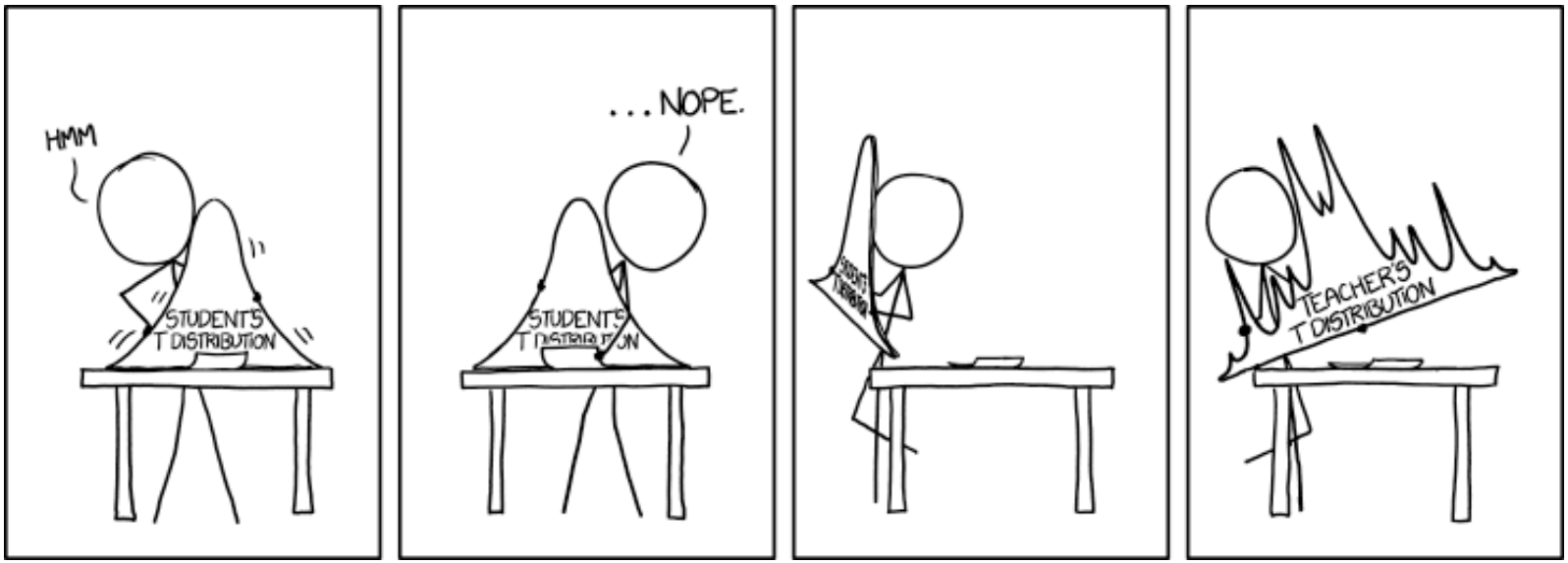
WASTE NOT, WANT NOT

Susan Holmes
http://www-stat.stanford.edu/~susan/

Bio-X and Statistics, Stanford University

August 9, 2015


HMM
STUDENT'S T DISTRIBUTION
...NOPE.
STUDENT'S T DISTRIBUTION
TEACHER'S T DISTRIBUTION

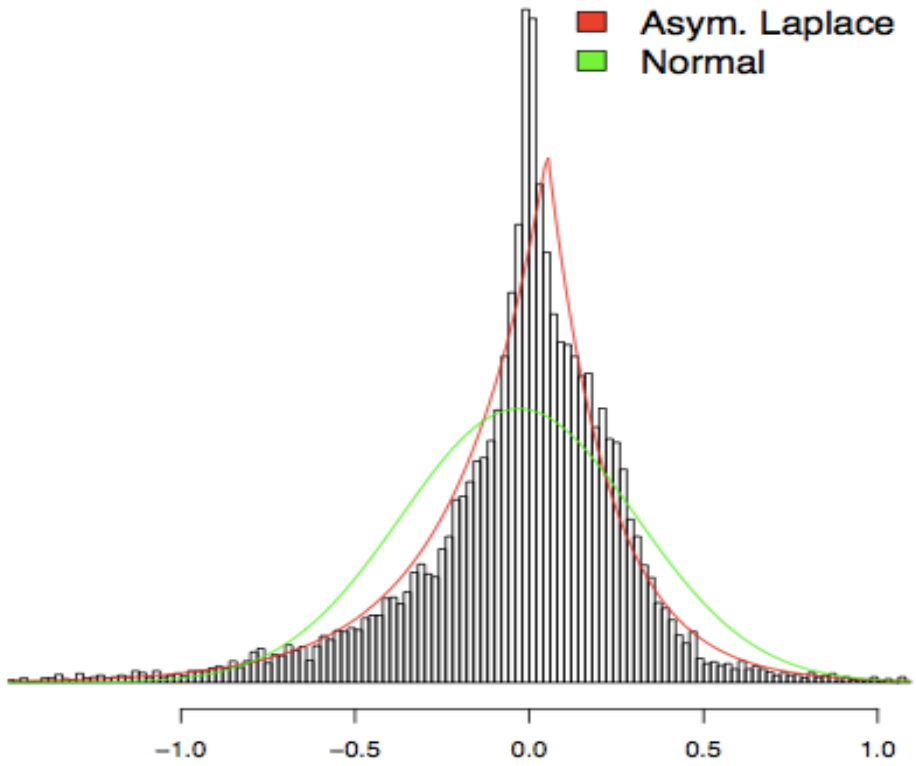
Challenges for those of us working from the ground up

- Heteroscedasticity.
- Information Leaks.

Part I

Heteroscedasticity: Mixtures and how to Normalize them

Statistical Applications in Genetics and Molecular Biology, Vol. 4 [2005], Iss. 1, Art. 16



Some real data (Caporoso et al, 2011)

```
> GlobalPatterns

phyloseq-class experiment-level object
otu_table()  OTU Table:      [ 19216 taxa and 26 samples ]
sample_data()Sample Data:     [ 26 samples by 7 sample varia
tax_table()Taxonomy Table:  [ 19216 taxa by 7 taxonomic ran
phy_tree() Phylogenetic Tree:[ 19216 tips and 19215 interna
```

