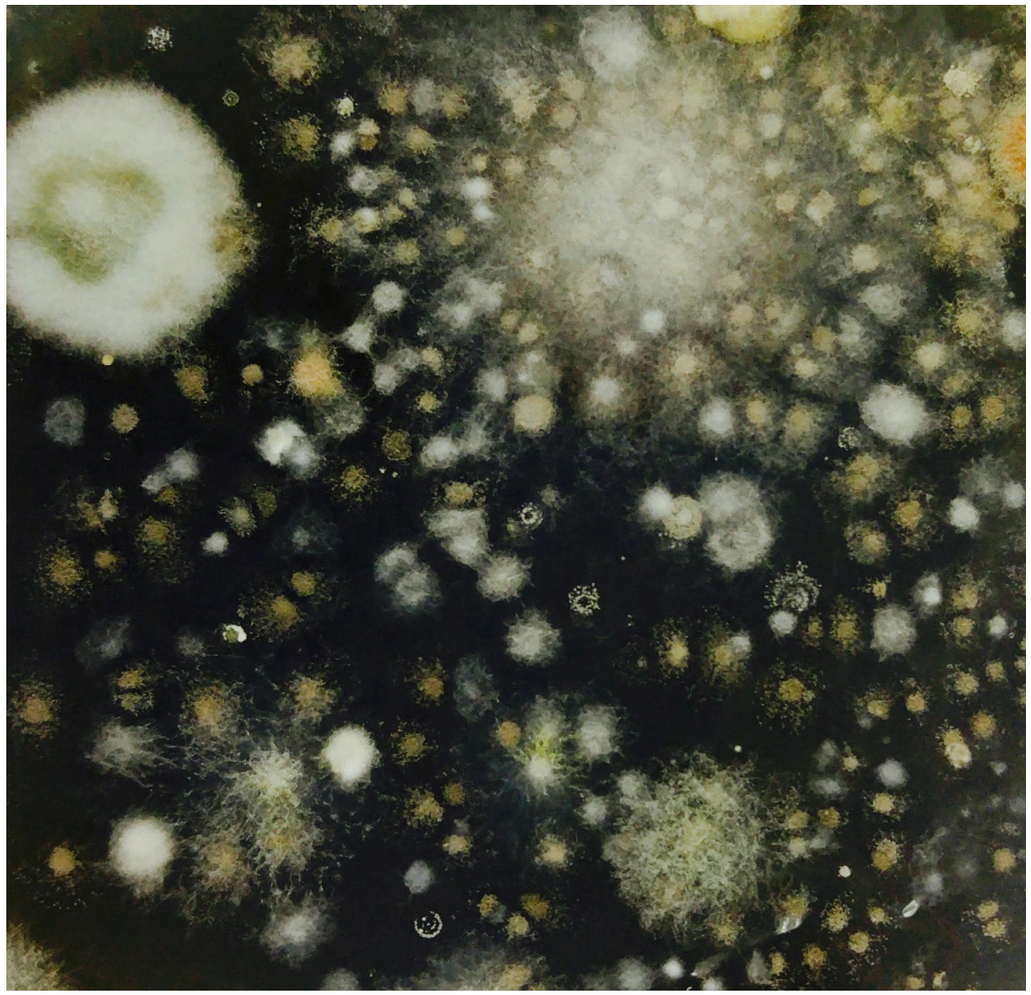


Modeling microbial abundances

Statistical Diversity Lab @ University of Washington

Amy Willis — @AmyDWillis — Associate Professor

Shirley Mathur — PhD Candidate

Sarah Teichman — PhD Candidate

María Valdez — PhD Candidate

Photo credit: T.D. Berry, Whitman lab, UW Madison

“How do I rigorously analyze my data?”

–Everyone, all the time

“It depends.”

–Stat Div Lab, all the time

Deciding on an analysis plan

- Your *scientific questions* should guide you in choosing your *analysis plan*
  - Many studies involve multiple analyses
  - These can answer *the same* or *different* questions
- What type of data you have may also constrain you

There is not **one** way to analyse your data!
You need to decide what is important to you!

Learning objectives

- Learning objectives

1. ~~Learn all the models~~

2. ~~Understand all their assumptions~~

3. ~~Resolve all confusion about statistical analysis of microbiome data~~

Learning objectives

- Learning objectives

    1. Learn *more* about *some* models
    2. Understand *some* of the *most important* assumptions and limitations of *some* methods
    3. Develop some facility using software to fit models
    4. Leave with more questions than ever

The plan

- Modeling with microbiome data
  - Abundance
    - 2 x lectures + 2 x labs
  - Diversity:
    - Lecture + lab
  - Experimental design
- Questions — throughout!

Now!

ask us about compositionality!

ask us about differential abundance!

Tomorrow morning!

ask us about rarefaction!

ask us about diversity metrics!

ask us about ordination!

ask us about replicates!

The pep talk

Modeling microbial abundances

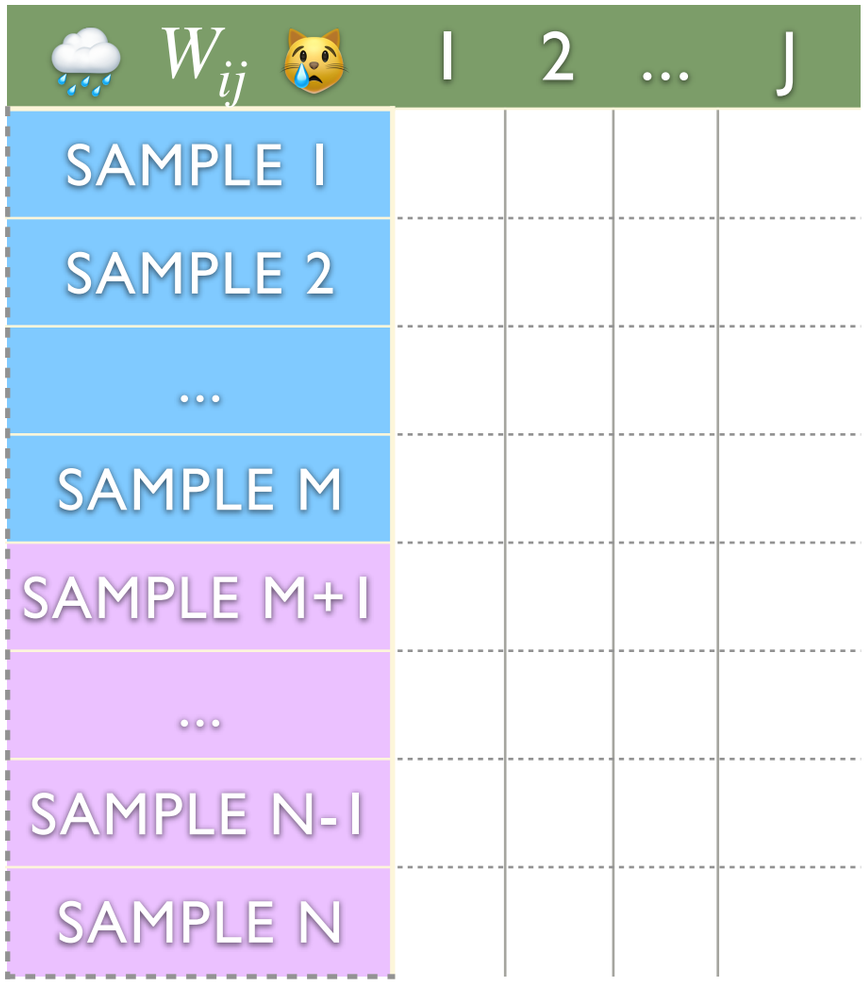
Data

- There are many different data/sequencing types that can be used to discuss “abundance”
  - amplicon - count tables
  - shotgun - coverage, proportion data…
  - qPCR / ddPCR - counts/concentrations…

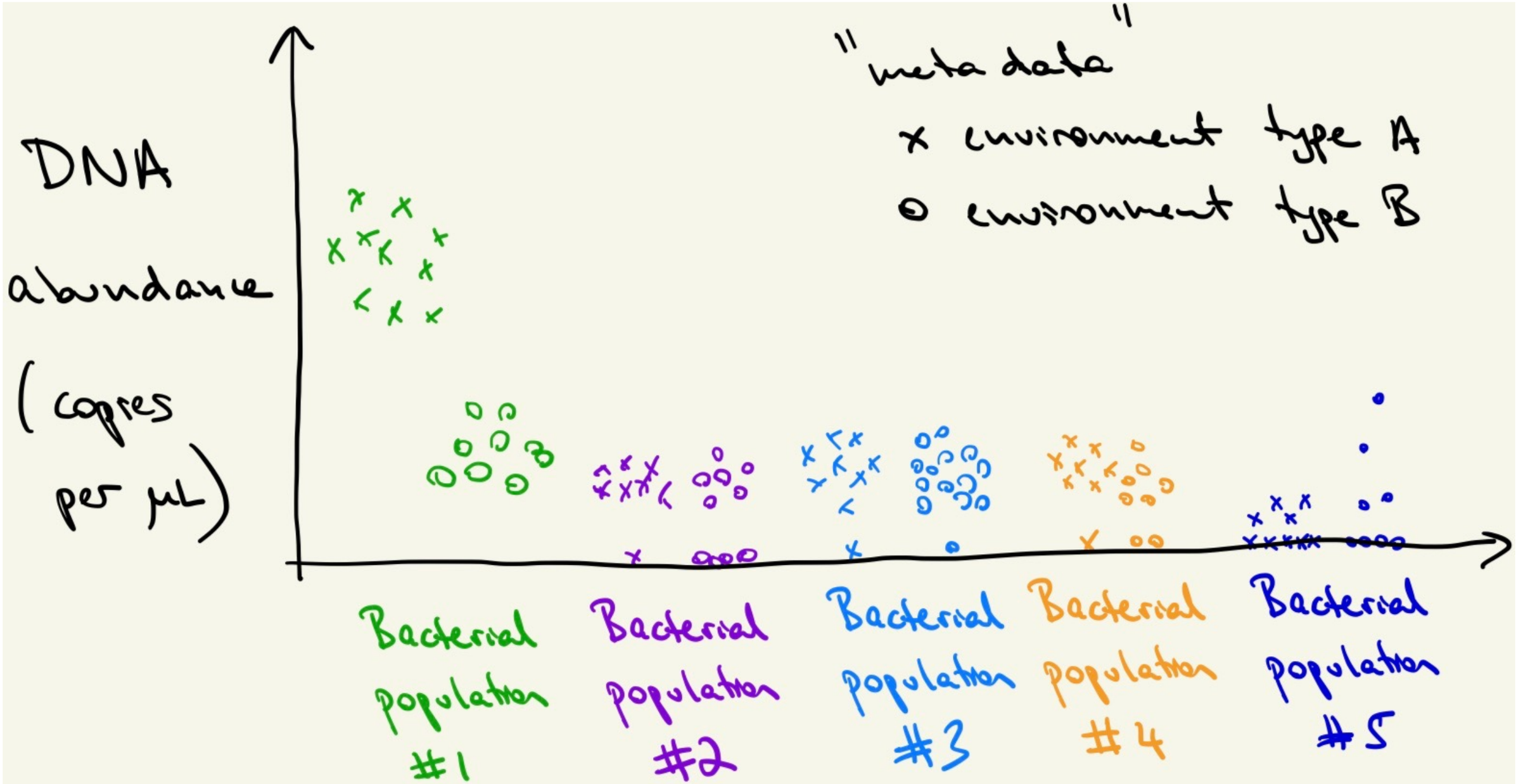
Data

- The *type of data* you have impacts the *approach you need*

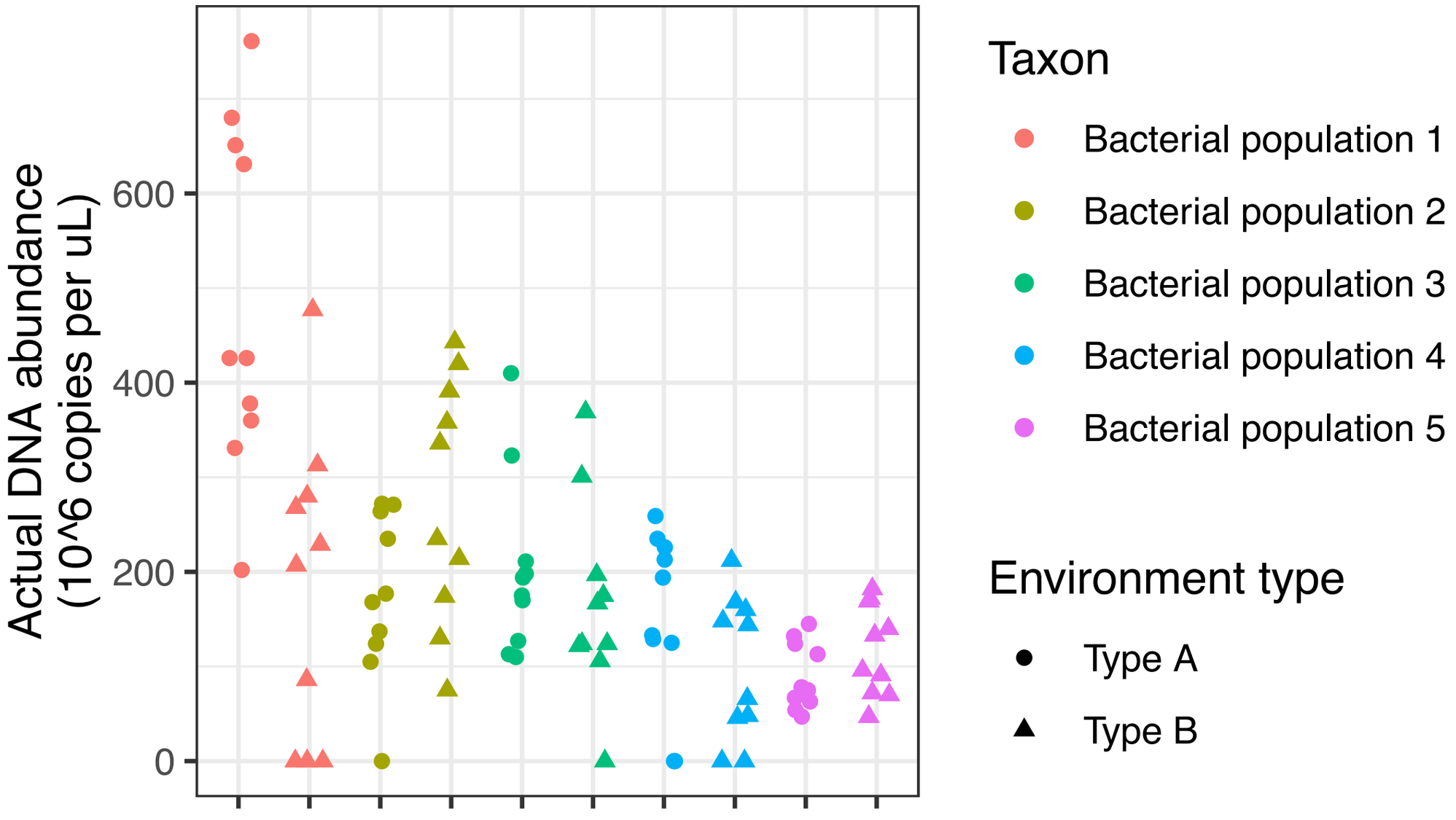
- You have to know the source of your data



