MB590-012
Microbiome Analysis
Community Alpha Diversity

Christine Hawkes

NC STATE UNIVERSITY

Wk	Date	Topic
1	12-Jan	Introduction – R, GitHub, Rmarkdown
2	19-Jan	Sequence prep, 16S ASV pipeline
3	26-Jan	Identification, normalization
4	2-Feb	Practicum – ITS ASV pipeline
5	9-Feb	Exploratory analysis 1 – alpha diversity
6	16-Feb	Exploratory analysis 2 – beta diversity
7	23-Feb	Dada2 on the HPC
8	2-Mar	Exploratory analysis 3 – core microbiomes
9	9-Mar	Practicum – full exploratory analysis
10	16-Mar	Spring break – no class
11	23-Mar	Hypothesis testing 1 – regression
12	30-Mar	Hypothesis testing 2 – permutation tests
13	6-Apr	Hypothesis testing 3 – TBD
14	13-Apr	Practicum – full hypothesis testing
15	20-Apr	Final project presentations

Bioinformatics complete!
Except HPC use on 2/23

Starting today: exploratory data analysis

Today's outline

- Alpha diversity
  - Definition
  - Indices
  - Rank abundance curves
- Phylogenetic alpha diversity

What is alpha diversity?

What is alpha diversity?

- Observed richness = number of ASVs detected

What is alpha diversity?

- Observed richness = number of ASVs detected

- Weighted richness
  - Abundance (evenness) → Shannon-Weaver, Simpson's
  - Phylogeny (branch lengths) → Faith's

Alpha diversity: Shannon-Weaver index (H')

- Captures both richness and evenness

$$H' = -\sum_{i=1}^{S} p_i \ln(p_i)$$

p_i = proportional abundance of the i^{th} ASV (abundance of ASV i/total abundance of all ASVs)
S = total number of ASVs

- Problems
  - Affected by sample size → limits comparability across studies
  - No intuitive interpretation of values
  - Can't distinguish if differences in H' are due to richness, abundance, or sampling effort

Alpha diversity: Simpson's index (D)

- Commonly used alternative to H'
- Probability that any two individuals drawn at random from a population belong to the same species:

$$D = [\sum_{i=1}^{S} n_i(n_i-1)]/N(N-1)$$

D = 1 = 100% probability of drawing the same species (num & denom are same); low diversity

D = 0 = 100% probability of drawing different species, i.e., each species has only one individual; high diversity

where
n_i = # of individuals of the i^{th} species
N = total # of individuals for all species in the sample
S = total number of species

- Problem of still confounding richness and evenness

Alpha diversity: Pielou's index (J)

A separate index of species evenness in a community is provided by:

$$J = H'/\ln(S)$$

J ranges from 0 to 1, where 1 is completely even

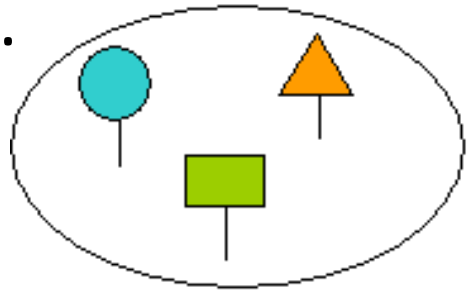
where

H' is the Shannon-Weaver diversity index

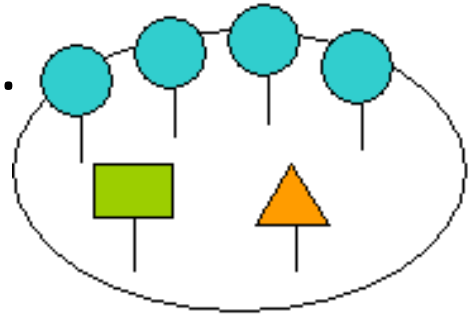
S is the number of species

Comparing alpha diversity metrics

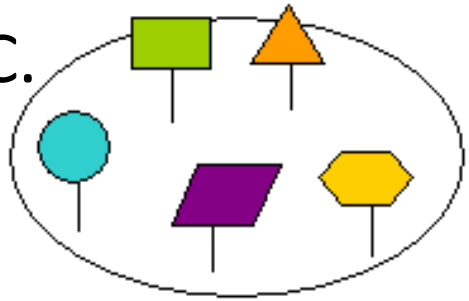
A.



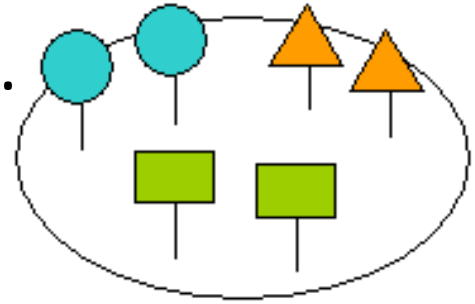
B.



C.



D.



Comparing alpha diversity metrics

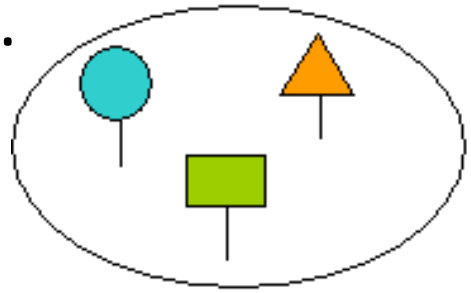
Spp = 3

H’ = 1.10

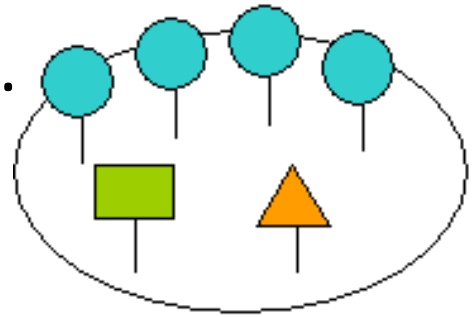
D = 0

J = 1

A.



