PRINCIPLES & METHODS FOR REASONABLE MICROBIOME ANALYSIS

Amy D Willis PhD

PI: Statistical Diversity Lab

Assistant Professor

Department of Biostatistics, University of Washington, USA

@AmyDWillis @paulinetrinh @BryanDMartin_ @davidandacat @Jake_in_the_Lab

adwillis@uw.edu

📷 ✅

📪 ✅

🐦 ✅

Slides available at github.com/adw96/presentations

Photo credit: Seattle Times

"How do I rigorously analyse my data?"

–Everyone, all the time

Joke credit: 🌽 Bryan Martin 🌽

"It depends."

–Stat Div Lab, all the time

Joke credit: Not a joke

PRINCIPLES

STATISTICAL THINKING

- Statisticians have a formal framework for thinking about uncertainty and error

- Key idea: *It's ok to be wrong, but state your uncertainty*

CONNECTING SCIENCE & STATISTICS

- You decide which parameter that you care about
  - relative abundance, concentration, diversity, functional potential…
- You have imperfect, noisy data
- You have a model that connects your data to your parameter
- You estimate your parameter from your data
- You estimate how wrong you could be

CONNECTING SCIENCE & STATISTICS

- You decide which parameter that you care about
  - relative abundance, concentration, diversity, functional potential…
- You have imperfect, noisy data
- You have a model that connects your data to your parameter
- You estimate your parameter from your data
- You estimate how wrong you could be

THINGS NON-STATISTICIANS *SHOULD* THINK ABOUT

- What parameter to care about
- What is sensible data to collect
- What is the best available model/method to estimate your parameter from your data

You need an understanding of your data <u>and</u> some understanding of statistics to make these decisions

THINGS NON-STATISTICIANS *SHOULD* THINK ABOUT

- What parameter to care about
- What is sensible data to collect
- What is the best available model/method to estimate your parameter from your data

You need an understanding of your data <u>and</u> some understanding of statistics to make these decisions

METHODS

ANALYSIS

- The questions that you have affect how you do your analysis
- Do you care about…
  - broad scale community structure?
  - granular detail?
  - Both/not sure?

ANALYSIS

- The questions that you have affect how you do your analysis
- Do you care about…
  - broad scale community structure? diversity analyses
  - granular detail? abundance
  - Both/not sure?

ANALYSIS

- The questions that you have affect how you do your analysis
- Do you care about…
  - broad scale community structure? diversity analyses
  - granular detail? abundance
  - Both/not sure?

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

RELATIVE ABUNDANCE

- Model & method for analysing **relative abundances** with covariates (environmental factors)


Bryan Martin

- Relative abundance of *what?*
  - ASVs/OTUs/genes/transcripts/metabolites
- Two types of hypothesis testing
  - covariates affecting relative abundance & covariates affecting overdispersion


Daniela Witten

github.com/bryandmartin/CORNCOB; Martin, Witten & Willis (2019) *Ann Appl Stat* 🌽🌽

RELATIVE ABUNDANCE

- Latent variable model: Beta-Binomial + link
- Maximum likelihood via accelerated gradient descent
- Testing via parametric bootstrap likelihood ratio test
- Handles zeroes, overdispersion, unequal depths of sampling, multiple testing
  - No need to rarefy

github.com/bryandmartin/CORNCOB; Martin, Witten & Willis (2019) *Ann Appl Stat*

RICHNESS

- The "species problem": how many species were missing from the sample

- *Idea:*
  - If many rare species in sample, likely there are many missing species
  - If few rare species in sample, likely there are few missing species
  - Use data on # rare species to predict # missing species

RICHNESS

- breakaway: *an R package for estimating species richness*
- *How?*
  - Fits the most flexible & ~~most realistic~~ least unrealistic* class of models for this problem
    - Details: Willis & Bunge (2015), *Biometrics*
- *Bonus*: betta is a method for comparing richness estimates and their uncertainities
  - Details: Willis, Bunge & Whitman (2017), *JRSS-C*

