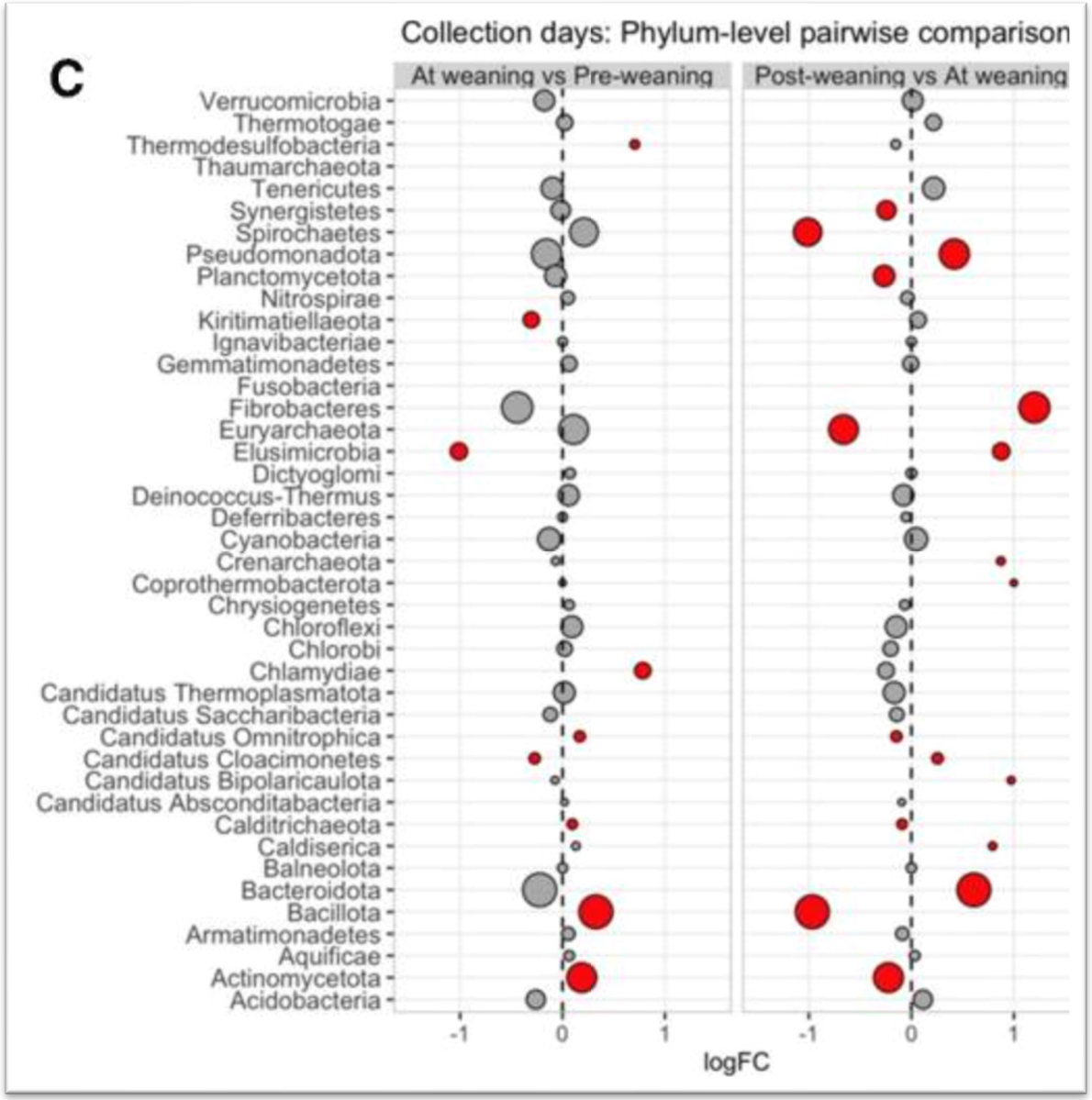
Differential abundance testing for bioinformatic data

What is differential abundance?