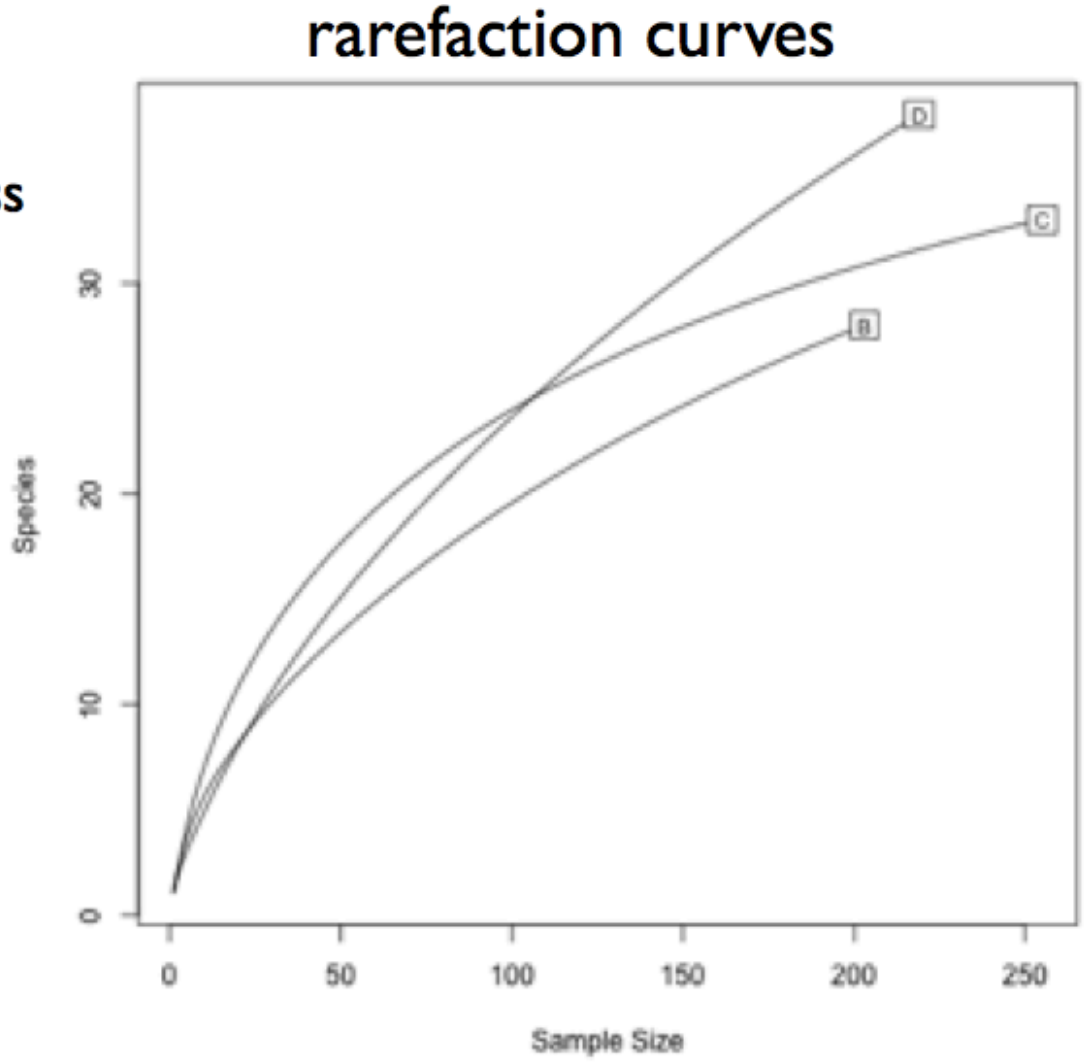
```
otu_table(GlobalPatterns)[45:55,1:10]
OTU Table:          [11 taxa and 10 samples]
                     taxa are rows
        CL3 CC1   SV1 M31Fcsw M11Fcsw M31Plmr M11Plmr F21
573586    0   0     0        0        0        0        0
568724    0   0     0        0        0        0        0
175045    0   0     0        0        1        0        0
552540    0   0     0        0        0        0        0
546313   72 153 11232        0        1        1        0
548602    0   0    16        0        0        0        0
564501    0   0     3        0        0        0        0
47778     1  14   207        0        0        0        5
54107     2  87   746        0        0        0        3
25116     1   4   169        0        0        0        0
71074    93 341 11788        1        0       23       48
```

```
> sample_sums(GlobalPatterns)
    CL3     CC1     SV1 M31Fcsw M11Fcsw M31Plmr M11Plmr F
 864077 1135457  697509 1543451 2076476  718943  433894
 .....
    NP3     NP5 TRRsed1 TRRsed2 TRRsed3    TS28    TS29   I
1478965 1652754   58688  493126  279704  937466 1211071

> summary(sample_sums(GlobalPatterns))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  58690  567100 1107000 1085000 1527000 2357000
```

How to deal with different numbers of reads?

- Sanders 1968
- non-parametric richness
- estimate coverage
- Normalize?



Sanders, H. L. (1968). Marine benthic diversity: a comparative study. *American Naturalist*

Very popular: qiime

Nature Methods **7**, 335–336 (1 May 2010) | doi:10.1038/nmeth.f.303

QIIME allows analysis of high-throughput community sequencing data

J Gregory Caporaso , Justin Kuczynski , Jesse Stombaugh , Kyle Bittinger , Frederic D Bushman , Elizabeth K Costello , Noah Fierer , Antonio Gonzalez Peña , Julia K Goodrich , Jeffrey I Gordon , Gavin A Huttley , Scott T Kelley , Dan Knights , Jeremy E Koenig , Ruth E Ley , Catherine A Lozupone , Daniel McDonald , Brian D Muegge , Meg Pirrung , Jens Reeder , Joel R Sevinsky , Peter J Turnbaugh , William A Walters , Jeremy Widmann , Tanya Yatsunenko , Jesse Zaneveld & Rob Knight

2,300 citations.

Current Method: Rarefying

Ad hoc library size normalization by random subsampling without replacement.

1. Select a minimum library size, $N_{L,min}$. This has also been called the rarefaction level.
2. Discard libraries (microbiome samples) that have fewer reads than $N_{L,min}$.
3. Subsample the remaining libraries without replacement such that they all have size $N_{L,min}$.

Often $N_{L,min}$ is chosen to be equal to the size of the smallest library that is not considered defective, and the process of identifying defective samples comes with a risk of subjectivity and bias. In many cases researchers have also failed to repeat the random subsampling step (3) or record the pseudorandom number generation seed/process --- both of which are essential for reproducibility.

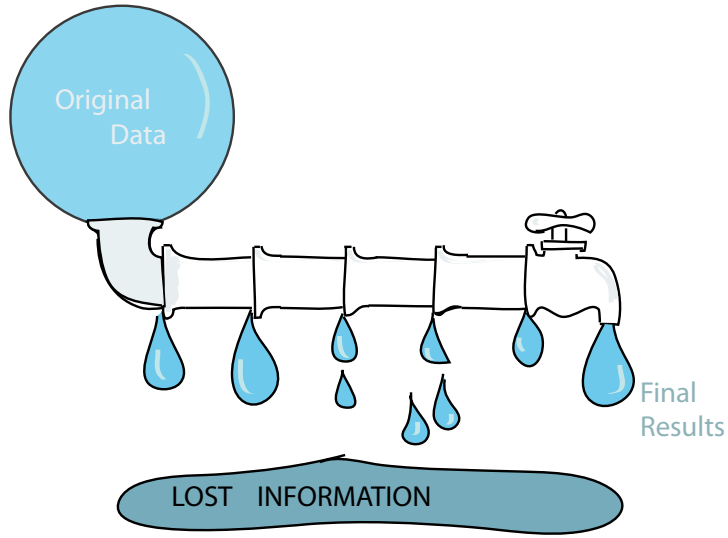
Reduction of Data to Proportions

Many software programs automatically reduce the data to relative proportions, losing the information about library sizes or read counts.
This makes comparisons very difficult.

