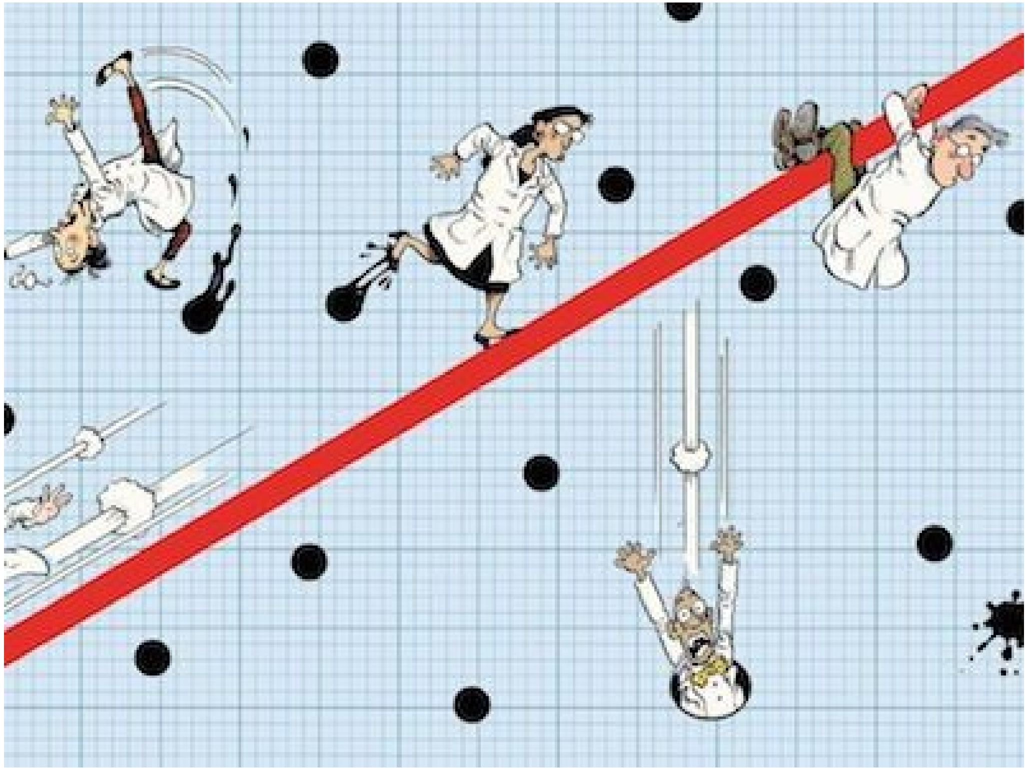
# Anatomy of taxonomic profiling data

- Patterns of variation in taxonomic profiling data

- Visualizing the data and statistical summaries

# Statistical properties: diving into data

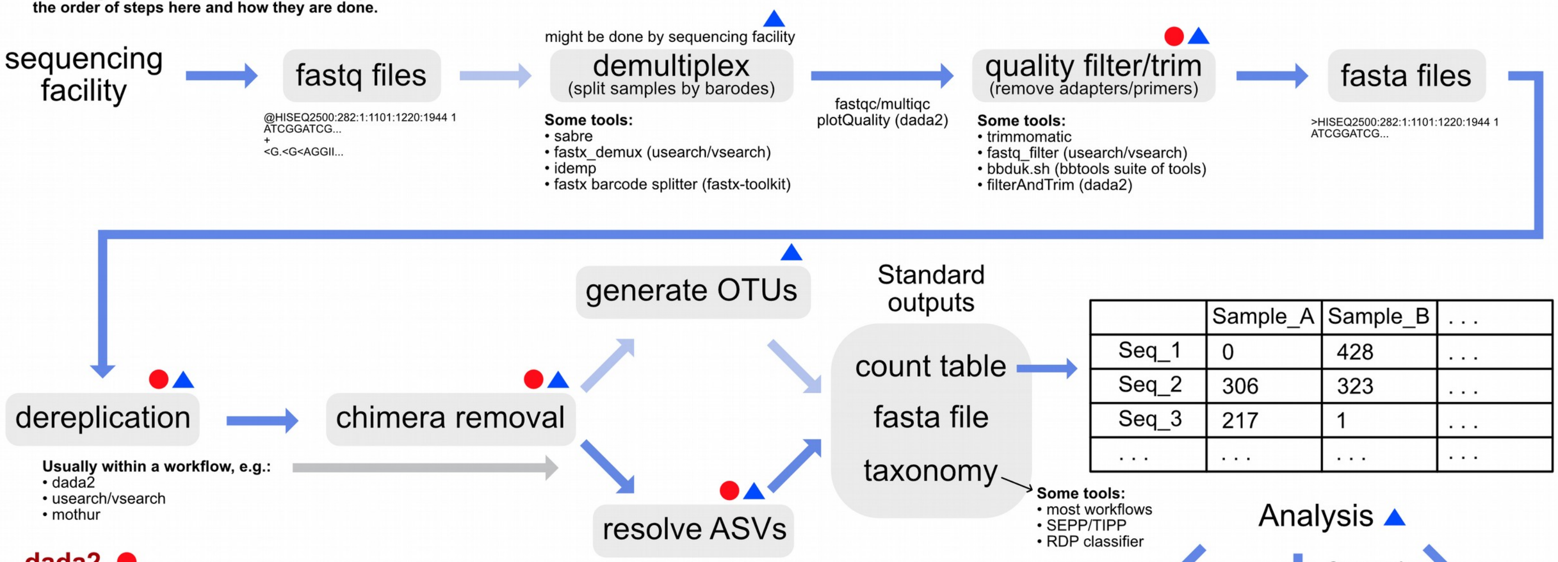


Picture: Nature Publishing Group

# Overview of generic* amplicon workflow

*This is generic; specific workflows can vary on the order of steps here and how they are done.

**When working with your own data you should never follow any pipeline blindly. There can be critical differences based on your data.**



sequencing facility → fastq files → demultiplex ▲ → quality filter/trim ● ▲ → fasta files

fastq files

```
@HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G.<G<AGGII...
```

might be done by sequencing facility

demultiplex ▲
(split samples by barodes)

**Some tools:**
- sabre
- fastx_demux (usearch/vsearch)
- idemp
- fastx barcode splitter (fastx-toolkit)

fastqc/multiqc
plotQuality (dada2)

quality filter/trim ● ▲
(remove adapters/primers)

**Some tools:**
- trimmomatic
- fastq_filter (usearch/vsearch)
- bbduk.sh (bbtools suite of tools)
- filterAndTrim (dada2)

fasta files

```
>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
```

dereplication ● ▲ → chimera removal ● ▲ → generate OTUs ▲ / resolve ASVs ● ▲ → Standard outputs

**Usually within a workflow, e.g.:**
- dada2
- usearch/vsearch
- mothur

Standard outputs
count table
fasta file
taxonomy

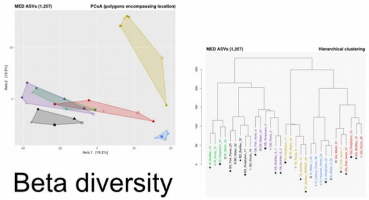
| | Sample_A | Sample_B | . . . |
|---|---|---|---|
| Seq_1 | 0 | 428 | . . . |
| Seq_2 | 306 | 323 | . . . |
| Seq_3 | 217 | 1 | . . . |
| . . . | . . . | . . . | . . . |

**Some tools:**
- most workflows
- SEPP/TIPP
- RDP classifier

**dada2** ●
**qiime2** ▲

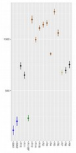
Analysis ▲

**Some tools:**
- phyloseq
- Breakaway
- DivNet
- CORNCOB
- SpiecEasi
- DESeq2



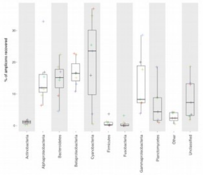
Beta diversity

e.g. dissimilarity metrics
ordination
hierarchical clustering



Alpha diversity

e.g. richness
evenness
diversity



Taxonomic summaries

**Some tools that provide whole workflows:**

**dada2** runs within R (ASVs)

**usearch/vsearch** runs at the command line (ASVs and OTUs)

**mothur** runs at the command line (OTUs only currently)

**qiime2** provides a multi-interface environment that employs processing tools like those above, infrastructure for easily documenting all processing performed, and interactive visualizations

**Happy Belly Bioinformatics**
JOSE 10.21105/jose.00053

AstrobioMike
Orcid: 0000-0001-7750-9145

Lee, (2019). Happy Belly Bioinformatics: an open-source resource dedicated to helping biologists utilize bioinformatics. Journal of Open Source Education, 4(41), 53, https://doi.org/10.21105/jose.00053

