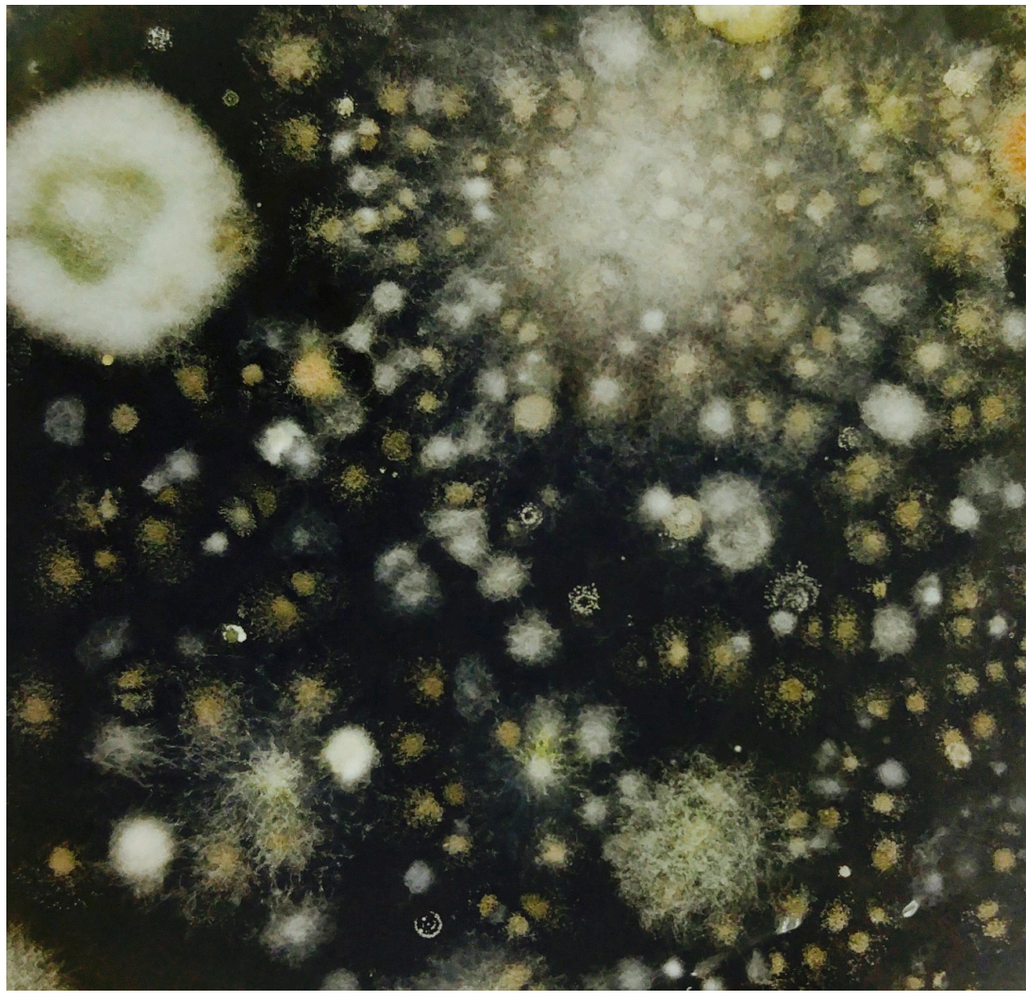


# Modeling microbial abundances

**Statistical Diversity Lab @ University of Washington**
Amy Willis — @AmyDWillis — Associate Professor
Sarah V Teichman — Research Scientist
María Valdez — Postdoctoral Scholar
and
Sarah J Tucker — Postdoctoral Scholar (MBL)

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
- Your data 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 (but with ideas on how to answer them)

# The plan

- Hypothesis testing
- Analyzing microbiome data
  - Abundance
  - Trees
  - Expression + abundance
  - Diversity
- Questions — throughout!

Now!

ask us about p values!

ask us about compositionality!

ask us about differential abundance!

Tomorrow!

ask us about phylogenetics

ask us about metatranscriptomics

ask us about rarefaction!

ask us about diversity metrics!

ask us about ordination!

# Inference

By Sarah

# Inference

- We have identified a parameter 🧈🫐
- We have found an estimator 👨‍🍳⏲️
- We have computed an estimate 🥧

**What can we *really* say about the parameter?**

# Confidence interval

- Estimate: most likely value of the parameter based on the data

- We still have uncertainty

- A confidence interval gives us a range of likely parameter values, taking into account that uncertainty

# Confidence interval example

- Question: what is the expected log fold difference in the abundance of *Cyanobacteria* between the Pacific and Atlantic oceans?
- Answer: we estimate it to be 3 with 95% confidence interval (1.5, 4.5)

# Confidence interval example

- Question: what is the expected log fold difference in the abundance of *Cyanobacteria* between the Pacific and Atlantic oceans?

- Answer: we estimate it to be 3 with 95% confidence interval (1.5, 4.5)

- Meaning: if we repeated this study, collecting data and constructing 95% intervals each time, 95% of those intervals would contain the true value of the parameter

# Why are confidence intervals useful?

- What would you conclude from $\hat{\beta} = 3$, where $\hat{\beta}$ is our estimated log fold difference in the abundance of *Cyanobacteria* between the Pacific and Atlantic oceans?
    - Estimated fold difference of $\exp(3) \approx 20$
- what would you conclude from $\hat{\beta} = 0.5$?
    - Estimated fold difference of $\exp(0.5) \approx 1.6$

# Why are confidence intervals useful?

**Which result do you think makes the strongest scientific claim? The weakest?**

- What about:
  - fold difference estimate: 20, CI = (12, 33)
  - fold difference estimate: 20, CI = (0.05, 8100)
  - fold difference estimate: 1.6, CI = (1, 2.7)
  - fold difference estimate: 1.6, CI = (1.3, 2.1)

# Hypothesis testing

- Question: do we have evidence to support the conclusion that the true abundance of *Cyanobacteria* is different between the Pacific and Atlantic oceans?

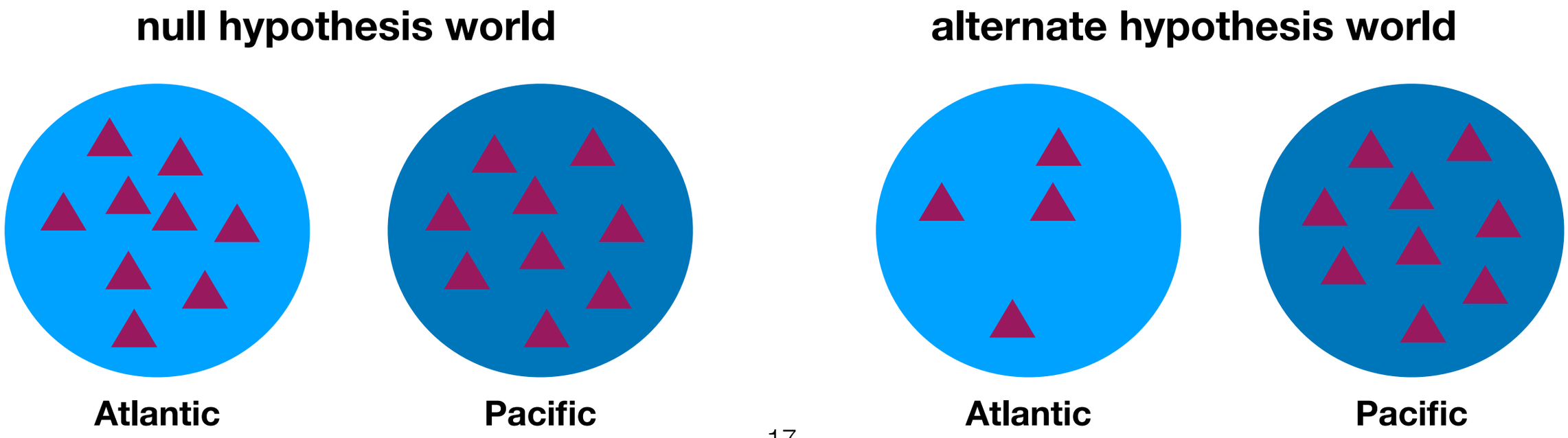
# Hypothesis testing

- Question: do we have evidence to support the conclusion that the true abundance of *Cyanobacteria* is different between the Pacific and Atlantic oceans?

- Hypothesis testing framework:
  - null hypothesis: the true abundance of *Cyanobacteria* is the same between these two oceans
  - alternate hypothesis: the true abundance of *Cyanobacteria* is different between these two oceans
  - central question: do we have evidence that conflicts with our null hypothesis?

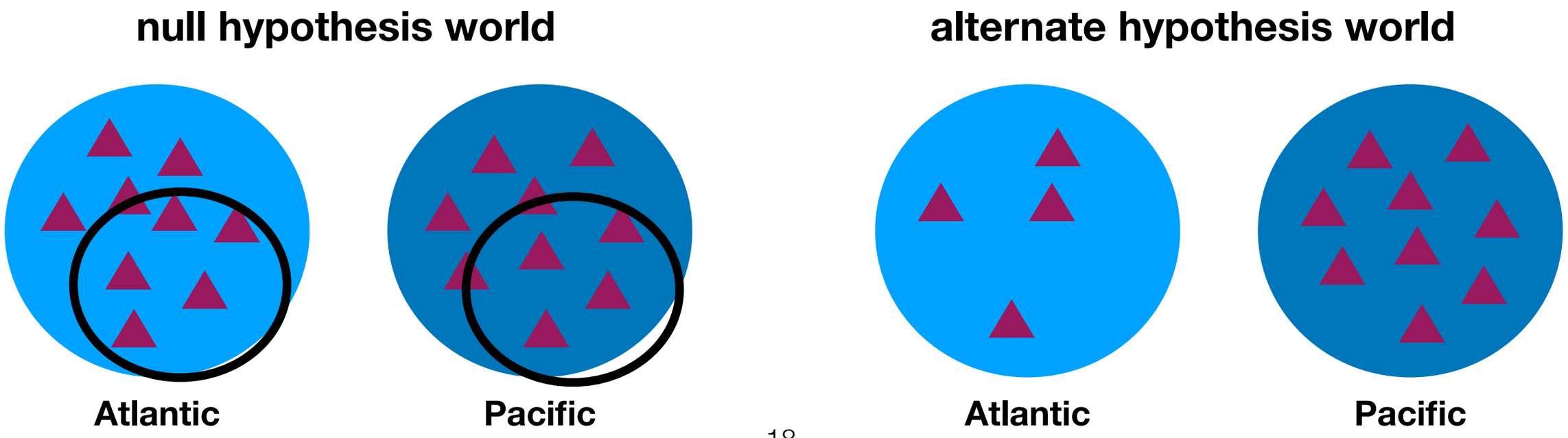
# The logic of hypothesis testing

- Assuming that we're in null hypothesis world, how likely is it that we would have collected this data?
  - If likely, then we do not have evidence against the null hypothesis



# The logic of hypothesis testing

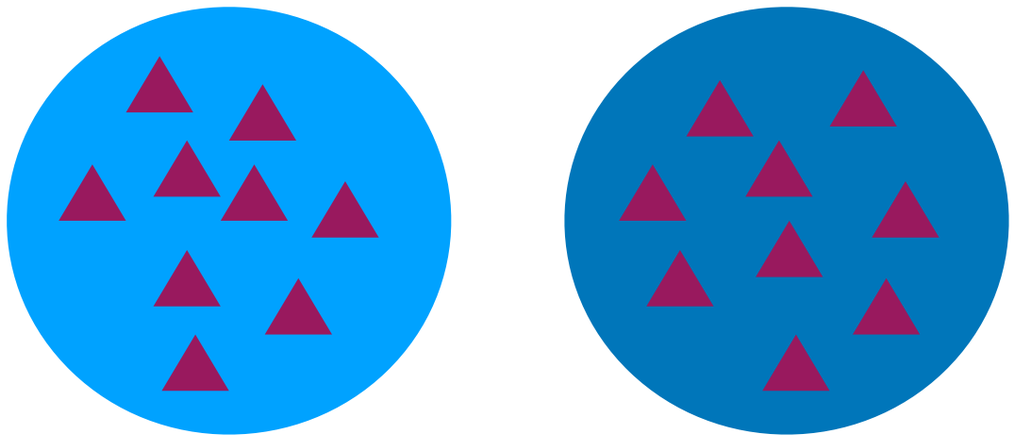
- Assuming that we're in null hypothesis world, how likely is it that we would have collected this data?
  - If likely, then we do not have evidence against the null hypothesis
  - maybe the null hypothesis holds



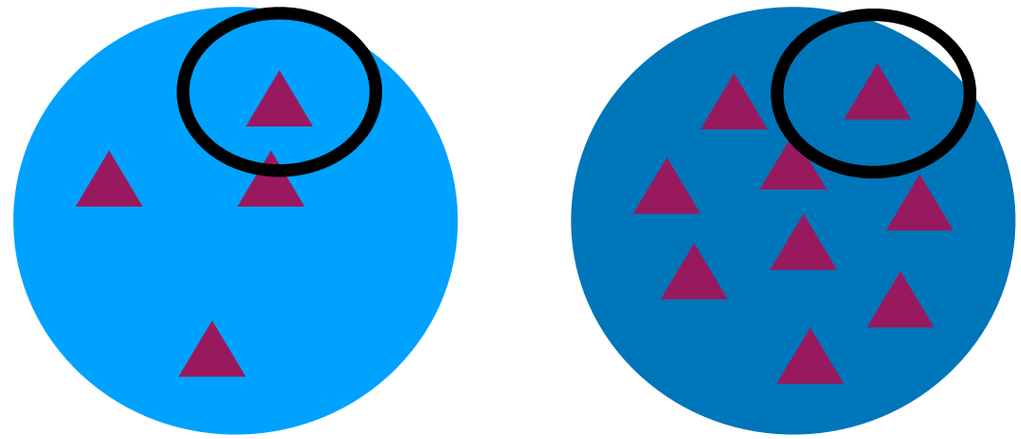
# The logic of hypothesis testing

- Assuming that we're in null hypothesis world, how likely is it that we would have collected this data?
  - If likely, then we do not have evidence against the null hypothesis
  - maybe we just didn't collect enough data

**null hypothesis world**




**alternate hypothesis world**

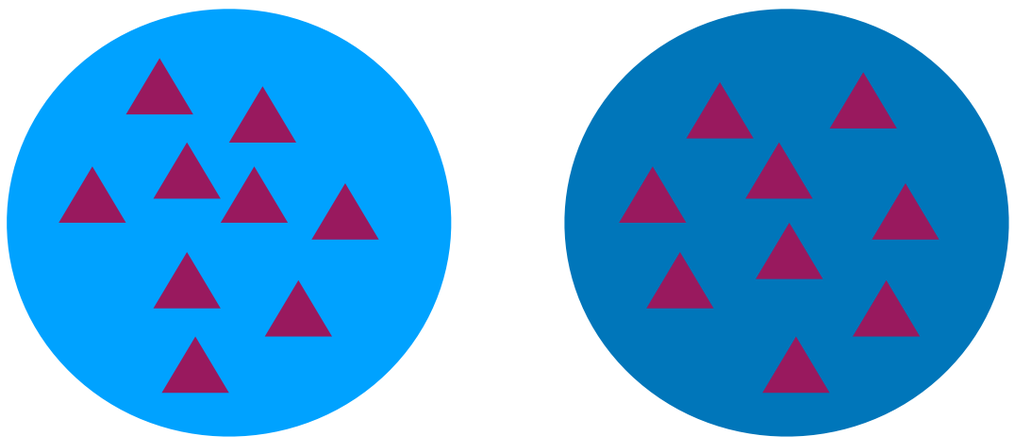


# The logic of hypothesis testing

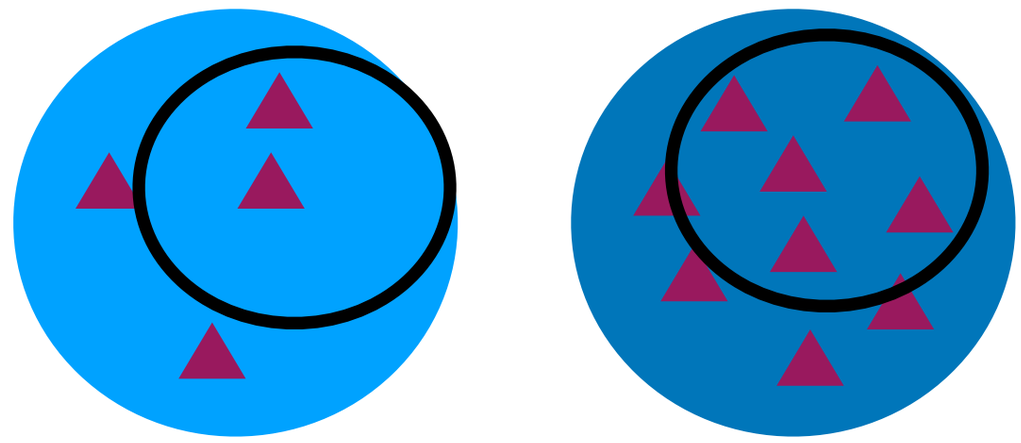
- Assuming that we're in null hypothesis world, how likely is it that we would have collected this data?

  - If unlikely, suggests that we might instead be in alternate hypothesis world

**null hypothesis world**




**alternate hypothesis world**



# Hypothesis testing example

- Question: Is giving a patient a new treatment for IBD associated with increased microbial diversity in the gut, compared to patients who receive a placebo?

- parameter you might care about:

- null hypothesis:

- alternate hypothesis:

# The life of a microbial ecologist

- Decide what statistical parameter to test, translate this into null hypothesis and alternate hypothesis

# p-value?

- Your friendly local statistician has come up with an exciting new procedure to test any hypothesis you want:

```
my_test <- function(data) {
      return(p = 0.001)
}
```

Is this a p-value? Why or why not?

# The life of a statistician

- Do the testing
  - calculate a test statistic based on data
  - determine how likely it would be to observe this test statistic in null hypothesis world
- p-value = probability that we would observe data as or more extreme as what we did, if we were in null hypothesis world
- we won’t focus on test statistic construction and p-value calculation, but you can ask us questions about it later!

# The life of a statistician

- A test statistic example: Wald test statistic
- $$t = \frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}}$$
    - What does it mean if this is big? If it is small?
- this is just one type of test statistic!

# Hypothesis testing interpretation

- Case 1:
  - $\hat{\beta} = 3$: We estimate the log fold difference to be 3.
  - p = 0.01: The probability that we would collect a sample leading to an estimate that is 3 or larger, given that the true log fold difference in abundance is 0, is 1%.

# Hypothesis testing interpretation

- Case 1:
  - $\hat{\beta} = 3$: We estimate the log fold difference to be 3.
  - p = 0.01: The probability that we would collect a sample leading to an estimate that is 3 or larger, given that the true log fold difference in abundance is 0, is 1%.
- Case 2:
  - $\hat{\beta} = 0.5$: We estimate the log fold difference to be 0.5.
  - p = 0.23: The probability that we would collect a sample leading to an estimate that is 0.5 or larger, given that the true log fold difference in abundance is 0, is 23%.

# Hypothesis testing interpretation

- Case 1:
  - $\hat{\beta} = 3$: We estimate the log fold difference to be 3.
  - p = 0.01: The probability that we would collect a sample leading to an estimate that is 3 or larger, given that the true log fold difference in abundance is 0, is 1%.
- Case 2:
  - $\hat{\beta} = 0.5$: We estimate the log fold difference to be 0.5.
  - p = 0.23: The probability that we would collect a sample leading to an estimate that is 0.5 or larger, given that the true log fold difference in abundance is 0, is 23%.

Can you reject the null hypothesis? Choose $\alpha$ level before running the test!

# Hypothesis testing interpretation

- Let's say we set $\alpha = 0.05$

- p = 0.01: We reject the null hypothesis that the true log fold difference in abundance of *Cyanobacteria* between these oceans is the same.

- p = 0.23: We fail to reject the null hypothesis that the true log fold difference in abundance of *Cyanobacteria* between these oceans is the same.

