https://github.com/AdamOss88

Data normalization and differential abundance

SEQUENCING DATA ANALYSIS

Adam Ossowicki

Why ? – Which forest has more squirrels ?
Which microbiome has more Pseudomonas ?

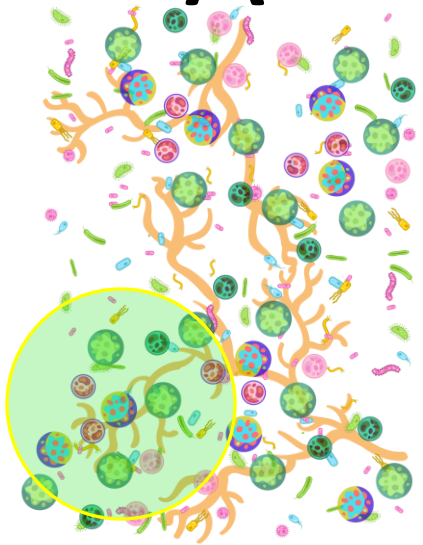
A

3x

B

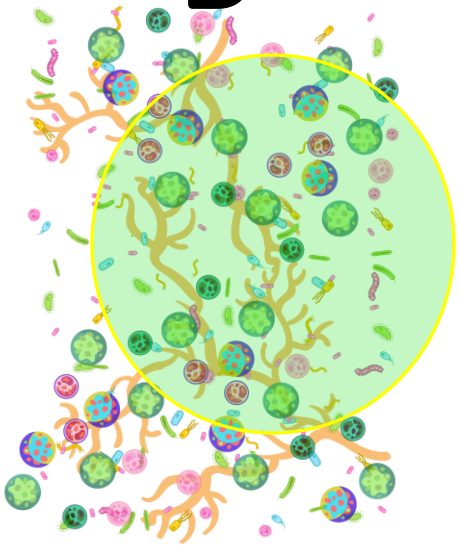
5x

A



20’000 reads

B



50’000 reads

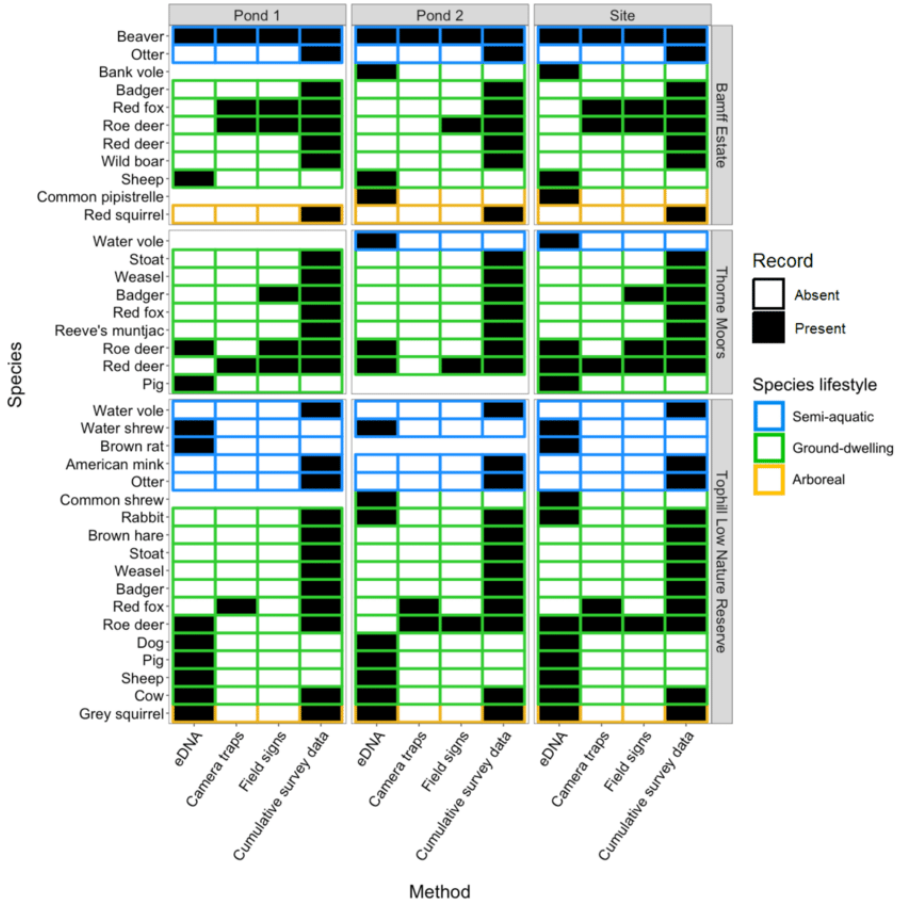
- Unequal sequencing dept
- PCR/sequencing biases
- Unequal detection

What is missing in those “experiments” ?

To normalize or not to normalize ?

