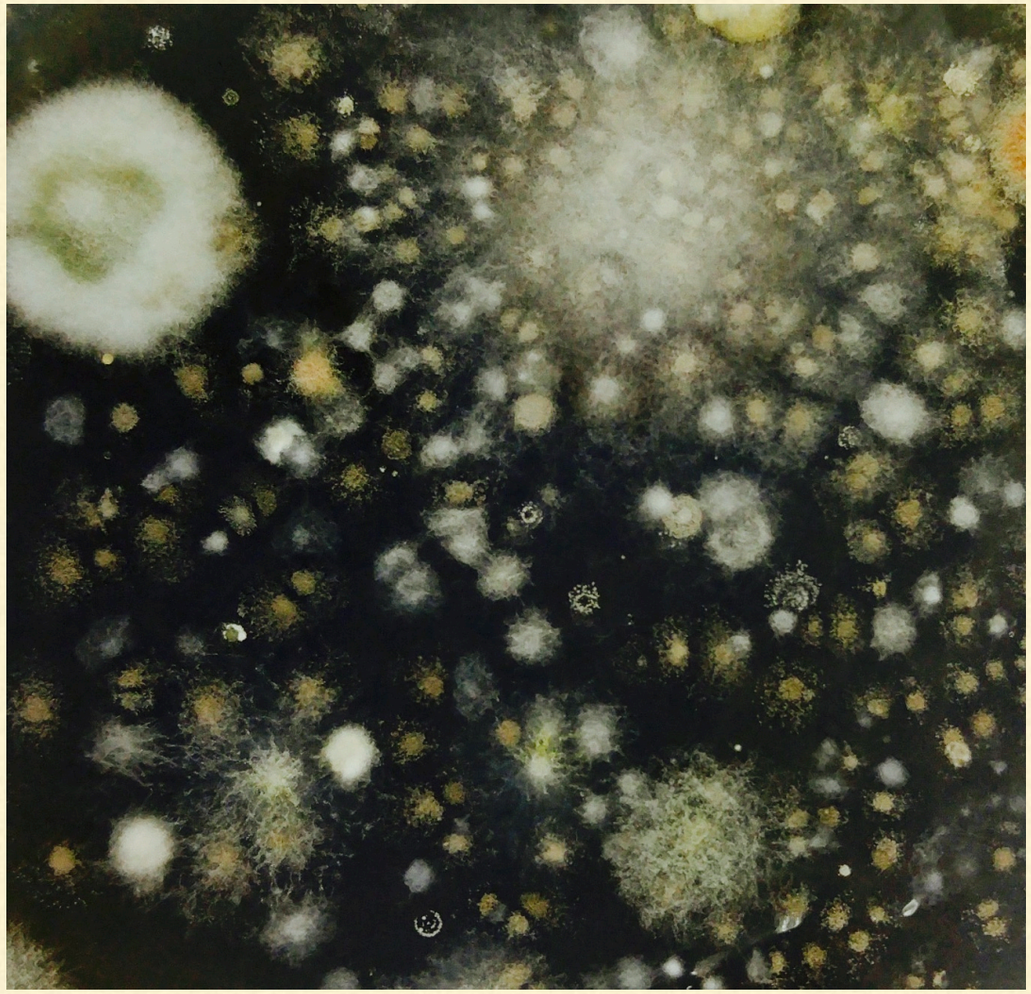


MODELING MICROBIOME DATA

Statistical Diversity Lab @ University of Washington

Amy Willis — @AmyDWillis — Associate Professor

Sarah Teichman — @sarah_teichman — 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 + 2x 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!

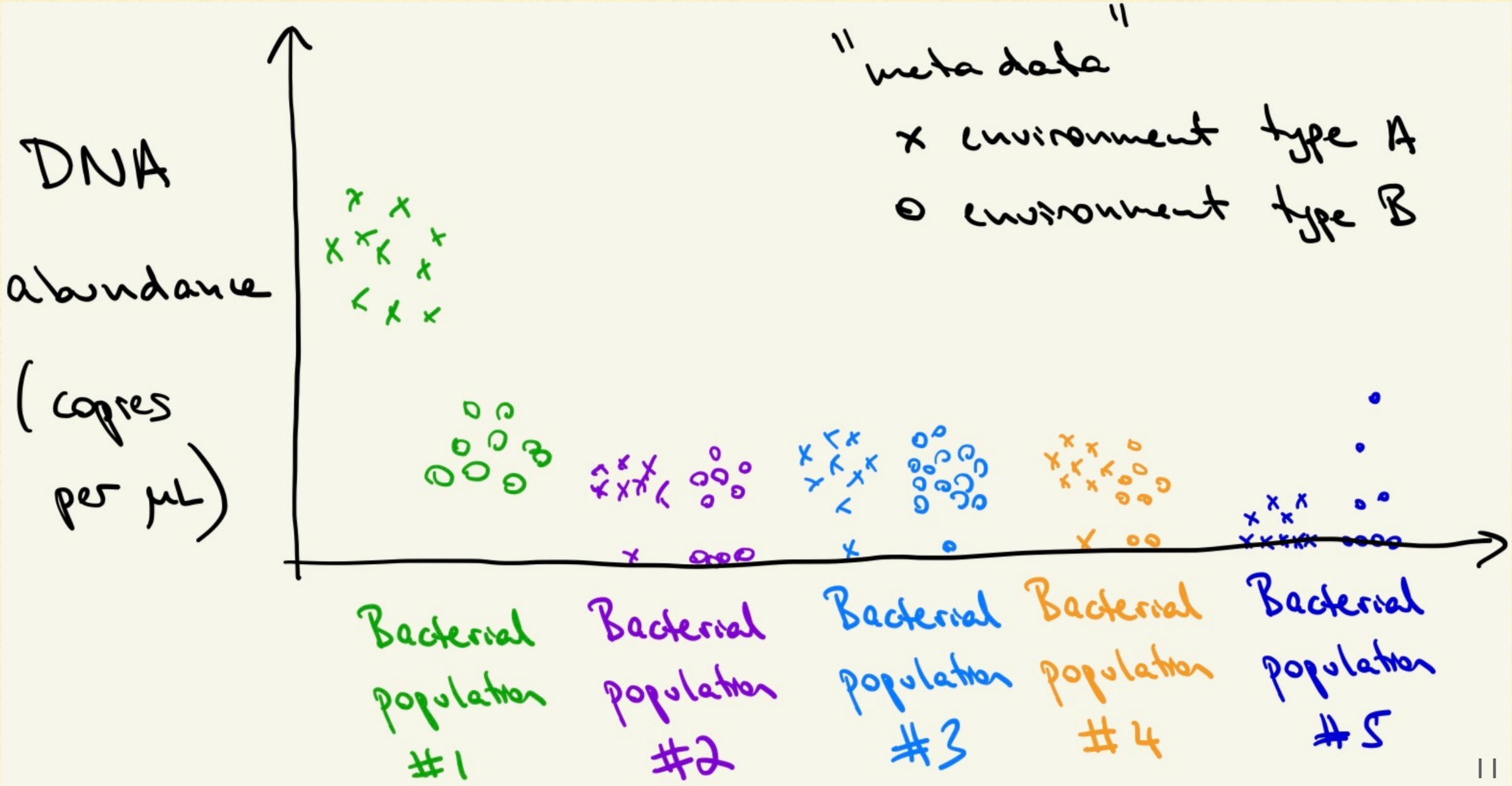
THE PEP TALK

MODELING ABUNDANCE

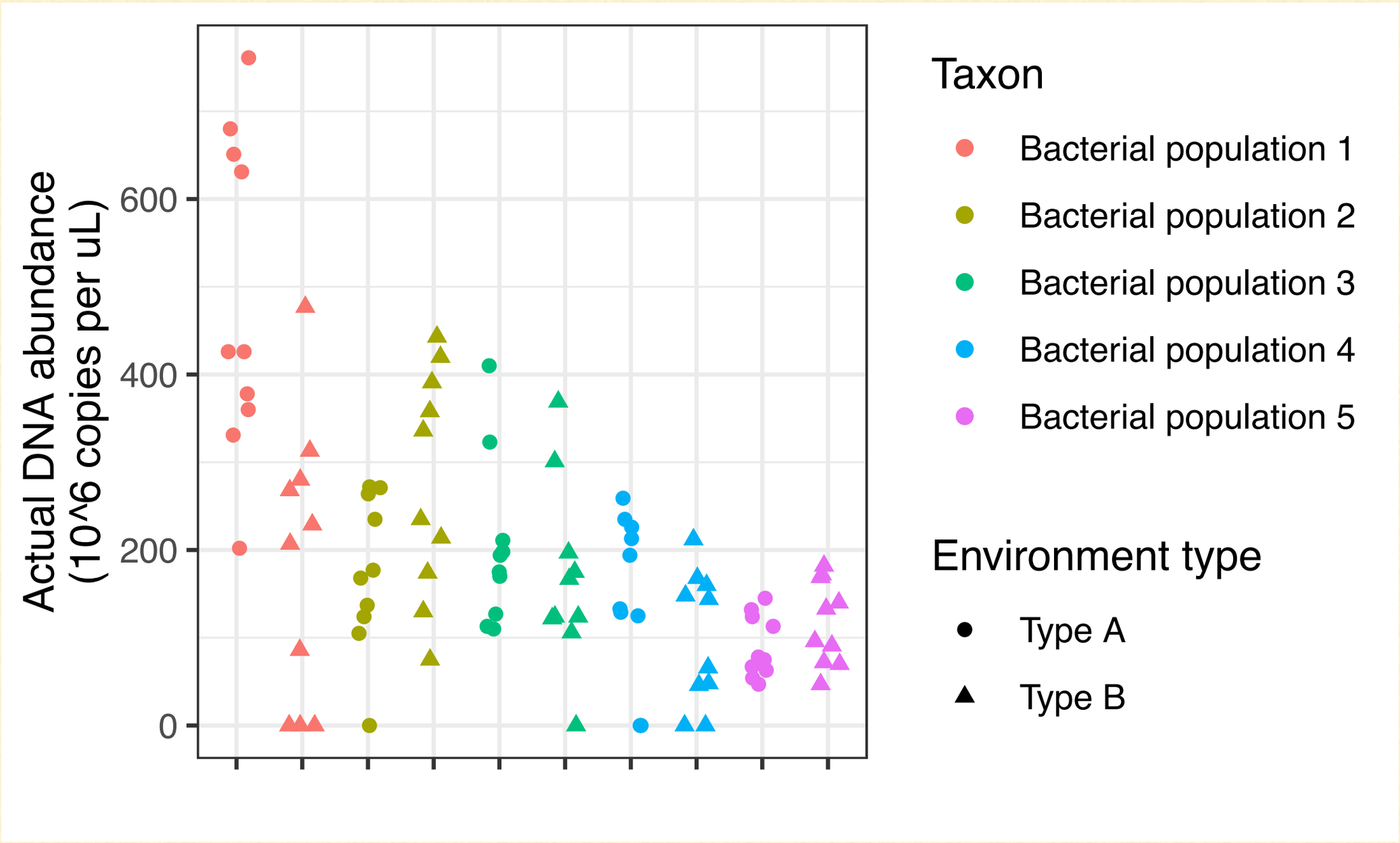
ANALYSIS

- There are many different data/sequencing types that can be used to model "abundance"
  - amplicon - count tables
  - shotgun - coverage, proportion data…
  - qPCR / ddPCR - counts/concentrations…
- The *type of data* you have impacts the *approach you need*