# What makes a good hypothesis test?

# What makes a good hypothesis test?

- reasonable assumptions or few assumptions
- valid
- high power

# Assumptions

- Assumptions required for good estimation are also required for good tests!

We like: 😺

- assumptions that are biologically plausible to you!
- methods that work well with small sample sizes

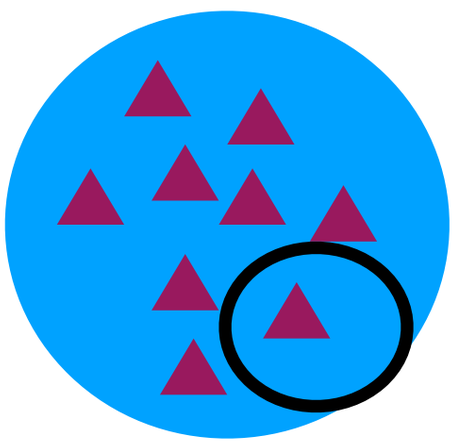
We like less: 😿

- (parametric) assumptions about the distribution of your data
- methods that assume your data are non-zero/integer-valued/etc.

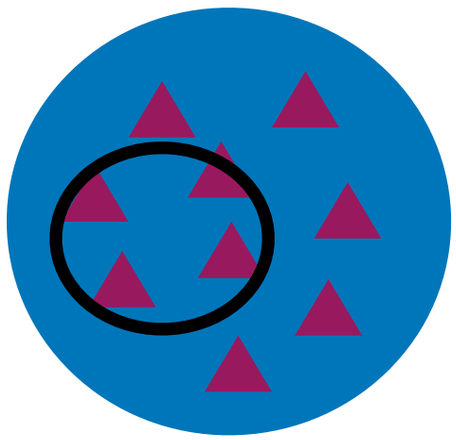
# Valid test

- type I error: rejecting null hypothesis when it is actually true

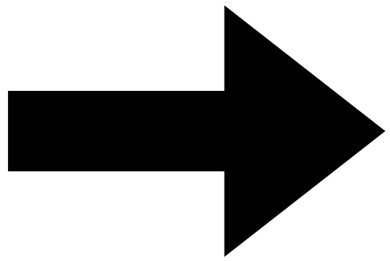
**null hypothesis world**



**Atlantic**



**Pacific**



**p = 0.03, reject the null hypothesis (incorrectly)**

# Valid test

- type I error: rejecting null hypothesis when it is actually true
- valid test: a testing procedure that will make type I errors $\alpha\,\%$ of the time when the null hypothesis is true, using a p-value threshold of $\alpha$

# Valid test

- type I error: rejecting null hypothesis when it is actually true

- valid test: a testing procedure that will make type I errors $\alpha\,\%$ of the time when the null hypothesis is true, using a p-value threshold of $\alpha$

- In hypothesis testing methods paper, authors should demonstrate that their controls type I error rate in simulation settings **that you care about**

# Power of test

- power: when the null hypothesis is false, how often do you have enough evidence to reject it?

# Power of test

- power: when the null hypothesis is false, how often do you have enough evidence to reject it?

- our rule of thumb: out of methods that have demonstrated type I error rate control in settings you care about, choose one with high power

# Which test do you like?

low variance data, n = 50:

| | T1 | T2 |
|---|---|---|
| % t1e* | 4% | 5% |
| power** | 60% | 65% |

* using 0.05 threshold

** to detect a moderate signal (ex: fold difference of 3)

# Which test do you like?

high variance, sparse data, n = 50:

| | T1 | T2 |
|---|---|---|
| % t1e* | 4% | 9% |
| power** | 30% | 60% |

* using 0.05 threshold

** to detect a moderate signal
(ex: fold difference of 3)

# Which test do you like?

low variance data, n = 50:

| | T1 | T2 |
|---|---|---|
| % t1e* | 4% | 5% |
| power** | 60% | 65% |

high variance, sparse data, n = 50:

| | T1 | T2 |
|---|---|---|
| % t1e* | 4% | 9% |
| power** | 30% | 60% |

* using 0.05 threshold

** to detect a moderate signal
(ex: fold difference of 3)

# Which test do you like?

low variance data, n = 50:

| | T1 | T2 |
|---|---|---|
| % t1e* | 4% | 5% |
| power** | 60% | 65% |

high variance, sparse data, n = 50:

| | T1 | T2 |
|---|---|---|
| % t1e* | 4% | 9% |
| power** | 30% | 60% |

* using 0.05 threshold

** to detect a moderate signal (ex: fold difference of 3)

**What if T1 only accepts positive integers?**

# Multiple Testing

- Your colleague tells you “I ran the test for our analysis for taxon 1 and got a p-value of 0.003!”

- how do you feel?

# Multiple Testing

- Your colleague tells you “I ran the test for our analysis for taxon 1 and got a p-value of 0.003!”

- you know that your analysis involved testing 1000 taxa. Does this change how you feel?

# Multiple Testing

- the hypothesis testing framework was developed when people were often running a single test at once

- it is common now to run many tests at a time

- for multiple tests, we need to adjust our criteria for statistical significance

# Multiple Testing

- 2 independent tests, null hypothesis is true for both:
    - Probability you don't reject $H_0$ for Test 1 $= .95$
    - Probability you don't reject $H_0$ for Test 2 $= .95$
    - Probability you don't reject $H_0$ for both tests
      $= .95 \times .95 = .9025$
- Probability you make at least one type 1 error: ~10%

# Multiple Testing

- 3 independent tests, null hypothesis true for all of them:
  - Probability you don't reject $H_0$ for all tests
    $= .95 \times .95 \times .95 = .8574$
- Probability you make at least one type 1 error: ~14%

# Multiple Testing

| $m$ | Probability of at least one Type I error |
|---|---|
| 10 | 0.40 |
| 100 | 0.994 |
| 1,000 | ~1 |

# Multiple Testing

- So, what can we do when we need multiple tests?

# Multiple Testing

- So, what can we do when we need multiple tests?

- Instead of controlling Type I error rate separately for each test, consider:

  - **Family-wise Error Rate (FWER):** probability of at least one type 1 error

  - **False Discovery Rate (FDR):** the expected proportion of type 1 errors among the rejected hypotheses

# Multiple Testing

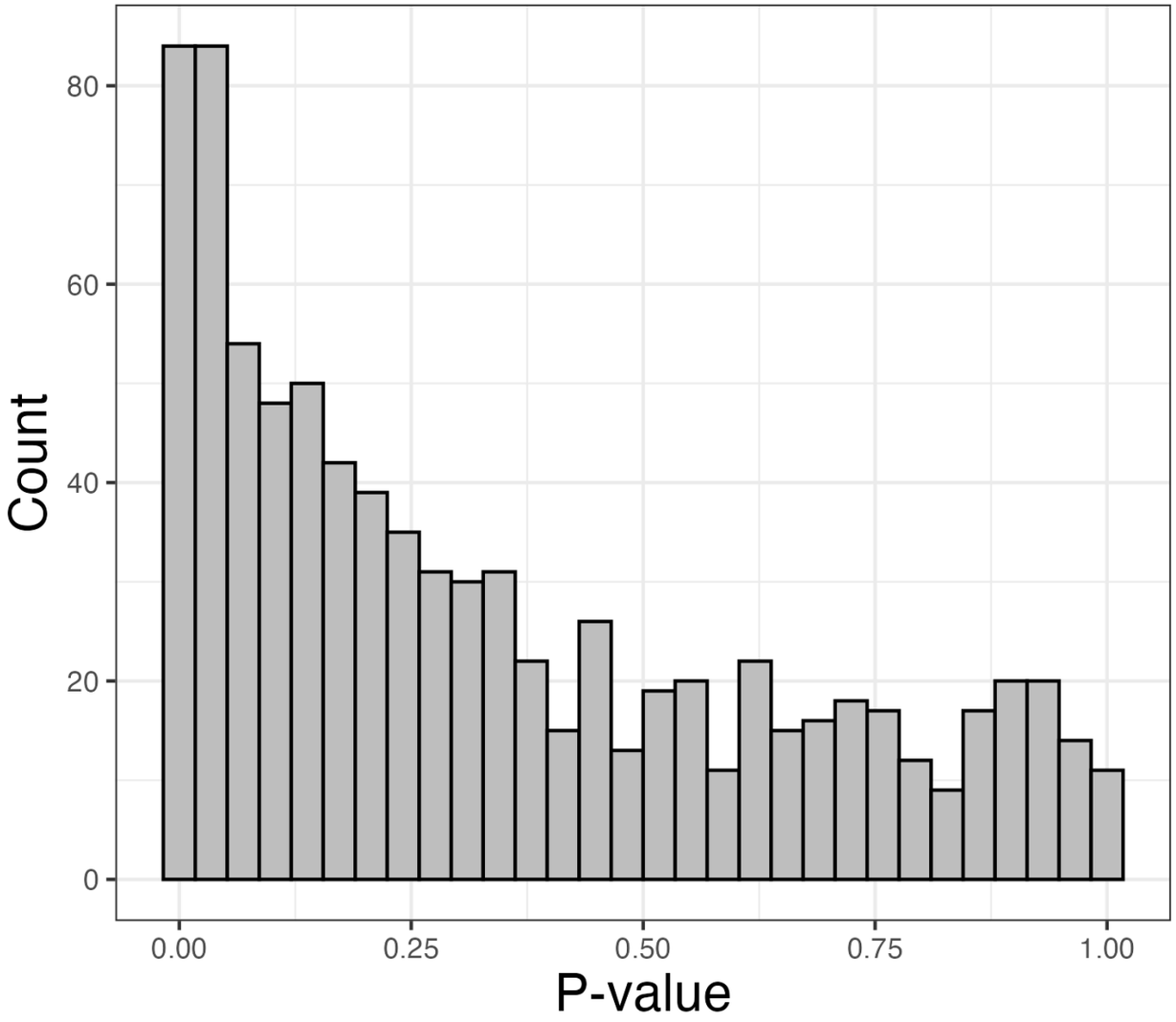
- Instead of controlling Type I error rate separately for each test, consider:
    - **Family-wise Error Rate (FWER):** probability of at least one type 1 error
        - Use Bonferroni correction, divide $\alpha$ by number of tests, use this as threshold for rejecting null hypothesis
    - **False Discovery Rate (FDR):** the expected proportion of type 1 errors among the rejected hypotheses
        - Can use q-values instead of p-values

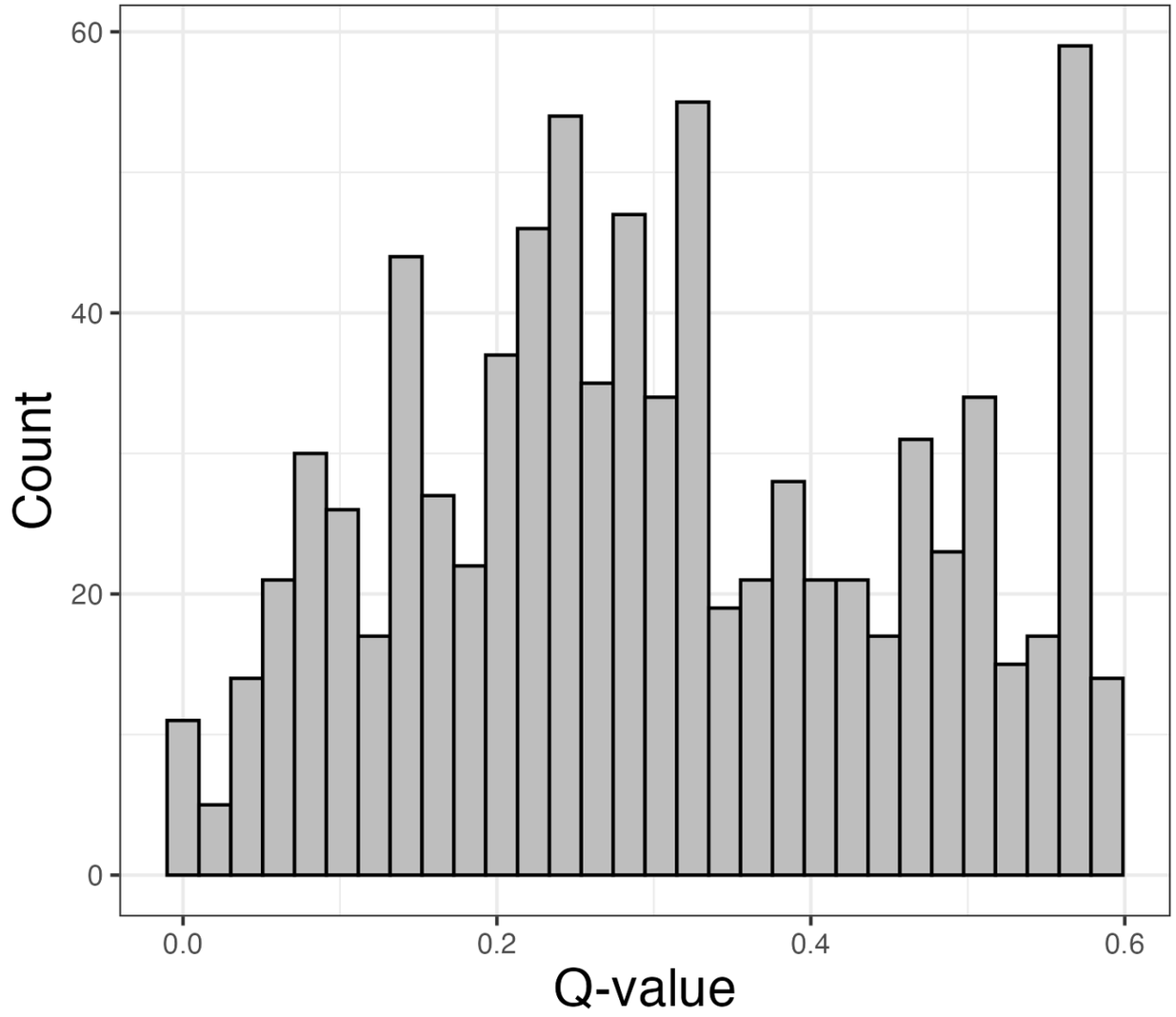
# Q-values

- q-values
    - Adjusted p-values to control FDR instead of Type I error rate
- In analysis of 1000 taxa, your colleague found one taxon with a p-value of 0.00005 and a q-value of 0.03
    - p-value: the probability they would see a test statistic as extreme as the one observed for a non-differentially abundant taxon is 0.00005
    - q-value: 3% of the taxa that were tested and had test statistics even more extreme than the one observed would be false positives

# Q-values

Distribution of p-values and q-values from analysis of data from Wirbel et al.




165/845 p-values <= 0.05

30/845 q-values <= 0.05

Wirbel et al. "Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer." *Nature medicine,* 25(4), 2019.

# Multiple Testing

- There are a number of other methods to avoid issues with multiple testing

- Your best bet is limiting formal testing to primary hypotheses when possible (for highest power!)

- It’s fine to screen for many signals, but adjust accordingly

# Interpreting hypothesis tests in context

- often we may use a series of hypothesis tests to identify signals that are the most statistically significant

- you may want to report the signals with smallest p/q-values

- ***don't forget context***, also report estimates and decide whether a statistically significant signal is also scientifically interesting

# Interpreting hypothesis tests in context

- often we may use a series of hypothesis tests to identify signals that are the most statistically significant

- you may want to report the signals with smallest p/q-values

- ***don't forget context***, also report estimates and decide whether a statistically significant signal is also scientifically interesting

- What is most interesting log fold-difference result?

| **Taxon A:** | **Taxon B:** | **Taxon C:** |
|---|---|---|
| $\hat{\beta} = 3,\ \mathrm{q} = 0.02$ | $\hat{\beta} = 9,\ \mathrm{q} = 0.33$ | $\hat{\beta} = 0.5,\ \mathrm{q} = 0.001$ |

# The life of a microbial ecologist

- Decide how statistical results (in terms of estimates, confidence intervals, and hypothesis tests) provide evidence to answer your scientific question(s)