B.



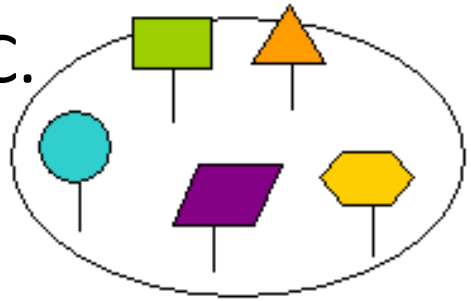
Spp = 3

H’ = 0.87

D = 0.40

J = 0.79

C.



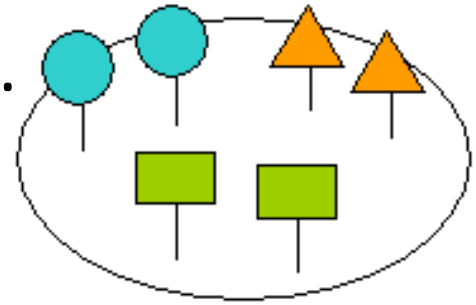
Spp = 5

H’ = 1.61

D = 0

J = 1

D.



Spp = 3

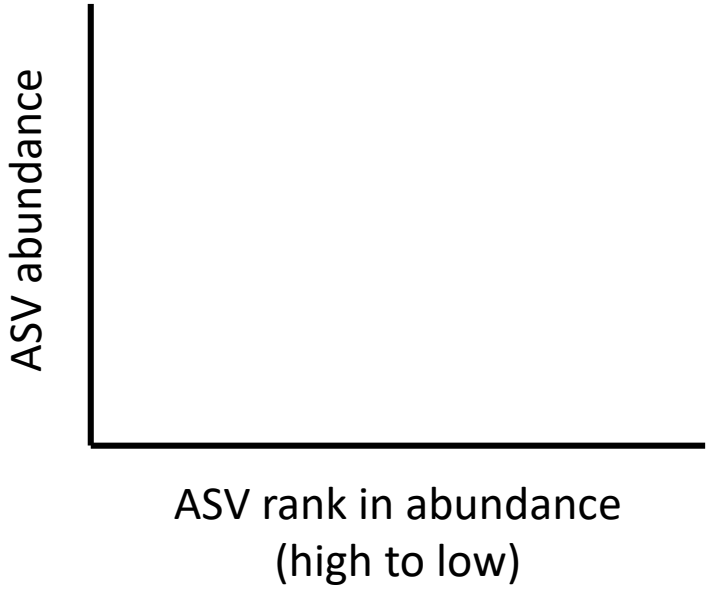
H’ = 1.10

D = 0.20

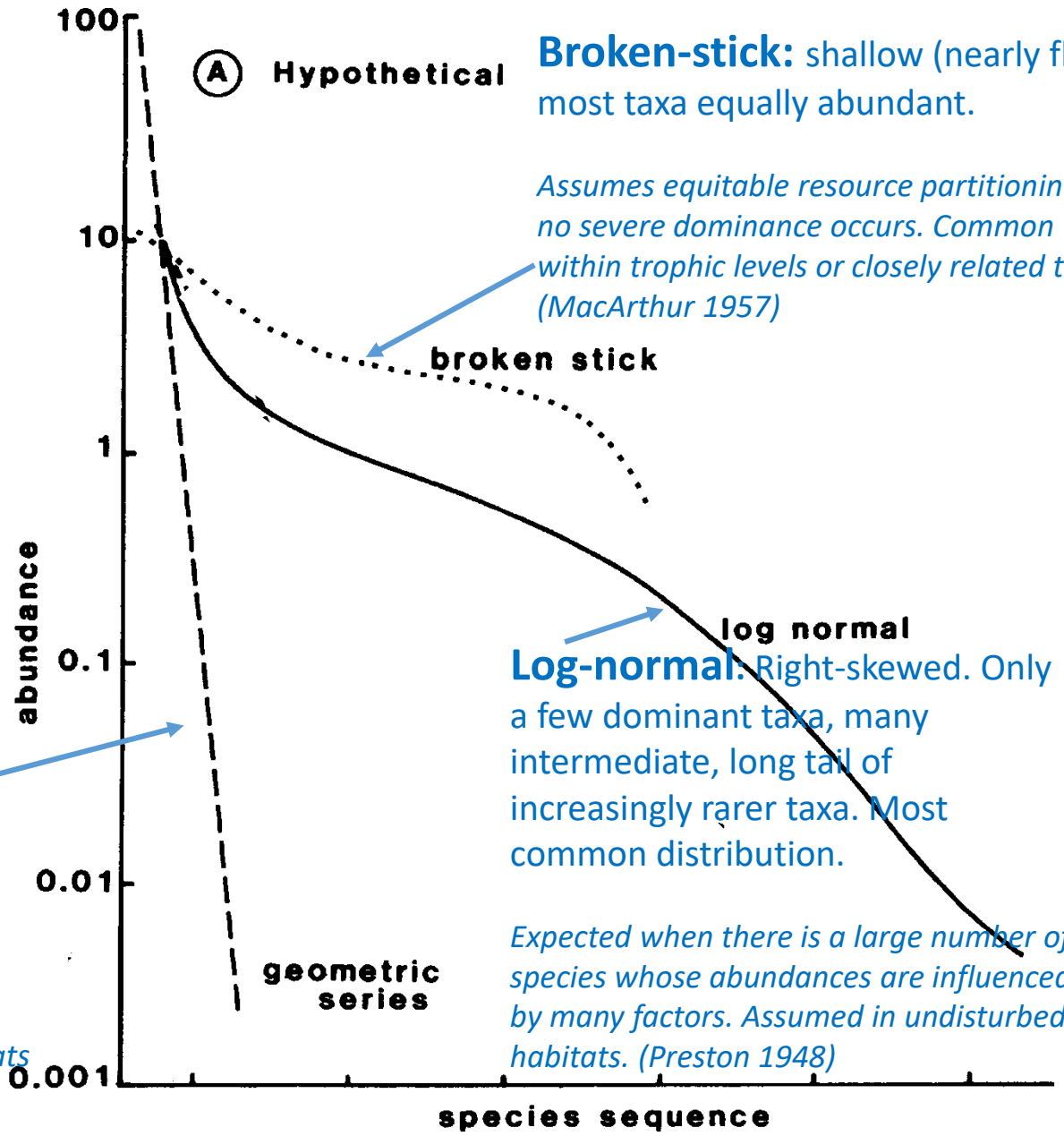
J = 1

Alpha diversity: rank-abundance

- Another way to look at evenness (and dominance) is to use rank abundance curves



Rank-abundance curves



Broken-stick: shallow (nearly flat), most taxa equally abundant.

Assumes equitable resource partitioning so no severe dominance occurs. Common within trophic levels or closely related taxa. (MacArthur 1957)