Statistical Formulation: When making a (testing) decision, reducing results from a Binomial distribution into a proportion does not give an **admissible** procedure.
Definition:An admissible rule is an optimal rule for making a decision in the sense that there is no other rule that is always better than it.

How to compress the data?

.



...without losing too much information?

The proportion is not a sufficient statistic for the Binomial.

 A statistic T(X) is called sufficient for θ if it contains all the information in X about θ.
Standard statistical viewpoint:
The joint probability distribution of the data conditional on the value of a sufficient statistic for a parameter, does not depend on that parameter: $P_\theta(X|T(X) = T)$ does not depend on θ. Wiki

Equivalent Definitions

Mutual Information:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} = K(P(x,y), P(x)P(y))$$

A function of the data $T(X)$ is a sufficient statistic for the distribution if

$$I(\theta, X) = I(\theta, T(X))$$

for all distributions on θ.

Note:

For a Bayesian, no matter what prior one uses, one only has to consider the sufficient statistic for making inference, because the posterior distribution given $T = T(x)$ is the same as the posterior given the data $X = x$.

Aim of the studies: Differential Abundance

Like differentially expressed genes, a species/OTU is considered differentially abundant if its mean proportion is significantly different between two or more sample classes in the experimental design.
Optimality Criteria:

Sensititivity or Power True Positive Rate.

Specificity True Negative Rate.

We have to correct for many sources of error (blocking, modeling, control, etc..)

Rarefaction and Reduction to Proportions are Inadmissible

The following is a minimal example to explain why rarefying is statistically inadmissible, especially with regards to variance stabilization.
Suppose we want to compare two different samples, called A and B, comprised of 100 and 1000 reads, respectively. In these hypothetical communities only two types of microbes have been observed, OTU1 and OTU2

According to Table 1, Left.

Table: **A minimal example of the effect of rarefying on power.**

Original Abundance		
	A	B
OTU1	62	500
OTU2	38	500
Total	100	1000

Rarefied Abundance		
	A	B
OTU1	62	50
OTU2	38	50
	100	100

Standard Tests for Difference			
P-value	χ^2	Prop	Fisher
Original	0.0290	0.0290	0.0272
Rarefied	0.1171	0.1171	0.1169

Hypothetical abundance data in its original (Top-Left) and rarefied (Top-Right) form, with corresponding formal test results for differentiation (Bottom).

Formally comparing the two proportions according to a standard test is done either using a χ^2 test (equivalent to a two sample proportion test here) or a Fisher exact test. This requires knowledge of the number of trials.
By rarefying (Table 1, top-right) so that both samples have the same number of counts, we are no longer able to differentiate between them.
This loss of power is completely attributable to reducing the size of B by a factor of 10, which also increases the confidence intervals corresponding to each proportion such that they are no longer distinguishable from those in A, even though they are distinguishable in the original data.
The variance of the proportion's estimate $\hat{p}$ is multiplied by 10 when the total count is divided by 10.

Equalization of variances

In this binomial example the variance of the proportion estimate is $\mathrm{Var}(\frac{X}{n}) = \frac{pq}{n} = \frac{q}{n}E(\frac{X}{n})$, a function of the mean.
This is a common occurrence and one that is traditionally dealt with in statistics by applying variance-stabilizing transformations.
However, in order to find the right transformation, we need a good model for the error.

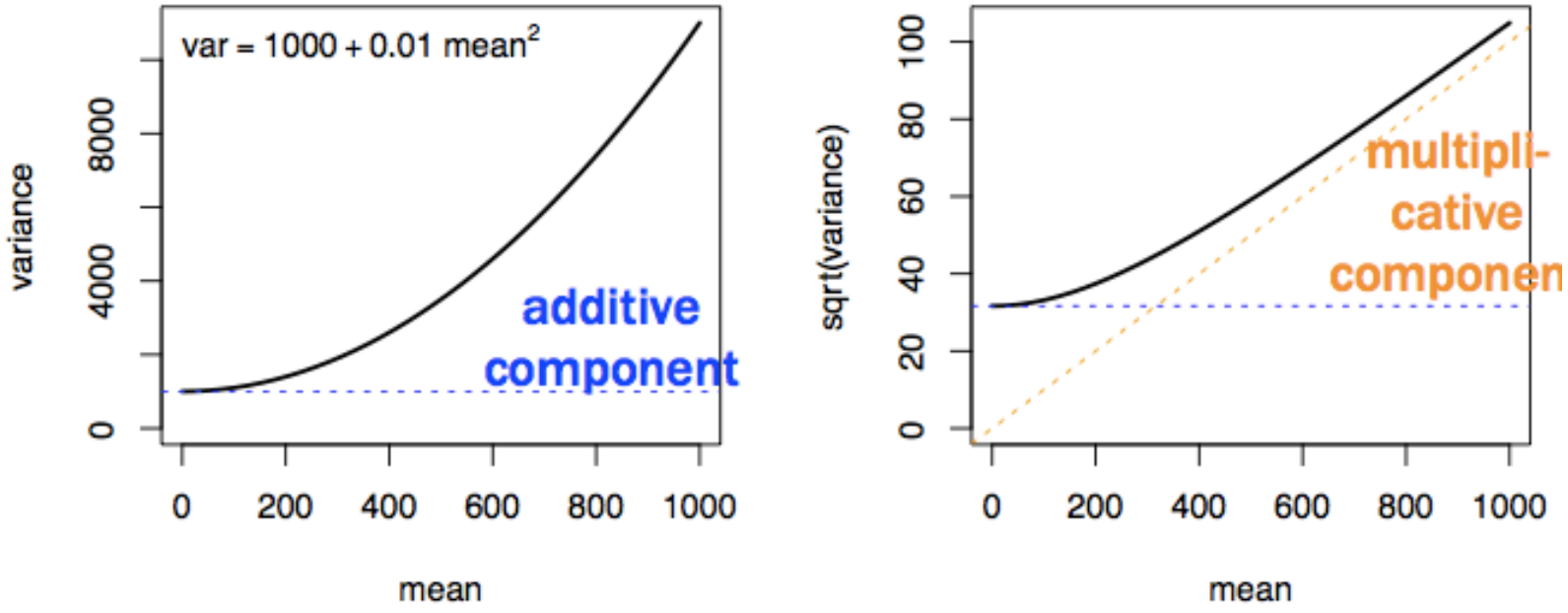
Variance Stabilization

Prefer to deal with errors across samples which are independent and identically distributed.
In particular homoscedasticity (equal variances) across all the noise levels.
This is not the case when we have unequal sample sizes and variations in the accuracy across instruments.
A standard way of dealing with heteroscedastic noise is to try to decompose the sources of heterogeneity and apply transformations that make the noise variance almost constant.
These are called variance stabilizing transformations.

Take for instance different Poisson variables with mean μ_i.
Their variances are all different if the μ_i are different.
However, if the square root transformation is applied to each of the variables, then the transformed variables will have approximately constant variance.
Actually if we take the transformation $x \longrightarrow 2\sqrt{x}$ we obtain a variance approximately equal to 1..