# Modeling microbial abundances

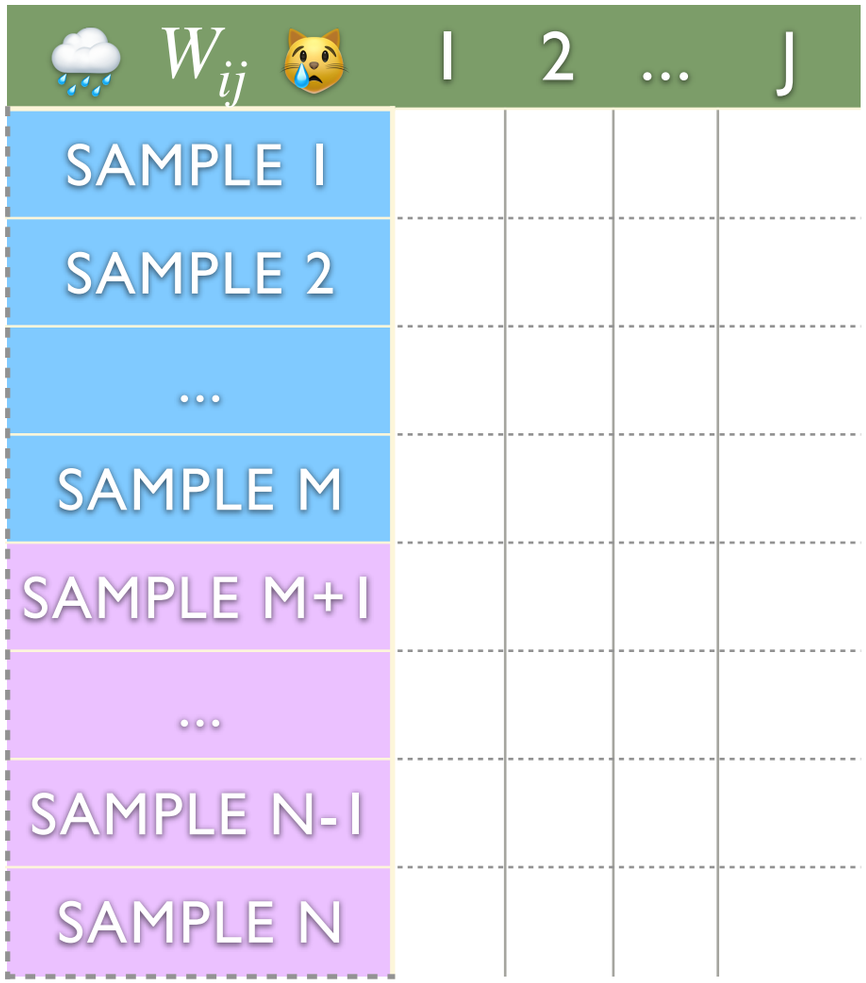
# Data

- 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…

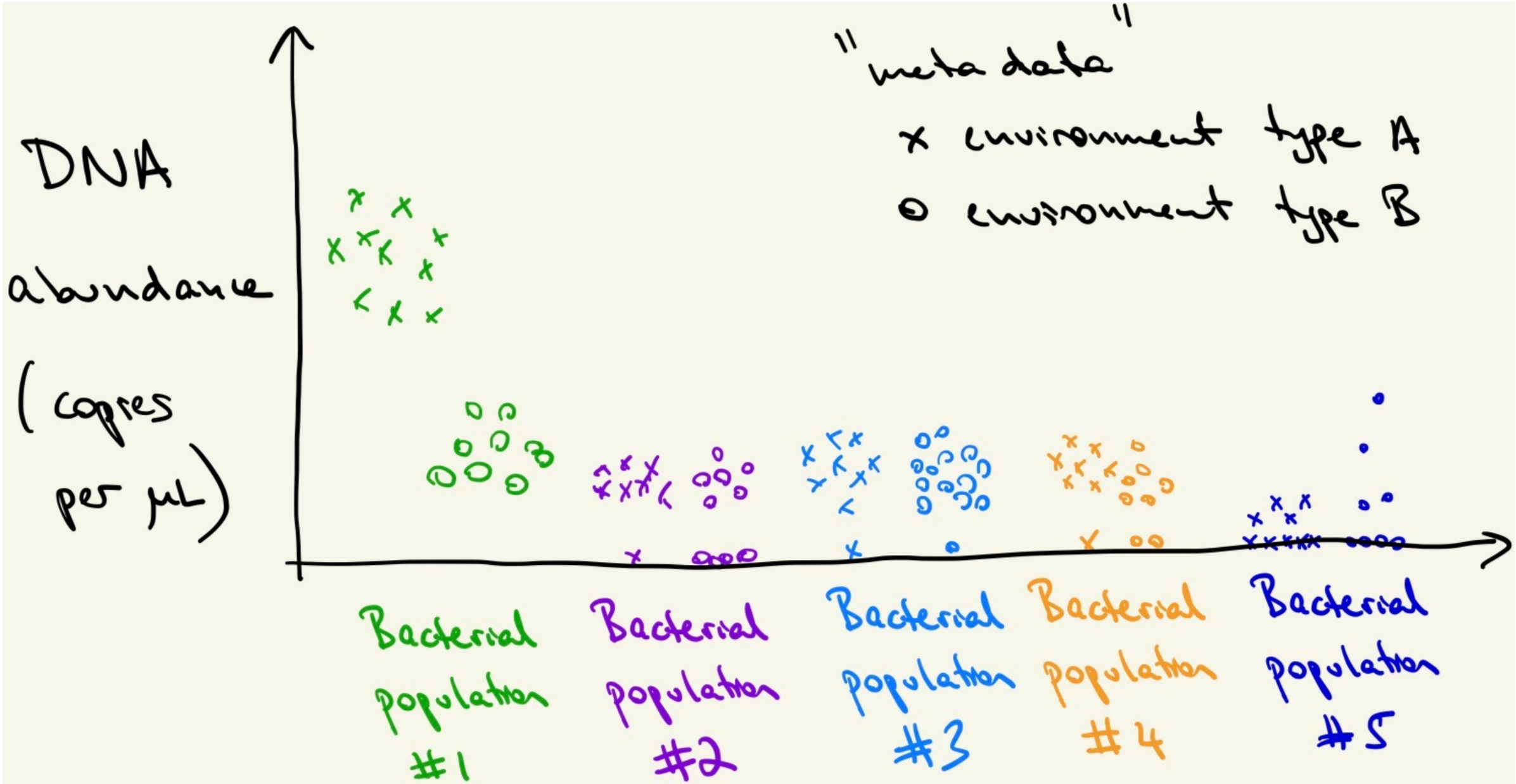
# Data

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

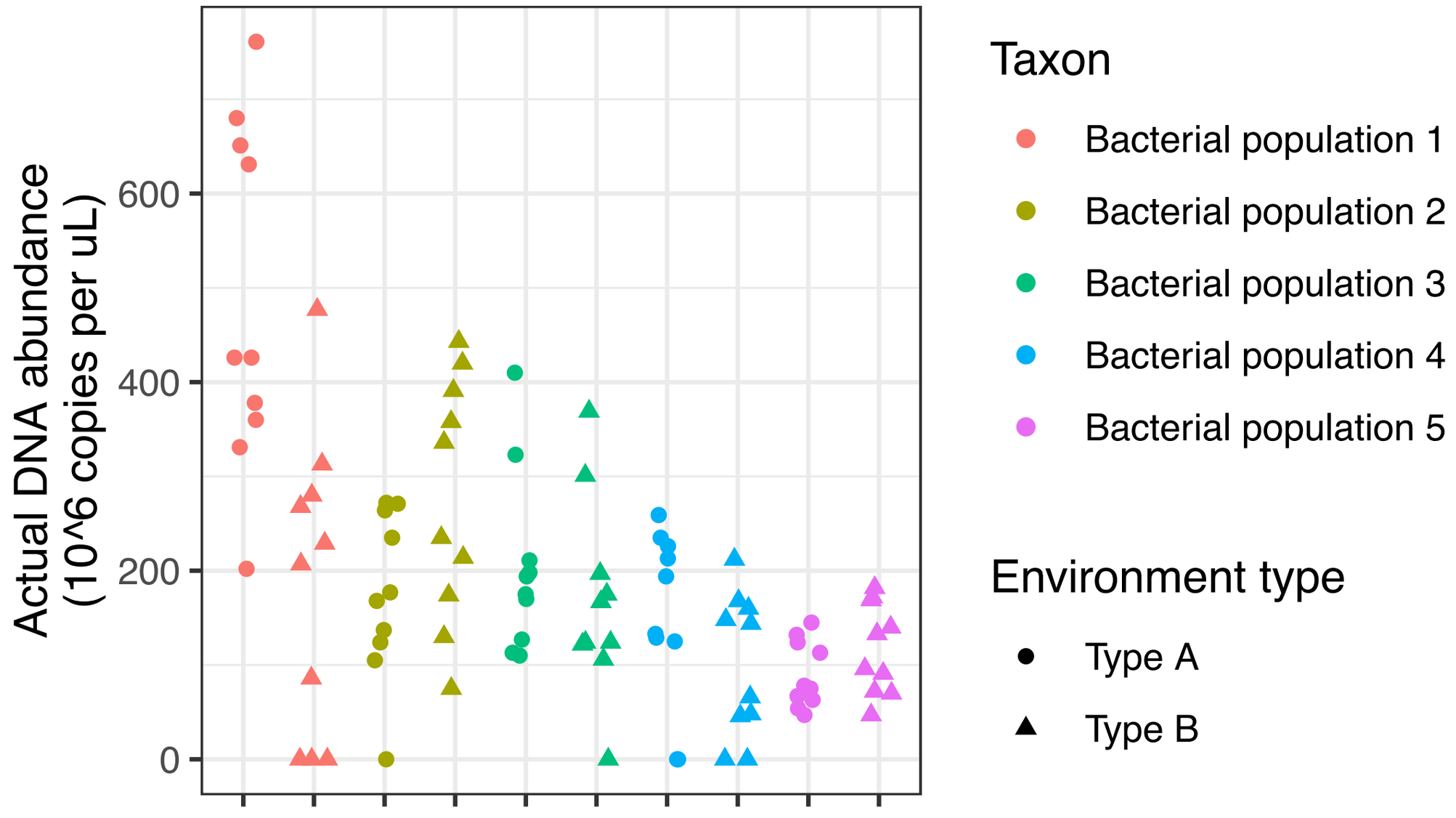
- You <u>must</u> know the source of your data



