Day 2: transformations & diversities

Leo Lahti

Workflow

Expanding out from the data container.

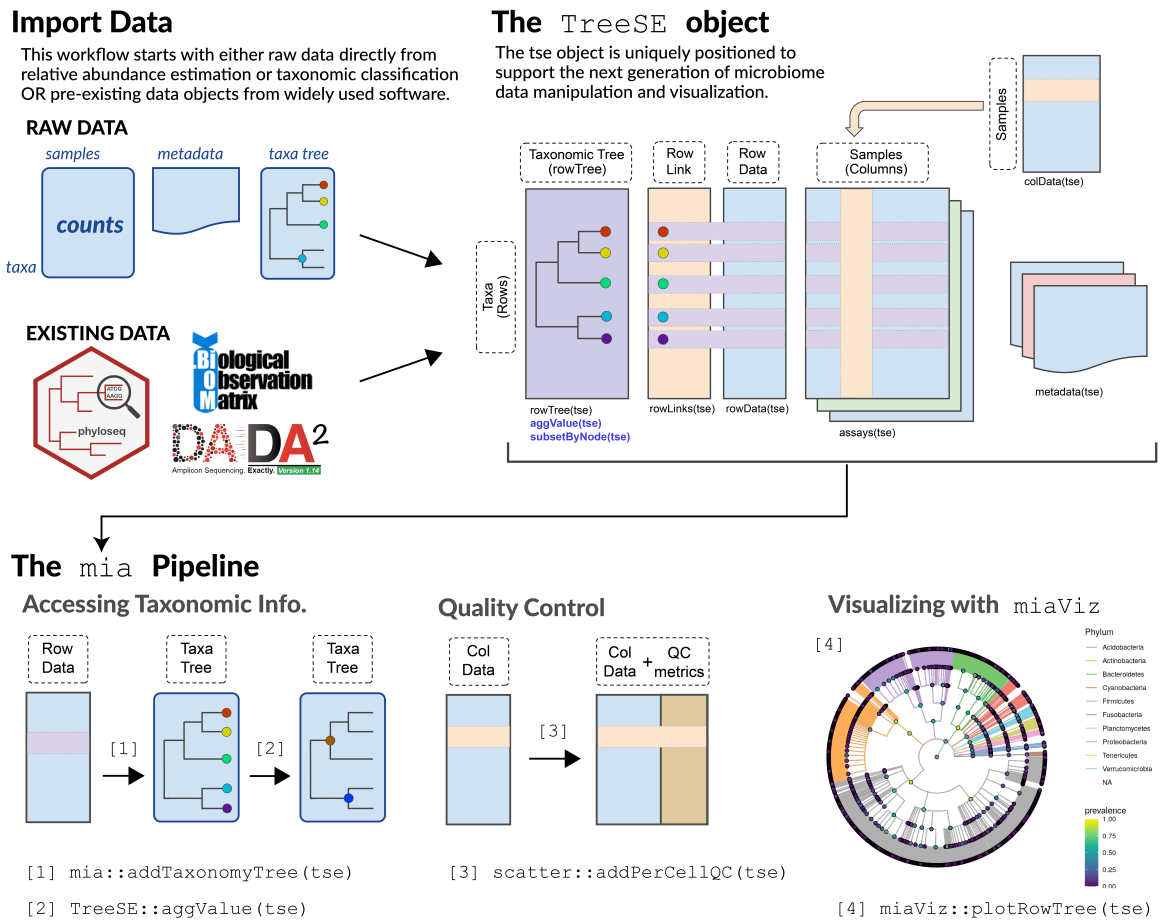


Figure 1: Domenick Braccia, EuroBioc 2020 (microbiome.github.io)

Overview of Day 2

Time	Theme
9-10	Overview & data import
10-11	Assays & transformations
11-12	Alpha & beta diversity
12-	Summary, Q & A

Data import

Data import from source files

- Construct a new TreeSE object from scratch, starting from original data files.
- Taxonomic profiling from 40 rat Cecum samples including 12706 OTUs from 318 species
- Diet comparison with High/Low fat diet and xylo-oligosaccaride supplementation.

Figure 2: International Journal of Environmental Research and Public Health 18(8):4049 https://doi.org/10.3390/ijerph18084049

Data import from source files

Construct a new TreeSE object from scratch, starting from original data files.

OMA, Chapter 18 (18.3.1-3) includes additional tips

- Read in the CSV files. You can use shared example data files on the cloud server; see in R: `dir("shared/data")`

- Data files include sample data (`coldata.csv`); taxonomic table (`rowdata_taxa.csv`); taxonomic abundance table (`assay_taxa.csv`). Load these in your RStudio session with e.g. `read.csv("shared/data/coldata.csv")`
- Construct TreeSE in R (see OMA Ch. 2)

Data import from other source formats

OMA, Chapter 18 (18.3.4)

- Follow the biom file example in OMA 2.3.2.1
- Example data files are available on the cloud server; see in R: `dir("shared/data")`

Data conversions

OMA, Chapter 18 (18.3.5)

TreeSummarizedExperiment and *phyloseq* are alternative containers for microbiome data in R. It is useful to know how to convert between these two formats.

- Convert your TreeSE into phyloseq
- Convert the phyloseq back to TreeSE

Data exploration

OMA, Chapter 18 (18.4.3 - 18.4.5)

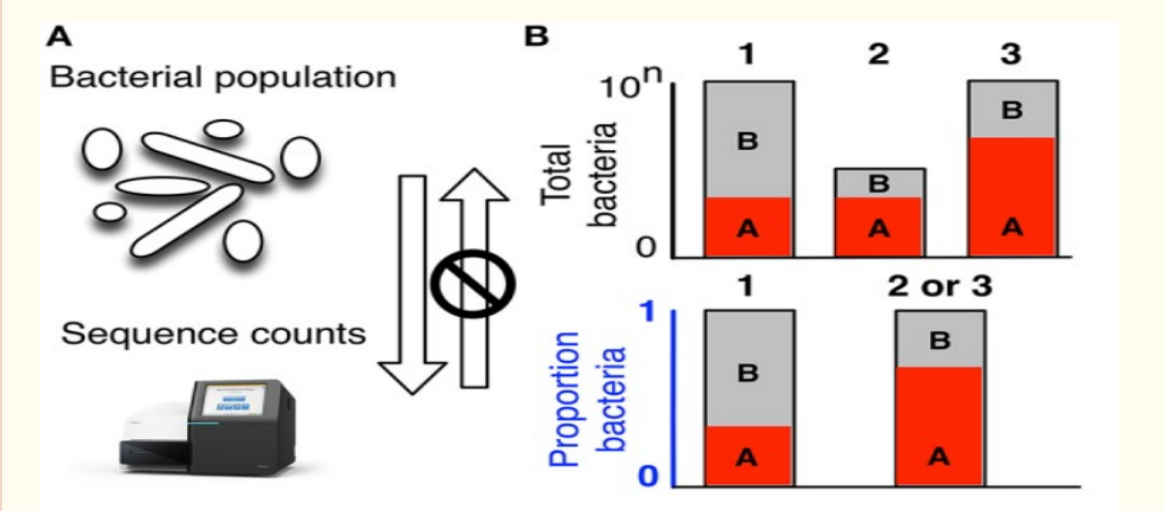
- Optional: if time will allow, you can try out these as well

Assays & transformations

Front Microbiol. 2017; 8: 2224. PMCID: PMC5695134
Published online 2017 Nov 15. doi: 10.3389/fmicb.2017.02224 PMID: 29187837

Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor,[1,*] Jean M. Macklaim,[1] Vera Pawlowsky-Glahn,[2] and Juan J. Egozcue[3]



Normalizing library size?