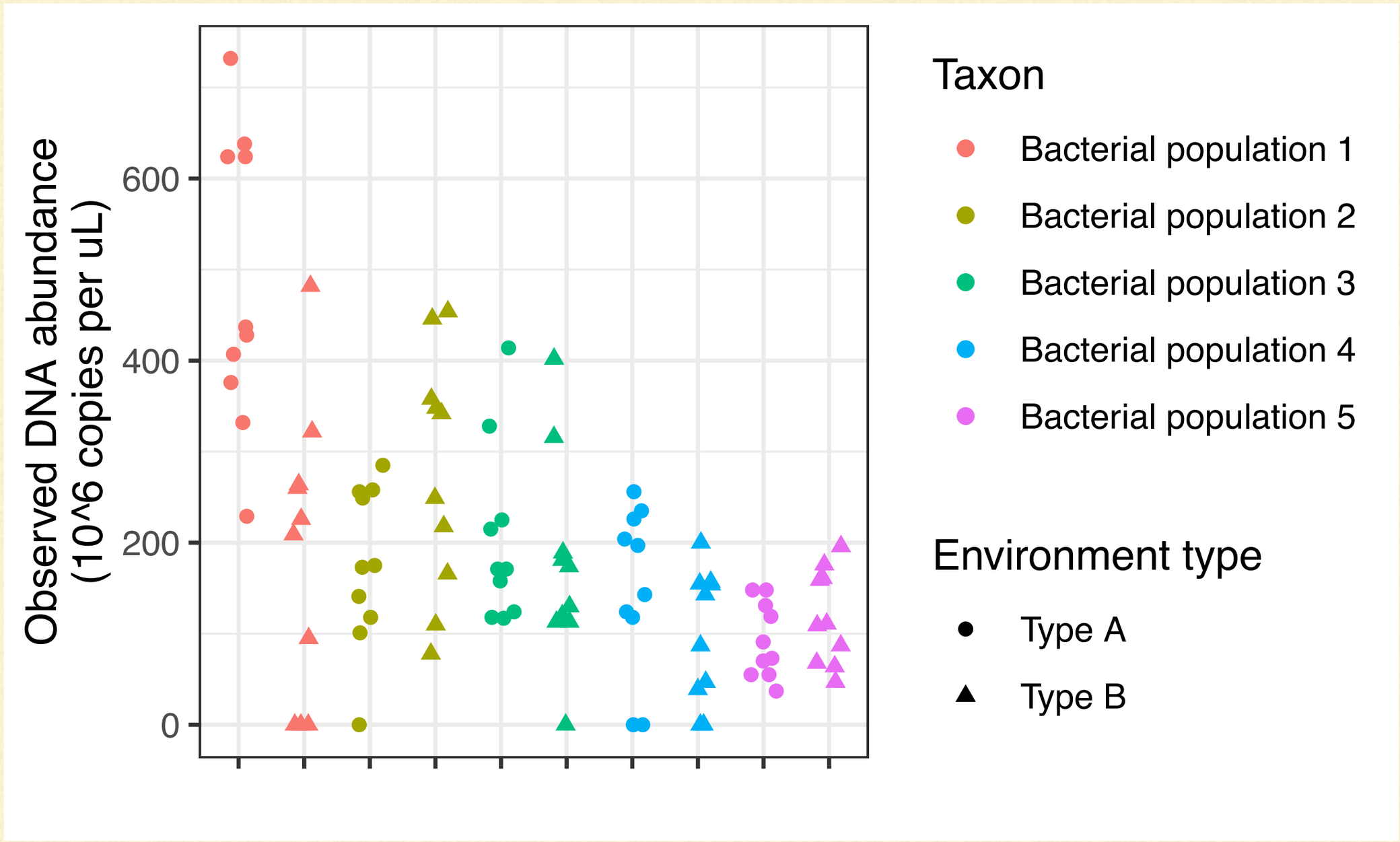
THE ENVIRONMENT



THE ENVIRONMENT



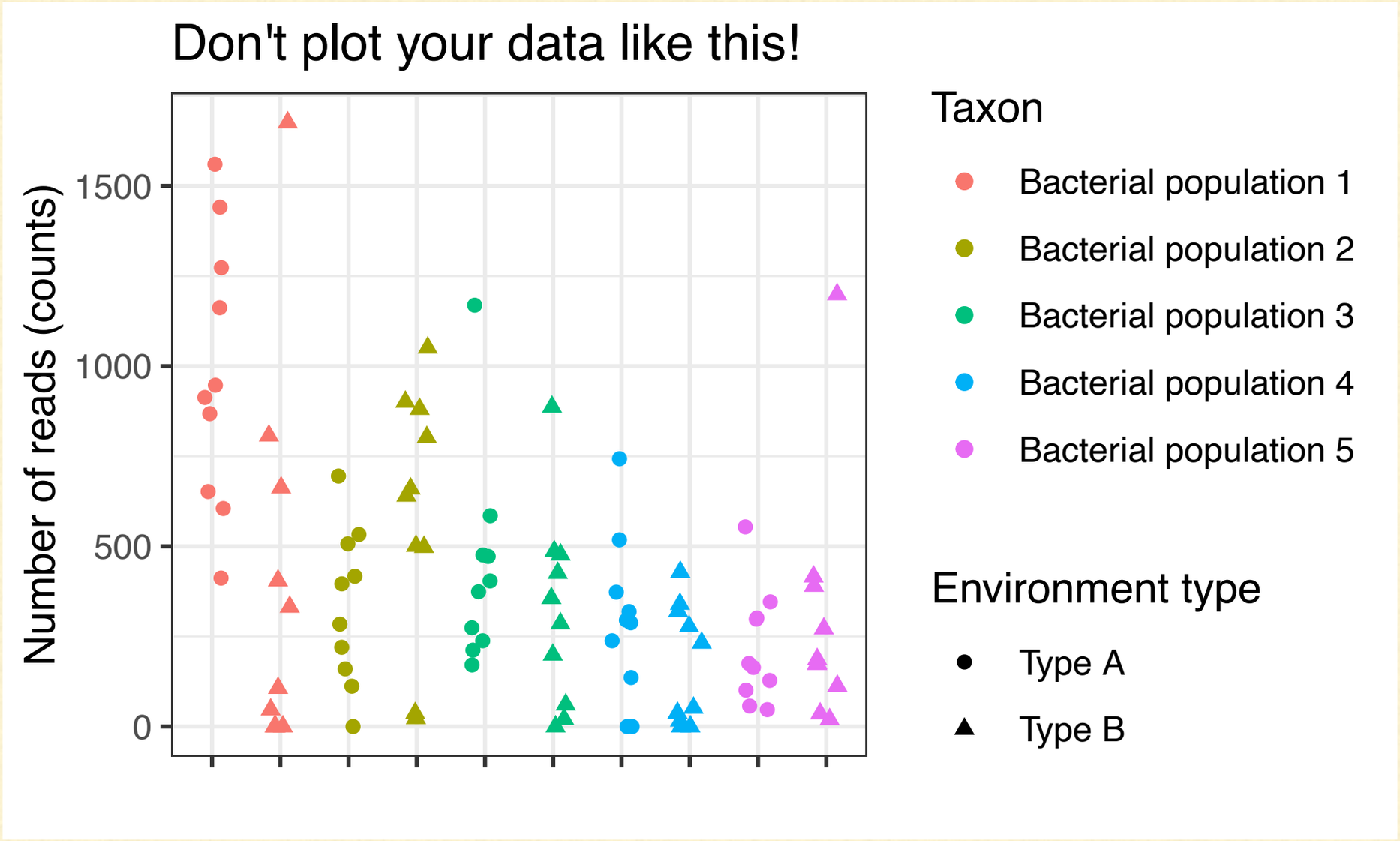
CONCENTRATION DATA*



*optimistically

HTS DATA*

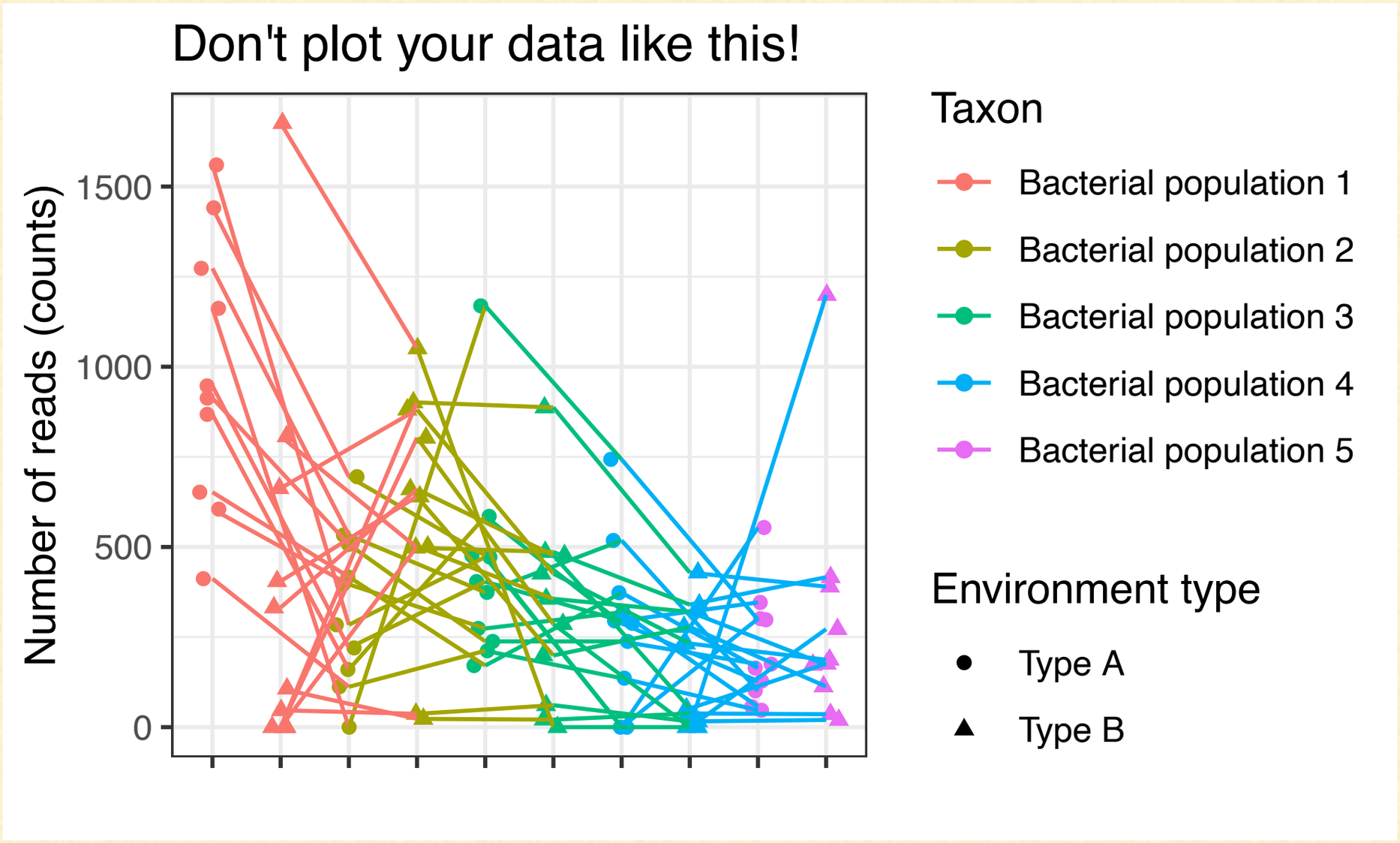
- We get a random number of reads per sample



*optimistically

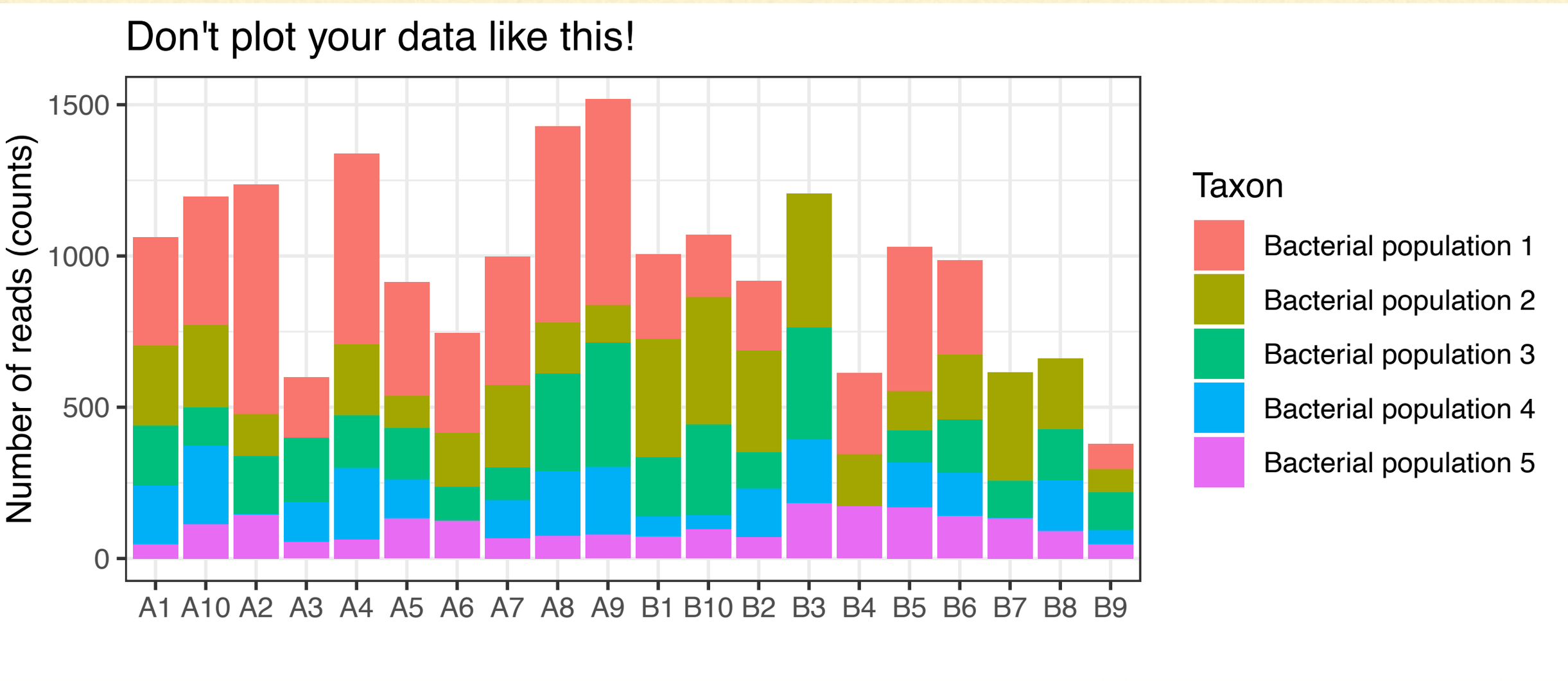
HTS DATA*

- We get a random number of reads per sample



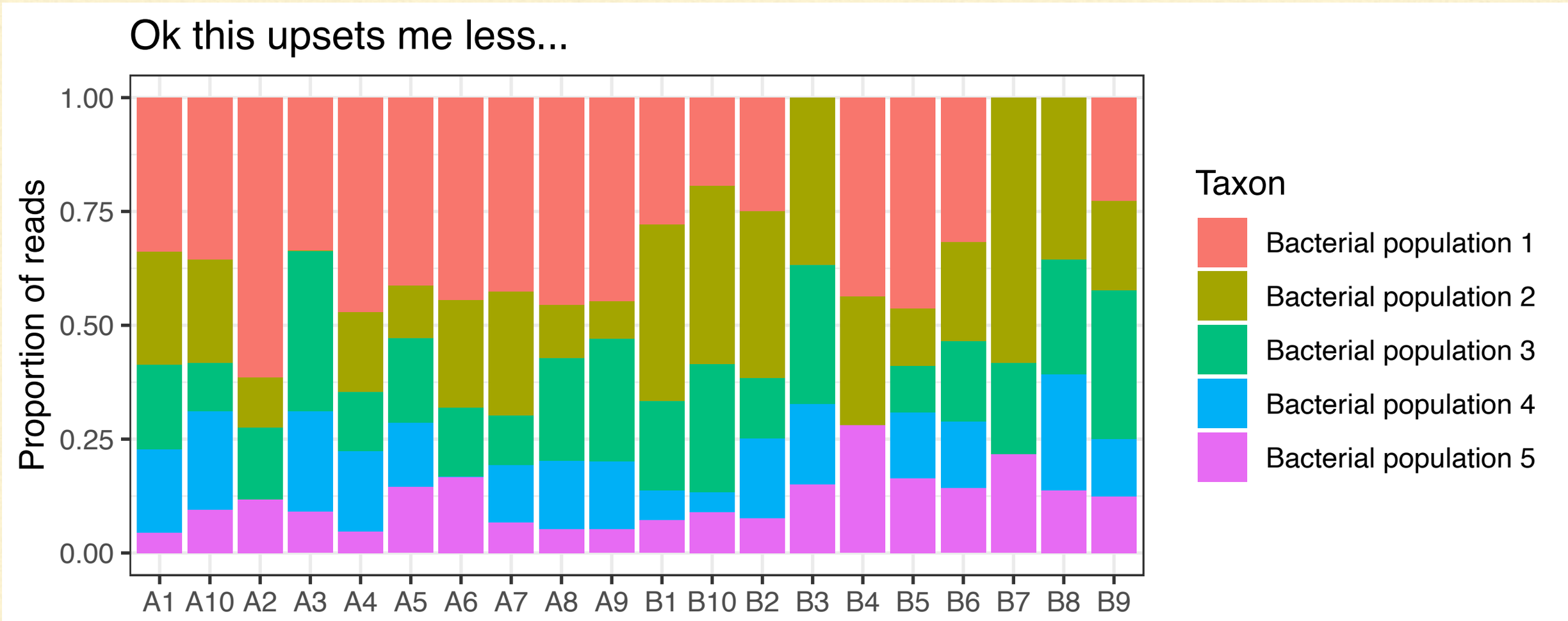
*optimistically

HTS DATA*



*optimistically

HTS DATA*



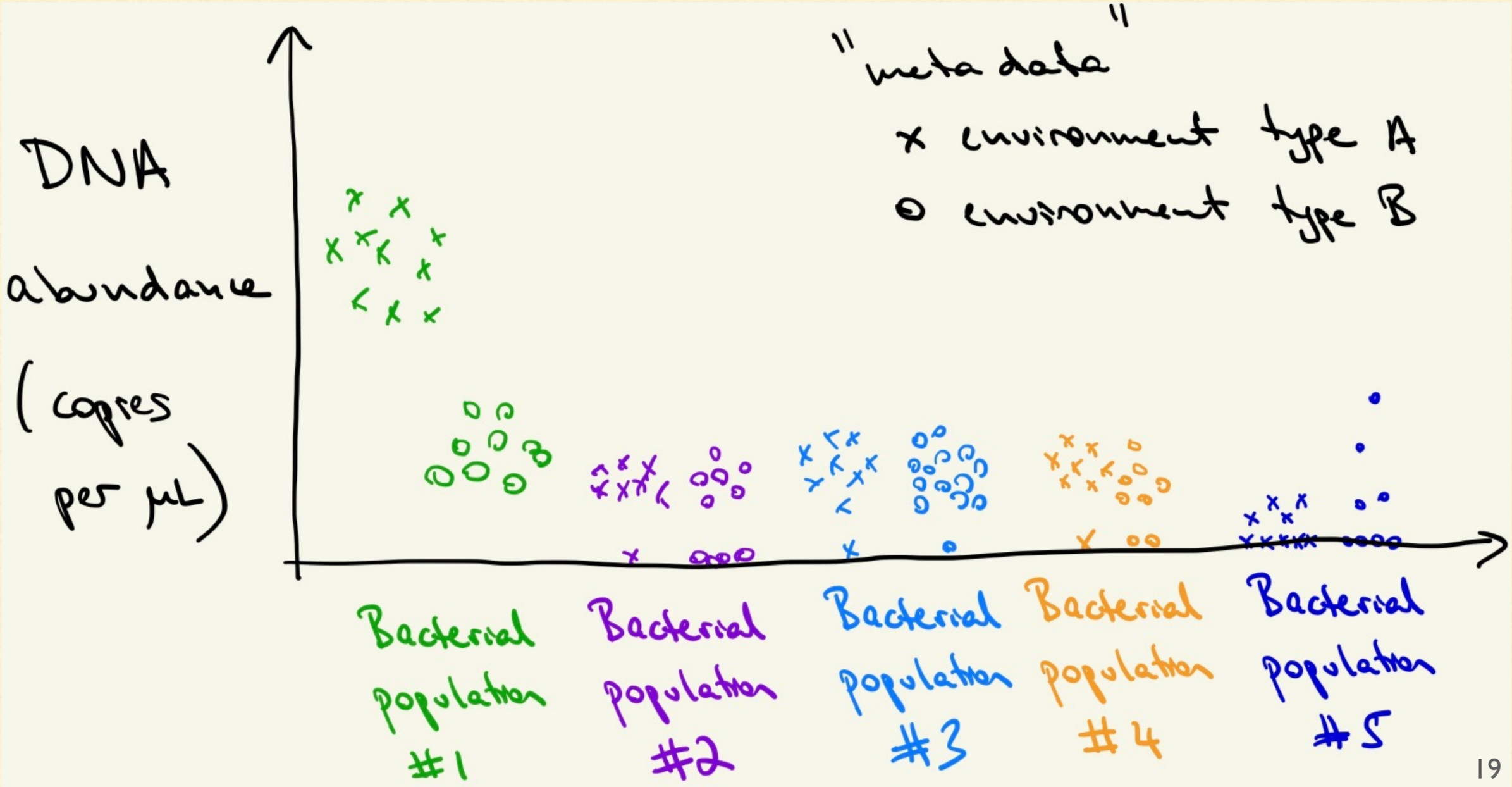
*optimistically

HTS DATA

- Some considerations

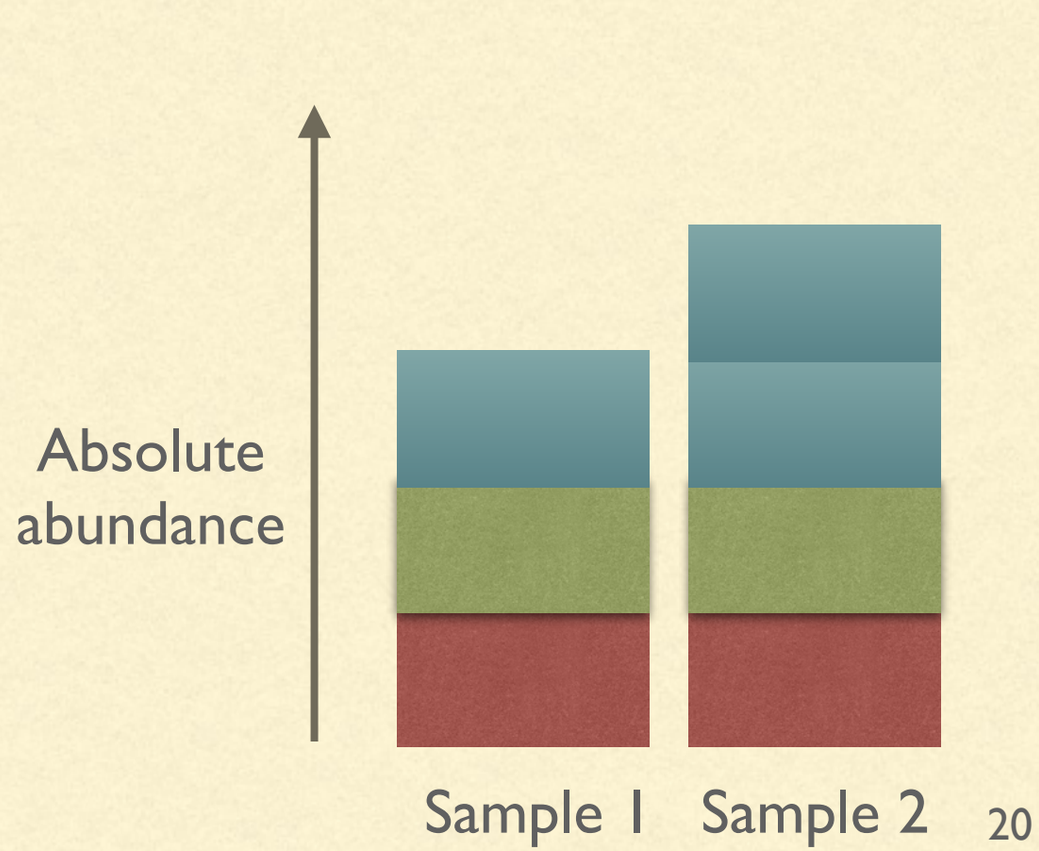
1. Total counts are random ✅
    - Modeling total counts directly is a bad idea
2. Proportions can be misleading
3. Taxa are unequally well-detected

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**
  - **0.25** / **0.25** / **0.50**
- This is an inherent limitation of *proportion-based parameters*