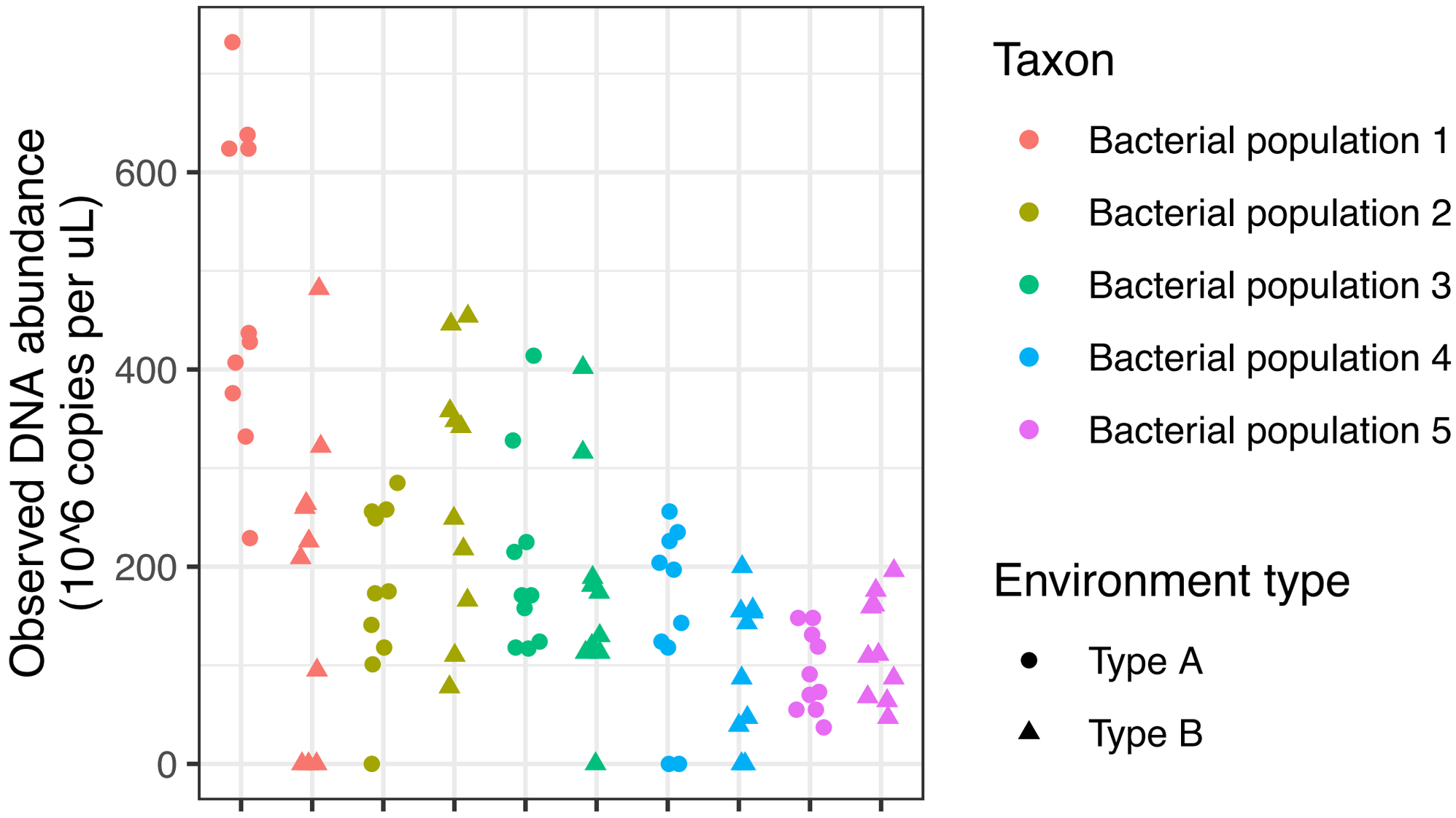
The environment



The environment



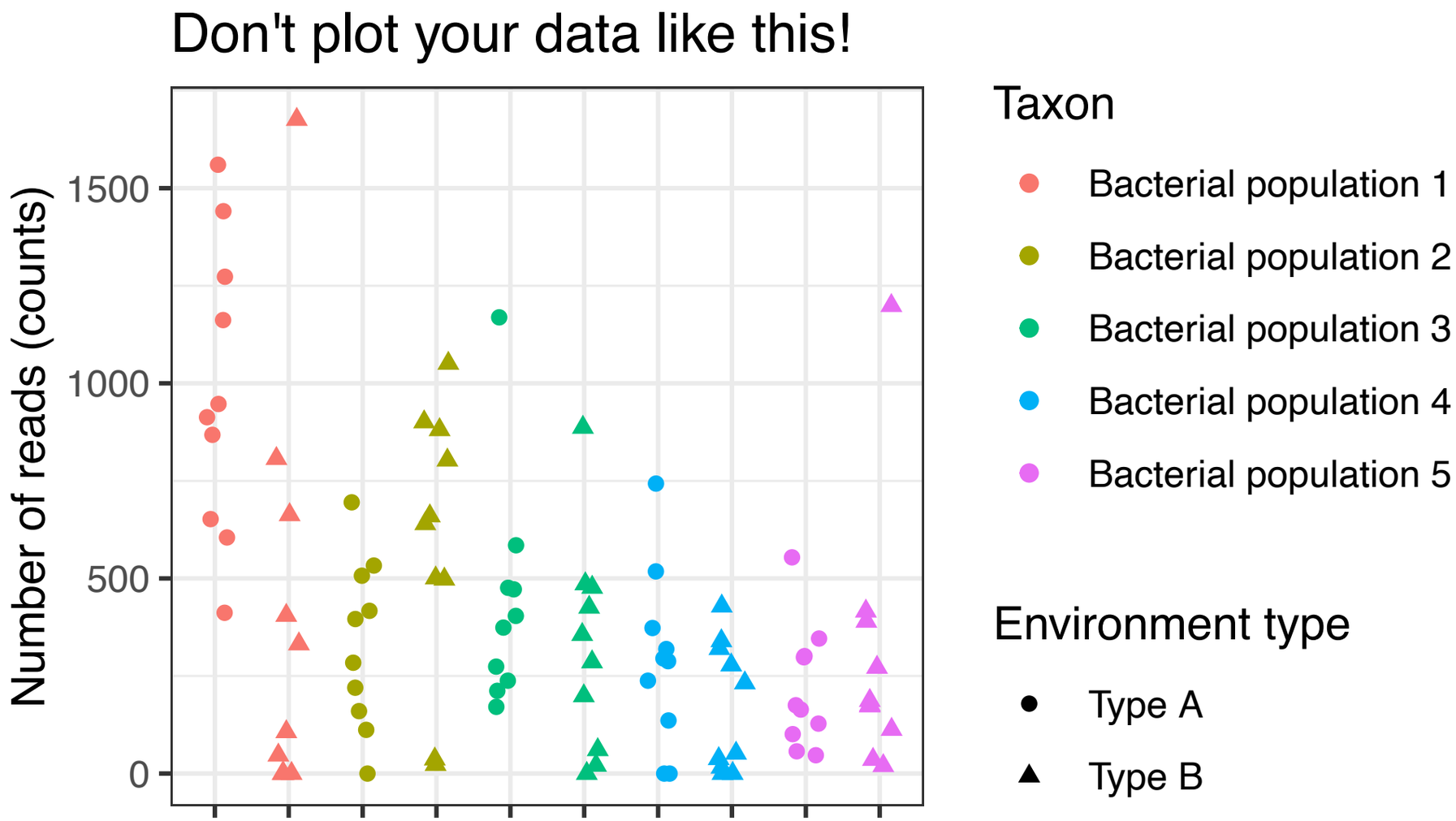
Concentration data*



*optimistically

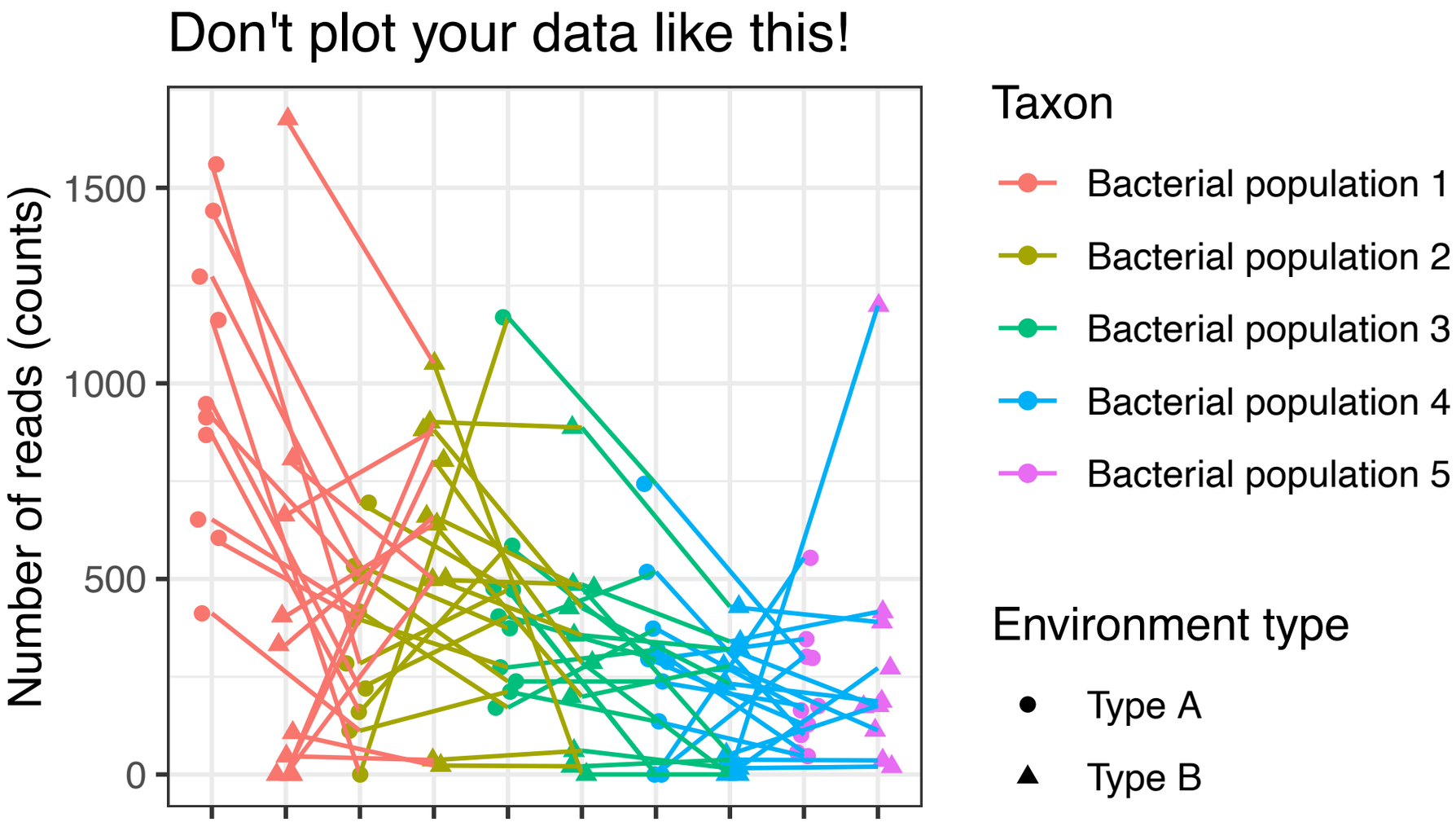
HTS data

- We get a random number of reads per sample

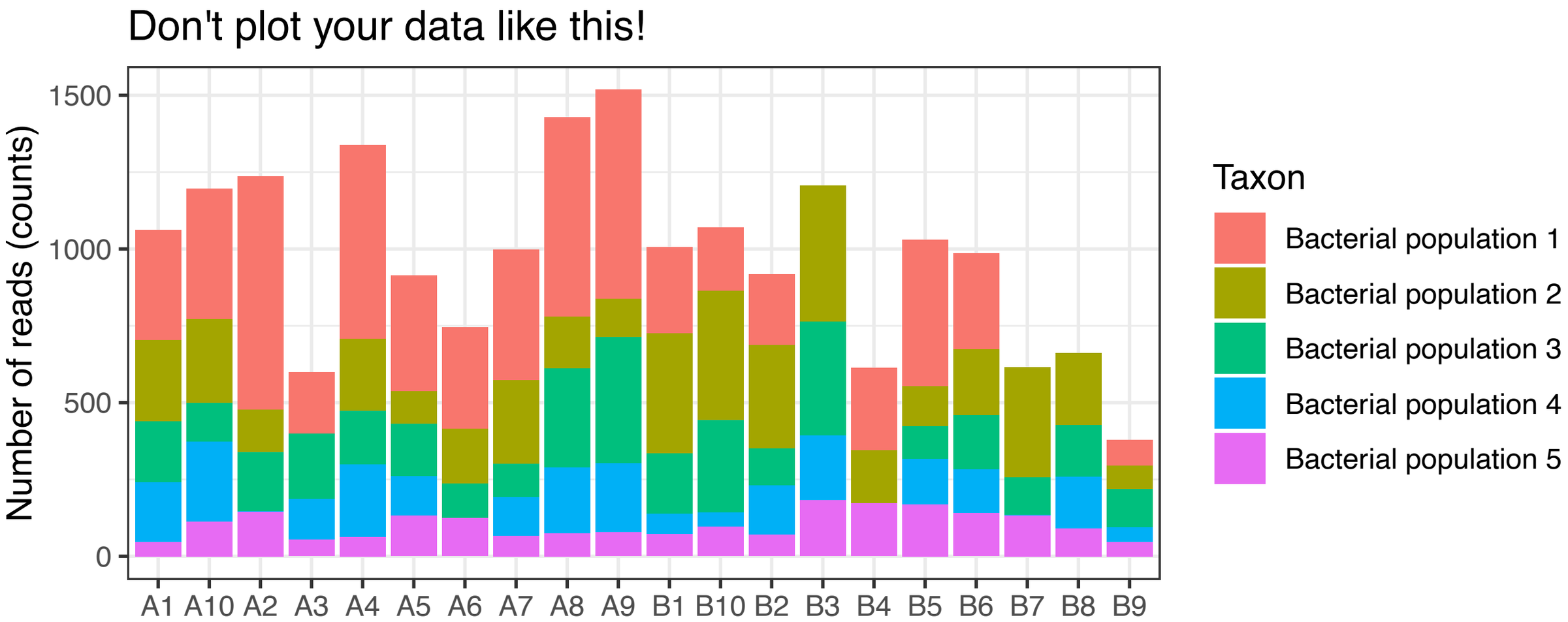


HTS data

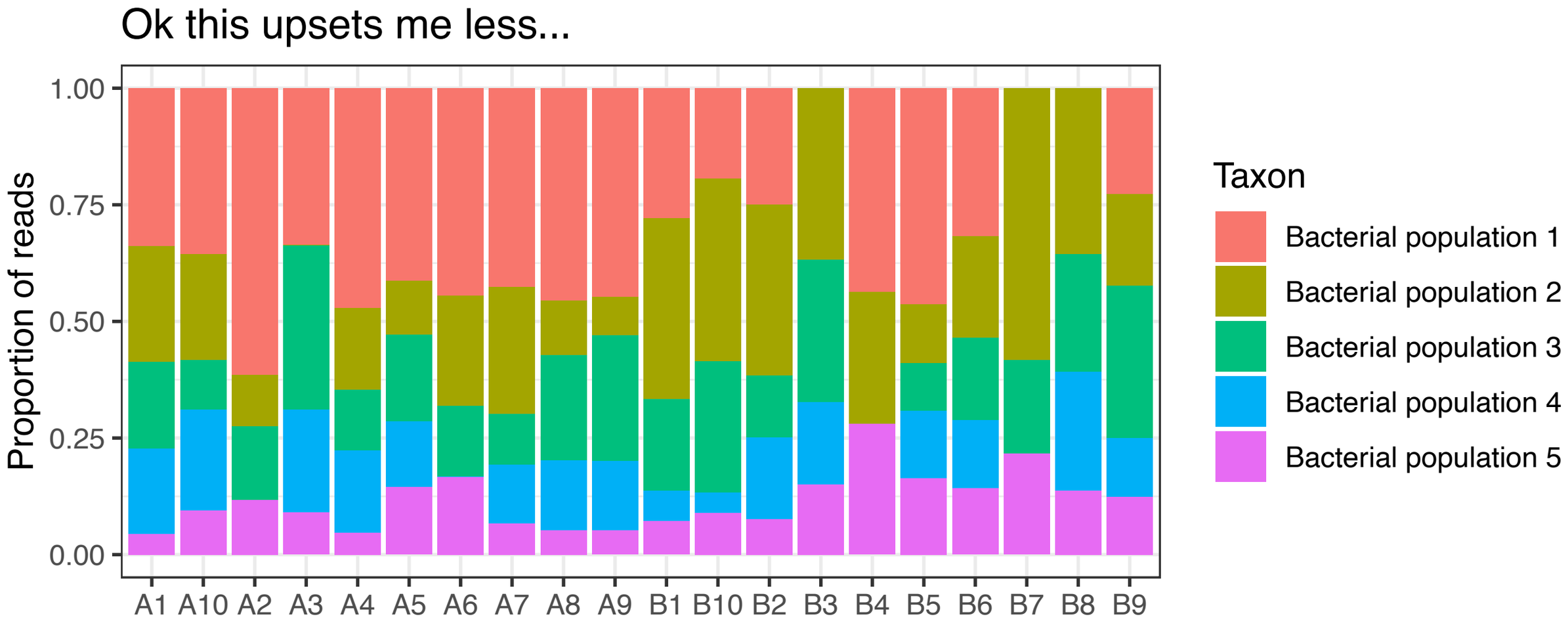
- We get a random number of reads per sample



HTS data


Don't plot your data like this!
Number of reads (counts)
1500
1000
500
0
A1 A10 A2 A3 A4 A5 A6 A7 A8 A9 B1 B10 B2 B3 B4 B5 B6 B7 B8 B9
Taxon
Bacterial population 1
Bacterial population 2
Bacterial population 3
Bacterial population 4
Bacterial population 5

HTS data

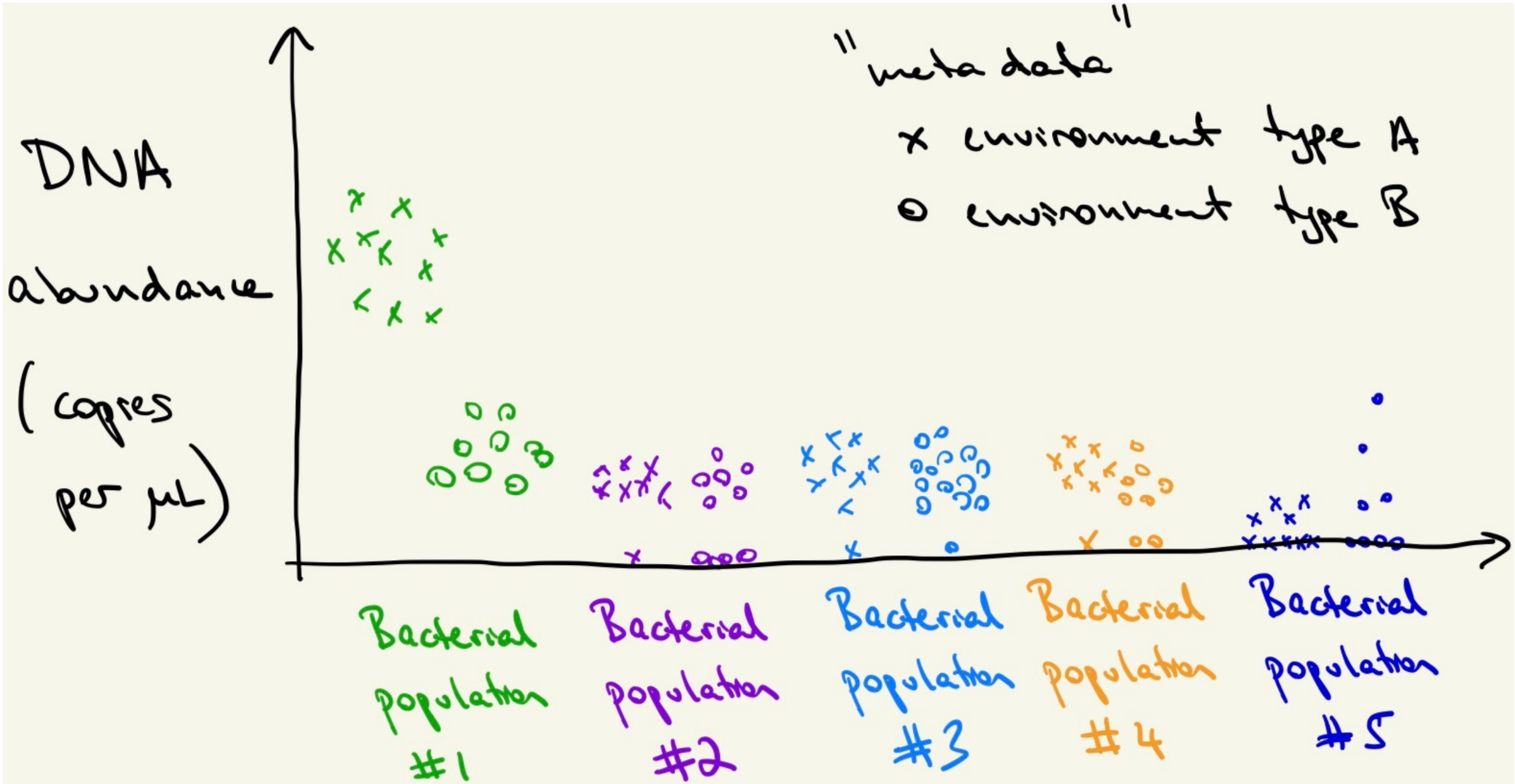

Ok this upsets me less...
Proportion of reads
1.00
0.75
0.50
0.25
0.00
A1
A10
A2
A3
A4
A5
A6
A7
A8
A9
B1
B10
B2
B3
B4
B5
B6
B7
B8
B9