* to my knowledge

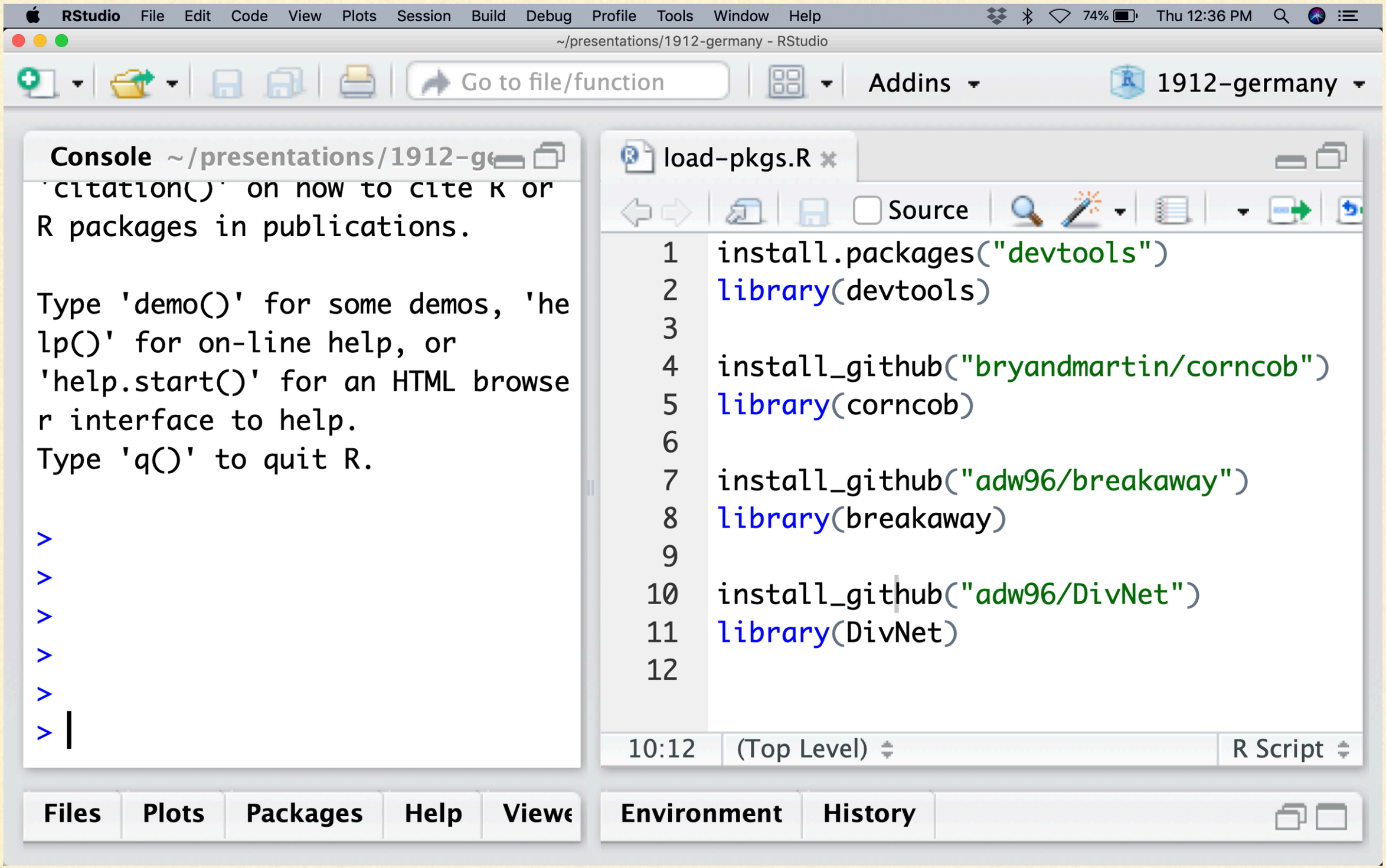
EVENNESS

- DivNet: *an R package for estimating & modelling "quantitative" alpha & beta diversity*

Bryan Martin

- Latent variable model: Multinomial + MVN + link
- EM algorithm for parameter estimation
- Deals with zeroes, overdispersion, unequal depths of sampling
  - No need to rarefy

Willis & Martin (2018+) *bioRxiv*



Example workflows are available for each package

CASE STUDY

- Canadian boreal forests contain ~10% of global terrestrial carbon stocks
- Microbes in soil fix carbon and mediate plant health
- Whitman et al (2019) investigated the response of soil microbial communities to the extreme wildfire season
  - 62 sites sampled