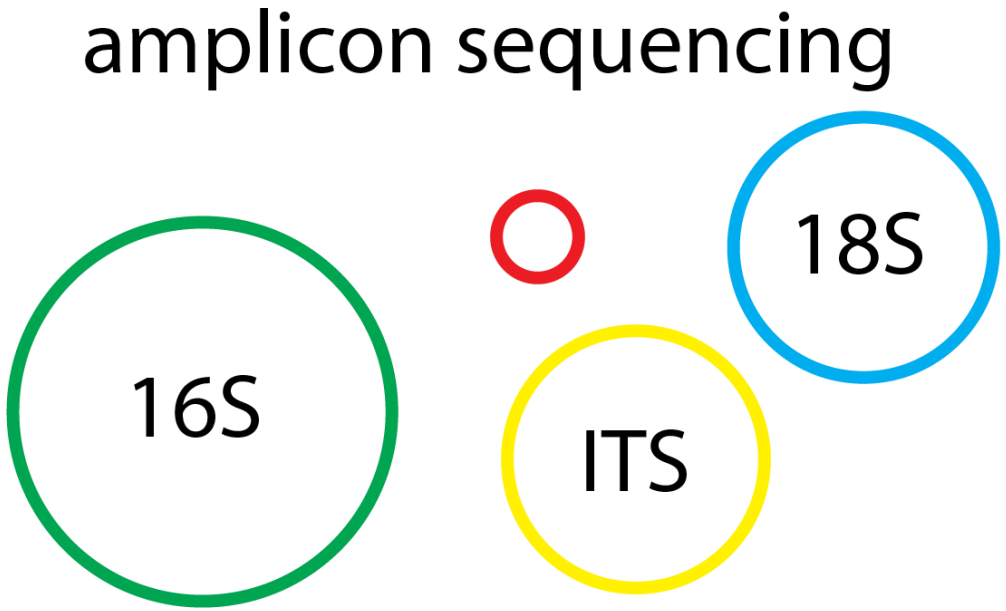
- Only presence absence patterns (be careful)
- Only observed number of species (be careful)
- Only phylogeny (ex. weighted unifrac)

- When comparing any relative abundance of taxa (amplicons) or any genes the normalization is <u>obligatory</u>



DOI:10.1101/546218

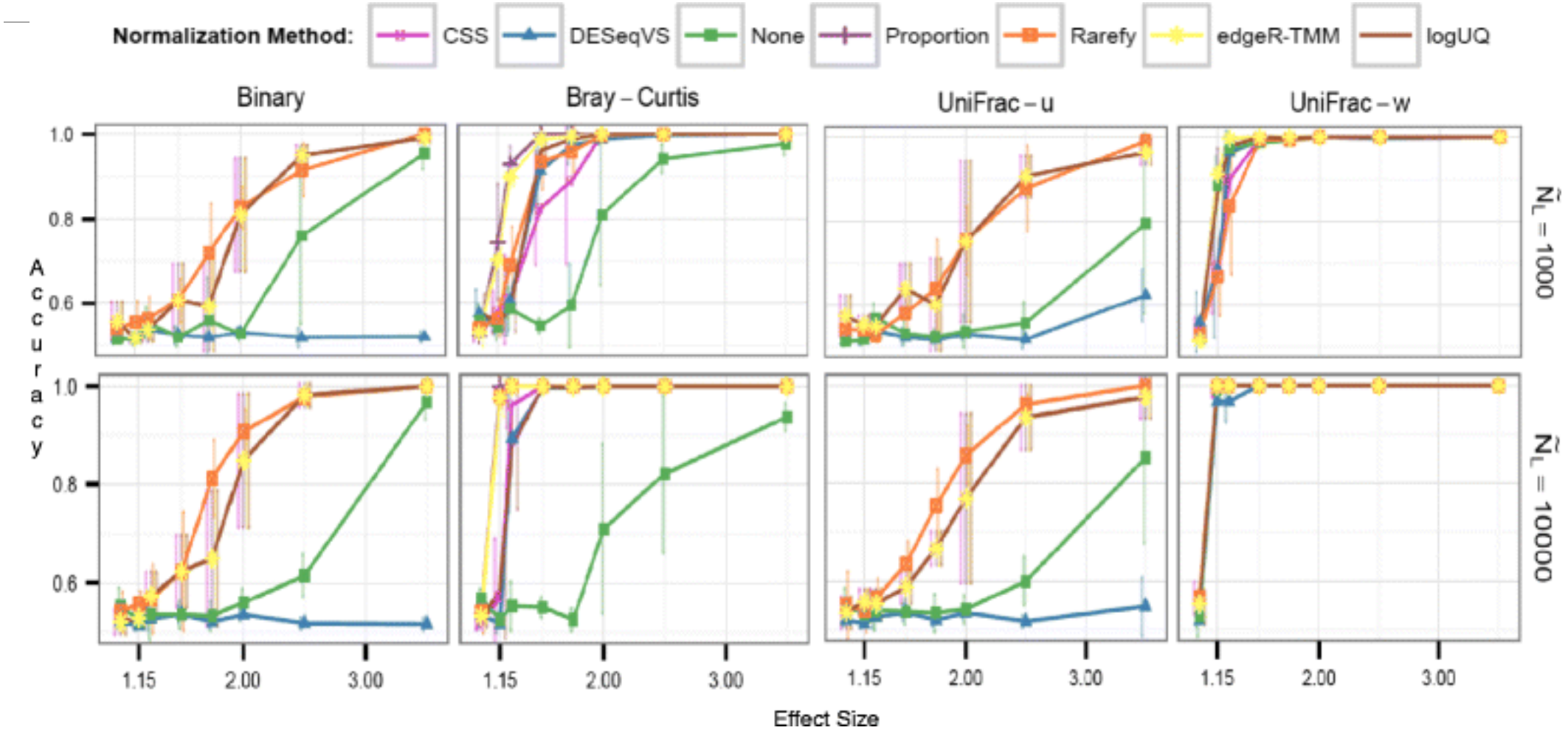
Normalization - amplicons



Method	Approach	Description
Rarefaction	Subsampling	Randomly subsamples reads to equal depth
CSS (Cumulative Sum Scaling)	Proportional	Divides each count by total reads per sample
Proportions (Percentage-based)	Proportional	Converts counts to fractions of total reads
Log Transformation	Transformation	Applies log(x + 1) to reduce skewness
Centered Log-Ratio (CLR)	Transformation	Log-ratio transformation for compositional data
And other more specific for certain functions: Deseq2, ANCOM II, edgeR …		

DOI:10.1101/546218

Comparison in normalization methods ?