astrobiomike.github.io

# Common study designs

**Cross-sectional**

population (cohort) studies

**Prospective**

long-term follow-ups

**Longitudinal**

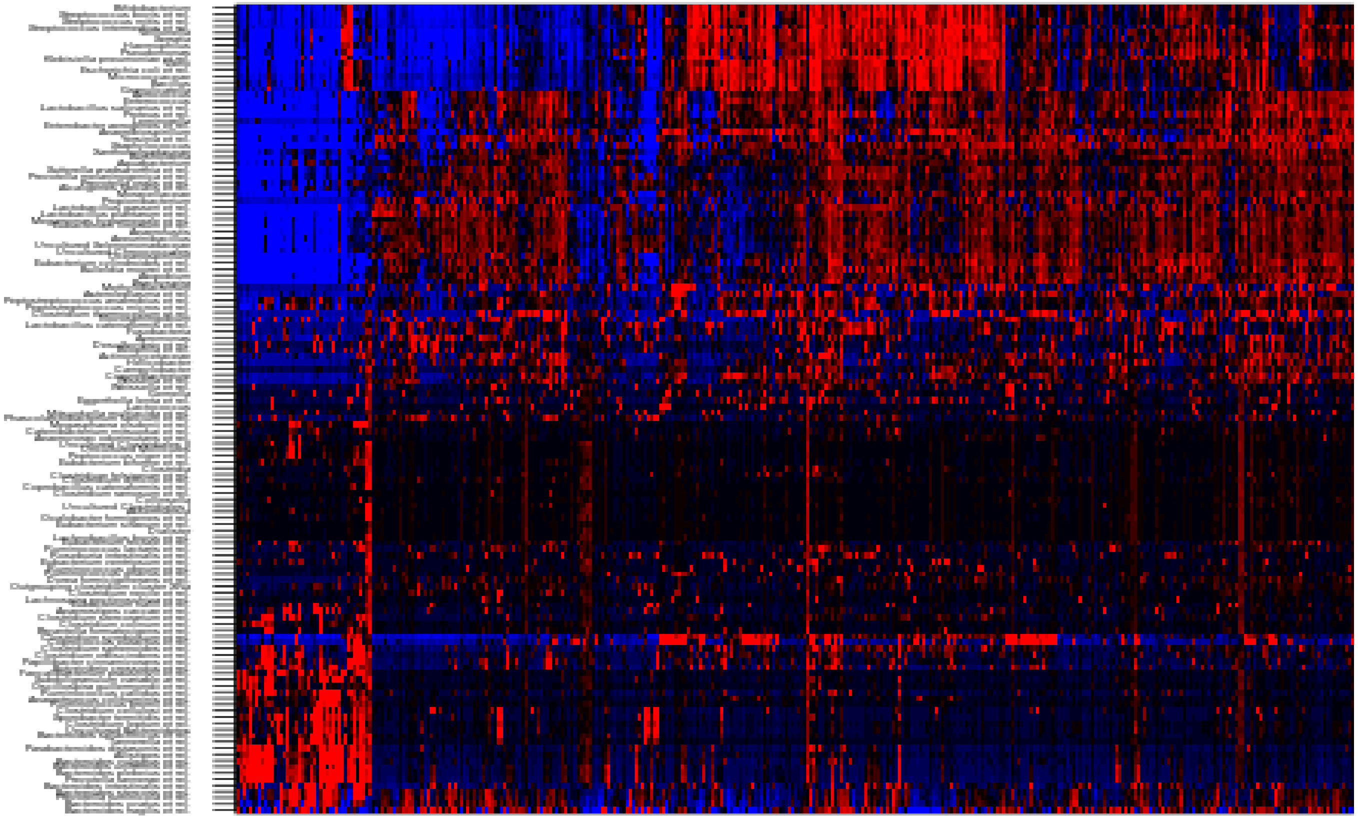
ecosystem dynamics

**Case-control & Intervention**

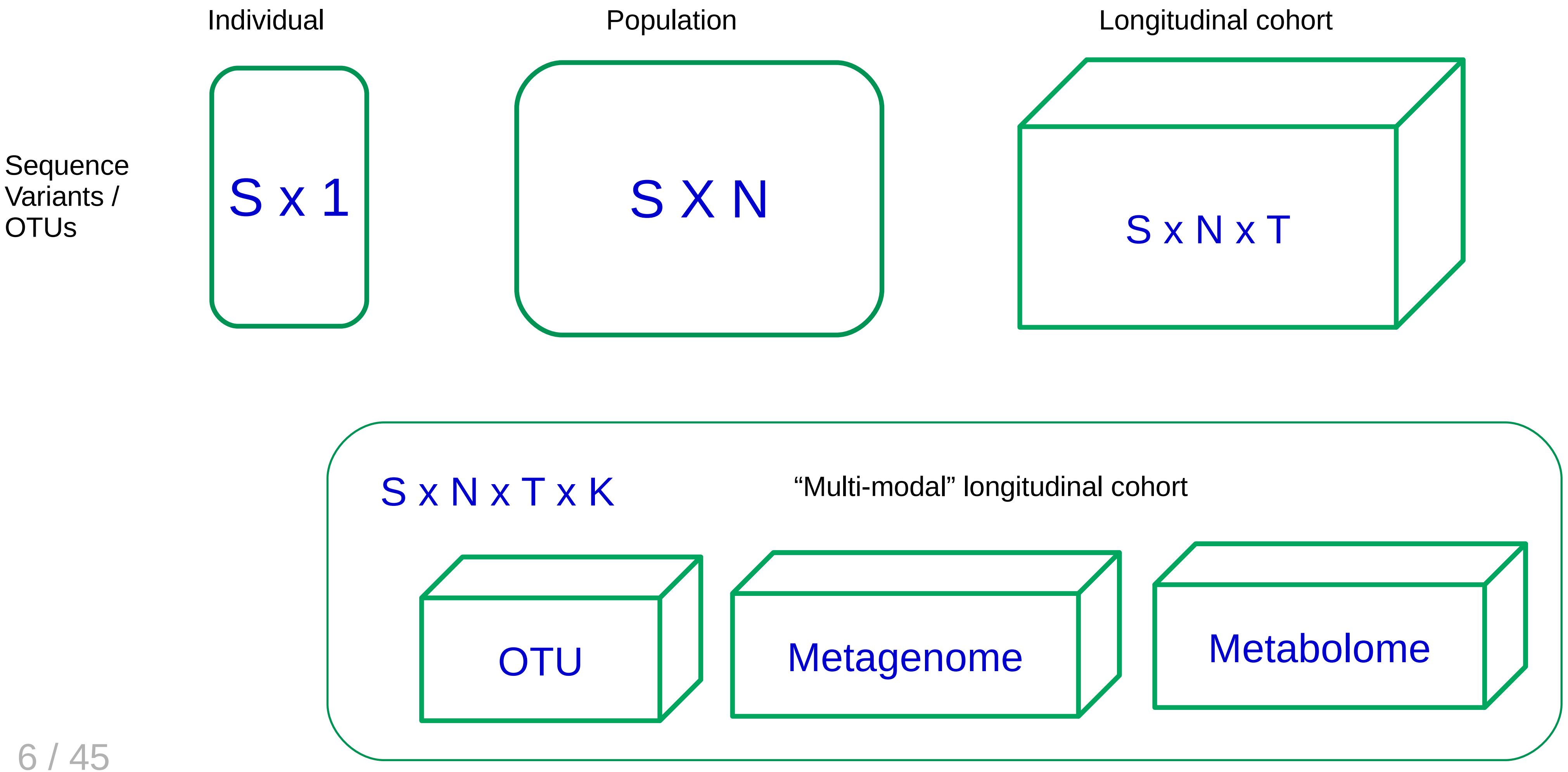
targeted experimental testing

# Organisms and samples are not independent

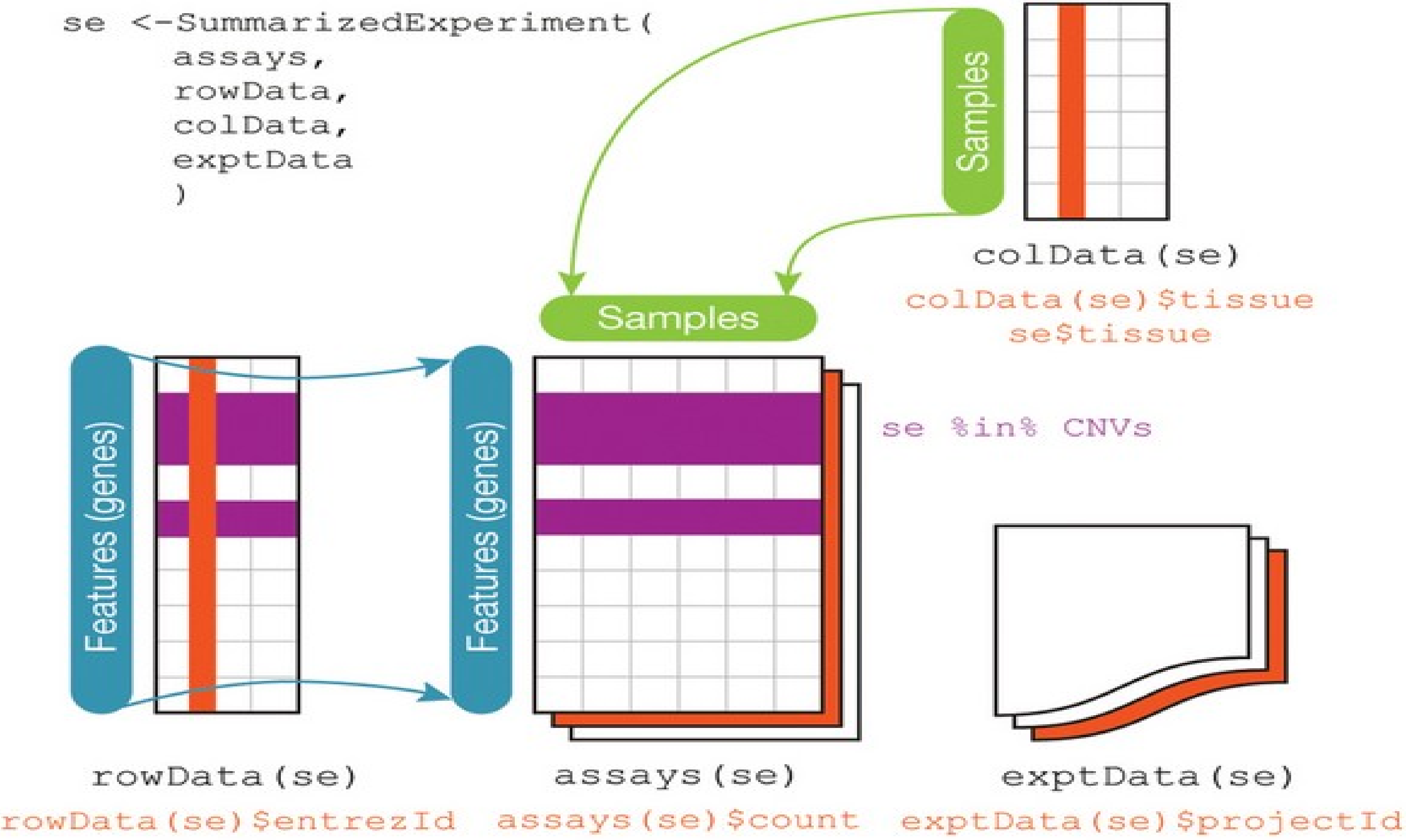
understanding & modeling the (latent) structure(s)



# From individuals to populations, follow-ups, and multimodal data



```
se <-SummarizedExperiment(
     assays,
     rowData,
     colData,
     exptData
     )
```



```
colData(se)
colData(se)$tissue
se$tissue
```

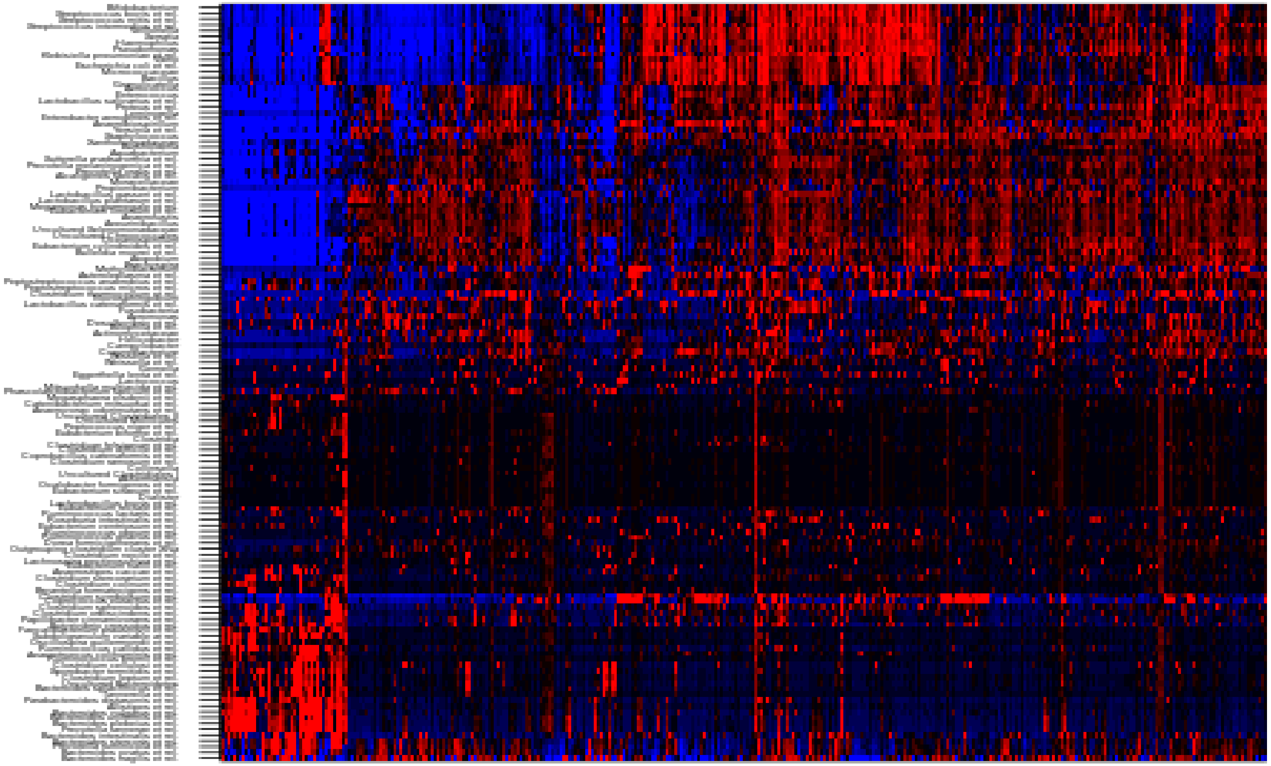
```
se %in% CNVs
```

```
rowData(se)
rowData(se)$entrezId
```

```
assays(se)
assays(se)$count
```

```
exptData(se)
exptData(se)$projectId
```

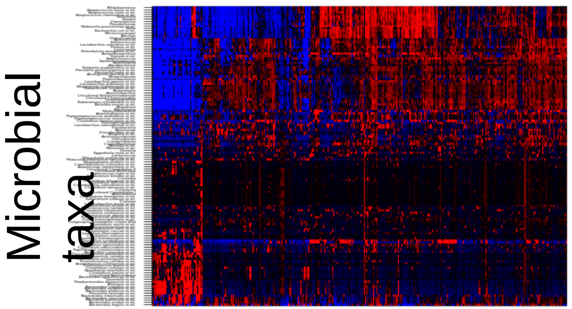
# Abundance matrix

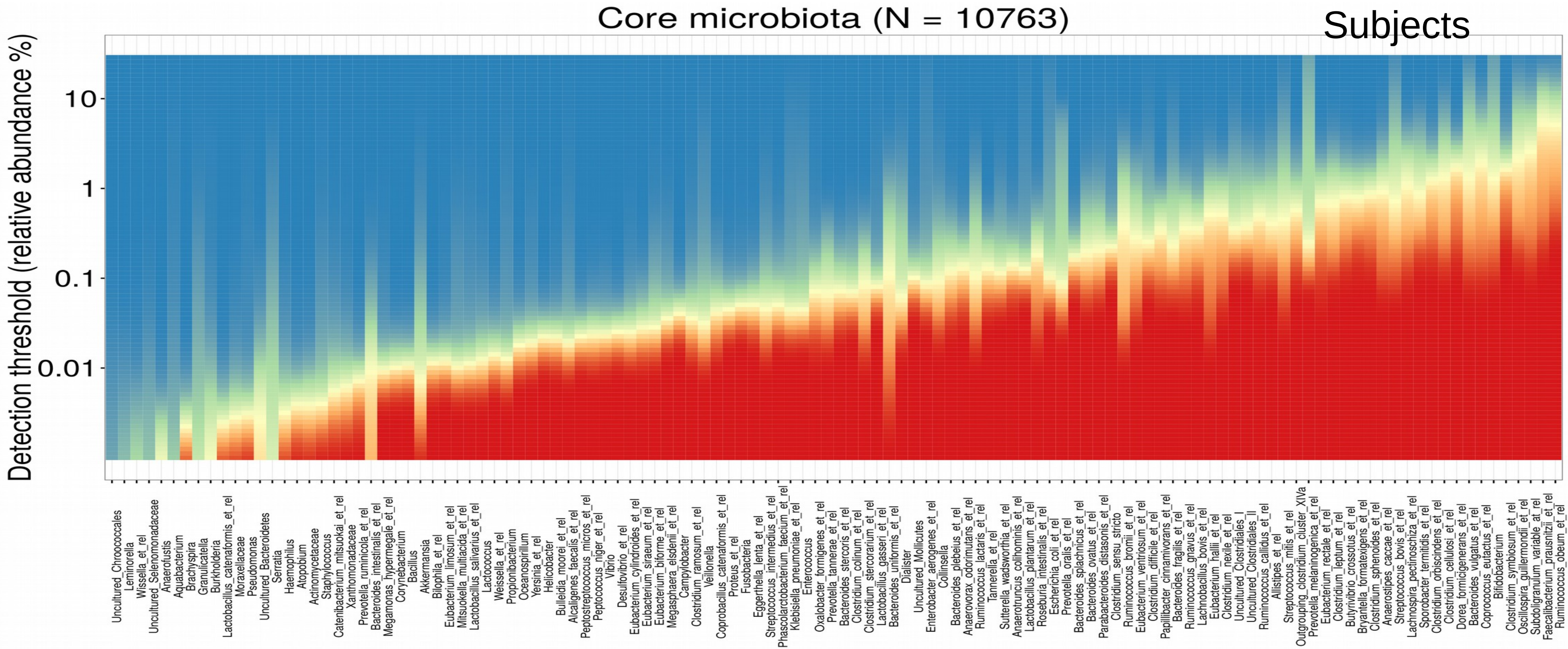


Open data:
Fecal microbiota in 1000 western adults (Lahti *et al.* Nature Comm. 2014)

# Core microbiota

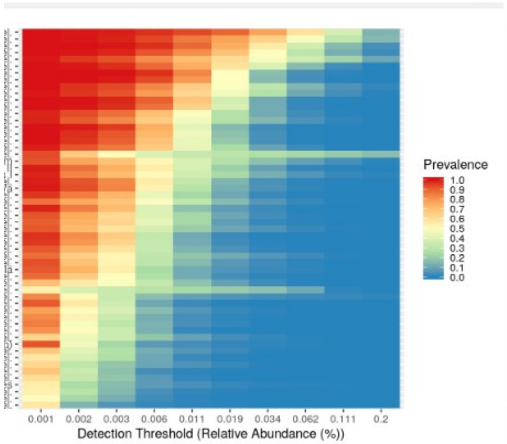
only few species are prevalent (shared) in population at a high abundance

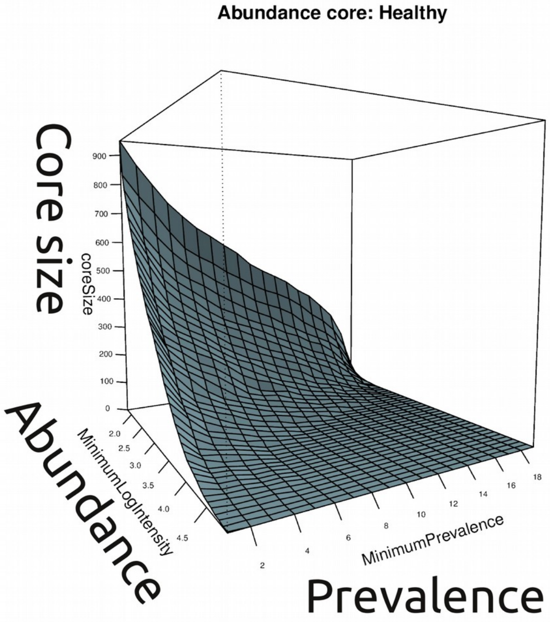




Data: HITChip Atlas

# Shared core microbiota in healthy adults depends on analysis depth and prevalence





N = 1488

"Blanket analysis"
github.com/microbiome

Estimate frequency in the core for each phylotype & bootstrap for confidence intervals

Jalanka-Tuovinen J et al. (2011) Intestinal microbiota in healthy adults: Temporal analysis reveals individual and common core and relation to digestive symptoms. PLoS One 6:e23035

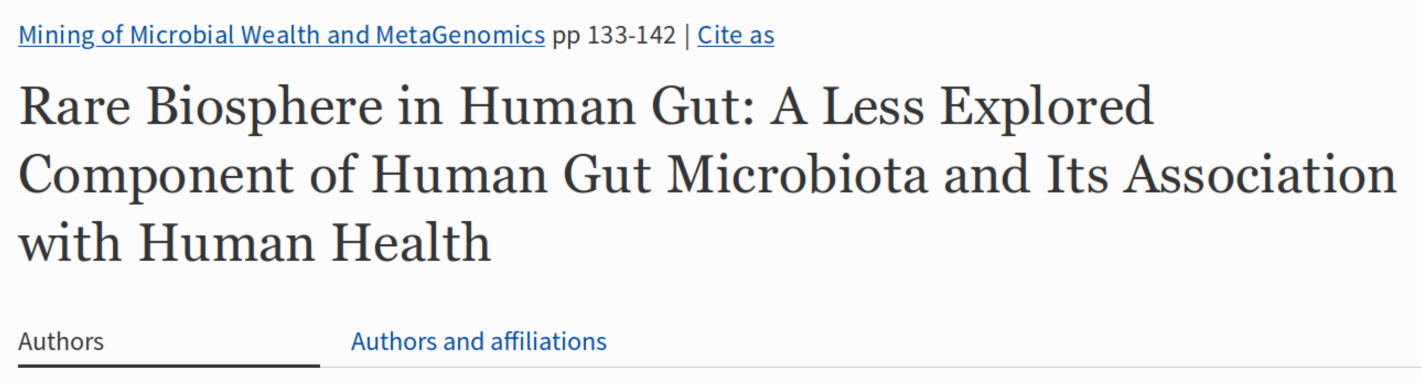
Salonen A et al. (2012) The adult intestinal core microbiota is determined by analysis depth and health status. Clinical Microbiology and Infection 18:16–20.