Taxon
Bacterial population 1
Bacterial population 2
Bacterial population 3
Bacterial population 4
Bacterial population 5

HTS data

- Some considerations

1. Total counts are random ✅
    - Modeling total counts directly is a bad idea
2. Proportions can be misleading
3. Taxa are unequally well-detected

The environment



#2 Proportions can be misleading

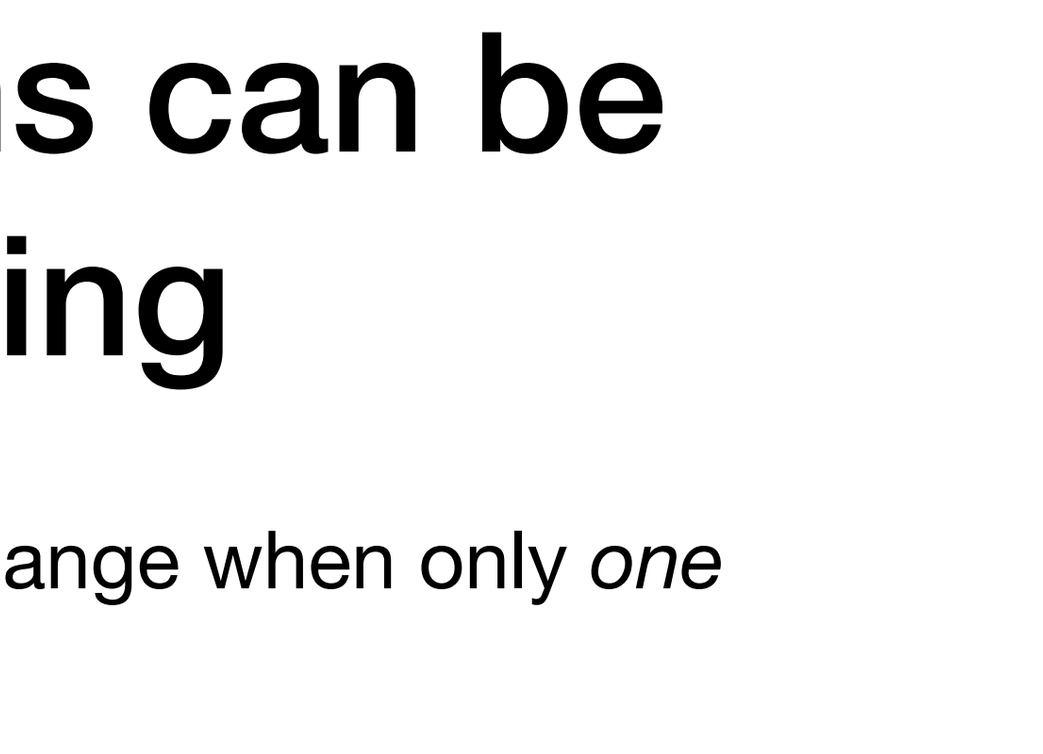
- Relative abundance of *all* taxa change when only *one* taxon's abundance changes

  - Not "spurious" but *misleading*

  - **0.33** / **0.33** / **0.33**

  - **0.50** / **0.25** / **0.25**

- This is an inherent limitation of *proportion-based parameters*



HTS data

- Some considerations

1. Total counts are random ✅
    - Modeling total counts directly is a bad idea
2. Proportions can be misleading ✅
3. **Taxa are unequally well-detected**