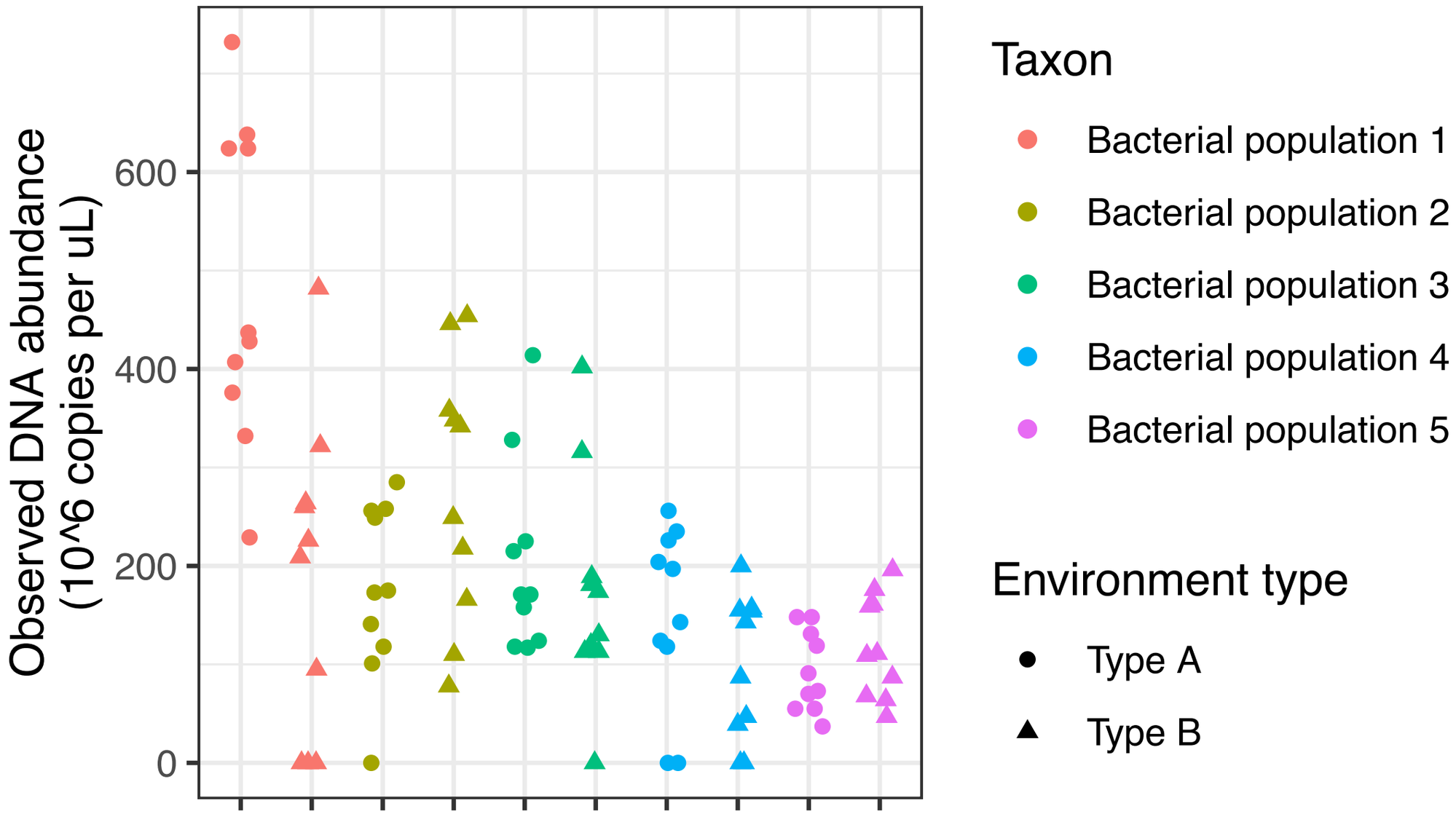
# The environment



# The environment



# Concentration data*

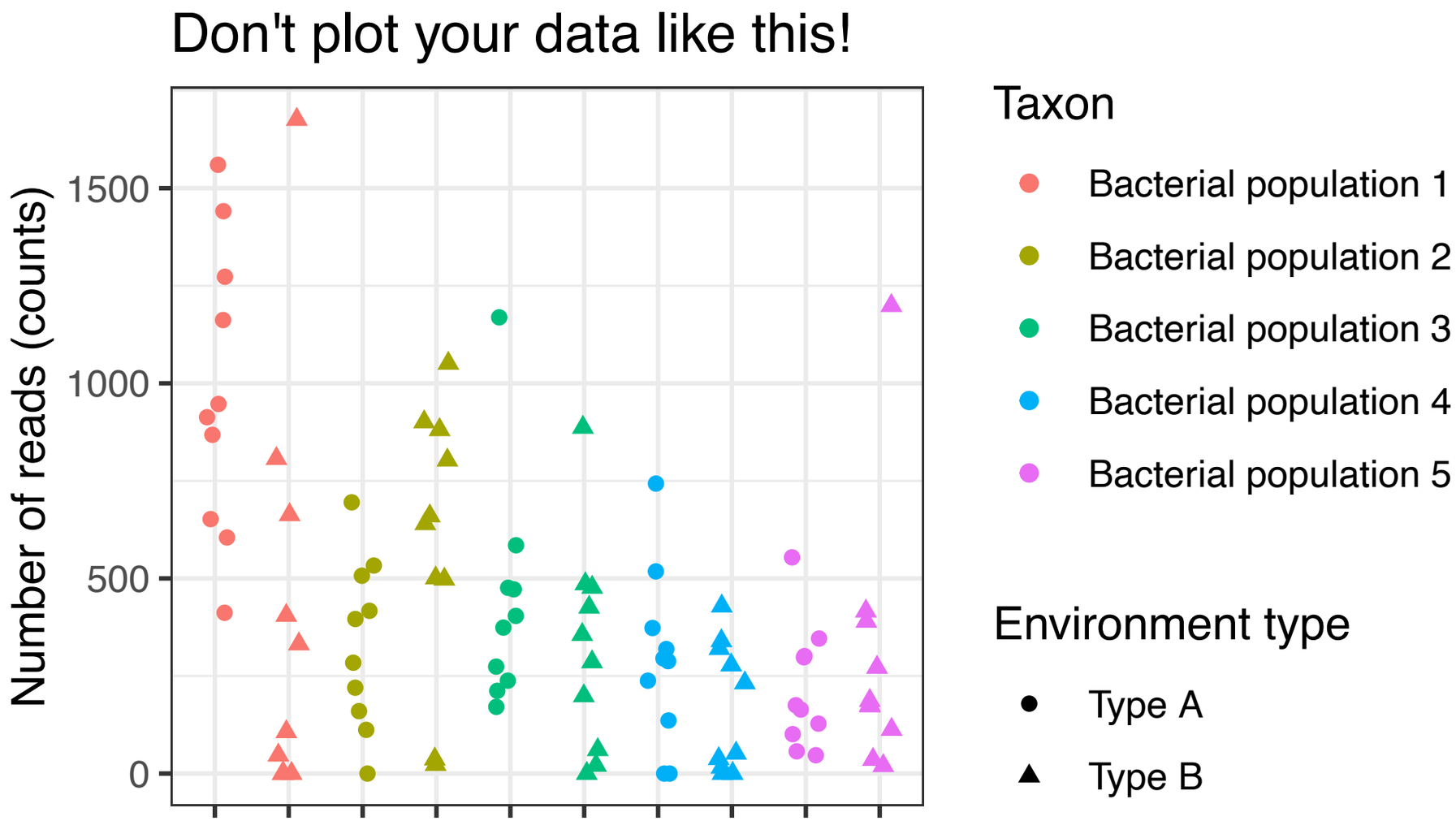





*optimistically

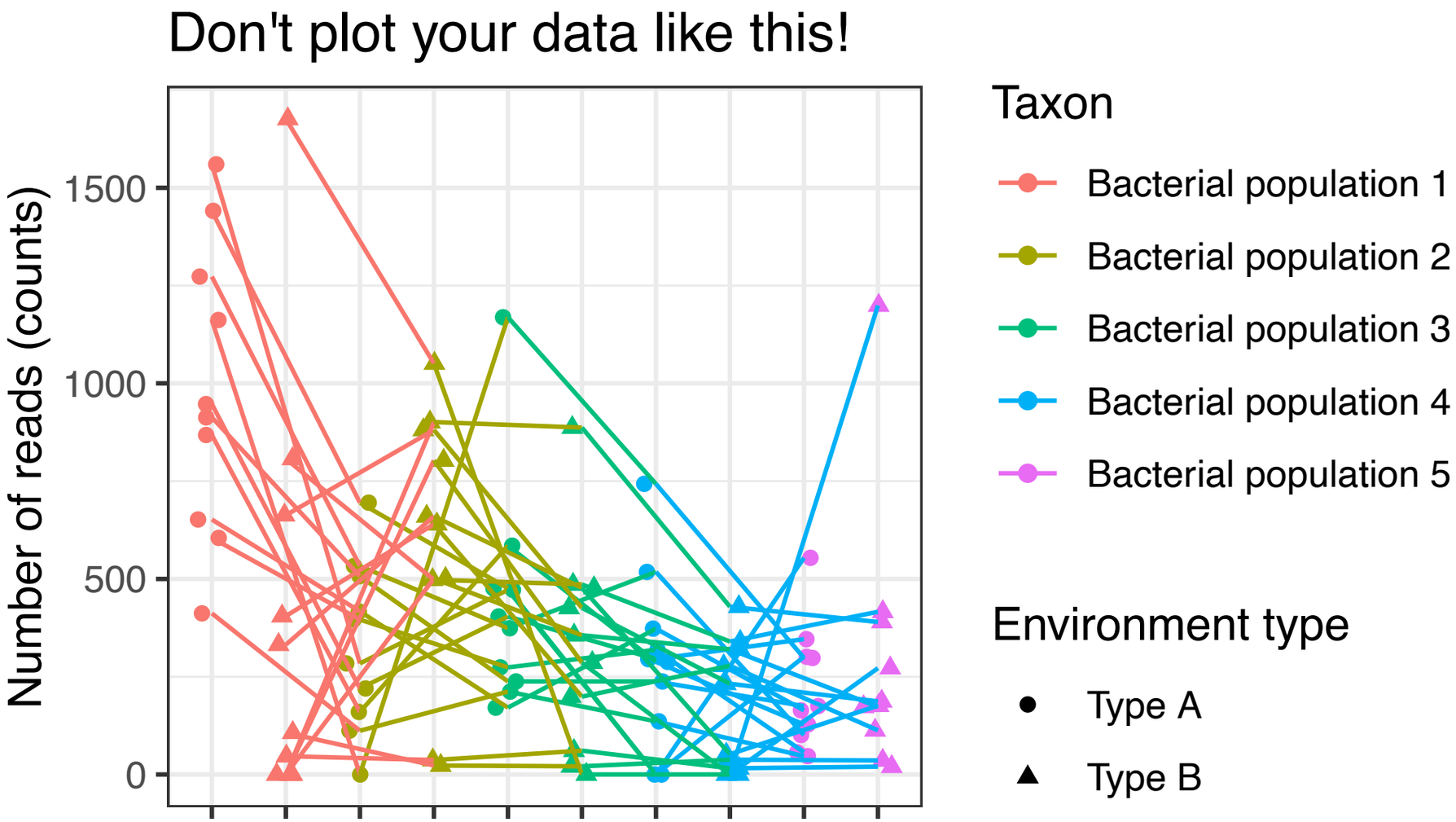
# HTS data

- We get a random number of reads per sample

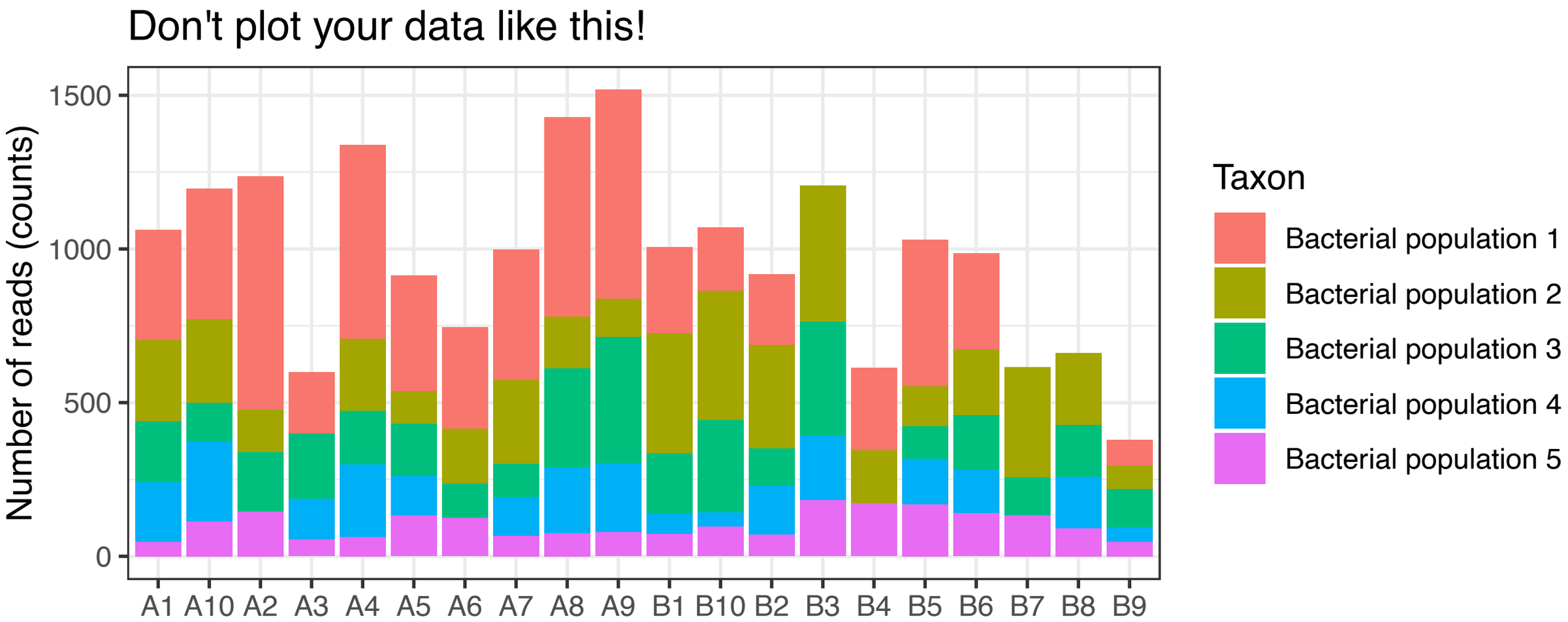


# HTS data

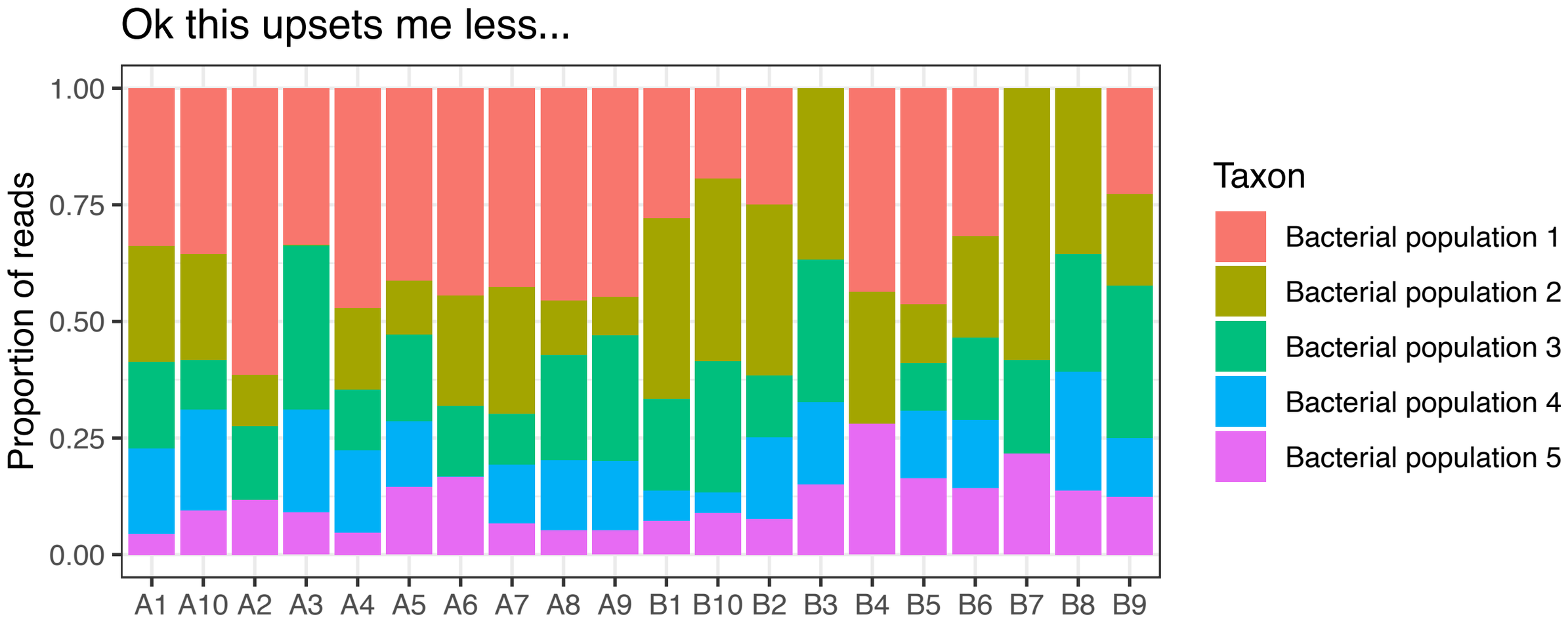
- We get a random number of reads per sample



# 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

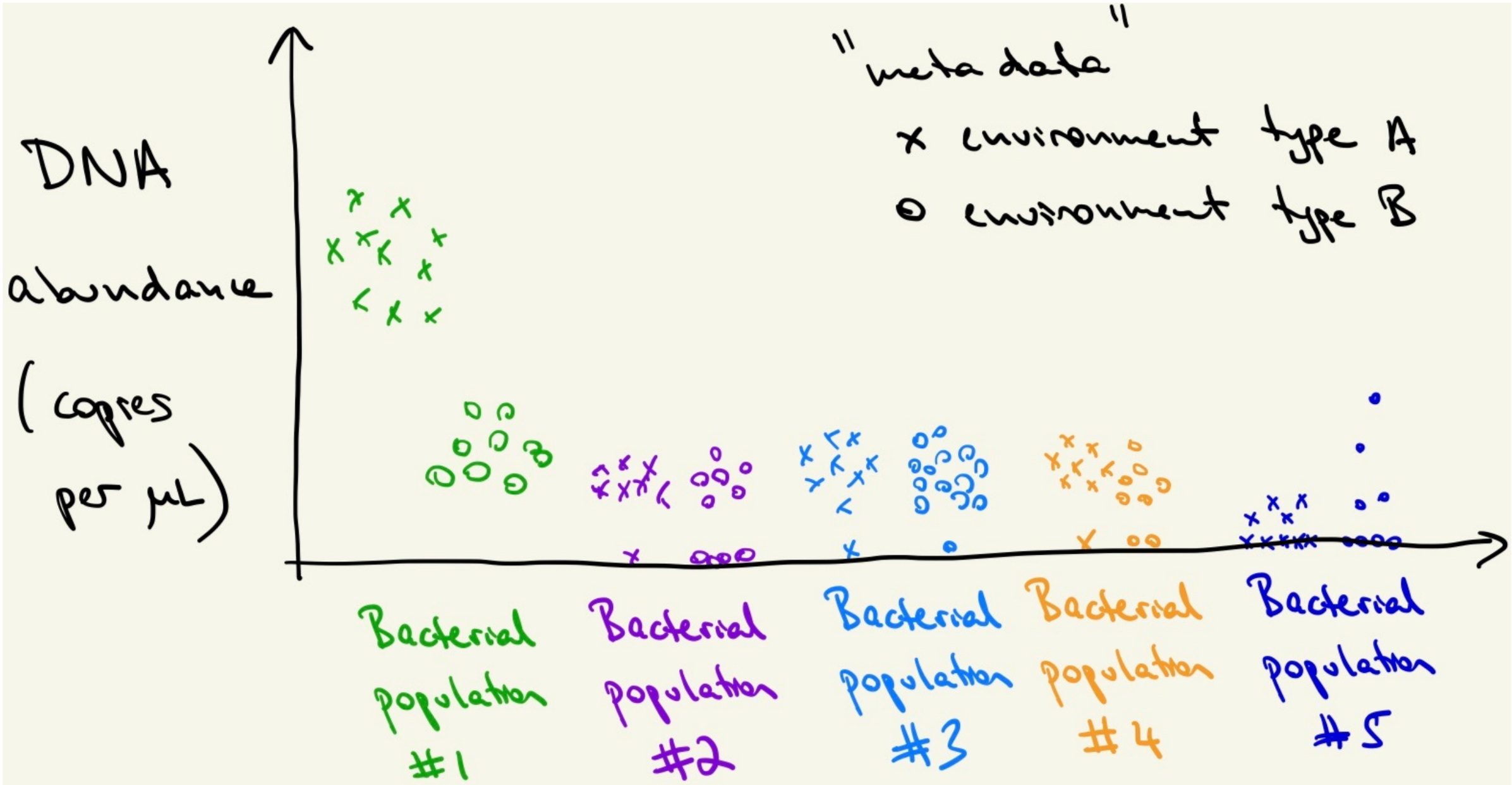

Ok this upsets me less...
Proportion of reads
1.00
0.75
0.50
0.25
0.00
A1
A10
A2
A3
A4
A5
A6
A7
A8
A9
B1
B10
B2
B3
B4
B5
B6
B7
B8
B9
Taxon
Bacterial population 1
Bacterial population 2
Bacterial population 3
Bacterial population 4
Bacterial population 5

# HTS data

- Some considerations

1. Total counts are random ✅
    - Analyzing 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

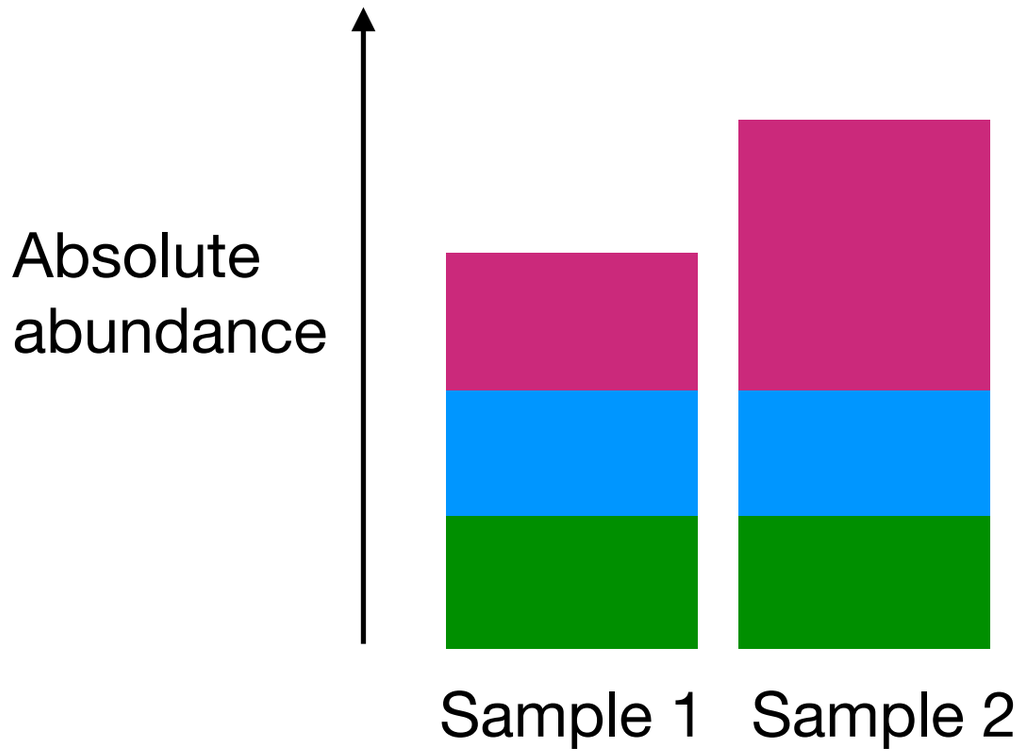
- Relative abundance of *all* taxa change when only *one* taxon's abundance changes

  - 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**

# Are taxa equally well-detected?

- If taxa were equally well-detected, we would have on average, $W_{ij} = c_i Y_{ij}$

$$\mathbb{E} W_{ij} = c_i Y_{ij}$$

- Let's evaluate the evidence!

# Are taxa equally well-detected?

- Mock community: An artificially constructed community of known composition

| | L.crispatus | L.iners | A.vaginae | S.agalactiae | G.vaginalis | S.amnii | P.bivia |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1.00 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0.5 | 0 | 0 | 0 | 0.5 |
| 3 | 0.33 | 0.33 | 0 | 0 | 0 | 0 | 0.33 |
| 4 | 0.33 | 0.33 | 0 | 0.33 | 0 | 0 | 0 |

# Are taxa equally well-detected?

| | L.crispatus | L.iners | A.vaginae | S.agalactiae | G.vaginalis | S.amnii | P.bivia |
|---|---|---|---|---|---|---|---|
| **1** | 19 | 4 | 2 | 51332 | 1 | 14 | 1 |
| **2** | 0 | 1 | 1424 | 0 | 0 | 7 | 21708 |
| **3** | 4775 | 11234 | 0 | 0 | 0 | 1 | 3249 |
| **4** | 1644 | 5497 | 1 | 4521 | 0 | 7 | 0 |

| | L.crispatus | L.iners | A.vaginae | S.agalactiae | G.vaginalis | S.amnii | P.bivia |
|---|---|---|---|---|---|---|---|
| **1** | 0 | 0 | 0 | 1.00 | 0 | 0 | 0 |
| **2** | 0 | 0 | 0.5 | 0 | 0 | 0 | 0.5 |
| **3** | 0.33 | 0.33 | 0 | 0 | 0 | 0 | 0.33 |
| **4** | 0.33 | 0.33 | 0 | 0.33 | 0 | 0 | 0 |



Data from Brooks *et al* (2015) BMC Microbiology

# Are taxa equally well-detected?

1. Despite equal mixing fractions, some taxa are observed many more times

2. Despite being purportedly absent, taxa are observed

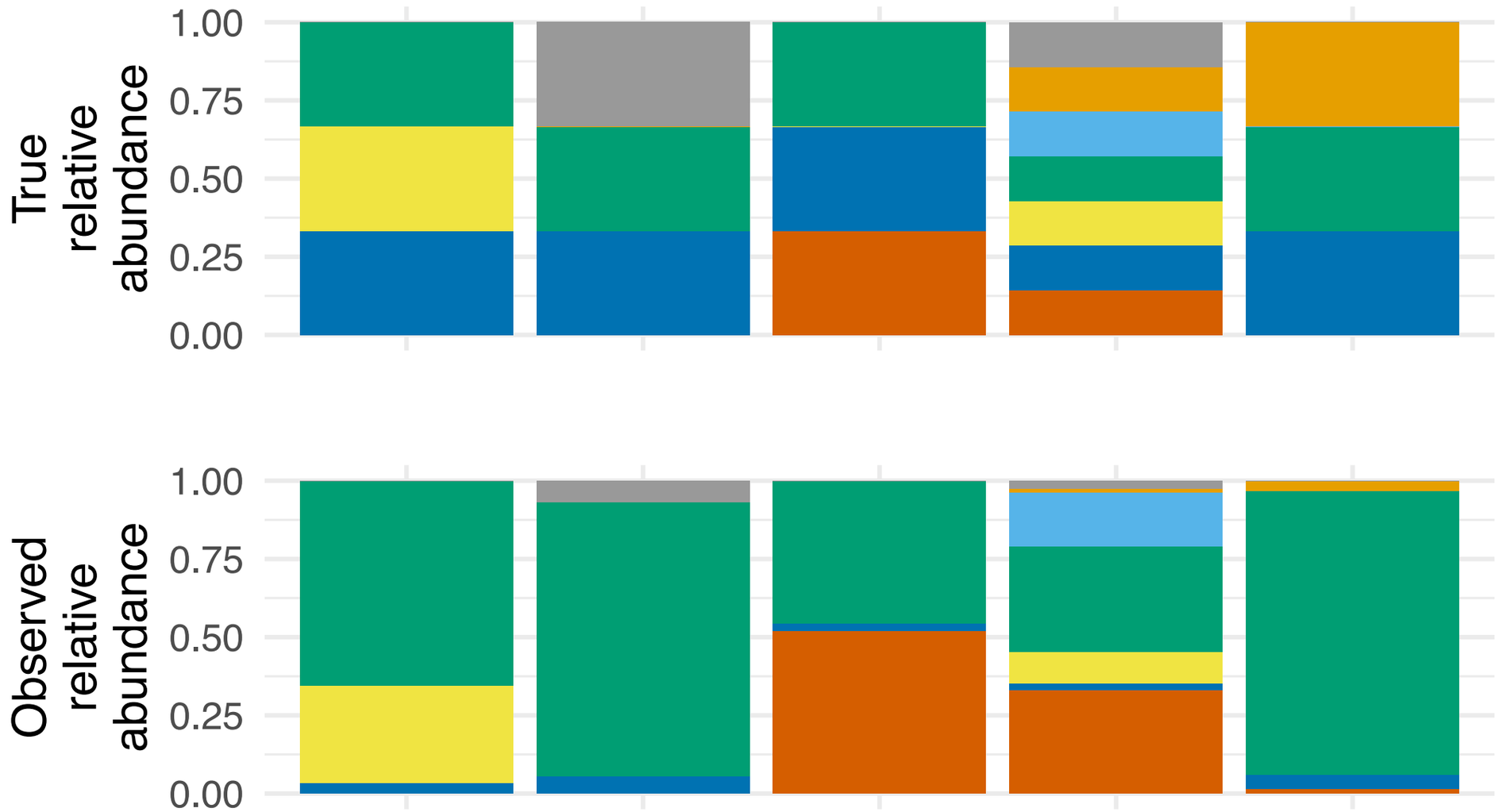
# #3 Taxa are unequally well-detected

- Despite the common assumption that

$$\mathbb{E}[W_{ij}] = c_i Y_{ij}$$

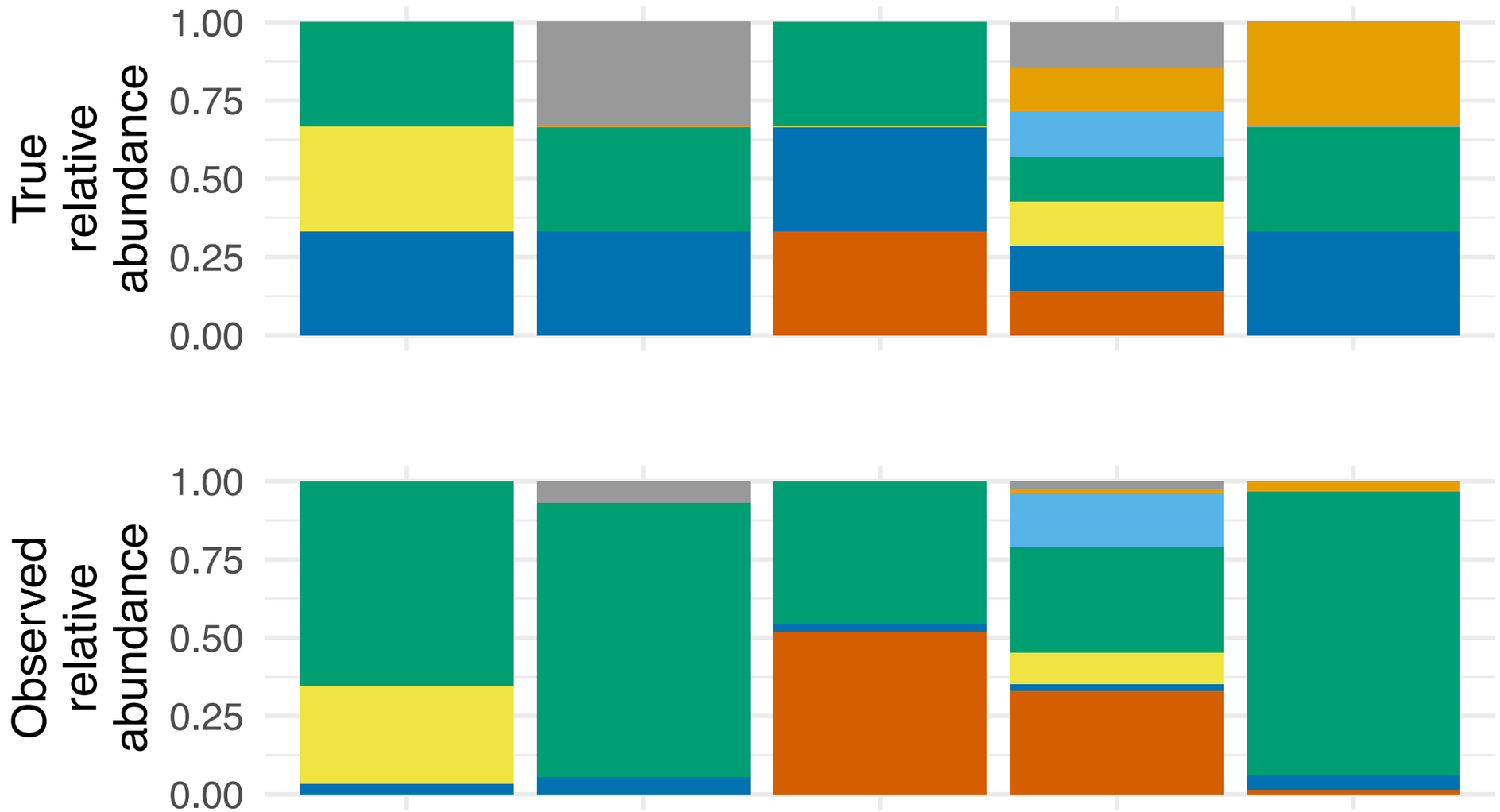
some taxa are *over-observed* for equal $c_i$ and $Y_{ij}$

- What model better explains this observation?