CASE STUDY

- Canadian boreal forests contain ~10% of global terrestrial carbon stocks

- Microbes in soil fix carbon and mediate plant health

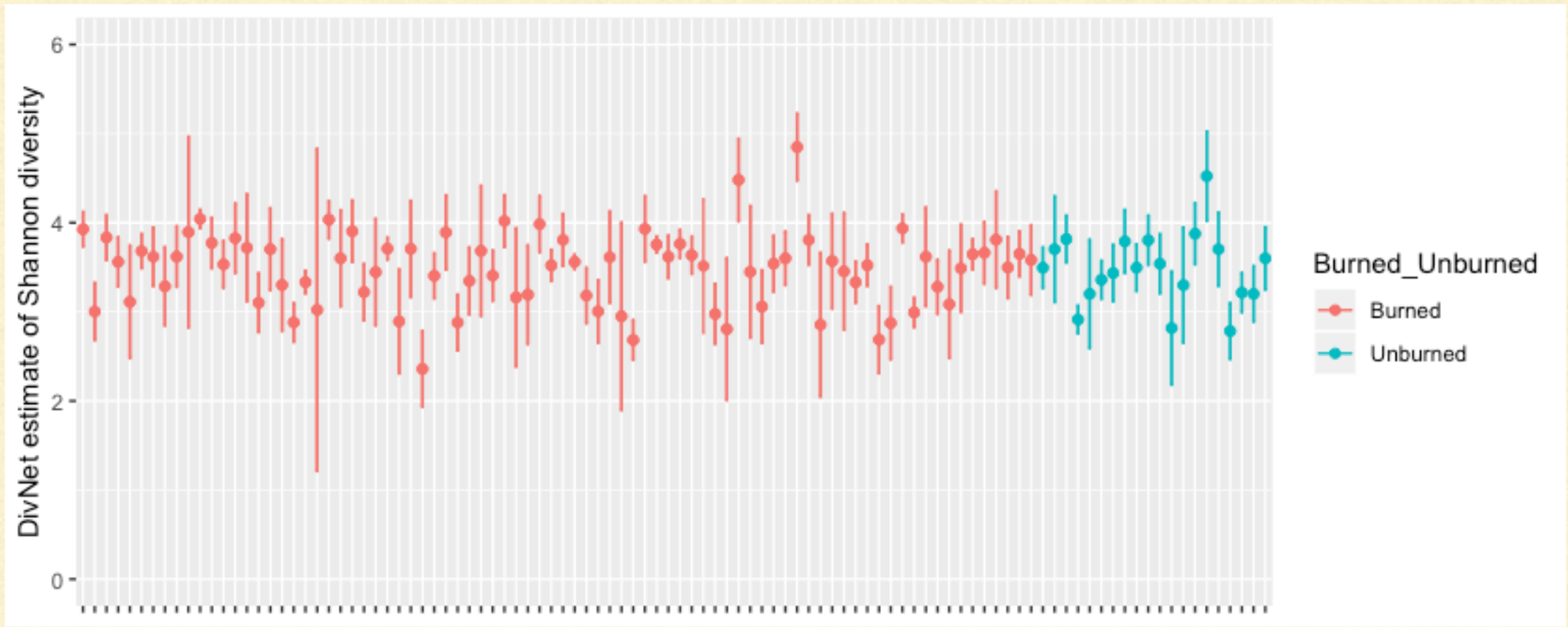
- Whitman et al (2019) investigated the response of soil microbial communities to the extreme wildfire season