Differential detection interlude

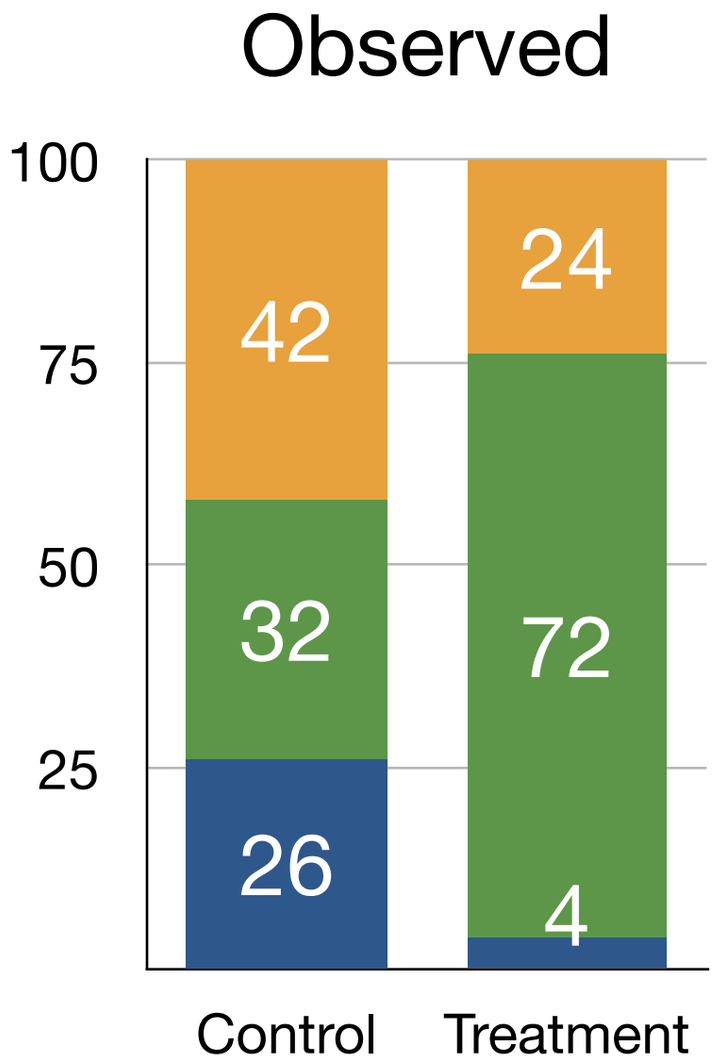
From Ben

#3 Taxa are unequally well-detected

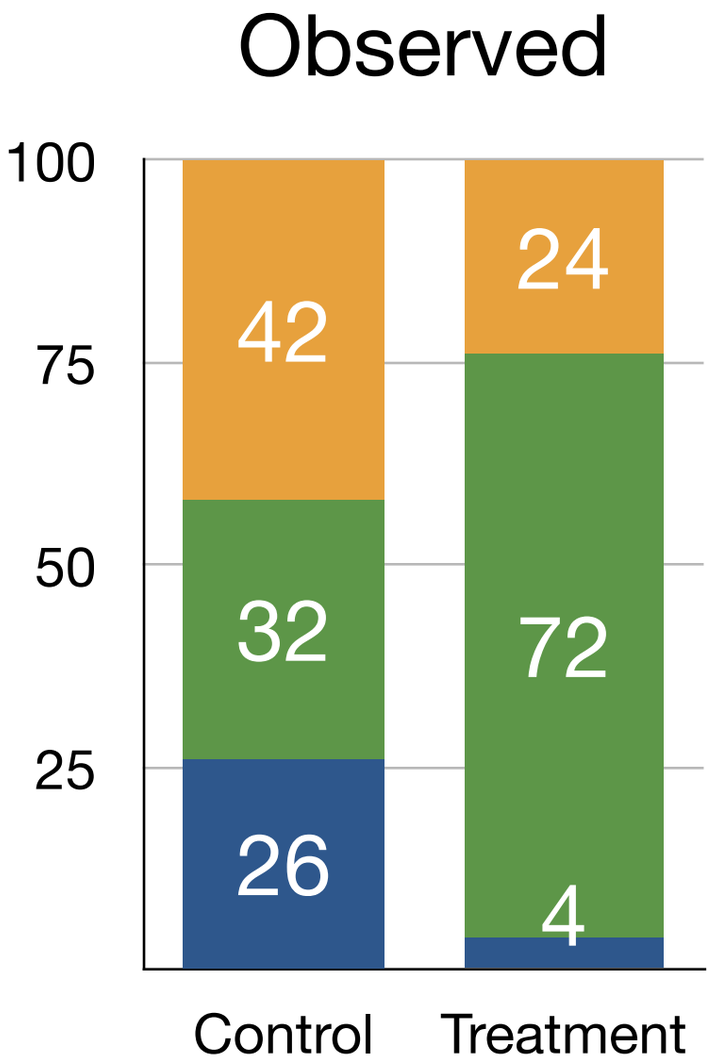
- As Ben talked about,

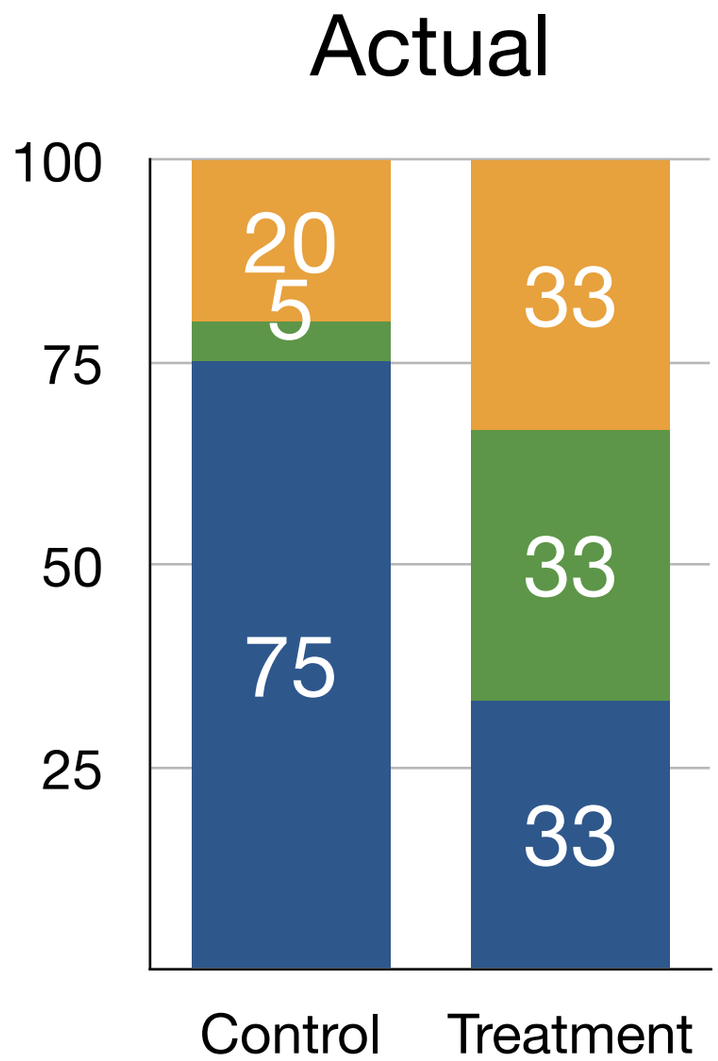
Observed relative abundance $\propto$ True relative abundance $\times$ Taxon-specific efficiencies

Expected value of $\frac{W_{ij}}{\sum_{j'} W_{ij'}} = \frac{p_{ij} e_j}{\sum_{j'} p_{ij'} e_{j'}}$

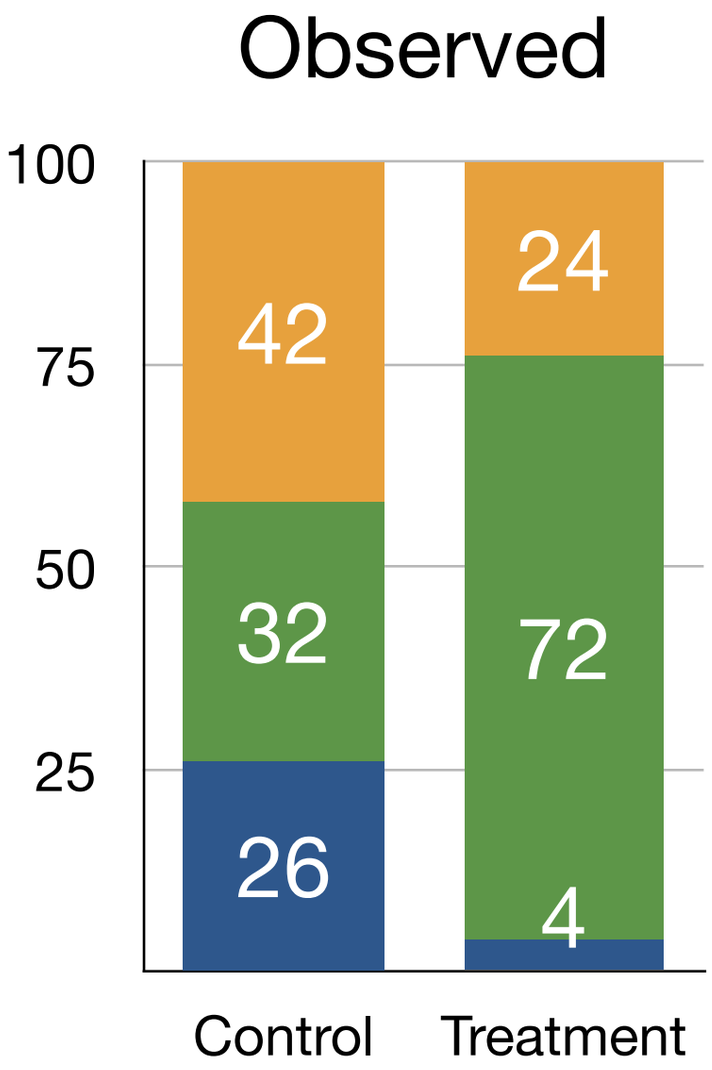


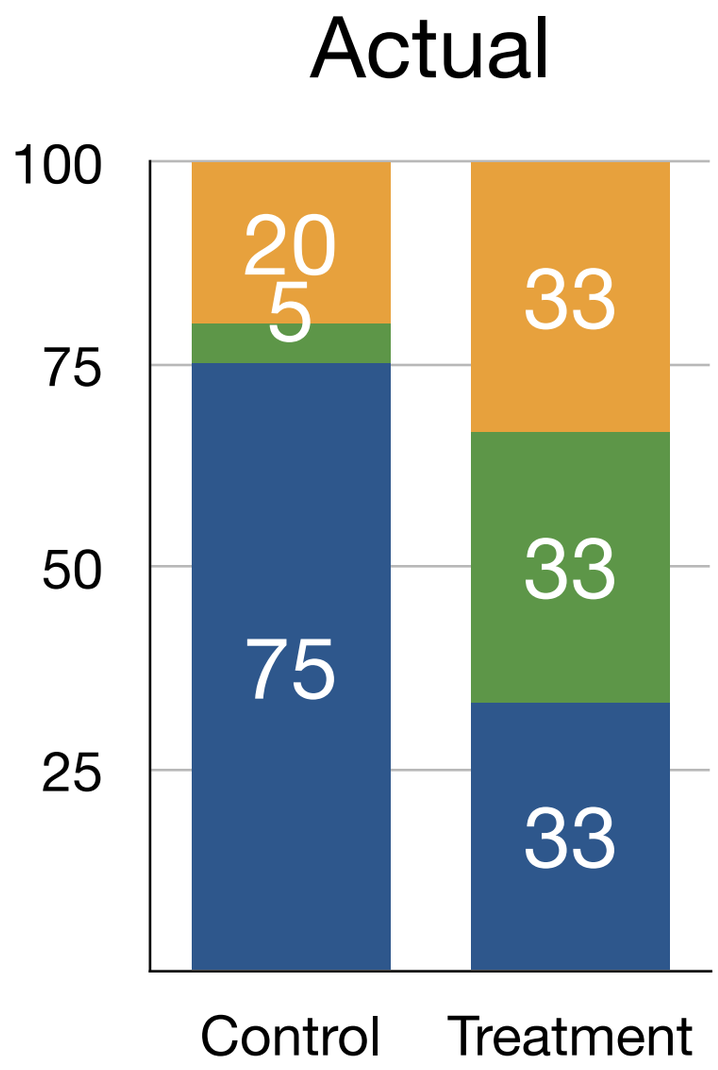
- A tempting conclusion:
  - The relative abundance of **taxon orange** *<u>decreased</u>* in the Treatment sample (right) compared to the Control sample (left)

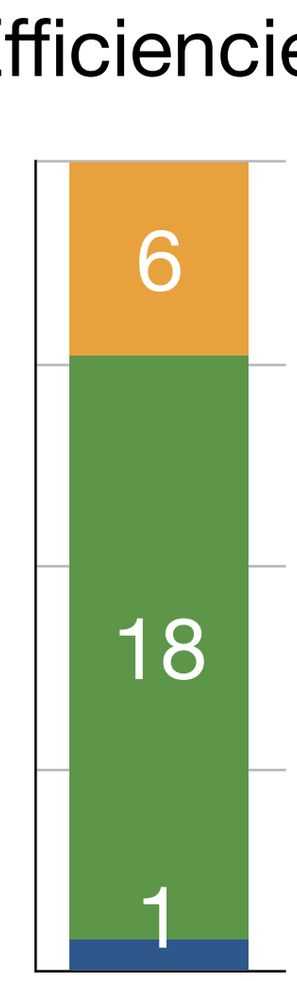




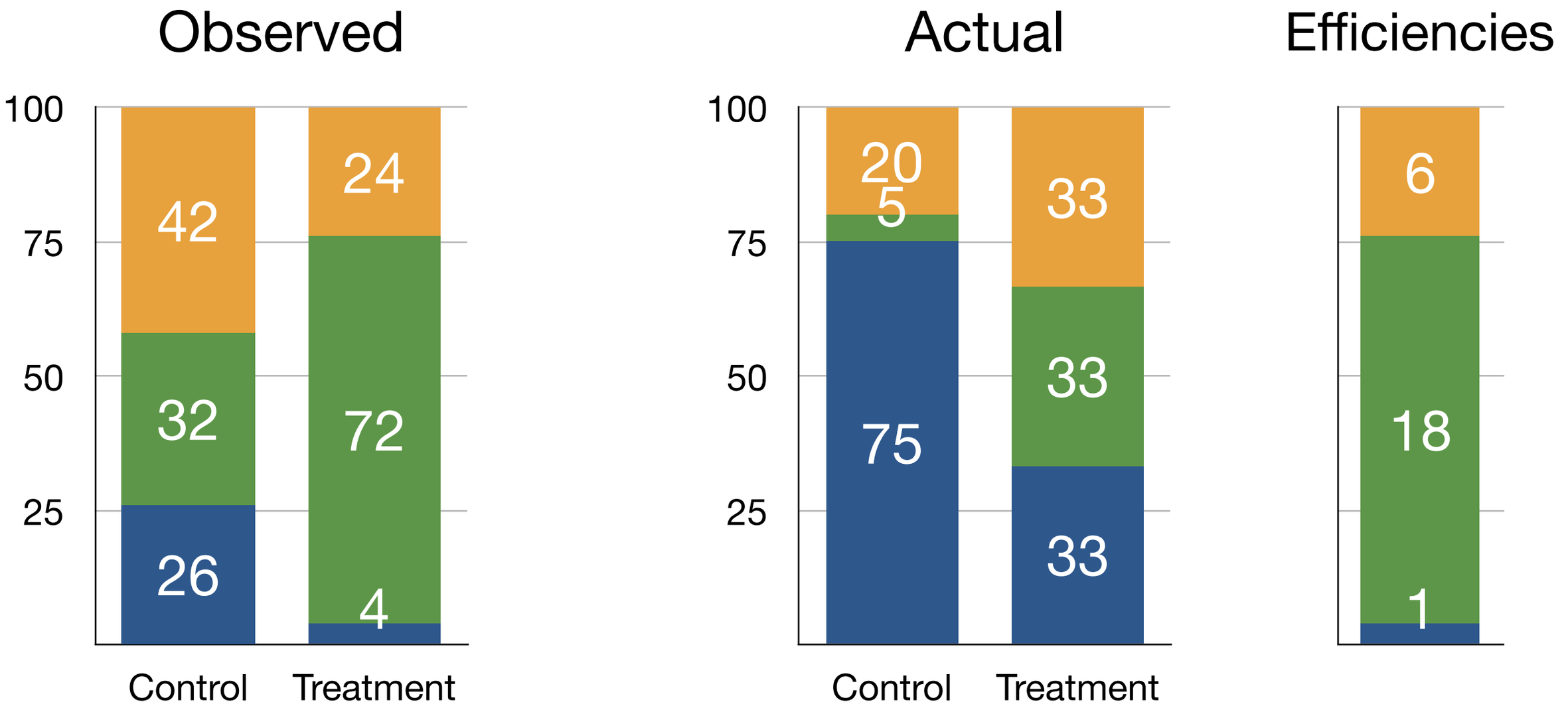
- In fact, the relative abundance of **taxon orange** *increased* in the Treatment sample compared to the Control sample







- In fact, the relative abundance of **taxon orange** *increased* in the Treatment sample compared to the Control sample



- In the absence of knowing e_j's, **you can't know** if relative abundances increase or decrease

For more details, see McLaren *et al* (2022+) "Implications of taxonomic bias..."