Source: https://doi.org/10.1186/s40168-017-0237-y

Data normalization - metagenomes

Normalization Method	Description	Best Use Case	Limitations
Reads Per Kilobase per Million (RPKM)	Normalizes read counts by gene length and total mapped reads per million.	Gene and transcript-level analysis.	Does not fully correct for compositional bias.
Transcripts Per Million (TPM)	Normalizes like RPKM but ensures total expression sums to 1 million, allowing better cross-sample comparison.	Functional profiling and gene abundance analysis.	Still compositional, does not adjust for sequencing depth differences across samples.
**Relative Abundance (RPM or %) **	Expresses counts as reads per million (RPM) or percent relative abundance.	Simple comparisons of microbial composition.	Sensitive to sequencing depth and compositional bias.

Normalization Method	Description	Best Use Case	Limitations
Median-of-Ratios Normalization (RLE, DESeq2)	Adjusts for library size and compositional biases using a median-based approach.	Differential abundance analysis.	Requires statistical modeling and proper filtering of low-abundance taxa.
Cumulative Sum Scaling (CSS, metagenomeSeq)	Adjusts for library size while down-weighting highly abundant features.	Differential abundance testing, reducing dominance of highly abundant species.	Can be sensitive to extreme low-abundance features.
Genome Relative Abundance (GRA)	Estimates microbial species abundance using genome coverage and sequencing depth.	Whole-genome taxonomic profiling.	Requires accurate genome binning and assembly.

Source: https://doi.org/10.3389/fgene.2024.1369628

Data normalization R tools and pipelines

Software	Normalization Methods
DESeq2 (R)	Relative Log Expression (RLE)
edgeR (R)	Trimmed Mean of M-values (TMM)
MetagenomeSeq (R)	Cumulative Sum Scaling (CSS)
Phyloseq (R)	Variance stabilizing transformation (VST)
ANCOM-BC (R)	Log-ratio transformations

Pipelines
SqueezeMeta
Atlas
nf-core
bioBakery4
JAMS
WGSA2

Highly recommended

Mock community taxonomic classification performance of publicly available shotgun metagenomics pipelines

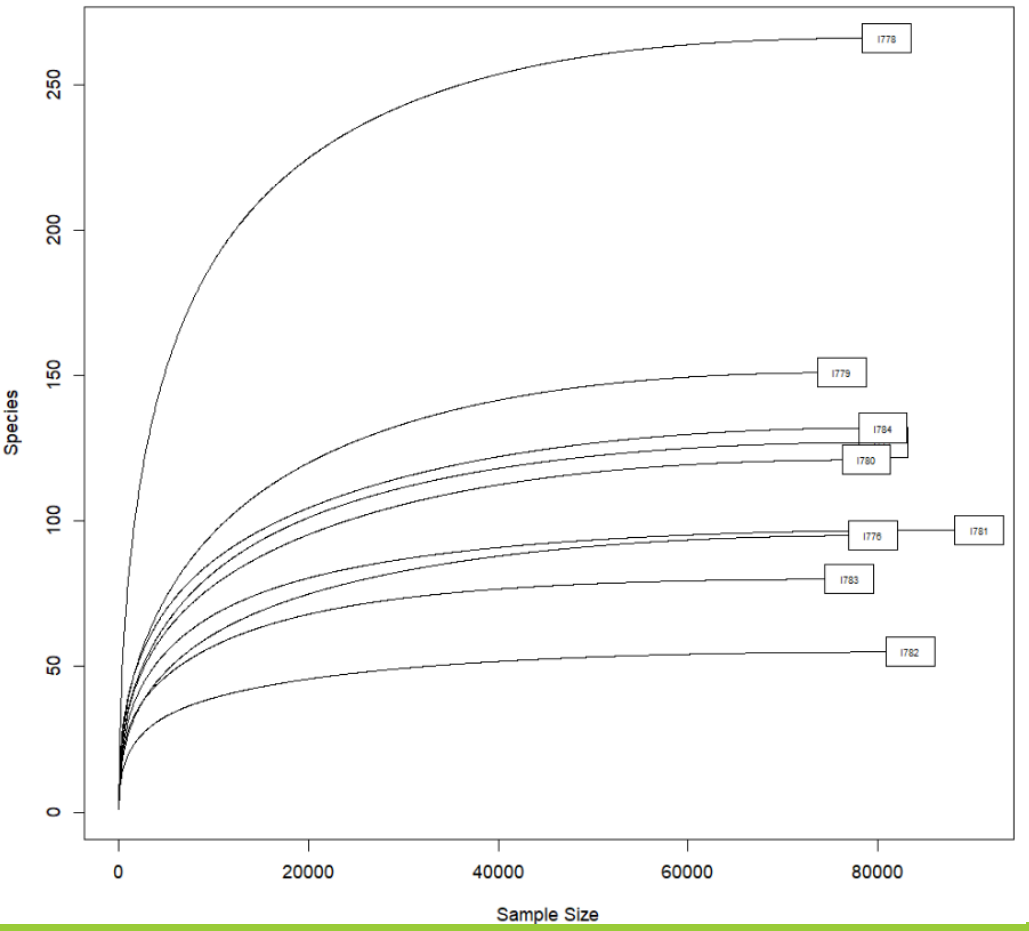
https://doi.org/10.1038/s41597-023-02877-7