**Core & prevalence**

prevalence(x)

core(x)

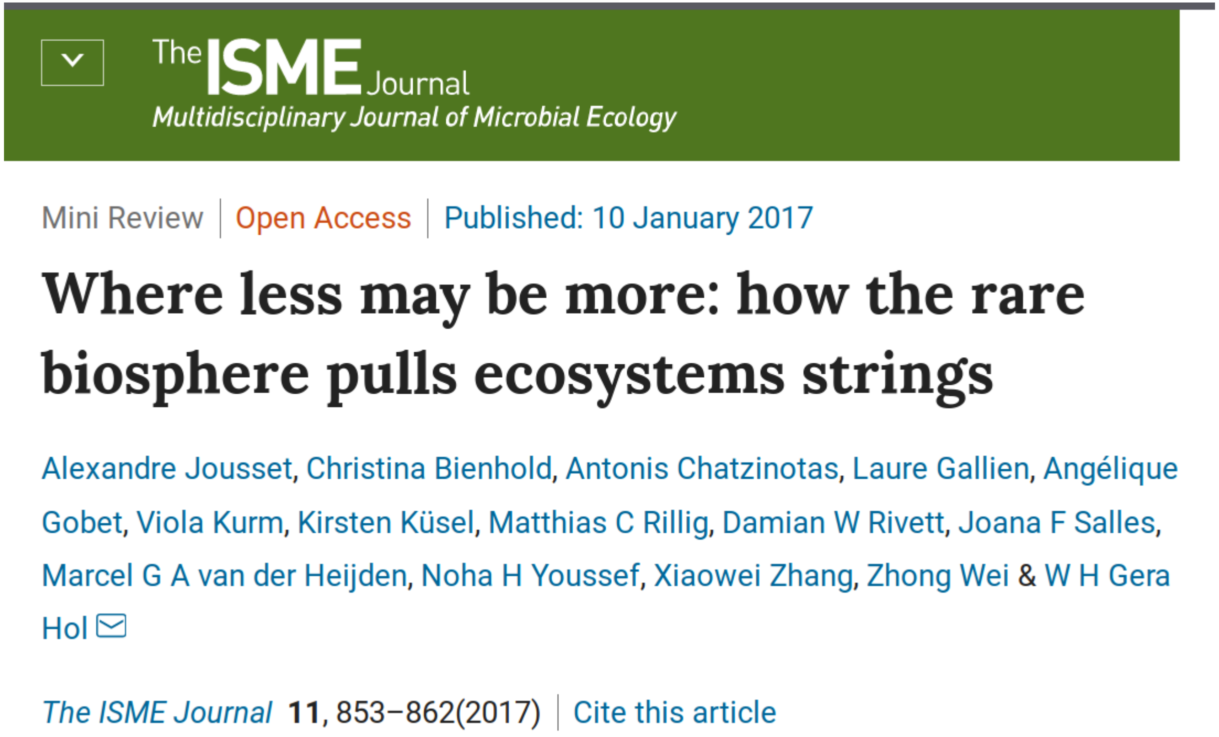
core_members(x)



Mining of Microbial Wealth and MetaGenomics pp 133-142 | Cite as

# Rare Biosphere in Human Gut: A Less Explored Component of Human Gut Microbiota and Its Association with Human Health

Shrikant S. Bhute, Saroj S. Ghaskadbi, Yogesh S. Shouche



The ISME Journal
Multidisciplinary Journal of Microbial Ecology

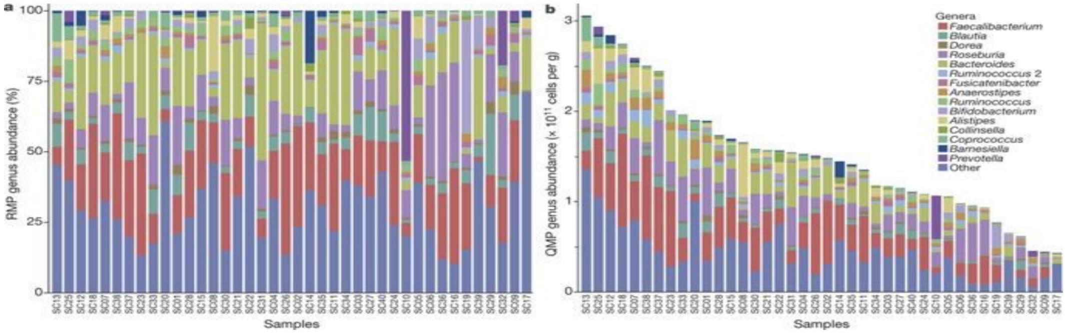
Mini Review | Open Access | Published: 10 January 2017

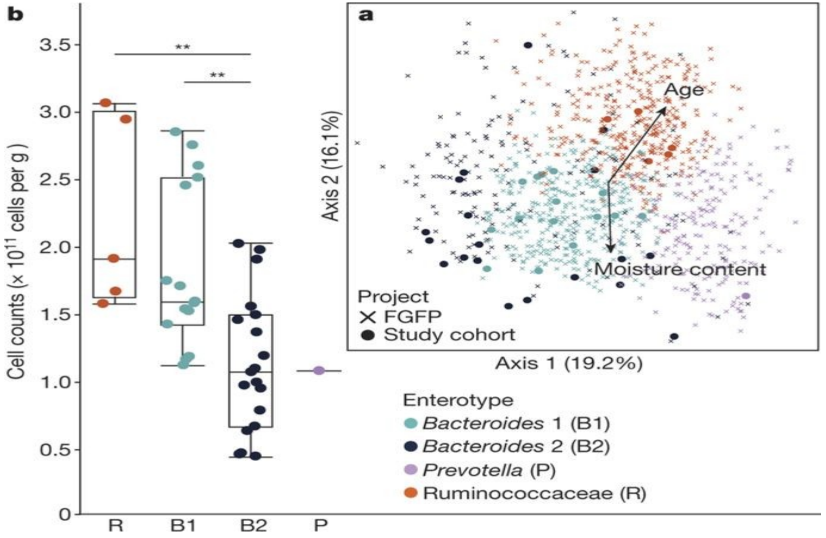
# Where less may be more: how the rare biosphere pulls ecosystems strings

Alexandre Jousset, Christina Bienhold, Antonis Chatzinotas, Laure Gallien, Angélique Gobet, Viola Kurm, Kirsten Küsel, Matthias C Rillig, Damian W Rivett, Joana F Salles, Marcel G A van der Heijden, Noha H Youssef, Xiaowei Zhang, Zhong Wei & W H Gera Hol

*The ISME Journal* **11**, 853–862(2017) | Cite this article

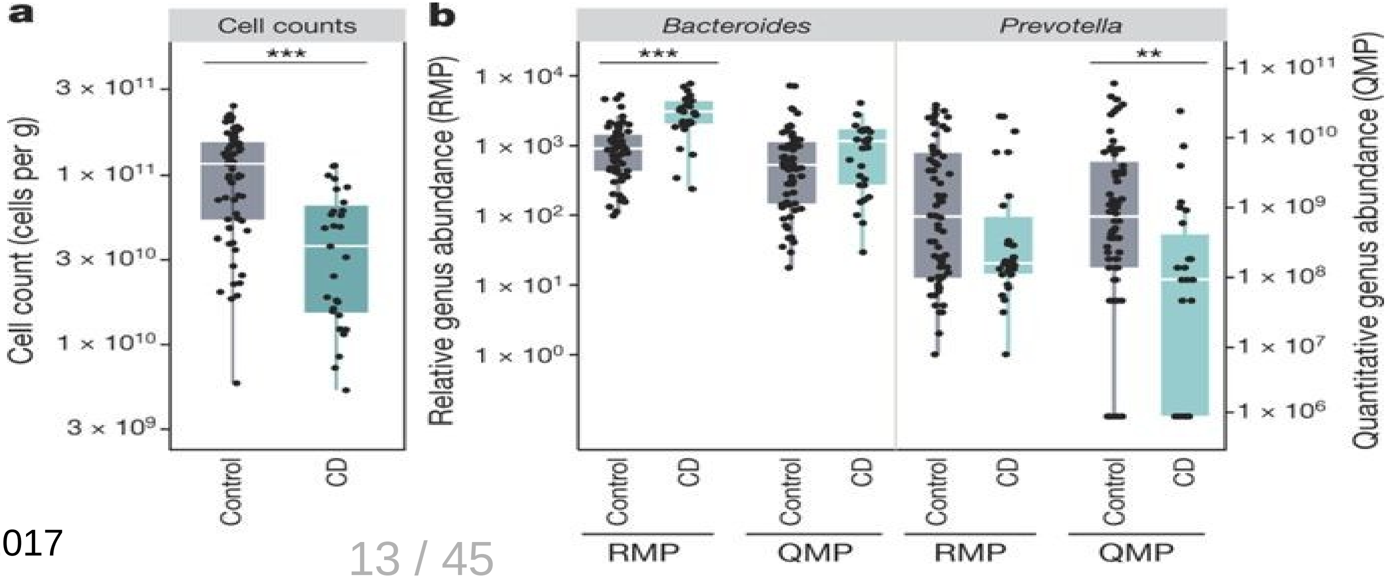
# Relative versus absolute abundance: quantitative microbiome profiling





RMP vs. QMP:

drastic effect on conclusions!



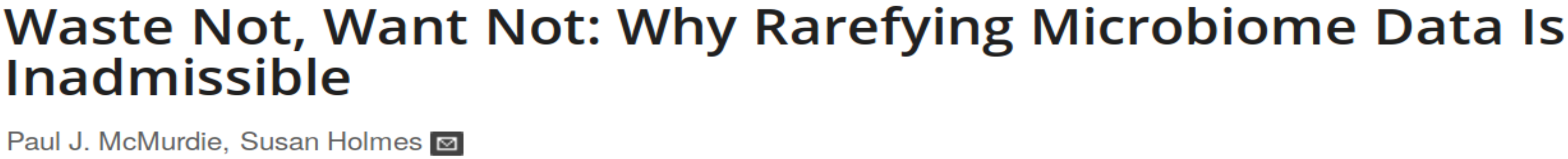
Vandeputte et al. Nature 551:507-511, 2017

# Normalizing library size?

If sample A has been sampled deeper than sample B, we the counts can be expected to be higher.

Compositional data: Divide by the total number of reads per sample (compositional abundance)

Problem: Abundant taxa may distort the ratios.

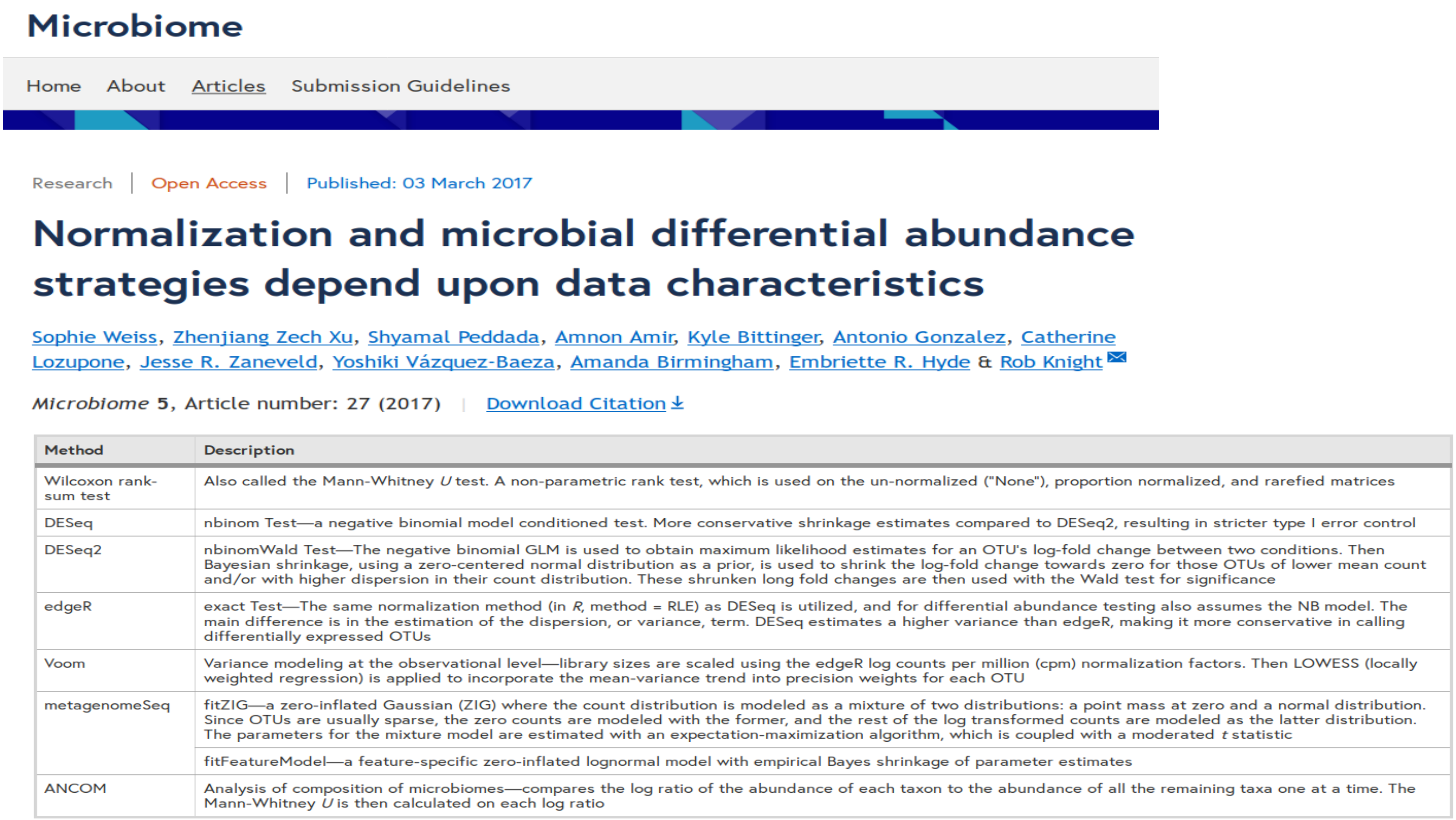


**Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible**

Paul J. McMurdie, Susan Holmes ✉

# Transformations

```
transform(x, “compositional”)
transform(x, “clr”)
transform(x, “log10p”)
transform(x, “hellinger”)
transform(x, “identity”)
```



Microbiome

Home About Articles Submission Guidelines

Research | Open Access | Published: 03 March 2017

# Normalization and microbial differential abundance strategies depend upon data characteristics

Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde & Rob Knight

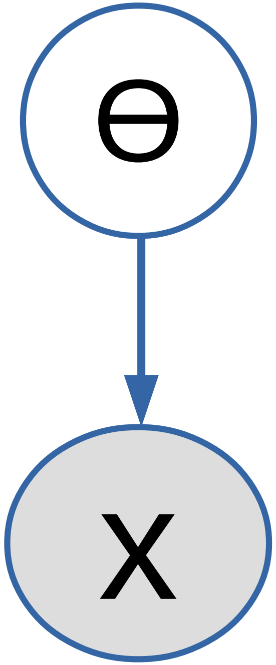
*Microbiome* **5**, Article number: 27 (2017) | Download Citation