Bias in compositional data:

- If sample A has been sampled deeper than sample B, also the counts can be expected to be higher.

Possible solutions:

- Divide by the total number of reads per sample (*compositional abundance*)
- Rarify (subsample) to even sampling depth

→ Problem: **Abundant taxa may distort the ratios**

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes

Relative vs. absolute abundances

-> Compositional data analysis (CoDa)

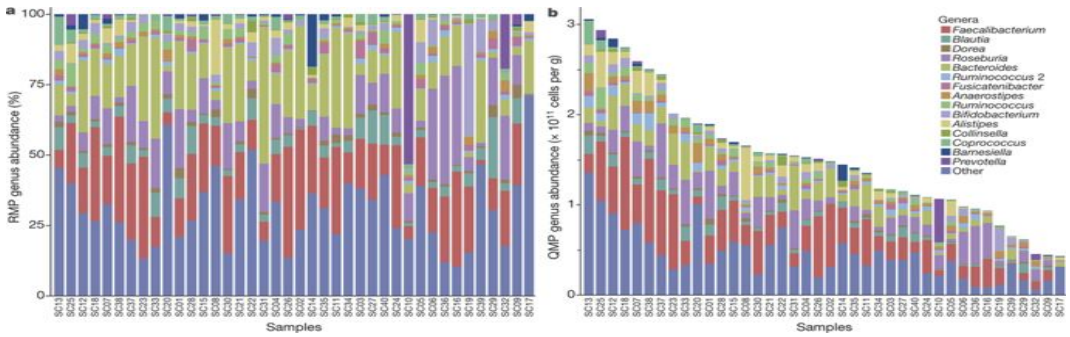


Figure 3: Vandeputte et al. Nature 551:507-511, 2017

RMP vs. QMP

Potentially drastic effect on conclusions!

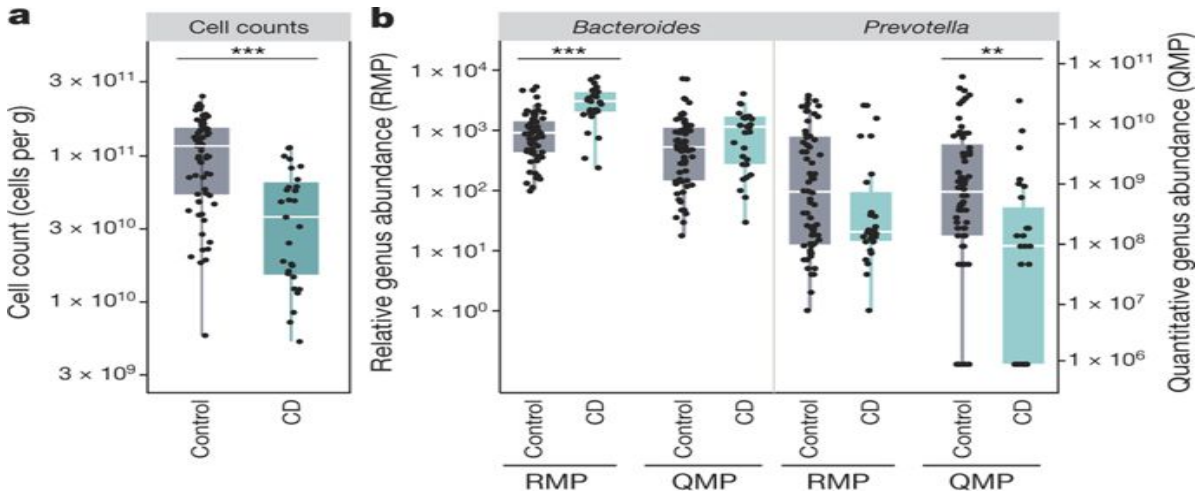


Figure 4: Vandeputte et al. Nature 551:507-511, 2017

Abundance along the community landscape

Aitchison transformations (e.g. CLR)

- Aitchison transformations are used to reduce compositional bias.

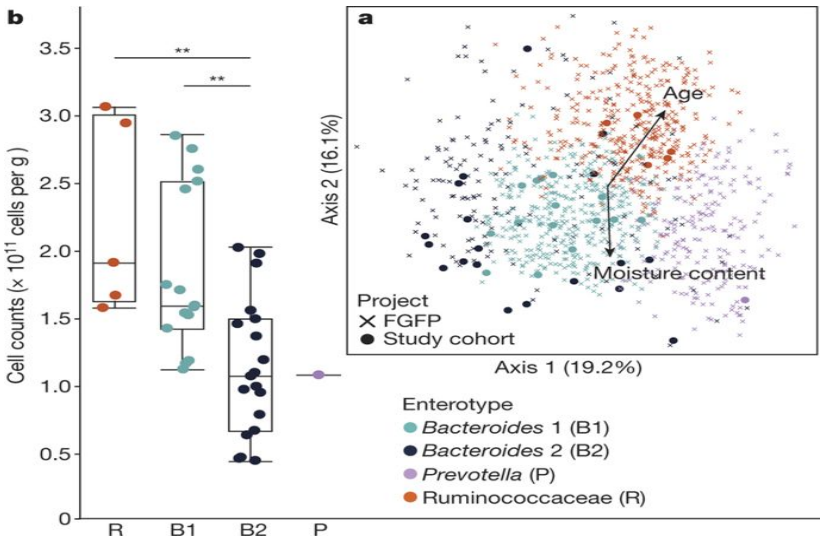


Figure 5: Vandeputte et al. Nature 551:507-511, 2017

- Balances, or ratios between taxa abundances, are conserved in compositional transformation: $\frac{x}{y} = \frac{cx}{cy}$

$$\text{clr}(\text{x}) = \left[\ln \frac{x_i}{g(\text{x})}; \ldots; \ln \frac{x_D}{g(\text{x})}\right]$$

$$\text{alr}(\text{x}) = \left[\ln \frac{x_i}{x_D}; \ldots; \ln \frac{x_{D-1}}{x_D}\right]$$