Source: https://doi.org/10.3389/fgene.2024.1369628

The case of rarefaction

Plot the rarefaction curves and check your data

vegan::rarecurve(asv_table, step = 50 , cex=0.5)

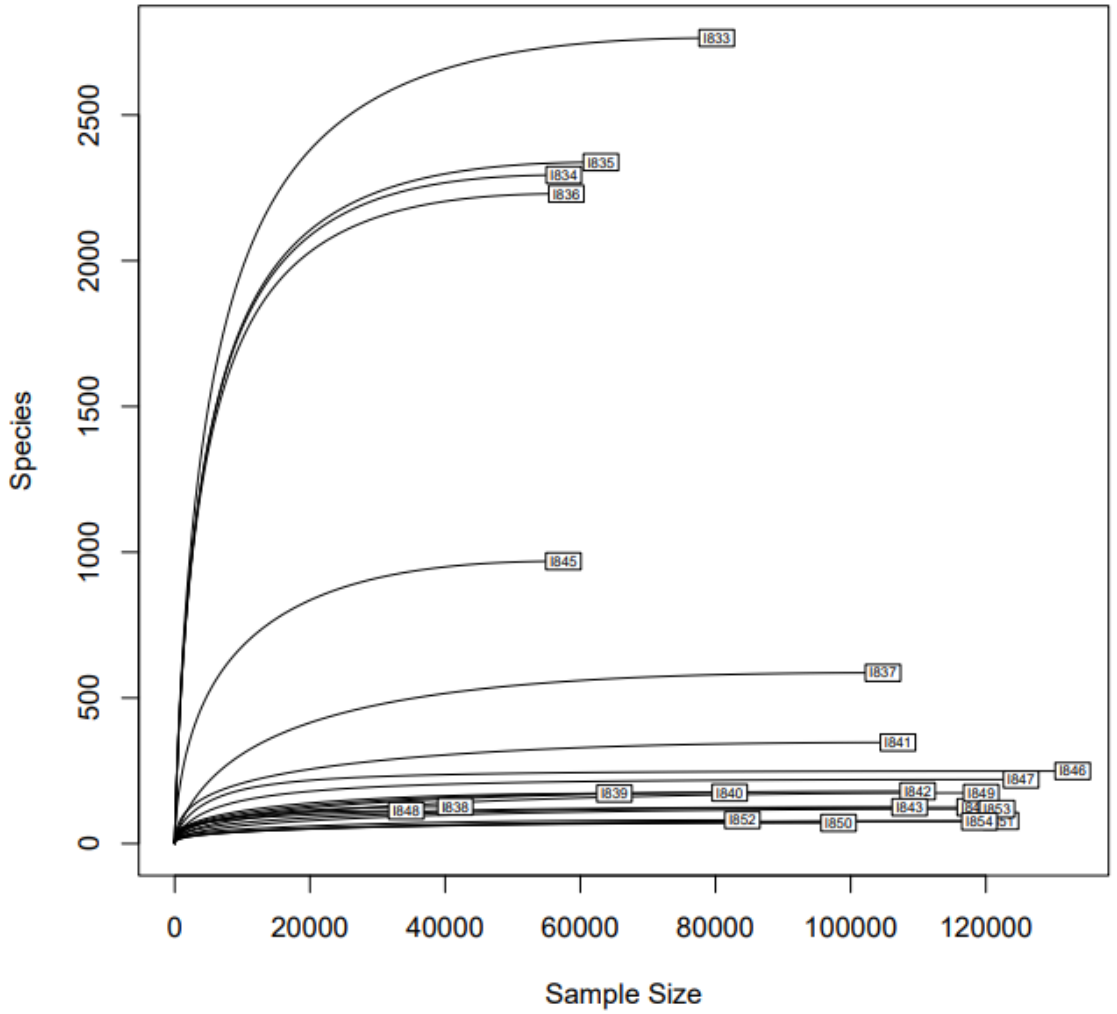


Rarefy within the phyloseq object

phyloseq::rarefy_even_depth(amplicon_data, rngseed=456, \
sample.size=min(sample_sums(amplicon_data)), replace=F)