HTS DATA

- Some considerations

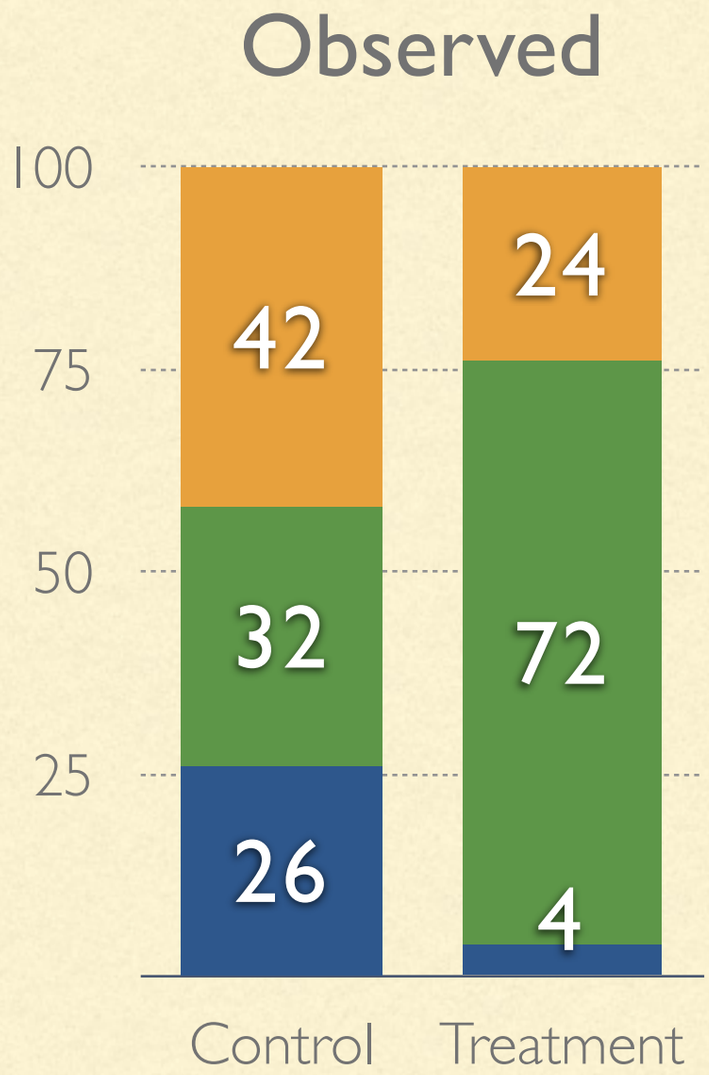
1. Total counts are random ✅
    - Modeling total counts directly is a bad idea
2. Proportions can be misleading ✅
3. **Taxa are unequally well-detected**

#3 TAXA ARE UNEQUALLY WELL-DETECTED

- As Ben talked about,

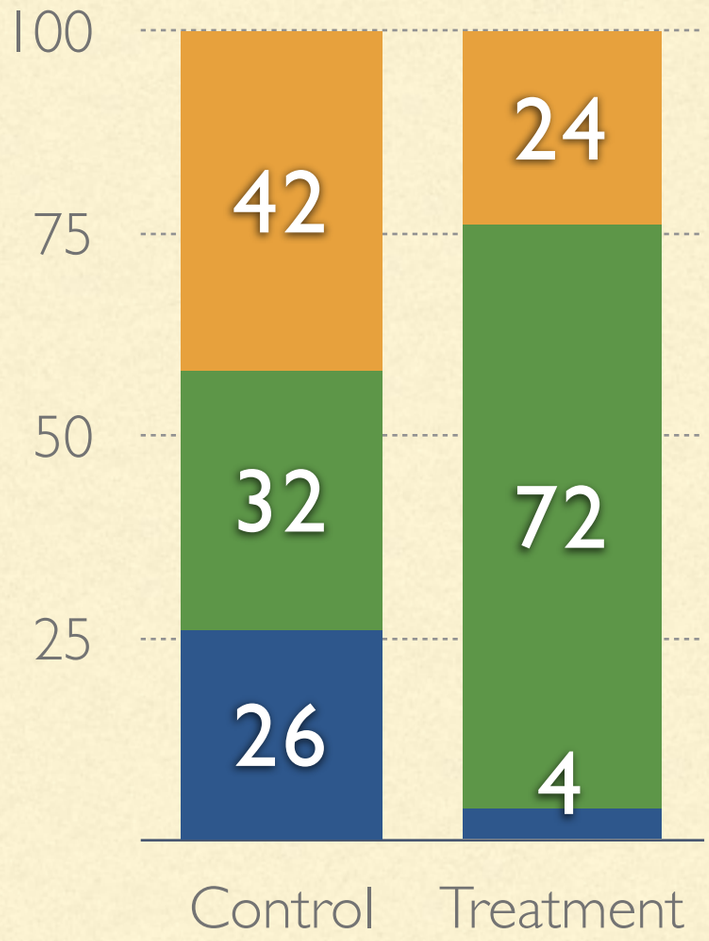
Observed relative abundance $\propto$ True relative abundance $\times$ Taxon-specific efficiencies

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

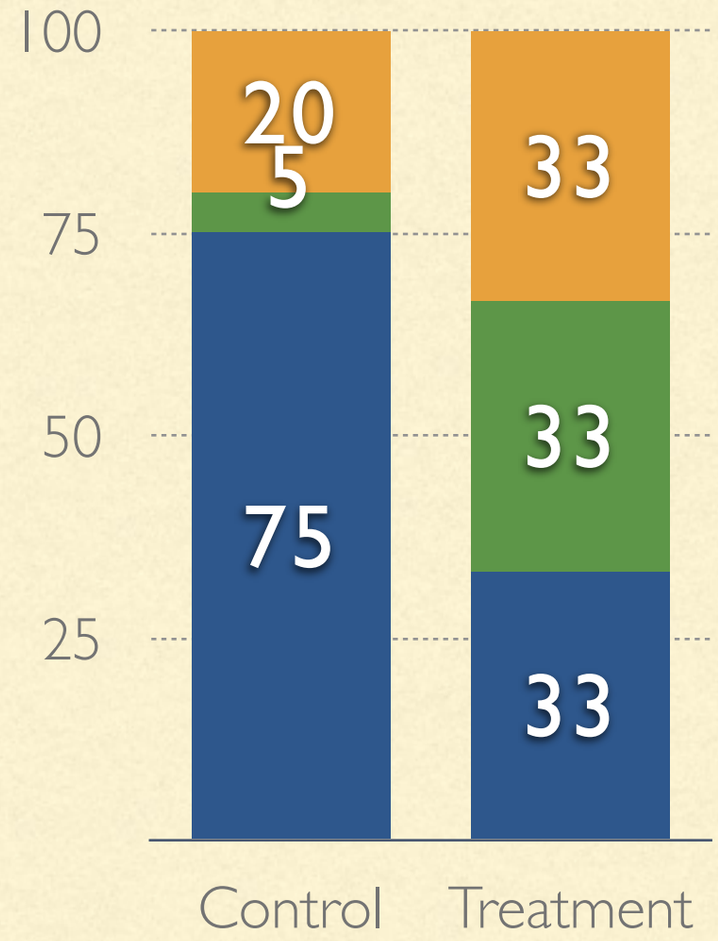


- A tempting conclusion:
  - The relative abundance of **taxon orange** *decreased* in the Treatment sample (right) compared to the Control sample (left)

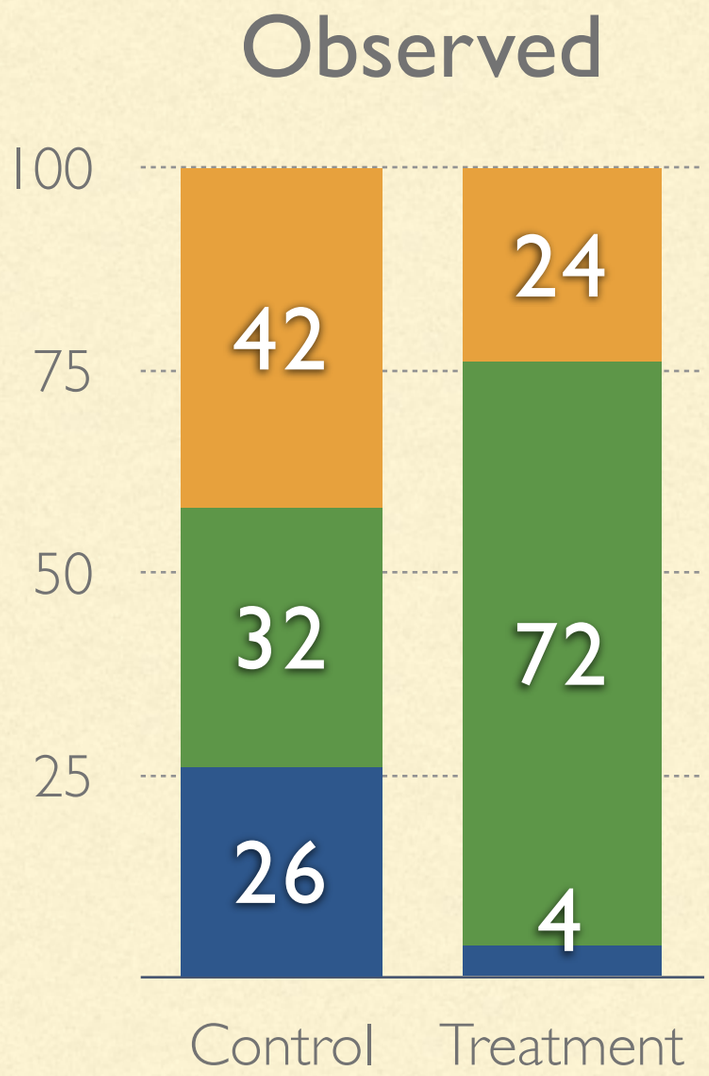
Observed

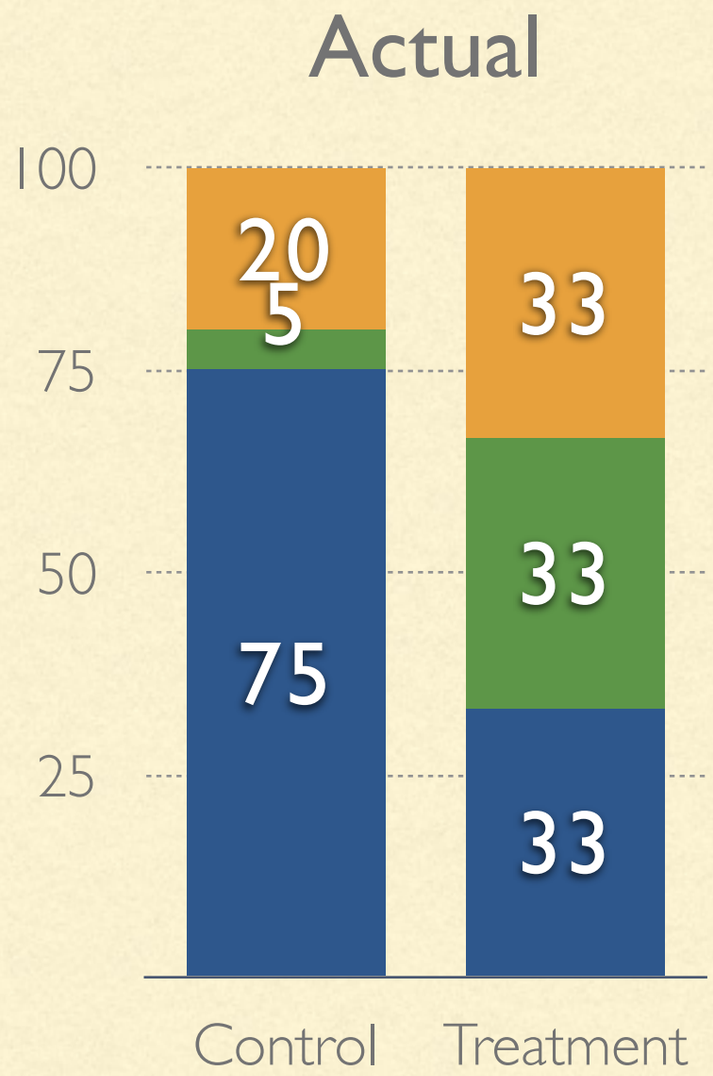


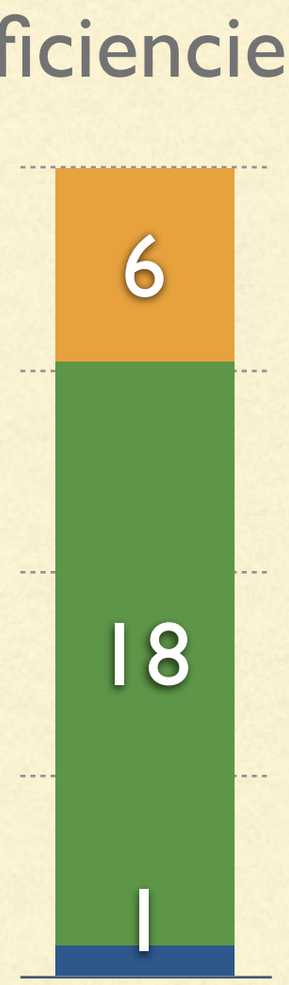
Actual



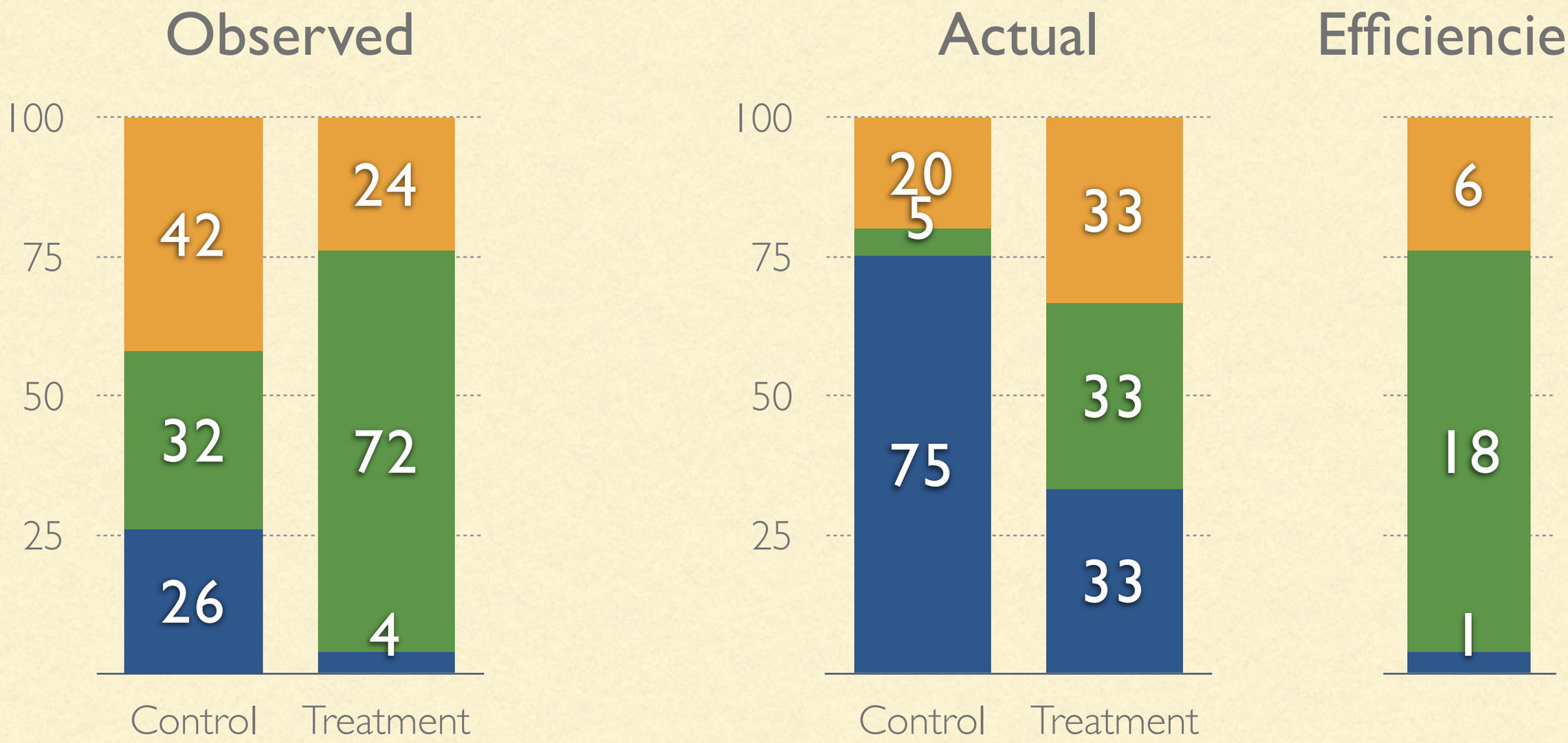
- In fact, the relative abundance of **taxon orange** *increased* in the Treatment sample compared to the Control sample







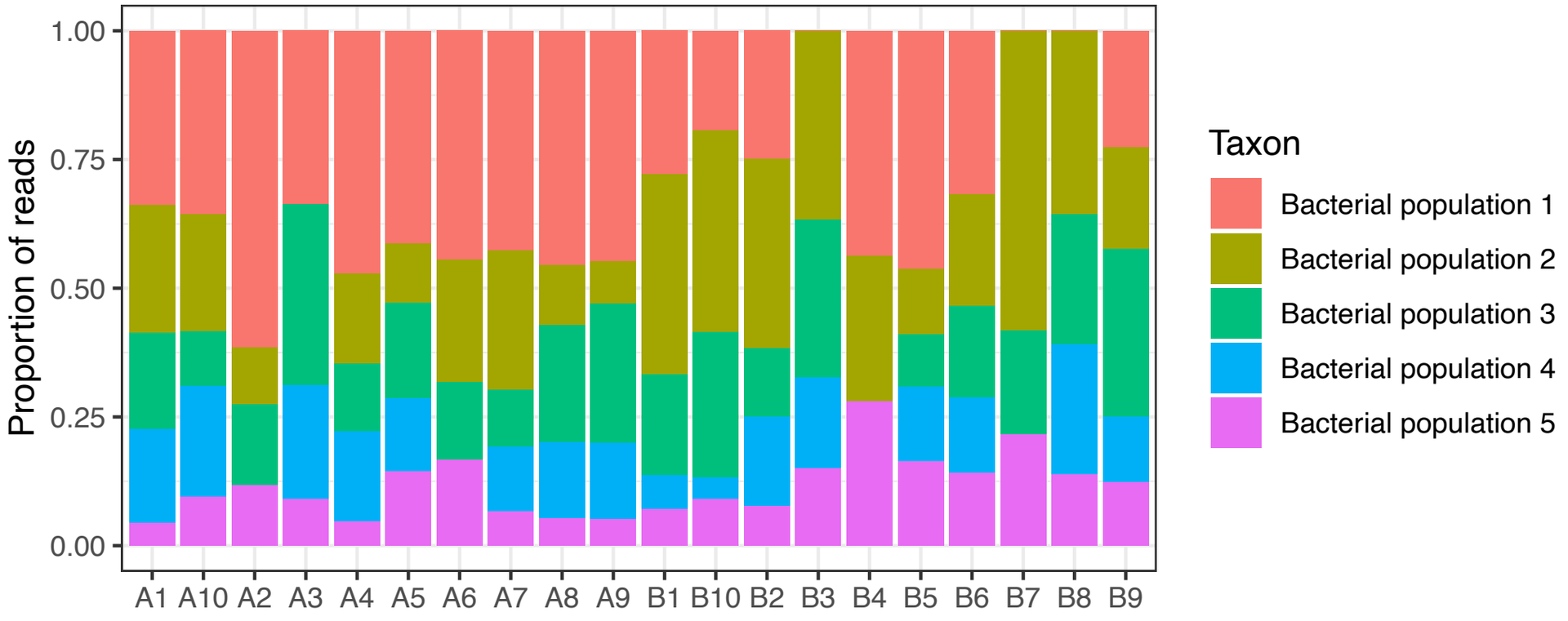
- In fact, the relative abundance of **taxon orange** *increased* in the Treatment sample compared to the Control sample

