q-scores and a 16-fold fold-change thresholds in the full dataset. This large absolute fold change is required as the within-group dispersion is enormous in this and other 16S rRNA gene sequencing datasets. In these plots all features that have a q score less than 0.1 also have an effect size greater than 1. Thus, the features in magenta are only identified as significantly different by q scores, those in orange are significantly different by both their q score and their effect size, and features in blue are significant by their q score, their effect size and their absolute difference. The dashed grey lines in the two plots demarcate the 16-fold difference location; note that the difference is in a log2 scale. The horizontal solid line in the Volcano plot indicates a FDR (q score) of 0.1. The diagonal solid lines in the Effect plot indicate the boundary where the difference equals the dispersion; ie, where the effect size is 1.

References

Aitchison,J. (1986) The statistical analysis of compositional data Chapman & Hall, London, England.

Anders,S. *et al.* (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc*, **8**, 1765–86.

Anders,S. *et al.* (2015) HTSeq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–9.

Bian,G. *et al.* (2017) The gut microbiota of healthy aged chinese is similar to that of the healthy young. *mSphere*, **2**, e00327–17.

Cohen,J. (2013) Statistical power analysis for the behavioral sciences Academic press.

Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, **4**, 210.1–210.10.

Fernandes,A.D. *et al.* (2013) ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-seq. *PLoS One*, **8**, e67019.

Fernandes,A.D. *et al.* (2014) Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, **2**, 15.1–15.13.

Gierliński,M. *et al.* (2015) Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics*, **31**, 3625–3630.

Halsey,L.G. *et al.* (2015) The fickle p value generates irreproducible results. *Nat Methods*, **12**, 179–85.

Hedges,L.V. and Olkin,I. (1985) Statistical methods for meta-analysis Academic Press, Orlando.

Jiang,L. *et al.* (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Res*, **21**, 1543–51.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nature methods*, **9**, 357–359.

Macklaim,J.M. *et al.* (2013) Comparative meta-RNA-seq of the vaginal microbiota and differential expression by lactobacillus iners in health and dysbiosis. *Microbiome*, **1**, 12.

Peter,F. *et al.* (2009) Principal component analysis for compositional data with outliers. *Environmetrics*, **20**, 621–632.

Schurch,N.J. *et al.* (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, **22**, 839–51.