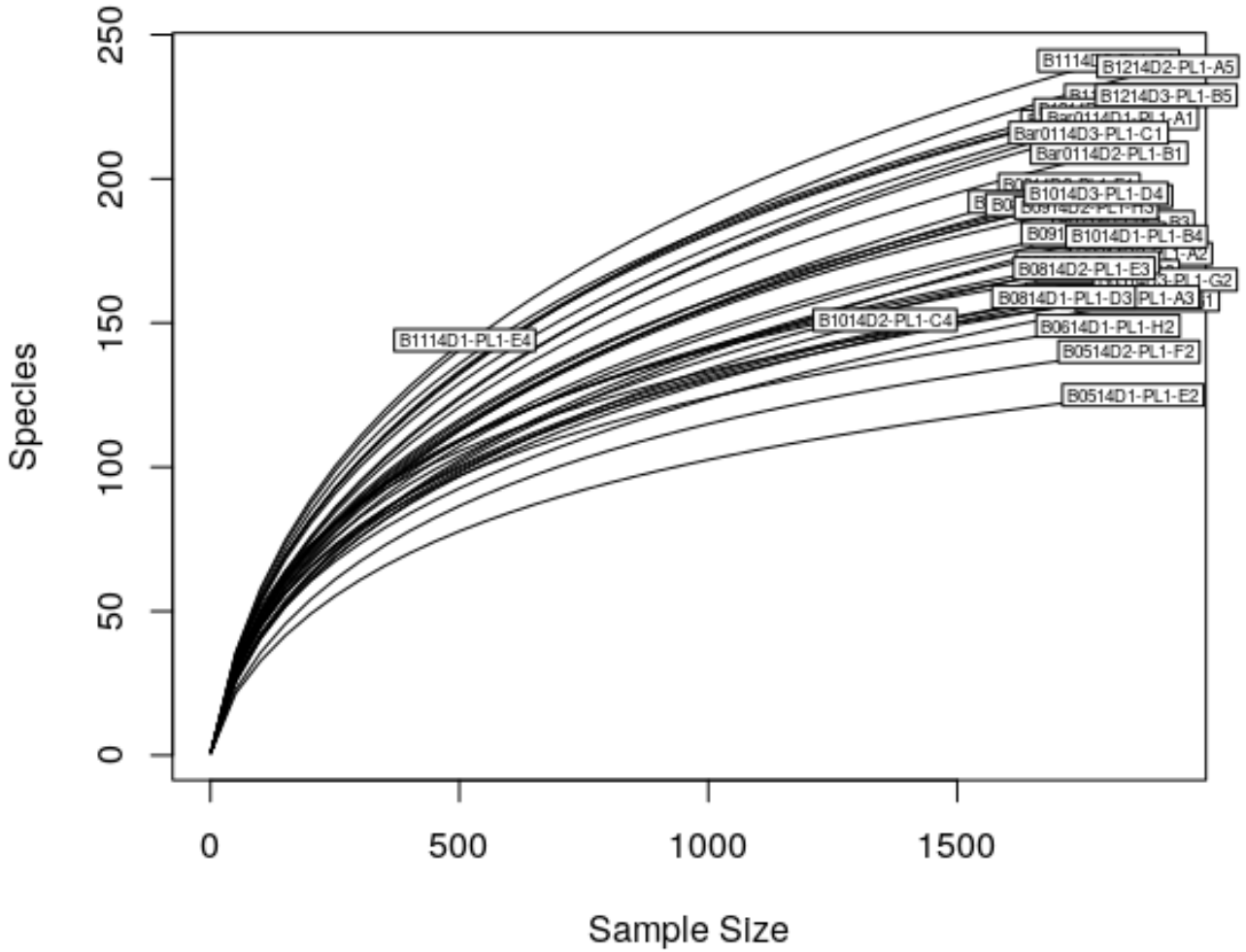
```
`set.seed(456)` was used to initialize repeatable random subsampling.
Please record this for your records so others can reproduce.
Try `set.seed(456); .Random.seed` for the full vector
...
20OTUs were removed because they are no longer
present in any sample after random subsampling

...
```

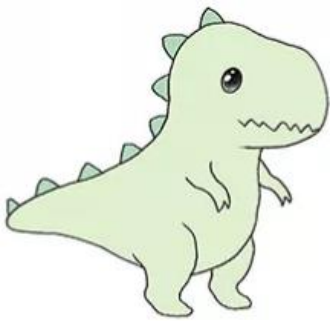


Hard decision:

- Should I discard samples ?
- What are the consequences of my actions ?
- Why I have just 3 replicates ….



