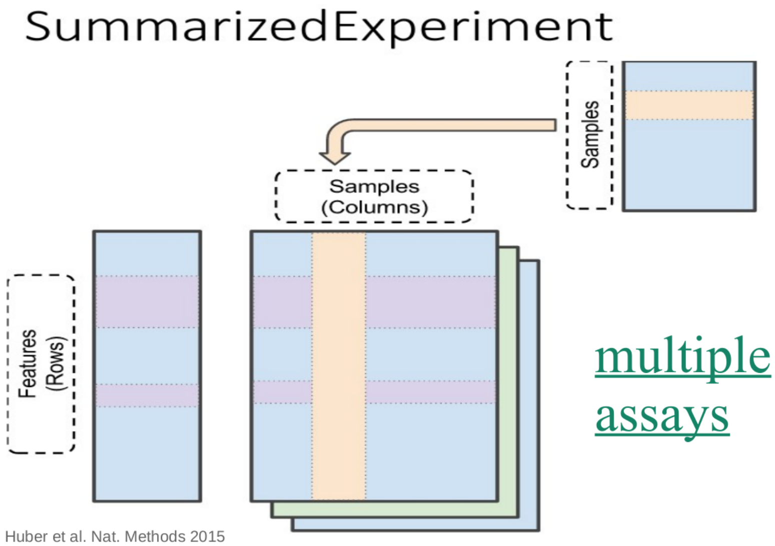
Tabular data analysis

Leo Lahti

Recap of Day 1

- **Data**: *SummarizedExperiment*: rowData, colData, assays
- **Methods**: subsetting, transformations



multiple assays

Huber et al. Nat. Methods 2015

Day 1: Basic data wrangling

- reproducible data science workflow
- data import
- data containers
- data manipulation (subsetting, transformations)

Today's learning goals

Expanding multi-assay analyses:
TreeSummarizedExperiment

- augmenting the data (add diversities)
- data agglomeration & *alternative experiments*
- tree information: *rowTree, colTree*

Today's program

Morning: data wrangling

Afternoon: data visualizations

Data enrichment

Visualizing colData

Task: visualize the abundance of a specific microbial Species against the measurement Site

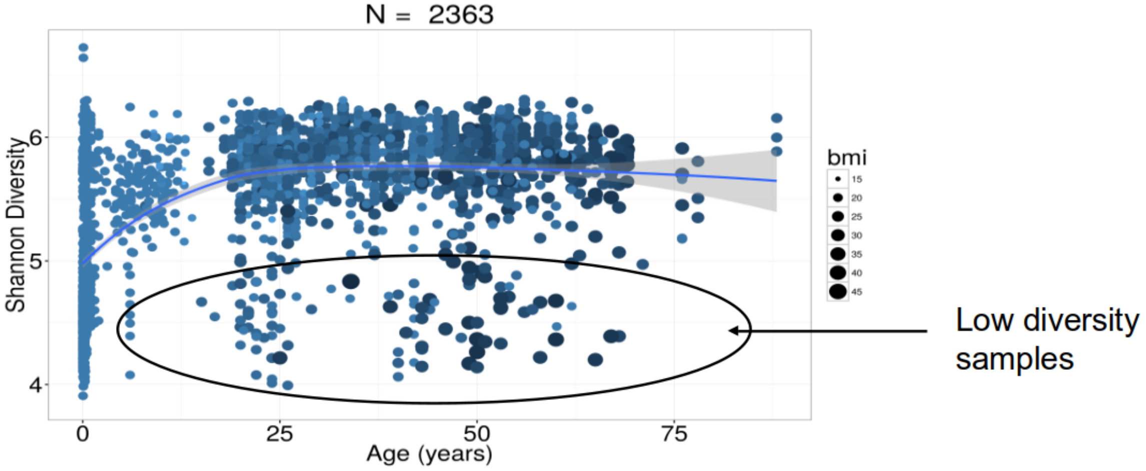
Alpha diversity task

Use the available tools to assess and visualize alpha diversity, and augment colData