Summary of challenges in modeling HTS data

- Broadly speaking…
  - Concentration data *can* be compared across samples and can't really be compared across taxa
  - HTS counts & coverages *cannot* be compared across samples nor taxa
  - HTS proportions *cannot* be compared across samples
  - What can be compared? Ratios and fold differences, more soon…

Modeling abundance

- **Modeling concentration data**
- Modeling high-throughput sequencing data

Absolute abundance data

- qPCR/ddPCR data can usually be modeled with techniques you learnt in Stats 101/102
    - Linear regression (or t-tests, ANOVA…)
        - Estimates additive differences in means
- Poisson regression
    - Estimates multiplicative/fold-differences in means

Absolute abundance data

- qPCR/ddPCR data can usually be modeled with techniques you learnt in Stats 101/102
    - Linear regression (or t-tests, ANOVA…)
        - Estimates additive differences in means
    - Poisson regression
        - Estimates multiplicative/fold-differences in means

Choose based on *parameters*, not *data characteristics*

Absolute abundance data

- **Linear models** most generally look like

$$\text{mean outcome}_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

- β's can be interpreted as either means or differences in means

- β_1 is the *difference in the mean* outcome between populations with the same values of $X_2, \ldots X_p$ but who differ in X_1 by 1 unit

$$\beta_1 = \text{mean outcome}(x_1 + 1, x_2, \ldots, x_p) - \text{mean outcome}(x_1, x_2, \ldots, x_p)$$

Absolute abundance data

- Examples:

mean bacterial load$_i = \beta_0 + \beta_1 \times \mathbf{1}_{\{\text{person } i \text{ is on antibiotics}\}}$

- $\hat{\beta}_0$ is an estimate of the mean/average/expected bacterial load for people not on antibiotics

- $\hat{\beta}_1$ is an estimate of the difference in mean bacterial load between people who are versus aren't on antibiotics

Absolute abundance data

mean bacterial load$_i = \beta_0 + \beta_1 \mathbf{1}_{\{\text{person } i \text{ is on antibiotics}\}} + \beta_2(\text{age}_i - 40)$

- $\hat{\beta}_0$ is an estimate of the mean bacterial load for 40 y.o.'s people *not* on antibiotics

- $\hat{\beta}_1$ is an estimate of the difference in mean bacterial load between people of the same age who *are* versus *aren't* on antibiotics

- $\hat{\beta}_2$ is an estimate of the difference in mean bacterial load between people who differ in age by 1 year who have the same antibiotics use

Absolute abundance data

$$\text{mean bacterial load}_i = \beta_0 + \beta_1 \mathbf{1}_{\{\text{person } i \text{ is on antibiotics}\}} + \beta_2(\text{age}_i - 40)$$

$$+\beta_3 \times \mathbf{1}_{\{\text{person } i \text{ is on antibiotics}\}} \times (\text{age}_i - 40)$$

- $\hat{\beta}_0$ is an estimate of the mean bacterial load for 40 y.o.'s people *not* on antibiotics

- $\hat{\beta}_1$ is an estimate of the difference in mean bacterial load between 40 y.o.'s who *are* versus *aren't* on antibiotics

- $\hat{\beta}_2$ is an estimate of the difference in mean bacterial load between people who differ in age by 1 year who *aren't* on antibiotics

- $\hat{\beta}_3$ is an estimate of the difference in mean bacterial load between people who differ in age by 1 year who *are* on antibiotics, compared to between people who differ in age by 1 year who *aren't* on antibiotics

Absolute abundance data

- Step 1: decide what you want to estimate

- Step 2: figure out how to fit the relevant model

Absolute abundance data

- Step 1: decide what you want to estimate

mean bacterial load$_i = \beta_0 + \beta_1 \mathbf{1}_{\text{person } i \text{ is on antibiotics}\}} + \beta_2 \mathbf{1}_{\text{person } i\text{'s sample is from sputum}\}}$

- Step 2: figure out how to fit the relevant model

```
> my_data %>%
+   lm(ddpcr ~ Treatment + `Sample Type`, data = .)

Call:
lm(formula = ddpcr ~ Treatment + `Sample Type`, data = .)

Coefficients:
        (Intercept)          TreatmentON
             996530              -409238
`Sample Type`Sputum
            1006955
```

Absolute abundance data

- Step 1: decide what you want to estimate

$$\text{mean bacterial load}_i = \beta_0 + \beta_1 \mathbf{1}_{\text{person } i \text{ is on antibiotics}\}} + \beta_2 \mathbf{1}_{\text{person } i\text{'s sample is from sputum}\}}$$

$$+\beta_3 \mathbf{1}_{\text{person } i \text{ is on antibiotics}\}} \mathbf{1}_{\text{person } i\text{'s sample is from sputum}\}}$$

- Step 2: figure out how to fit the relevant model

```
> my_data %>%
+   lm(ddpcr ~ Treatment * `Sample Type`, data = .)

Call:
lm(formula = ddpcr ~ Treatment * `Sample Type`, data = .)

Coefficients:
                      (Intercept)                       TreatmentON
                           968202                         -234550
          `Sample Type`Sputum  TreatmentON:`Sample Type`Sputum
                          1063610                         -349375
```

Absolute abundance data

- **Poisson models** most generally look like

$$\log(\text{mean outcome})_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

e^{β_1} is the fold-difference in the mean outcome between populations with the same values of $X_2, \ldots X_p$ but who differ in X_1 by 1 unit

$$e^{\beta_1} = \text{mean outcome}(x_1 + 1, x_2, \ldots, x_p) \Big/ \text{mean outcome}(x_1, x_2, \ldots, x_p)$$

$$\beta_1 = \log \text{mean outcome}(x_1 + 1, x_2, \ldots, x_p) - \log \text{mean outcome}(x_1, x_2, \ldots, x_p)$$

Accessing `lm` Lab

1. Go to Schedule on Wiki to Wednesday afternoon, click on "Statistics Labs"

2. *Copy the command* under LM LAB

3. *Run the copied command* in your RStudio Server console or locally

```
> download.file("https://raw.githubusercontent.com/statdivlab/stamps2024/main/stats
  -labs/lm-lab/lm-ggplot-lab.Rmd", "lm-ggplot-lab.Rmd")
```

Modeling abundance

- Modeling concentration data
- **Modeling high-throughput sequencing data**

Differential abundance

- What are different ways the abundance of unit j could be "different" across the environment types treatment and control?

😻	Y_{ij} 💸	1	2	…	J
SAMPLE 1					
SAMPLE 2					
…					
SAMPLE M					
SAMPLE M+1					
…					
SAMPLE N-1					
SAMPLE N					

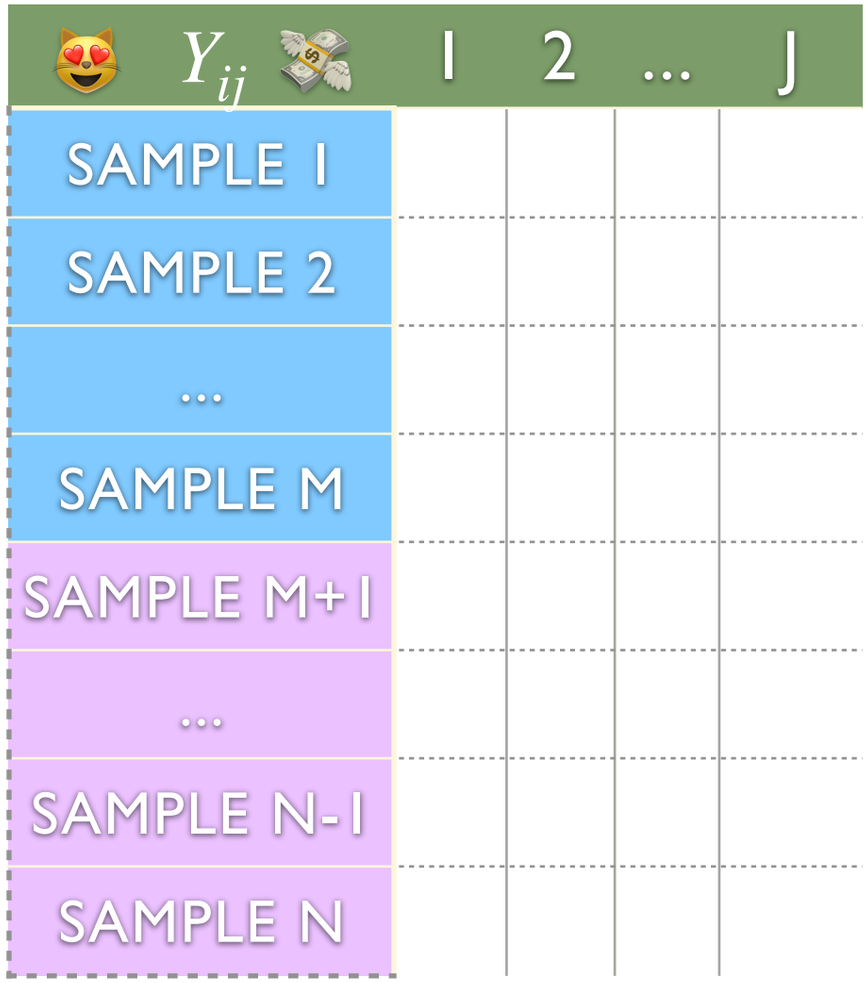
Dream big! Remember, you know *everything*!

Differential abundance

- …

- …

- …

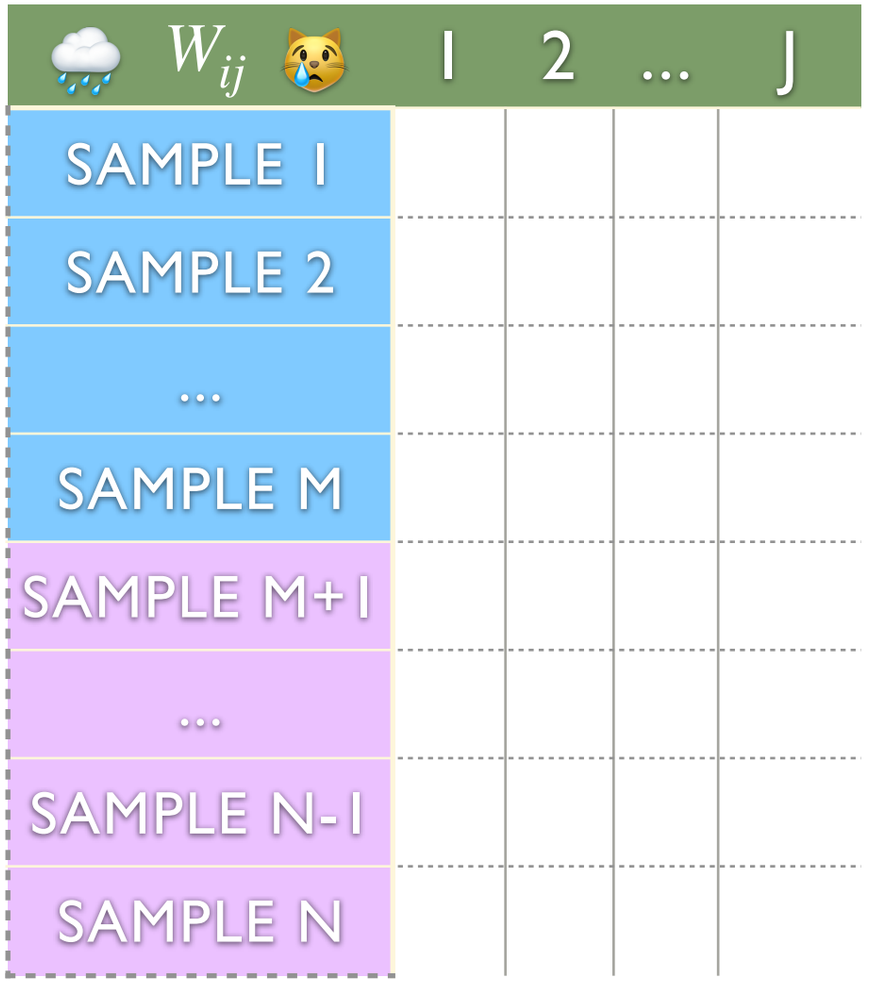


Differential abundance & identifiability

- Under reasonable models for our data W_{ij}, many of these parameters are not *identifiable**