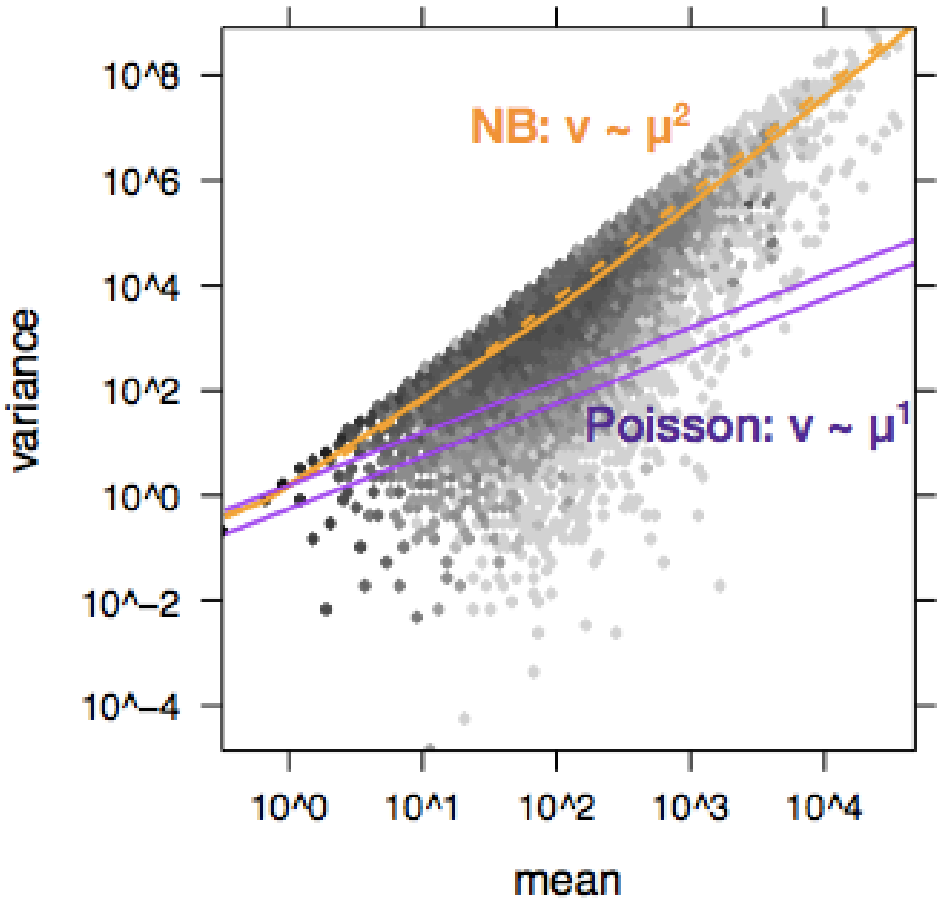
$$\text{var} = \mu + c\mu^2$$

The additive-multiplicative error model



Trey Ideker et al.: JCB (2000)
David Rocke and Blythe Durbin: JCB (2001), Bioinformatics (2002)
For robust affine regression normalisation: W. Huber et al. Bioinformatics (2002)
For background correction in RMA: R. Irizarry et al. Biostatistics (2003)

Two component error models



Microarrays

$\mathrm{var}(\mu) = b + c \cdot \mu^2$

b: background

c: asymptotic coefficient of variation

Sequencing counts

early edgeR:

$\mathrm{var}(\mu) = \mu + \alpha \cdot \mu^2$

μ: from Poisson

α: dispersion

DESeq

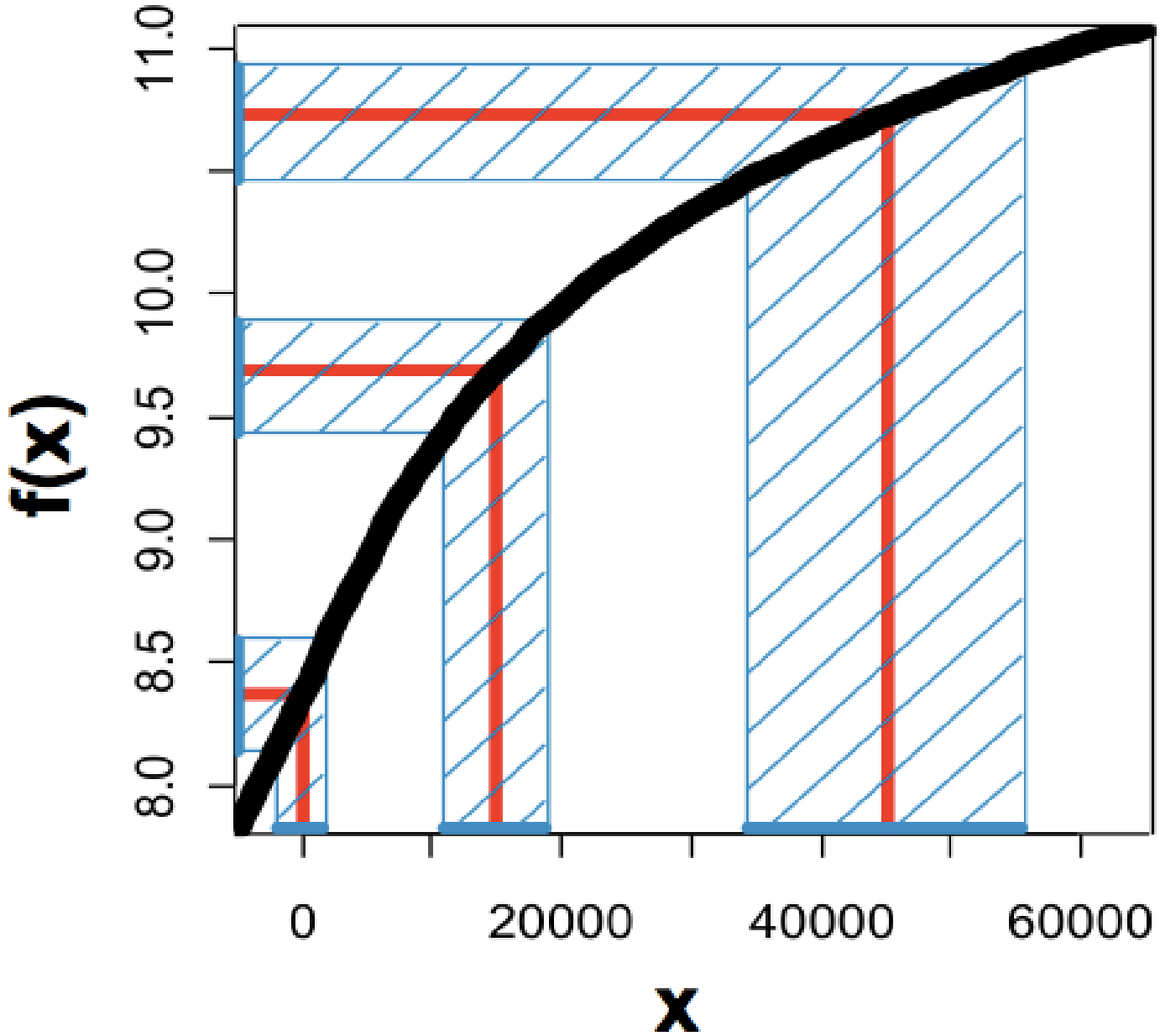
$\mathrm{var}(\mu) = \mu + \alpha(\mu) \cdot \mu^2$