“All models are wrong, some are useful.”
–George Box

Learning objectives

- Describe how differential abundance (DA) modeling tests which taxa are significantly different between groups (e.g. treatment, host, etc)
- Be able to describe how count data is “**compositional**” and “**sparse**”, and why this influences our decision making
- Understand the balance between type 1 and type 2 error based on your model selection for DA
- Describe the 3 most common categories of statistical methods used

Outline

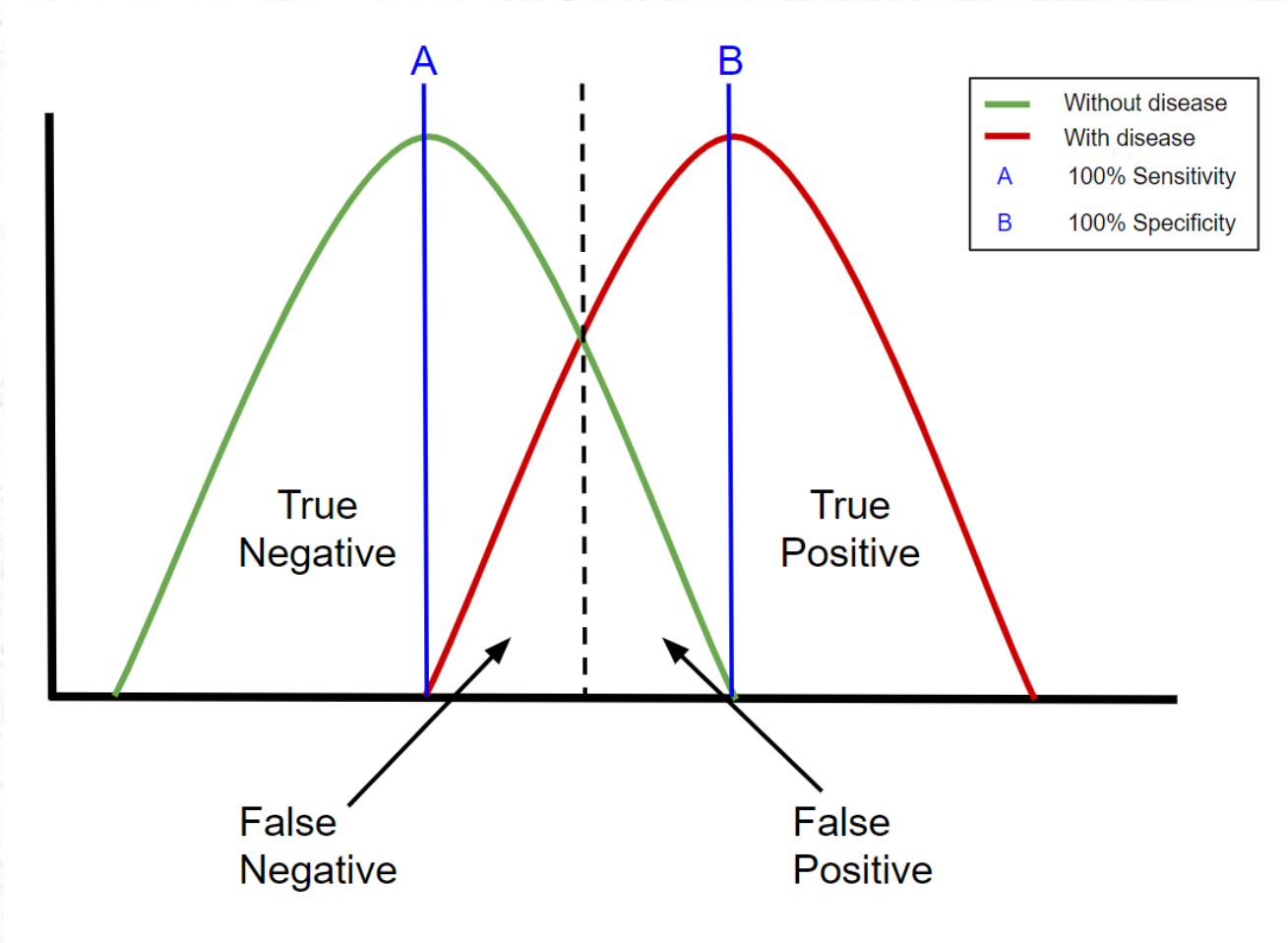
1. Review types of analyses you've learned so far
2. Review typical model selection
3. Sequencing data considerations
4. My recommendation : ANCOM-BC2
5. Hands-on activity in R

Starting from count data, we want to analyze:

Analysis type	Goal	Response variable	Example test
Alpha diversity	Compare richness and evenness	1 diversity index per sample	Wilcox
Beta diversity	Compare community structure	Differences between samples	PERMANOVA
Differential abundance testing	Compare differences in taxa abundance	Counts/abundance of taxa	ANCOMBC

What do we want from a statistical test?