Transformations

- Count data
- Presence/absence
- Compositional (percentages)
- Log_{10}
- *Aitchison* family of transformations (CLR, ALR, ILR)
- Phylogenetic transformations (e.g. phILR)
- Custom transformations

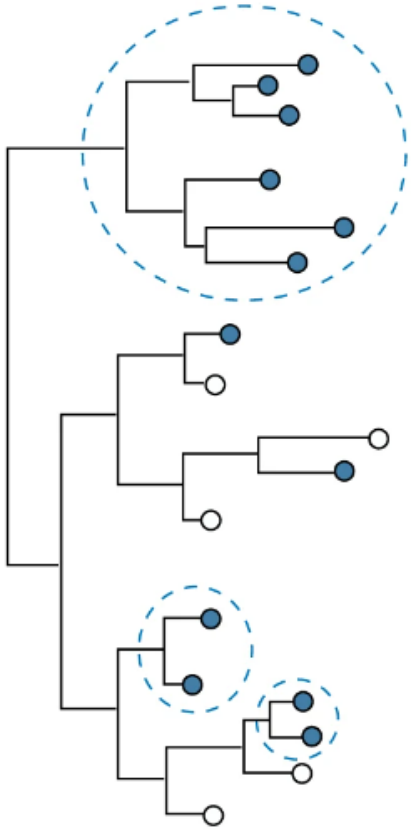
Transformations

Create and add new assays in the data:

- OMA exercise 18.5

Phylogenetic transformations

- - there exists no single taxonomic resolution at which taxonomic variation unambiguously reflects functional variation, and at which environmental selection of certain functions - - unambiguously translates to a selection of specific taxa (Louca et al. 2018).



Function and functional redundancy in microbial systems

Stilianos Louca ✉, Martin F. Polz, Florent Mazel, Michaeline B. N. Albright, Julie A. Huber, Mary I. O'Connor, Martin Ackermann, Aria S. Hahn, Diane S. Srivastava, Sean A. Crowe, Michael Doebeli & Laura Wegener Parfrey

Nature Ecology & Evolution **2**, 936–943 (2018) | Cite this article

12k Accesses | **463** Citations | **191** Altmetric | Metrics

Phylogenetic balances: phILR transformation

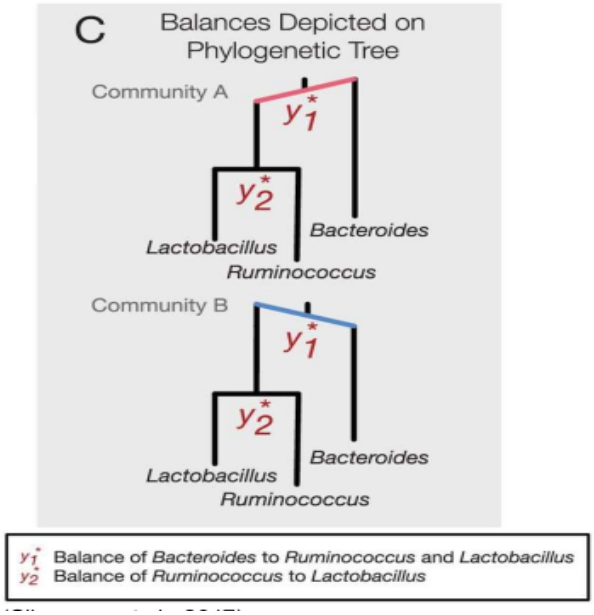
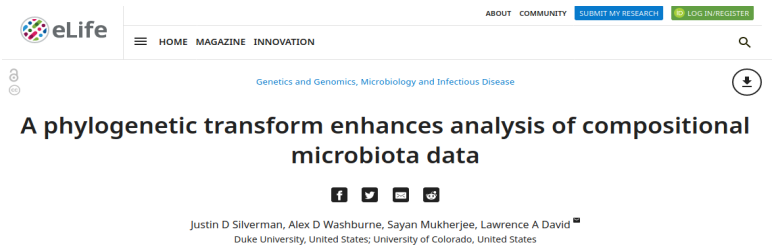


Figure 6: Silverman et al. 2017



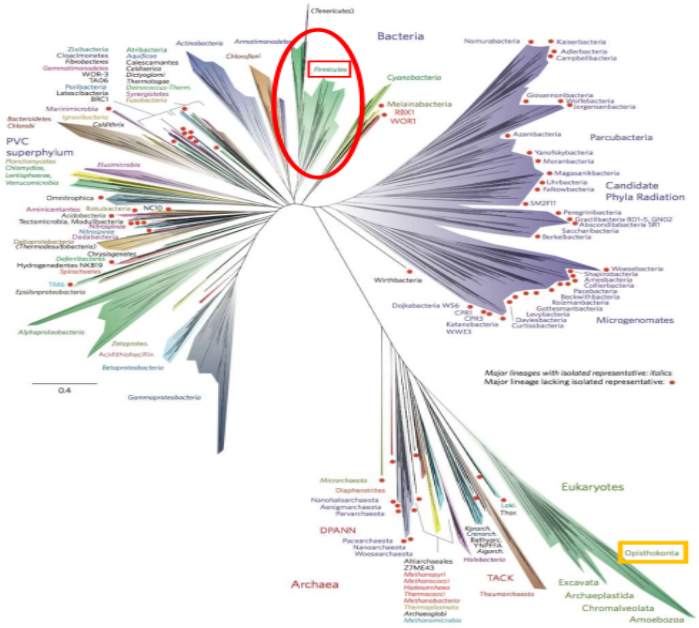
Phylogenetic balances as features for FLI predictions

(Based on phILR; Silverman et al. 2017)

Pathways in representative bacterial genomes of Clostridium subclusters IV and XIVa indicated the presence of e.g., ethanol fermentation pathways → endogenous ethanol producers associated with fatty liver?

In addition to age and sex, the models included differences in 11 microbial groups from class Clostridia, mostly belonging to orders Lachnospirales and Oscillospirales. Previously NAFLD-associated Clostridia XIVa group members were detected. Two species in Clostridia IV group were not previously associated with fatty liver disease.

Key associations validated in another Finnish cohort (N=258).