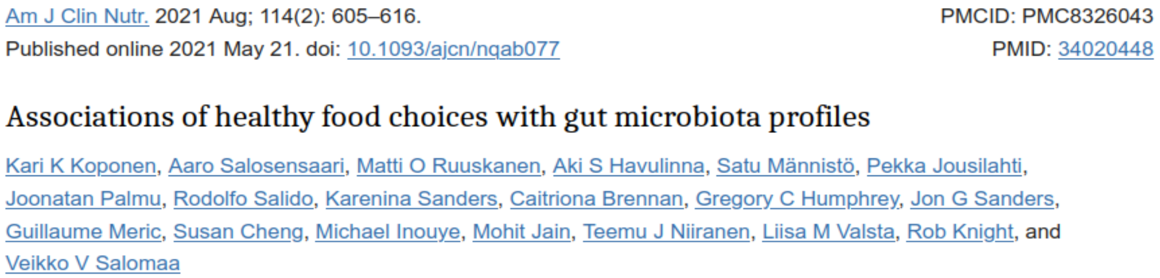
- Exercises 17.5.1-17.5.2
- Add Shannon diversity in colData
- Visualize diversity differences between sample groups

Alpha diversity & aging

Healthy & normal obese subjects.



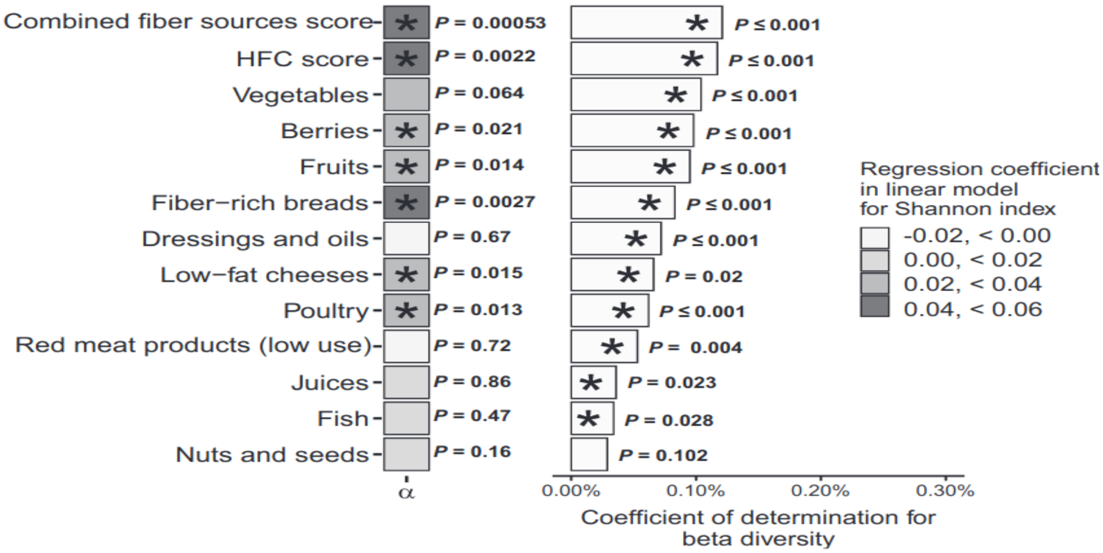
Alpha diversity and diet



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

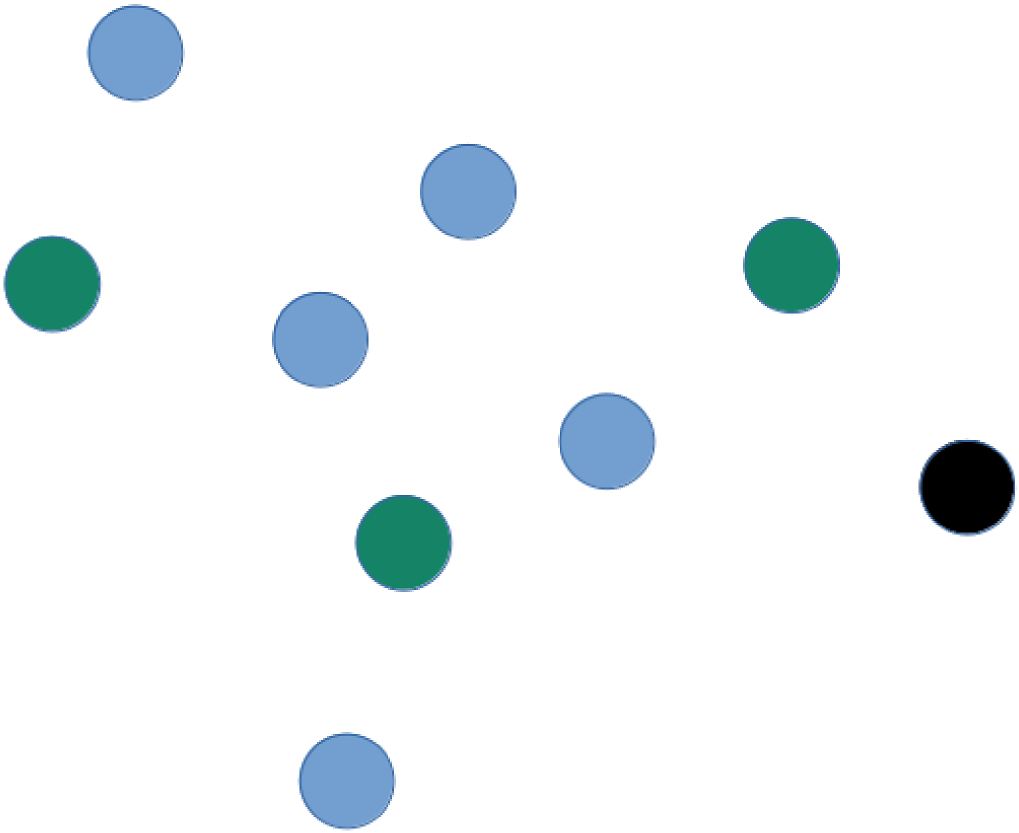
Associations of healthy food choices with gut microbiota profiles