Log-normal: Right-skewed. Only a few dominant taxa, many intermediate, long tail of increasingly rarer taxa. Most common distribution.

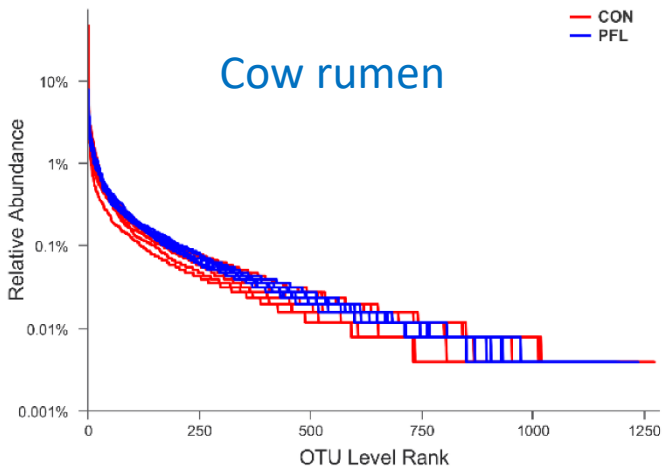
Expected when there is a large number of species whose abundances are influenced by many factors. Assumed in undisturbed habitats. (Preston 1948)

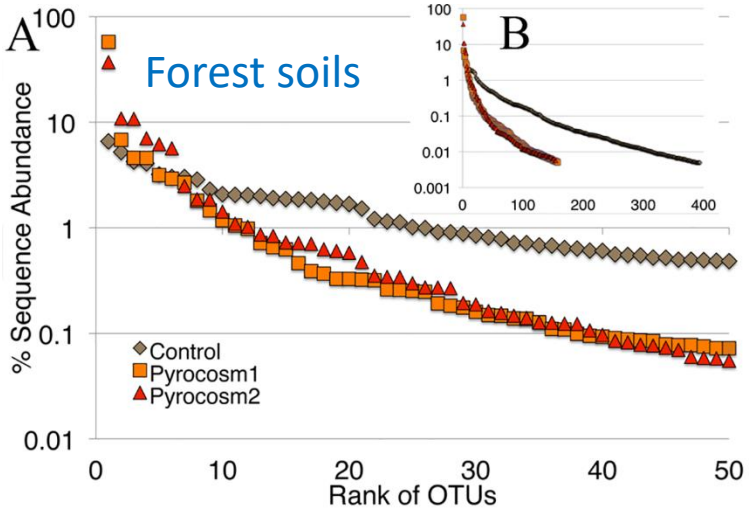
Geometric: very steep, strong dominance by 1 or a few taxa.

Assumes early-arriving taxa pre-empt resources and each successive arrival occupies a constant proportion of remainder. Common in disturbed habitats or harsh envs. (Motomura 1932)

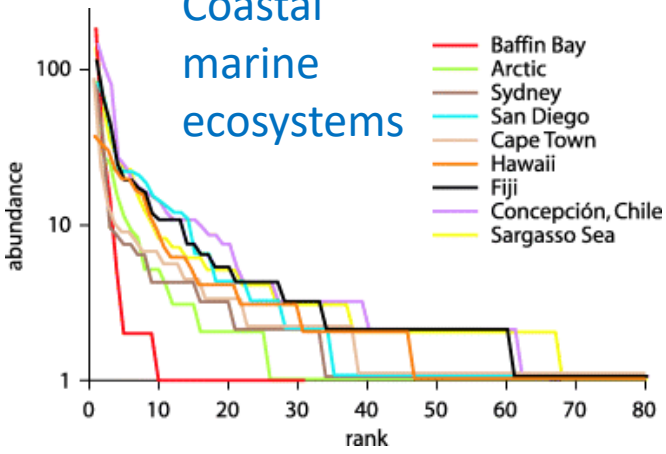
Rank abundance curves in bacteria communities