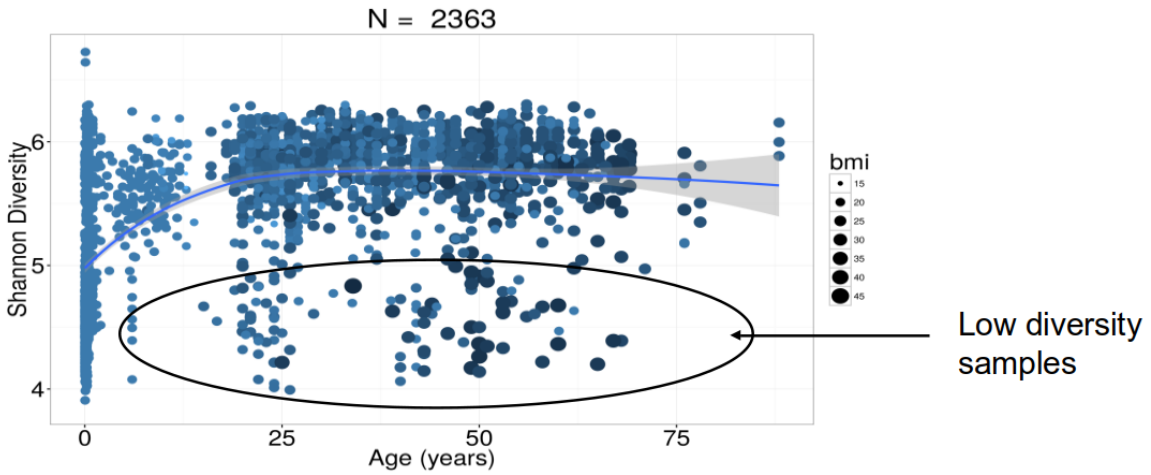
Optional task

- If the data has phylogenetic tree, perform the phILR transformation (e.g. `GlobalPatterns` data set in the **mia** R package)

Alpha diversity

Alpha diversity & aging

Healthy & normal obese subjects.

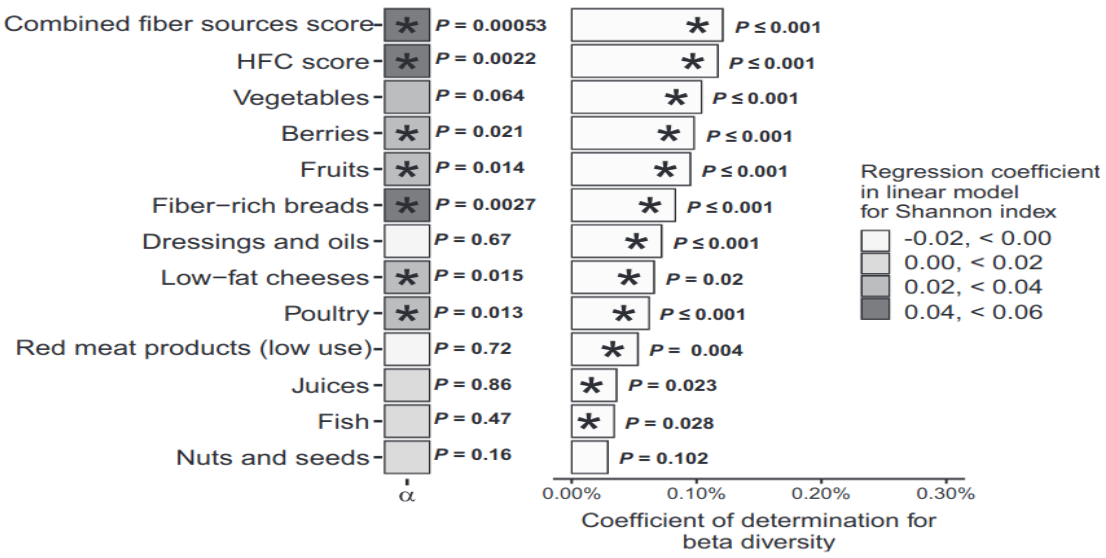


Alpha diversity and diet

Am J Clin Nutr. 2021 Aug; 114(2): 605–616.
Published online 2021 May 21. doi: 10.1093/ajcn/nqab077
PMCID: PMC8326043
PMID: 34020448

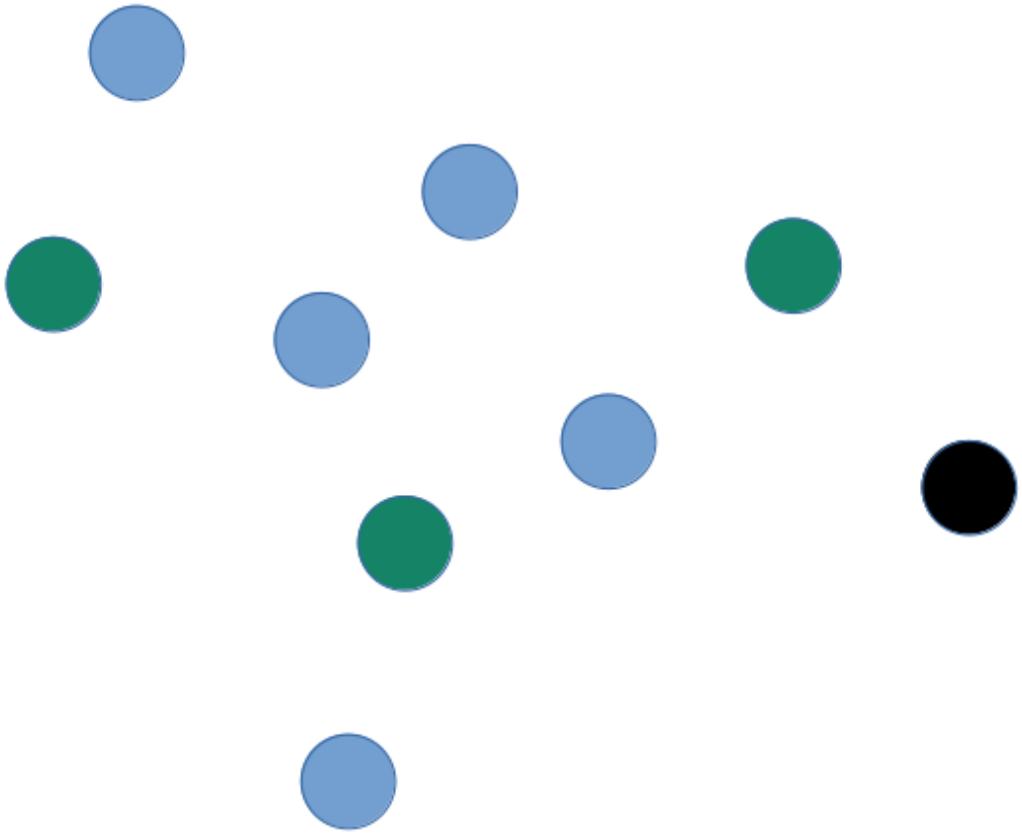
Associations of healthy food choices with gut microbiota profiles

Kari K Koponen, Aaro Salosensaari, Matti O Ruuskanen, Aki S Havulinna, Satu Männistö, Pekka Jousilahti, Joonatan Palmu, Rodolfo Salido, Karenina Sanders, Caitriona Brennan, Gregory C Humphrey, Jon G Sanders, Guillaume Meric, Susan Cheng, Michael Inouye, Mohit Jain, Teemu J Niiranen, Liisa M Valsta, Rob Knight, and Veikko V Salomaa