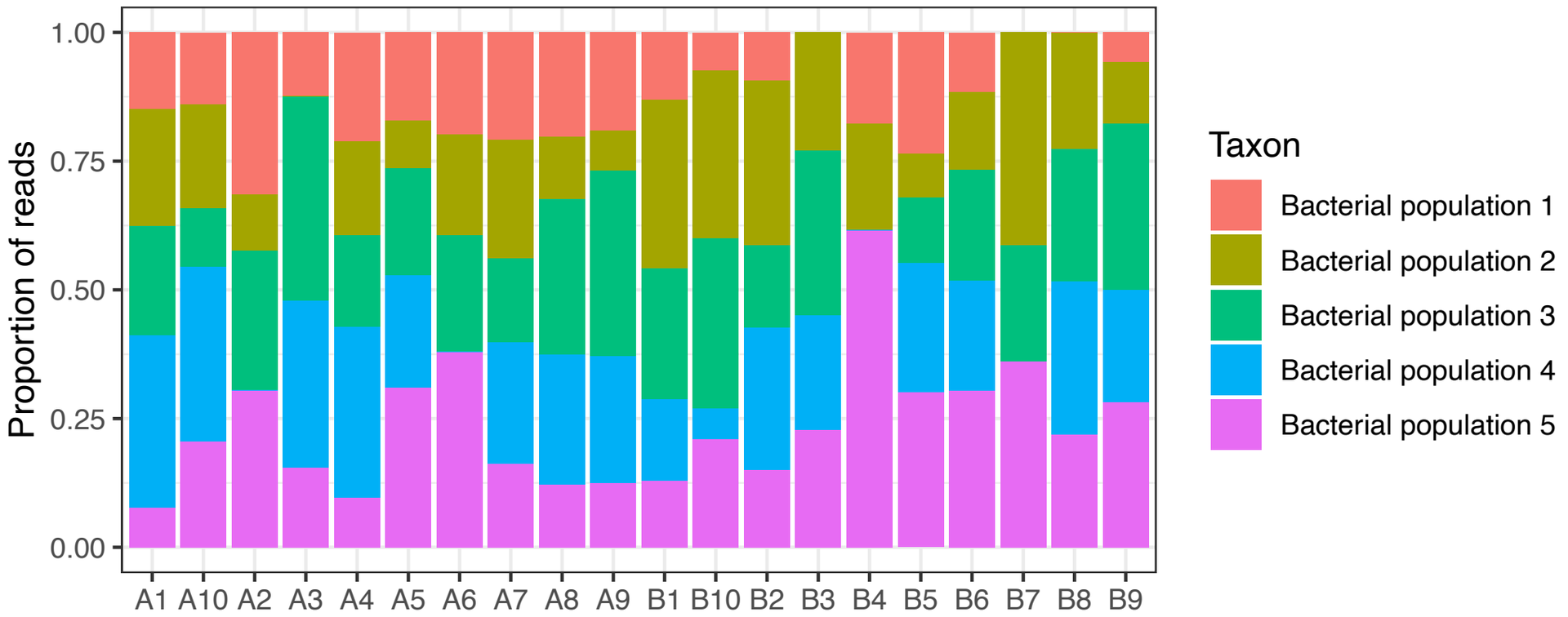


- **Taxon green** is high efficiency; its abundance increased (C vs T).
  Taxon blue is low efficiency, and its abundance decreased.

- **Taxon orange**'s abundance depends on the abundance of **the other taxa**.

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


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


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

SUMMARY OF CHALLENGES IN MODELING HTS DATA

- Some modest simplifications
  - Concentration data *can* be compared across samples
    - Can't really be compared across taxa
  - HTS counts & coverages *cannot* be compared across samples nor taxa
  - HTS proportions *cannot* be compared across samples
    - In the absence of calibration control data
  - What can be compared? Ratios and fold differences…
    - More discussion after the break ☕🍪

MODELING ABUNDANCE

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

ABSOLUTE ABUNDANCE DATA

- qPCR and ddPCR data can usually be modeled with regression techniques you can about learn in Applied Stats 101

- **Linear models** most generally look like

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

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

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

- 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
```

ACCESSING `LM` LAB

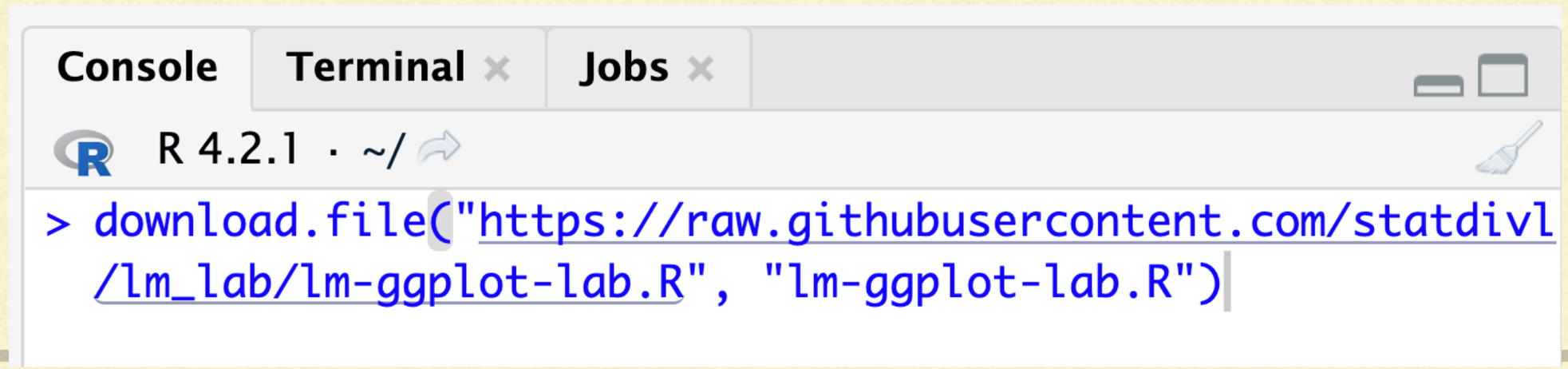
1. Go to Schedule on Wiki to Wednesday afternoon, click on "Statistics Labs"

2. *Copy the command* under LM LAB

```
lm lab:

download.file("https://raw.githubusercontent.com/statdivlab/stamps2023/main/labs/lm-lab/lm-g
```

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



MODELING ABUNDANCE

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

MODELING HTS DATA

- A common scientific 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."*

MODELING HTS DATA

- Many methods exist
  - ALDEx2
  - ANCOM
  - 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

MODELING HTS DATA

- Many methods exist
  - ALDE
  - ANC
  - cornc
  - DESe
- multiple methods

These methods estimate different parameters.

They are *not* directly comparable.

You need to decide what parameters *you* care about.

om

/t-tests on

ns

ratios

most all

MODELING HTS DATA

- Some flavors
  - Proportions
  - Ratios-of-ratios
  - Fold-changes

MODELING HTS DATA

- Many methods exist
  - ALDEx2
  - ANCOM
  - 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

MODELING HTS DATA

- We don't have time to talk about *all* of these, so I'm going to talk in depth about *my favourite*…
  - …. and the lab will walk you through some less modern, but popular alternatives!

MODELING HTS DATA

- Many methods exist
  - **ALDEx2**
  - **ANCOM**
  - 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

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

- radEmuAbPill

- 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 is a differential abundance method that estimates fold differences in "absolute abundance" from sequencing data

$$\text{Fold difference in F. prauznitzii} = \frac{\text{mean DNA conc. F. prauznitzii in cases}}{\text{mean DNA conc. F. prauznitzii in controls}}$$

RADEMU

- Advantages
  - Estimates something about the *environment*, not something about *sequencing*
  - Robust to differential detection
  - Controls Type I 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

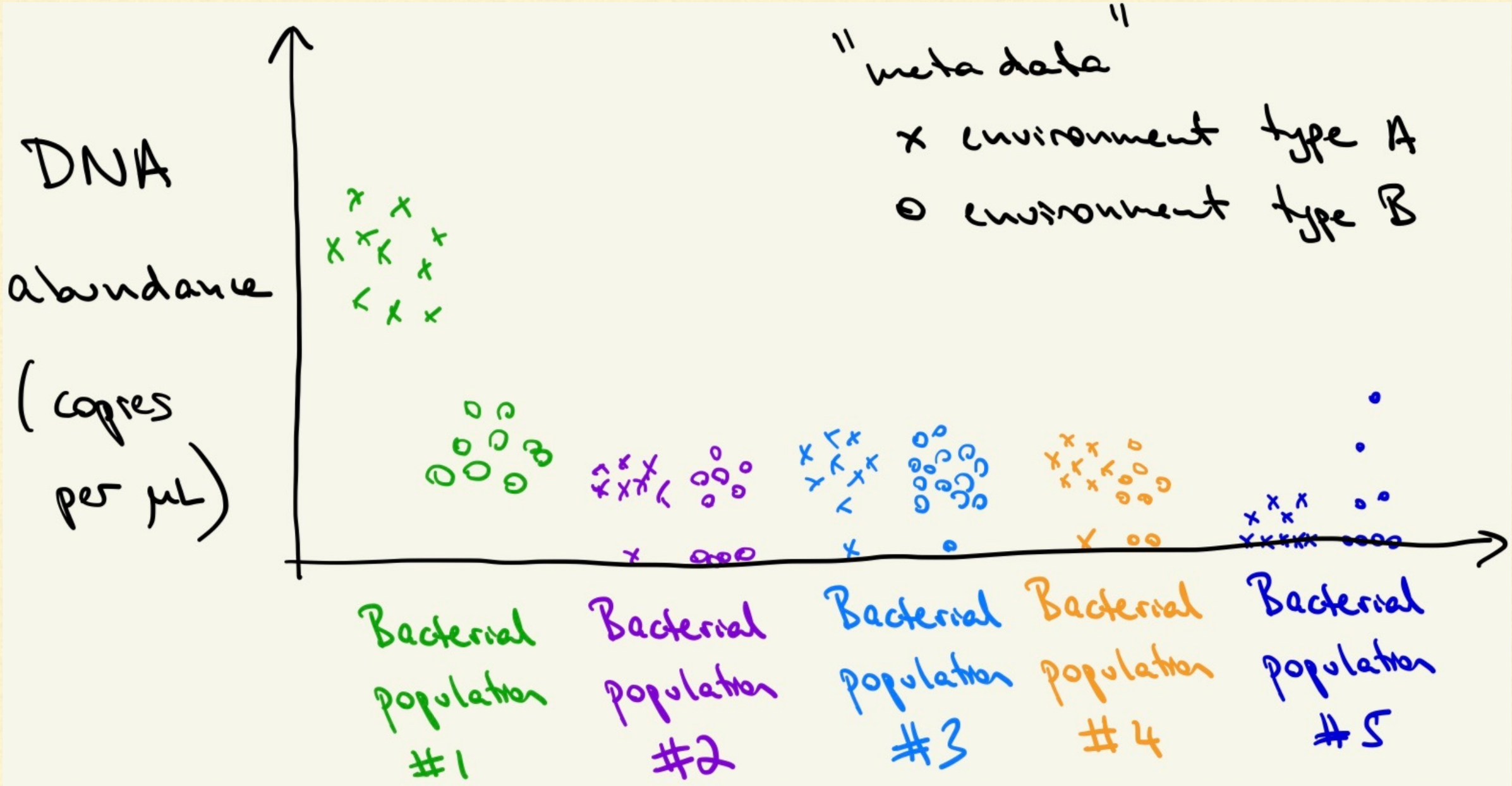
RADEMU

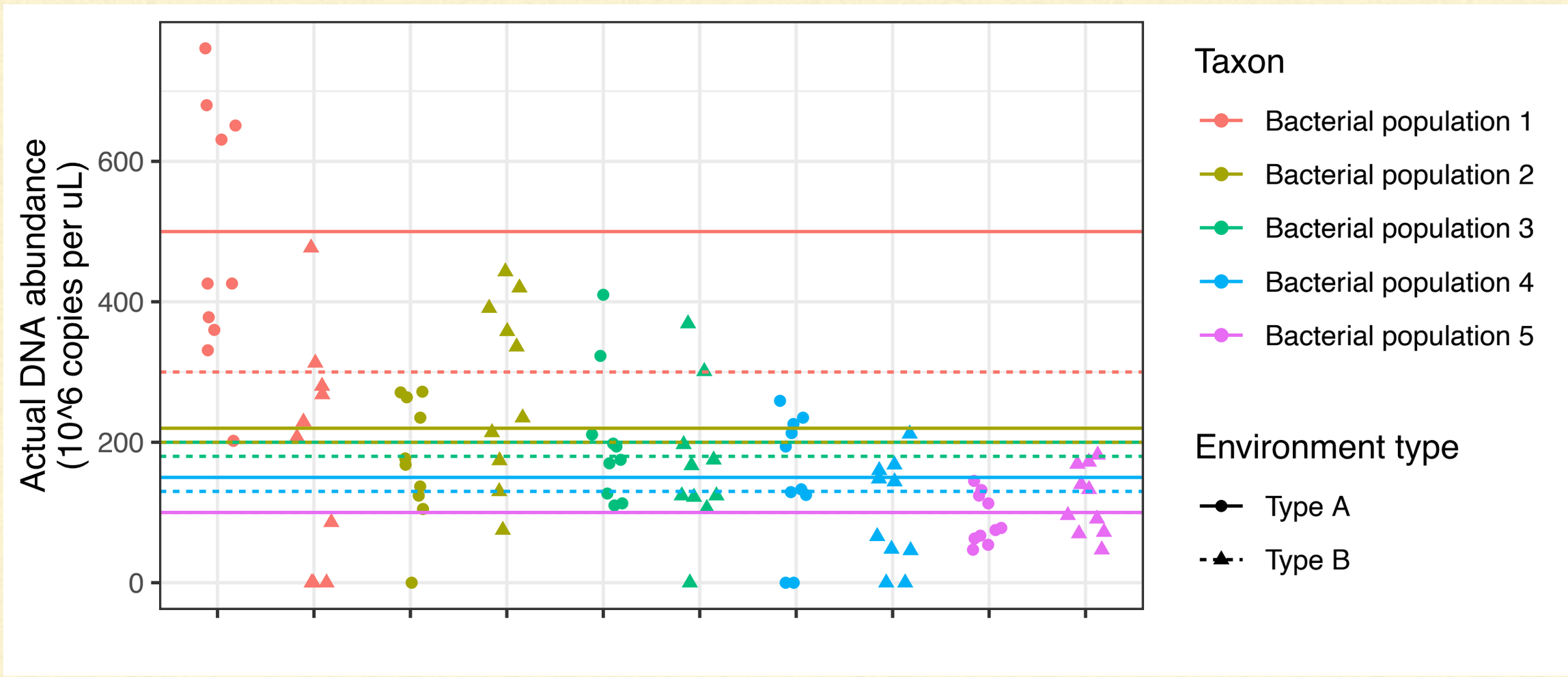
- Assumption
  - The average taxon isn't changing in its abundance
    - Advantage: you decide what an "average" taxon means
    - (Personal opinion — this assumption is valid in almost all relevant comparisons)
- Limitation
  - Doesn't scale to 1000's of taxa… *yet…*

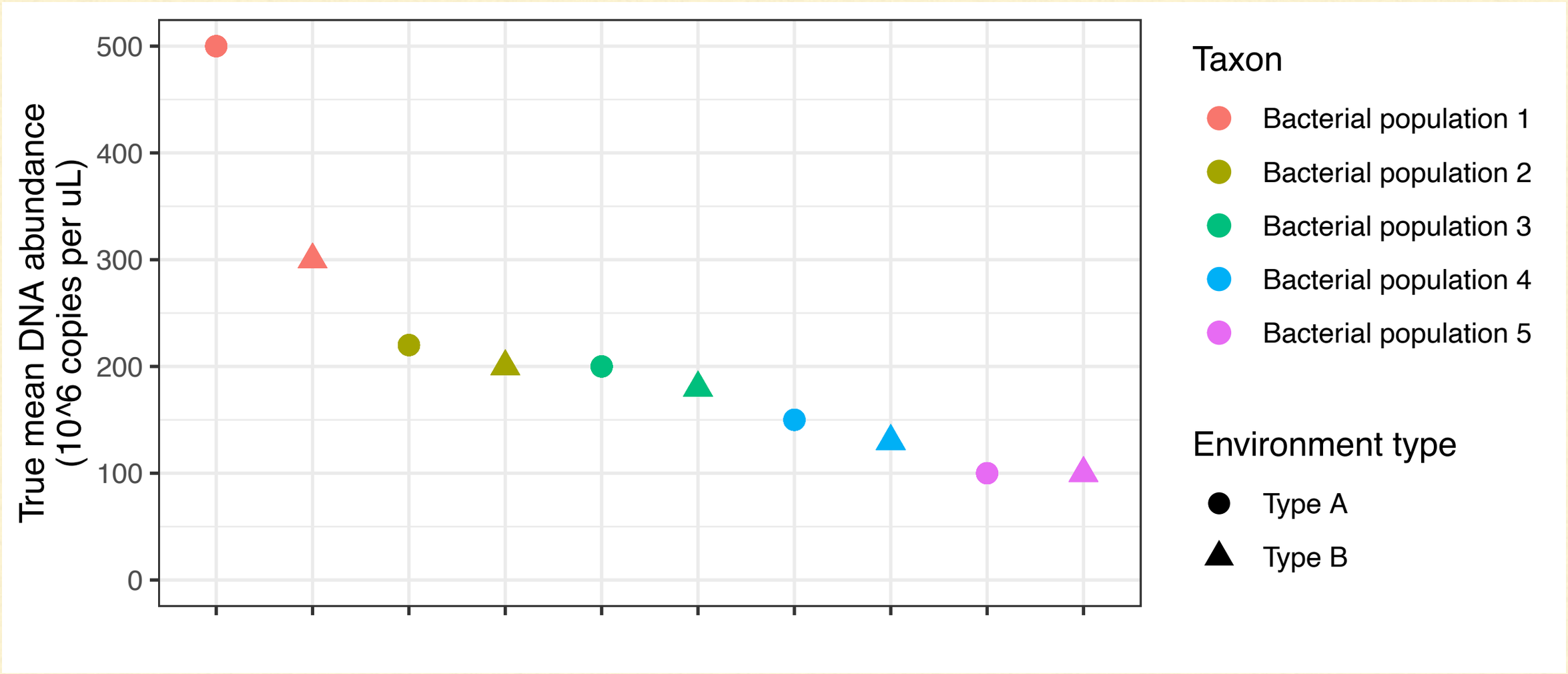
RADEMU

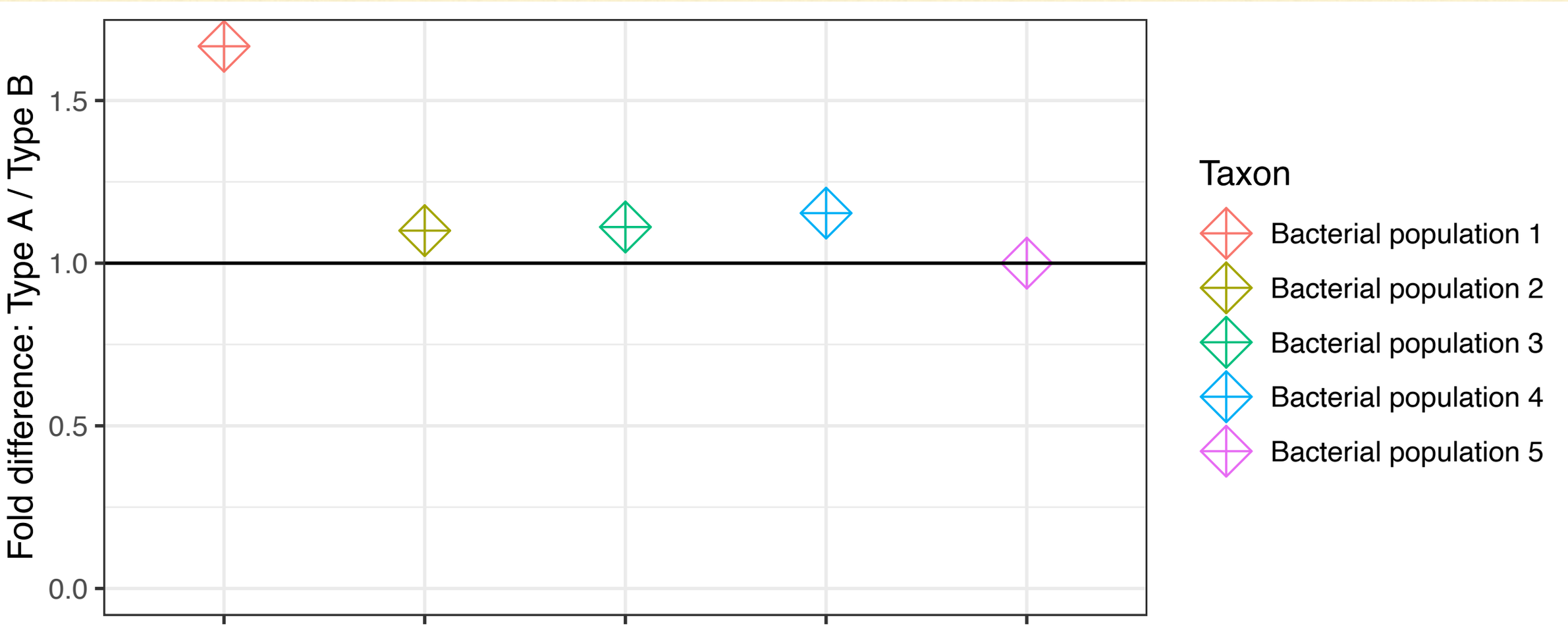
- Basic idea