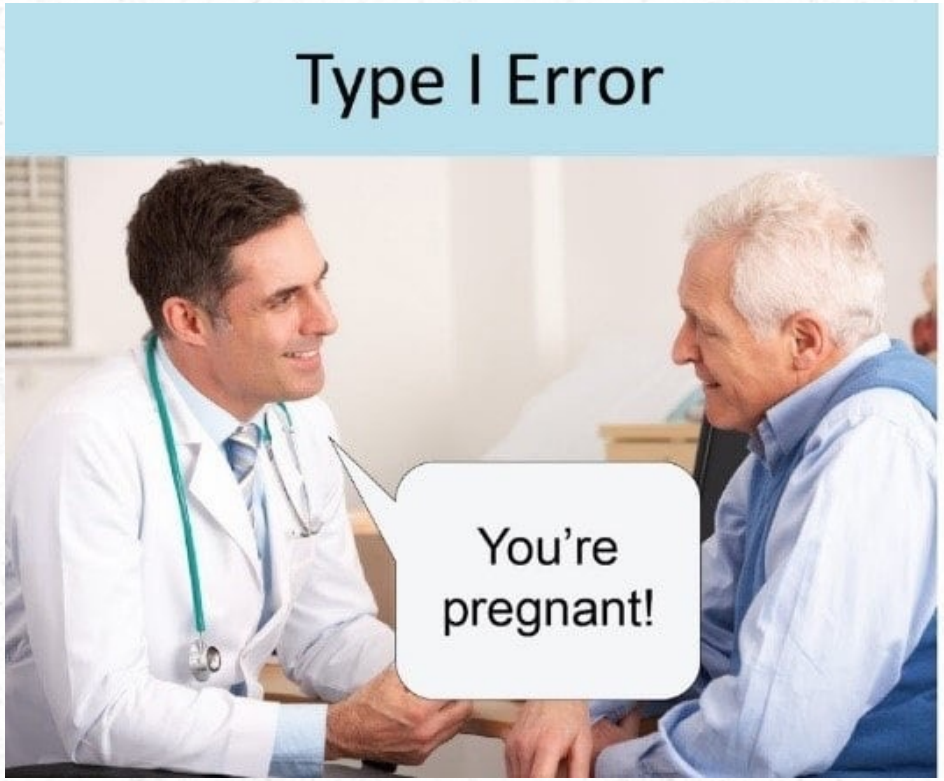
- Accounts for data structure
- Controls for multiple comparisons (100s of taxa)
- Must balance **Sensitivity** vs **Specificity**



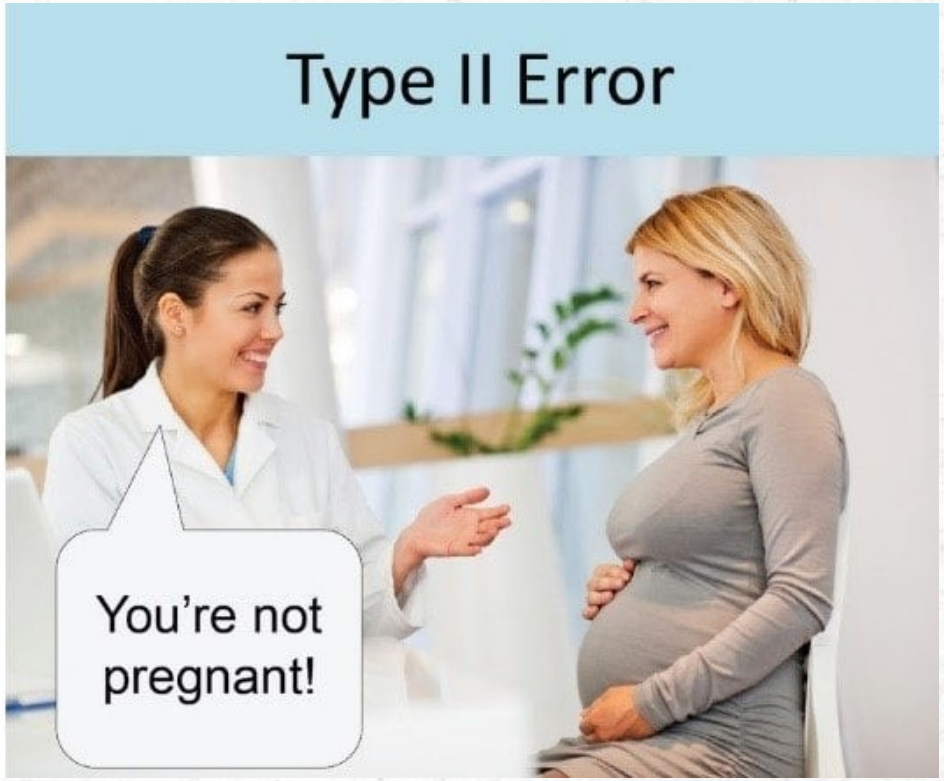
What do you prioritize?