  - 62 sites sampled

Whitman et al (2019) *Soil Biology & Biochemistry*

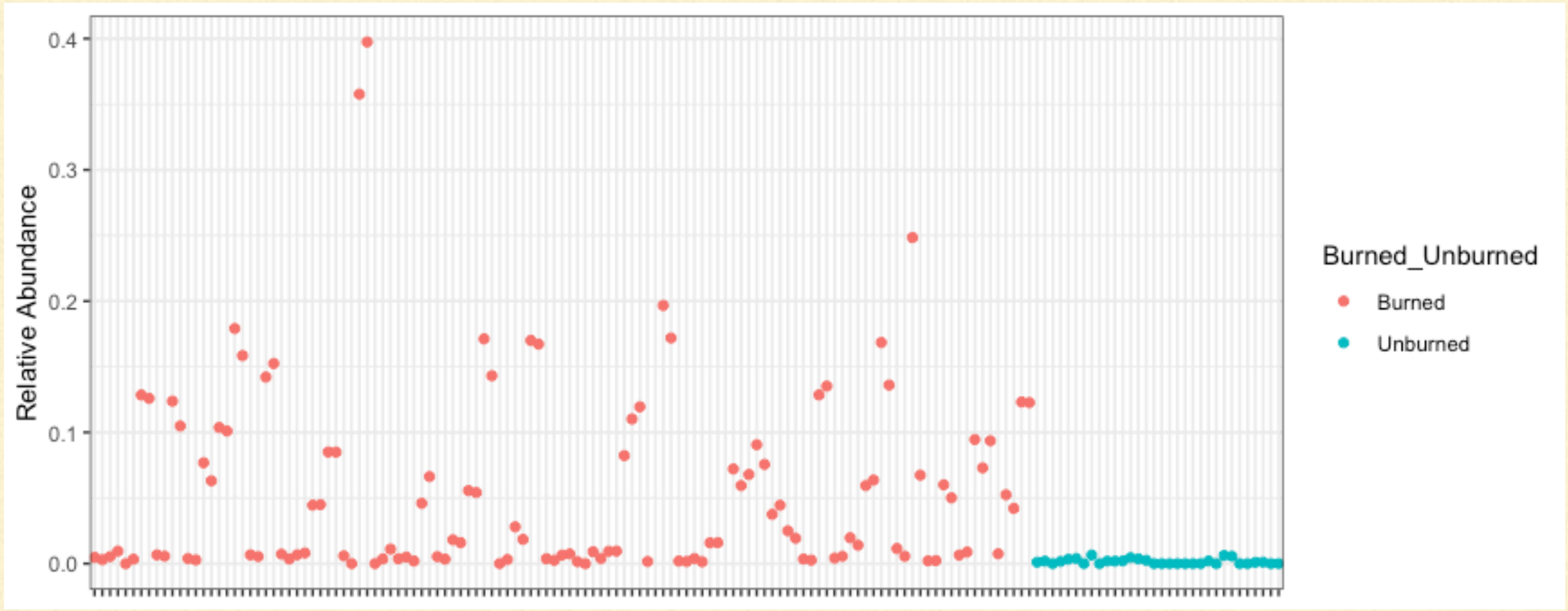
SHANNON DIVERSITY

- H_0: $Shannon_{unburned}$ = $Shannon_{burned}$ is *not* rejected: p = 0.46



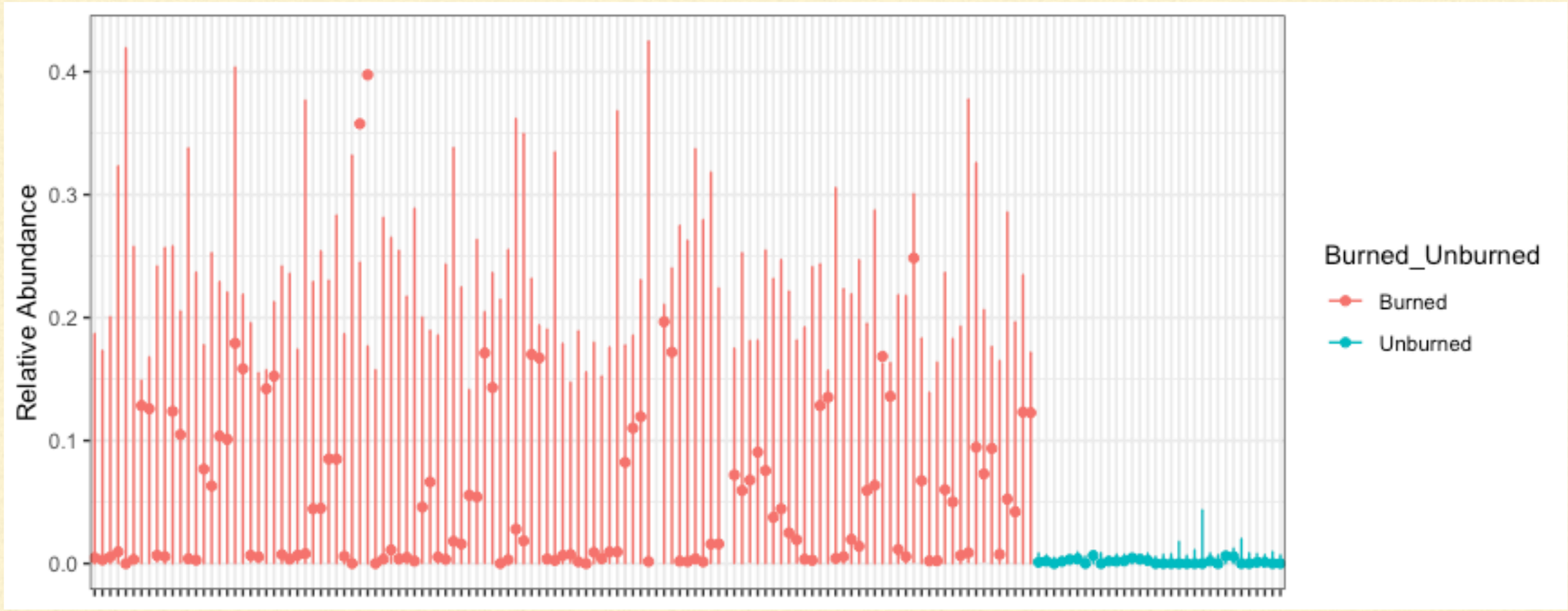
Data from Whitman et al (2019) *Soil Biology & Biochemistry*