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



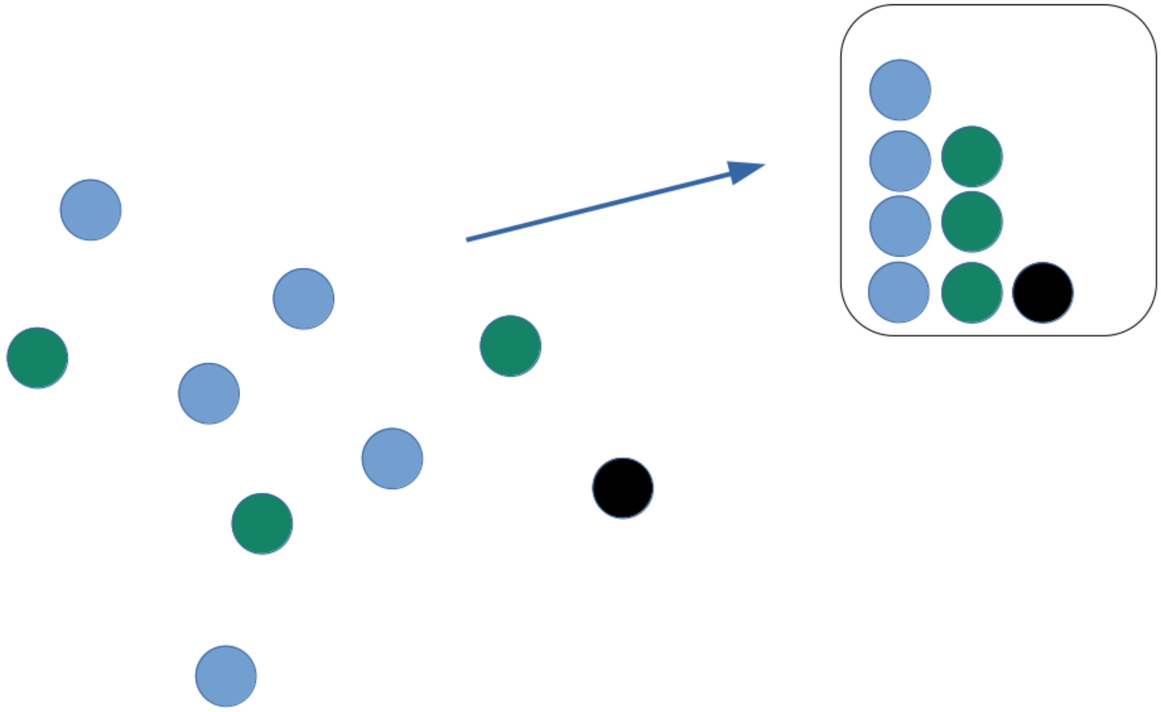
Alpha diversity

- 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