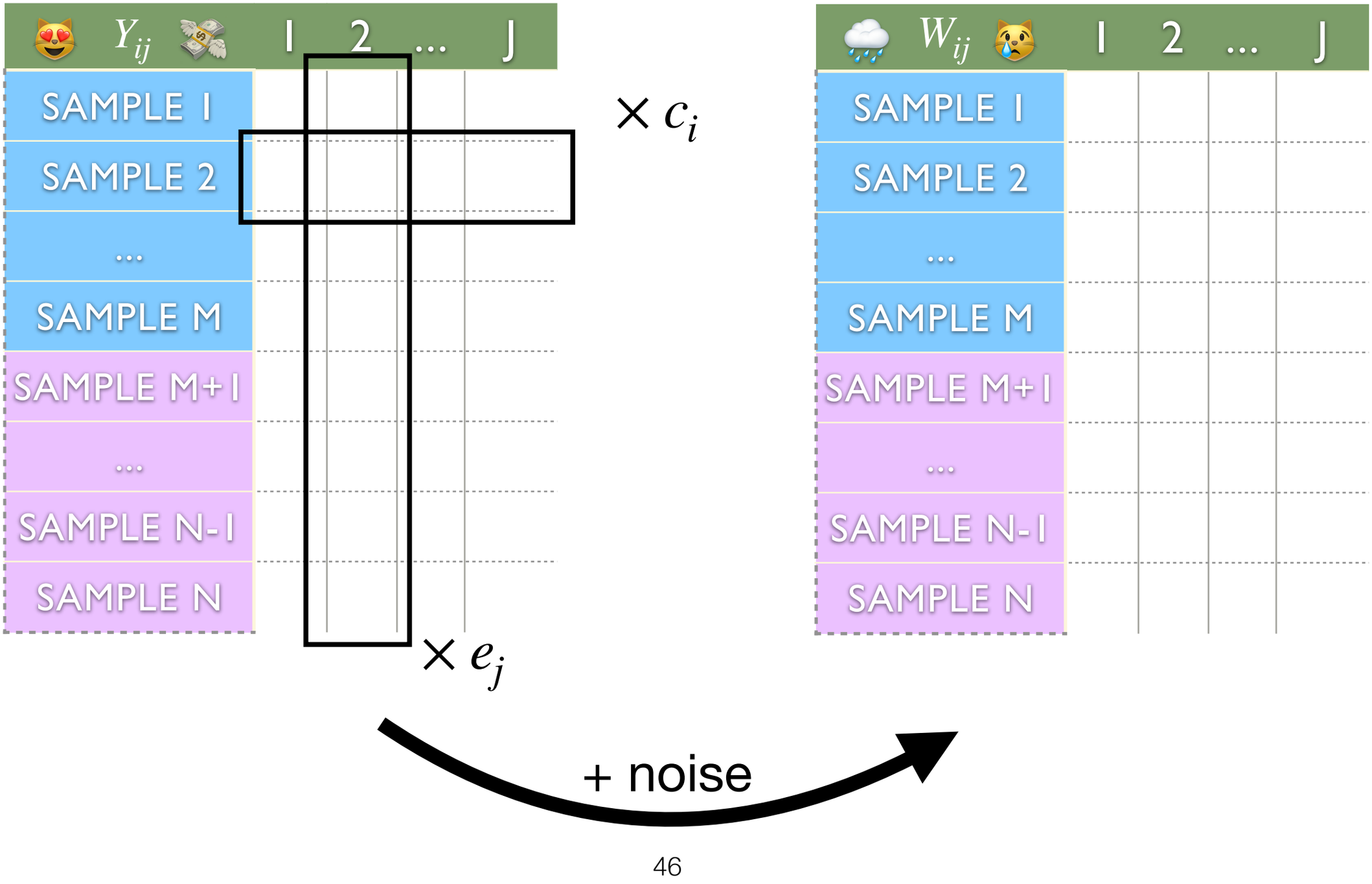
- not *identifiable* = we can't learn them from the $W'_{ij}s$

- 💔



*It took me *years* to figure out that identifiability is the problem

Reasonable models



Reasonable models

1. Total counts are random ✅
2. Proportions can be misleading ✅
3. Taxa are unequally well-detected ✅

- My reasonable model is

$$\text{expected } W_{ij} \approx c_i \times e_j \times Y_{ij}$$

- c_i are unknown sample-specific observation intensities
- e_j are unknown taxon-specific detection efficiencies
- W_{ij} are random observations; Y_{ij} are unknown true abundances

Un-identifiable parameters

$$\text{expected } W_{ij} \approx c_i \times e_j \times Y_{ij}$$

- Some parameters that are *not identifiable* include
  - average $Y_{\text{g1},j}$
  - average $Y_{\text{g1},j}$ − average $Y_{\text{g2},j}$
  - average $p_{\text{g1},j}$ − average $p_{\text{g2},j}$
  - average $Y_{\text{g1},j}$ / average $Y_{\text{g2},j}$

Identifiable parameters

- One parameter that is *identifiable* is

$$\frac{\mathbb{E}Y_{\text{group 1},j} / \mathbb{E}Y_{\text{group 2},j}}{\mathbb{E}Y_{\text{group 1},j'} / \mathbb{E}Y_{\text{group 2},j'}}$$

Identifiable parameters

$$\text{expected } W_{ij} \approx c_i \times e_j \times Y_{ij}$$

- Intuitively,

$$\frac{Y_{\text{group 1},j} / Y_{\text{group 2},j}}{Y_{\text{group 1},j'} / Y_{\text{group 2},j'}} \approx \frac{W_{\text{group 1},j} / W_{\text{group 2},j}}{W_{\text{group 1},j'} / W_{\text{group 2},j'}}$$

- (Remember: identifiable = we can learn about it)

Identifiable parameters

- Since

$$\frac{\mathbb{E}Y_{\text{group 1},j}}{\mathbb{E}Y_{\text{group 2},j}} \Big/ \frac{\mathbb{E}Y_{\text{group 1},j'}}{\mathbb{E}Y_{\text{group 2},j'}}$$

is identifiable, so is its logarithm:

$$\log\left(\frac{\mathbb{E}Y_{\text{group 1},j}}{\mathbb{E}Y_{\text{group 2},j}}\right) - \log\left(\frac{\mathbb{E}Y_{\text{group 1},j'}}{\mathbb{E}Y_{\text{group 2},j'}}\right)$$

Identifiable parameters

$$\log\left(\frac{\mathbb{E}Y_{\text{group 1},j}}{\mathbb{E}Y_{\text{group 2},j}}\right) - \log\left(\frac{\mathbb{E}Y_{\text{group 1},j'}}{\mathbb{E}Y_{\text{group 2},j'}}\right)$$

- Great if you have a "reference category" j'
  - One unchanging in abundance, or
  - One that you're happy to compare to
- What if you don't?

Identifiable parameters

- Another parameter that is *identifiable* is

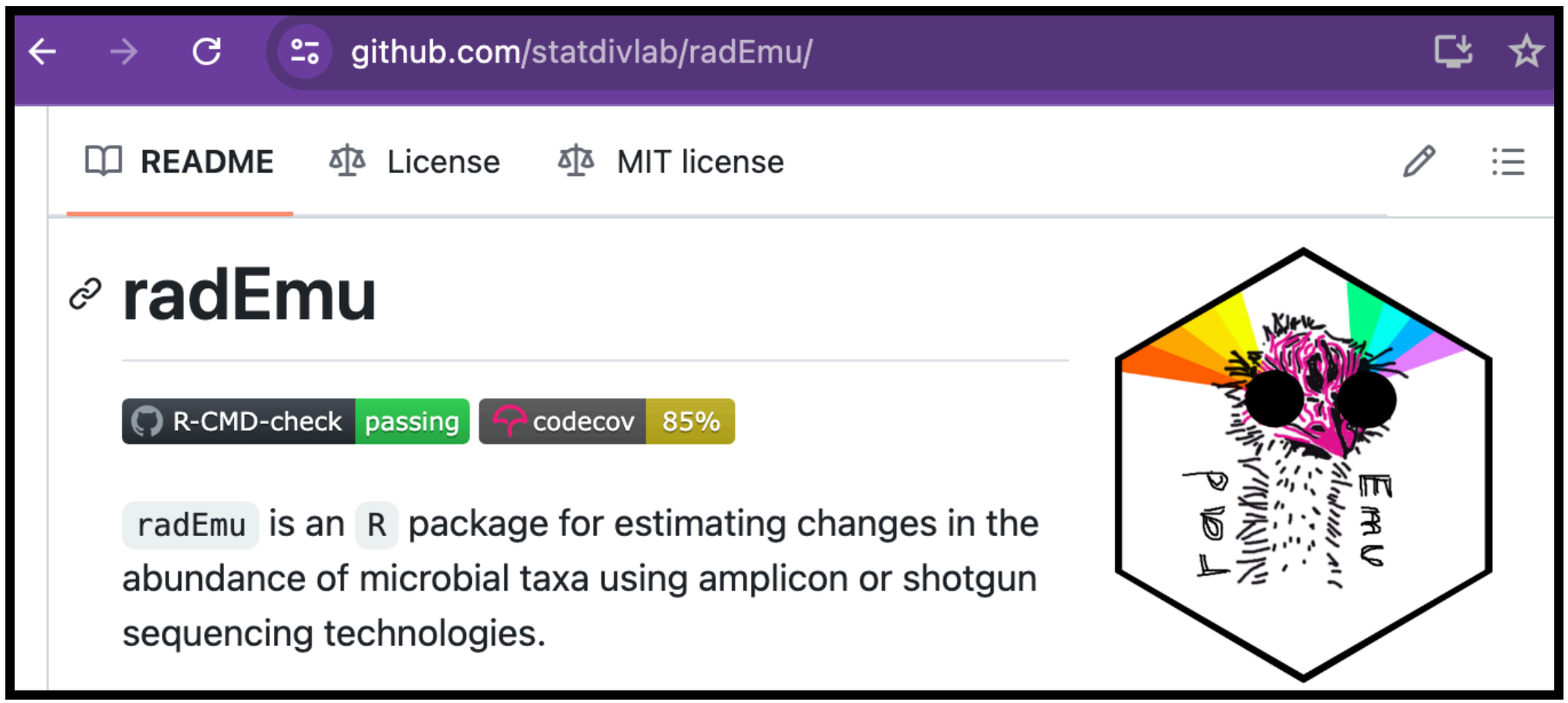
$$\log \frac{\mathbb{E}Y_{\text{group 1},j}}{\mathbb{E}Y_{\text{group 2},j}} - \text{average}_{j'} \log \left(\frac{\mathbb{E}Y_{\text{group 1},j'}}{\mathbb{E}Y_{\text{group 2},j'}} \right)$$

- Great if no reference category

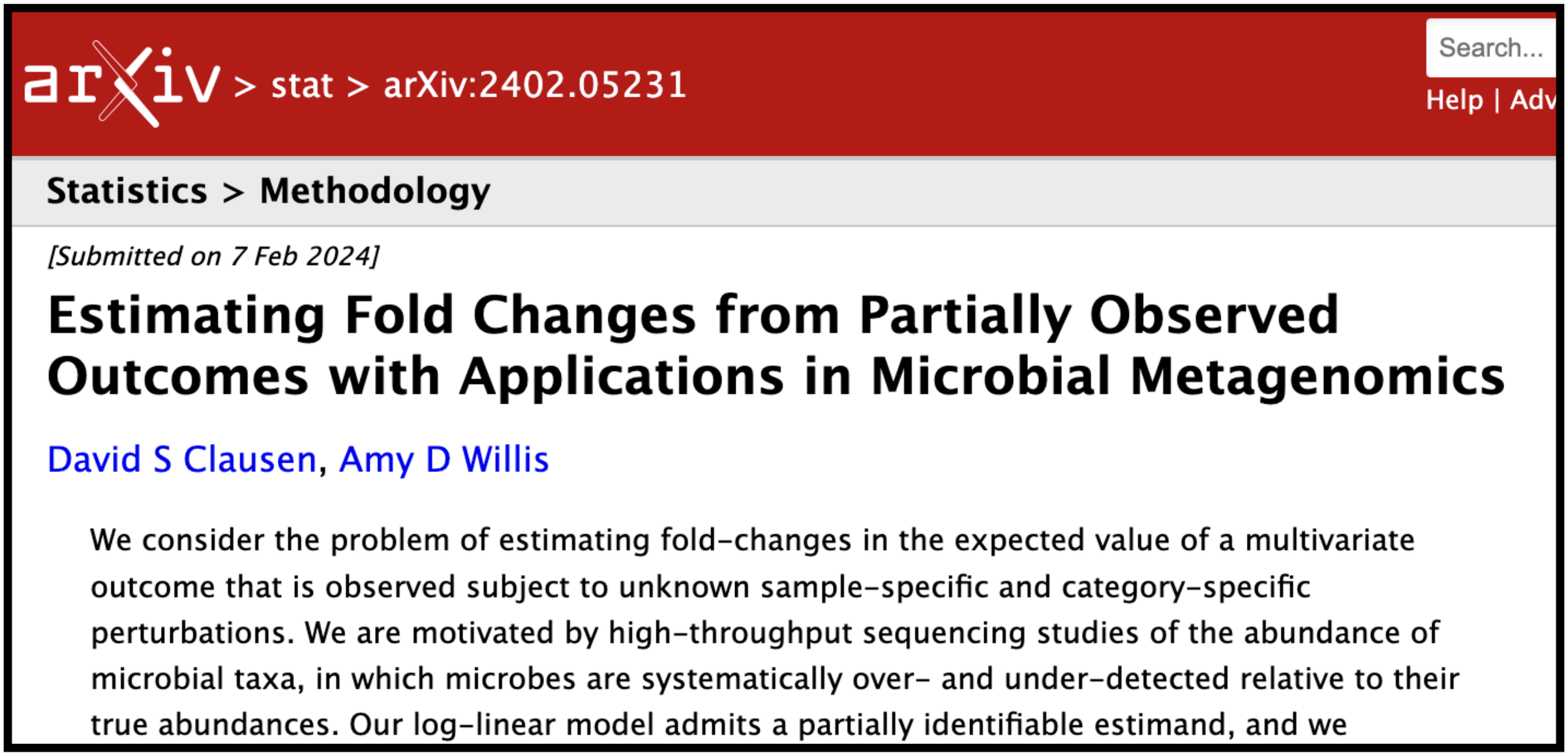
- average = mean or smoothed median

My favourite method…

- 🦤radEmu🦤 can estimate whichever you prefer!