Type 1 error: reporting **a difference**, when there isn't

Type 2 error: reporting **no difference**, when there is



"it's significant!"

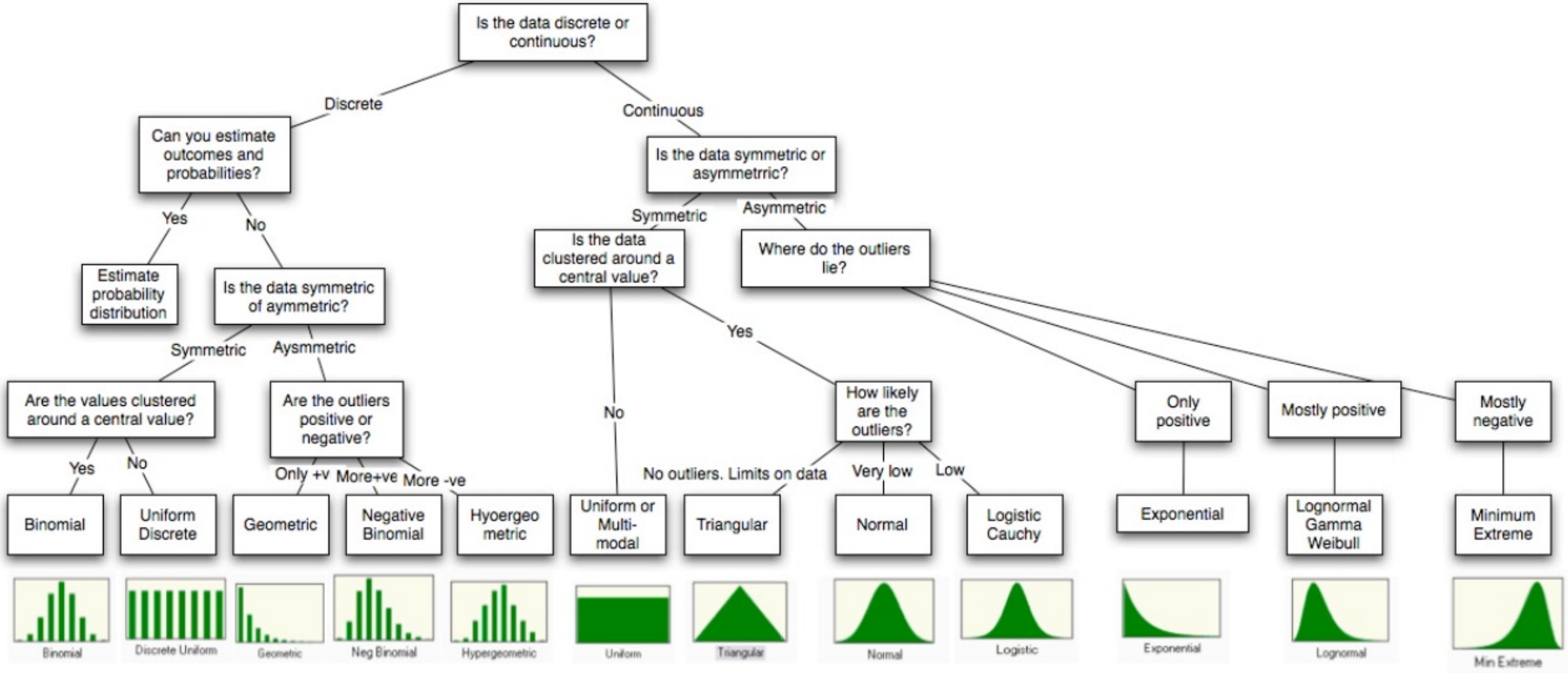


"no significant findings"

VERO
VETERINARY EDUCATION, RESEARCH, & OUTREACH

Image source: unbiasedresearch.blogspot.com

What model should we choose?