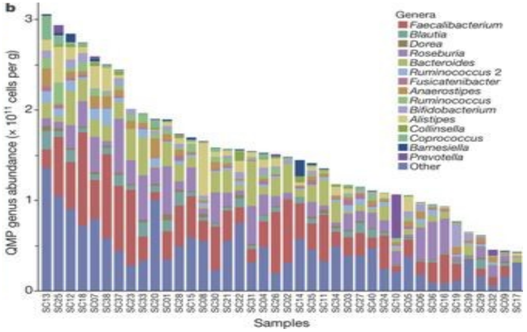
| Method | Description |
|---|---|
| Wilcoxon rank-sum test | Also called the Mann-Whitney *U* test. A non-parametric rank test, which is used on the un-normalized ("None"), proportion normalized, and rarefied matrices |
| DESeq | nbinom Test—a negative binomial model conditioned test. More conservative shrinkage estimates compared to DESeq2, resulting in stricter type I error control |
| DESeq2 | nbinomWald Test—The negative binomial GLM is used to obtain maximum likelihood estimates for an OTU's log-fold change between two conditions. Then Bayesian shrinkage, using a zero-centered normal distribution as a prior, is used to shrink the log-fold change towards zero for those OTUs of lower mean count and/or with higher dispersion in their count distribution. These shrunken long fold changes are then used with the Wald test for significance |
| edgeR | exact Test—The same normalization method (in *R*, method = RLE) as DESeq is utilized, and for differential abundance testing also assumes the NB model. The main difference is in the estimation of the dispersion, or variance, term. DESeq estimates a higher variance than edgeR, making it more conservative in calling differentially expressed OTUs |
| Voom | Variance modeling at the observational level—library sizes are scaled using the edgeR log counts per million (cpm) normalization factors. Then LOWESS (locally weighted regression) is applied to incorporate the mean-variance trend into precision weights for each OTU |
| metagenomeSeq | fitZIG—a zero-inflated Gaussian (ZIG) where the count distribution is modeled as a mixture of two distributions: a point mass at zero and a normal distribution. Since OTUs are usually sparse, the zero counts are modeled with the former, and the rest of the log transformed counts are modeled as the latter distribution. The parameters for the mixture model are estimated with an expectation-maximization algorithm, which is coupled with a moderated *t* statistic |
| | fitFeatureModel—a feature-specific zero-inflated lognormal model with empirical Bayes shrinkage of parameter estimates |
| ANCOM | Analysis of composition of microbiomes—compares the log ratio of the abundance of each taxon to the abundance of all the remaining taxa one at a time. The Mann-Whitney *U* is then calculated on each log ratio |

# Data is not compositional!

Model

Observations (Data)





# State diagnosis & manipulation:
# from specific targets to the overall ecosystem

Diet

Life style

Antibiotics

Probiotics

Prebiotics

Fecal transplants

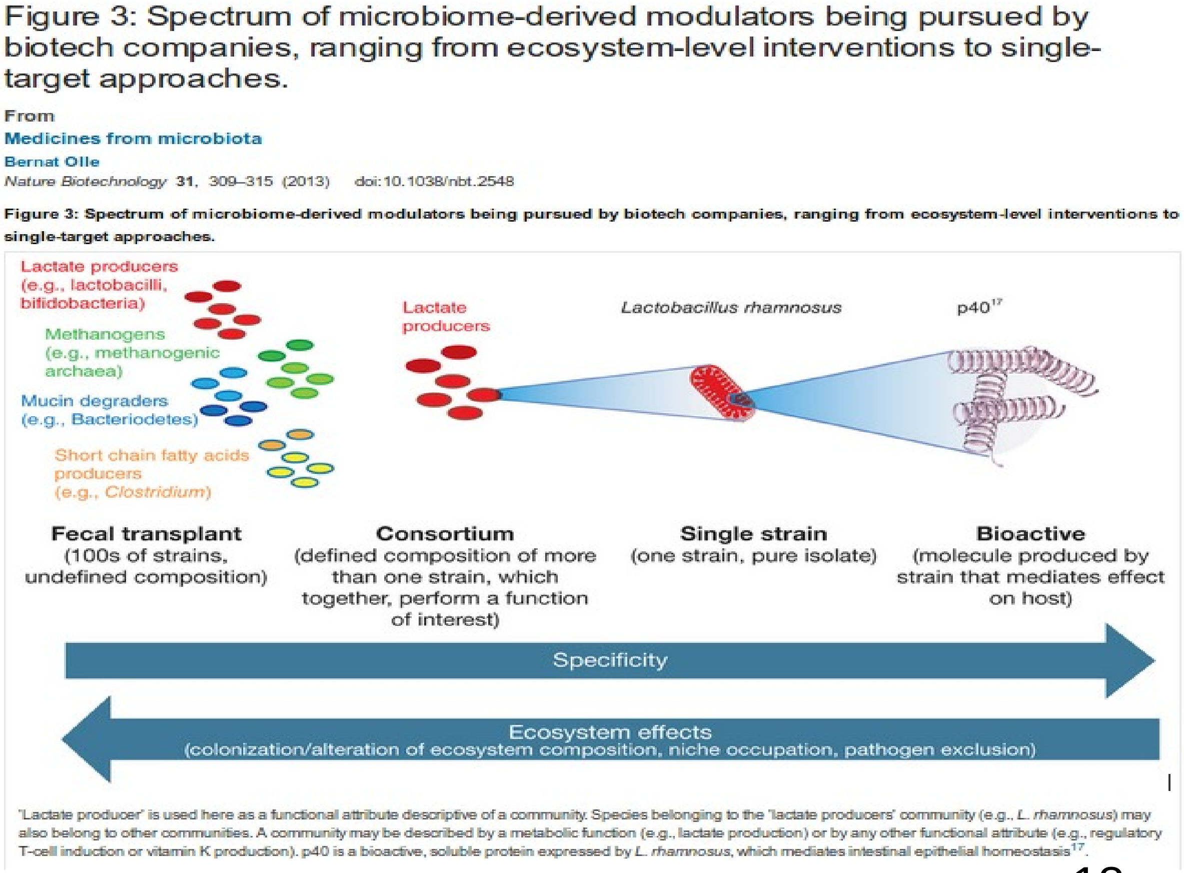


Figure 3: Spectrum of microbiome-derived modulators being pursued by biotech companies, ranging from ecosystem-level interventions to single-target approaches.

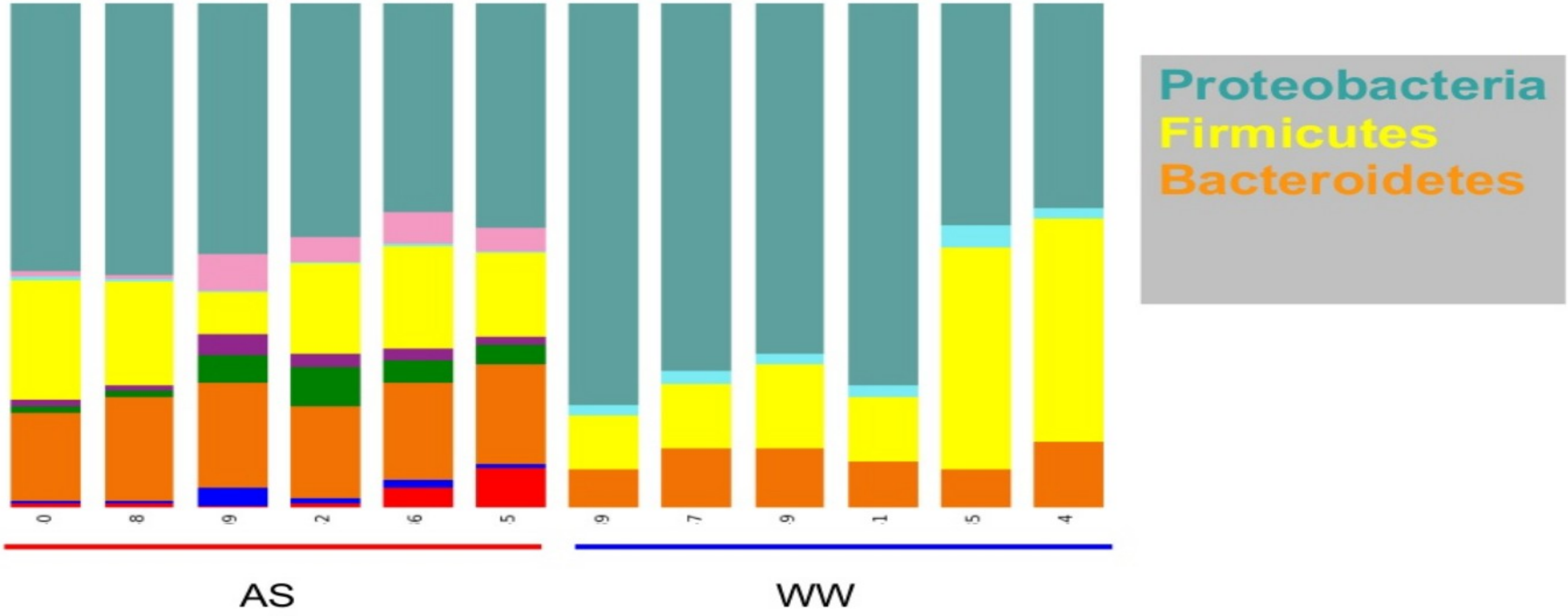
From
Medicines from microbiota
Bernat Olle
*Nature Biotechnology* **31**, 309–315 (2013) doi:10.1038/nbt.2548

**Figure 3: Spectrum of microbiome-derived modulators being pursued by biotech companies, ranging from ecosystem-level interventions to single-target approaches.**

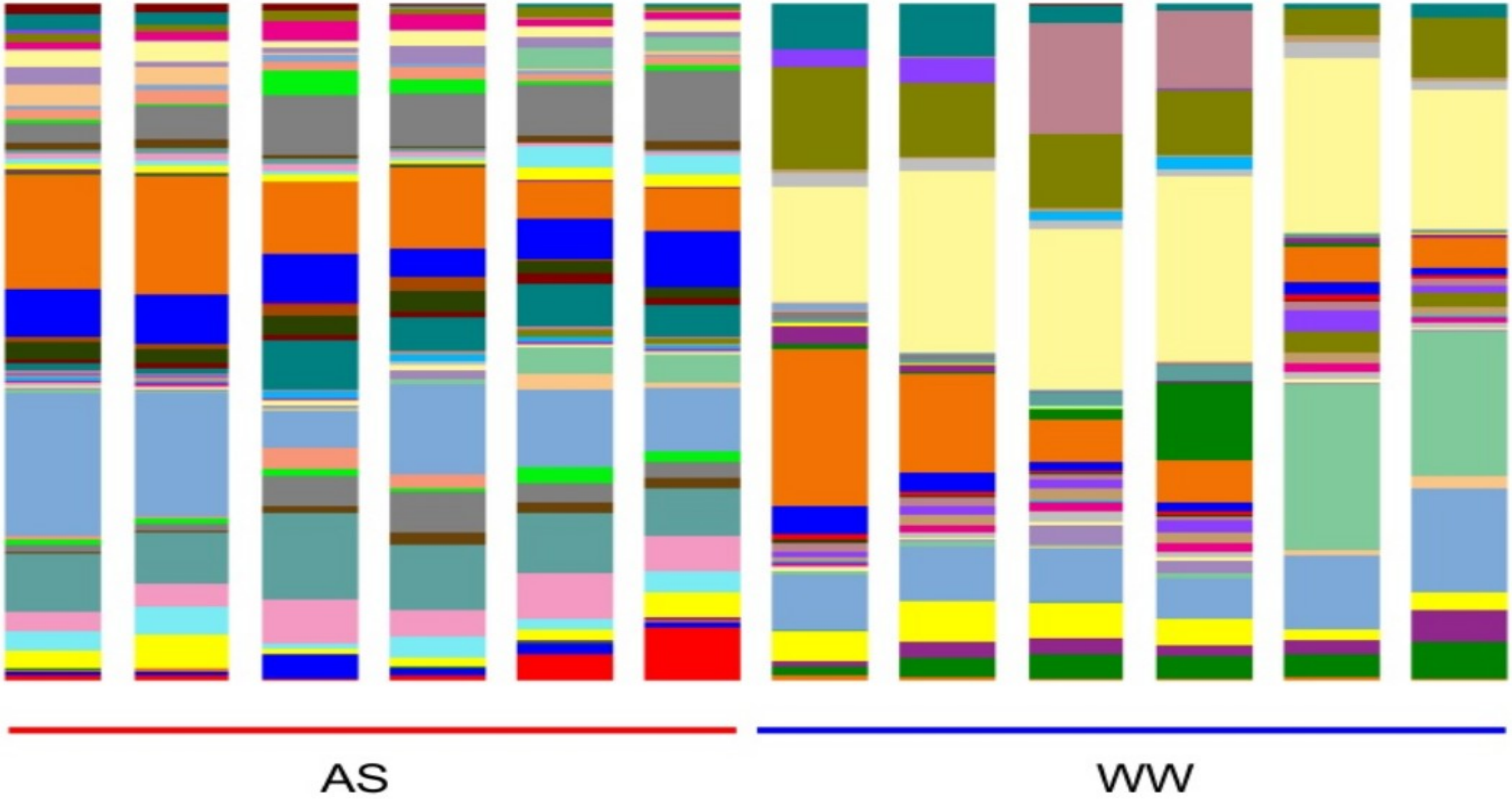
Lactate producers (e.g., lactobacilli, bifidobacteria)
Methanogens (e.g., methanogenic archaea)
Mucin degraders (e.g., Bacteriodetes)
Short chain fatty acids producers (e.g., *Clostridium*)
Lactate producers
*Lactobacillus rhamnosus*
p40[17]
**Fecal transplant**
(100s of strains, undefined composition)
**Consortium**
(defined composition of more than one strain, which together, perform a function of interest)
**Single strain**
(one strain, pure isolate)
**Bioactive**
(molecule produced by strain that mediates effect on host)
Specificity
Ecosystem effects
(colonization/alteration of ecosystem composition, niche occupation, pathogen exclusion)

'Lactate producer' is used here as a functional attribute descriptive of a community. Species belonging to the 'lactate producers' community (e.g., *L. rhamnosus*) may also belong to other communities. A community may be described by a metabolic function (e.g., lactate production) or by any other functional attribute (e.g., regulatory T-cell induction or vitamin K production). p40 is a bioactive, soluble protein expressed by *L. rhamnosus*, which mediates intestinal epithelial homeostasis[17].