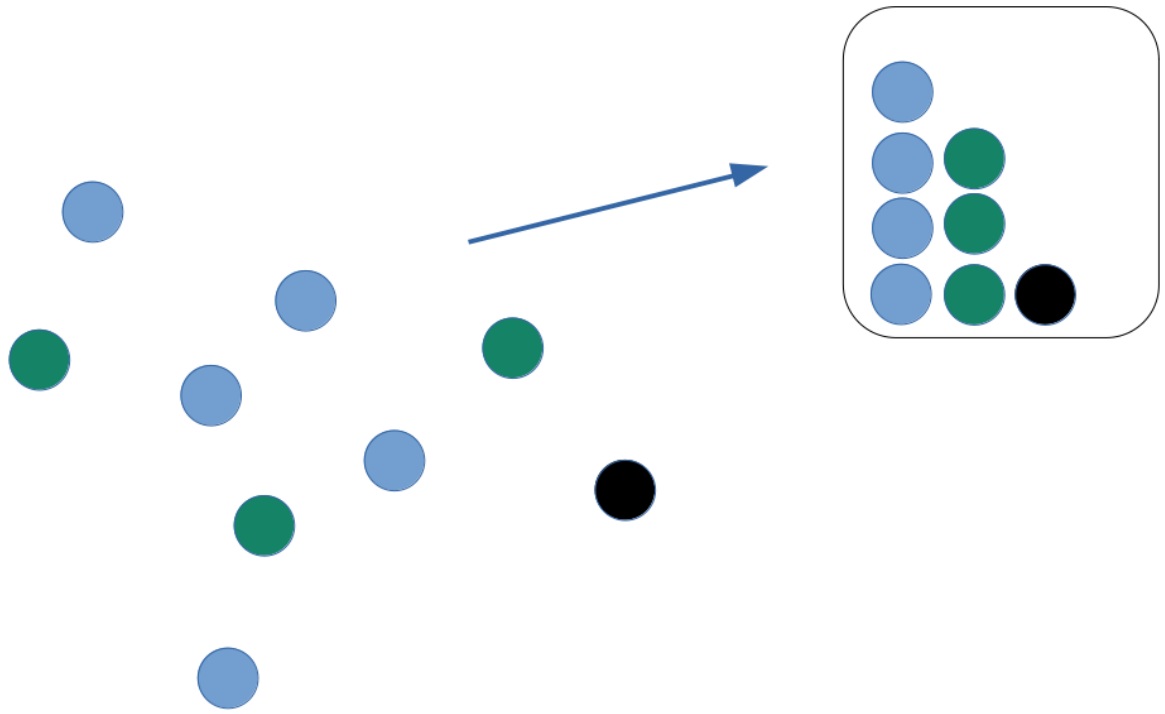
Alpha diversity

- How many types?
- Distribution of types?
- Dominance of types?



Alpha diversity

- How many types?
- Distribution of types?
- Dominance of types?



Alpha diversity indices

Richness

- number of types
- Eetimates of true richness based on finite sample sizes (Howard Sanders 1968); see e.g. Chao1

Evenness

- distribution of sizes (even or uneven?)

Diversity

- Combining richness & evenness

Dominance

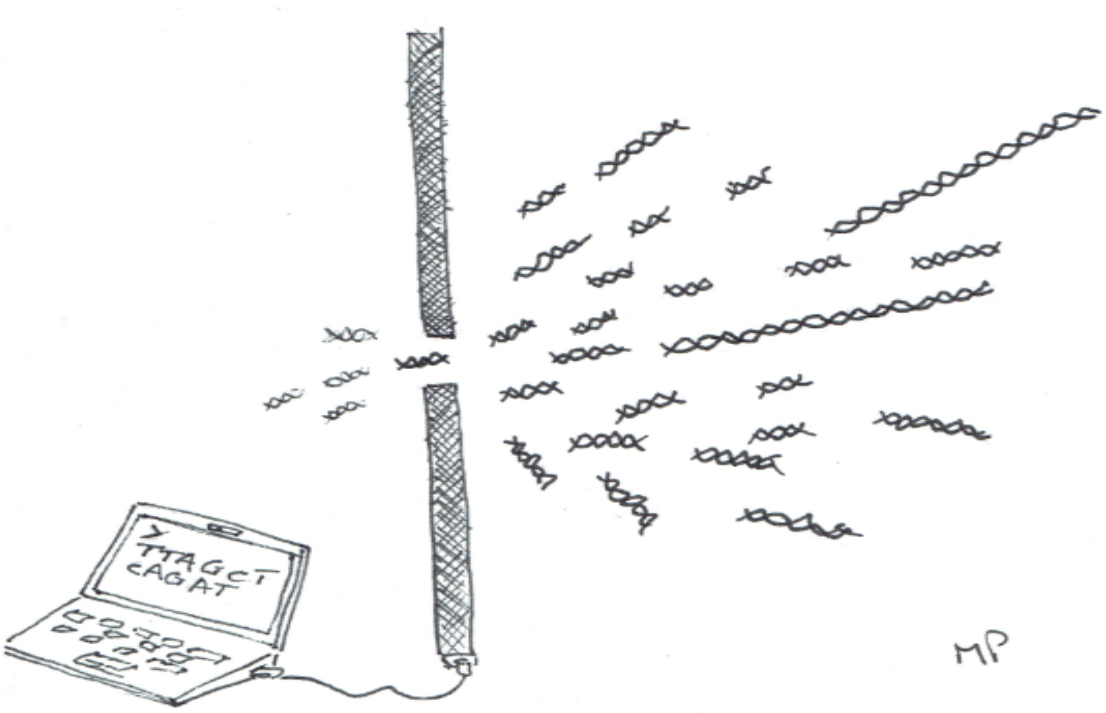


Figure 7: https://github.com/mblstamps/stamps2019/blob/master/STAMPS2019_overview_Pop.pdf

Finite sampling

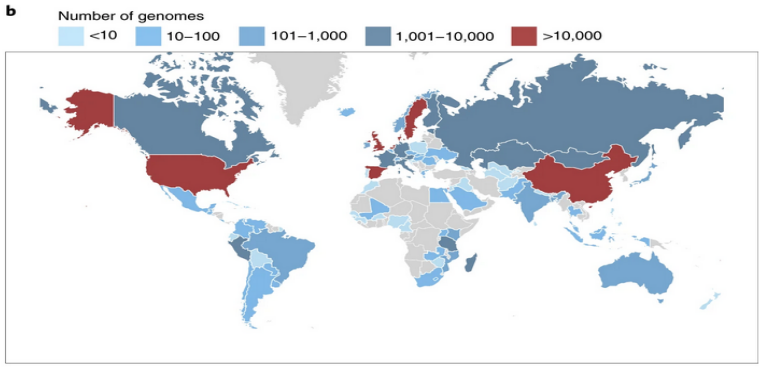
Resource | Open Access | Published: 20 July 2020