  - We *have* high throughput sequencing data
    - We *want* to say something about an "absolute" quantity
      - e.g., cell concentration, DNA concentration…
    - Can we?

THE ENVIRONMENT

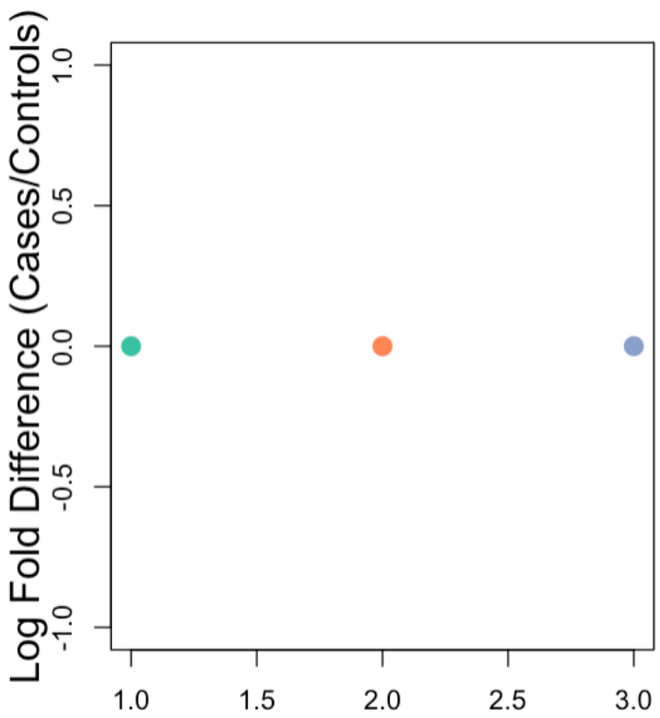


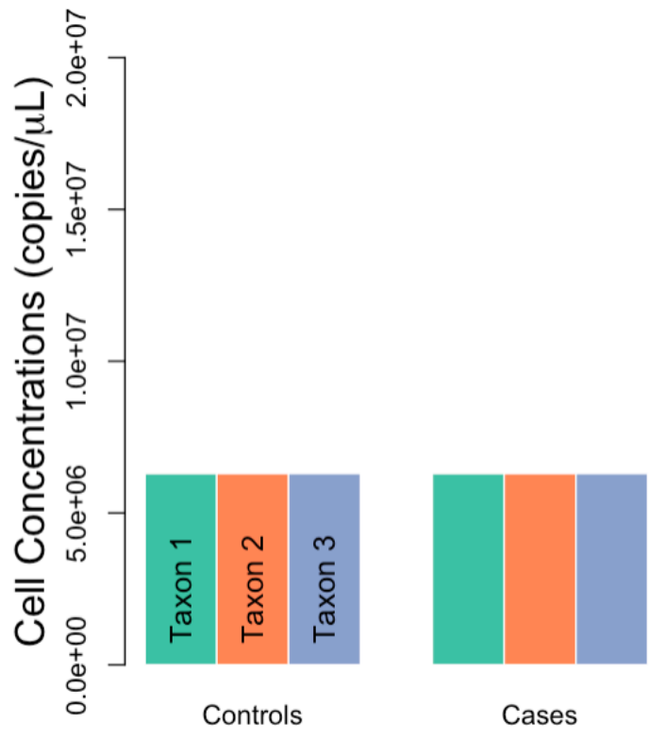

Actual DNA abundance
(10^6 copies per uL)
600
400
200
0
Taxon
Bacterial population 1
Bacterial population 2
Bacterial population 3
Bacterial population 4
Bacterial population 5
Environment type
Type A
Type B

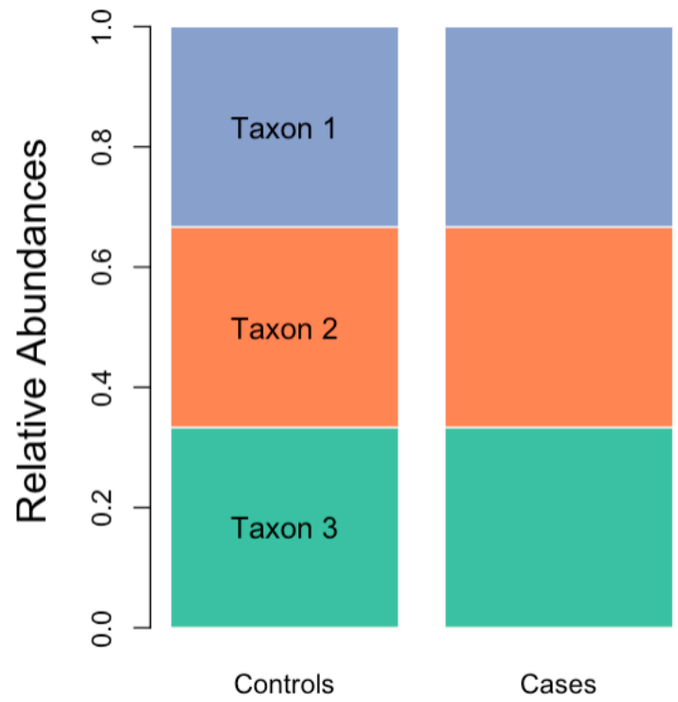

True mean DNA abundance
(10^6 copies per uL)
500
400
300
200
100
0
Taxon
Bacterial population 1
Bacterial population 2
Bacterial population 3
Bacterial population 4
Bacterial population 5
Environment type
Type A
Type B


Fold difference: Type A / Type B
1.5
1.0
0.5
0.0
Taxon
Bacterial population 1
Bacterial population 2
Bacterial population 3
Bacterial population 4
Bacterial population 5

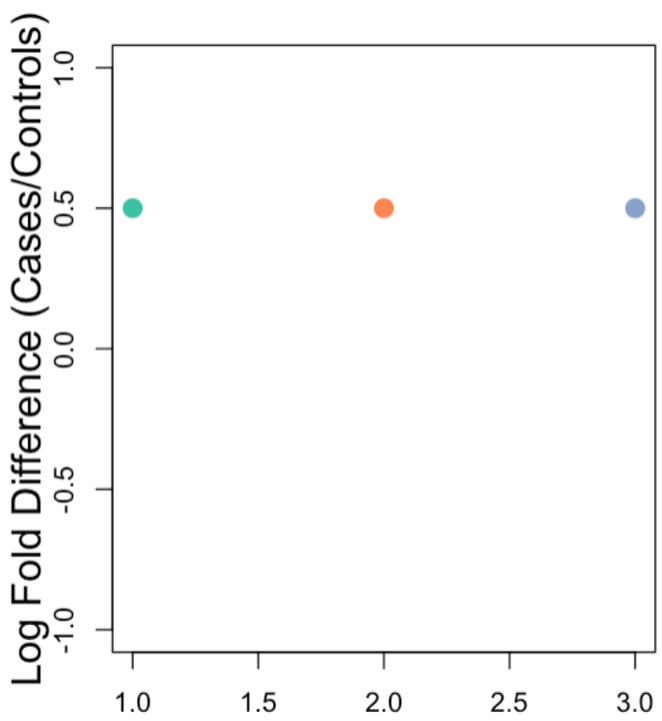
FOLD DIFFERENCES

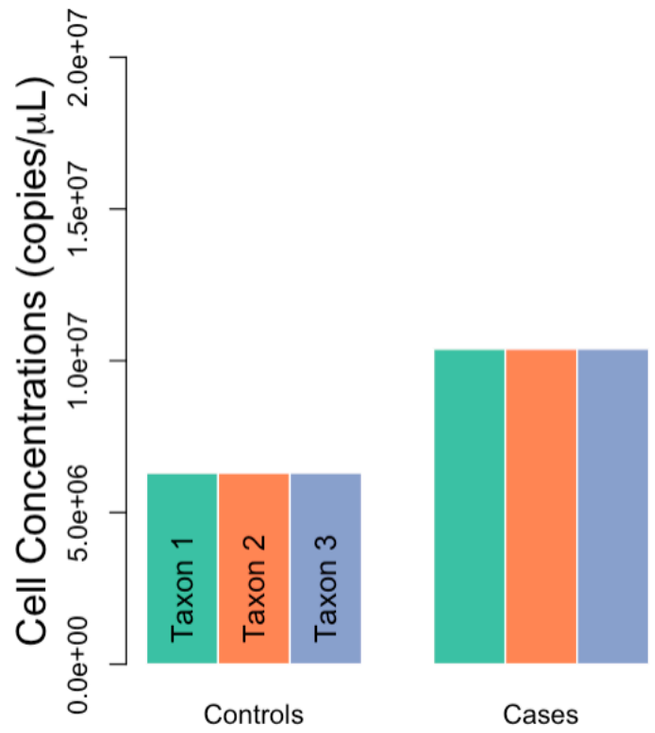


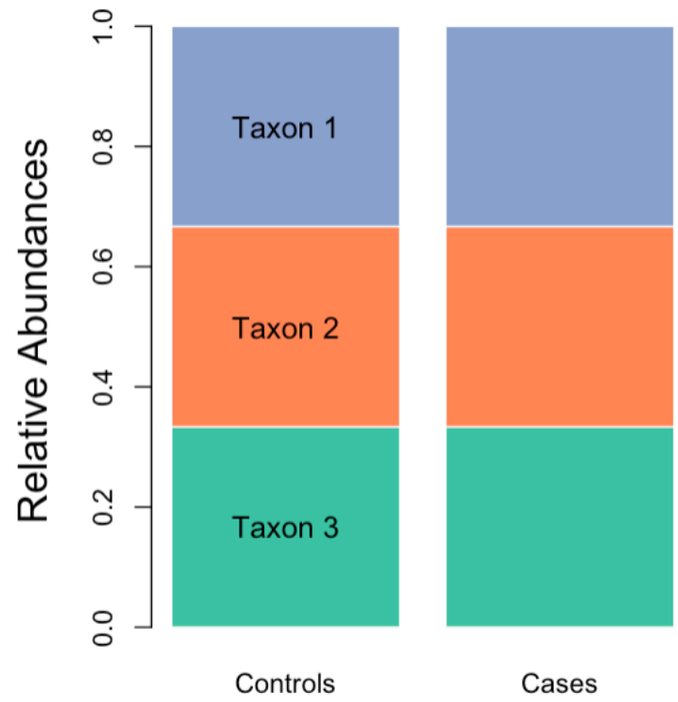




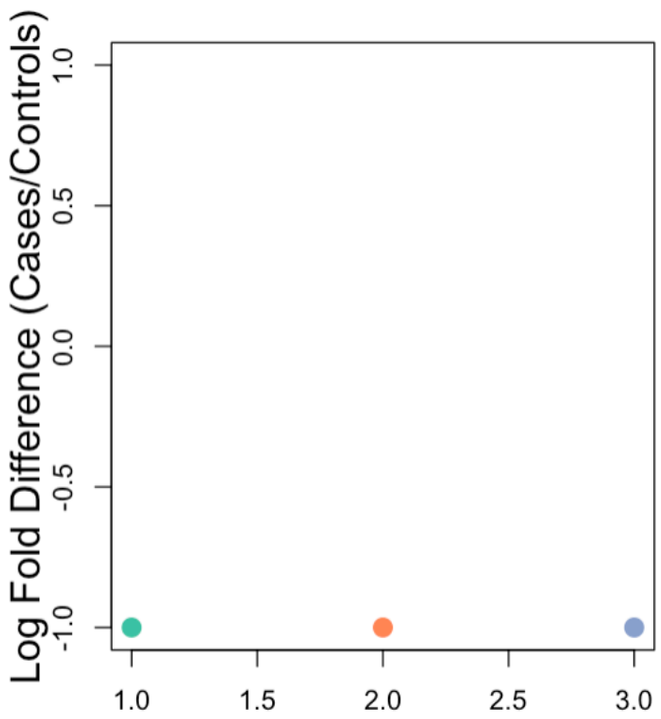
FOLD DIFFERENCES

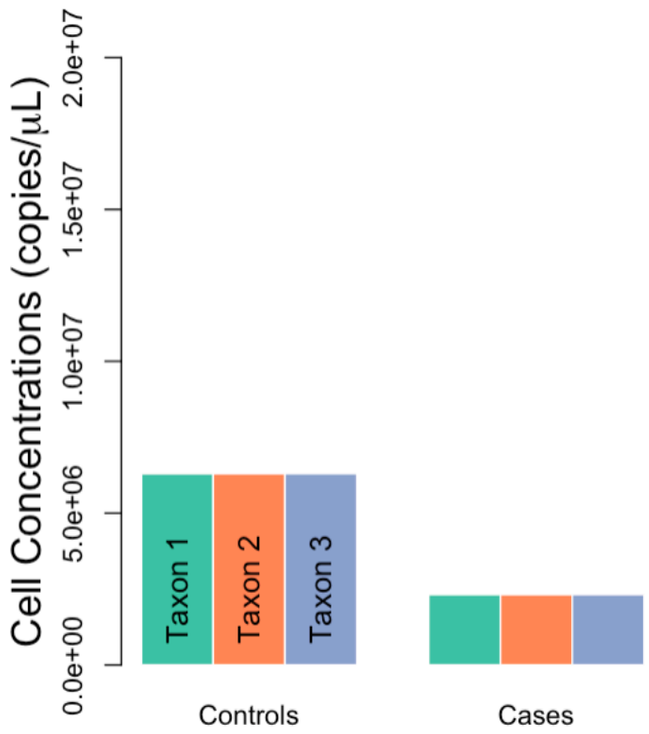


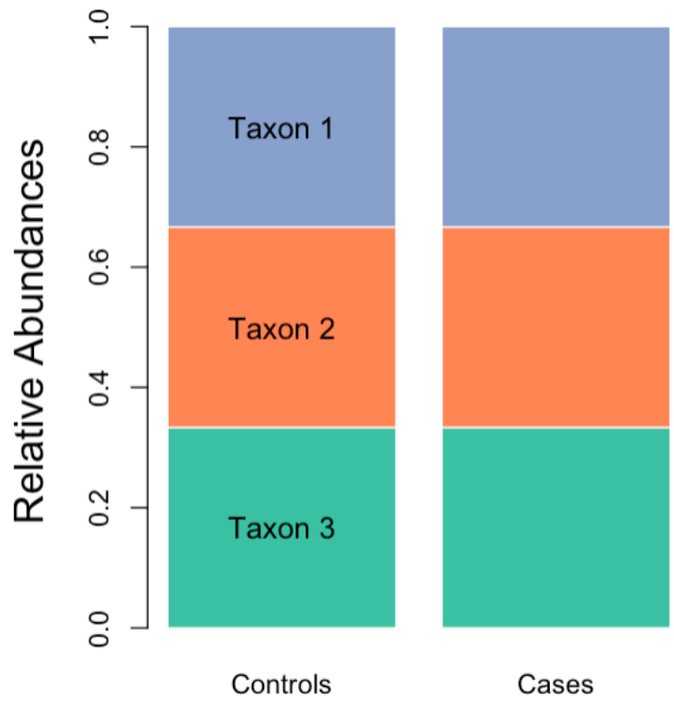




FOLD DIFFERENCES

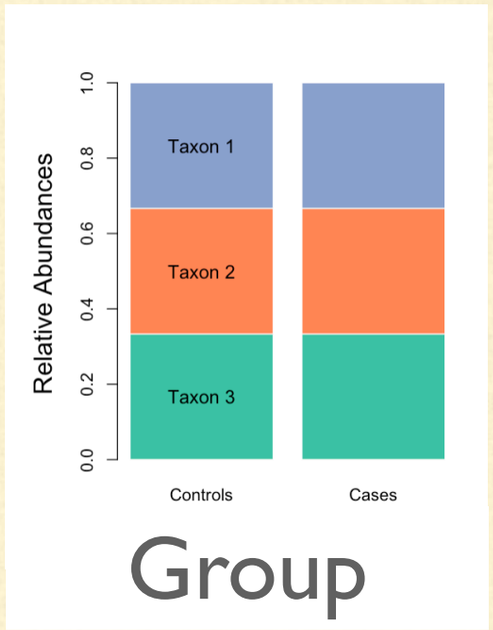


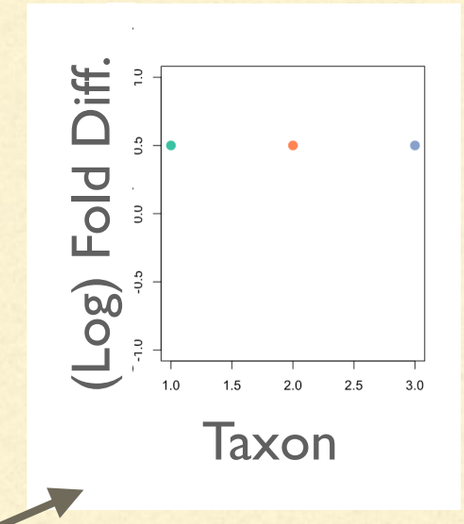




STARTING FROM PROPORTIONS

We observe

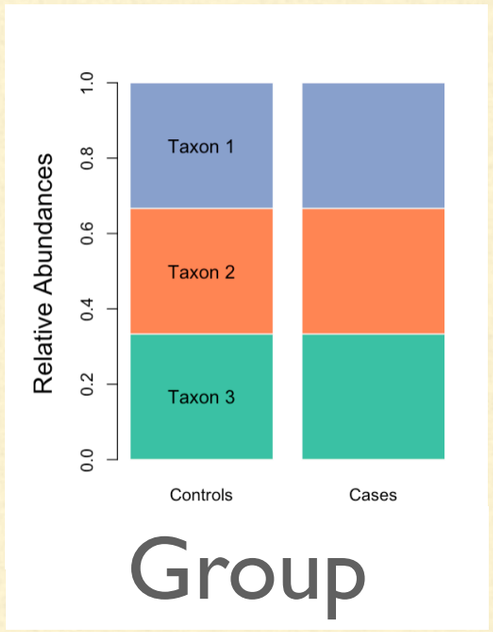




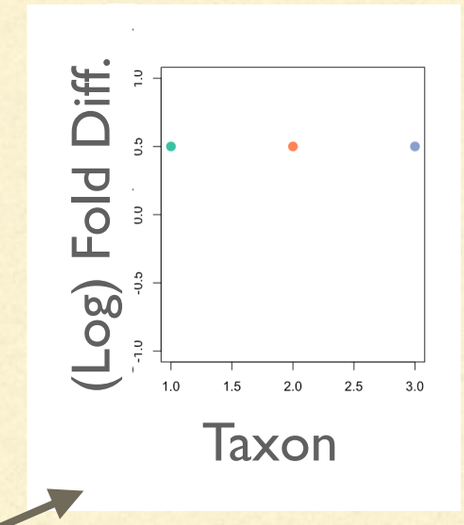
Consistent with this

STARTING FROM PROPORTIONS

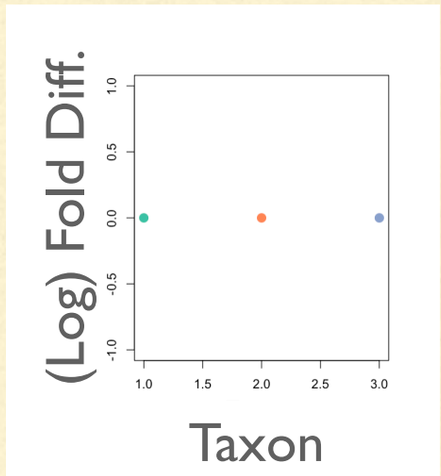
We observe



Consistent with this

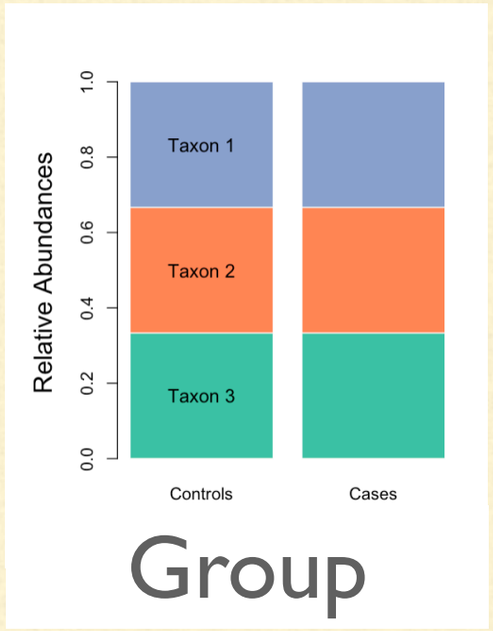


Or this

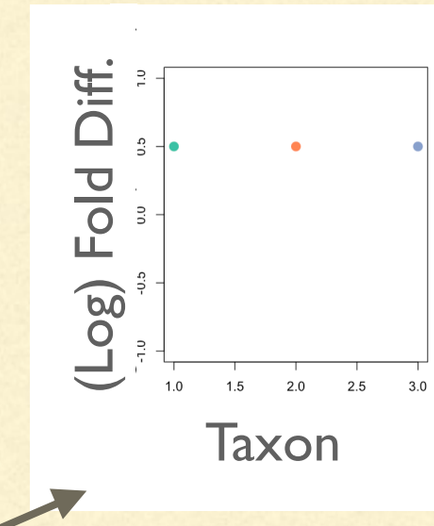


STARTING FROM PROPORTIONS

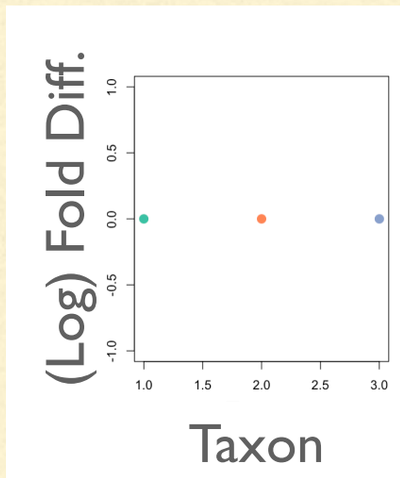
We observe



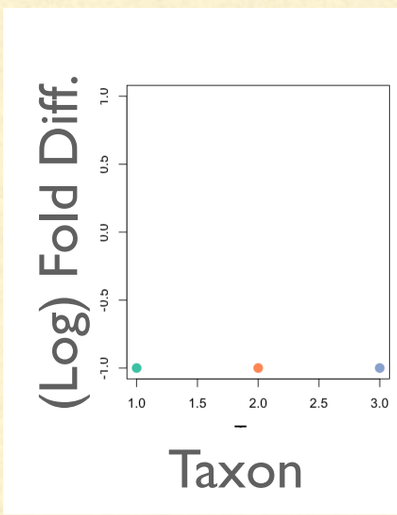
Consistent with this



Or this

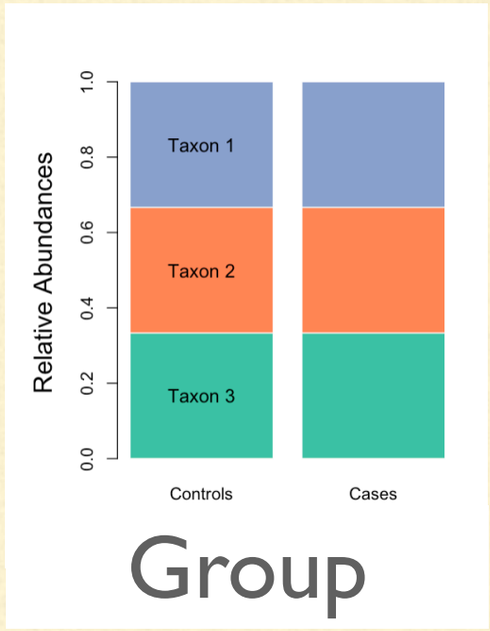


Or this

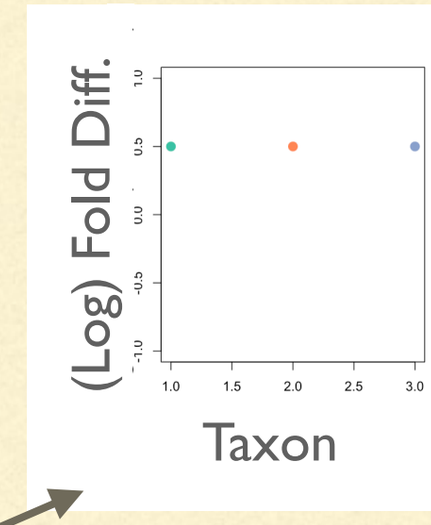


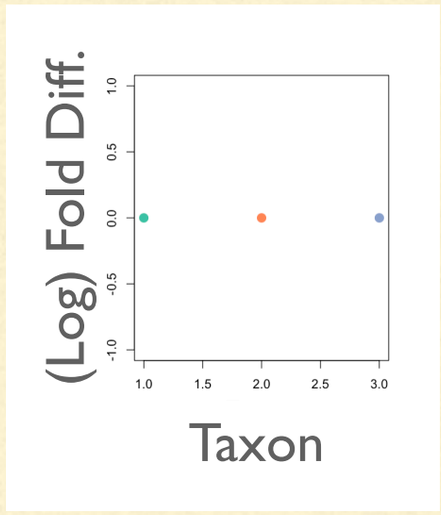
STARTING FROM PROPORTIONS

We observe



Consistent with this



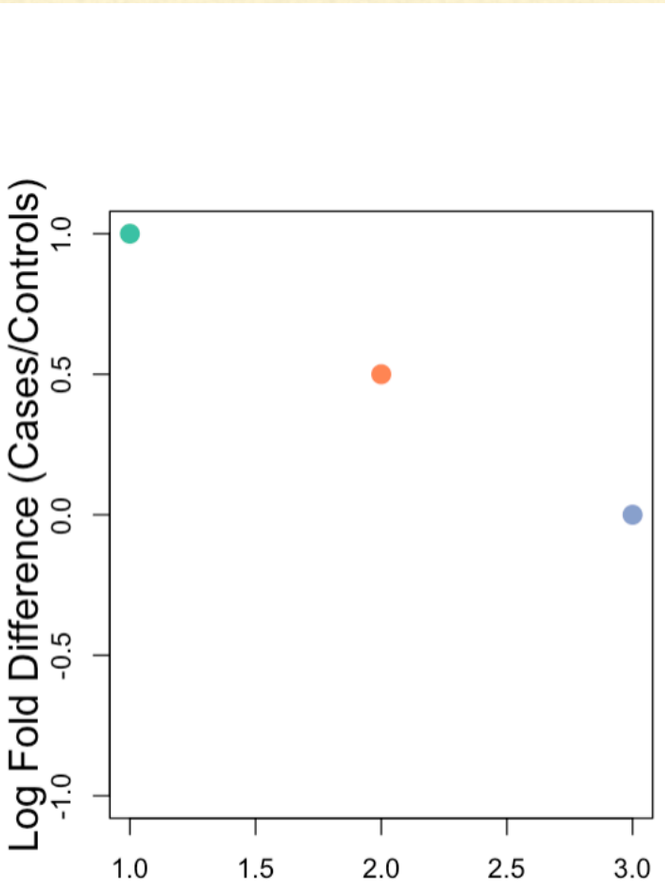


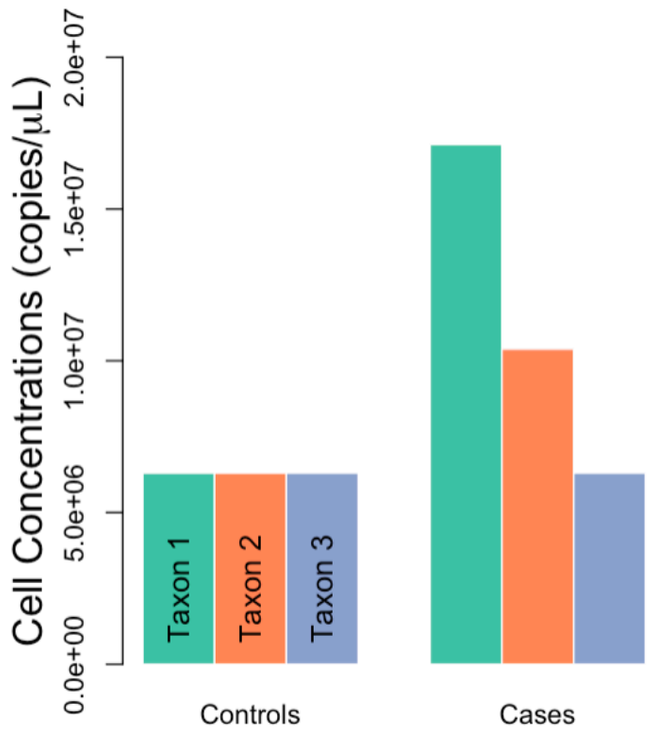
Or this

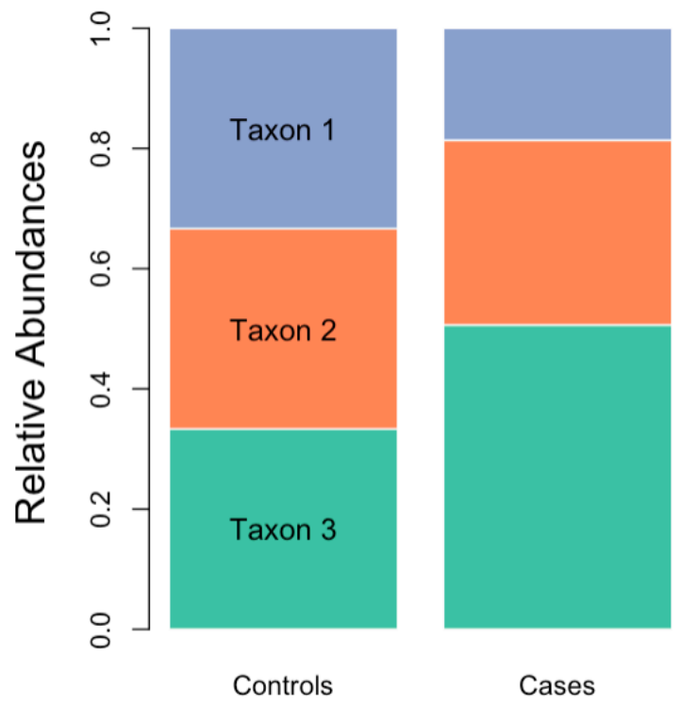
Or…

…

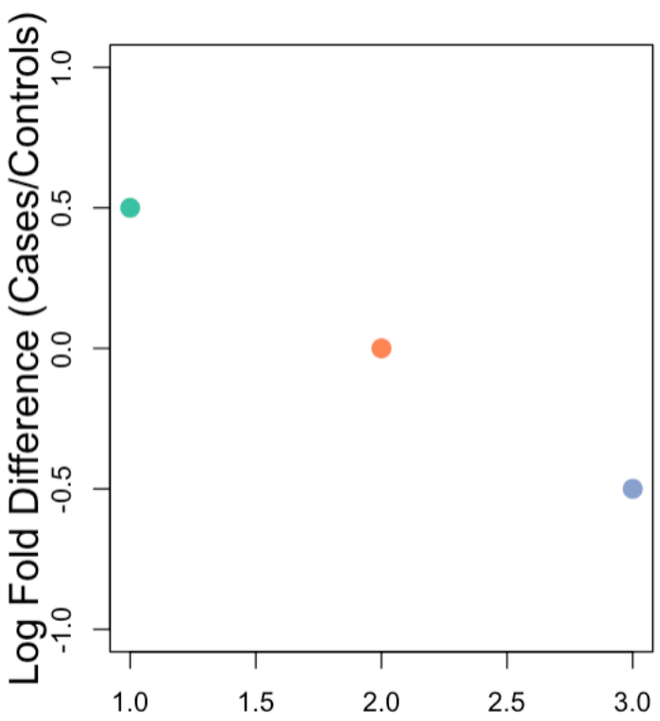
FOLD DIFFERENCES

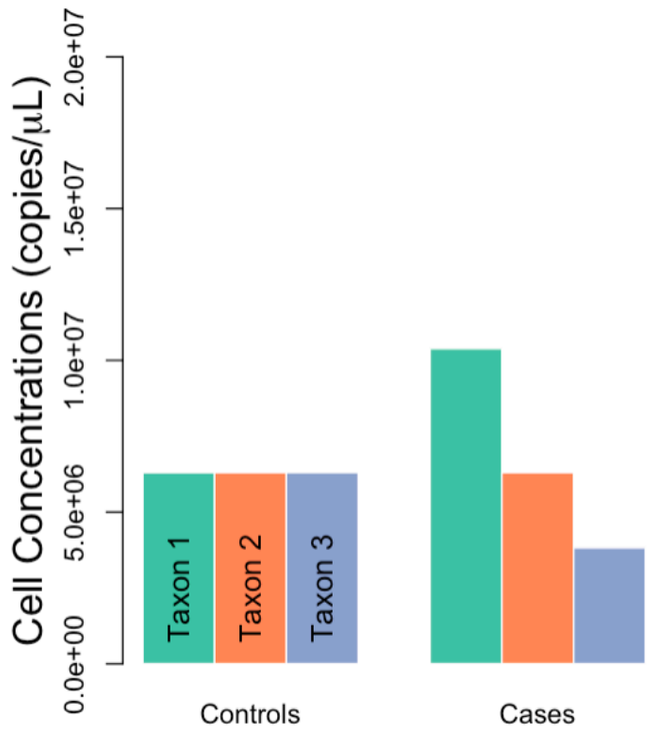


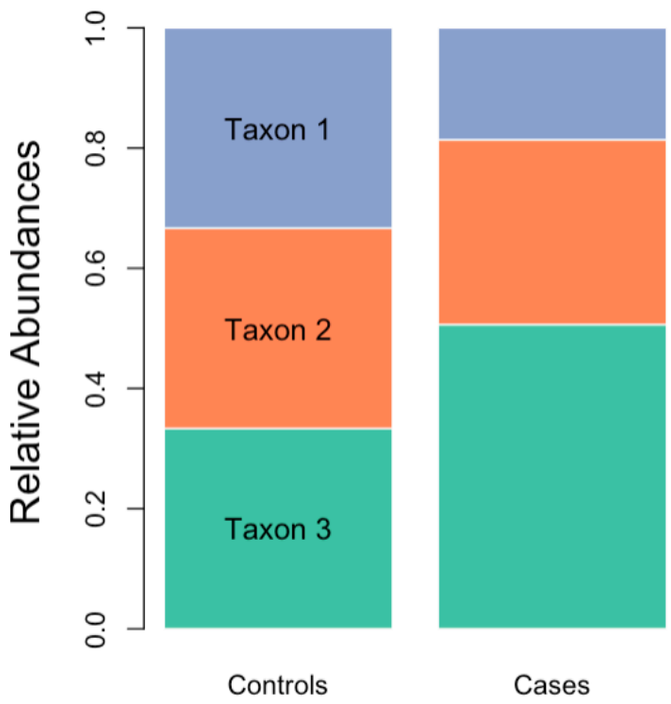




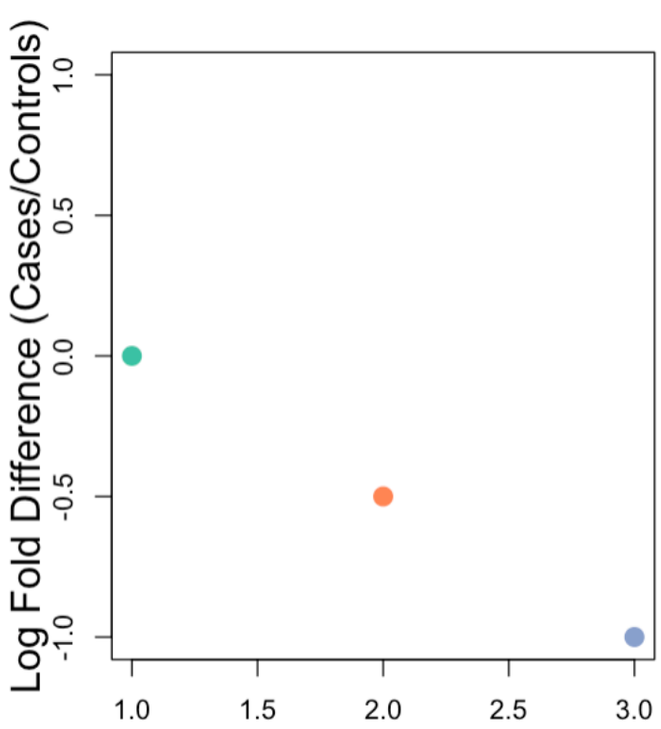
FOLD DIFFERENCES