Zhiqiang et al. 2019 Microorganisms
doi.org/10.3390/microorganisms7110562

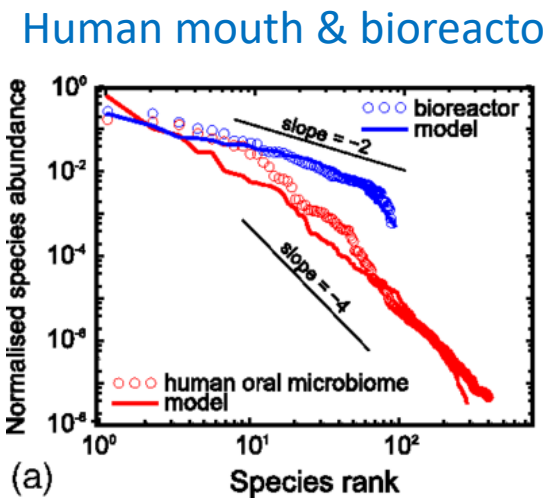




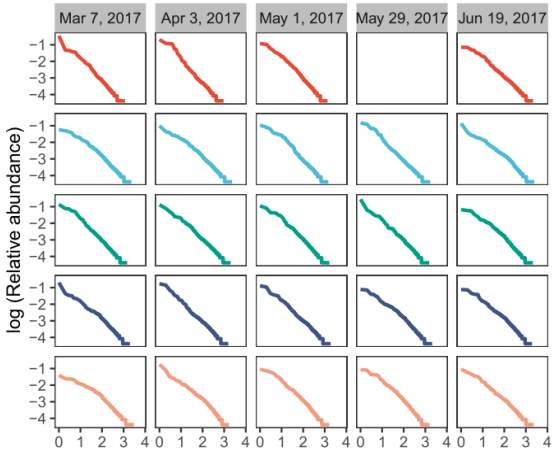
Bruns et al. 2020 PLoS1
doi.org/10.1371/journal.pone.0222691



Pommier et al. 2006 Molec Ecol
doi.org/10.1111/j.1365-294X.2006.03189.x



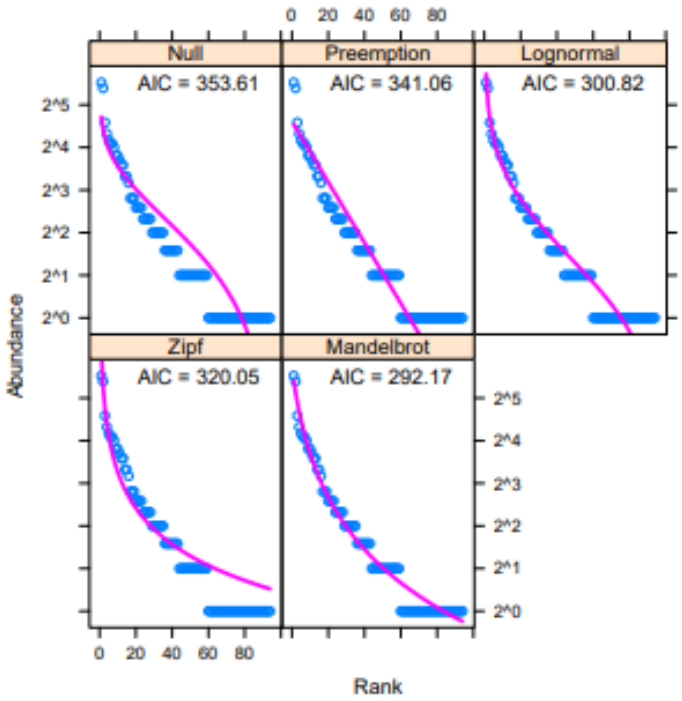
Goyal & Maslov 2018 Phys Rev Letts
doi.org/10.1103/PhysRevLett.120.158102