My favourite method...



arXiv > stat > arXiv:2402.05231

Search...

Help | Adv

Statistics > Methodology

[Submitted on 7 Feb 2024]

Estimating Fold Changes from Partially Observed Outcomes with Applications in Microbial Metagenomics

David S Clausen, Amy D Willis

We consider the problem of estimating fold-changes in the expected value of a multivariate outcome that is observed subject to unknown sample-specific and category-specific perturbations. We are motivated by high-throughput sequencing studies of the abundance of microbial taxa, in which microbes are systematically over- and under-detected relative to their true abundances. Our log-linear model admits a partially identifiable estimand, and we

radEmu

- radEmuAbPiIl
- Using **r**elative **a**bundance **d**ata
  - to **e**stimate **mu**ltiplicative differences in **ab**solute abundances
    - with **p**artially **i**dentified **l**og-**l**inear models

radEmu

- radEmuAbPiIl

- Using **r**elative **a**bundance **d**ata

  - to **e**stimate **mu**ltiplicative differences in **ab**solute abundances

    - with **p**artially **i**dentified **l**og-**l**inear models

radEmu

- radEmu estimates fold-differences in mean *absolute abundance* Y_{ij} using *sequencing data…*

… compared to typical fold-differences

radEmu

- radEmu is not…
  - estimating absolute abundances because they're not identifiable
    - estimating fold-differences in absolute abundances across groups also not identifiable
- radEmu is…
  - estimating fold-differences in absolute abundances across groups relative to typical differences

radEmu

- radEmu is *most similar* to linear regression methods on CLR-transformed abundances

    - i.e., comparing the average of

$$\text{clr}\left(W_{ij}\right) = \log W_{ij} - \frac{1}{J}\sum_{j'=1}^{J} \log W_{ij'}$$

across groups

What problems arise in with this approach?

radEmu

- How people often deal with zeroes
  - transform their data e.g., replace zeroes with small values
  - throw out samples with “too much” sparsity
- These are *unnecessary* and *suboptimal*

radEmu

- Replacing

$$\text{mean}\left(\log W_{ij}\right) \text{ not well-defined}$$

with

$$\log\left(\text{mean} W_{ij}\right) \text{ well-defined}$$

is *one* of the things happening under the hood

- You can think about radEmu as an alternative to transforming your data to deal with zeroes

radEmu

- **Advantages**
  - Estimates something about the *environment*, not something about *sequencing*
  - Robust to differential detection
  - Controls Type 1 error
  - Handles lots of zeroes without pseudocounts
  - Robust to “overdispersion”
  - Adjusts for differential sequencing depth i.e., don’t rarefy
  - Handles any experimental design
  - Assumption-light

radEmu

- **Limitations**
  - Slower than other methods might run overnight
  - Sarah has a new method requiring a “reference set” that’s much faster… you can demo it today!

radEmu

- radEmu estimates fold-differences in mean absolute abundance using sequencing data

$$\text{fold diff. in F. prauznitzii} = \frac{\text{mean abs. abundance F. prauznitzii in cases}}{\text{mean abs. abundance F. prauznitzii in controls}}$$

- radEmu estimates

log (fold diff. in F. prauznitzii) − average log (fold diff. across all taxa)

```
emuFit(formula = ~ cases,
       data = my_metadata,
       Y = my_counts)
```

radEmu

```
emuFit(formula = ~ cases + age + sex,
       data = my_metadata,
       Y = my_counts)
```

fold difference in F. prauznitzii

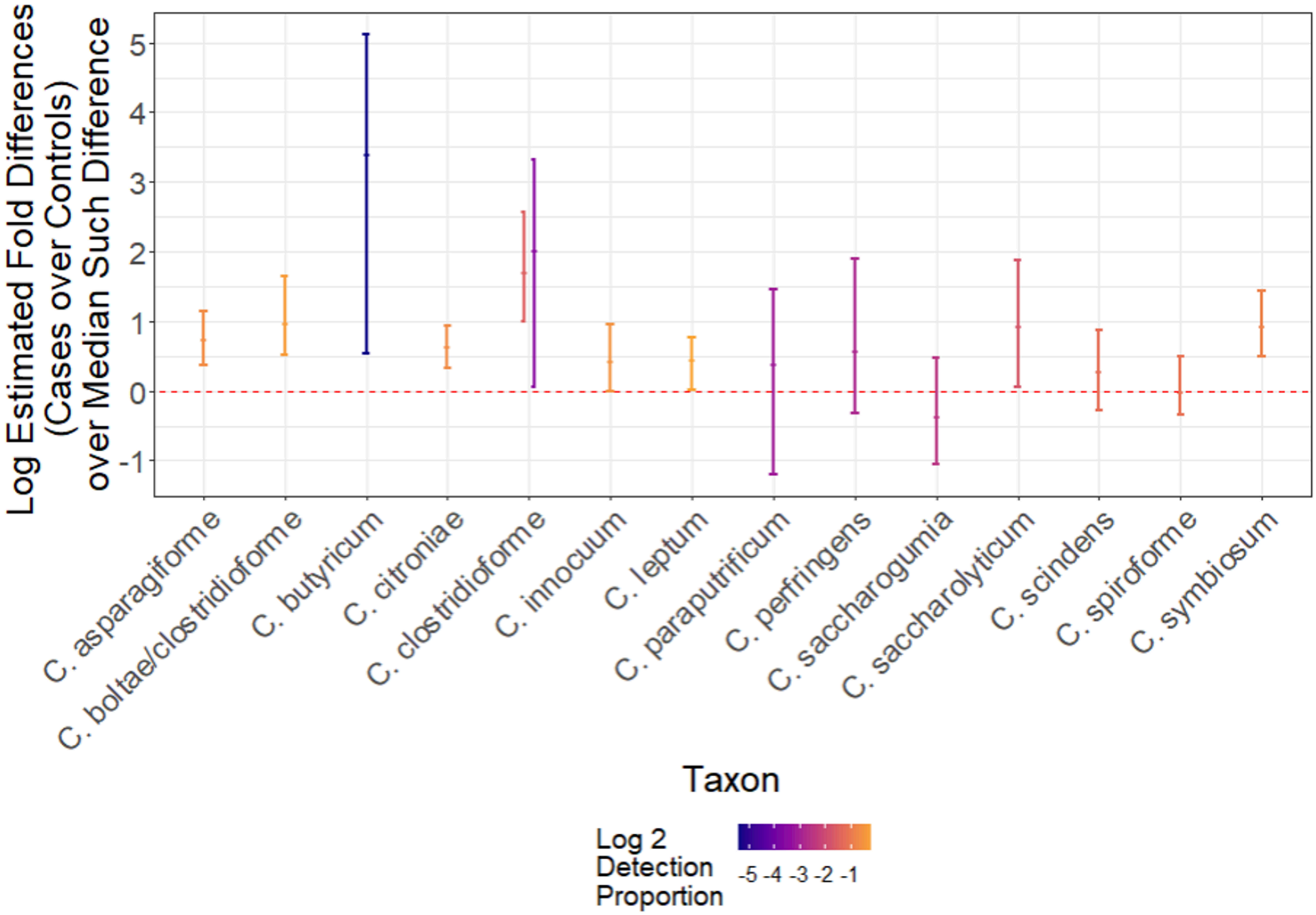
$$= \frac{\text{mean abs. abundance F. prauznitzii in cases of age } a \text{ and sex } s}{\text{mean abs. abundance F. prauznitzii in controls of age } a \text{ and sex } s}$$

- Goal: identify strains enriched/depleted in CRC samples compared to otherwise similar controls

radEmu

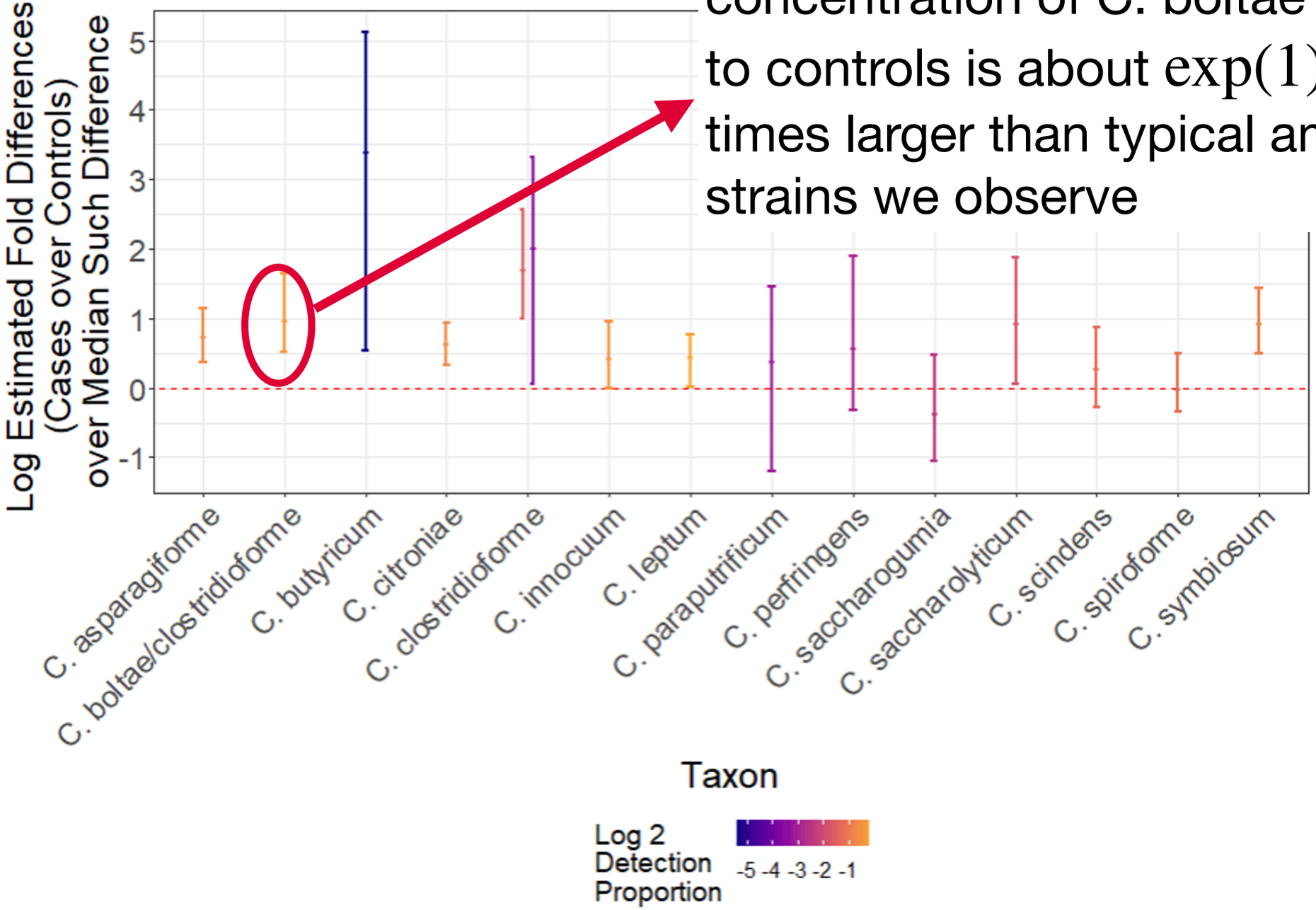
- Typical use case
  - Identify strains enriched/depleted in one type of samples compared to otherwise similar samples that differ in their type
- "Making reasonable comparisons"
- In practice: *Look at all taxa, identify the most enriched/depleted taxa*
  - Typical to look at FDR (q-values), not Type 1 error (p-values)

radEmu: Example



radEmu: Example

We estimate that the ratio of mean concentration of C. boltae in cases to controls is about $\exp(1) \approx 2.7$ times larger than typical among the strains we observe

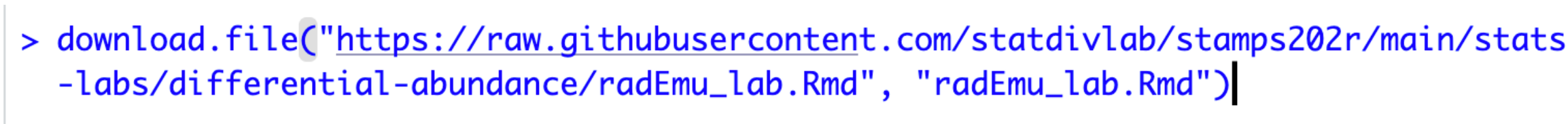


Accessing `radEmu` lab

1. Go to Schedule on Wiki to Wednesday afternoon, click on “Statistics Labs”

2. *Copy the command* under radEmu lab

3. *Run the copied command* in RStudio



Comparing radEmu to other methods

- I like radEmu… why?