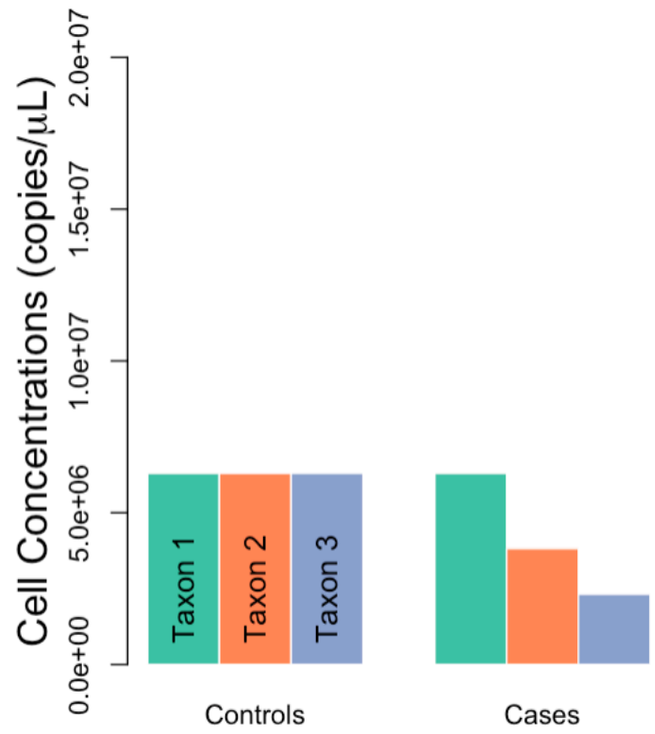


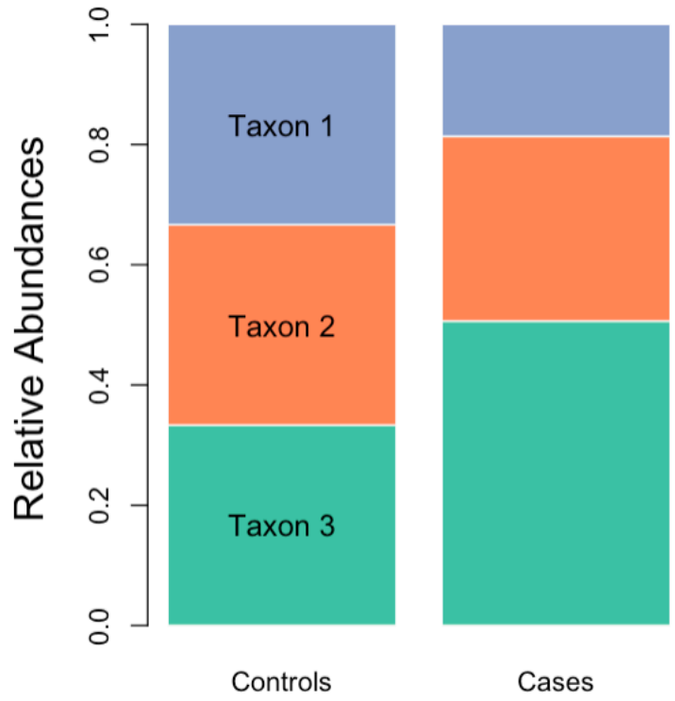




FOLD DIFFERENCES

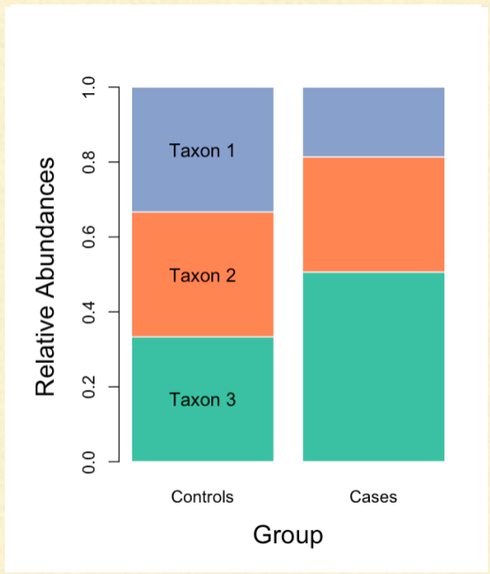


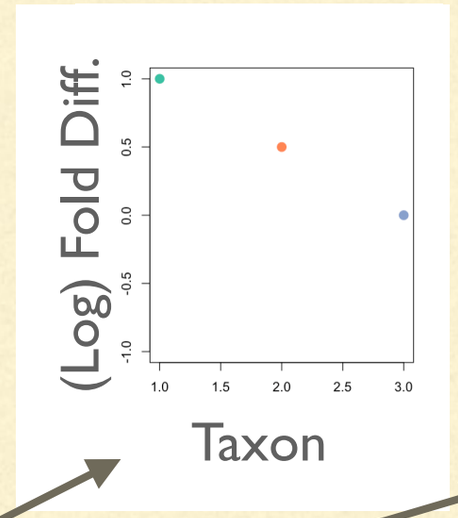




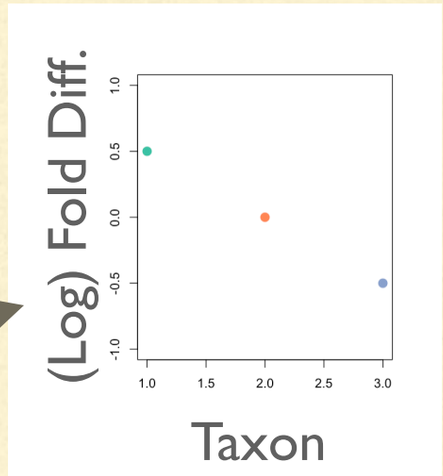
STARTING FROM HTS

We observe

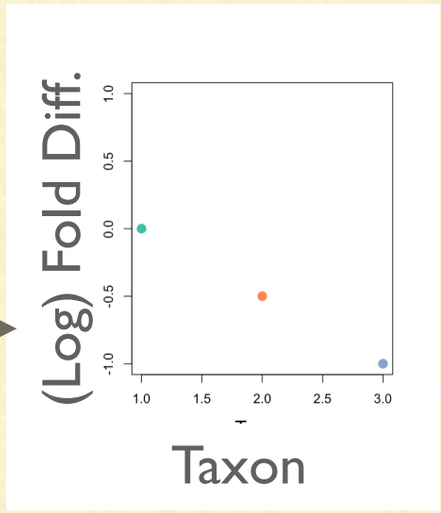




Consistent with this



Or this

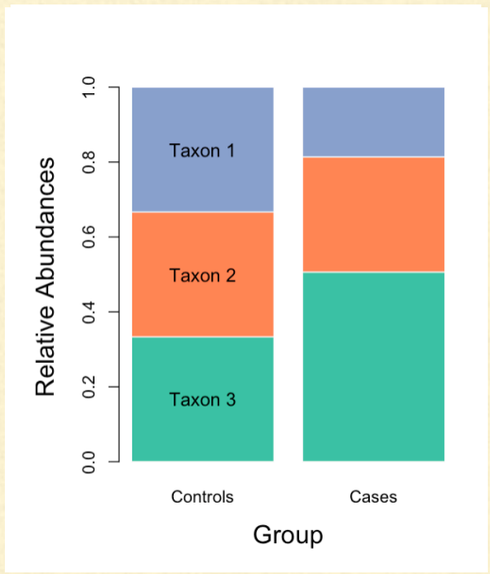


Or…

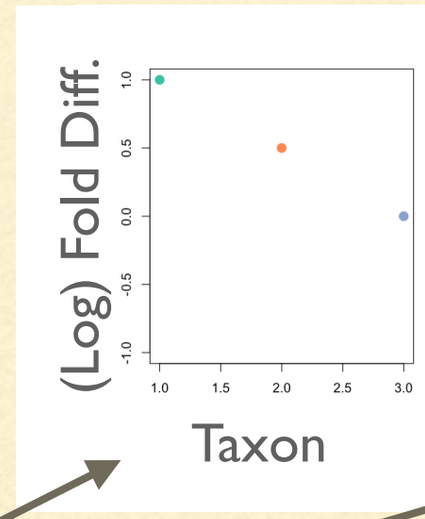
…

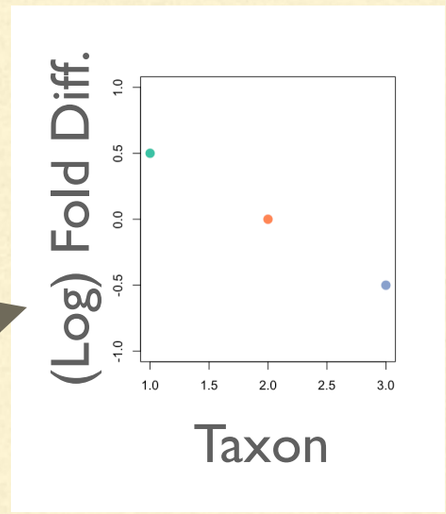
STARTING FROM HTS

We observe

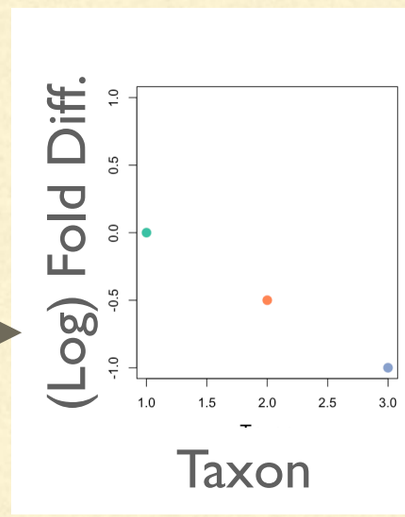


Consistent with this





Or this

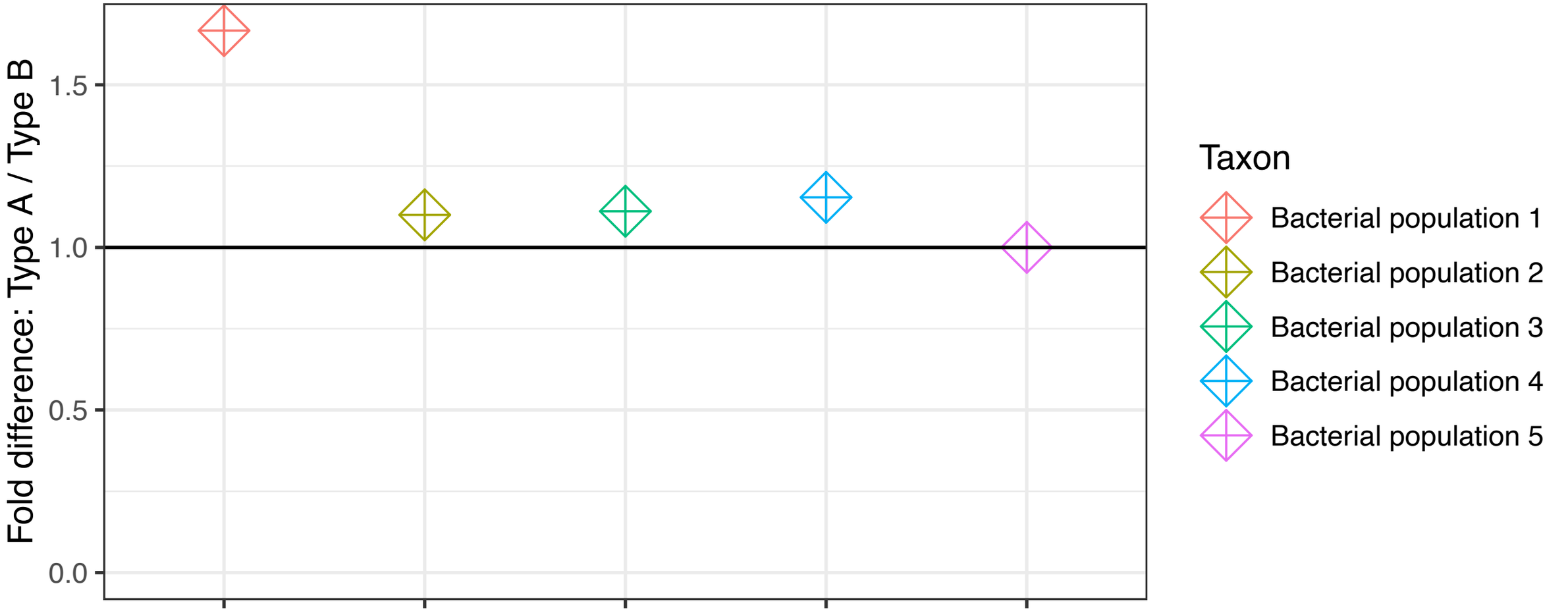


Or this

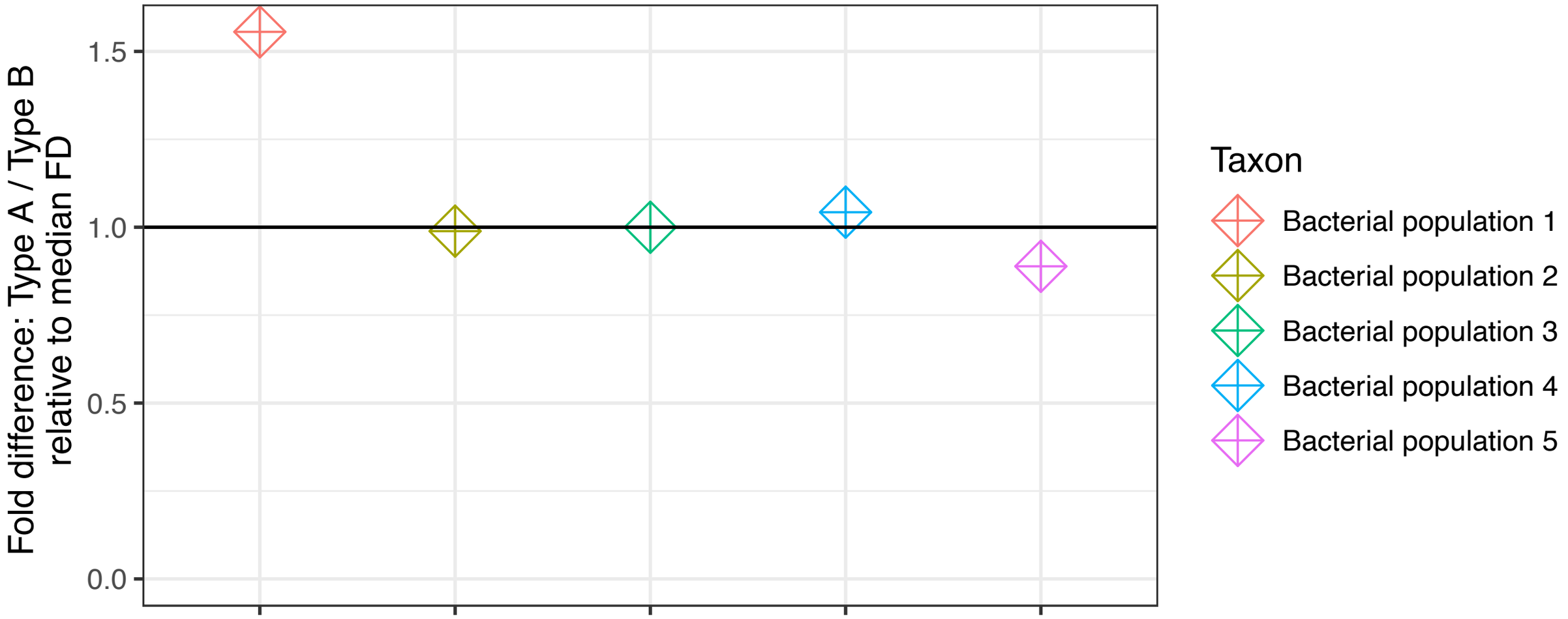
Or…

…

We *cannot* pin the height of this pattern down…
But we *can* estimate which taxa have *larger* fold-differences!



We can identify that the FD for taxon 1 is much larger than for the others…



We estimate

FD in taxon 1 - average FD over *all* taxa

and identify taxa with unusually large fold differences

RADEMU: EXAMPLE

- Wirbel et al. (2019): meta-analysis of 4 case-control studies investigating gut microbiome & colorectal cancer
    - Cases = participants diagnosed with colorectal cancer
    - Controls = participants without a colorectal cancer diagnosis
    - Data: mOTU = "metagenomic OTU" tables
        - Clustering based on a set of highly conserved marker genes
- Goal: characterize which mOTUs unusually over/under-represented in cases

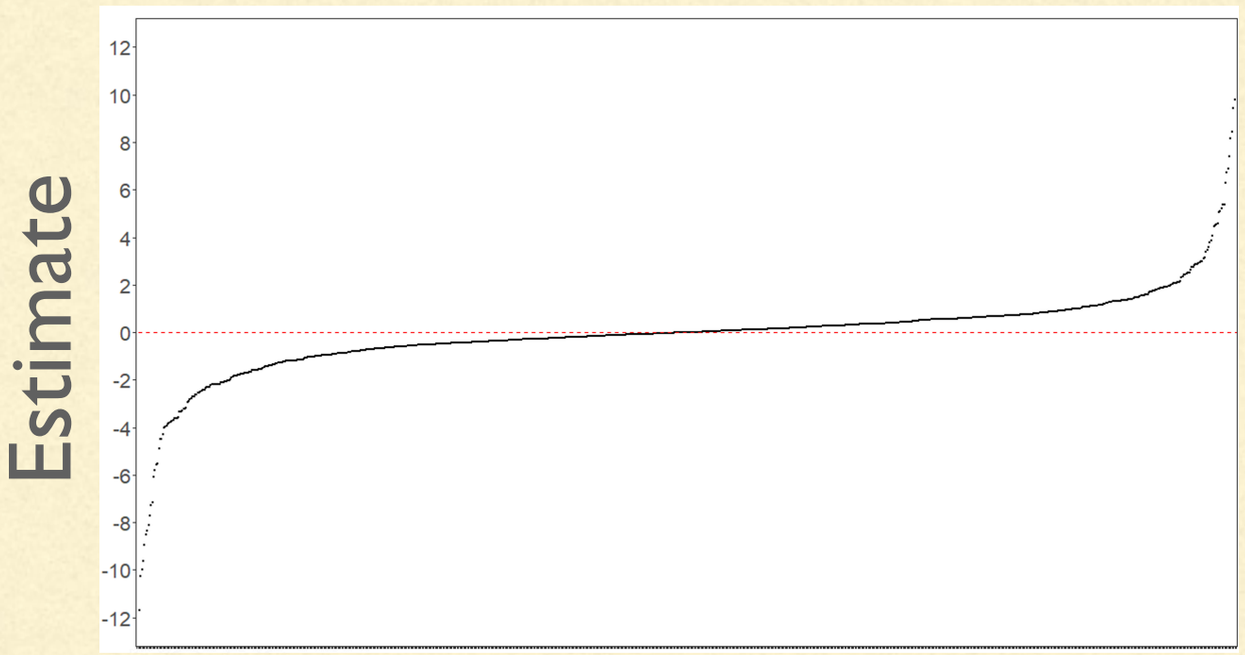
Wirbel, J, et al. *Nature Medicine* (2019)

RADEMU: EXAMPLE

Estimate =

$$\log \frac{\text{mean cell conc. taxon j in cases}}{\text{mean cell conc. taxon j in controls}}$$

but constrained: we force median to be zero



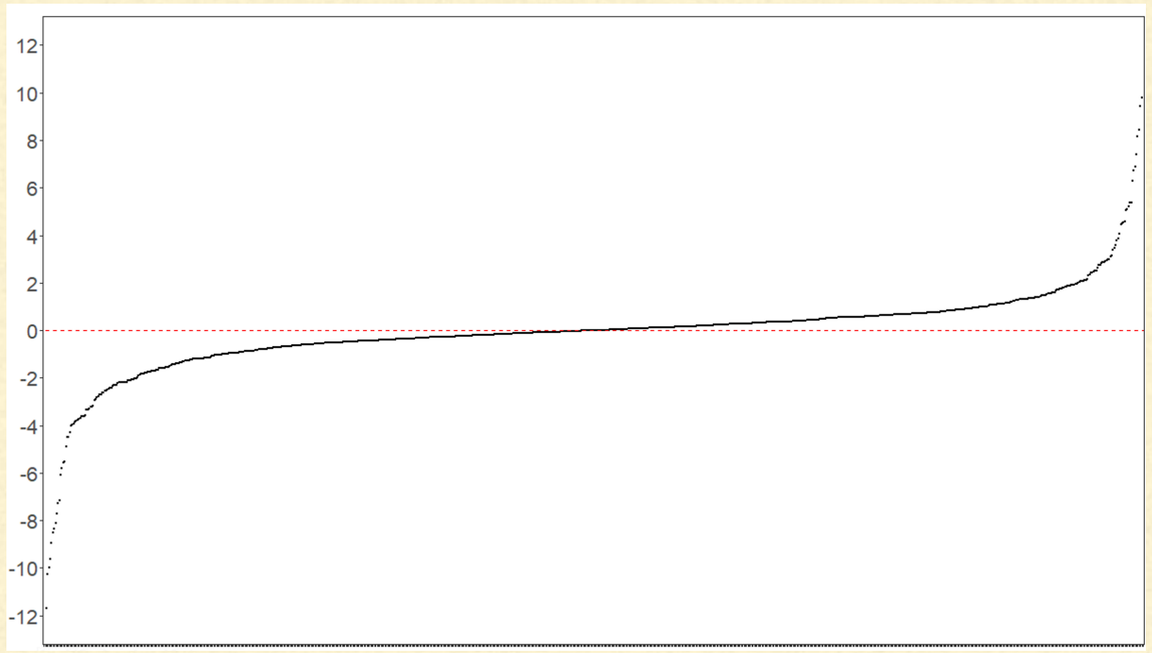
RADEMU: EXAMPLE

Estimate =

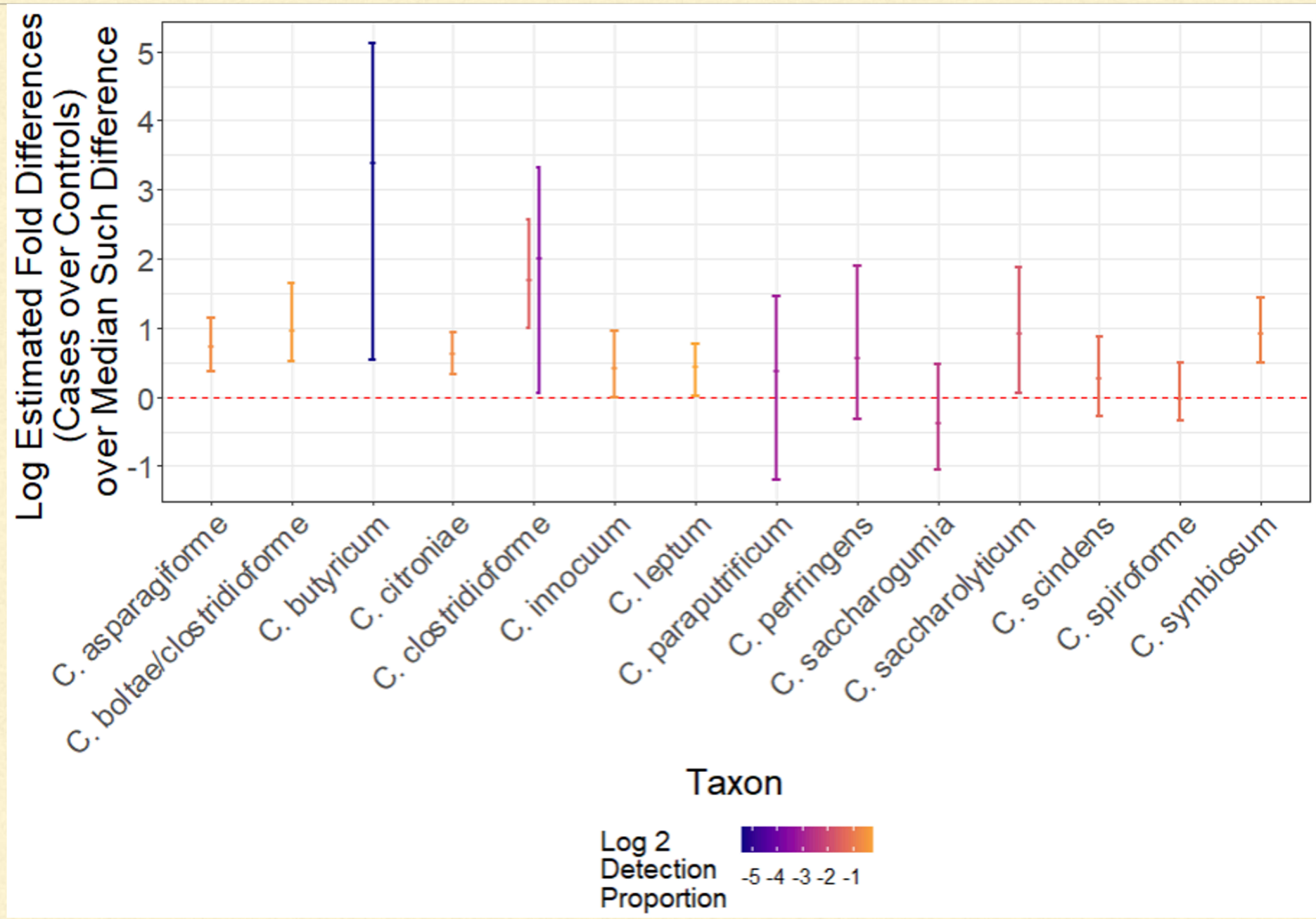
$$\log \frac{\text{mean cell conc. taxon j in cases}}{\text{mean cell conc. taxon j in controls}}$$

but constrained: we force median to be zero

How different are cell concentrations across groups… relative to the *typical* difference?

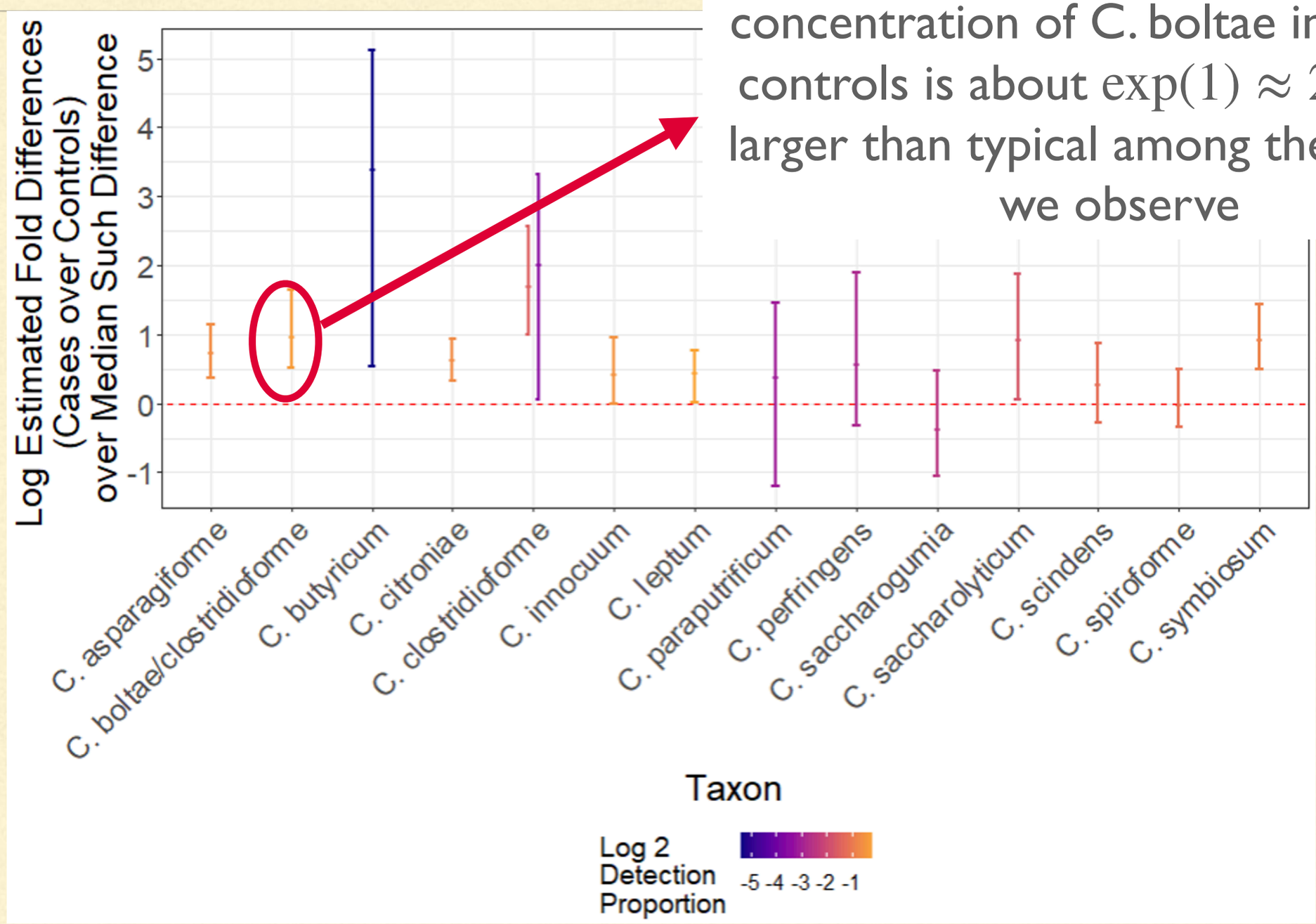


RADEMU: EXAMPLES



RADEMU: EXAMPLES

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 mOTUs we observe



BREAK… OR LAB?

ACCESSING `RADEMU` LAB

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

2. *Copy the command* under radEmu lab

`lm` lab:

```
download.file("https://raw.githubusercontent.com/statdivlab/stamps2023/main/labs/lm-lab/lm-g
```

3. *Run the copied command* in RStudio

```
> download.file("https://raw.githubusercontent.com/statdivlab
  /stamps2023/main/labs/radEmu-lab/rademu_lab.R", "rad-emu-lab
  .R")
```

BREAK

CLOSING THOUGHTS

PEOPLE WORRY ABOUT THE WRONG THINGS

- Examples
  - Is data compositional?
  - Where to rarefy to? We'll talk about this tomorrow
  - Which beta diversity metric? We'll sort of talk about this tomorrow

PEOPLE DON'T WORRY ABOUT THE RIGHT THINGS

- Examples
  - Am I analyzing something I care about?
  - Can I estimate what I care about?

HTS DATA

- Some considerations

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

- Differential abundance methods estimate different parameters; they are *not* directly comparable; *you* need to decide what parameters *you* care about
- Various methods handle 1, 2, and 3 more or less well

IMPLICATIONS

- Now we know about 1, 2, 3 exists, what can we do?
  - Cautious: model relative abundance, but understand limitations
  - Cynical: Only rely on qPCR to discuss abundance
  - Aspirational: Use mock communities for calibration…
  - Aware: model ratios
  - Progressive: radEmu

IMPLICATIONS

- Now we know about 1, 2, 3 exists, what can we do?
  - Cautious: model relative abundance, but understand limitations
  - Cynical: Only rely on qPCR to discuss abundance
  - Aspirational: Use mock communities for calibration…
  - Aware: model ratios
  - Progressive: radEmu

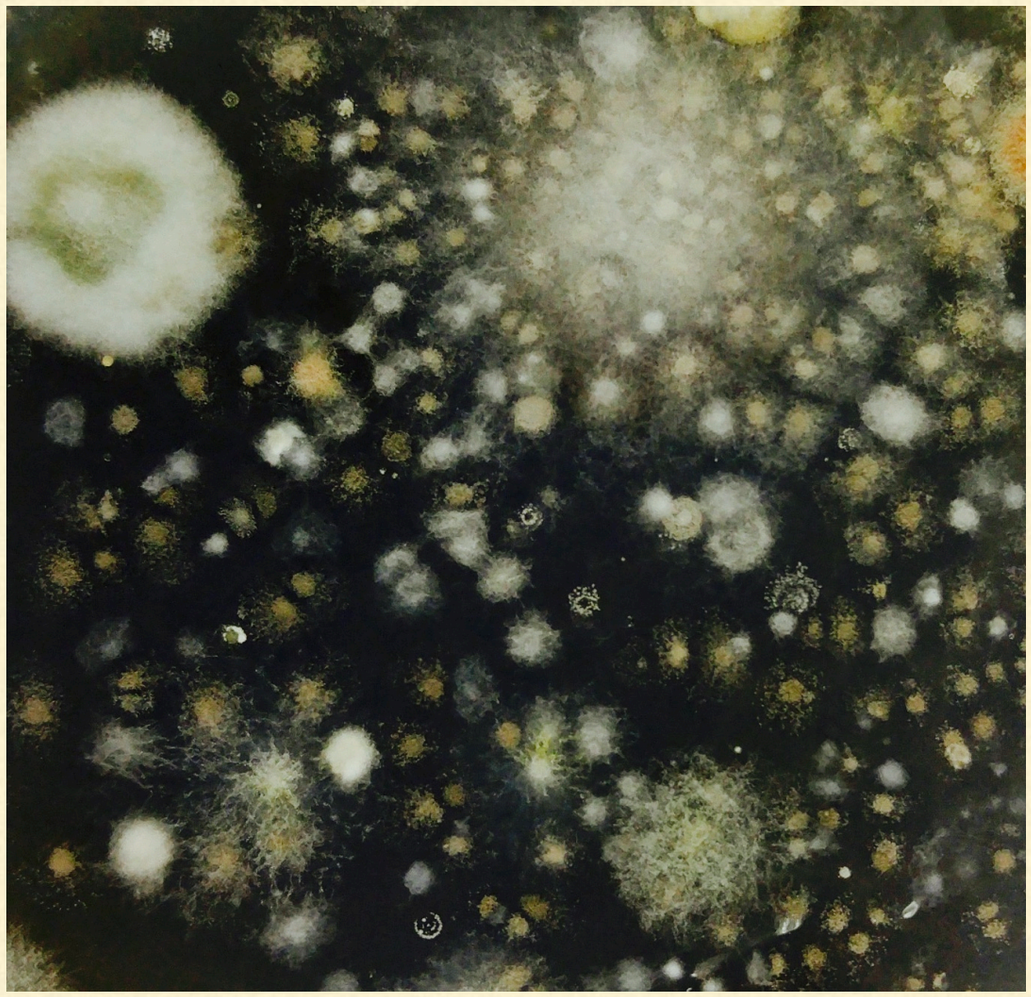
CLOSING THOUGHTS

- Methods for modeling microbiome data is a fast-moving field, and new methods are constantly emerging
- Talk to lots of people
  - "What's the biggest limitation of this?"
- Stay critical but open-minded

WE MAKE TOOLS!

- Estimating and modeling species richness 💰breakaway💰 & 🐠betta🐠
- Estimating and modeling Shannon diversity 🕸️DivNet🕸️
- Estimating and modeling relative abundances 🌽corncob🌽
- **Estimating and modeling presence/absence 🔵happi🟣**
- Estimating and modeling fold differences 🦤radEmu🦤
- Estimating detection efficiencies of HTS relative to qPCR data ⛑️paramedic⛑️
- **Decontaminating relative abundance & estimating differential detection w/ mock communities 🦹 tinyvamp🦹**
- General purpose regression models with robust hypothesis testing 📈rigr📈
- **Investigating gene-phylogenies alongside your phylogenomic tree 🌴groves🎄**

- Modeling where a mammoth that lived 17,000 years ago went 🦣KikWalk🦣



MODELING MICROBIOME DATA

Statistical Diversity Lab @ University of Washington

Amy Willis — @AmyDWillis — Associate Professor

Sarah Teichman — @sarah_teichman — PhD Candidate

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