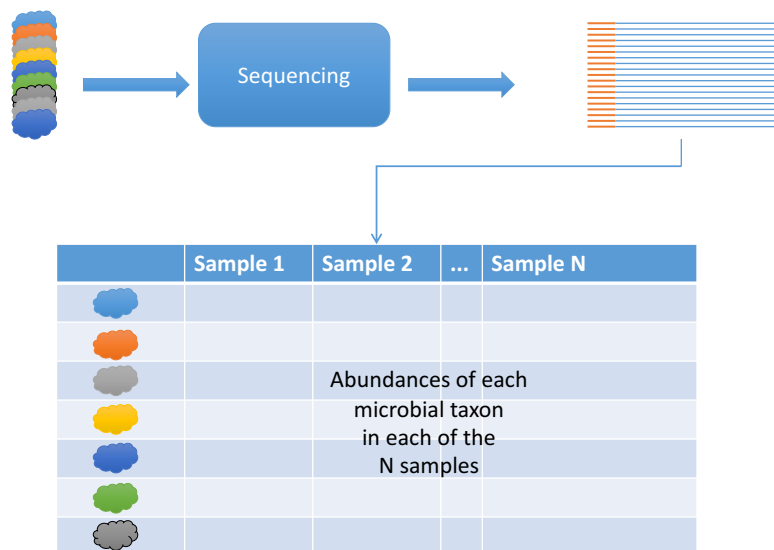


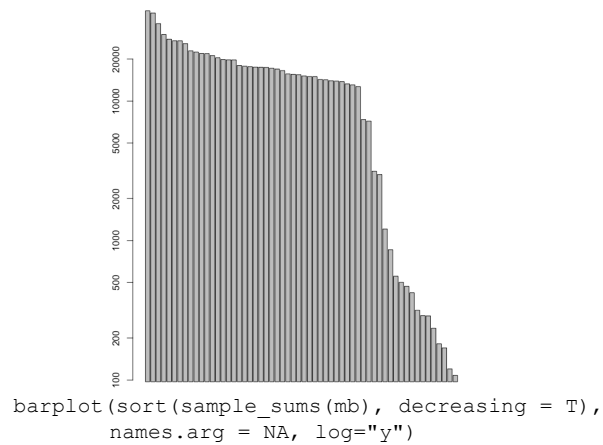
# Normalization and alpha-diversity

Derived from slides by  
Alexander V. Alekseyenko & Paul "Joey" McMurdie

From sequences to OTU table










## Number of reads per sample










3

## Normalizing OTU tables for sequencing effort

### Raw Counts

	Sample 1	...	Sample N
	$n_{11}$		$n_{1N}$
	$n_{21}$		$n_{2N}$
	$n_{31}$		$n_{3N}$
	$n_{41}$		$n_{4N}$
	$n_{51}$		$n_{5N}$
	$n_{61}$		$n_{6N}$
	$n_{71}$		$n_{7N}$
	$n_{\cdot 1}$		$n_{\cdot N}$

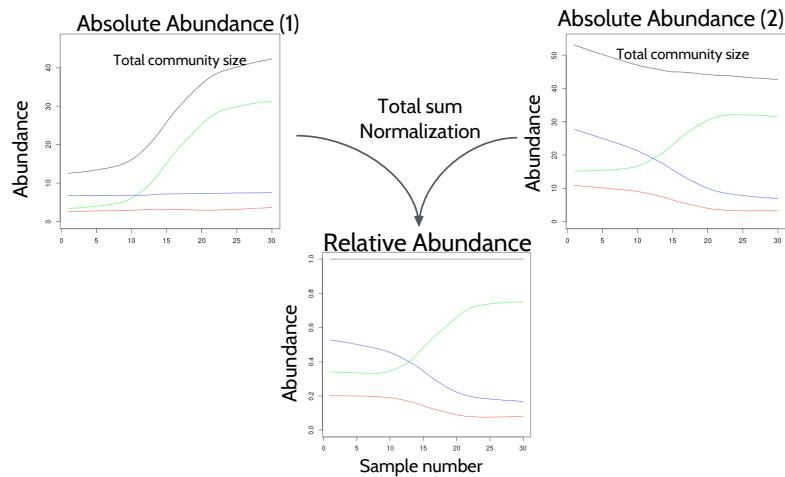
### Proportions

	Sample 1	...	Sample N
	$p_{11}$		$p_{1N}$
	$p_{21}$		$p_{2N}$
	$p_{31}$		$p_{3N}$
	$p_{41}$		$p_{4N}$
	$p_{51}$		$p_{5N}$
	$p_{61}$		$p_{6N}$
	$p_{71}$		$p_{7N}$
	1		1

$$p_{ij} = n_{ij} / n_{\cdot j}$$

4

## Potential problem with relative abundance



5

## Negative correlation of the relative abundances

- The proportions are negatively correlated by design.
- If one (or more) OTUs were to increase in absolute abundance, the relative abundances of all other OTUs will decrease to accommodate the additive constraint.

