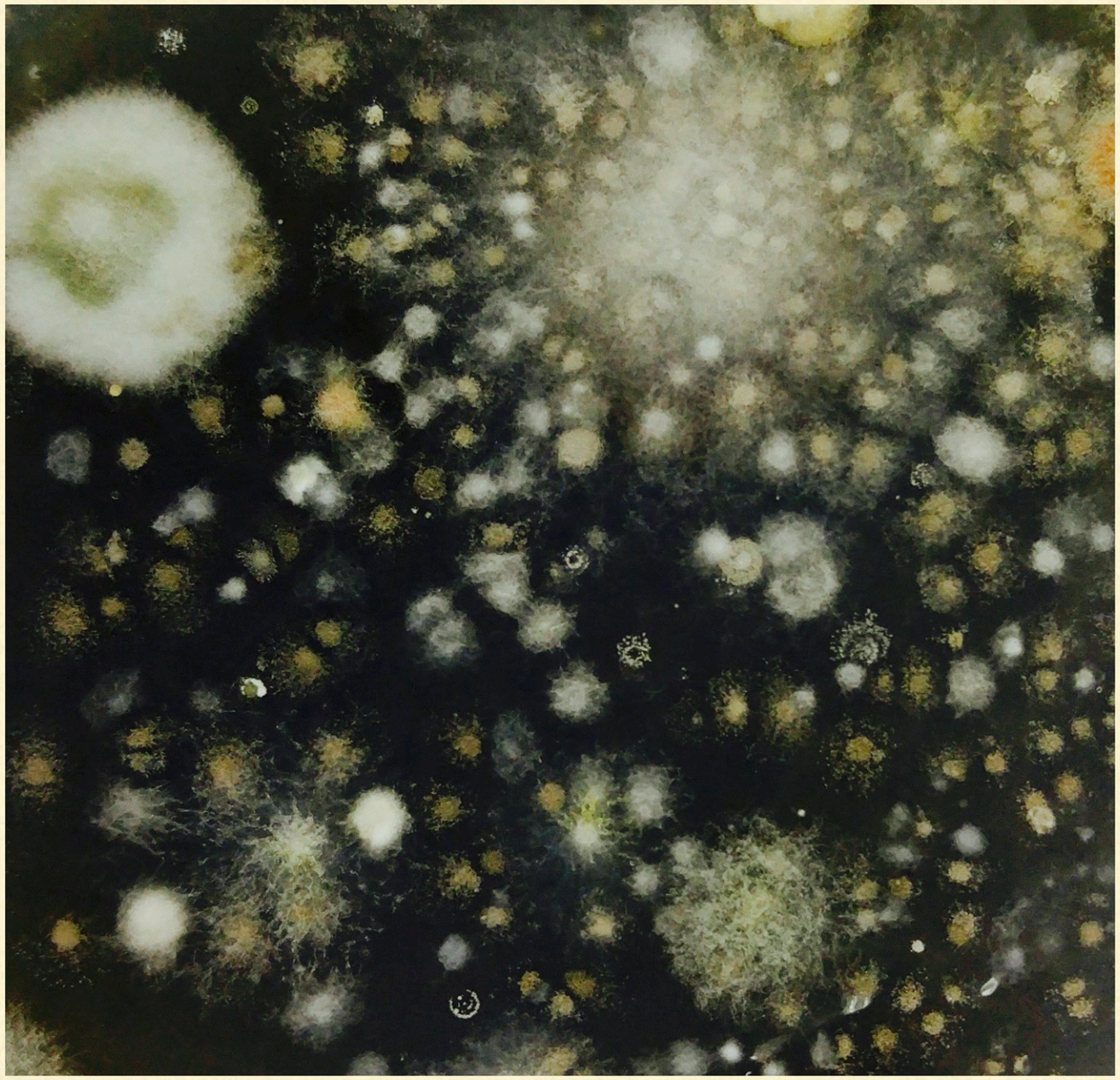


STATISTICAL ANALYSIS OF ALPHA DIVERSITY

Research Group: Statistical Diversity Lab (new!)