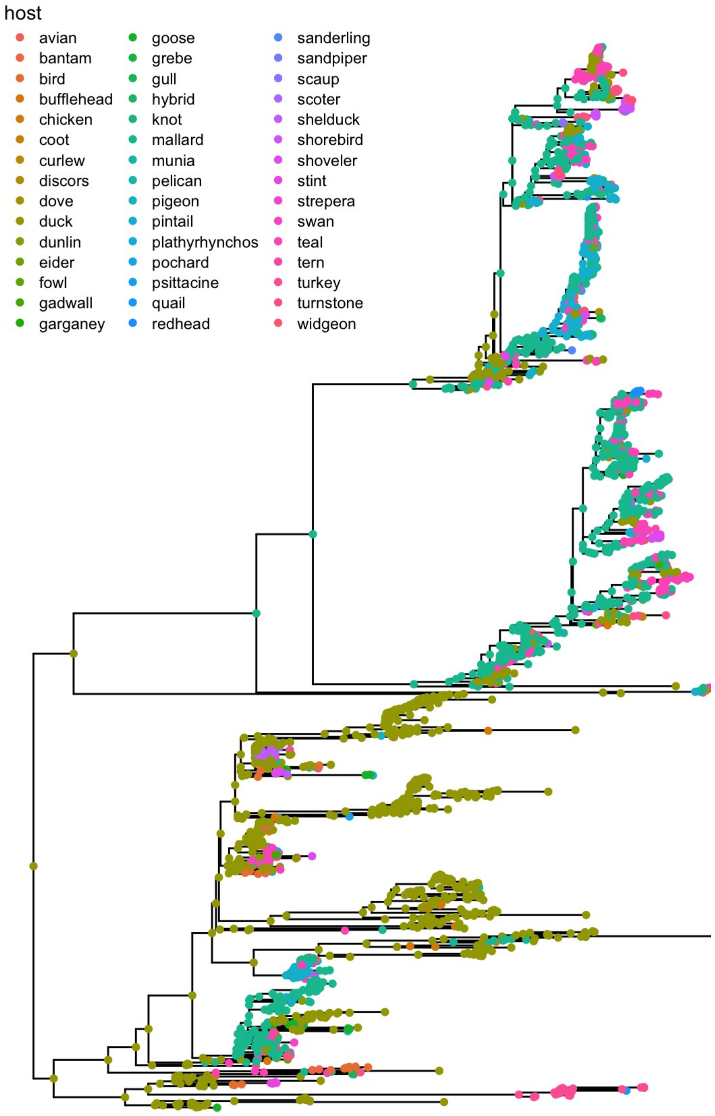
# Phylum level Classification

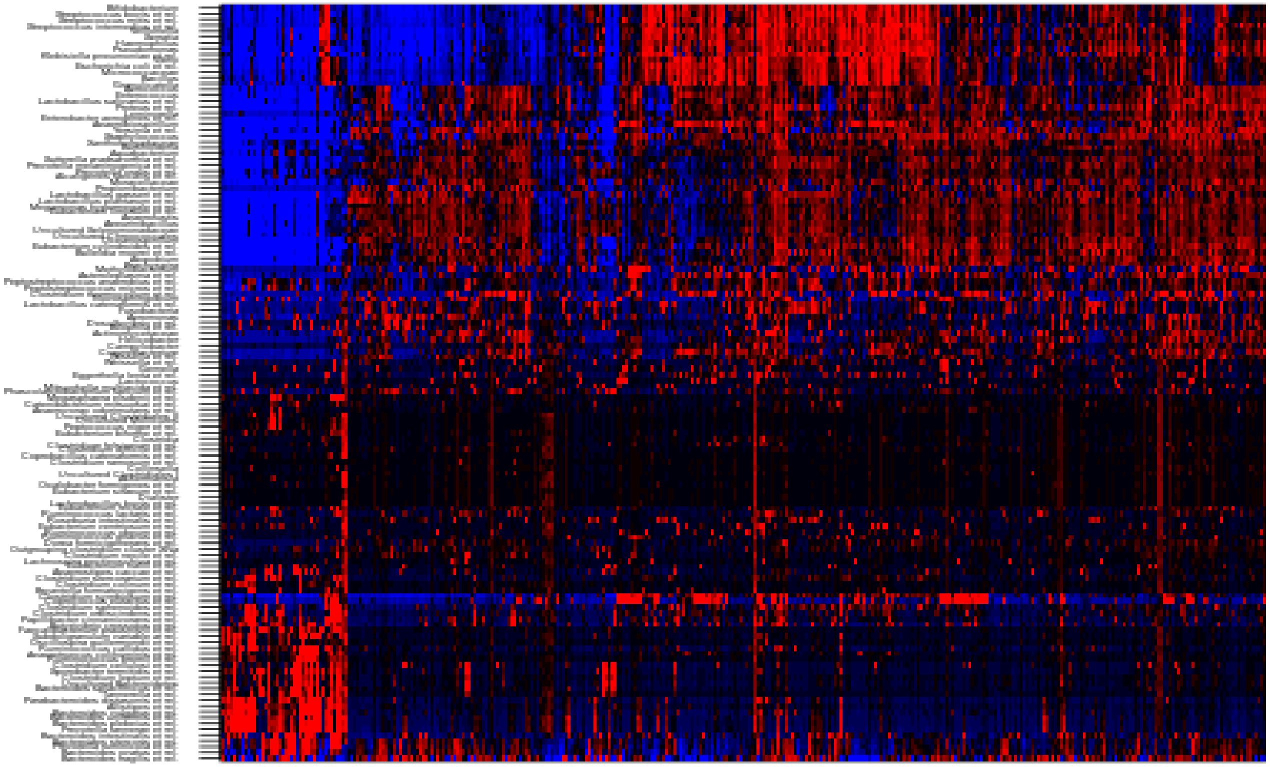


# Genus level Classification



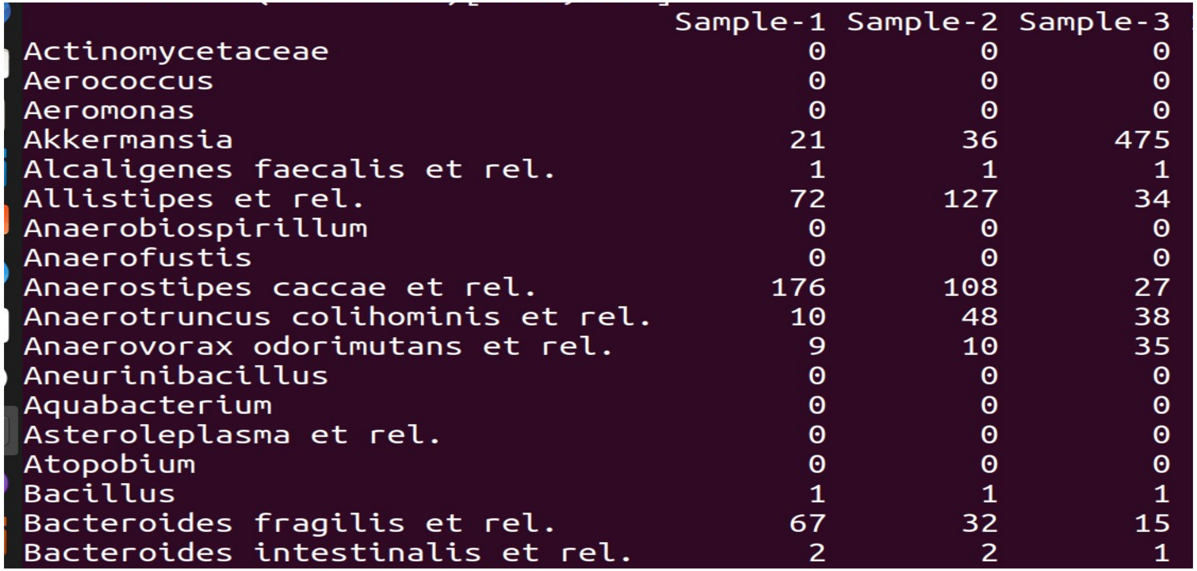
# Phylogenetic trees



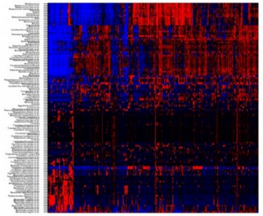


# Abundance matrix

- Sparse
- Non-Gaussian
- Overdispersed
- Compositional
- Complex
- Stochastic
- Hierarchical

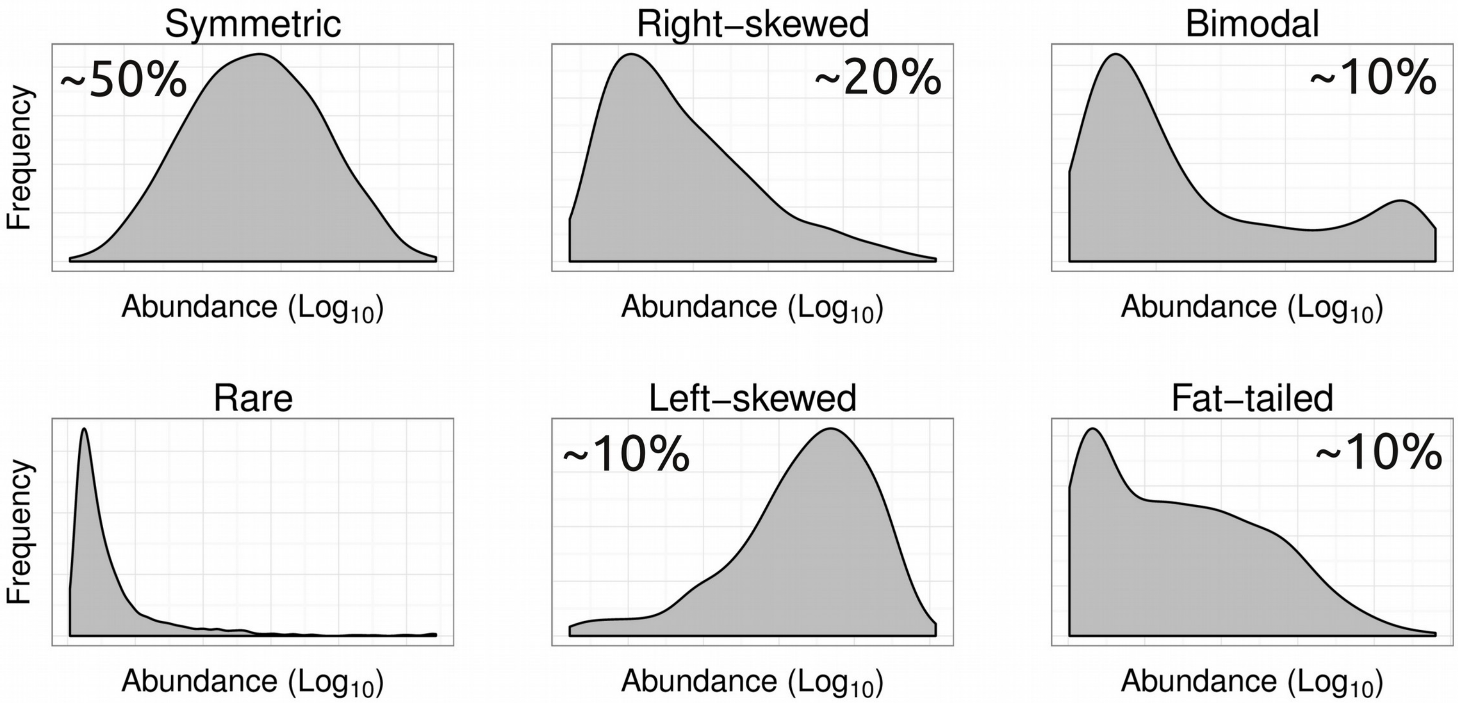


```
                                   Sample-1 Sample-2 Sample-3
Actinomycetaceae                          0        0        0
Aerococcus                                0        0        0
Aeromonas                                 0        0        0
Akkermansia                              21       36      475
Alcaligenes faecalis et rel.              1        1        1
Allistipes et rel.                       72      127       34
Anaerobiospirillum                        0        0        0
Anaerofustis                              0        0        0
Anaerostipes caccae et rel.             176      108       27
Anaerotruncus colihominis et rel.        10       48       38
Anaerovorax odorimutans et rel.           9       10       35
Aneurinibacillus                          0        0        0
Aquabacterium                             0        0        0
Asteroleplasma et rel.                    0        0        0
Atopobium                                 0        0        0
Bacillus                                  1        1        1
Bacteroides fragilis et rel.             67       32       15
Bacteroides intestinalis et rel.          2        2        1
```

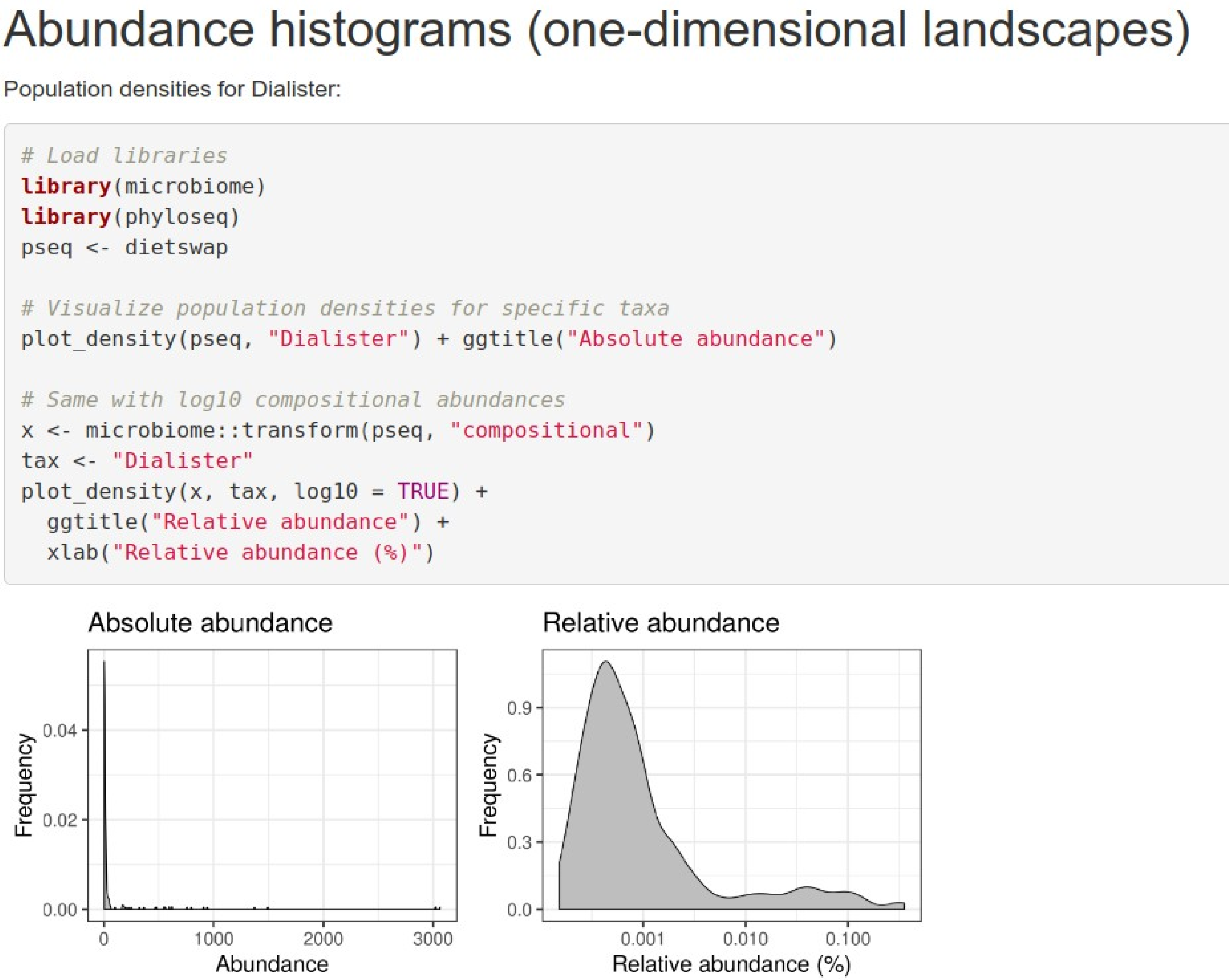


# Bacterial 'abundance types' in 1000 western adults:

~% indicates proportion among prevalent taxa



Lahti et al. Nat. Comm. 5:4344, 2014



# Abundance histograms (one-dimensional landscapes)

Population densities for Dialister:

```
# Load libraries
library(microbiome)
library(phyloseq)
pseq <- dietswap

# Visualize population densities for specific taxa
plot_density(pseq, "Dialister") + ggtitle("Absolute abundance")

# Same with log10 compositional abundances
x <- microbiome::transform(pseq, "compositional")
tax <- "Dialister"
plot_density(x, tax, log10 = TRUE) +
  ggtitle("Relative abundance") +
  xlab("Relative abundance (%)")
```

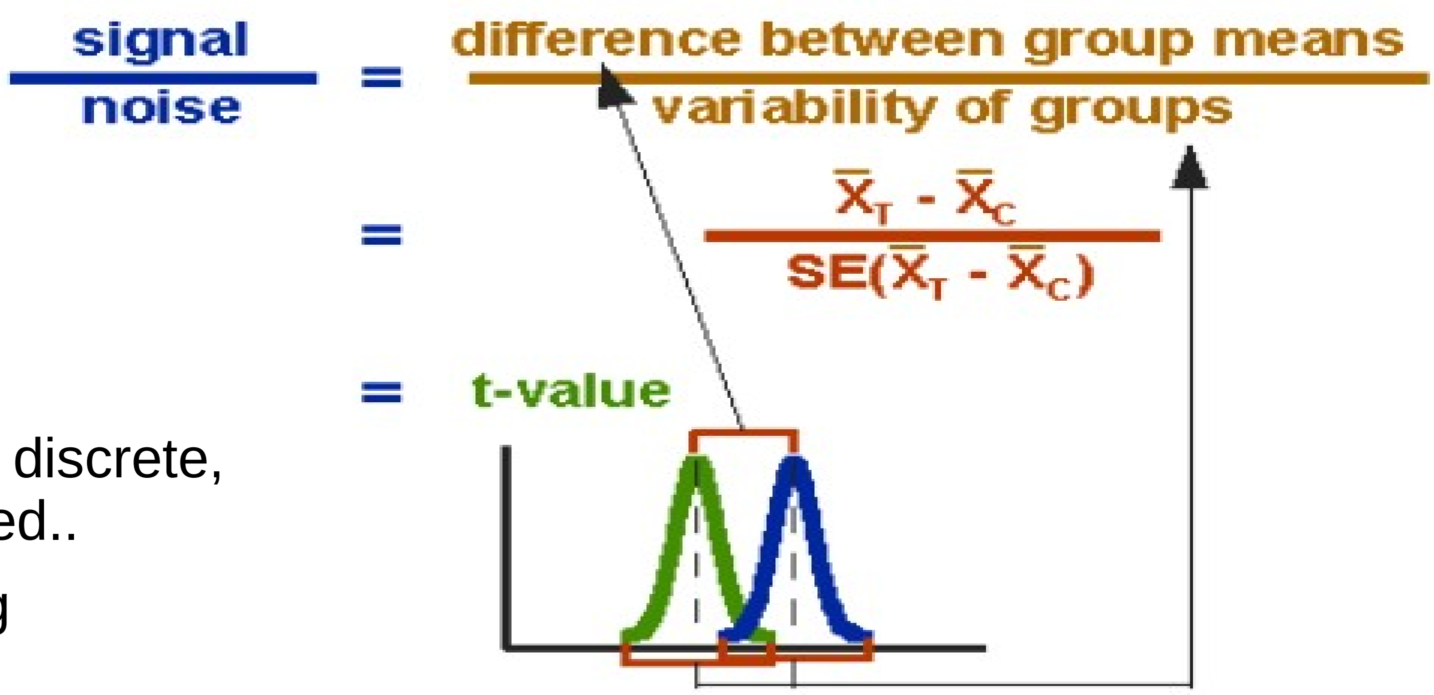
Absolute abundance

Relative abundance

# Standard t-test for two-group comparison?

<u>Problems</u>:

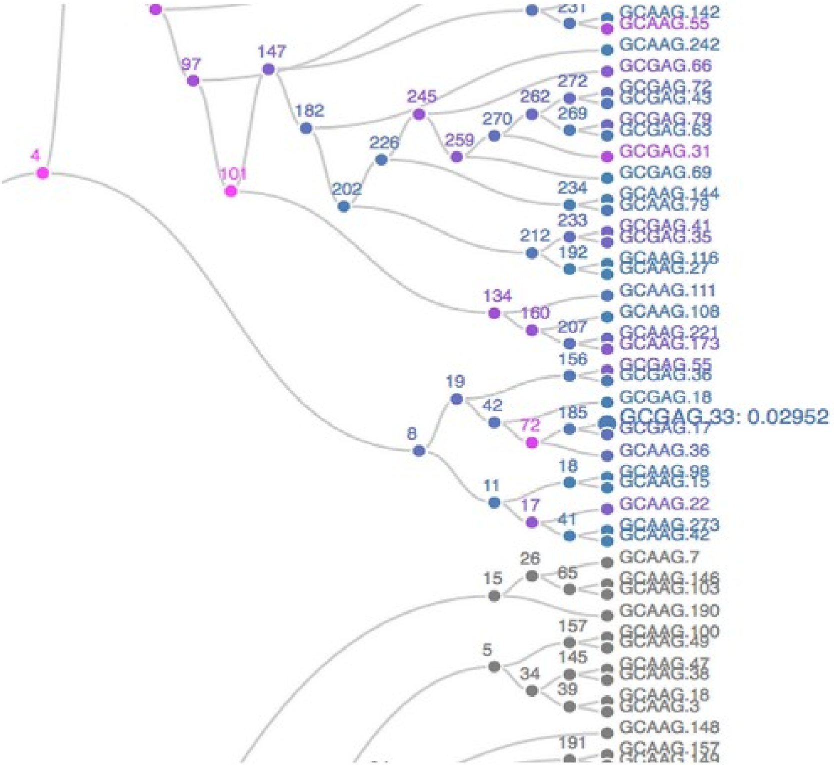
- Few replicates
- Non-gaussian, discrete, positive, skewed..
- Multiple testing



$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}} = \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)} = \text{t-value}$$

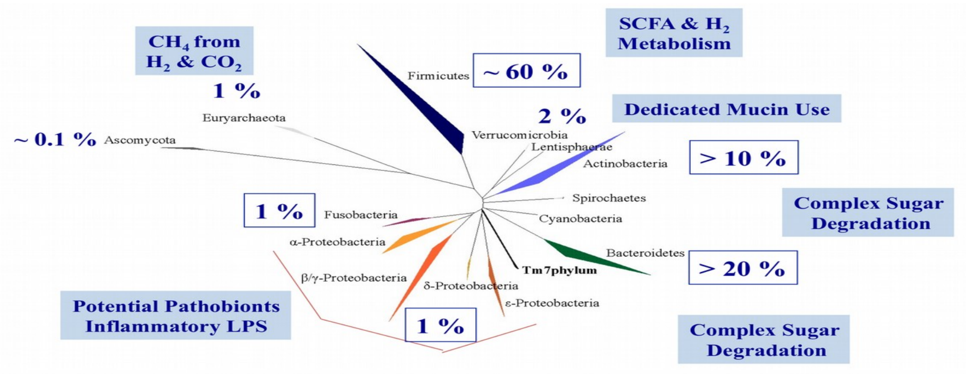
http://www.socialresearchmethods.net/kb/stat_t.php

# Hierarchical testing (Kris Sankaran)



Taking account of the phylogenetic tree when testing:

- CRAN package: structSSI
- Journ. Stat. Software paper JSS link

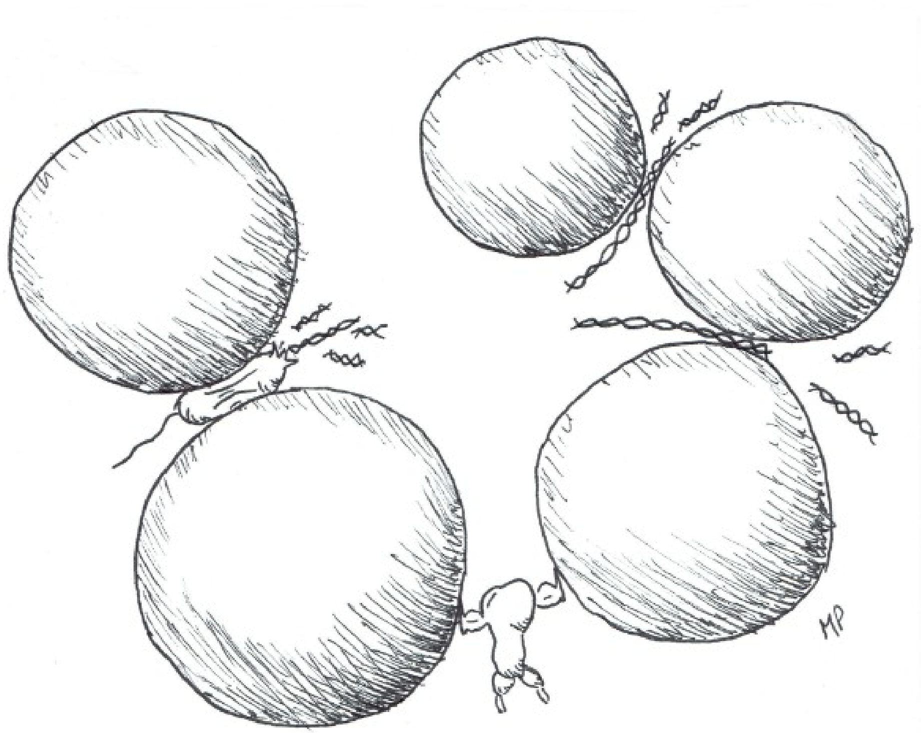


Tree-based methods

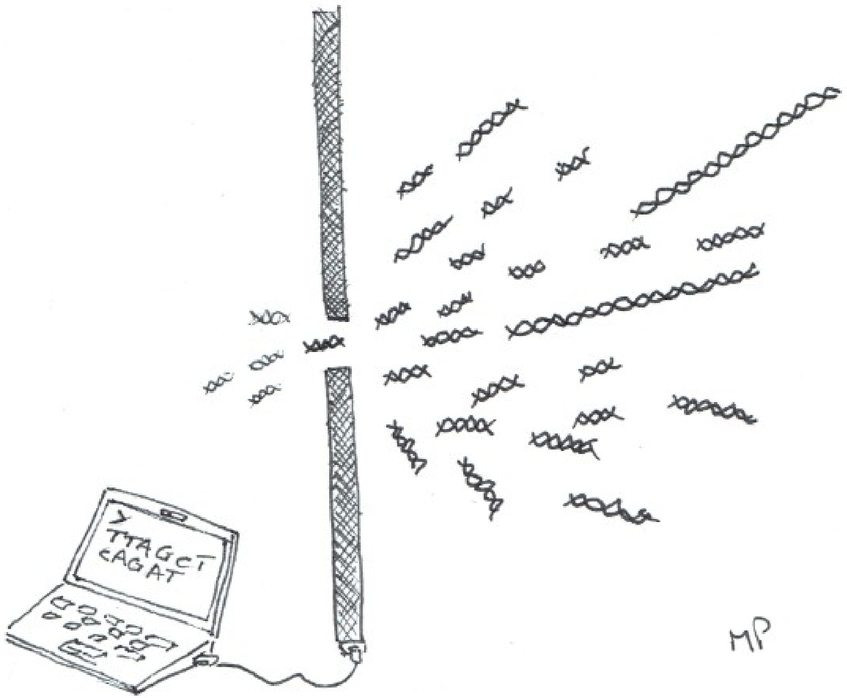
- StructSSI
- phylofactor
- tree-PCA
- UniFrac

Source: Susan Holmes | http://web.stanford.edu/class/bios221/Short-Phyloseq-Resources.html

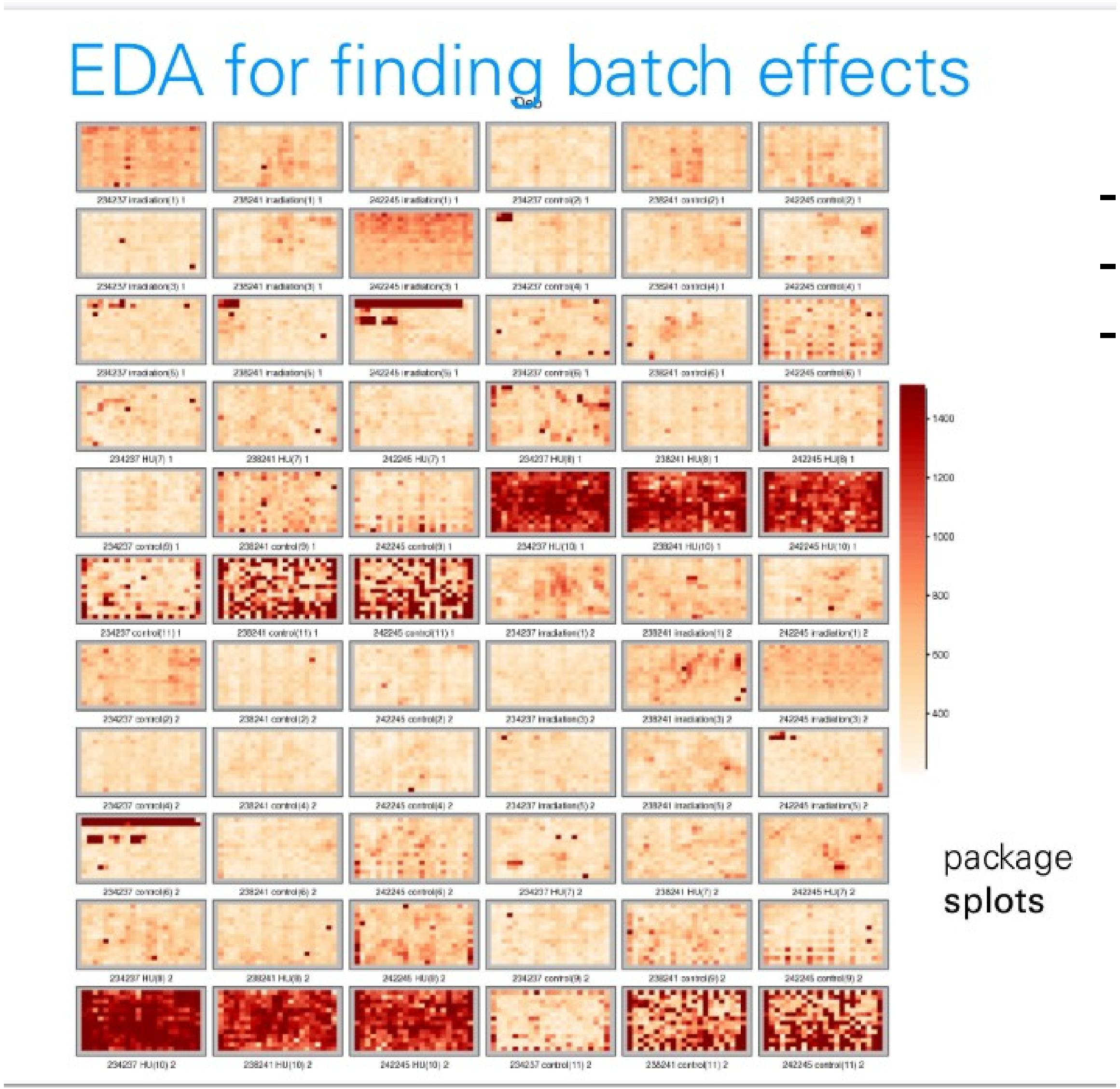
## Biased cell lysis



## Biased sequencing



https://github.com/mblstamps/stamps2019/blob/master/STAMPS2019_overview_Pop.pdf



# EDA for finding batch effects

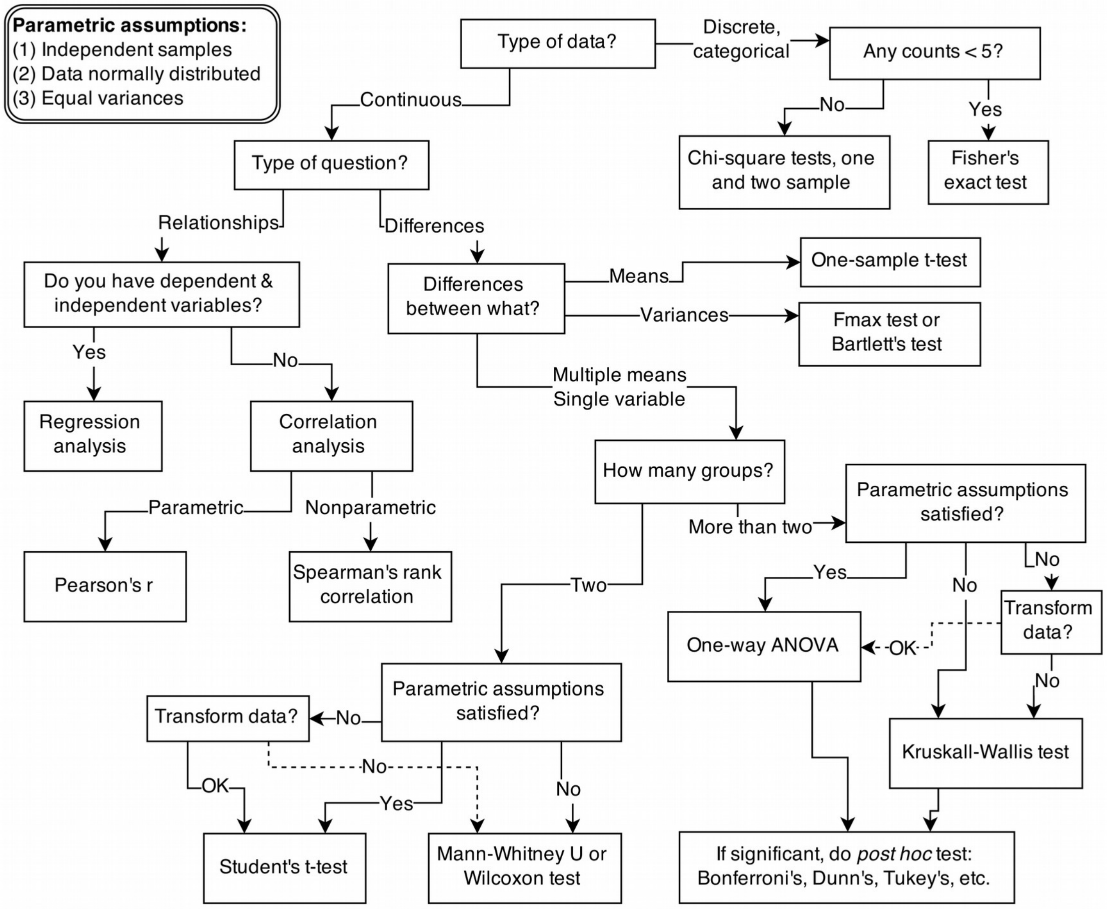
package
**splots**

- negative controls
- positive controls
- batch..