Guo et al. 2021 Sci Tot Env
doi.org/10.1016/j.scitotenv.2021.149169

Rank-abundance curves

- How do you determine shape?
- vegan::radfit



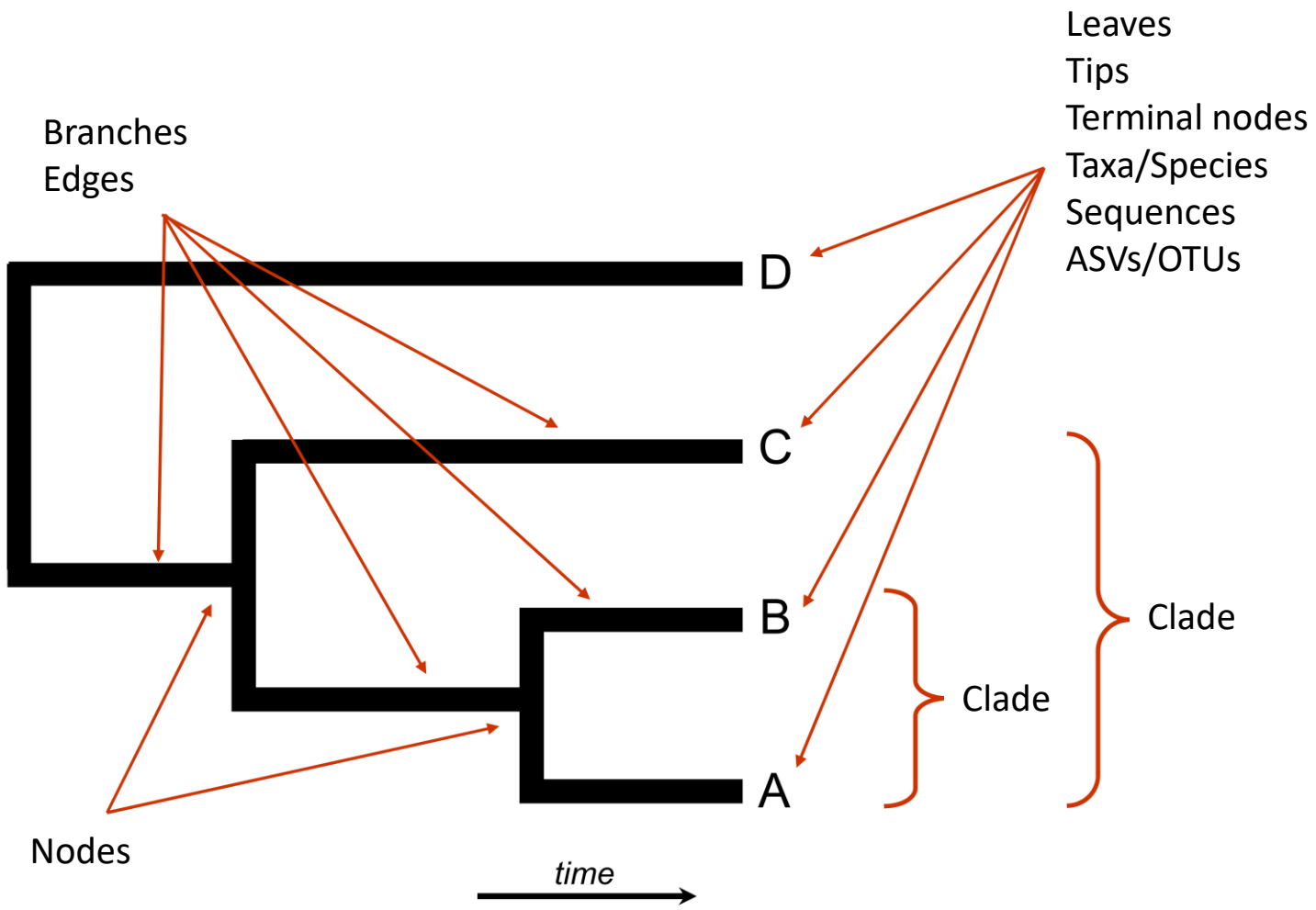

```
RAD models, family poisson
No. of species 94, total abundance 483

                par1      par2     par3    Deviance AIC      BIC
Null                                        77.2737 353.6126 353.6126
Preemption   0.048132                       62.7210 341.0598 343.6031
Lognormal    0.97341    1.1723              20.4770 300.8158 305.9024
Zipf         0.14073   -0.84897             39.7066 320.0454 325.1320
Mandelbrot   1.9608    -1.522    6.7247     9.8353 292.1741 299.8040
```

Broken stick →

Geometric →

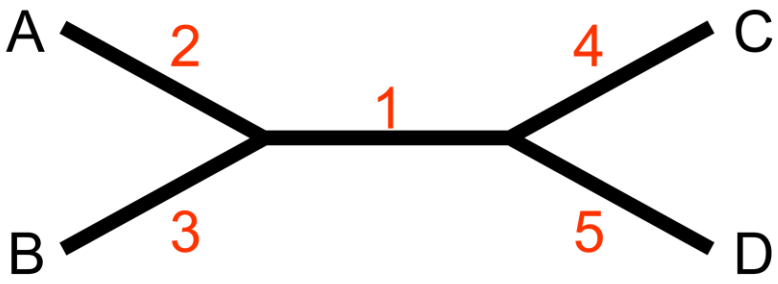
Lower values = better fit

Phylogenetic alpha diversity: tree structure



Phylogenetic alpha diversity: rooted vs. unrooted trees

Unrooted Tree



Rooted Trees – have one node from which all other nodes descend
– imply direction corresponding to evolutionary time

(r(A,B)(C,D))

(r(A(B(C,D))))

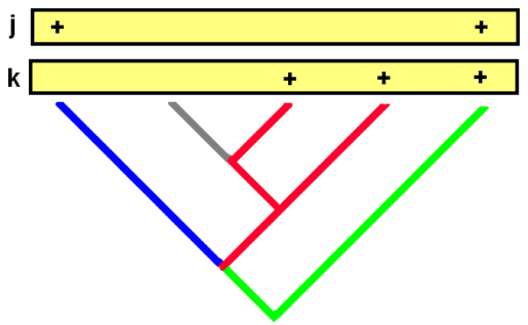
(r(B(A(C,D))))

(r(C(D(A,B))))

(r(D(C(A,B))))

Phylogenetic alpha diversity: Faith's PD

- When branch lengths are known:
  - PD = sum of total branch lengths for one or more samples