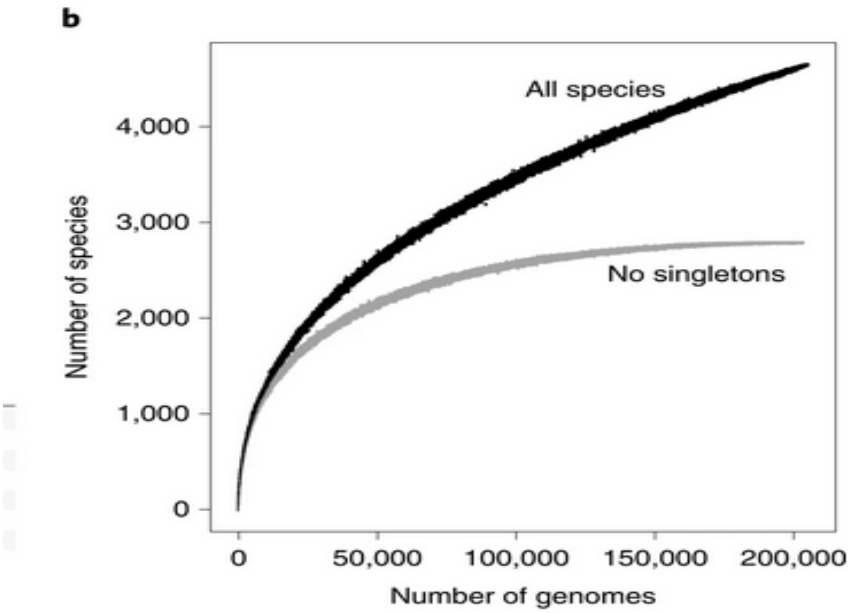
A unified catalog of 204,938 reference genomes from the human gut microbiome

Alexandre Almeida ✉, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, Ekaterina Sakharova, Donovan H. Parks, Philip Hugenholtz, Nicola Segata, Nikos C. Kyrpides & Robert D. Finn ✉

Nature Biotechnology **39**, 105–114 (2021) | Cite this article

51k Accesses | **153** Citations | **680** Altmetric | Metrics





High-quality reference genomes are required for functional characterization and taxonomic assignment of the human gut microbiota.

Unified Human Gastrointestinal Genome (UHGG):

- 4,644 gut prokaryotes (>70% lack cultured representatives)
- 204,938 nonredundant genomes
- Encode >170 million protein sequences, collated into Unified Human Gastrointestinal Protein (UHGP) catalog.

UHGP more than doubles the number of gut proteins in comparison to those present in the Integrated Gene Catalog.

- 40% of the UHGP lack functional annotations
- Intraspecies genomic variation analyses revealed a large reservoir of accessory genes and single-nucleotide variants, many of which are specific to individual human populations.

The UHGG and UHGP collections enable studies linking genotypes to phenotypes in the human gut microbiome.

Estimating species content

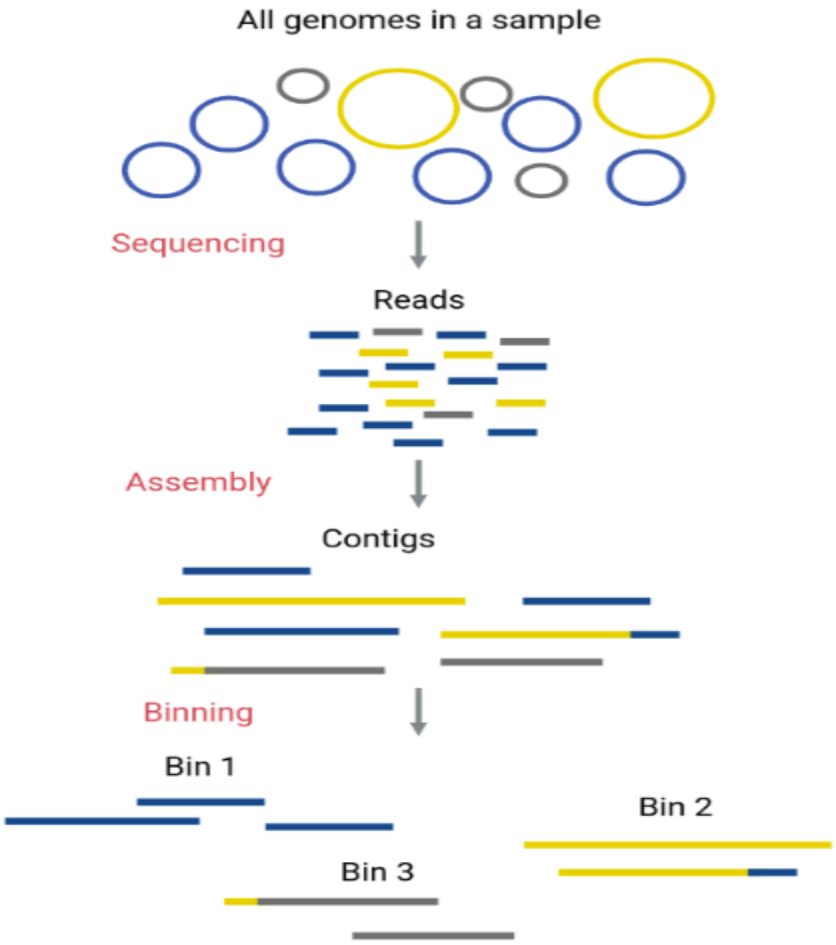
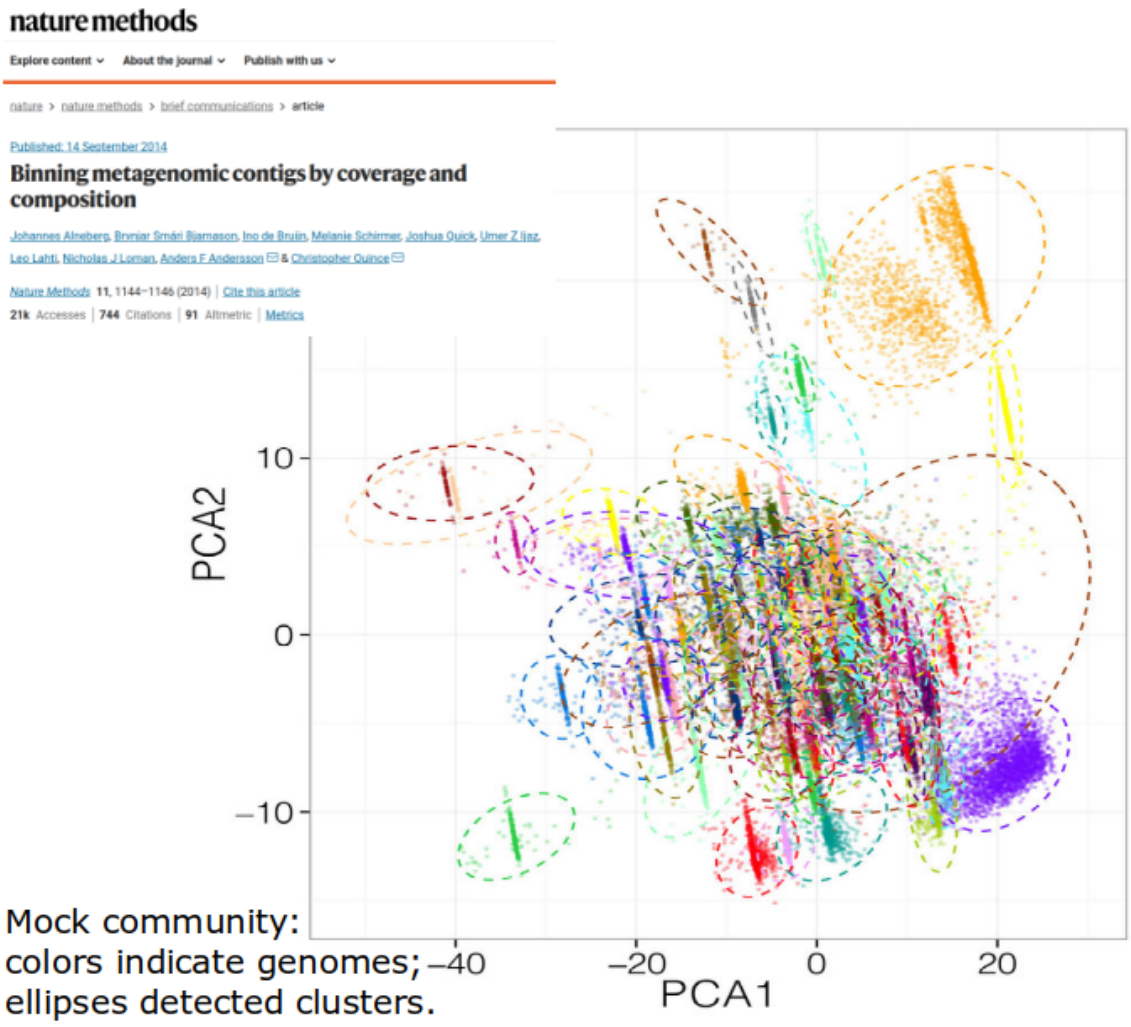