DESeq parametric option

$\alpha(\mu) = a_1/\mu + a_0 \quad \Leftrightarrow$

$\mathrm{var}(\mu) = \mu + a_1 \cdot \mu + a_0 \cdot \mu^2$

variance stabilizing transformation



Modeling read counts

If technical replicates have same number of reads: s_j.
Poisson variation with mean $\mu = s_j u_i$.
Taxa i having an incidence proportion u_i.
Number of reads for the sample j and taxa i would be

$$K_{ij} \sim \text{Poisson}\,(s_j u_i)$$

Negative Binomial with the two parameters: the mean m and $r = \frac{1-p}{p}m$, then the probability is:

$$X \sim NB(m; r)$$

$$\begin{aligned} P(X = k) &= \binom{k+r-1}{k} \left(\frac{r}{r+m}\right)^r \left(\frac{m}{r+m}\right)^k \\ &= \frac{\Gamma(k+r)}{k!\Gamma(r)} \left(\frac{r}{r+m}\right)^r \left(\frac{m}{r+m}\right)^k \end{aligned}$$

The variance is $Var(X) = \frac{m(m+r)}{r} = m + \frac{m^2}{r}$, we will also use $\phi = \frac{1}{r}$ and call this the overdispersion parameter, giving $Var(X) = m + \phi m^2$. When $\phi = 0$ the distribution of X will be Poisson(m). This is the (mean=m,overdispersion=ϕ) parametrization we will use from now on.

Modeling Counts

For biological replicates within the same group -- such as treatment or control groups or the same environments -- the proportions u_i will be variable between samples.
Call the two parameters r_i and $\phi = \frac{p_i}{1-p_i}$.
So that U_{ij} the proportion of taxa i in sample j is distributed according to Gamma$(r_i, \phi = \frac{p_i}{1-p_i})$.
K_{ij} have a Poisson-Gamma mixture of different Poisson variables each with its own parameter generated from the Poisson.
This gives the Negative Binomial with parameters $(m = u_i s_j)$ and ϕ_i as a satisfactory model of the variability.

Different Conditions

Samples belong to different conditions such as treatment and control or different environments.
Estimate the values of the parameters separately for each of the different biological replicate conditions/classes.
Use the index c for the different conditions, we then have the counts for the taxa i and sample j in condition c having a Negative Binomial distribution with $m_c = u_{ic}s_j$ and ϕ_{ic} so that the variance is written

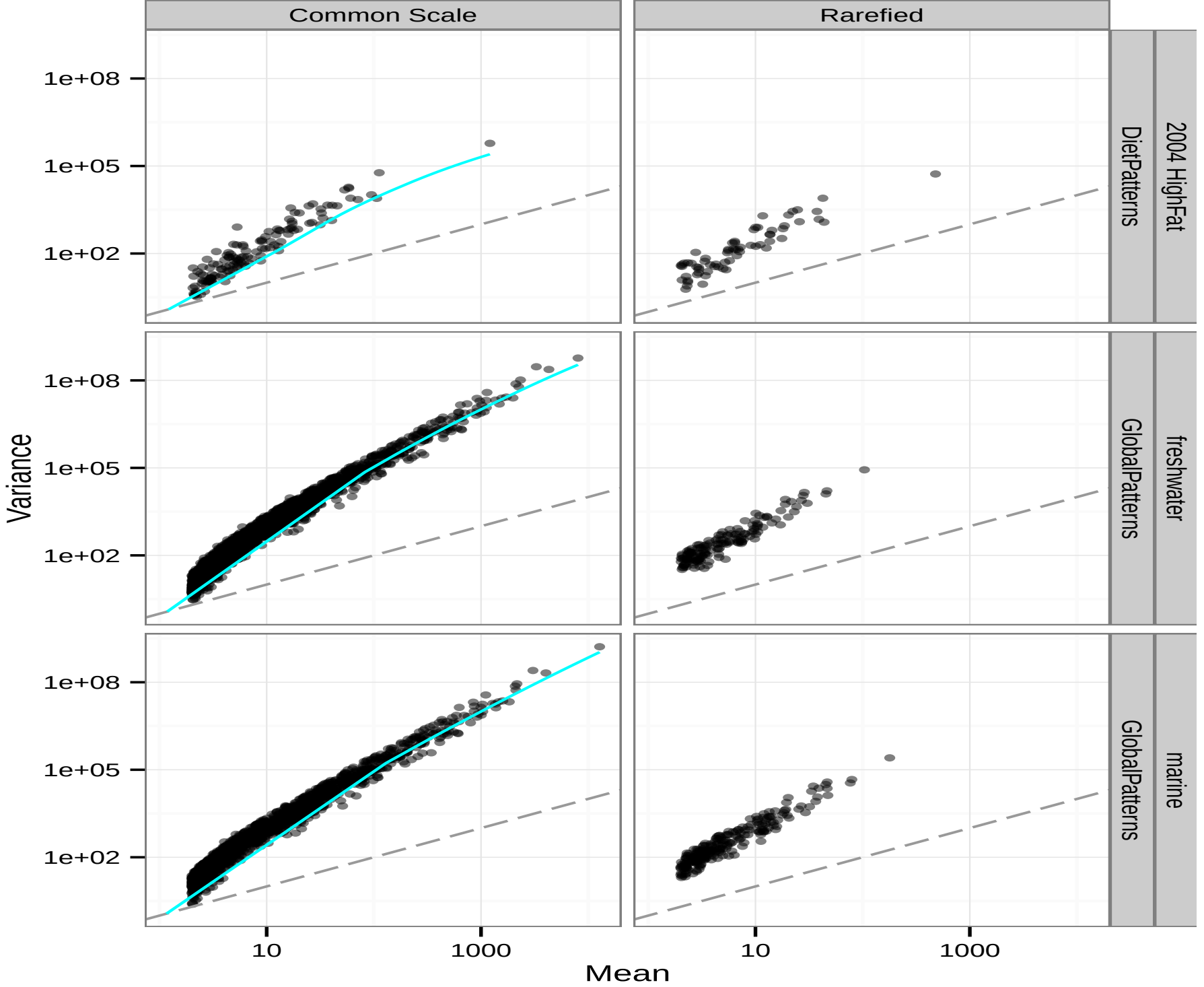
$$u_{ic}s_j + \phi_{ic}s_j^2 u_{ic}^2 \tag{1}$$