Differential abundance



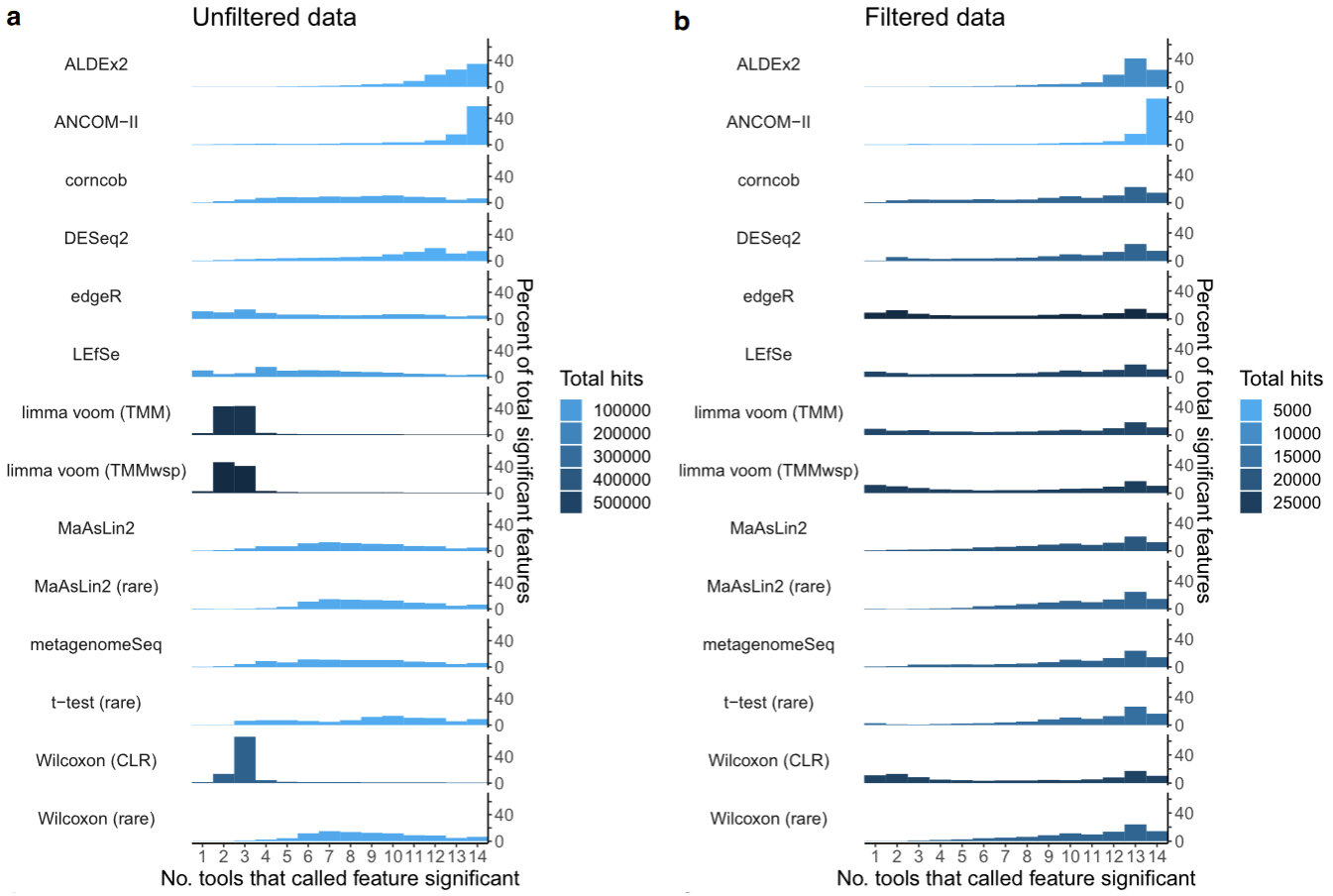
Microbiome differential abundance methods produce different results across 38 datasets

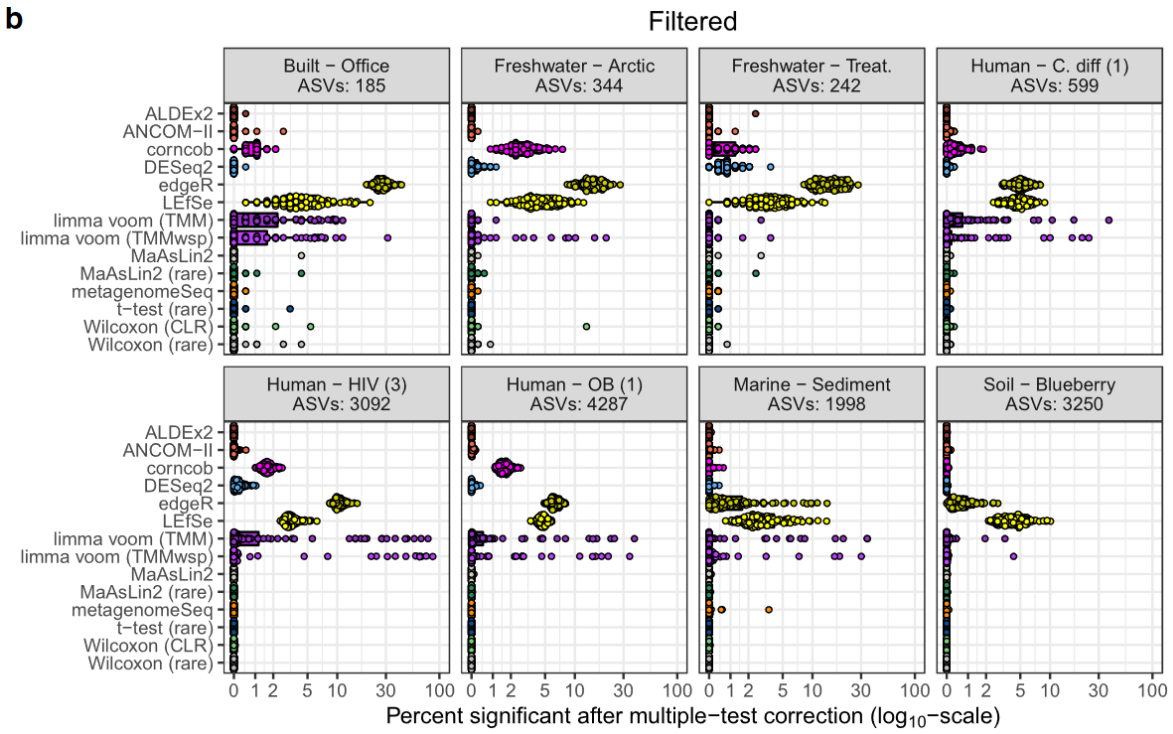
Jacob T. Nearing[1,7✉], Gavin M. Douglas[1,7], Molly G. Hayes[2], Jocelyn MacDonald[3], Dhwani K. Desai[4], Nicole Allward[5], Casey M. A. Jones[6], Robyn J. Wright[6], Akhilesh S. Dhanani[4], André M. Comeau[4] & Morgan G. I. Langille[4,6]