Data from Brooks *et al* (2015) BMC Microbiology




| L.iners/S.agalactiae | | | | | | |
|---|---|---|---|---|---|---|
| | 19.5 | 16.1 | 18.9 | 17.4 | 19.9 | **18.3 (1.6)** |

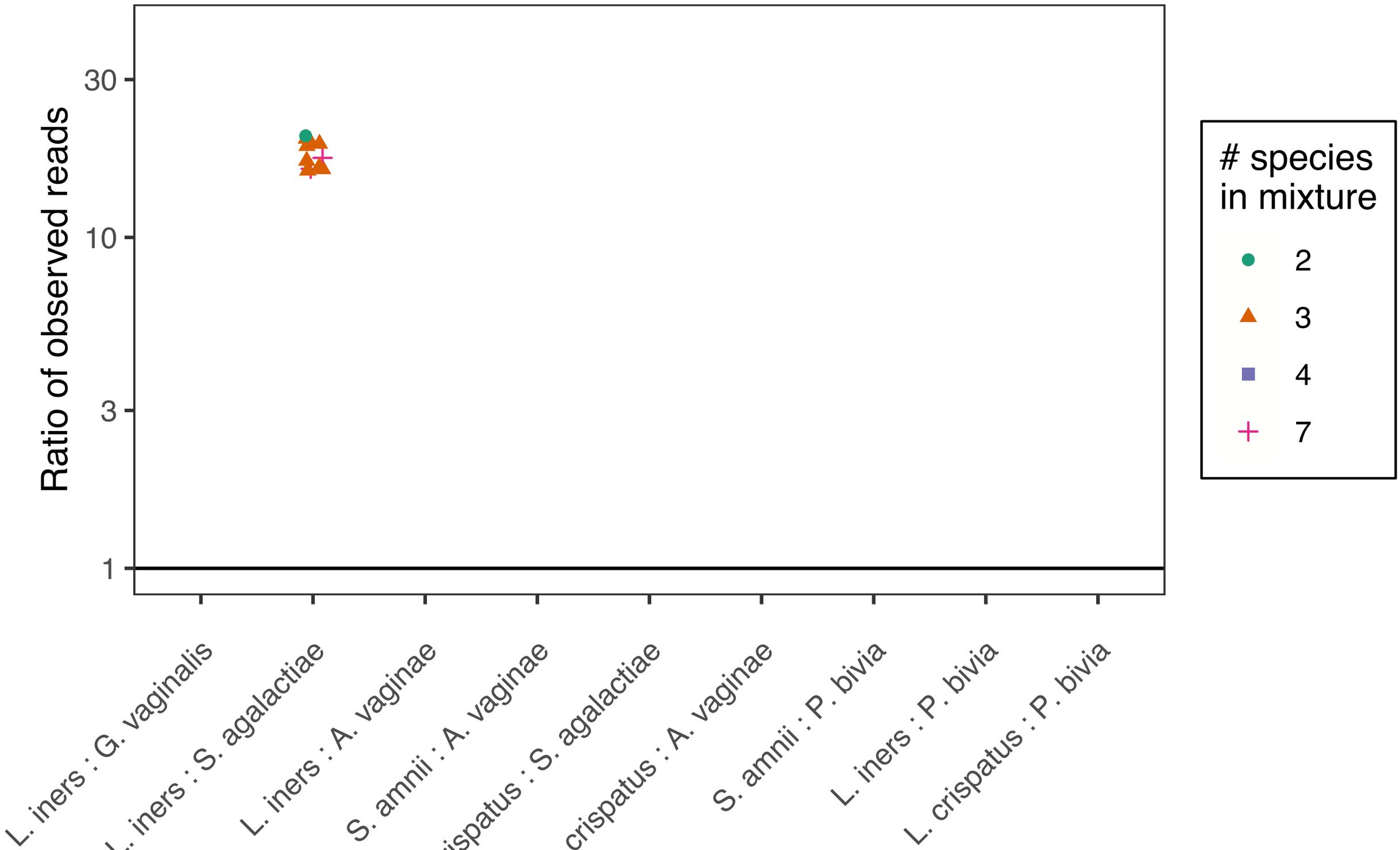



 Figure adapted from McLaren, Willis & Callahan, 2019, *eLife*

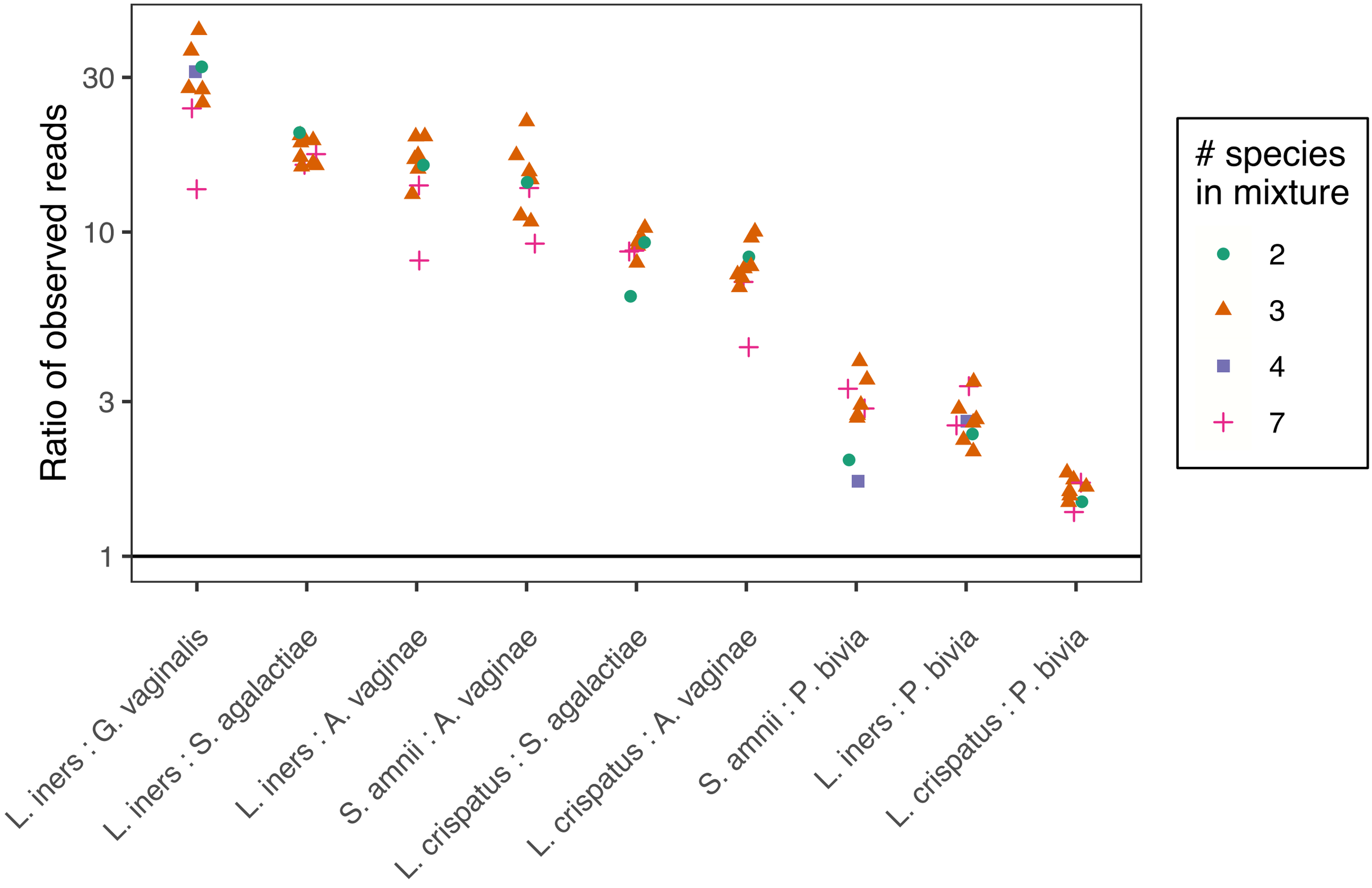



 Figure adapted from McLaren, Willis & Callahan, 2019, *eLife*

# #3 Taxa are unequally well-detected

- Evidence *against*

$$\mathbb{E}[W_{ij}] = c_i \times Y_{ij}$$

- Better support for

$$\mathbb{E}[W_{ij}] = c_i \times e_j \times Y_{ij}$$

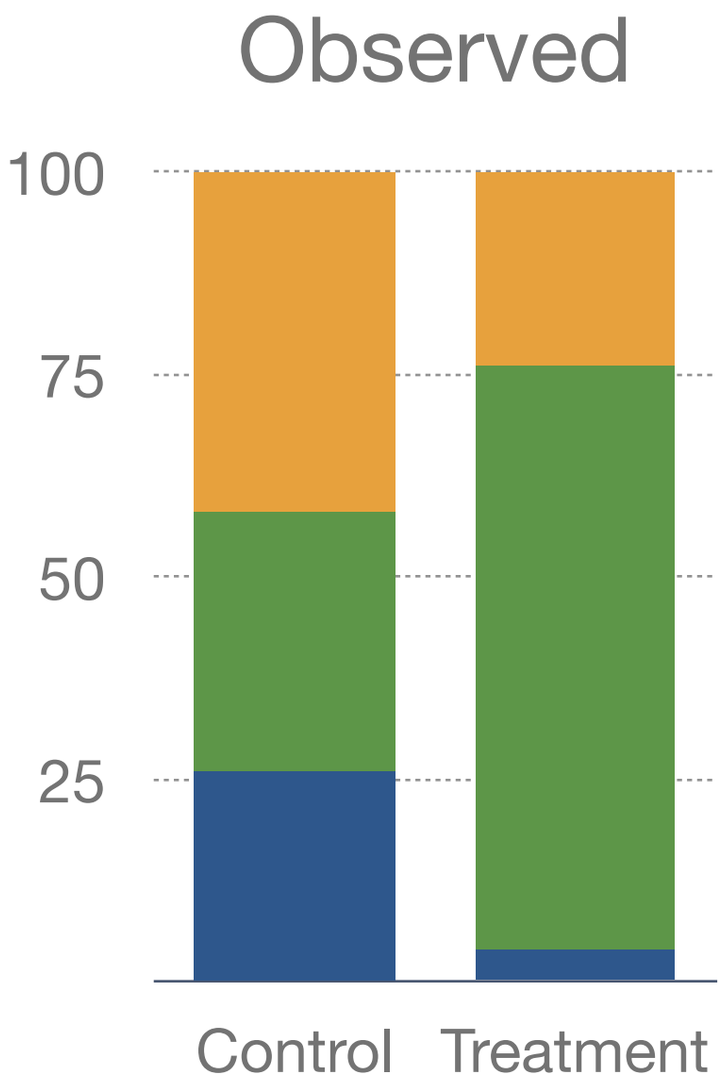
Why is this so important for data analysis?

# #3 Taxa are unequally well-detected

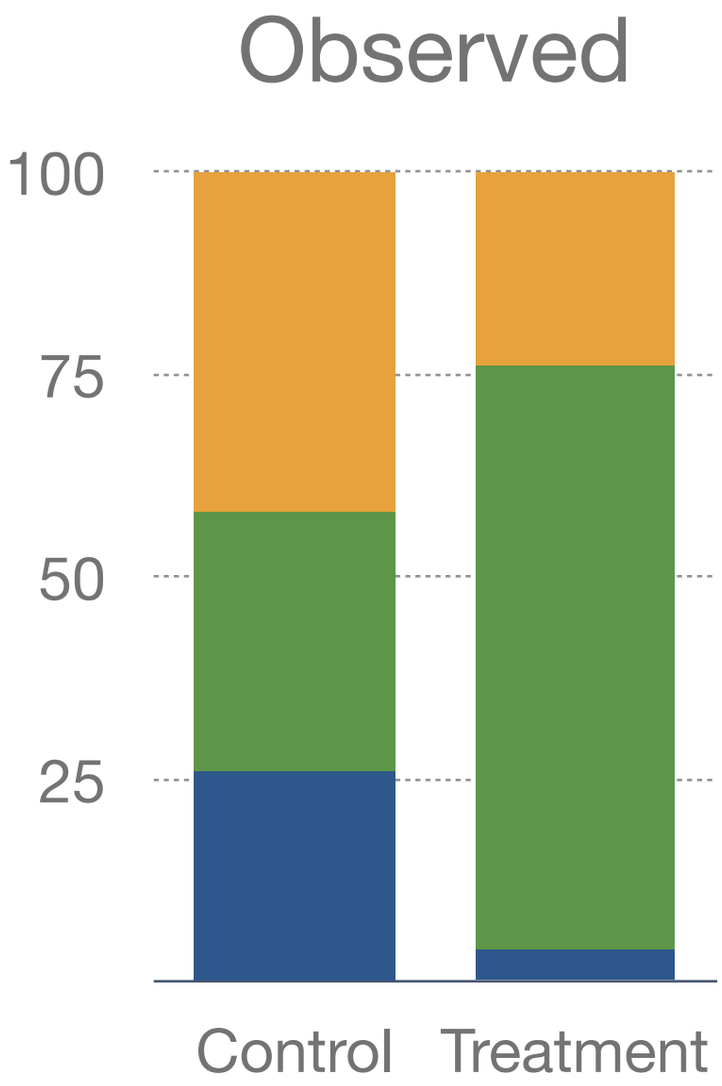
- Stated differently,

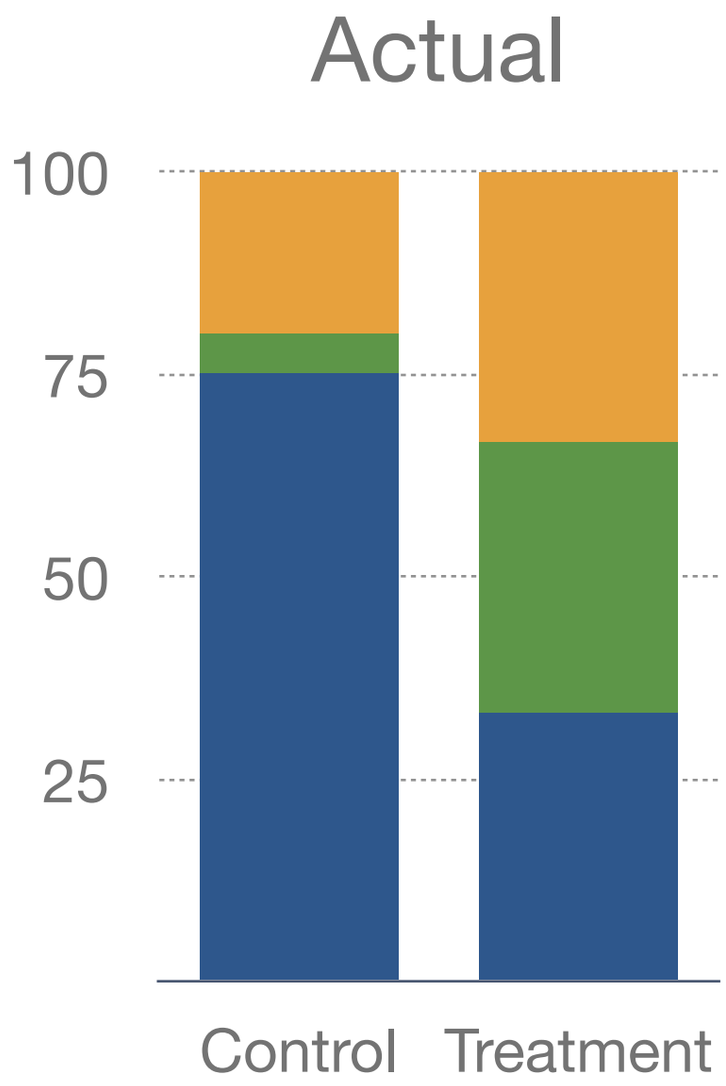
Observed relative abundance $\propto$ True relative abundance $\times$ Taxon-specific efficiencies

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



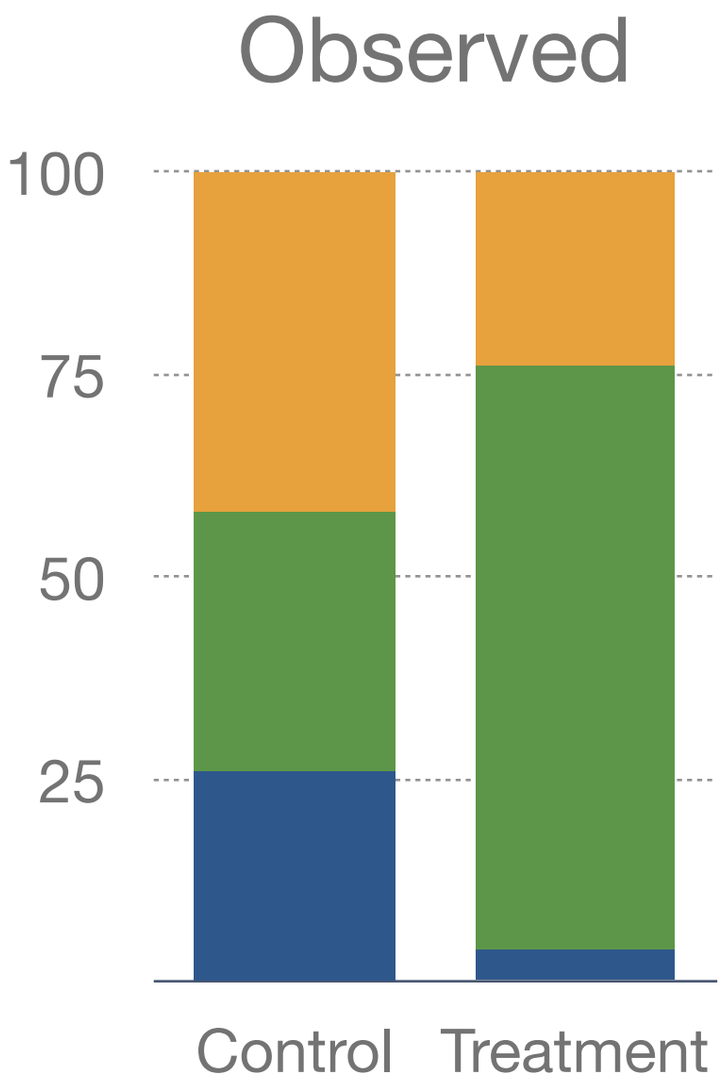

- A tempting conclusion:
  - The relative abundance of **orange** *<u>decreased</u>* in the Treatment sample (right) compared to the Control sample (left)