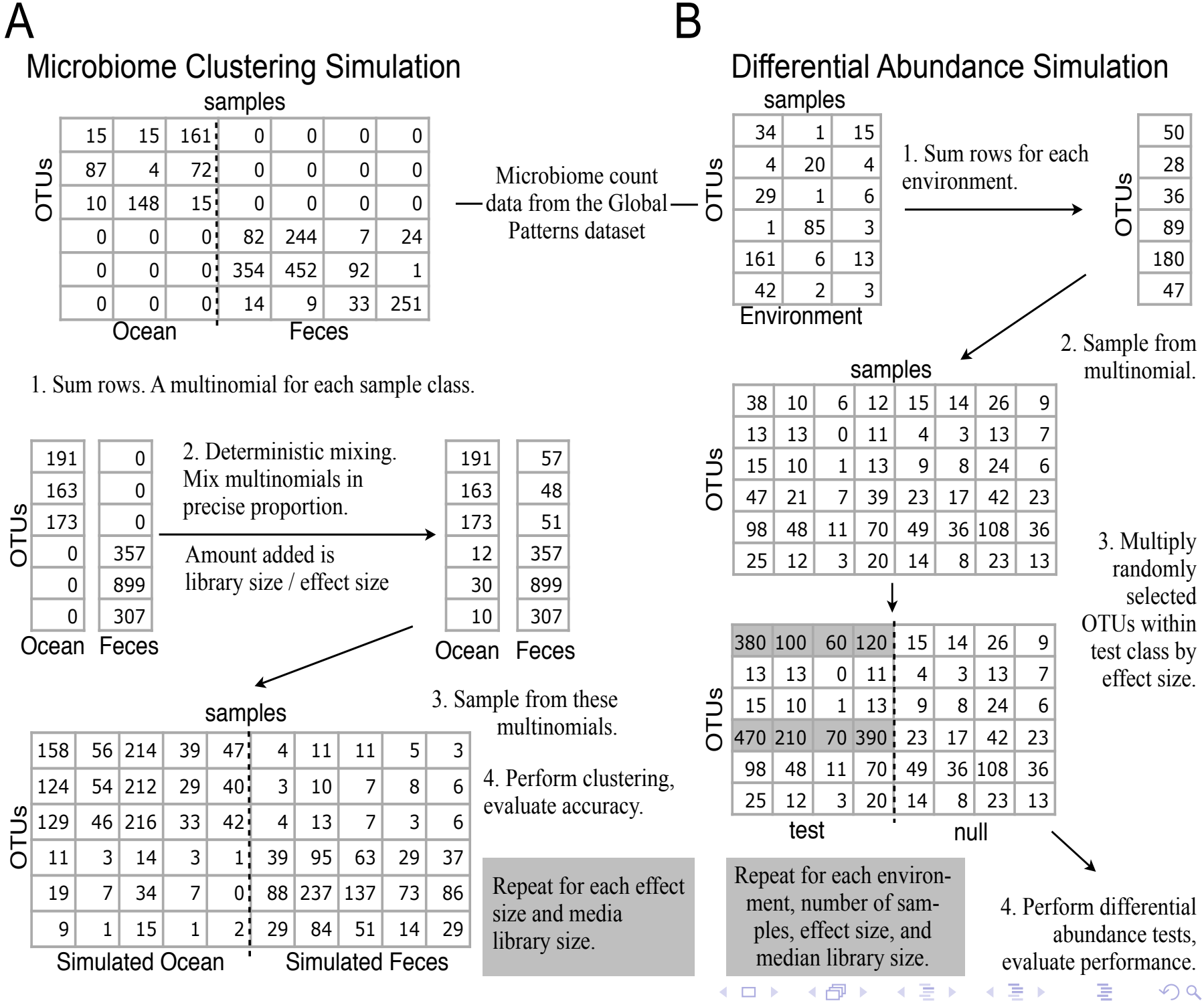
Estimate the parameters u_{ic} and ϕ_{ic} from the data for each OTU and sample condition.
The end result provides a variance stabilizing transformation of the data that allows a statistically efficient comparisons between conditions.
This application of a hierarchical mixture model is very similar to the random effects models used in the context of analysis of variance.

Using RNA-seq implementation : DESeq2

McMurdie and Holmes (2014) ``Waste Not, Want Not: Why rarefying microbiome data is inadmissible", PLOS Computational Biology, Methods.
Examples of Overdispersion in Microbiome Data.
Common-Scale Variance versus Mean for Microbiome Data. Each point in each panel represents a different OTU's mean/variance estimate for a biological replicate and study. The data in this figure come from the Global Patterns surveyand the Long-Term Dietary Patterns study(Right)
Variance versus mean abundance for rarefied counts. (Left) Common-scale variances and common-scale means, estimated according to the DESeq2 package.
The dashed gray line denotes the $\sigma^2 = \mu$ case (Poisson; $\phi = 0$). The cyan curve denotes the fitted variance estimate using DESeq.

Code


Common Scale
Rarefied
2004 HighFat
freshwater
marine
DietPatterns
GlobalPatterns
GlobalPatterns
Variance
Mean
1e+08
1e+05
1e+02
10
1000


A
Microbiome Clustering Simulation
samples
OTUs
15 15 161 0 0 0 0
87 4 72 0 0 0 0
10 148 15 0 0 0 0
0 0 0 82 244 7 24
0 0 0 354 452 92 1
0 0 0 14 9 33 251
Ocean
Feces
Microbiome count data from the Global Patterns dataset
1. Sum rows. A multinomial for each sample class.
191 163 173 0 0 0
0 0 0 357 899 307
Ocean Feces
2. Deterministic mixing. Mix multinomials in precise proportion.
Amount added is library size / effect size
191 163 173 12 30 10
57 48 51 357 899 307
Ocean Feces
3. Sample from these multinomials.
samples
158 56 214 39 47 4 11 11 5 3
124 54 212 29 40 3 10 7 8 6
129 46 216 33 42 4 13 7 3 6
11 3 14 3 1 39 95 63 29 37
19 7 34 7 0 88 237 137 73 86
9 1 15 1 2 29 84 51 14 29
Simulated Ocean
Simulated Feces
4. Perform clustering, evaluate accuracy.
Repeat for each effect size and media library size.
B
Differential Abundance Simulation
samples
34 1 15
4 20 4
29 1 6
1 85 3
161 6 13
42 2 3
Environment
1. Sum rows for each environment.
50 28 36 89 180 47
2. Sample from multinomial.
samples
38 10 6 12 15 14 26 9
13 13 0 11 4 3 13 7
15 10 1 13 9 8 24 6
47 21 7 39 23 17 42 23
98 48 11 70 49 36 108 36
25 12 3 20 14 8 23 13
3. Multiply randomly selected OTUs within test class by effect size.
380 100 60 120 15 14 26 9
13 13 0 11 4 3 13 7
15 10 1 13 9 8 24 6
470 210 70 390 23 17 42 23
98 48 11 70 49 36 108 36
25 12 3 20 14 8 23 13
test
null
Repeat for each environment, number of samples, effect size, and median library size.
4. Perform differential abundance tests, evaluate performance.

Normalizations in Simulation

For each simulated experiment we used the following normalization methods prior to calculating sample-wise distances.

1. **DESeqVS**. Variance Stabilization implemented in the DESeq package.
2. **None**. Counts not transformed. Differences in total library size could affect the values of some distance metrics.
3. **Proportion**. Counts are divided by total library size.
4. **Rarefy**. Rarefying is performed as defined in the introduction, using `rarefy_even_depth` implemented in the phyloseq package. with $N_{L,min}$ set to the 15^{th}-percentile of library sizes within each simulated experiment.
5. **UQ-logFC**. The Upper-Quartile Log-Fold Change normalization implemented in the edgeR package, coupled with the top-MSD distance.

Distances in Simulation

For each of the previous normalizations we calculated sample-wise distance/dissimilarity matrices using the following methods, if applicable.