# Statistical aspects: summary

- Biased
- Sparse
- Non-Gaussian
- Overdispersed
- Compositional
- Complex
- Stochastic
- Hierarchical

# How to choose a correct model?



**Parametric assumptions:**
(1) Independent samples
(2) Data normally distributed
(3) Equal variances

Type of data?

Continuous

Discrete, categorical

Any counts < 5?

No

Yes

Chi-square tests, one and two sample

Fisher's exact test

Type of question?

Relationships

Differences

Do you have dependent & independent variables?

Yes

No

Regression analysis

Correlation analysis

Parametric

Nonparametric

Pearson's r

Spearman's rank correlation

Differences between what?

Means

Variances

One-sample t-test

Fmax test or Bartlett's test

Multiple means Single variable

How many groups?

Two

More than two

Parametric assumptions satisfied?

Yes

No

No

Transform data?

OK

No

One-way ANOVA

Kruskall-Wallis test

If significant, do *post hoc* test: Bonferroni's, Dunn's, Tukey's, etc.

Parametric assumptions satisfied?

No

Yes

No

Transform data?

OK

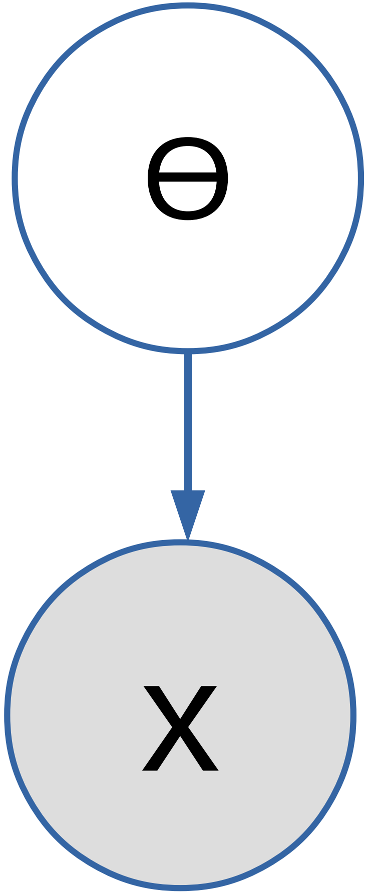
No

Student's t-test

Mann-Whitney U or Wilcoxon test

# Generative models

Model



Observations
(Data)

# Generative models

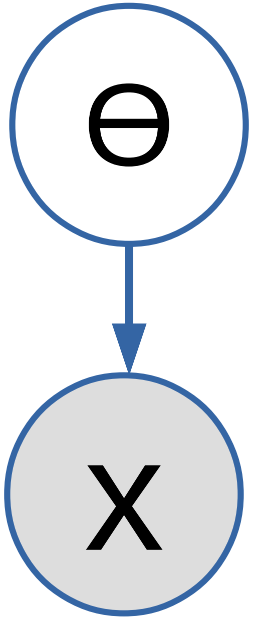
**Construct a model**

- Incorporate prior knowledge
- Learn the model with some data

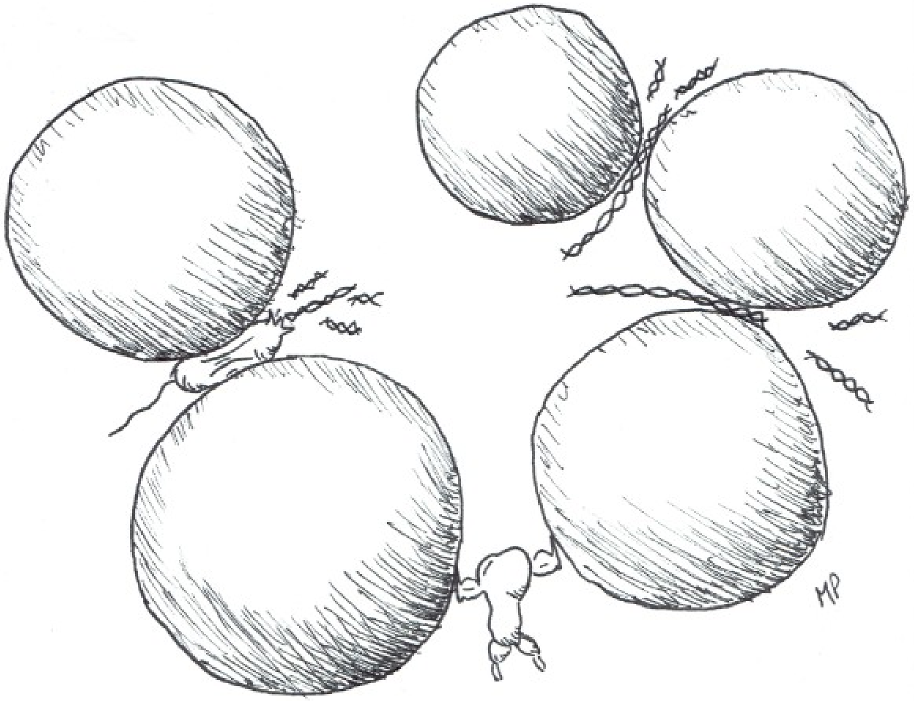
**Criticize the model**

- Generate artificial data
- Compare to real data
- Revise the model
- Regularize overfitting!

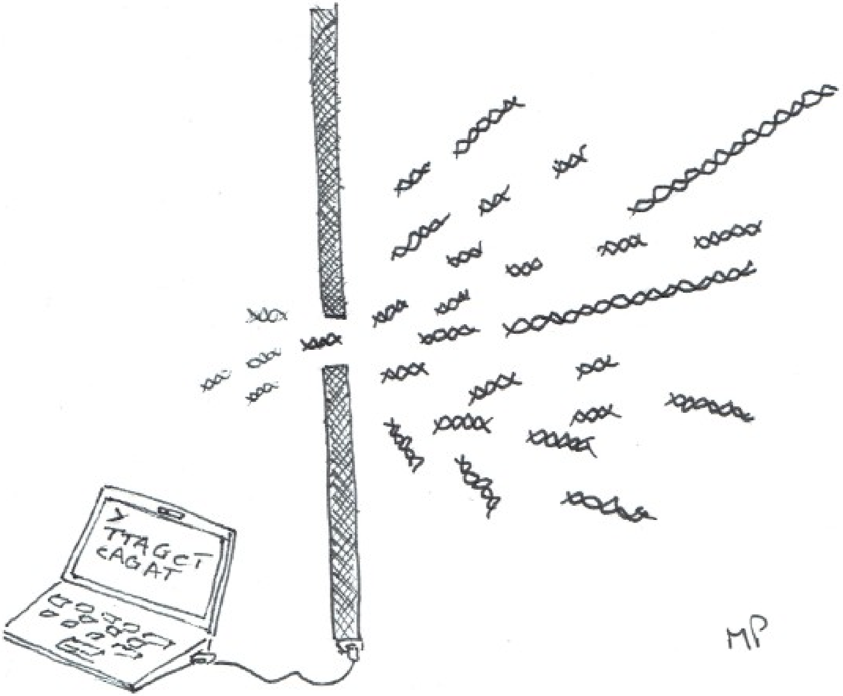
**Validate the model**



## Biased cell lysis



## Biased sequencing

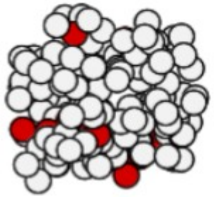


https://github.com/mblstamps/stamps2019/blob/master/STAMPS2019_overview_Pop.pdf

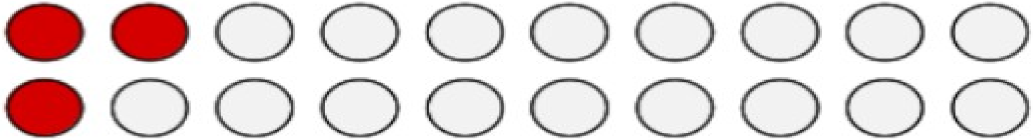
# The Poisson distribution

- This bag contains very many small balls, 10% of which are red.

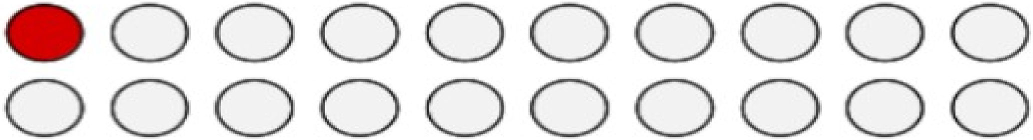
- Several experimenters are tasked with determining the percentage of red balls.



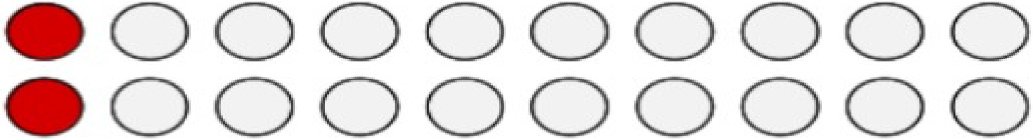
- Each of them is permitted to draw 20 balls out of the bag, without looking.



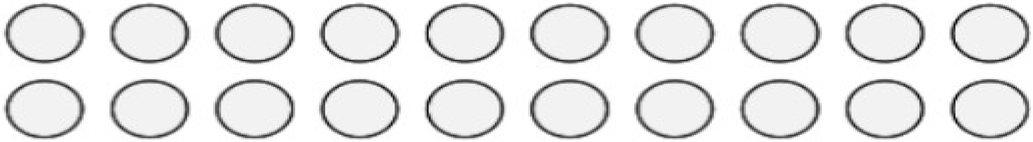
3 / 20 = 15%



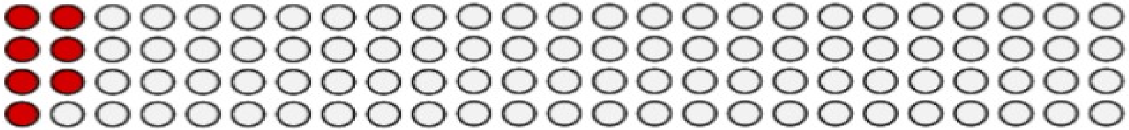
1 / 20 = 5%



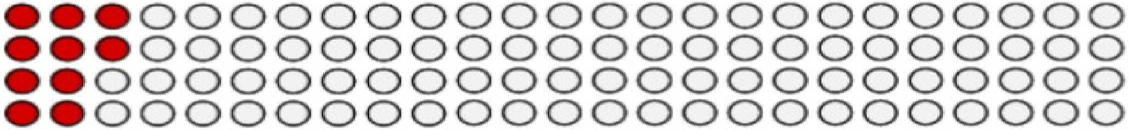
2 / 20 = 10%



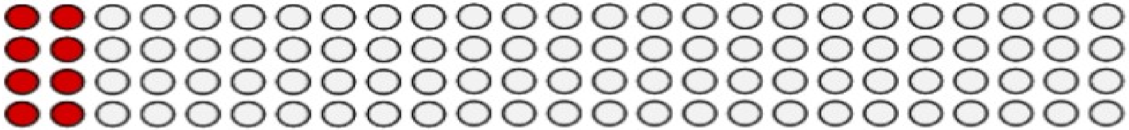
0 / 20 = 0%



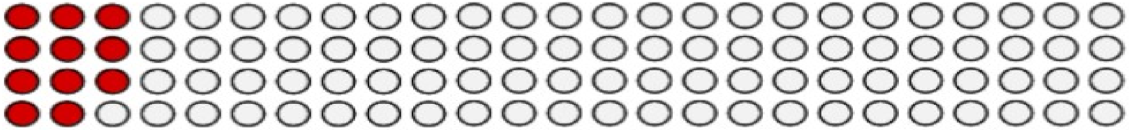
7 / 100 = 7%



10 / 100 = 10%



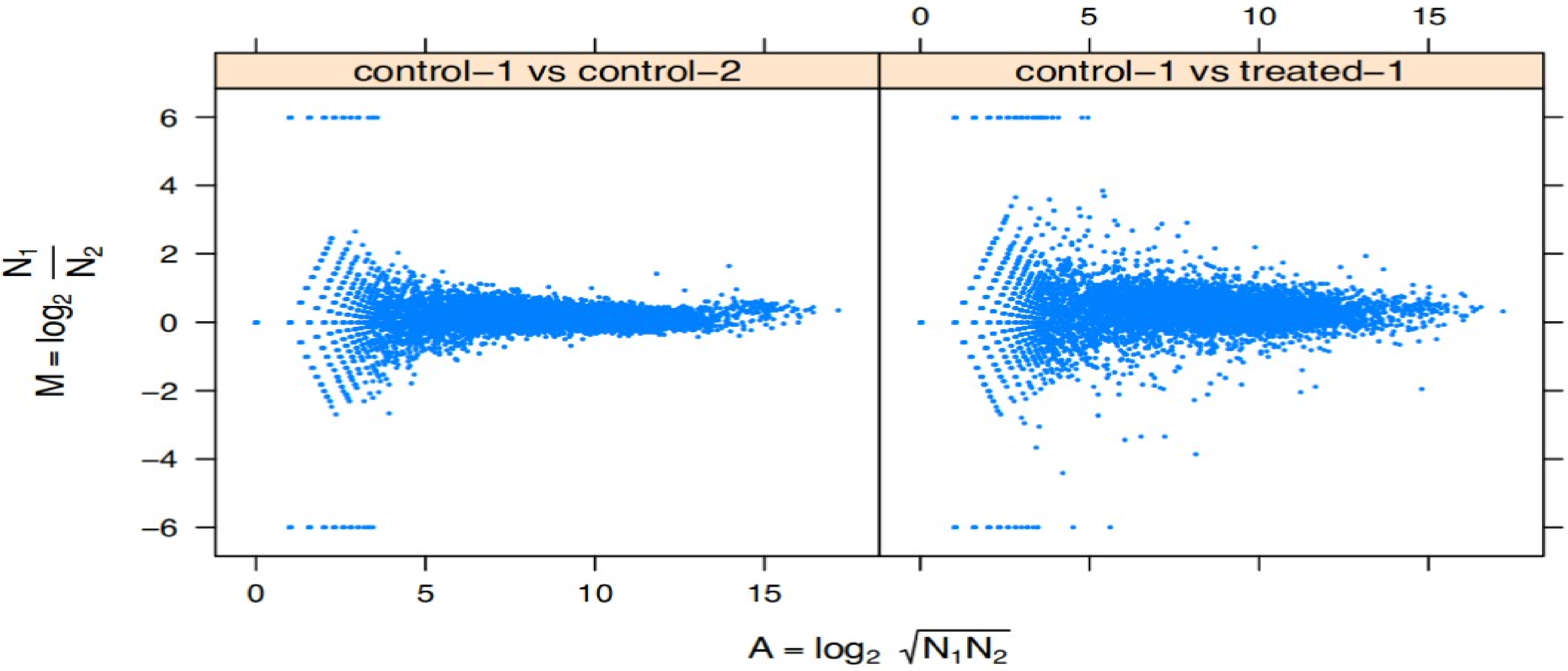
8 / 100 = 8%



11 / 100 = 11%

# Poisson distribution: Counting uncertainty

| expected number of red balls | standard deviation of number of red balls | relative error in estimate for the fraction of red balls |
|---:|---:|---:|
| 10 | √10 = 3 | 1 / √10 = 31.6% |
| 100 | √100 = 10 | 1 / √100 = 10.0% |
| 1,000 | √1,000 = 32 | 1 / √1000 = 3.2% |
| 10,000 | √10,000 = 100 | 1 / √10000 = 1.0% |