PI: Amy D Willis PhD, Assistant Professor, Department of Biostatistics, UW

@AmyDWillis adwillis@uw.edu

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

STATISTICIANS VS DOCTORS

- Problems we do have answers to:
  - Statisticians
    - Fitting correlative models between observed random variables
    - How to conservatively adjust for multiple testing
  - Doctors
    - Influenza vaccines
    - Why blood clots

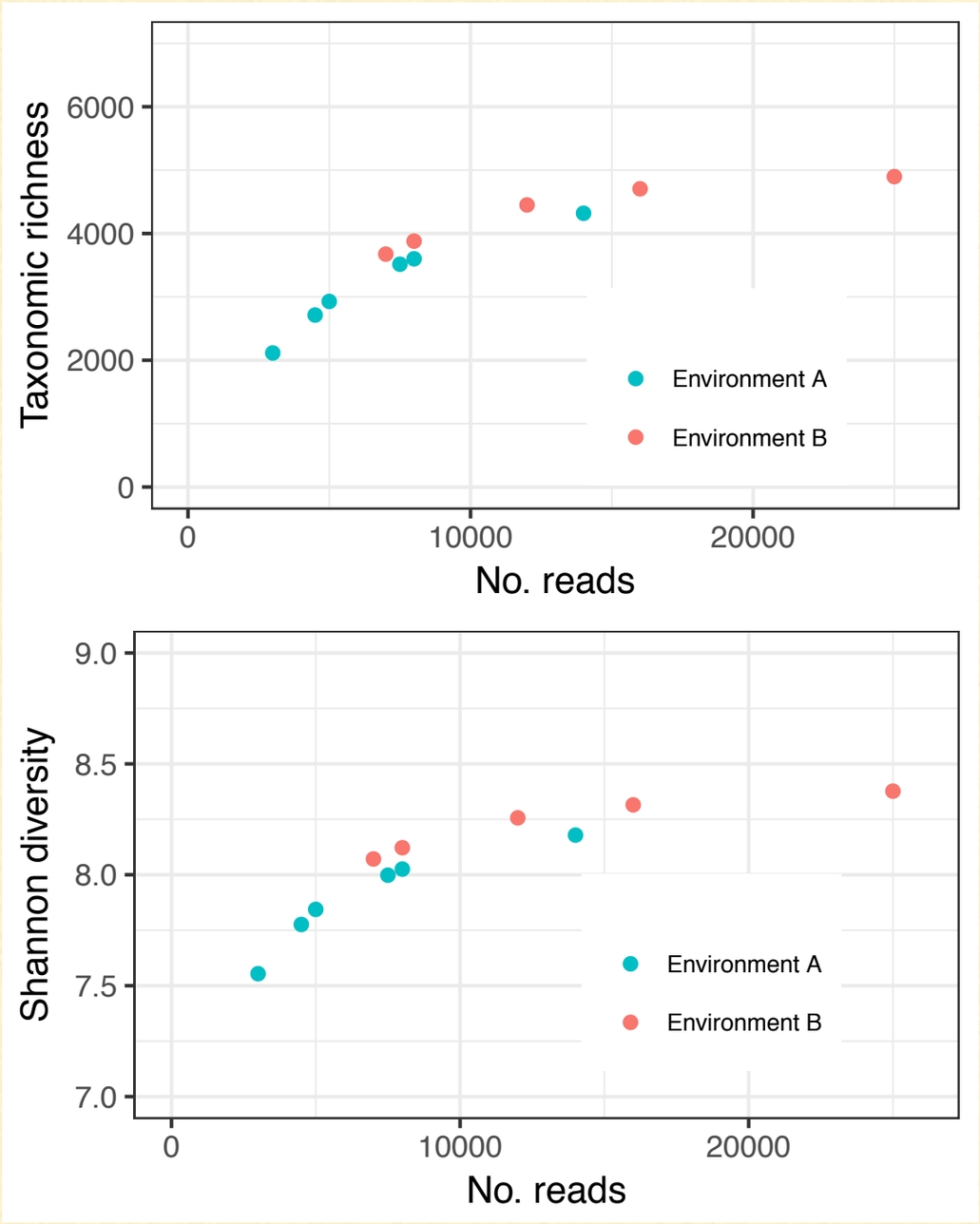
STATISTICIANS VS DOCTORS

- Problems we don't have answers to:
  - Statisticians
    - How to estimate alpha diversity in microbiome studies
    - How to incorporate microbial population structures into hypothesis testing
  - Doctors
    - How to cure cancer
    - How to develop an HIV vaccine