Amy’s wish list

- You choose a meaningful parameter to estimate
- You choose a sensible way to estimate the parameter
- You choose tests that control Type 1 error

Comparing radEmu to other methods

- radEmu…
  - Estimates a parameter I *understand* and *care about*
    - Biggest fold-changes in absolute abundance
- Estimates it well
- Inference is correct = p-values trustworthy

Comparing radEmu to other methods

- Only makes sense to compare methods that estimate the same parameter
- None, but
    - ANCOM-BC2
    - ALDEx2
    - DESeq2
    - t-test on CLR transformed data

    target *similar* parameters

Comparing radEmu to other methods

- Make W_{ij}'s realistic lots of zeroes, high-variance

- Ask

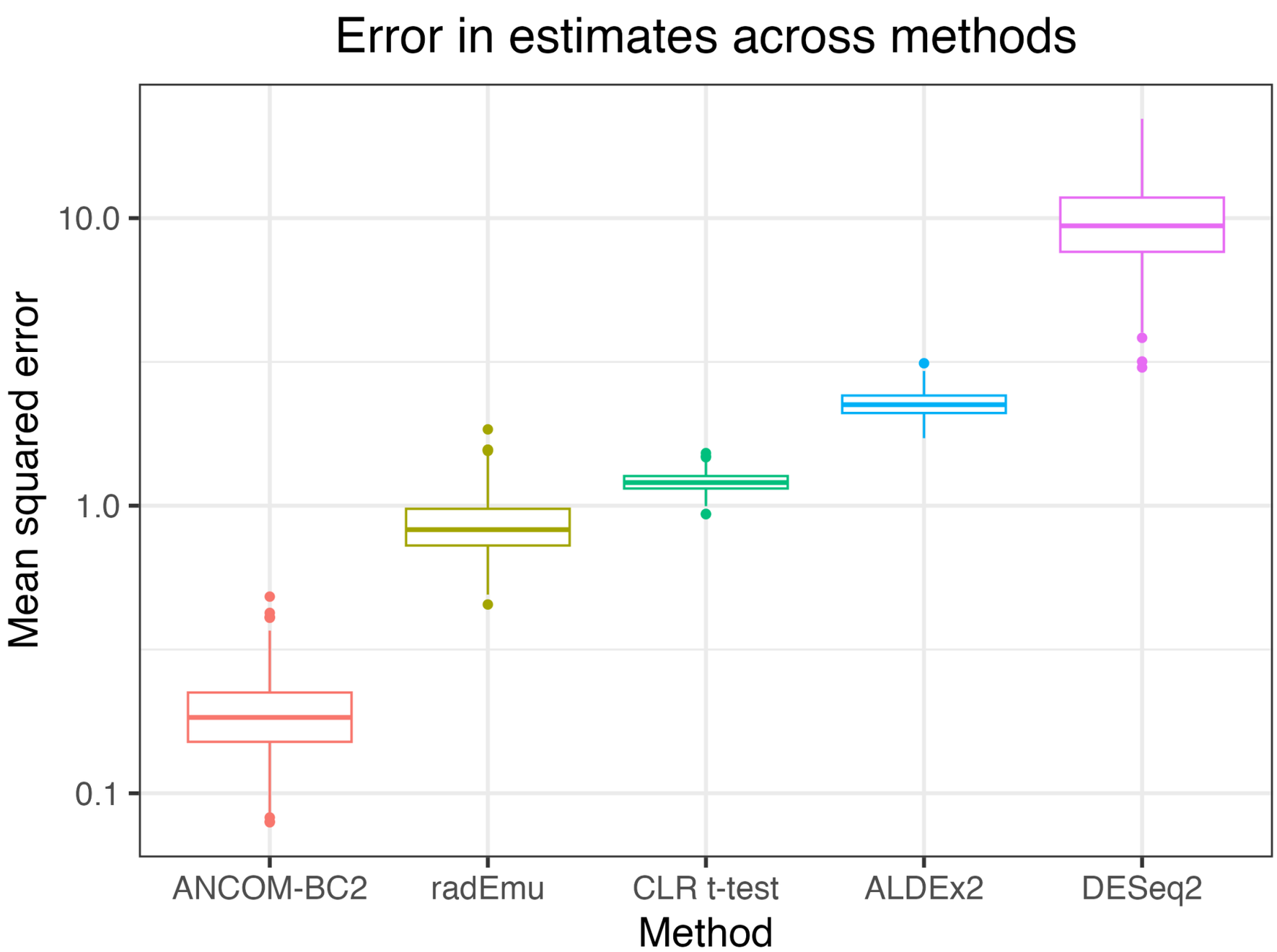
1. "How good are our estimates?"

2. "Do we have error rate control?"

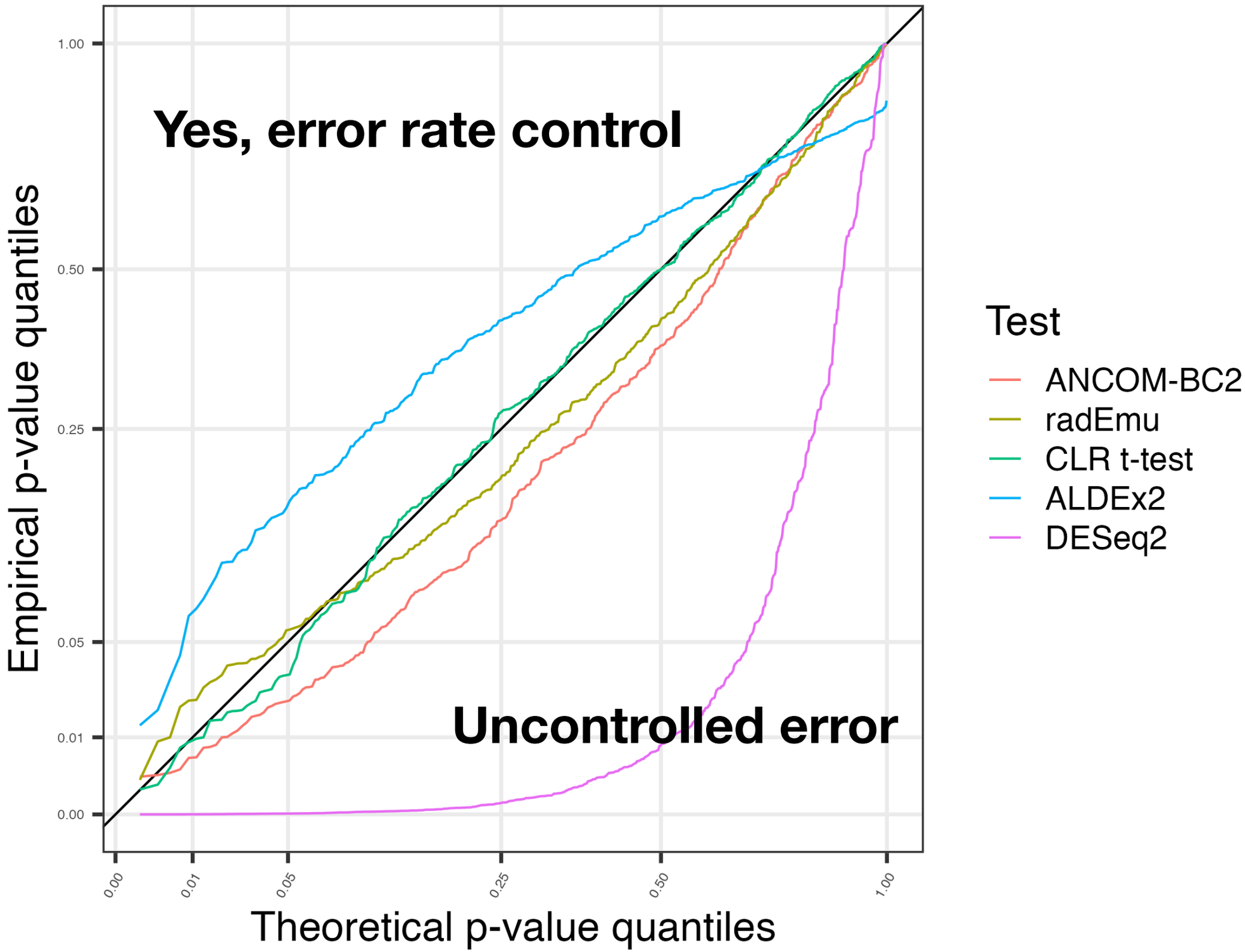
    - Null hypothesis: "Fold difference (cases vs. controls) in *F. praus* is equal to typical fold difference across taxa"

Special thanks to Sarah! 😻🍺

“How good are our estimates?”



“Do we have error rate control?”



Type I error rate control results

Method	1% Type 1 error	5% Type 1 error rate
ALDEx2	0.00	0.01
ANCOM-BC2	0.02	0.11
CLR t-test	0.01	0.06
DESeq2	0.52	0.67
radEmu	0.00	0.04

Simulation takeaways

- TL;DR In a realistically pathological setting,
    - radEmu has the lowest error in estimation out of all methods that control error rates

Simulation takeaways

- Under our simulation settings:
    - ALDEx2 controls the Type I error rate (is very conservative) BUT has the second highest MSE
    - ANCOM-BC2 has the lowest MSE BUT fails to control Type I error
    - CLR t-test almost controls Type I error but has the third highest MSE (and requires a pseudo count)
    - DESeq2 fails to control Type I error rate and has the highest MSE
    - radEmu controls Type I error rate and has the second lowest MSE

Closing thoughts

Differential abundance

- A common goal:
  - Determine which taxa are present in greater abundance in one group compared to another
- “Differential abundance [is] a category subject to some controversy in part on account of the fact that no unambiguous definitions of ‘differential’ or ‘abundance’ are widely agreed upon.”

Quoted with somewhat questionable taste

Differential abundance

- Many methods exist for “differential abundance”
  - ALDEx2
  - ANCOM-BC2
  - corncob
  - DESeq2
  - edgeR
  - metagenomeSeq
  - MaAsLin2
  - LEfSE
  - limma voom
  - radEmu
  - Wilcoxon/t-tests on proportions
  - t-tests on ratios
- multiple versions of almost all methods; multiple options for almost all methods

Differential abundance & HTS

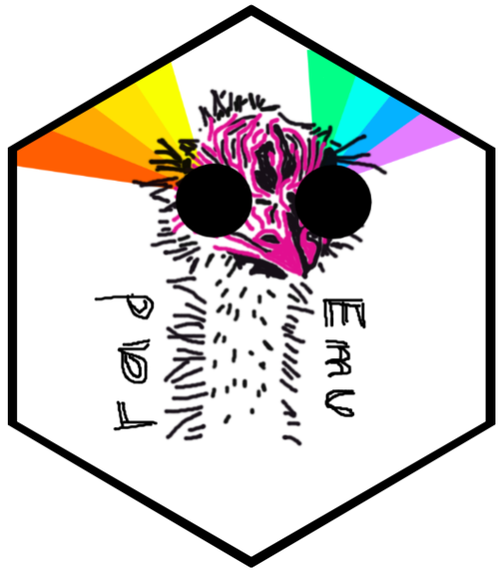
- We are restricted in what we can learn from HTS, because

1. Total counts are random ✅
2. Proportions can be misleading ✅
3. Taxa are unequally well-detected ✅

Amy’s wish list

- You choose a meaningful parameter to estimate
- You choose a sensible way to estimate the parameter
- You choose tests that control Type 1 error

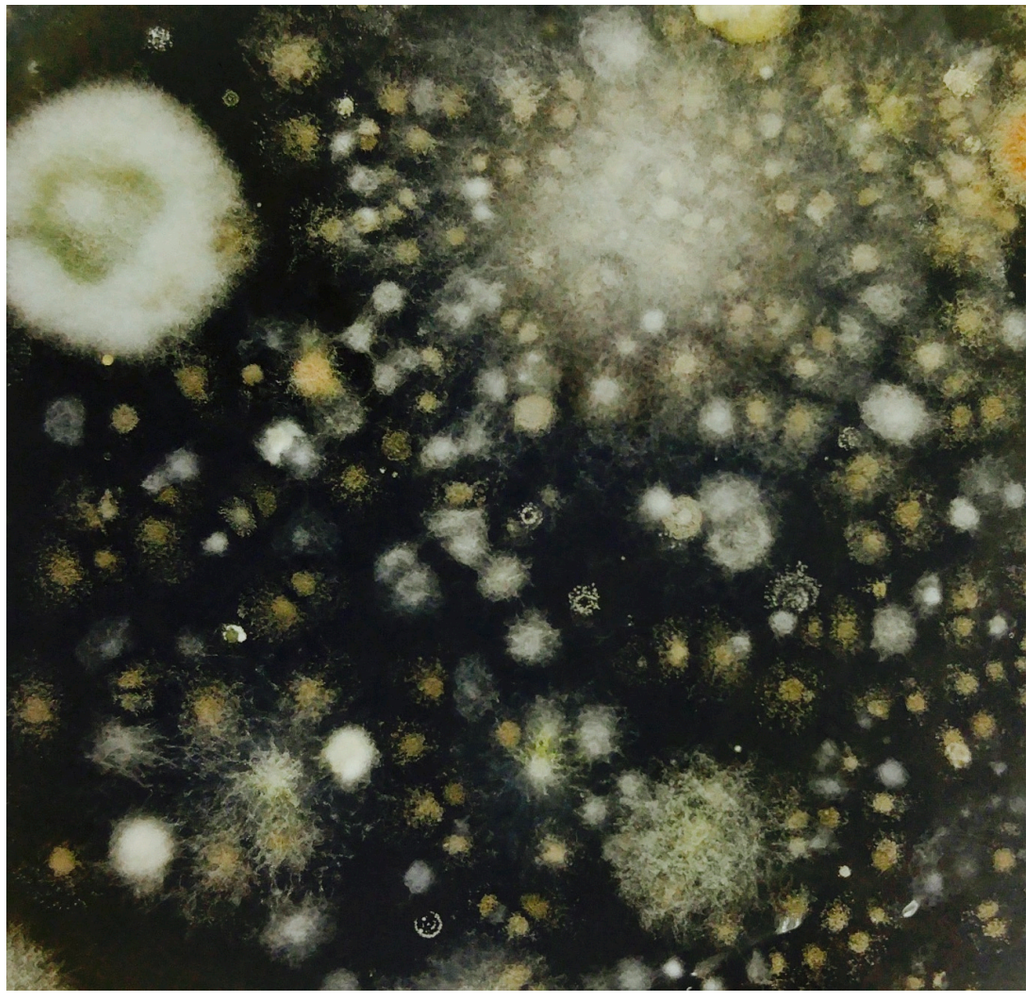
Amy's wish list



- You choose a meaningful parameter to estimate
- You choose a sensible way to estimate the parameter
- You choose tests that control Type 1 error

I like radEmu because it meets these criteria!

https://github.com/statdivlab/radEmu/



Modeling microbial abundances

Statistical Diversity Lab @ University of Washington

Amy Willis — @AmyDWillis — Associate Professor

Shirley Mathur — PhD Candidate

Sarah Teichman — PhD Candidate

María Valdez — PhD Candidate

Photo credit: T.D. Berry, Whitman lab, UW Madison