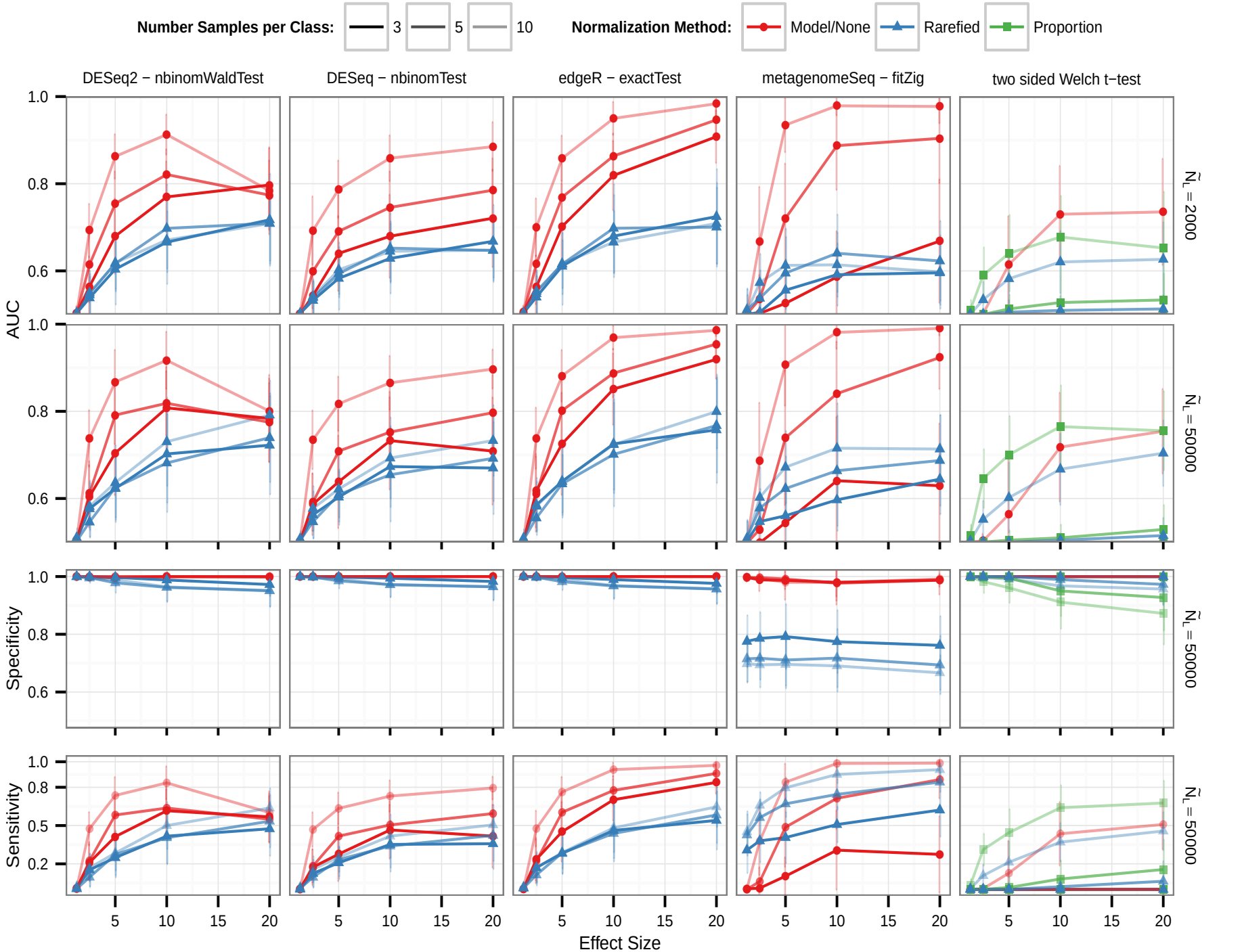
1. **Bray-Curtis**. The Bray-Curtis dissimilarity first defined in 1957 for forest ecology.
2. **Euclidean**. The euclidean distance treating each OTU as a dimension. $\sqrt{\sum_{i=1}^{n}(K_{i1} - K_{i2})^2}$, is the distance between samples 1 and 2,n the number of distinct OTUs.
3. **PoissonDist**. Our abbreviation of `PoissonDistance`, a sample-wise distance implemented in the PoiClaClu package (Witten,2011).
4. **top-MSD**. The mean squared difference of top OTUs, as implemented in edgeR.
5. **UniFrac-u**. The Unweighted UniFrac distance (Lozupone,2005).
6. **UniFrac-w**. The Weighted UniFrac distance (Lozupone,2007).

In order to consistently evaluate performance in this regard, we generated microbiome counts by sampling from two different multinomials that were based on either the Ocean or Feces microbiomes of the Global Patterns empirical dataset. An equal number of simulated microbiome samples was generated from each multinomial. The Ocean and Feces sample classes have negligible overlapping OTUs.
Mixing them by a defined proportion allows control over the difficulty of the clustering task from trivial (no mixing) to impossible (both multinomials evenly mixed).
Clustering was performed independently for each combination of simulated experiment, normalization method, and distance measure using partitioning around medoids (PAM).
The accuracy is the fraction of simulated samples correctly clustered; worst possible accuracy is 50% if all samples are clustered. (Rarefying procedure omits samples, so its accuracy can be below 50%)