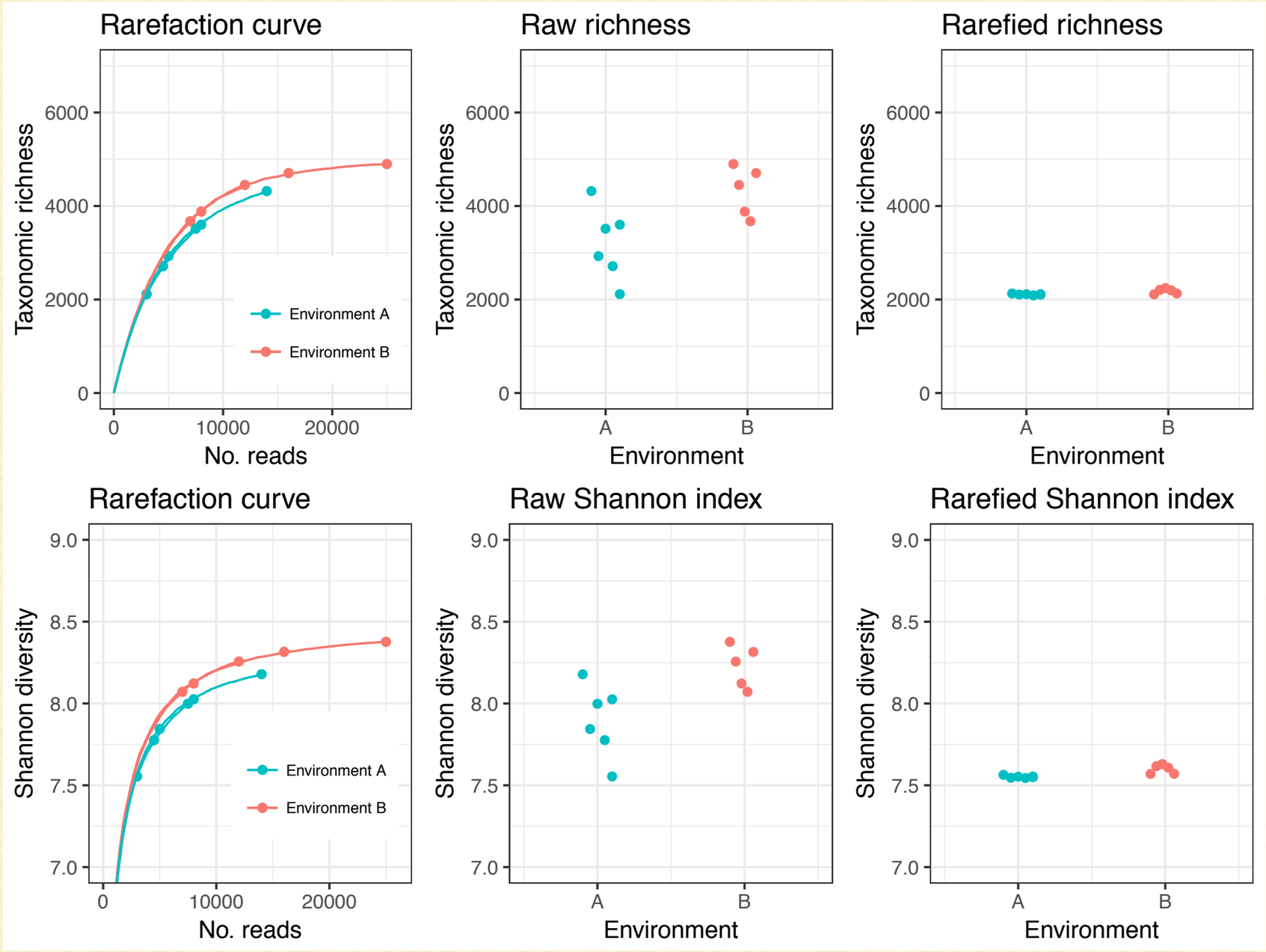
ALPHA DIVERSITY

- alpha diversity metrics ("indices") are low-dimensional summaries of microbial composition data
- A community of C microbes with abundances p_1, p_2, ..., p_C
- Taxonomic richness/total diversity: C
- Shannon diversity: $-\sum_{i=1}^{C} p_i \ln p_i$
- Simpson diversity: $\sum_{i=1}^{C} p_i^2$
- ...