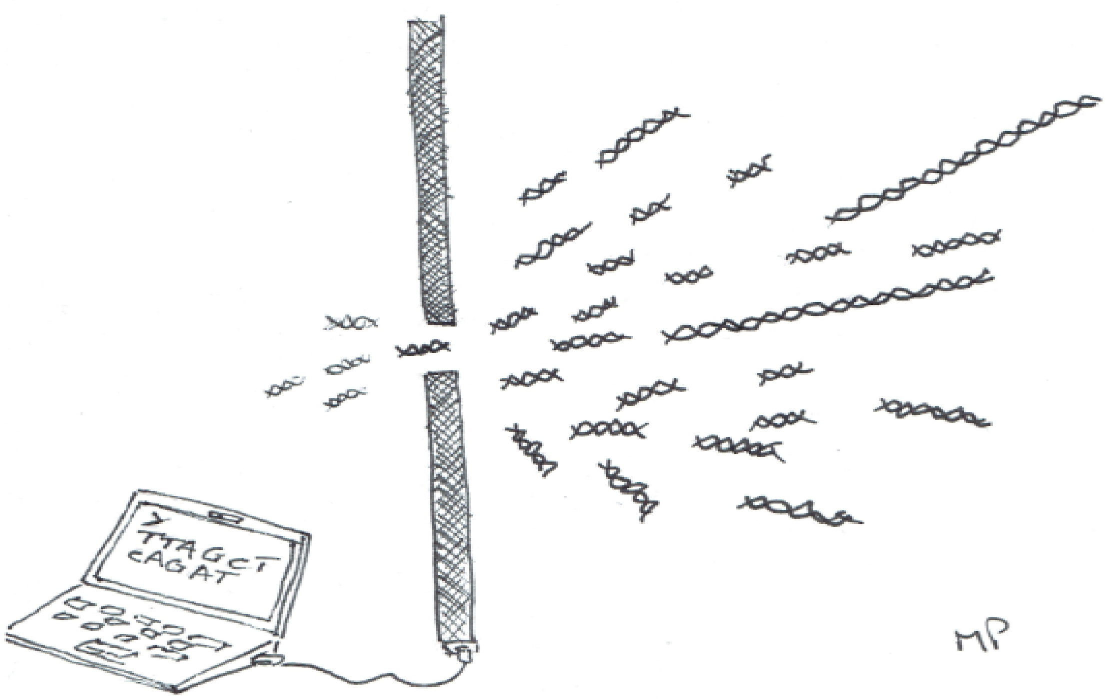
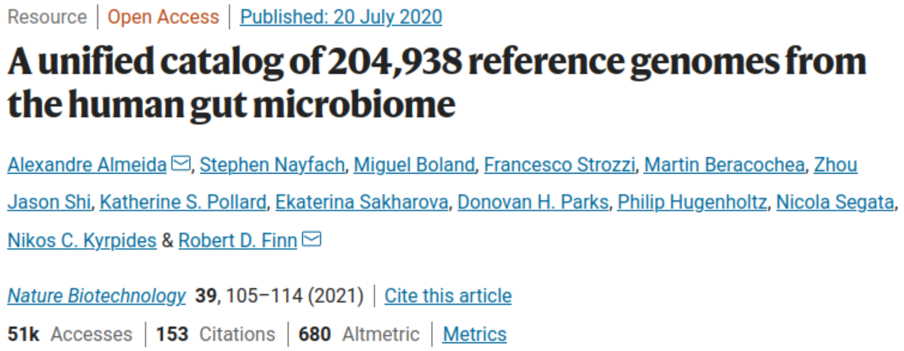


Figure 1: https://github.com/mblstamps/stamps2019/blob/master/STAMPS2019_overview_Pop.pdf