6

## Compositional data analysis: log ratios

- Main idea: ratios of absolute and compositional data are preserved
- $\log \frac{x_i}{x_j} = \log \frac{\omega_i/M}{\omega_j/M} = \log \frac{\omega_i}{\omega_j}$ ,
- Where
  - $M = \text{total community size}$
  - $i, j = \text{microbe/OTU}$
- More details is Aitchison, J. (1986). The statistical analysis of compositional data.

7

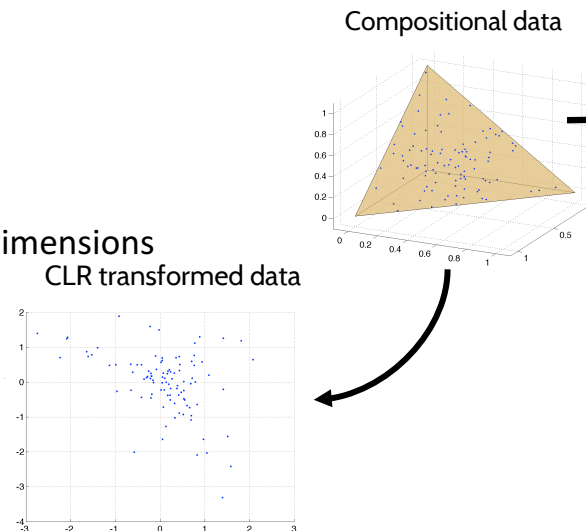
## Other normalizations

- Normalized by 1 component,  $n_d$ 
  - $y_{ij} = \log\left(\frac{n_{ij}}{n_{dj}}\right) = \log(n_{ij}) - \log(n_{dj})$
  - $n_{dj} > 0$  for all  $d$
  - Assuming the true abundance of OTU  $d$  is the same across all samples
- Normalized by geometric mean (centered)
  - $y_{ij} = \log\left(\frac{n_{ij}}{g(n_{1j}, \dots, n_{Tj})}\right) = \log(n_{ij}) - \log(g(n_{1j}, \dots, n_{Tj}))$
  - $g(n_{1j}, \dots, n_{Tj}) = (\prod_{i=1}^T n_{ij})^{1/T}$
- Note:  $\log[0] \rightarrow -\infty$ ; so often we add 'pseudo-counts' before these transformations.

8

## CLR: Centered Log-ratio transformation

- $clr(x) = \log \frac{x}{g(x)}$
- $g(x) = \sqrt[N]{x_1 x_2 \dots x_N}$
- Transformed data are unconstrained in N-1 dimensions



9

## Other normalizations

- DESeq2: normalizes by estimating the negative binomial distribution for each taxon in each sample;
- MetagenomeSeq: uses sample quantiles to normalize accounting for undersampling.

10

## Describing microbiome community is alike to taking a demographic census

	Town1	...	TownN
carpenter	$p_{11}$		$p_{1N}$
banker	$p_{21}$		$p_{2N}$
student	$p_{31}$		$p_{3N}$
teacher	$p_{41}$		$p_{4N}$
doctor	$p_{51}$		$p_{5N}$
police	$p_{61}$		$p_{6N}$
chef	$p_{71}$		$p_{7N}$
	1		1

- How many professions are represented?
- How well represented are the different professions?
- Are some professions more popular?

11

## Alpha diversity definition(s)

- Alpha diversity describes the diversity of a single community (specimen).
- In statistical terms, it is a scalar statistic computed for a single observation (column) that represents the diversity of that observation.
- There are many statistics that can describe diversity: e.g. taxonomical richness, evenness, dominance, etc.

12

## Species richness

- Suppose we observe a community that can contain up to  $k$  'species'.
- The relative proportions of the species are  $P = \{p_1, \dots, p_k\}$ .
- Richness is computed as
$$R = \mathbf{1}(p_1) + \mathbf{1}(p_2) + \dots + \mathbf{1}(p_k),$$
where  $\mathbf{1}(\cdot)$  is an indicator function, i.e.  $\mathbf{1}(x) = 1$  if  $p_i \neq 0$ , and 0 otherwise.
- Higher  $R$  means greater diversity
- Very dependent upon depth of sampling and sensitive to presence of rare species

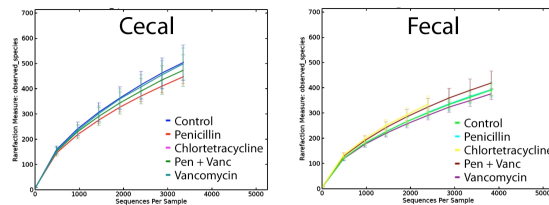
13

## Rarefaction curves

- Note: rarefaction as a means for normalization is from statistical standpoint a bad idea. Don't throw away information!
- Rarefaction curves are not the same!
- Useful to assess sensitivity of sample size to observed alpha-diversity estimates.
- Idea:
  - Let  $N_1, \dots, N_k$  be a set of numbers  $N_i < N_{i+1}$ ;
  - Let  $n'_{ij}^{(k)}$  be abundance of taxon  $i$  in sample  $j$  subsampled to  $N_k$  total counts per sample;
  - Estimate average alpha diversity for each  $N_k$  over a several repeated subsamplings;
  - Plot the average alpha diversity as a function of sample size.