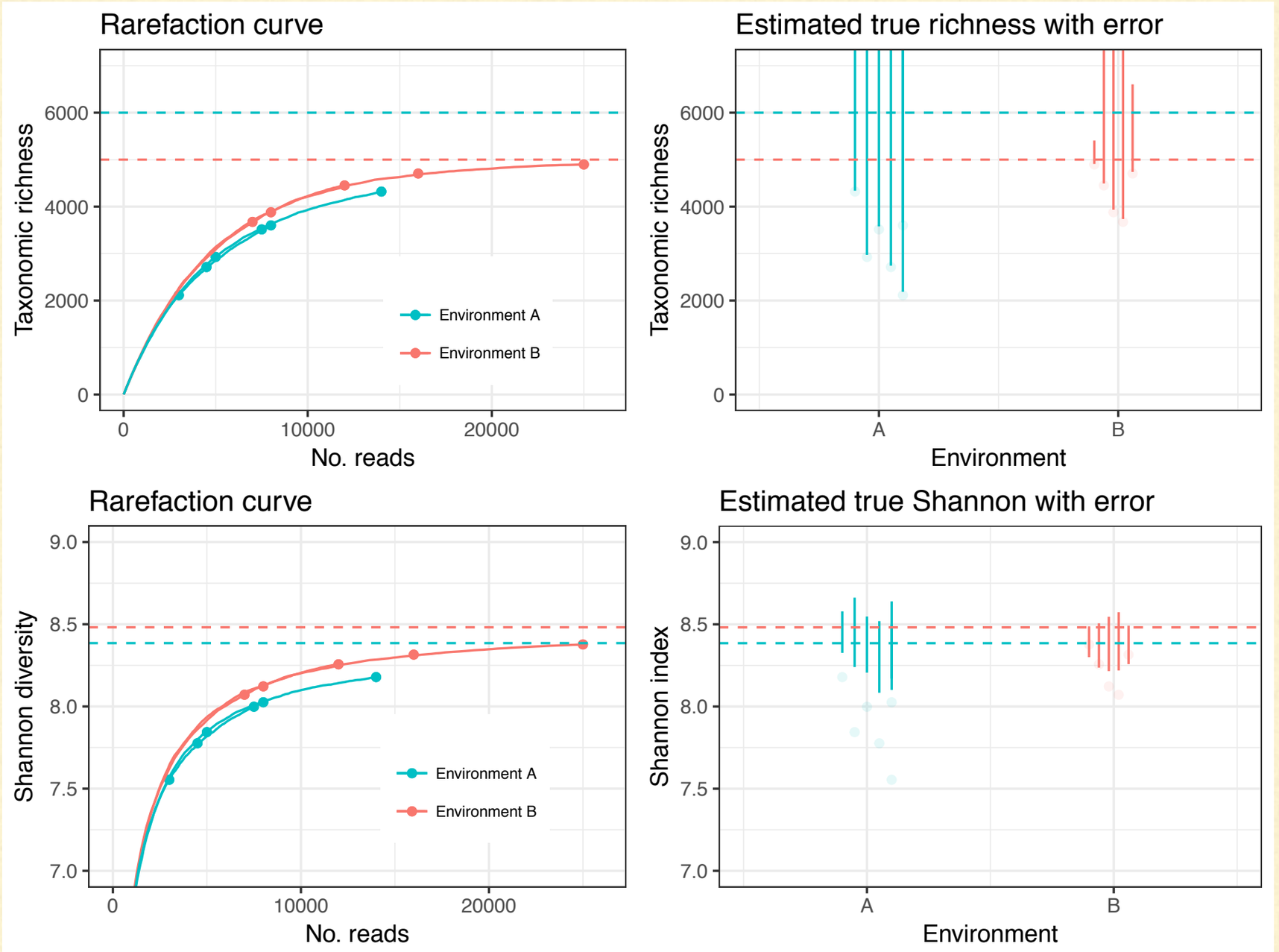
ALPHA DIVERSITY & LIBRARY SIZE



ALPHA DIVERSITY & LIBRARY SIZE



ALPHA DIVERSITY DOES NOT DEPEND ON LIBRARY SIZE!



STATISTICAL PARADIGM

- We did not exhaustively sequence the environment of interest
- alpha diversity is a property of the environment, not the sample
- We can use alpha diversity of the sample to estimate the alpha diversity of the environment
  - This approach implicitly adjusts for inexhaustive sampling and unobserved taxa

STATISTICAL LITERATURE

- Taxonomic richness: lots!
  - Chao1: valid only when all taxa are equally abundant
  - Chao-Bunge: simple negative binomial model
  - CatchAll: suite of mixed-Poisson models
  - breakaway: adapted for high diversity microbiomes
- Shannon diversity: none!
- Simpson diversity: none!
- Others: none!

NEW WORK

- Estimating alpha diversity metrics in a statistically rigorous way
- Adjusting for inexhaustive sampling
- Adding standard errors... on each sample's estimate

NEW WORK

- Setting: Not all taxa observed, have estimate of the number of unobserved taxa
- Assumptions:
  - Combined abundance of the observed taxa in the environment is close to the fraction of taxa observed
- Constraints:
  - Taxa observed with the same abundance have the same estimated abundance
  - For observed taxa, the estimated abundance is proportional to the observed abundance