GENUS MASSILIA



Data from Whitman et al (2019) *Soil Biology & Biochemistry*

GENUS MASSILIA



- Highly significant increase in relative abundance
- Highly significant increase in overdispersion

Data from Whitman et al (2019) *Soil Biology & Biochemistry*

RESOURCES

Content advisory/disclaimer: this sections contains opinions

WHERE TO START

- New to bioinformatics?
  - Happy Belly Bioinformatics

astrobiomike.github.io

- New to R?
  - My BIOST 509 course (UW): lectures & homeworks public

github.com/adw96/biost509

WHERE TO START

- New to statistics?
  - Be open to learning: read widely, think critically, take courses
  - Consider involving a statistician in your project *from the beginning*
  - Some review papers better than others; not all (any?) statisticians agree. 🙏?

WHERE TO START

- New to statistics?
  - Be open to learning: read widely, think critically, take courses
  - Consider involving a statistician in your project *from the beginning*
  - Some review papers better than others; not all (any?) statisticians agree. 🙏?

So many of you have excellent intuition for statistical problems — design, extrapolation, uncertainty… — refreshing to see!

WHERE TO START

- My STAMPS @ MBL lectures from 2019
  - Statistics Bootcamp: bridging STAT101 & microbiome science
  - Estimation: More details on breakaway, corncob, DivNet; tutorials & examples

github.com/statdivlab/stamps2019

Consider applying to attend STAMPS @ MBL 2020!

AMY'S FAVOURITE THINGS

- Get your hands dirty! Look at your data *closely*
  - R & tidyverse
  - DADA2 & exact sequence variants
  - anvi'o & manual genome curation/refinement

STATISTICAL & SCIENTIFIC THINKING

- Please maintain
  - an awareness of the extent to which your samples reflect your population
    - Tebbe: why do you think you can extrapolate from 1 cm^3 soil to all the soil in a field? Or globally?
  - an awareness of the limitations of your data
  - a kind but critical lens to your own & others' work ❤️

STATISTICAL & SCIENTIFIC THINKING

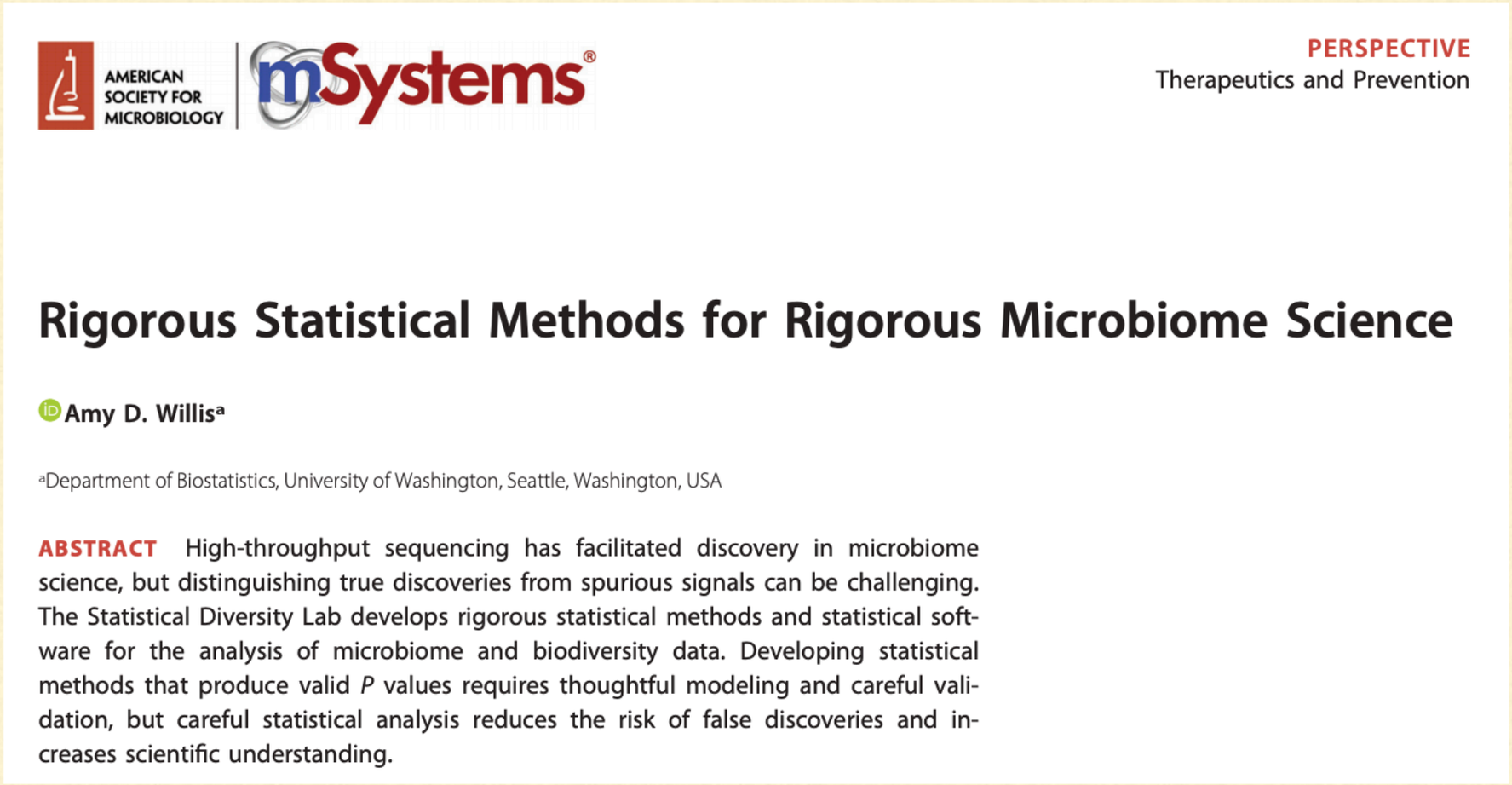
You need multiple approaches to validate your findings!