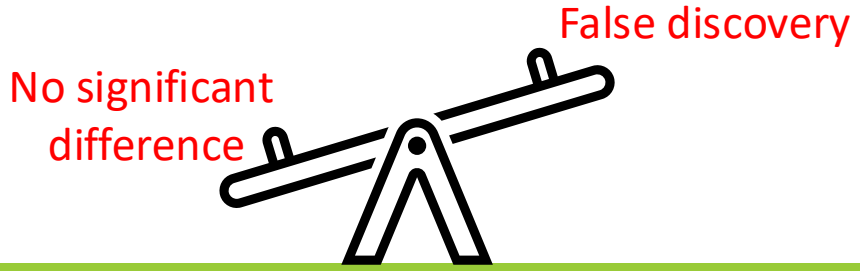
Table 1 Differential abundance tools compared in this study.

Tool (version)	Input	Norm.	Trans.	Distribution	Covariates	Random effects	Hypothesis test	FDR Corr.	CoDa	Dev. For
ALDEx2 (1.18.0)	Counts	None	CLR	Dirichlet-multinomial	Yes*	No	Wilcoxon rank-sum	Yes	Yes	RNA-seq, 16S, MGS
ANCOM-II (2.1)	Counts	None	ALR	Non-parametric	Yes	Yes	Wilcoxon rank-sum	Yes	Yes	MGS
Corncob (0.1.0)	Counts	None	None	Beta-binomial	Yes	No	Wald (default)	Yes	No	16S, MGS
DESeq2 (1.26.0)	Counts	Modified RLE (default is RLE)	None	Negative binomial	Yes	No	Wald (default)	Yes	No	RNA-seq, 16S, MGS
edgeR (3.28.1)	Counts	RLE (default is TMM)	None	Negative binomial	Yes*	No	Exact	Yes	No	RNA-seq
LEFse	Rarefied Counts	TSS	None	Non-parametric	Subclass factor only	No	Kruskal-Wallis	No	No	16S, MGS
MaAsLin2 (1.0.0)	Counts	TSS	AST (default is log)	Normal (default)	Yes	Yes	Wald	Yes	No	MGS
MaAsLin2 (rare) (1.0.0)	Rarefied counts	TSS	AST (default is log)	Normal (default)	Yes	Yes	Wald	Yes	No	MGS
metagenomeSeq (1.28.2)	Counts	CSS	Log	Zero-inflated (log-) Normal	Yes	No	Moderated t	Yes	No	16S, MGS
limma voom (TMM) (3.42.2)	Counts	TMM	Log; Precision weighting	Normal (default)	Yes	Yes	Moderated t	Yes	No	RNA-seq
limma voom (TMMwsp) (3.42.2)	Counts	TMMwsp	Log; Precision weighting	Normal (default)	Yes	Yes	Moderated t	Yes	No	RNA-seq
t-test (rare)	Rarefied Counts	None	None	Normal	No	No	Welch's *t*-test	Yes	No	N/A
Wilcoxon (CLR)	CLR abundances	None	CLR	Non-parametric	No	No	Wilcoxon rank-sum	Yes	Yes	N/A
Wilcoxon (rare)	Rarefied counts	None	None	Non-parametric	No	No	Wilcoxon rank-sum	Yes	No	N/A

*The tool supports additional covariates if they are provided. ANCOM-I automatically performs ANOVA in this case. ALDEx2 requires that users select the test, and edgeR requires use of a different function (glmFit or glmQLFit instead of exactTest).
ALR additive log-ratio, *AST* arcsine square-root transformation, *CLR* centered log-ratio, *CoDa* compositional data analysis, *CSS* cumulative sum scaling, *FDR Corr.* false-discovery rate correction, *MGS* metagenomic sequencing, *RLE* relative log expression, *TMM* trimmed mean of M-values, *Trans.* transformation, *TSS* total sum scaling