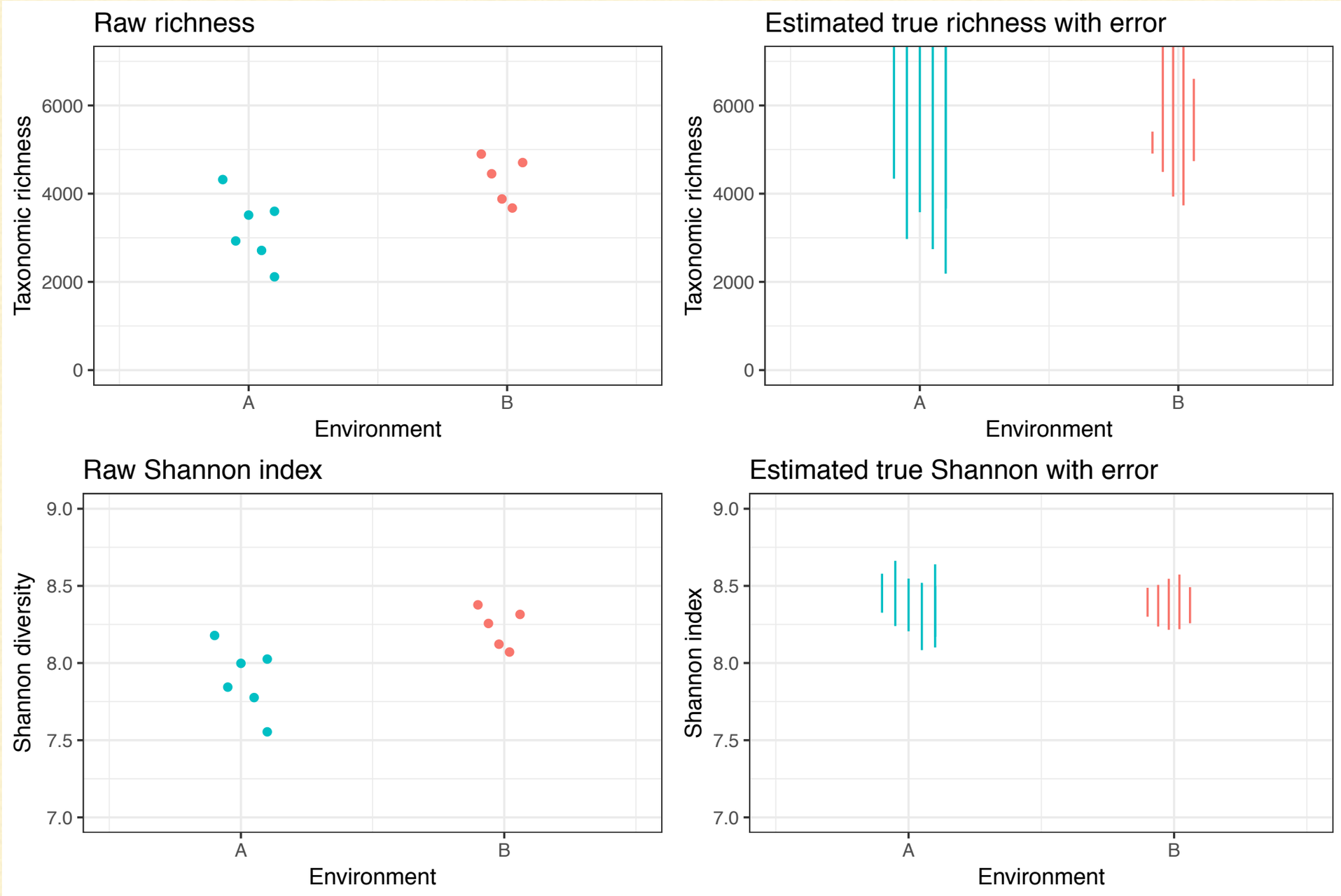
ADJUSTING FOR MISSING TAXA

- When not all taxa were observed, relative abundances are positively biased
- Observed taxa: lower relative abundance than observed
- Unobserved taxa: higher relative abundance than observed