Finite sampling



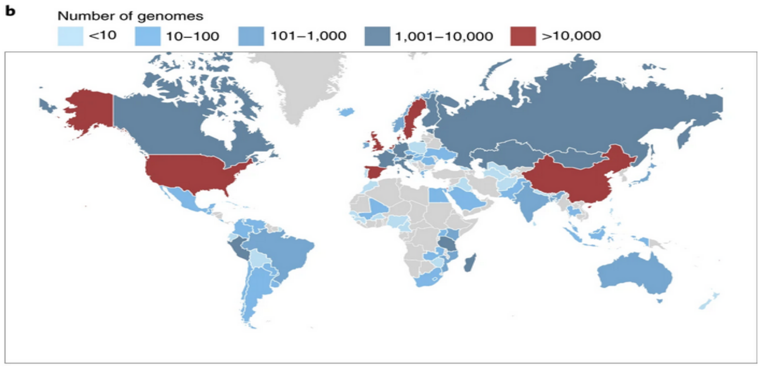
Resource | Open Access | Published: 20 July 2020

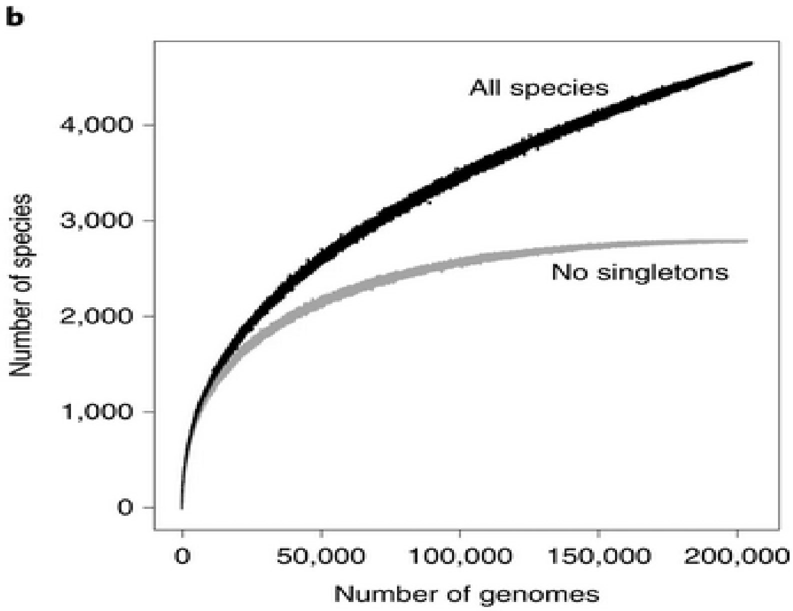
A unified catalog of 204,938 reference genomes from the human gut microbiome