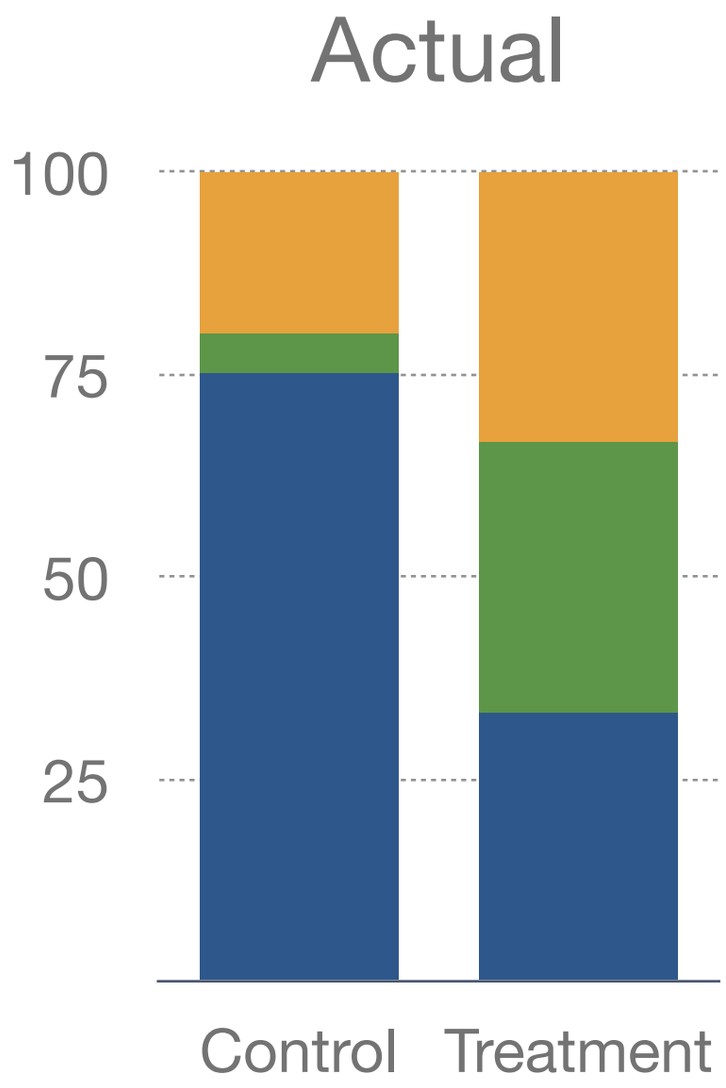


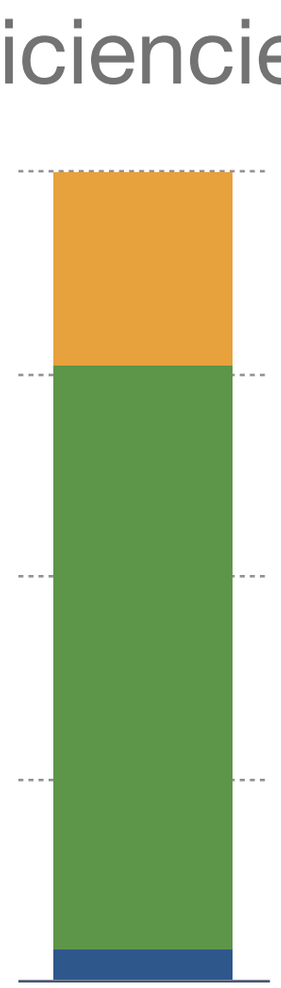




- In fact, the relative abundance of **orange** *increased* in the Treatment sample compared to the Control sample

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

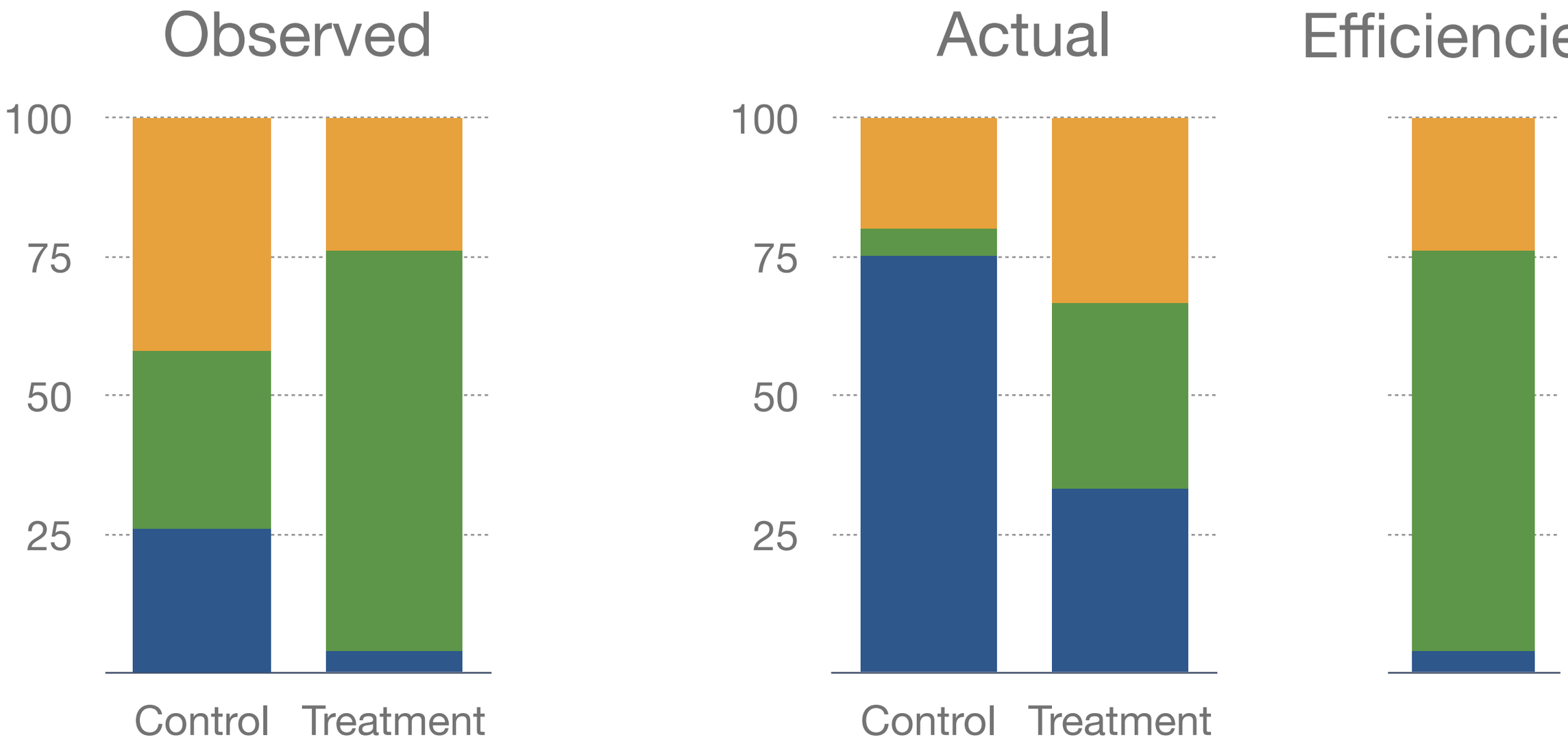



- In fact, the relative abundance of **orange** *increased* in the Treatment sample compared to the Control sample

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




- **Green** is high efficiency; its abundance increased.
  **Blue** is low efficiency, and its abundance decreased.

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

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

# HTS data

- Some considerations

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

# 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 *shouldn't* be compared across samples
  - What can be compared? Ratios and fold differences, more soon…

# Absolute abundance data

- qPCR/ddPCR data can usually be modeled with techniques you already know!
- Linear regression to estimate additive differences in means
- Poisson regression to estimate fold-differences in means

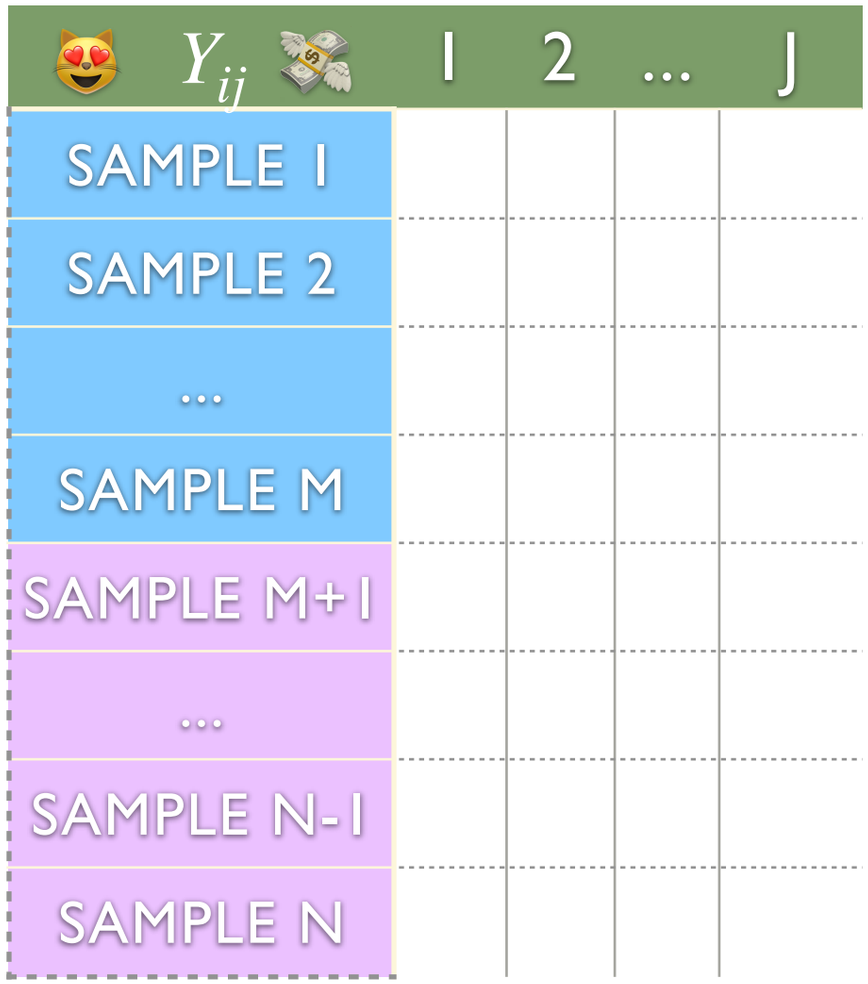
Choose based on *parameters*, not *data characteristics*

# Modeling abundance

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

# Differential abundance

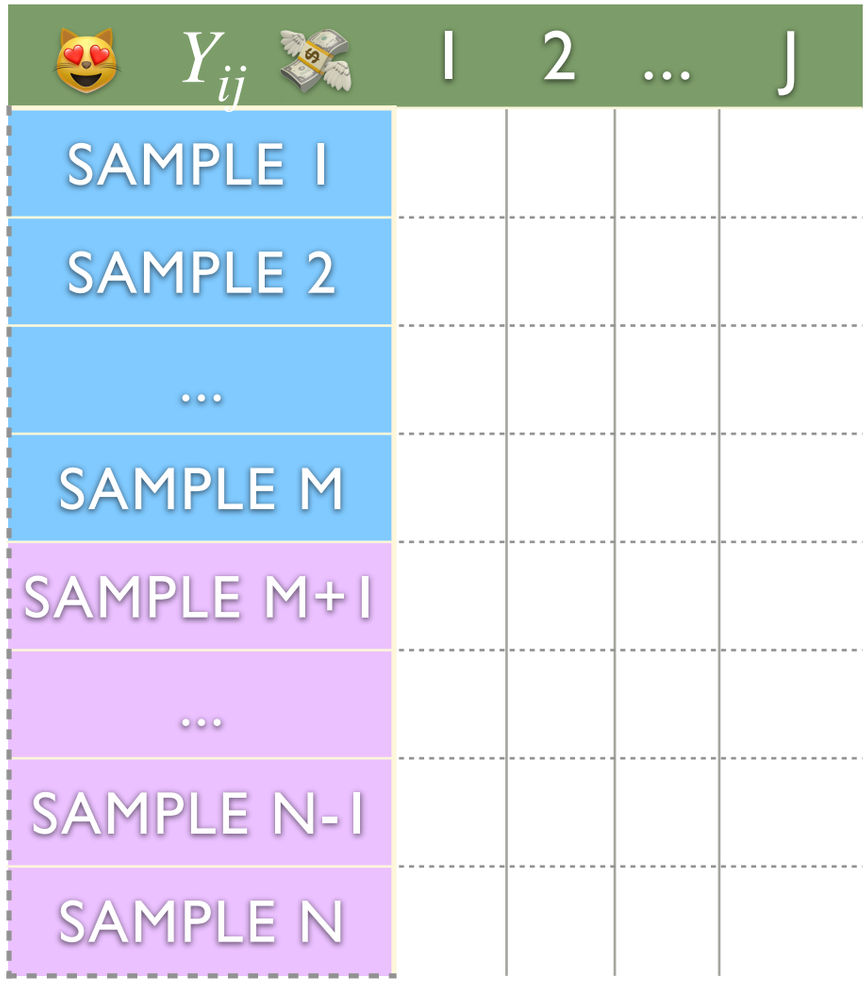
- What are different ways the abundance of unit $j$ could be "different" across the environment types treatment and control?




Dream big! Remember, you know *everything*!

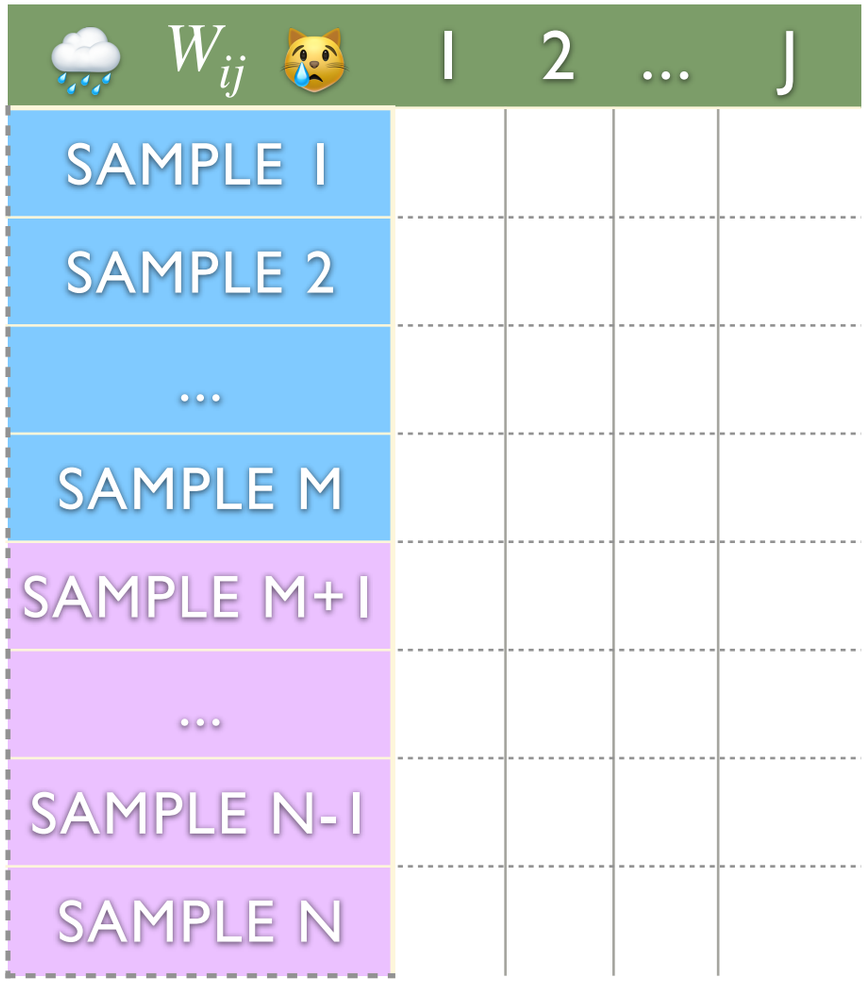
# Differential abundance

- …



# Differential abundance & identifiability

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



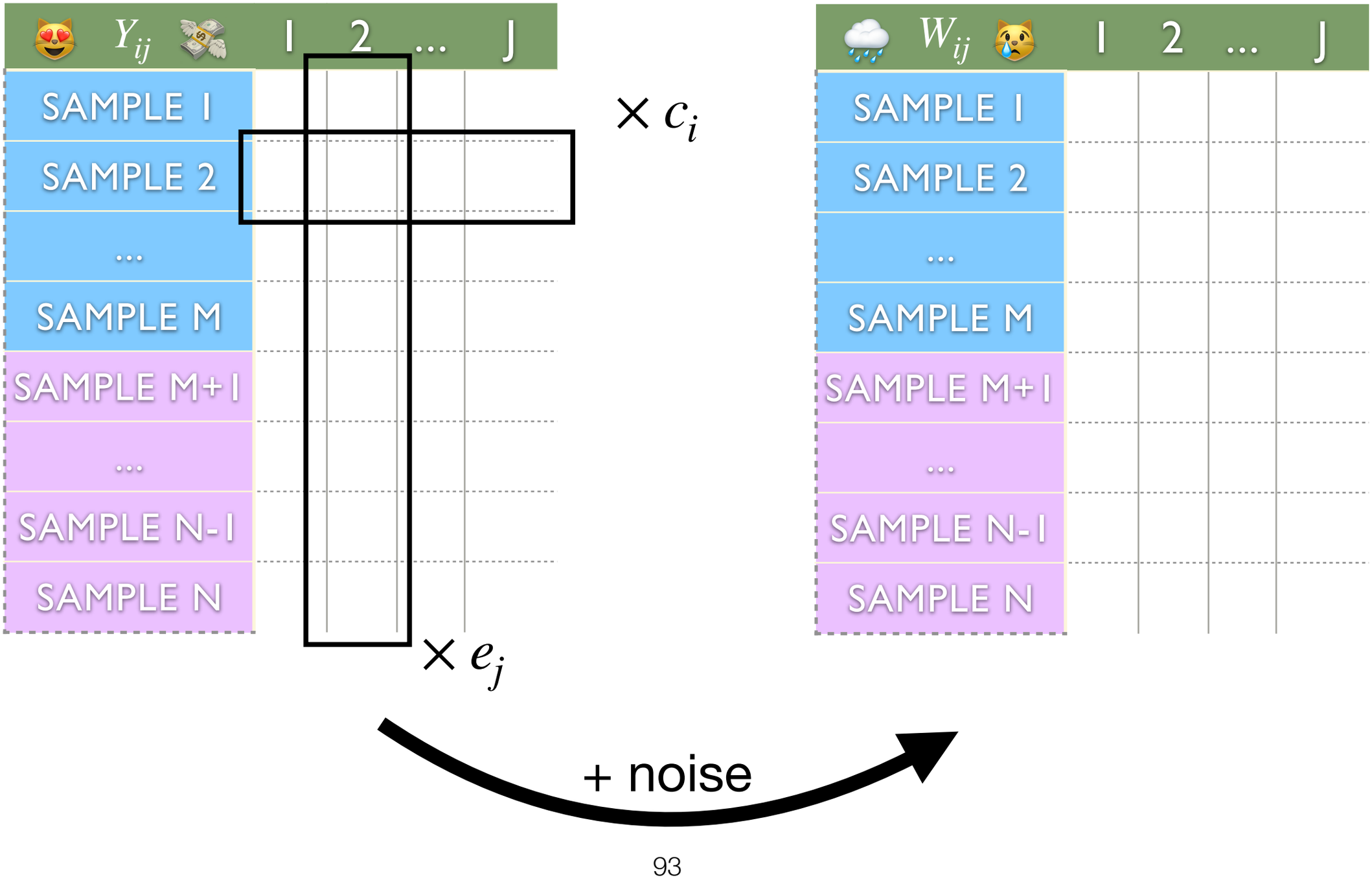

- not *identifiable* = we can't learn them from the $W'_{ij}s$ under reasonable assumptions