A single approach (16S/qPCR/shotgun/proteome/long read) is not sufficient to make conclusions!

WHO, AGAIN?

statistical diversity lab

- We are the statistical diversity lab, and we develop…



AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

PERSPECTIVE
Therapeutics and Prevention

Rigorous Statistical Methods for Rigorous Microbiome Science

Amy D. Willis[a]

[a]Department of Biostatistics, University of Washington, Seattle, Washington, USA

ABSTRACT High-throughput sequencing has facilitated discovery in microbiome science, but distinguishing true discoveries from spurious signals can be challenging. The Statistical Diversity Lab develops rigorous statistical methods and statistical software for the analysis of microbiome and biodiversity data. Developing statistical methods that produce valid *P* values requires thoughtful modeling and careful validation, but careful statistical analysis reduces the risk of false discoveries and increases scientific understanding.

statisticaldiversitylab.com

@AmyDWillis adwillis@uw.edu

WHO, AGAIN?

statistical diversity lab

We work on what we believe to be the most critical methodological needs in microbial science and the most series shortcomings of existing analysis methods. Along with our research, we see outreach, education, and collaboration as a core part of this mission.

statisticaldiversitylab.com

@AmyDWillis adwillis@uw.edu

PRINCIPLES & METHODS FOR REASONABLE MICROBIOME ANALYSIS

Amy D Willis PhD

PI: Statistical Diversity Lab

Assistant Professor

Department of Biostatistics, University of Washington, USA

@AmyDWillis @paulinetrinh @BryanDMartin_ @davidandacat @Jake_in_the_Lab

adwillis@uw.edu

📷 ✅

📫 ✅

🐦 ✅

Slides available at github.com/adw96/presentations

frontiers in Microbiology

PERSPECTIVE
published: 23 October 2019
doi: 10.3389/fmicb.2019.02407

Check for updates

Rarefaction, Alpha Diversity, and Statistics

Amy D. Willis*

Department of Biostatistics, University of Washington, Seattle, WA, United States

Understanding the drivers of diversity is a fundamental question in ecology. Extensive literature discusses different methods for describing diversity and documenting its effects on ecosystem health and function. However, it is widely believed that diversity depends on the intensity of sampling. I discuss a statistical perspective on diversity, framing the diversity of an environment as an unknown parameter, and discussing the bias and variance of plug-in and rarefied estimates. I describe the state of the statistical literature for addressing these problems, focusing on the analysis of microbial diversity. I argue

CONNECTING SCIENCE & STATISTICS

- You decide which parameter that you care about
  - relative abundance, concentration, diversity, functional potential…
- You have imperfect, noisy data
- You have a model that connects your data to your parameter
- You estimate your parameter from your data
- You estimate how wrong you could be