14

# Rarefactions



**Supplementary Figure 6. Rarefaction curves measuring alpha diversity in fecal and cecal communities.** The vertical axis shows the number of OTUs observed after sampling the number of tags or sequences shown on the horizontal axis. Curvature toward horizontal indicates that increased sequencing effort is required to observe novel OTUs, when only rare OTUs remain to be discovered. Rarefaction curves were based on the V3 16S rRNA sequences and analyzed at OTU-level phylotypes, defined by  $\geq 97\%$  identity. Values represent the Mean  $\pm$  95% confidence interval.

Cho, I., Meth, BA., Nondorf, L., Li, K., Alekseyenko, AV., Blaser, MJ. "Subtherapeutic antibiotics alter the murine colonic microbiome and early life adiposity", *Nature* 488, 621 -- 626 (30 August 2012).

15

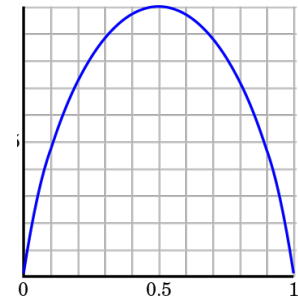
## Chao1 index

- Species richness index is often too sensitive to depth of sampling,
- Chao1 index overcomes this problem by applying a correction
- $R_C = S_{obs} + \left(\frac{f_1^2}{2f_2}\right),$
- Where  $f_1$  is the number of taxa with a single observation (singletons),  $f_2$  is the number of taxa with exactly two observations.
- If a sample contains a lot of singleton taxa, then there is a greater chance that this sample is undersampled.



## Shannon index

- Suppose we observe a community that can contain up to  $k$  'species'.
- The relative proportions of the species are  $P = \{p_1, \dots, p_k\}$ .
- Shannon index is related to the notion of information content from information theory. It roughly represents the amount of information that is available for the distribution of  $P$ .
- When  $p_i = p_j$  for all  $i$  and  $j$ , then we have no information about which species a random draw will result in. As the inequality becomes more pronounced, we gain more information about the possible outcome of the draw. The Shannon index captures this property of the distribution.
- Shannon index is computed as
 
$$S_k = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \dots - p_k \log_2 p_k$$
 Note as  $p_i \rightarrow 0$ ,  $\log_2 p_i \rightarrow -\infty$ , we therefore define  $p_i \log_2 p_i = 0$ .
- Higher  $S_k$  means higher diversity



17

## From Shannon to Evenness

- Shannon index for a community of  $k$  species has a maximum at  $\log_2 k$
- We can make different communities more comparable if we normalize by the maximum
- Evenness index is computed as
 
$$E_k = S_k / \log_2 k$$
- $E_k = 1$  means total evenness

18

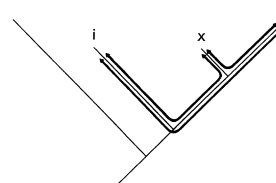
## Simpson index

- Suppose we observe a community that can contain up to  $k$  'species'.
- The relative proportions of the species are  $P = \{p_1, \dots, p_k\}$ .
- Simpson index is the probability of resampling the same species on two consecutive draws with replacement.
- Suppose on the first draw we picked species  $i$ , this event has probability  $p_i$ , hence the probability of drawing that species twice is  $p_i * p_i$ .
- Simpson index is thus computed as
 
$$D = 1 - (p_1^2 + p_2^2 + \dots + p_k^2)$$
- $D = 0$  means no diversity (1 species is completely dominant)
- $D = 1$  means complete diversity

19

## Phylogenetic Diversity (Faith's D)

- Faith (Biological Conservation 1992, 61, 1-10) considered the problem of selecting species for conservation so as to preserve diversity.
- Faith defines PD (phylogenetic diversity) as the sum of all the branch lengths. PD is analogous to total information in the tree.
- The marginal contribution of a tip  $x$  is then  $\min_{i,j}(D_{x,i} + D_{x,j} - D_{i,j})$ . Higher value suggest a greater impact on conservation.



20

## Numbers equivalent diversity

- Often it is convenient to talk about alpha diversity in terms of equivalent units:
  - How many equally abundant taxa will it take to get the same diversity as we see in a given community?
- For richness there is no difference in statistic
- For Shannon, remember that  $\log_2 k$  is the maximum which is attained when all species are equally represented. Hence the diversity in equivalent units is  $2^{S_k}$
- For Simpson the equivalent units measure of diversity is  $1/(1-D)$