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

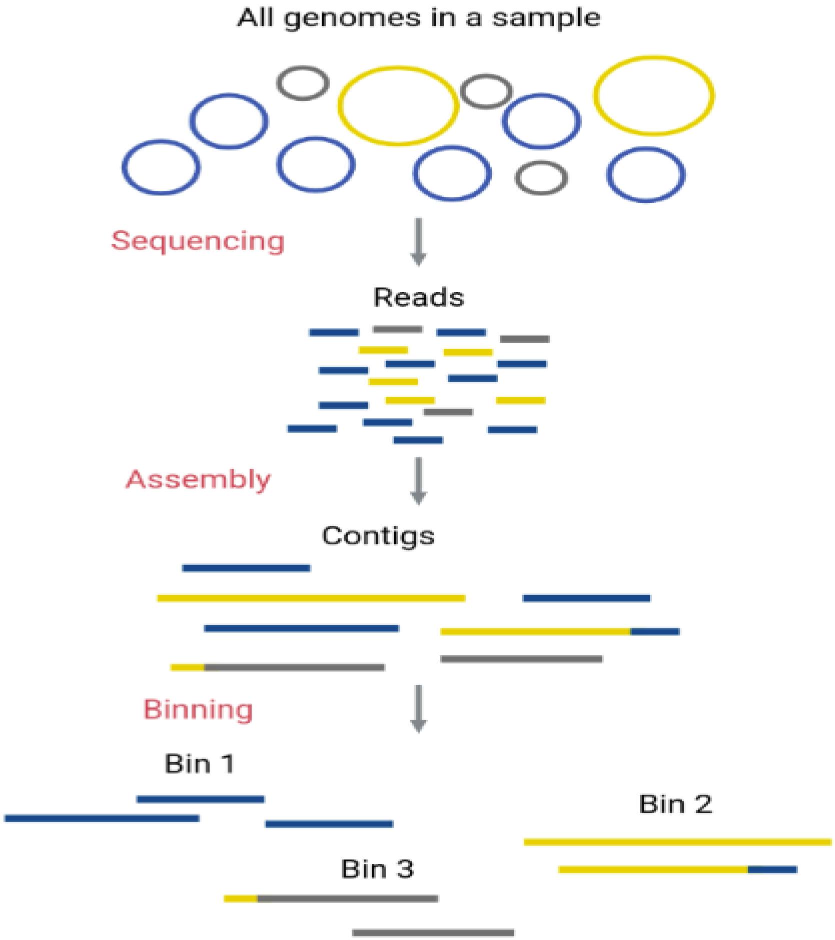
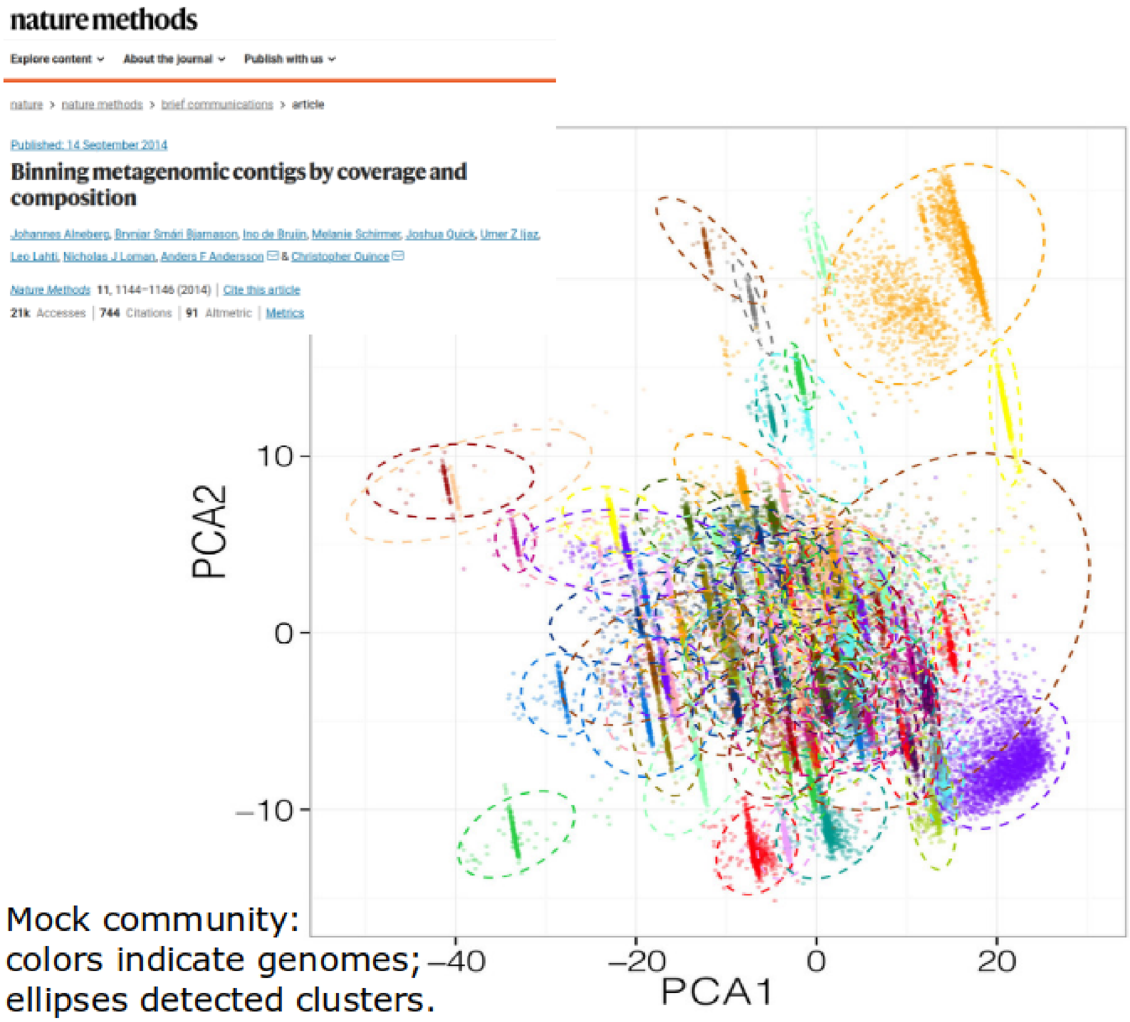