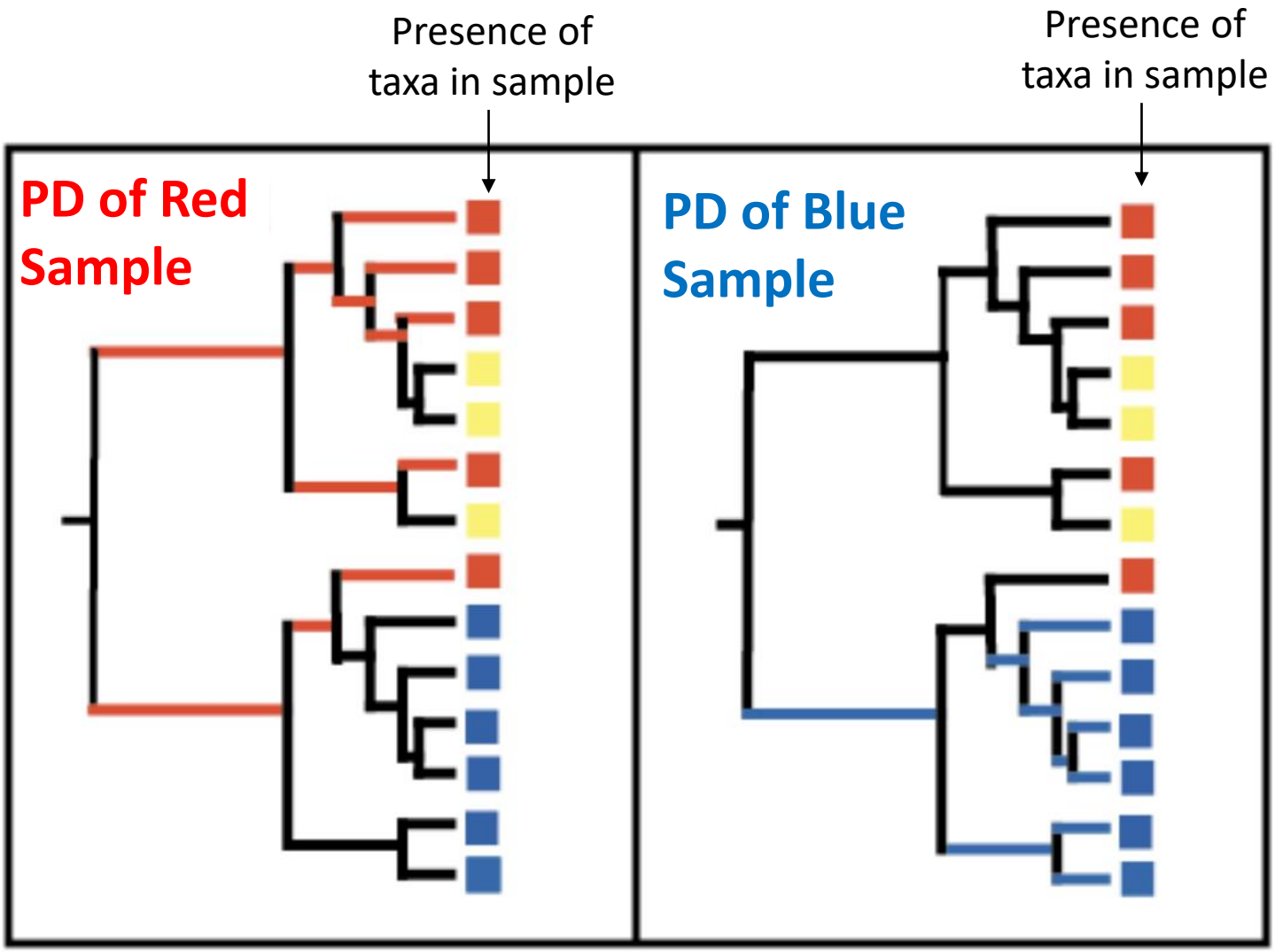


- When branch lengths are unknown:
  - PD = (N - 1) + no. of internal nodes (branching points) on the minimum spanning path
  - Where N = number of taxa

- In picante, include.root=TRUE/FALSE will include/exclude branch lengths from the calculation
- Today, we'll call picante via btools::estimate_pd to allow us to work directly with a ps object
  - Default is include.root=FALSE

Faith et al. 2009 Int J Mol Sci doi.org/10.3390/ijms10114723

Phylogenetic alpha diversity: Faith's PD



Faith 1992 Biol Conservation DOI: 10.1016/0006-3207(92)91201-3

Practical considerations

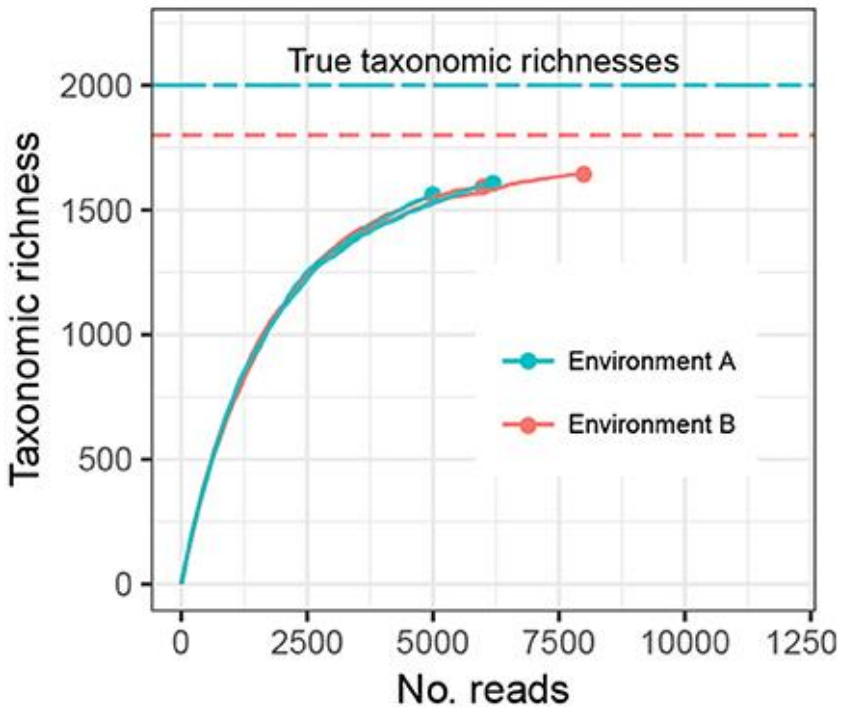
- Diversity and evenness indices and RA curves assume abundances are meaningful
- When should you calculate alpha diversity?
- Have you sampled enough to capture the true alpha diversity in your samples/system?