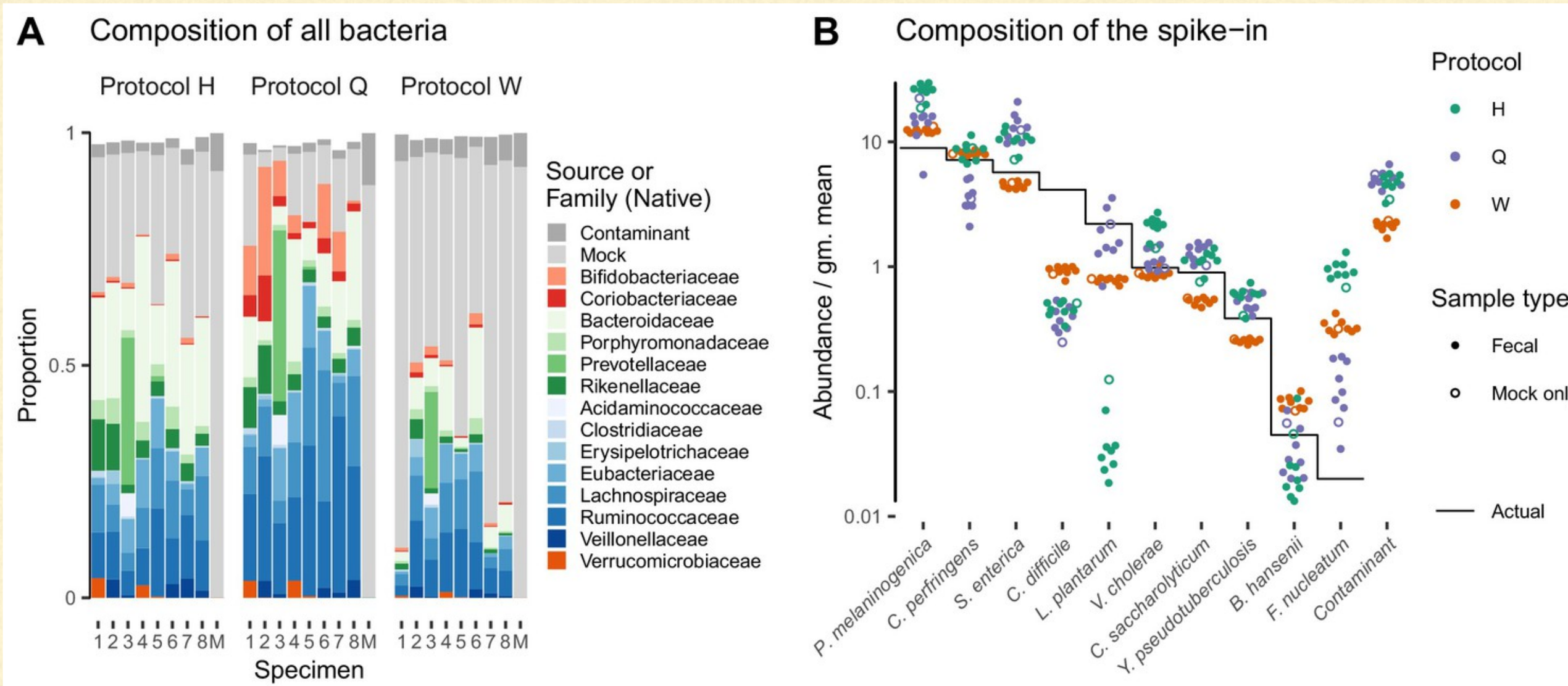


Figure from McLaren et al., 2019, *eLife*

EVENNESS

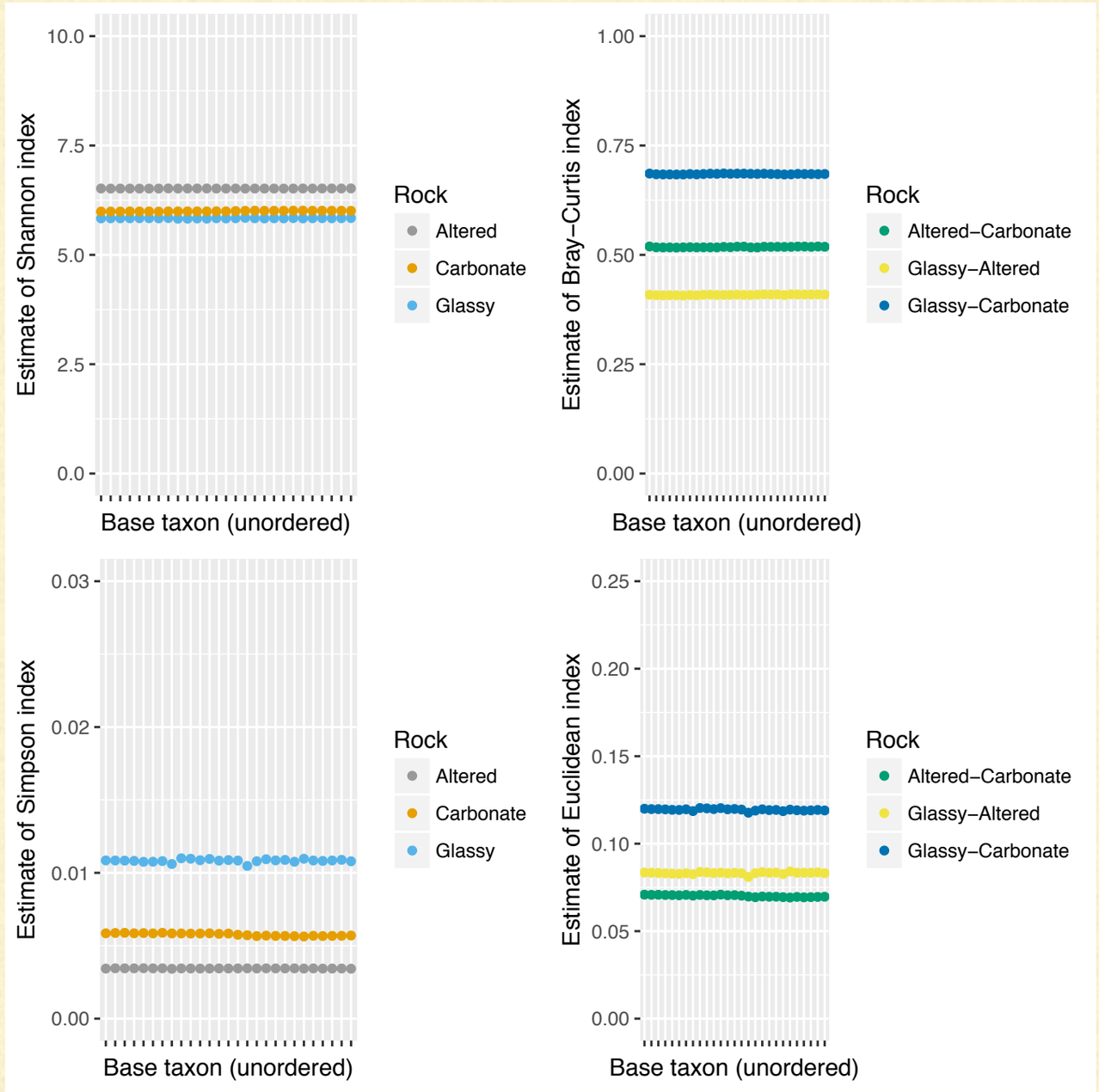
$$(W_{i1}, \ldots, W_{iq}) \sim Multinomial(M_i, (Z_{i1}, \ldots, Z_{iq}))$$

$$\left(\log\left(\frac{Z_{i1}}{Z_{iD}}\right), \ldots, \log\left(\frac{Z_{iq}}{Z_{iD}}\right)\right) \sim \mathcal{N}_{q-1}(X_i\beta, \Sigma)$$

- Fit model to obtain $\left(\hat{\beta}, \hat{\Sigma}\right)$
- Calculate fitted proportions $\hat{Z}_{ik} \propto e^{X_i^T \hat{\beta}_k}$
- Estimate e.g. Shannon diversity index by $\hat{\alpha}_{i,Shannon} = -\sum_{k=1}^{q} \hat{Z}_{ik} \log \hat{Z}_{ik}$
- Estimate uncertainty in $\hat{\alpha}_i$ using non/parametric bootstrap

Willis & Martin (2018+) *bioRxiv*

DIVNET:VARYING DENOMINATOR TAXON



RELATIVE ABUNDANCE

- For a specific strain, let Z_i be the relative abundance of a strain/gene/metabolite in sample i

- 🌽corncob🌽 fits the latent variable model

$$W_i|Z_i, M_i \sim \text{Binomial}(M_i, Z_i),$$
$$Z_i \sim \text{Beta}(a_{1,i}, a_{2,i})$$

and links covariates to a_{1i} and a_{2i}

github.com/bryandmartin/CORNCOB; Martin, Witten & Willis (2019) *Ann Appl Stat* 🌽🌽

RELATIVE ABUNDANCE

- Parametrisation:

$$\mu_i = \frac{a_{1,i}}{a_{1,i} + a_{2,i}},$$ expected relative abundance

$$\phi_i = \frac{1}{a_{1,i} + a_{2,i} + 1}$$ overdispersion/correlation

- Link parameters to covariates:
  - $g(\mu_i) = X_i^T \beta$ and $h(\phi_i) = X_i^{*T} \beta^*$
  - General link g(·) and h(·) permitted; log link by default