- 💔



*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

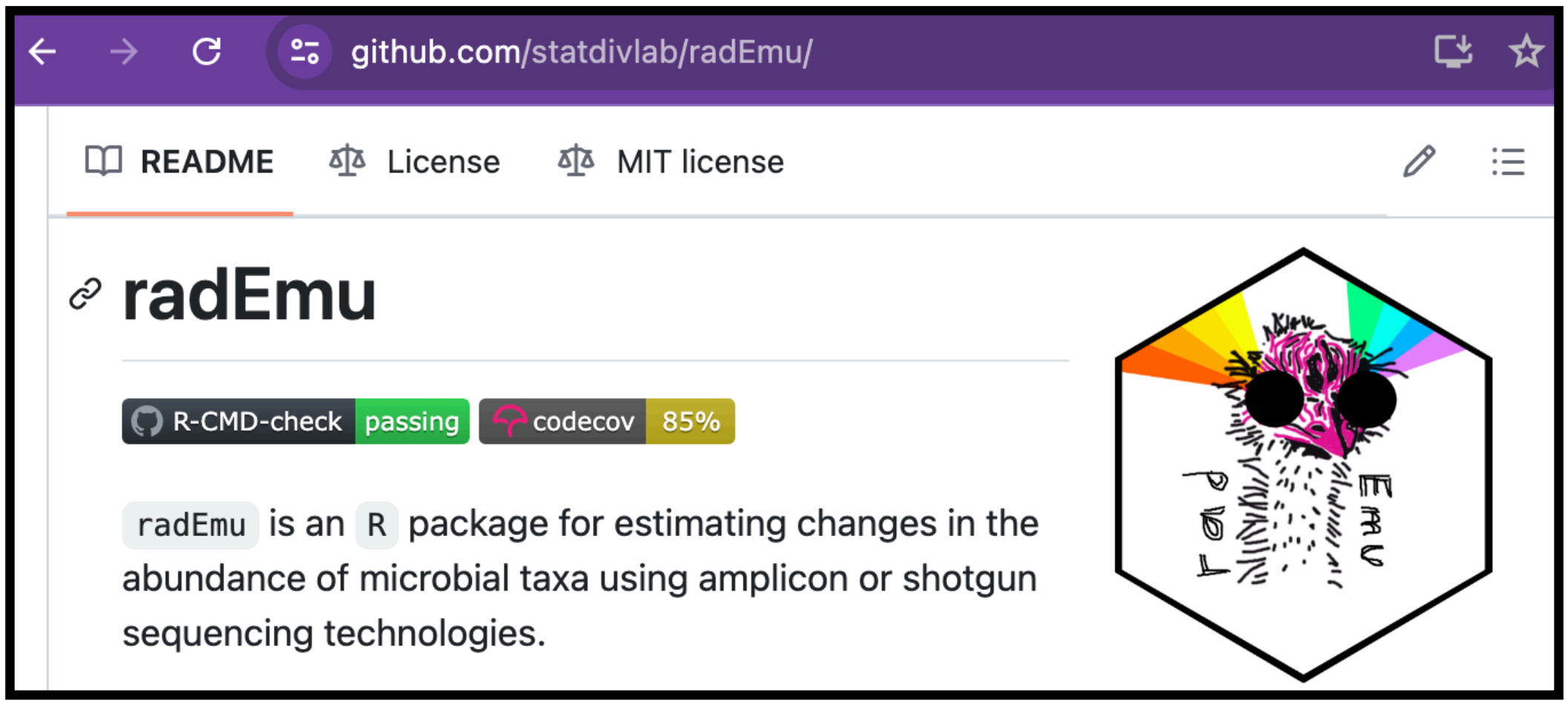
$$\log \frac{\mathbb{E}Y_{\text{group 1},j}}{\mathbb{E}Y_{\text{group 2},j}} - \text{average}_{j'} \log \left( \frac{\mathbb{E}Y_{\text{group 1},j'}}{\mathbb{E}Y_{\text{group 2},j'}} \right)$$

- Great if no reference category

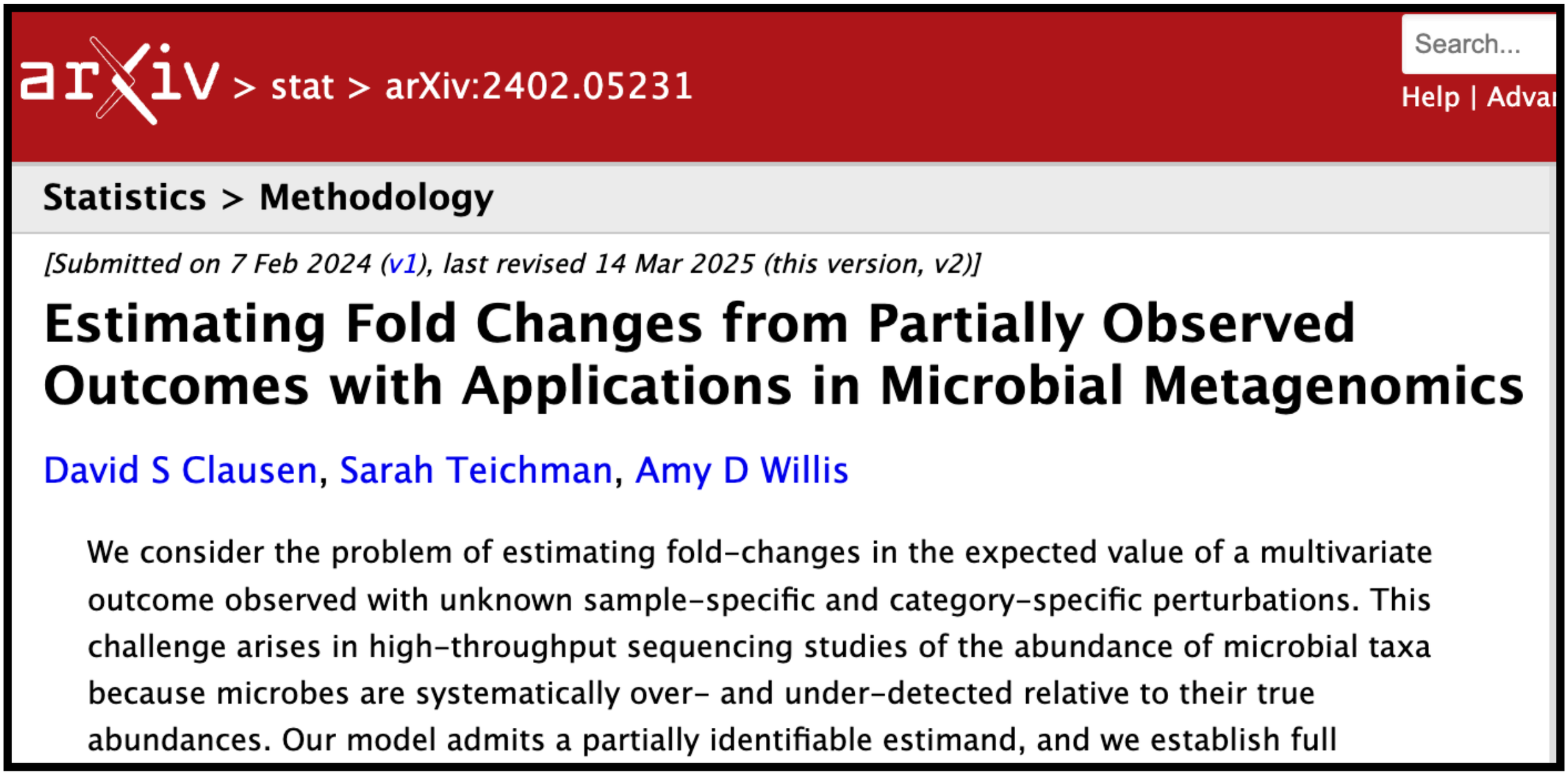
- average = mean or smoothed median

# My favourite method…

- 🦤radEmu🦤 can estimate whichever you prefer!



# My favourite method...



# 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

- **Limitations**
  - Slower than other methods might run overnight
    - Sarah has a 🤩 new method 😻 that chooses a “reference set” for you and is *up to 1,000x* faster…
      - You can demo 🍡 fastEmu 🍡 today!
  - New?
- Anything else? Tell us!

# 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

# 🦤 Now: radEmu lab 🦤

1. Wiki ➡️ Schedule ➡️ “Statistics labs”
2. Download the .zip file “differential-abundance.zip”
3. Upload the .zip file to the RStudio server
4. Work through “radEmu_lab.Rmd”

🌴 *This is an ordinary quest* 🌴

🟩 *Green sticky = “I’m finished”* 🟩

💗 *Pink sticky = “I am joining the queue with a question.”* 💗

👋 *Hands up / waving / audible crying = “I am stuck, help me now plz”* 👋

# 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

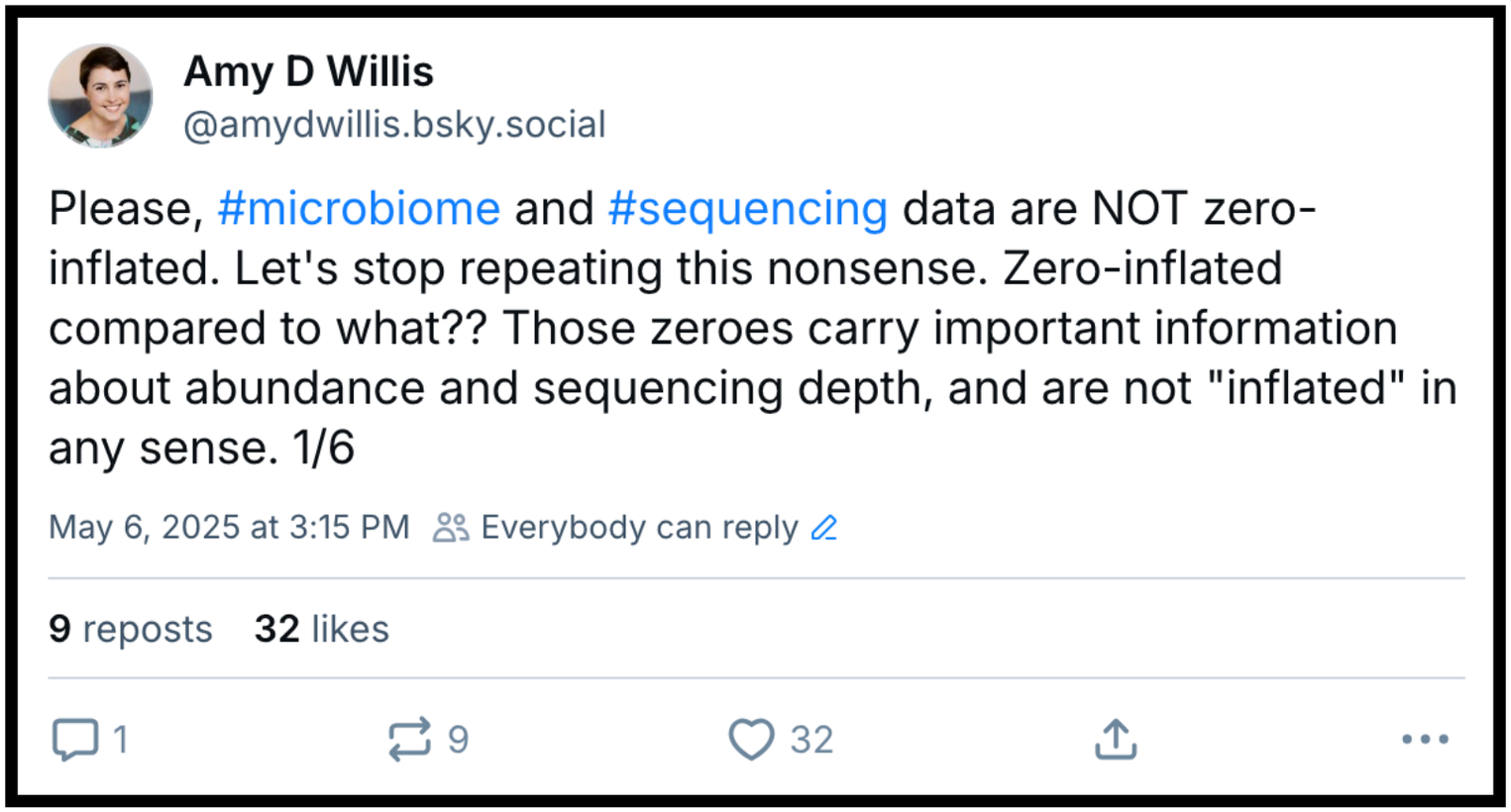
- Many methods exist for “differential abundance”
  - ALDEx2
  - ANCOM-BC2
  - corncob
  - DESeq2
  - edgeR
  - metagenomeSeq
  - MaAsLin3
  - LEfSE
  - limma voom
  - LinDA
  - radEmu
  - Wilcoxon/t-tests on proportions
  - t-tests on ratios
  - many others!
- multiple versions of almost all methods; 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 ✅

# HTS data


Amy D Willis
@amydwillis.bsky.social
Please, #microbiome and #sequencing data are NOT zero-inflated. Let's stop repeating this nonsense. Zero-inflated compared to what?? Those zeroes carry important information about abundance and sequencing depth, and are not "inflated" in any sense. 1/6
May 6, 2025 at 3:15 PM
Everybody can reply
9 reposts
32 likes
1
9
32

# HTS data


Amy D Willis
@amydwillis.bsky.social
Please, #microbiome and #sequencing data are NOT zero
inflated.
compare
about ab
any sens
May 6, 2025
9 reposts
1
9
32
Amy D Willis @amydwillis.bsky.social · 2mo
In fact, if you look at blanks and other control data, you see a lot of incorrect detections. There's better evidence that microbiome data is NON-ZERO inflated than zero-inflated.
2/6

# HTS data

**Amy D Willis**
@amydwillis.bsky.social

Please, #microbiome and #sequencing data are NOT zero-
inflated.
compare
about ab
any sens

May 6, 2025

**9** reposts

1

**Amy D Willis** @amydwillis.bsky.social · 2mo

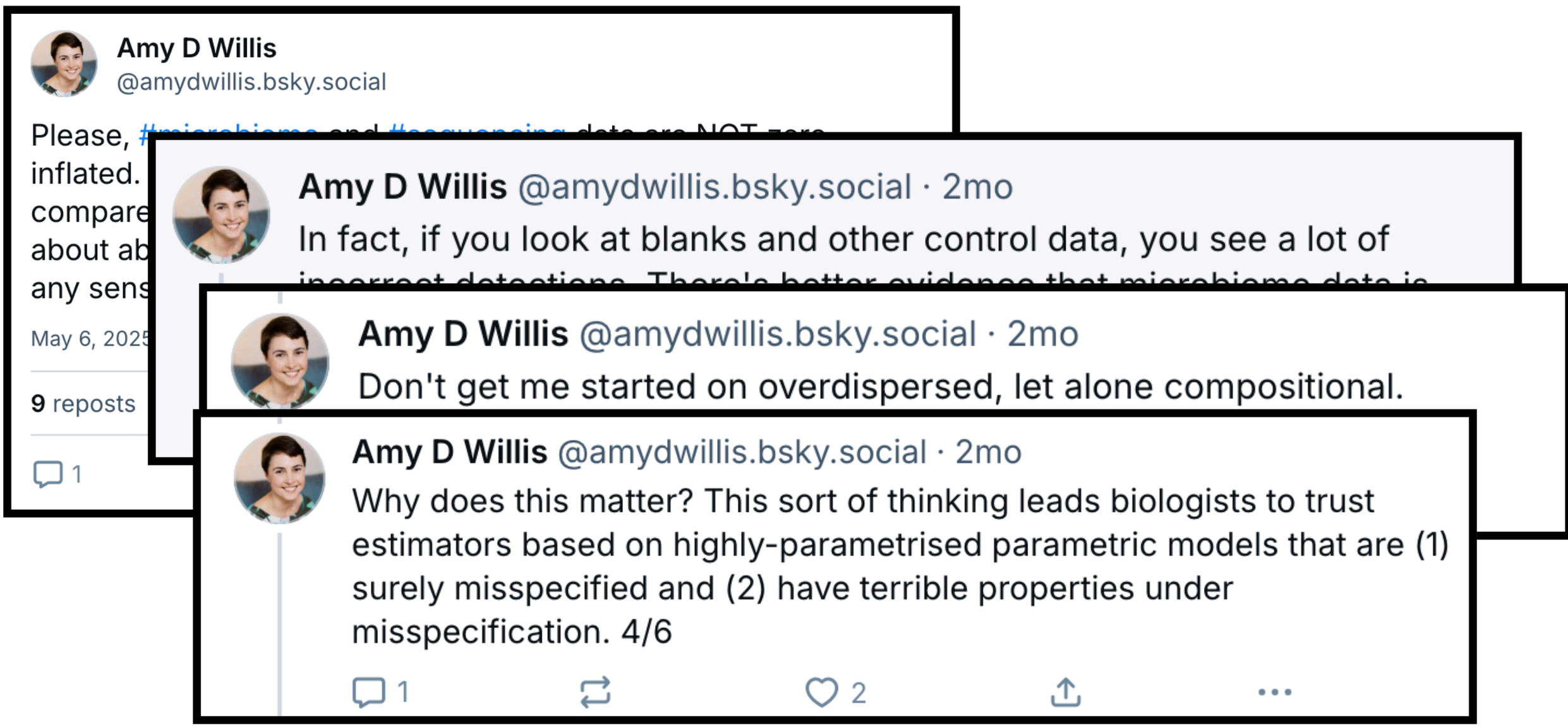
In fact, if you look at blanks and other control data, you see a lot of
incorrect detections. There's better evidence that microbiome data is

**Amy D Willis** @amydwillis.bsky.social · 2mo

Don't get me started on overdispersed, let alone compositional.
Microbiome data is none of these, and I'm not new to this field. 3/6

1 2

# HTS data


Amy D Willis
@amydwillis.bsky.social
Please, #microbiome and #sequencing data are NOT zero-
inflated.
compare
about ab
any sens
May 6, 2025
9 reposts
1
Amy D Willis @amydwillis.bsky.social · 2mo
In fact, if you look at blanks and other control data, you see a lot of
incorrect detections. There's better evidence that microbiome data is
Amy D Willis @amydwillis.bsky.social · 2mo
Don't get me started on overdispersed, let alone compositional.
Amy D Willis @amydwillis.bsky.social · 2mo
Why does this matter? This sort of thinking leads biologists to trust estimators based on highly-parametrised parametric models that are (1) surely misspecified and (2) have terrible properties under misspecification. 4/6
1
2

# HTS data

**Amy D Willis**
@amydwillis.bsky.social

Please, #microbiome and #sequencing data are NOT zero-inflated. [illegible] compare[illegible] about ab[illegible] any sens[illegible]

May 6, 2025

**9** reposts

1

**Amy D Willis** @amydwillis.bsky.social · 2mo

In fact, if you look at blanks and other control data, you see a lot of incorrect detections. There's better evidence that microbiome data is

**Amy D Willis** @amydwillis.bsky.social · 2mo

Don't get me started on overdispersed, let alone compositional.

**Amy D Willis** @amydwillis.bsky.social · 2mo

Why does this matter? This sort of thinking leads biologists to trust estimators based on highly-parametrised parametric models that are (1)

**Amy D Willis** @amydwillis.bsky.social · 2mo

Saying "microbiome data is zero-inflated" leads people to seek out "zero-inflated models." Usually, these are bad methods with bad properties. Stay away. 5/6

1 2

# HTS data

**Amy D Willis**
@amydwillis.bsky.social

Please, #microbiome and #sequencing data are NOT zero-
inflated.
compare
about ab
any sens

May 6, 2025

**Amy D Willis** @amydwillis.bsky.social · 2mo

In fact, if you look at blanks and other control data, you see a lot of
incorrect detections. There's better evidence that microbiome data is

**Amy D Willis** @amydwillis.bsky.social · 2mo

ompositional.

gists to trust
nodels that are (1)

people to seek out
ethods with bad

**Amy D Willis** @amydwillis.bsky.social · 2mo

I leave you with the StatDivLab mantra:

1. choose something meaningful to estimate
2. choose a sensible way to estimate it
3. choose tests that control Type 1 error

That's what we will keep doing, even if anonymous reviewers insist on buzzwords.