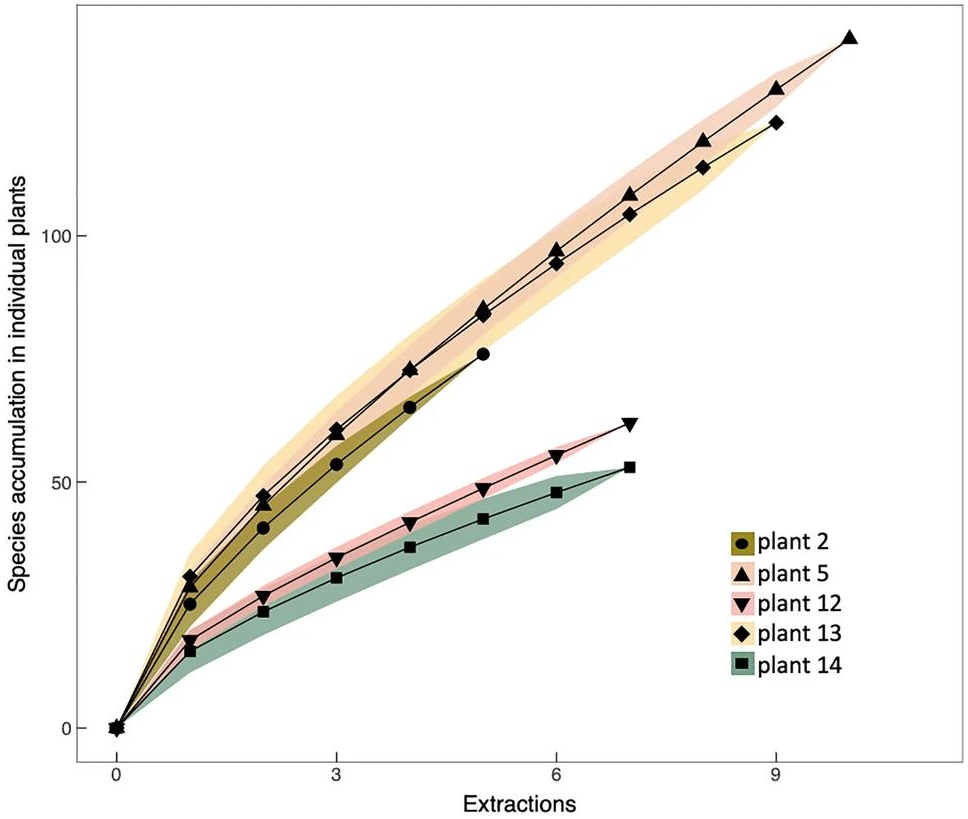
When to calculate alpha diversity

- Before singleton or rare taxa removal
- Singletons are ASVs that only occur once in the data
  - Common to remove singletons from microbial datasets
    - Assumed to be sequencing errors
    - Can use synthetic sequences in your Illumina run to estimate the true low-abundance cut off
- Others remove low abundance taxa to improve statistical properties of the data

- Remove singletons today with phyloseq::prune_taxa based on phyloseq::taxa_sums greater than 1
- Note: for practical (computing time!) purposes, we will also limit today's data to the top 100 most abundant taxa