VERO
VETERINARY EDUCATION, RESEARCH, & OUTREACH

Image source: nyu.edu/adamodar

Extra considerations for sequencing data

- "Count data"
  - Count data is typically modeled with a Poisson distribution
  - Sequencing count data is different

    Microbiome Datasets Are Compositional: And This Is Not Optional

    1. **Zero-inflated/sparse**
       - Lots of features have 0's
    2. **Compositional** due to a finite amount of sequencing
       - Changes in one taxa, affect the others

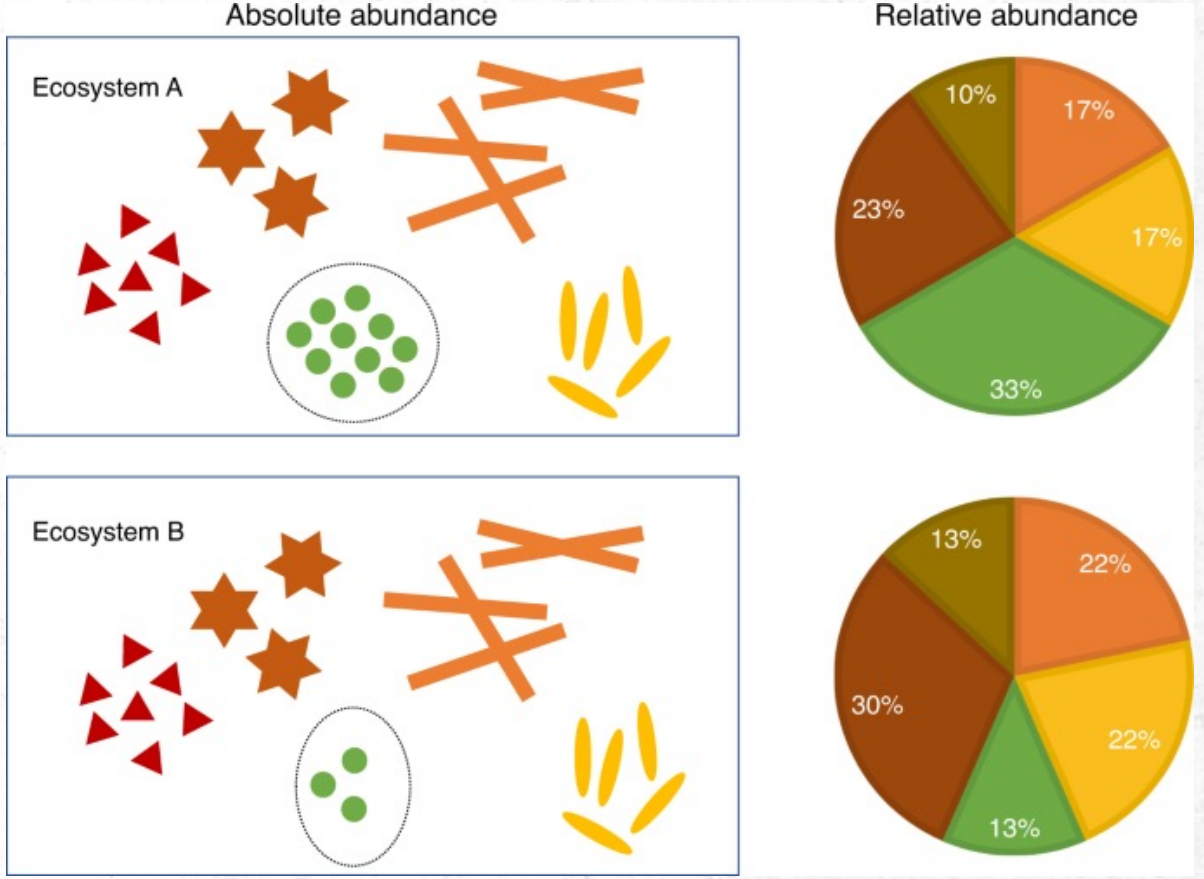


Image source: https://www.nature.com/articles/s41467-020-17041-7

What options do we have?

- Microbiome differential abundance methods produce different results across 38 datasets (Langille et. al. 2022)
- In my opinion, I prefer a conservative approach that minimizes Type 1 error and accounts for sequencing count data structure

- Current recommendation: Analysis of Compositions of Microbiomes with Bias Correction 2 (**ANCOM-BC2**)
  - Takes raw counts (not normalized)
  - Models the log ratios between features
  - Runs sensitivity analysis to reduce false positive results
  - Flexible model creation (repeated measures, random variables, interactions, etc)

ANCOMBC2

- Takes raw counts (not normalized)
- Models the log ratios between features (instead of counts)
- Flexible model creation (repeated measures, random variables, interactions, etc)
- Reduces Type1 error with sensitivity test and FDR
- Identifies “structural zeros” (only present in one group)

Hands-on activity for differential abundance

1. We'll use the R package "testDA" to run multiple types of differential abundance tests
2. Then, we'll run the ANCOM-BC2 model on it's own
3. Finally, we'll explore the effect of removing "sparse" features