Improvement in Power and FDR

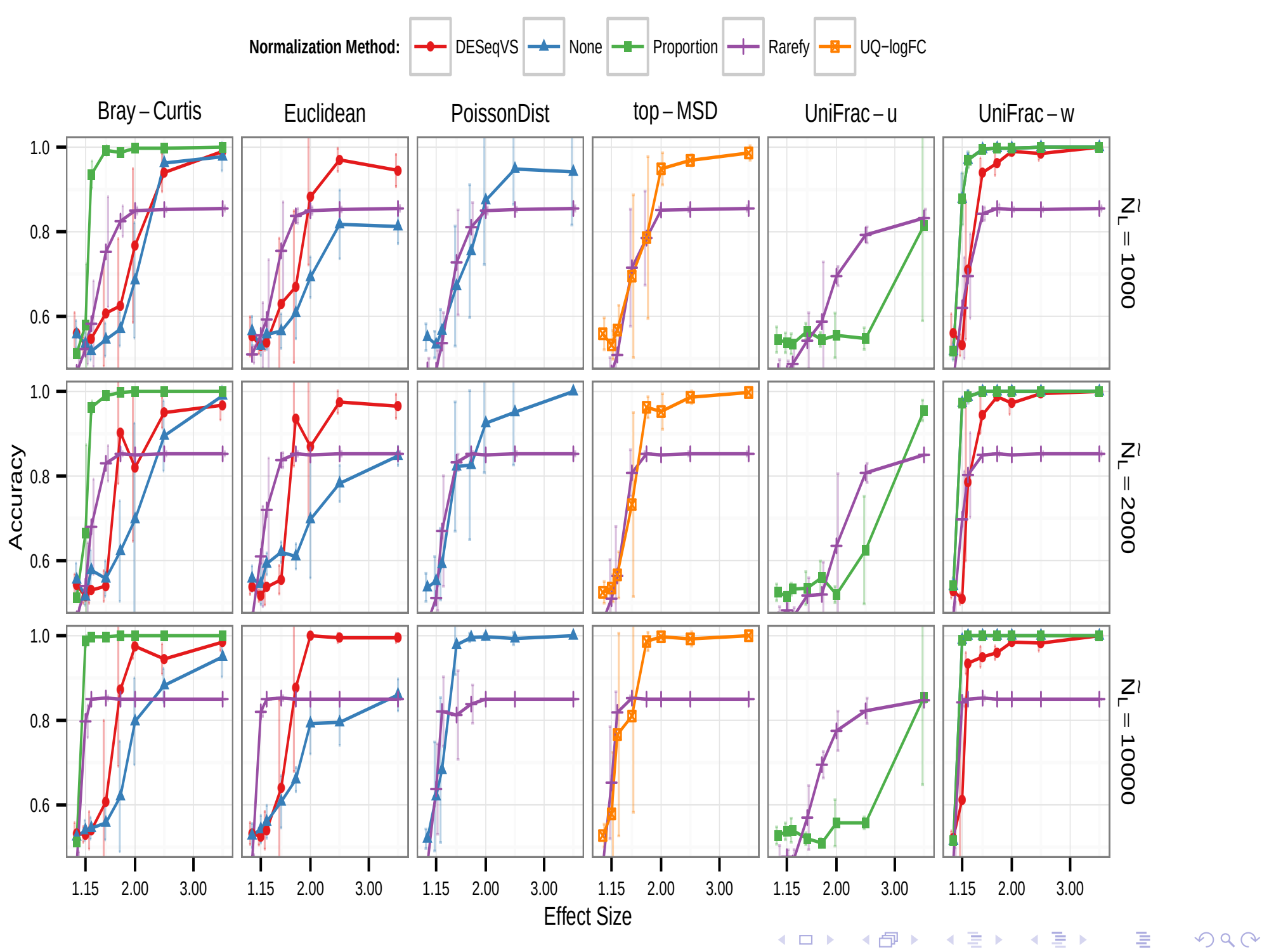
Performance of differential abundance detection with and without rarefying summarized by "Area Under the Curve" (AUC) metric of a Receiver Operator Curve (ROC) (vertical axis).
Briefly, the AUC value varies from 0.5 (random) to 1.0 (perfect).
The horizontal axis indicates the effect size, shown as the factor applied to OTU counts to simulate a differential abundance.
Each curve traces the respective normalization method's mean performance of that panel, with a vertical bar indicating a standard deviation in performance across all replicates and microbiome templates.

The right-hand side of the panel rows indicates the median library size, N, while the darkness of line shading indicates the number of samples per simulated experiment.
Color shade and shape indicate the normalization method.
Detection among multiple tests was defined using a False Discovery Rate (Benjamini-Hochberg) significance threshold of 0.05.


Number Samples per Class:
3
5
10
Normalization Method:
Model/None
Rarefied
Proportion
DESeq2 − nbinomWaldTest
DESeq − nbinomTest
edgeR − exactTest
metagenomeSeq − fitZig
two sided Welch t−test
AUC
Specificity
Sensitivity
Effect Size
$\tilde{N}_L = 2000$
$\tilde{N}_L = 50000$
$\tilde{N}_L = 50000$
$\tilde{N}_L = 50000$

Improvements of Distance based clustering

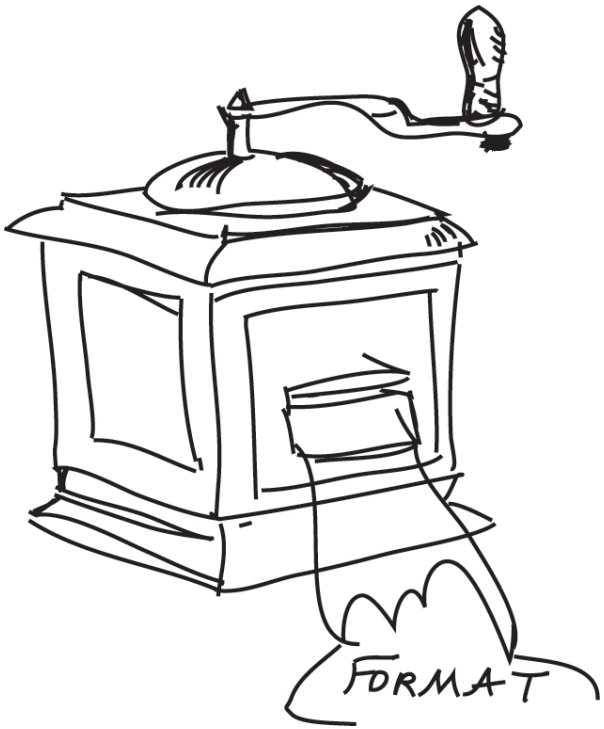
Clustering accuracy in simulated two-class mixing.
Clustering accuracy (with PAM:vertical axis) following different normalization and distance methods.
Points denote the mean values of replicates, with a vertical bar representing one standard deviation above and below.
The horizontal axis is the effect size.
Each multinomial is derived from two microbiomes that have negligible overlapping OTUs (Fecal and Ocean microbiomes in the Global Patterns dataset).
Higher values of effect size indicate an easier clustering task.


Normalization Method:
DESeqVS
None
Proportion
Rarefy
UQ−logFC
Bray – Curtis
Euclidean
PoissonDist
top – MSD
UniFrac – u
UniFrac – w
$\tilde{N}_L = 1000$
$\tilde{N}_L = 2000$
$\tilde{N}_L = 10000$
Accuracy
1.0
0.8
0.6
1.15
2.00
3.00
Effect Size

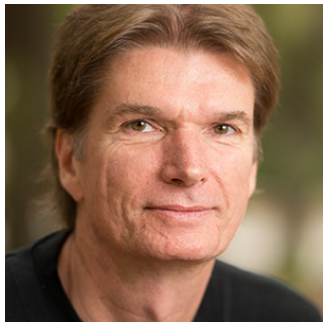
Examples using Phyloseq:
http://joey711.github.io/phyloseq-extensions/DESeq2.html

Benefitting from the tools and schools of Statisticians.....

Thanks to the R community:
Chessel, Jombart, Dray, Thioulouse ade4 , Wolfgang Huber, Michael Love for DESeq2, Gordon Smyth and his team for edgeR and Emmanuel Paradis for ape.



Collaborators:


David Relman Alfred Spormann Elizabeth Purdom
Justin Sonnenburg and Persi Diaconis.

Postdoctoral Fellows Paul (Joey) McMurdie, Alex Alekseyenko (NYU), Ben Callahan.
Students: Miling Shen, Diana Proctor, Alden Timme, Katie Shelef, Yana Hoy, John Chakerian, Julia Fukuyama, Kris Sankaran.
Funding from NIH-TR01, NIH/ NIGMS R01, NSF-VIGRE and NSF-DMS.

phyloseq

Joey McMurdie (joey711 on github).
Available in Bioconductor.
How can I (my students, my postdocs...) learn more?
Google: wiki phyloseq deseq2
`http://www-stat.stanford.edu/~susan/`