Difficult to completely sample microbial alpha diversity in natural environments





Willis 2019 Fr Micro doi.org/10.3389/fmicb.2019.02407

Bullington et al. 2021 Sci Rep doi.org/10.1038/s41598-021-83153-9

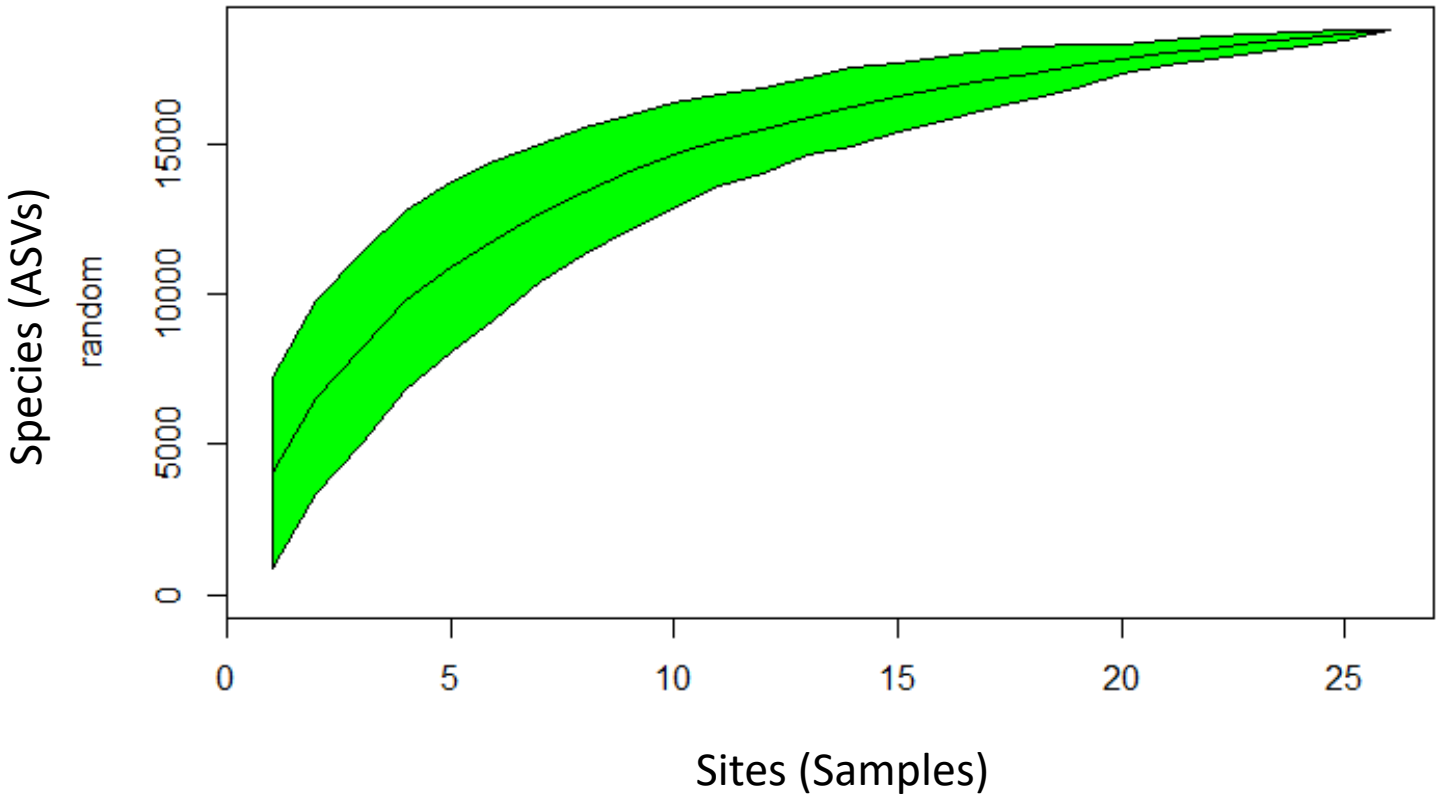
Two assessments of sampling effort

- Species accumulation curves (vegan::specaccum)
  - Accumulation of species as number of sites/samples increases
  - Focus is the site/sample level (# samples)
  - Uses re-sampling to find mean accumulation
  - New species (and SD) decreases with increasing number of sites
  - Methods options – random, exact, collector

- Rarefaction curves (vegan::rarecurve)
  - Expected number of species in random subsamples
  - Focus is on sampling effort (# reads) within sites/samples
  - Uses re-sampling random of subsamples from the data
  - Methods options – subsample size (1 to total), step size (interval)

- Does number of taxa plateau? (no new taxa expected)

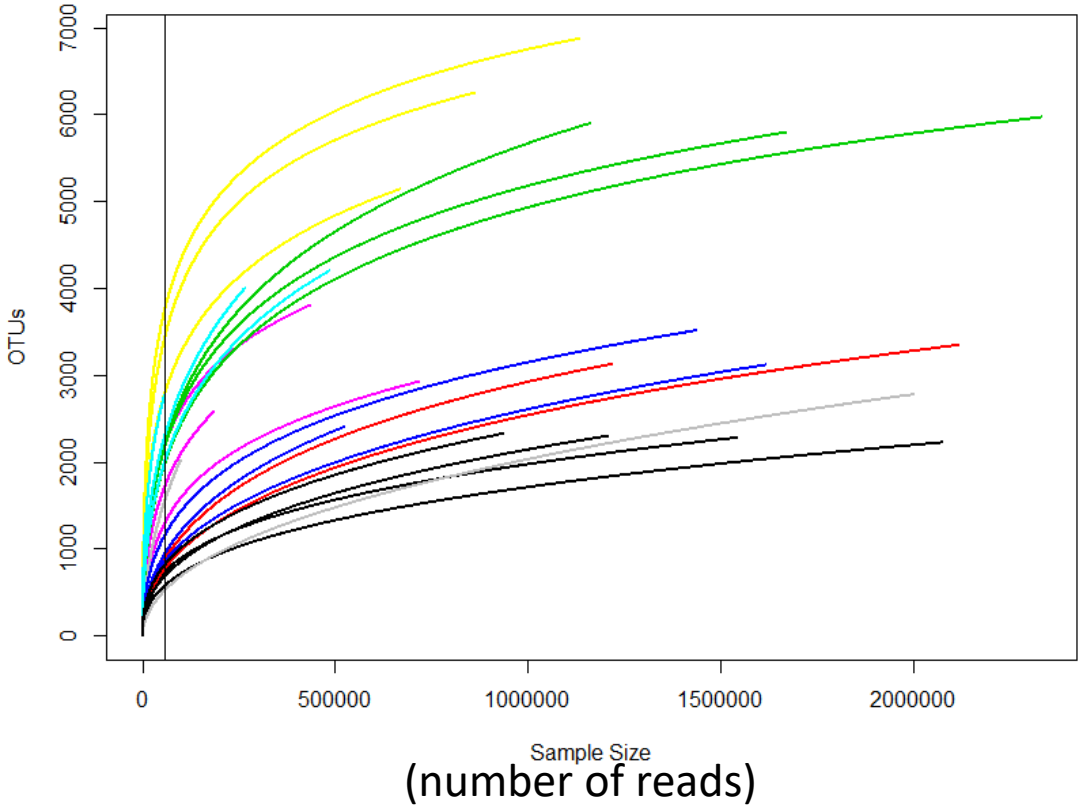
Species accumulation curve for GP all bacteria in all samples



vegan::specaccum random method

- uses random permutations of the data by subsampling sites (rows) without replacement
- number of permutations can be specified
- can also provide weights for sampling effort at each site

Rarefaction curves for GP bacteria in each sample



vegan::rarecurve method

- Draws a rarefaction curve for each sample (row) of the data using reads across columns
- Curves are evaluated using the specified “step”
  - Interval of sample size
  - Always includes 1 and total sample size (= total # sequence reads per sample)

Alpha diversity estimates for incomplete sampling: Chao1

- Use data to estimate true number of taxa
- Chao1 uses singletons and doubletons to estimate species that were undetected in the sample

$$S_1 = S_{obs} + (F_1^2)/(2F_2)$$

S_{obs} = number of species measured in sample

F_1 = number of singletons (species that occurred only once)

F_2 = number of doubletons (species that occurred only twice)

Chao 1984 Scand J Stats 11:265-270

Let’s practice!

- Switch to .html

- GlobalPatterns data (ps object built in to phyloseq!)
- Practice subsetting the data
- Alpha diversity metrics and comparisons
- Rank abundance curves
- Species accumulation and rarefaction sampling curves
- Coding exercises!