Figure 2: Copyright © Claudia Zirion, Diego Garfias, Vanessa Arellano, Aaron Jaime, Abel Lovaco, Daniel Díaz, Abraham Avelar, Nelly Sélem https://carpentries-incubator.github.io/metagenomics-workshop/)



nature methods

nature > nature methods > brief communications > article

Published: 14 September 2014

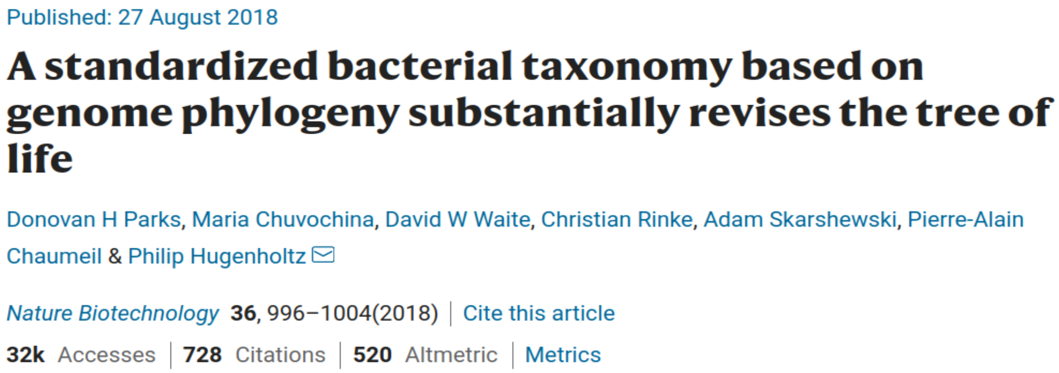
Binning metagenomic contigs by coverage and composition

Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson ✉ & Christopher Quince ✉

Nature Methods 11, 1144–1146 (2014) | Cite this article

21k Accesses | 744 Citations | 91 Altmetric | Metrics

Mock community: colors indicate genomes; ellipses detected clusters.



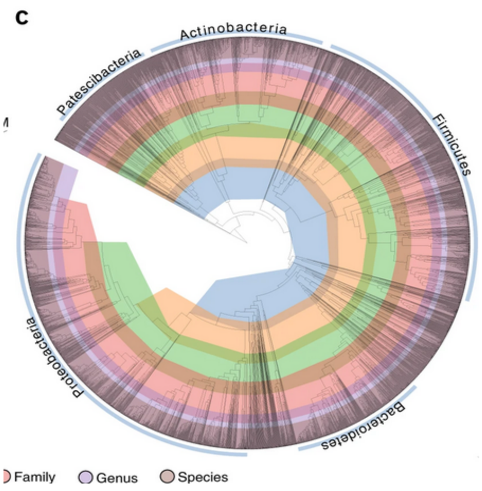
Published: 27 August 2018

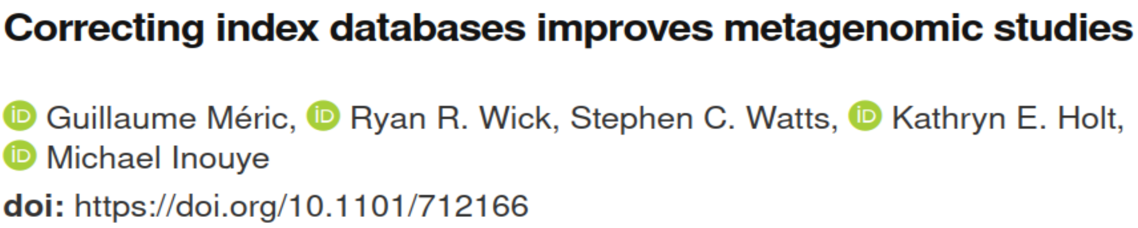
A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life