BENEFITS

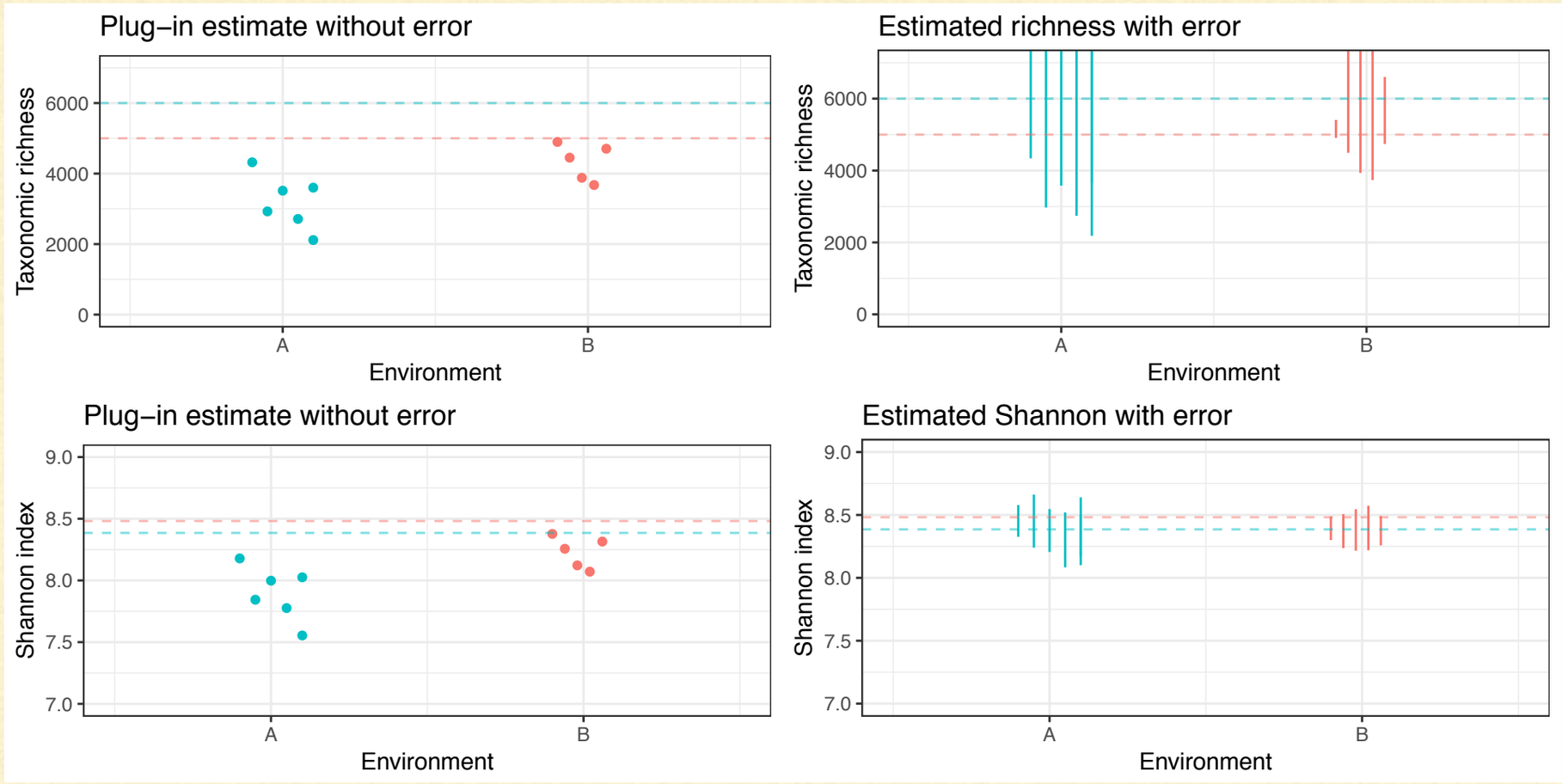
- First solution to open problem: alpha diversity with unobserved taxa (equivalently, *alpha diversity for the microbiome*)
- Applicable to any alpha diversity metric
  - Phylogeny-weighted coming soon!
- Weak, nonparametric assumptions
  - No assumptions of independence!
- Software, examples and vignettes available

WHAT DOES THIS LOOK LIKE?



MODELLING ALPHA DIVERSITY

- When do we model and plot points versus lines?