Figure 8: Copyright © Claudia Zirion, Diego Garfias, Vanessa Arellano, Aaron Jaime, Abel Lovaco, Daniel Díaz, Abraham Avelar, Nelly Sélem https://carpentries-incubator.github.io/metagenomics-workshop/)



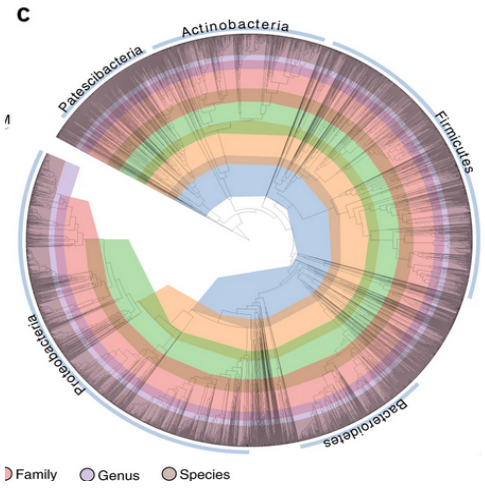
Mock community: colors indicate genomes; ellipses detected clusters.

Published: 27 August 2018

A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life

Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil & Philip Hugenholtz

Nature Biotechnology **36**, 996–1004(2018) | Cite this article



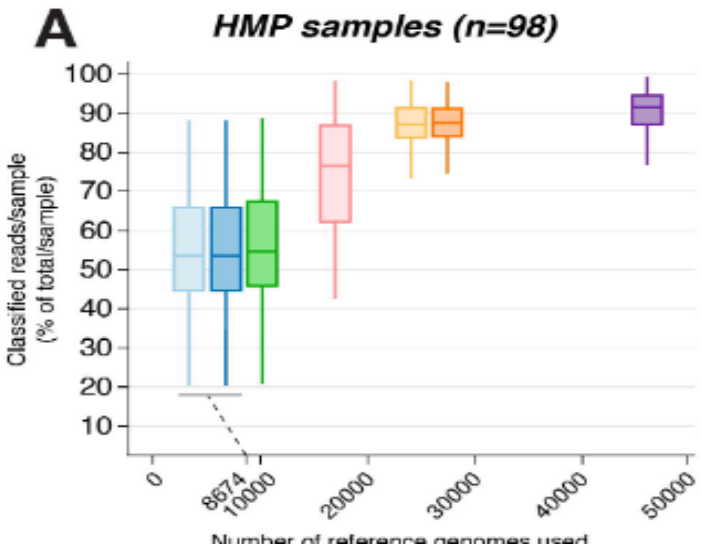
Correcting index databases improves metagenomic studies

Guillaume Méric, Ryan R. Wick, Stephen C. Watts, Kathryn E. Holt, Michael Inouye

doi: https://doi.org/10.1101/712166

Legend:

Index name	Taxonomic definitions	# genomes
NCBI_r86	Default NCBI RefSeq r86	8,674
GTDB_r86_8.6k	NCBI_r86 with GTDB definitions	8,674
NCBI_r88	Default NCBI RefSeq r88	10,089
NCBI_r88_Human17k	NCBI_r88 + 70 corrected human-associated genera	16,908
GTDB_r86_noMAGs	GTDB_r86 without MAGs	25,660
GTDB_r86	Representative genomes from GTDB	28,560
GTDB_r86_46k	GTDB_r86 + more genomes / taxon	46,006



Alpha diversity task

Use the available tools to assess and visualize alpha diversity:

- OMA exercise 18.7

Common alpha diversity indices

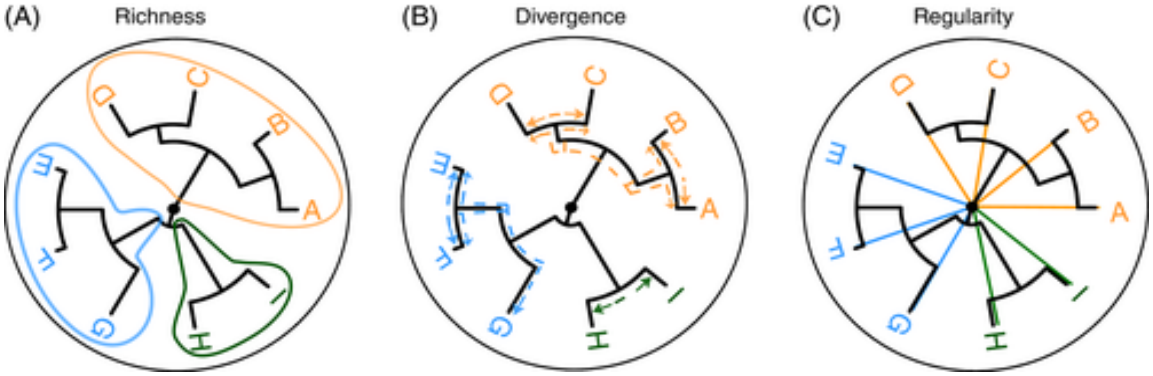
Phylogenetically neutral diversities:

- Richness (observed, Chao1, ACE)
- Evenness (Pielou's evenness)
- Diversity (inverse Simpson, Shannon)

Phylogeny-aware diversities:

- Faith diversity index

Phylogenetic diversity indices



A guide to phylogenetic metrics for conservation, community ecology and macroecology

Caroline M. Tucker, Marc W. Cadotte, Silvia B. Carvalho, T. Jonathan Davies, Simon Ferrier, Susanne A. Fritz, Rich Grenyer, Matthew R. Helmus, Lanna S. Jin ... See all authors

First published: 20 January 2016 | https://doi.org/10.1111/brv.12252 | Cited by: 147

Inverse Simpson

$$DI = \frac{N(N-1)}{\sum n(n-1)}$$

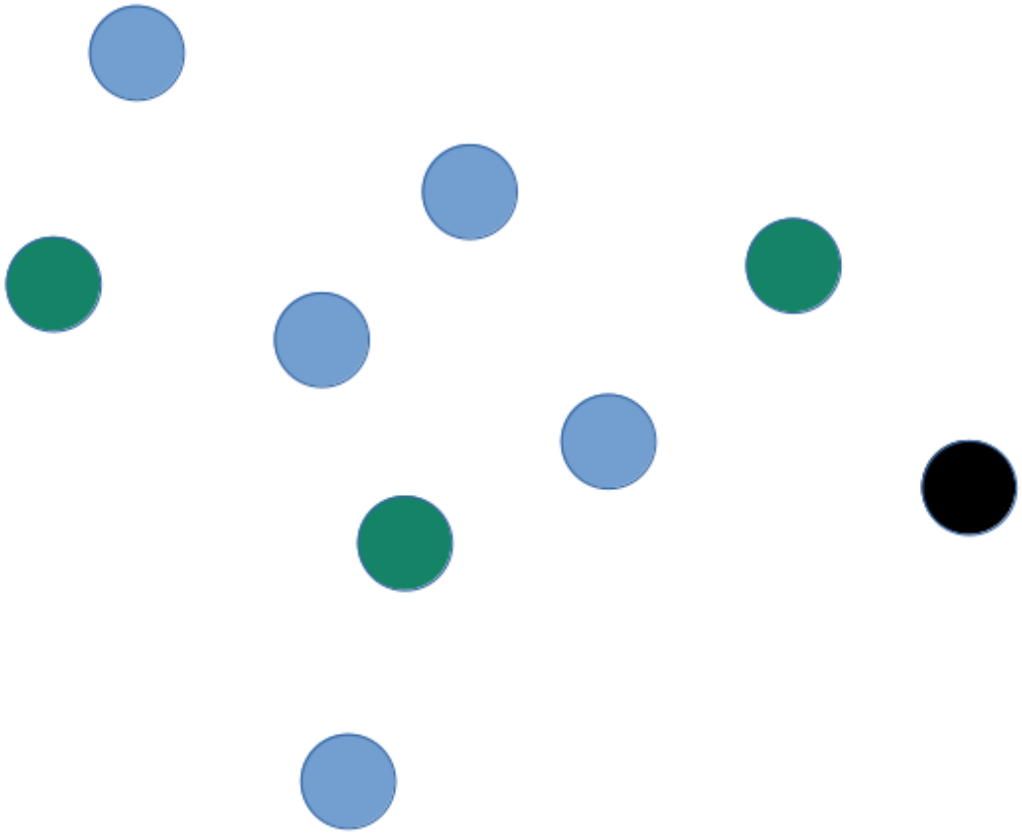
KEY

N = Total number of individuals collected

n_i = Number of individuals of a species

DI = Simpson Diversity Index

How likely it is to pick two members of the same species at random?



Inverse Simpson

Beware the variants:

- Simpson (λ)
- reciprocal Simpson ($1 - \lambda$)
- inverse Simpson ($\frac{1}{\lambda}$)

Shannon diversity

Shannon Index:

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

True Richness:

$$exp(H)$$

True diversity, or the effective number of types, refers to the number of equally abundant types needed for the average proportional abundance of the types to equal what is observed in the dataset of interest.

Evenness

H / ln(S)

- H: Shannon diversity
- S: Species richness

Hill's Diversity as a unifying concept

$$^{q}D = (\sum_{i}^{R} p_i^q)^{\frac{1}{1-q}} \tag{1}$$

Hill's alpha diversities

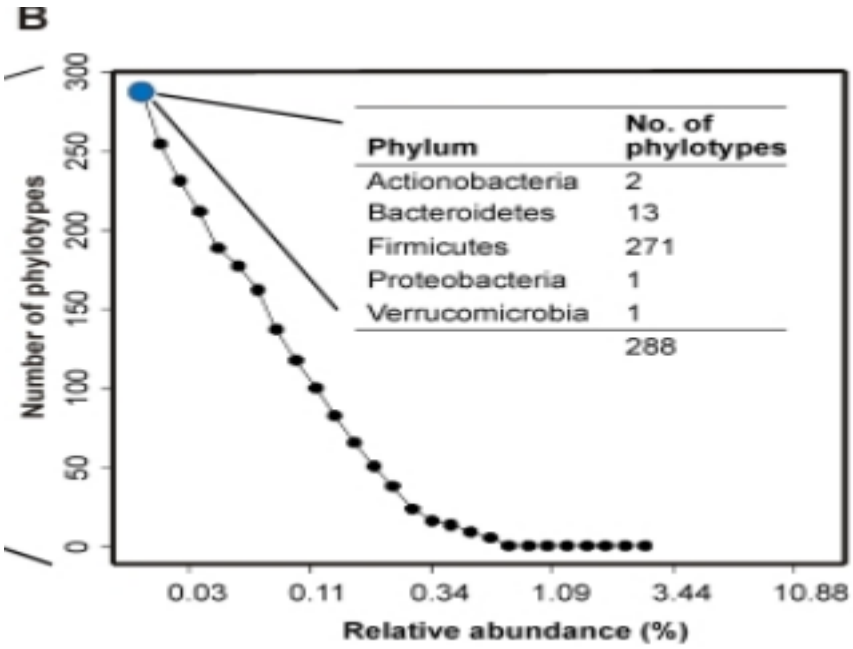
R: richness (number of distinct types)

pi: proportion of type I

Order of diversity:

- q = 0 : Species Richness
- q = 1 : Shannon diversity
- q = 2 : (Inverse) Simpson diversity
- q 1 : Renyi entropy

Hill's Diversity as a unifying concept



Hill's alpha diversities

- Richness
- inverse Simpson
- Shannon