**two biological replicates**

**treatment vs control**

# Two component noise model

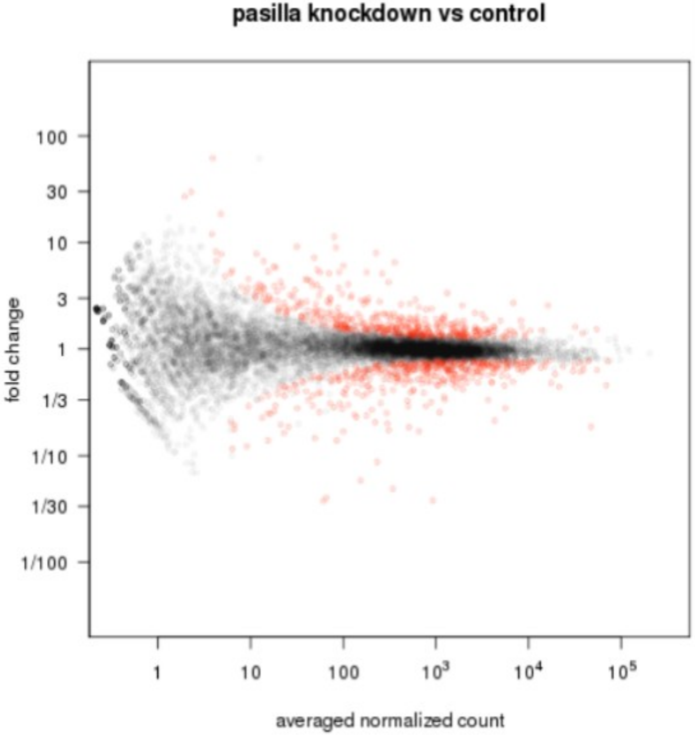
$$\mathrm{var} = \mu + c\,\mu^2$$

shot noise (Poisson) biological noise

## Small counts

Sampling noise dominant

Improve power: deeper coverage



## Large counts

Biological noise dominant

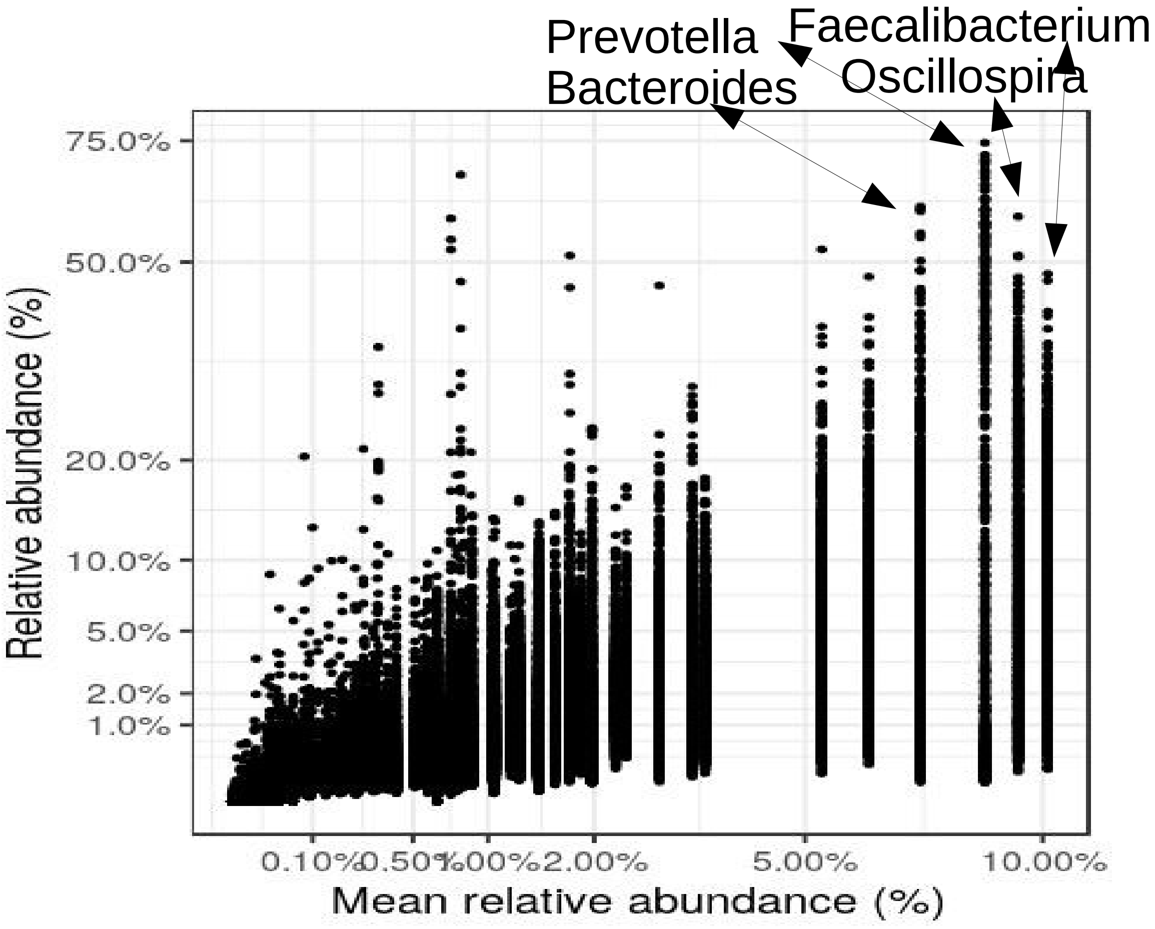
Improve power: more biol. replicates

# Taylor’s law (in HITChip Atlas)

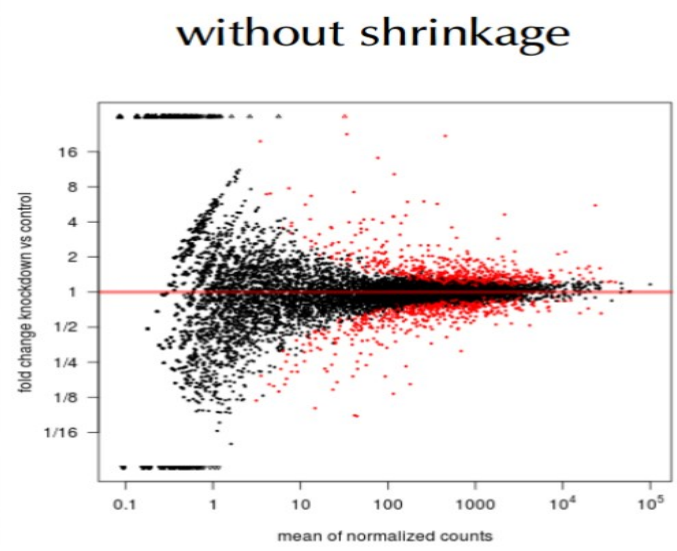
Heteroschedasticity: Variance increases with the mean

Overdispersion: Variance increases faster than proposed by the model

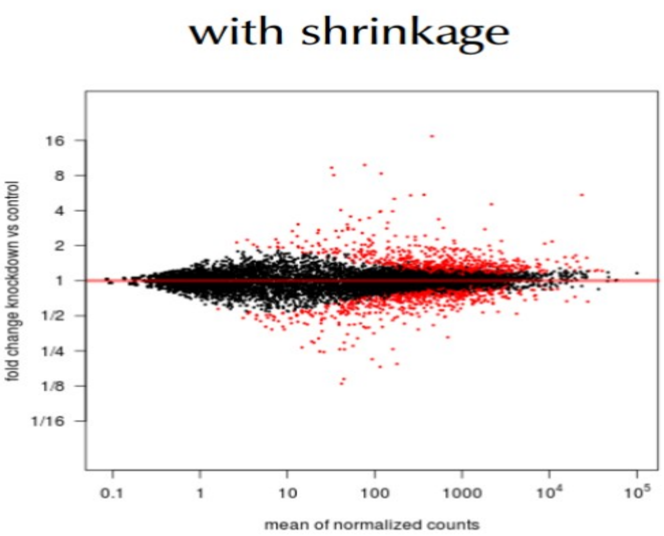
Data: HITChip Atlas



# Effect of shrinkage of log fold-change estimates



without shrinkage



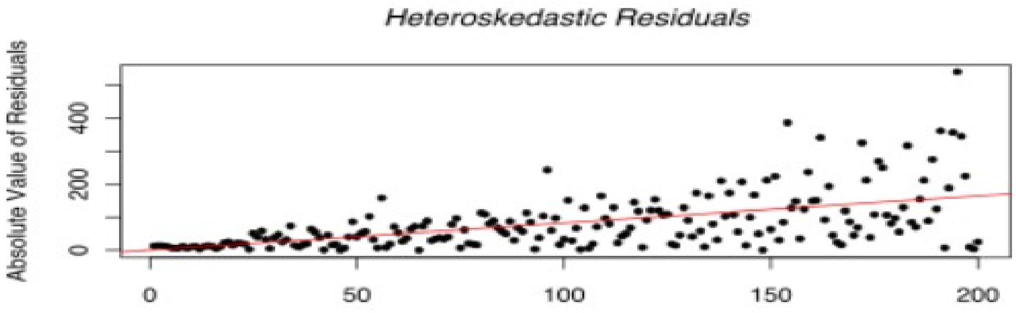
with shrinkage

Key assumption:

**Taxa with similar abundances have similar sample variances**

→ Variance can be estimated with a higher precision

# Dispersion and overdispersion



- Minimum variance of count data:

  $v = \mu$ (Poisson)

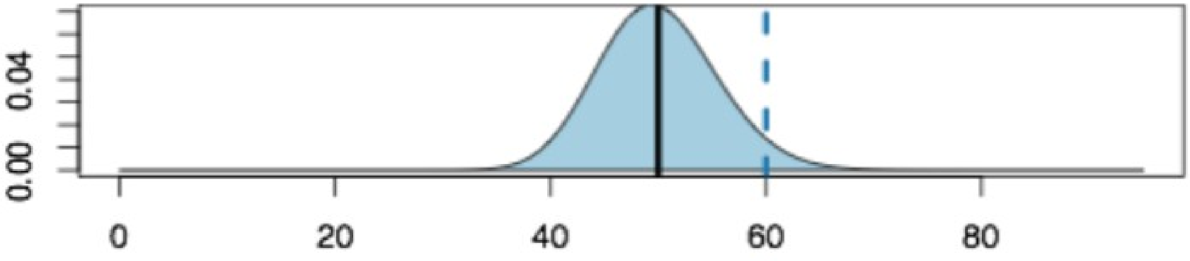
- Actual variance:

  $v = \mu + \alpha \mu^2$

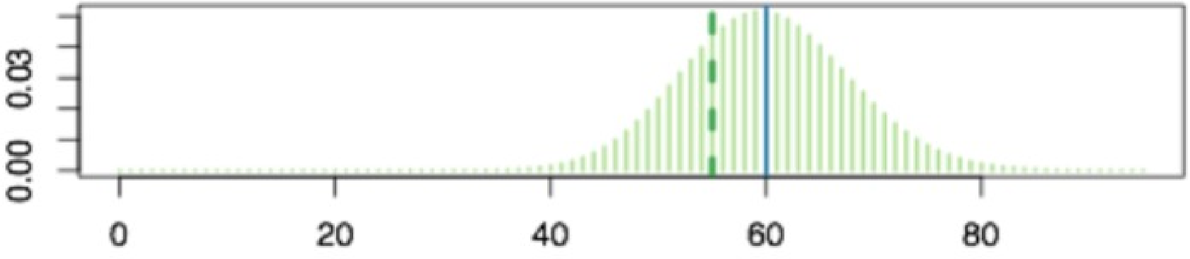
- $\alpha$ : "dispersion" $\alpha = (\mu - v) / \mu^2$

  (squared coefficient of variation of extra-Poisson variability)

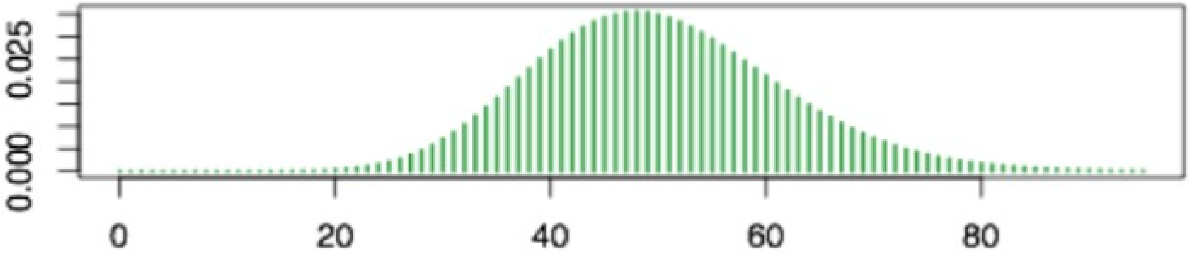
# The NB from a hierarchical model



Biological sample with mean $\mu$ and variance $v$



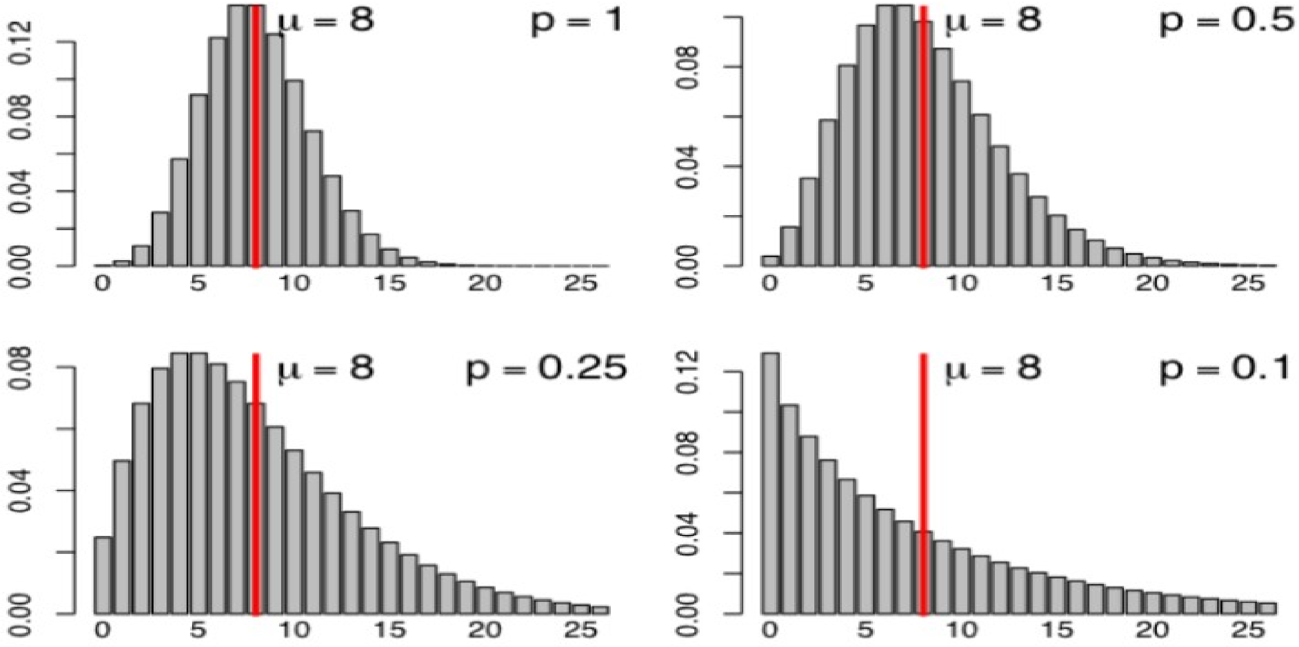
Poisson distribution with mean $q$ and variance $q$.



Negative binomial with mean $\mu$ and variance $q+v$

# The negative binomial distribution

A commonly used generalization of the Poisson distribution with *two* parameters



$$\Pr(Y = k) = \binom{k+r-1}{r-1} p^r (1-p)^k \quad \text{for } k = 0, 1, 2, \ldots$$