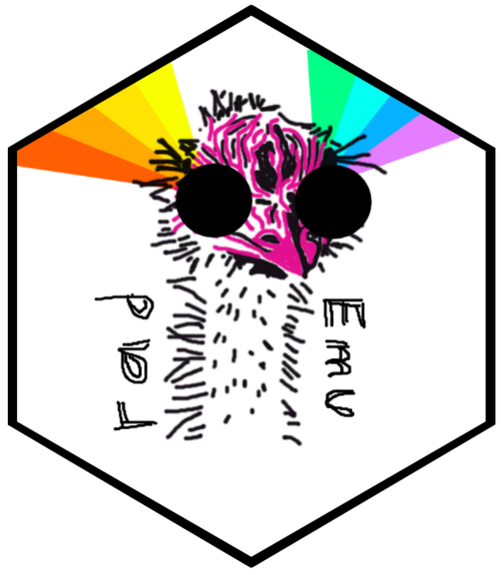
6/6

1 5

# 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

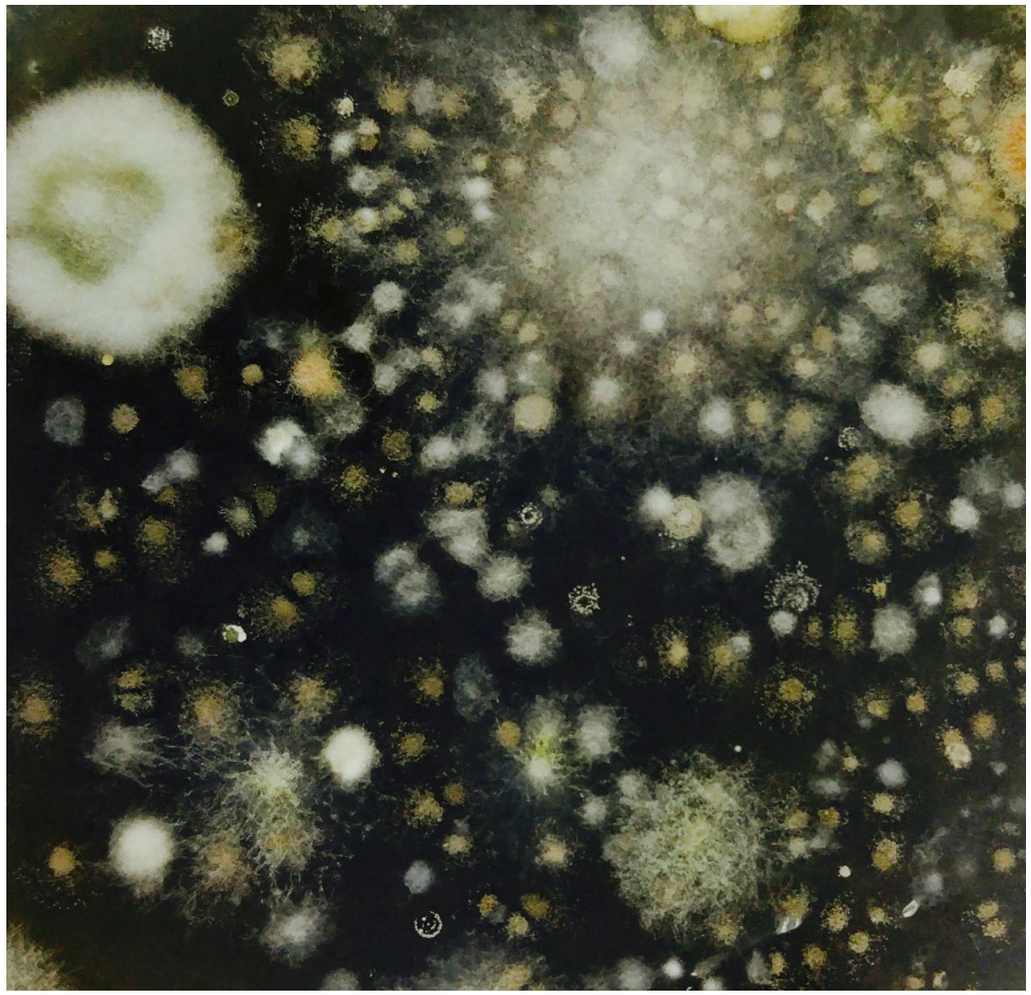
# Amy's wish list




- 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
Sarah V Teichman — Research Scientist
María Valdez — Postdoctoral Scholar
and
Sarah J Tucker — Postdoctoral Scholar (MBL)

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

# Supplementary slides

# 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

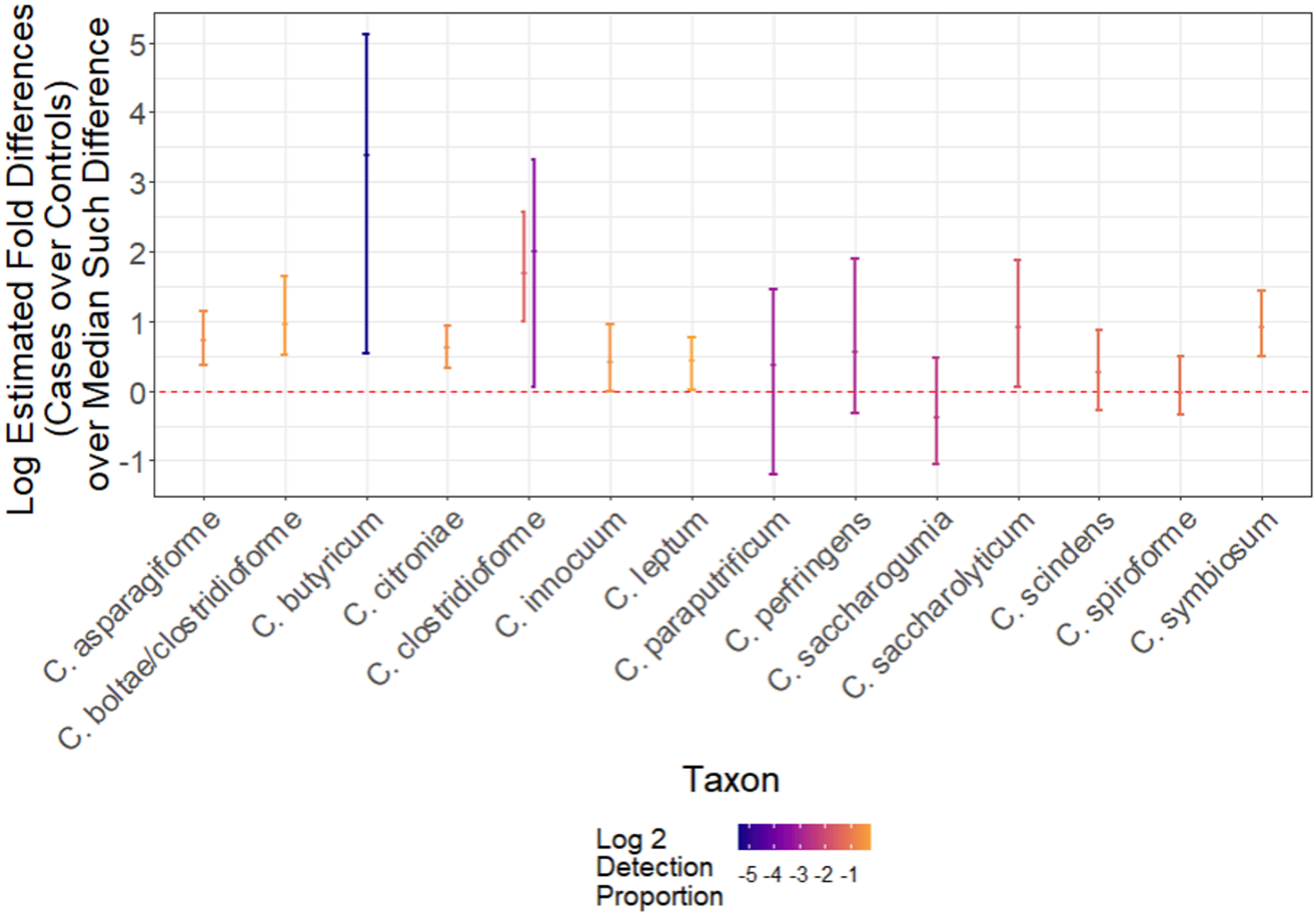
$$= \frac{\text{mean abs. abundance F. prauznitzii in cases of age } a \text{ and sex } s}{\text{mean abs. abundance F. prauznitzii in controls of age } a \text{ and sex } s}$$

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

# radEmu

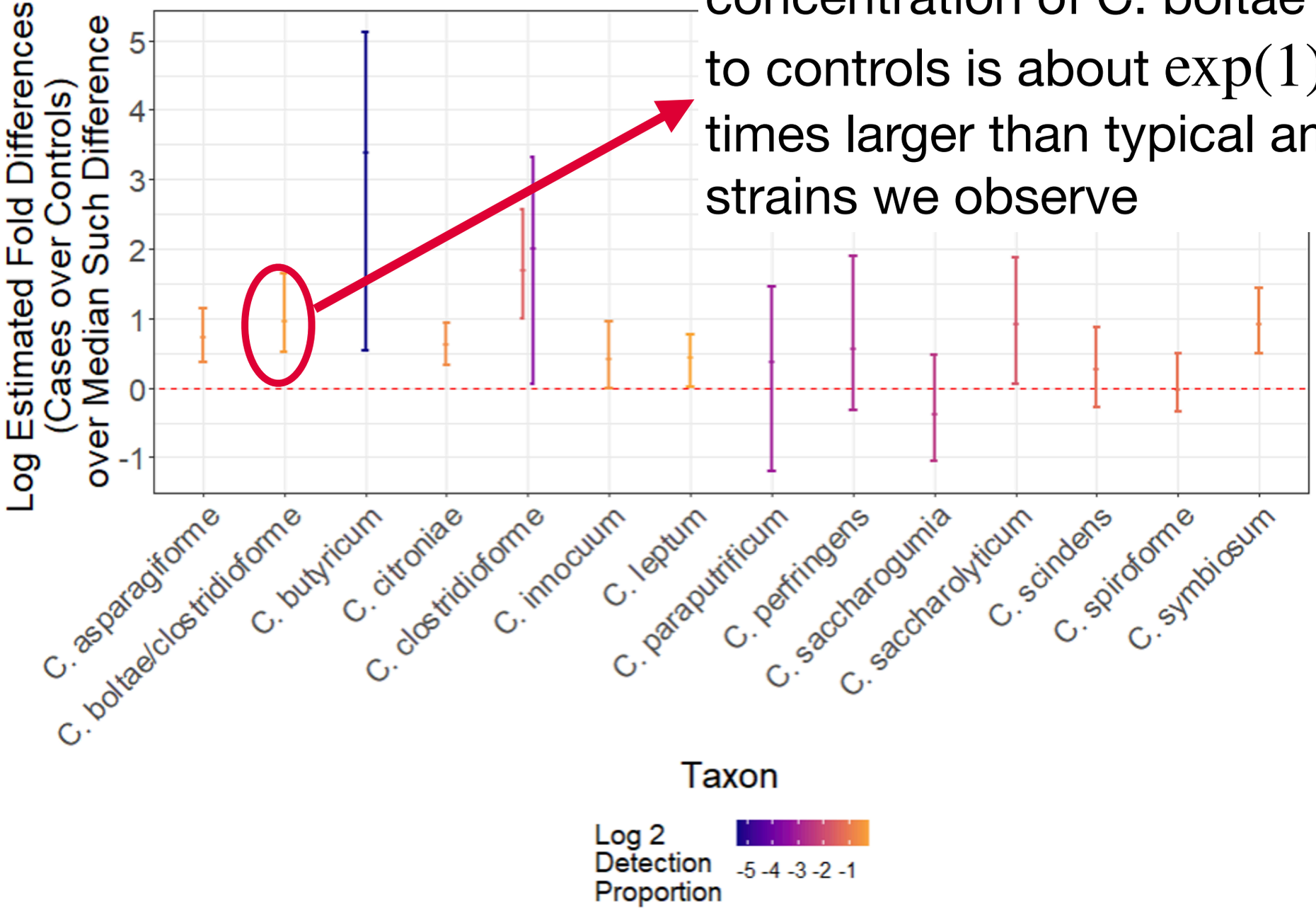
- Typical use case
  - Identify strains enriched/depleted in one type of samples compared to otherwise similar samples that differ in their type
- "Making reasonable comparisons"
- In practice: *Look at all taxa, identify the most enriched/depleted taxa*
  - Typical to look at FDR (q-values), not Type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



# 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
  - One that makes reasonable assumptions
- 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 are 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?"

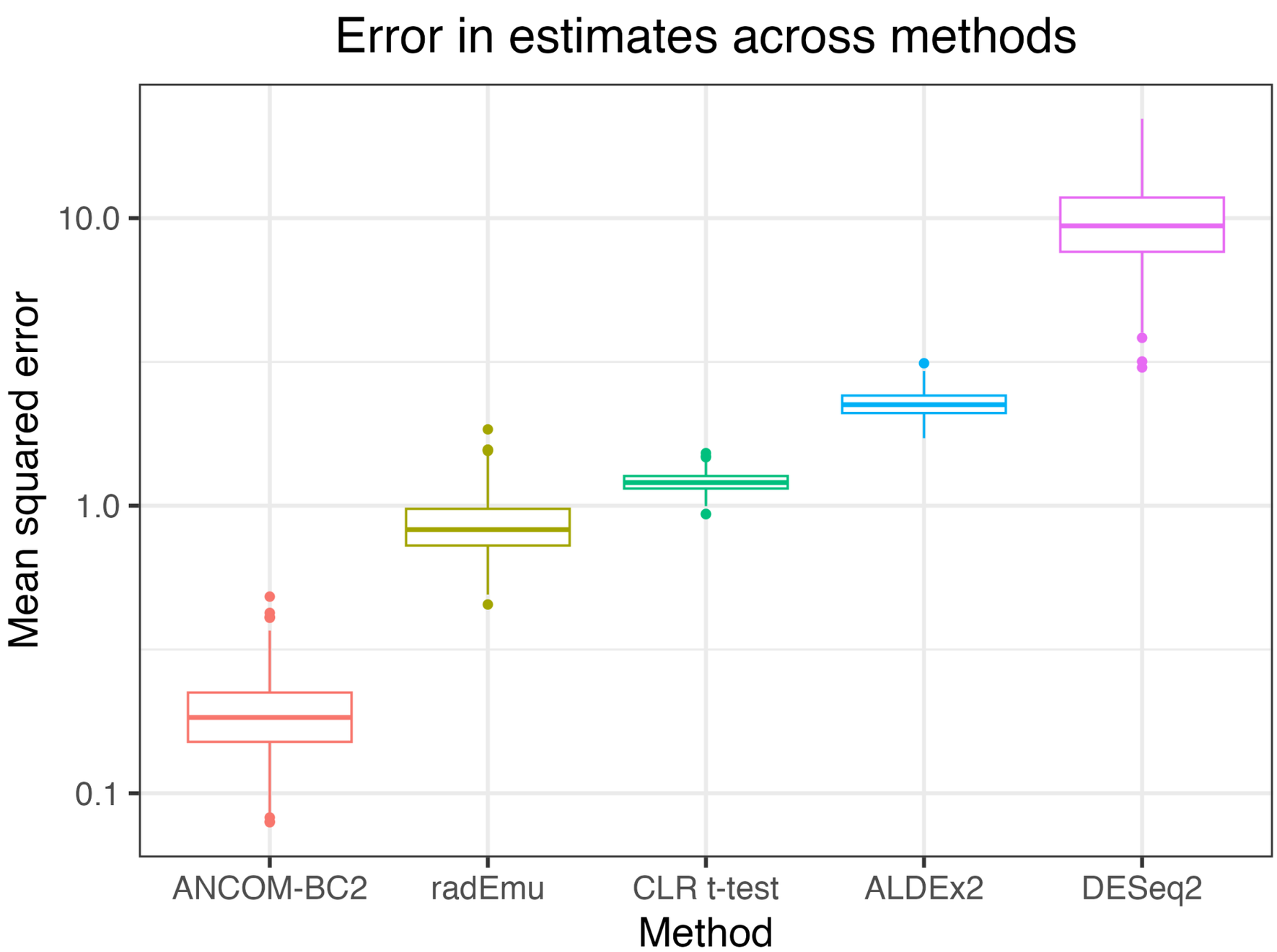
2. "Do we have error rate control?"

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

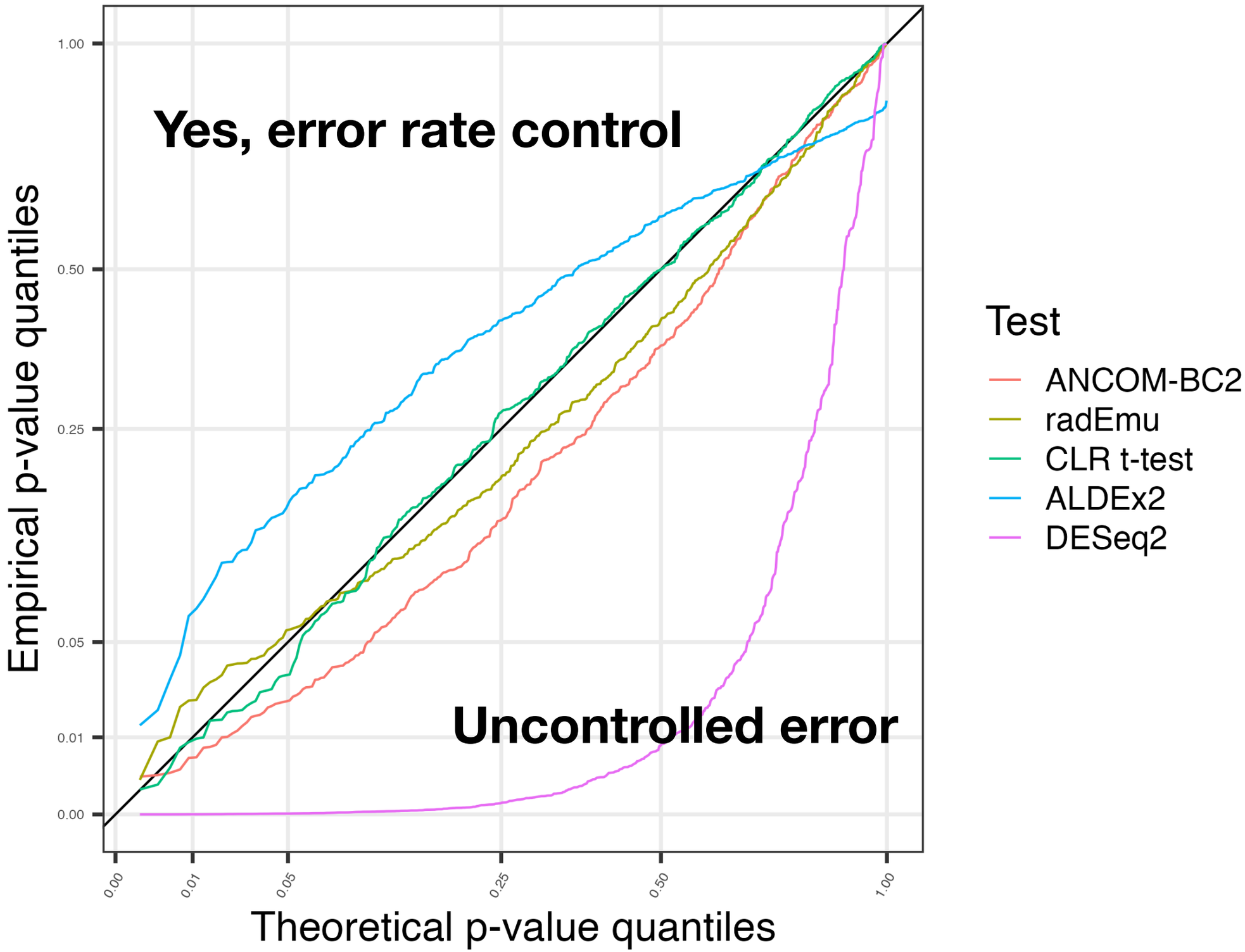


Special thanks to Sarah! 😻🍺

# “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 Type 1 error rate

# 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