Take home: There is no perfect tool … but the more data the better

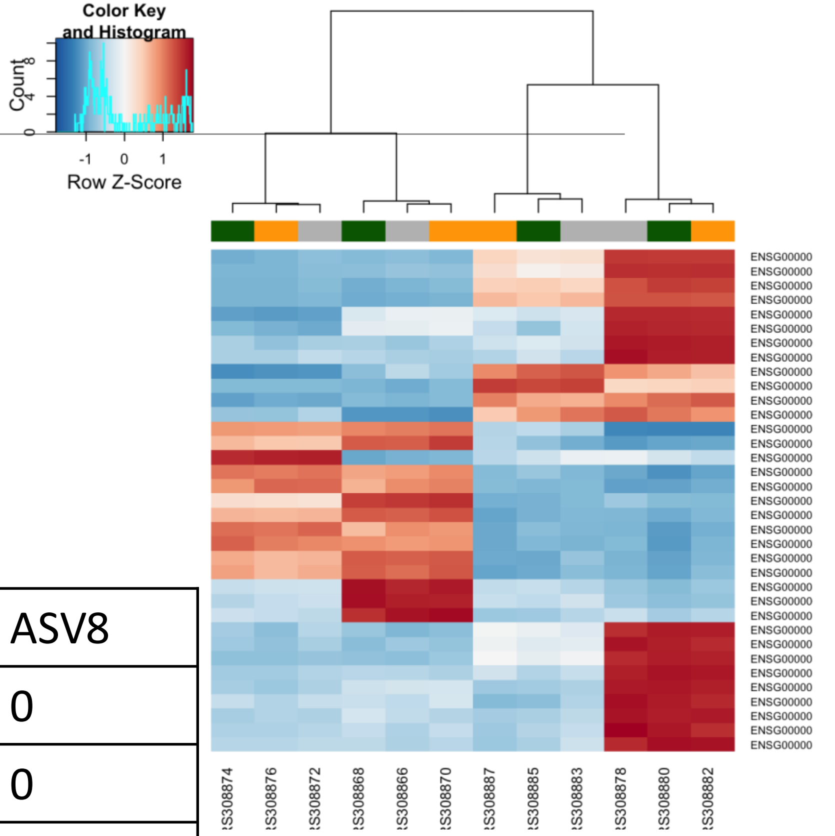
Deseq2

- use non-normalized data !

- design to work with RNAseq data

- uses negative binominal distribution model

- zeros are often a problem, but they always are – sparse matrix



	ASV1	ASV2	ASV3	ASV4	ASV5	ASV6	ASV7	ASV8
sample1	0	0	0	0	0	0	0	0
sample2	0	0	0	0	0	0	0	0
sample3	0	0	0	0	0	0	0	0
sample4	0	0	0	0	0	0	0	0
sample5	0	0	0	0	0	0	0	0

ANCOM-II

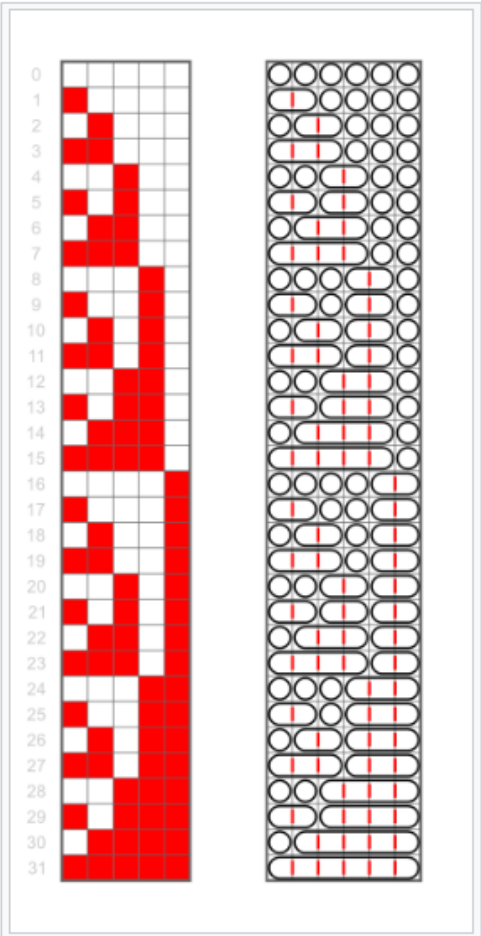
Analysis of Compositions of Microbiomes with Bias Correction

Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor[1*], Jean M. Macklaim[1], Vera Pawlowsky-Glahn[2] and Juan J. Egozcue[3]

[1] Department of Biochemistry, University of Western Ontario, London, ON, Canada
[2] Departments of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona, Girona, Spain
[3] Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain

- Original paper entitled: "Analysis of Microbiome Data in the Presence of Excess Zeros"
- ANCOM-BC estimates the unknown sampling fractions, corrects the bias induced by their differences through a log linear regression model including the estimated sampling fraction as an offset terms



The 32 compositions of 6

1 + 1 + 1 + 1 + 1 + 1
2 + 1 + 1 + 1 + 1
1 + 2 + 1 + 1 + 1
. . .
1 + 5
6

wikipedia

Practically speaking…

- look at number of reads
- look at taxonomic level
- look at replication
- look what it is …. What if it is *Rickettsiales* ?

Multivariate statistics

	Kingdom	Phylum	Class	Order
CGGAATTACTGGGCGTAAAGC...	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacterales
CGGAATTACTGGGCGTAAAGC...	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales
CGGAATTACTGGGCGTAAAGC...	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacterales
GGATTTATTGGGCGTAAAGCG...	Bacteria	Firmicutes	Bacilli	Lactobacillales
CGGAATTACTGGGCGTAAAGC...	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacterales
CGGAATTACTGGGCGTAAAGC...	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacterales
CGGAATTACTGGGCGTAAAGC...	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales
CGGAATCATTGGGCGTAAAGA...	Bacteria	Actinobacteriota	Actinobacteria	Micrococcales
CGGAATTACTGGGCGTAAAGC...	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales
CGGAATTACTGGGCGTAAAGC...	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales
CGGAATTACTGGGCGTAAAGC...	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales
CGGAATTATTGGGCGTAAAGG...	Bacteria	Firmicutes	Desulfitobacteriia	Desulfitobacteriales
CGGAATTACTGGGCGTAAAGC...	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacterales

	TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTCTGTTAAGTCAGATGTGAAATCCCC
1776	27767
1777	26603
1778	10539
1779	52375
1780	56209
1781	70882
1782	49657
1783	41002
1784	37660