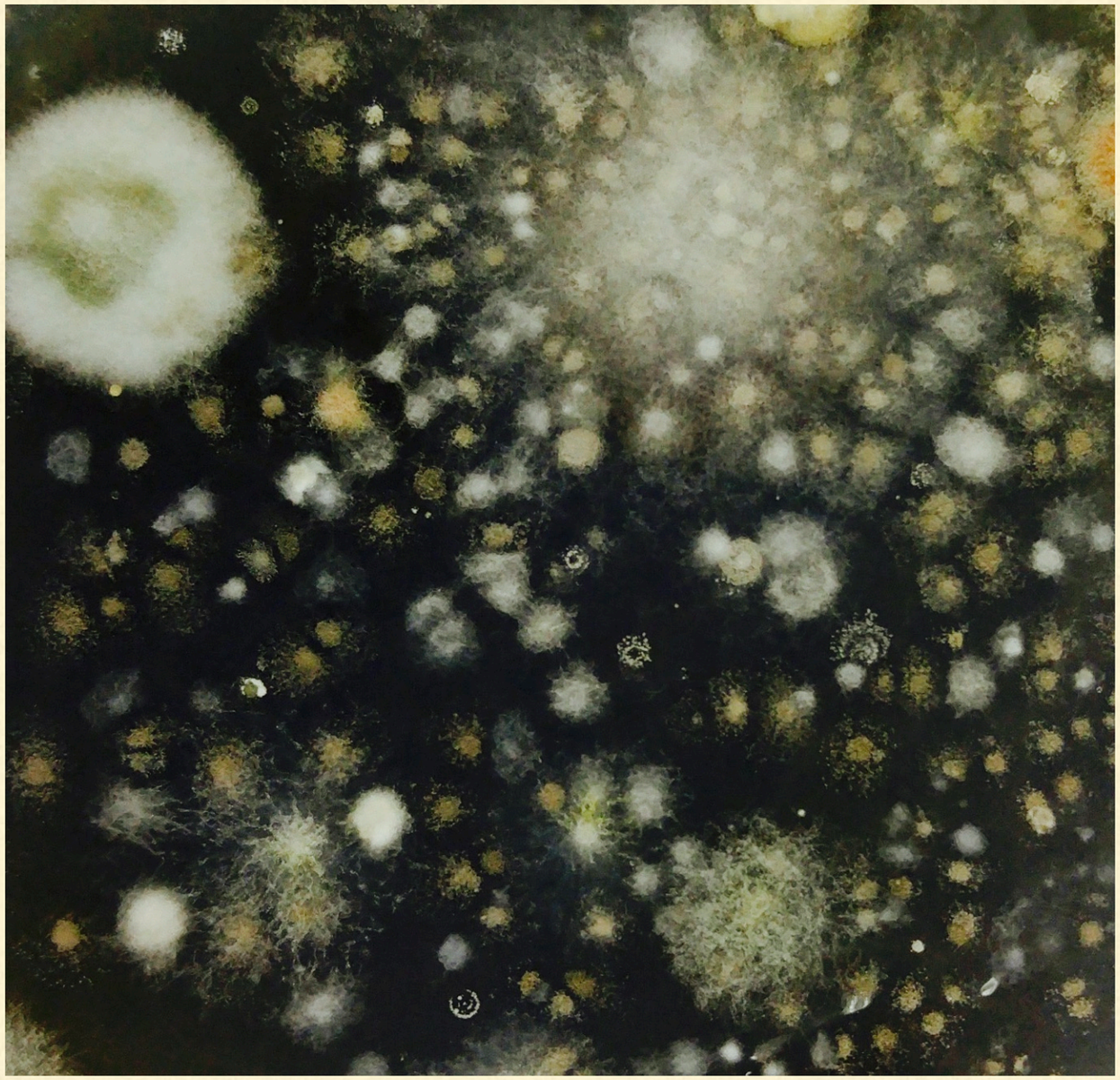
- betta( ) should be used instead of lm( ) or aov( ) for microbiome summary statistics!

SUMMARY

- You don't "calculate" alpha diversity, you "estimate" it
- This needs a correction for inexhaustive sampling
- Historically this has been done with rarefaction; by throwing away data
- This should be done with *statistics*
- Idea: estimate number of unobserved taxa, incorporate this into alpha diversity estimates

RESOURCES

- breakaway: alpha diversity, better
  - adw96.github.io/breakaway
- The new Statistical Diversity Lab @ UW
  - http://faculty.washington.edu/adwillis/
    - new site coming soon...
- CMiST!
- STAMPS: Strategies and Techniques for Analyzing Microbial Population Structures at the MBL (Marine Biological Laboratory)



STATISTICAL ANALYSIS OF ALPHA DIVERSITY

Research Group: Statistical Diversity Lab (new!)

PI: Amy D Willis PhD, Assistant Professor, Department of Biostatistics, UW

@AmyDWillis adwillis@uw.edu

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