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

32k Accesses | **728** Citations | **520** Altmetric | Metrics

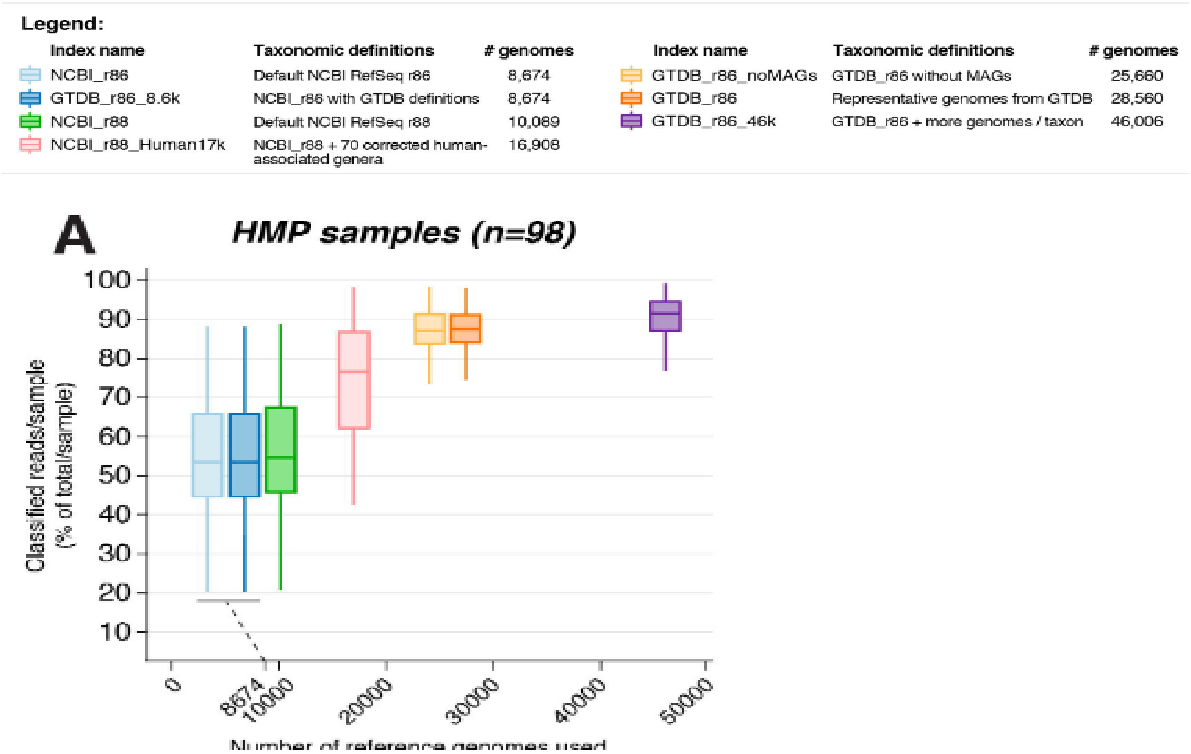




Correcting index databases improves metagenomic studies

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

A
HMP samples (n=98)
Classified reads/sample (% of total/sample)
100
90
80
70
60
50
40
30
20
10
0
8674
10000
20000
30000
40000
50000
Number of reference genomes used

Common alpha diversity indices

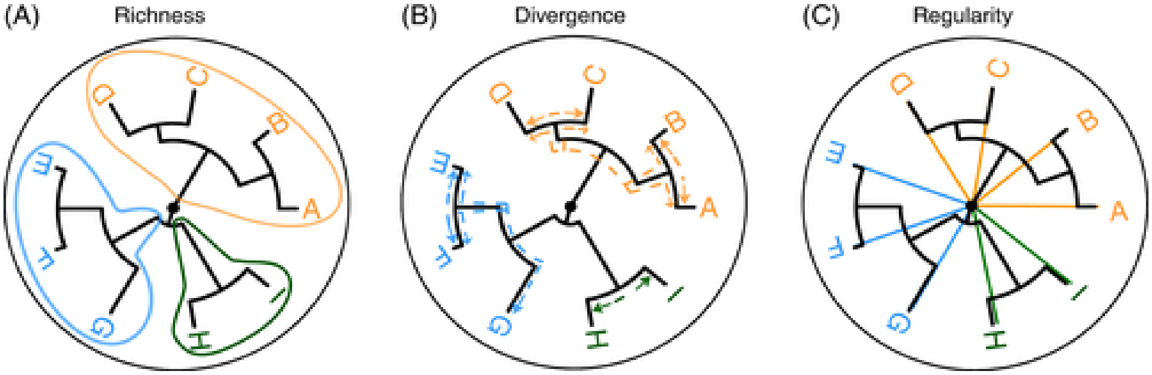
Phylogenetically neutral diversities:

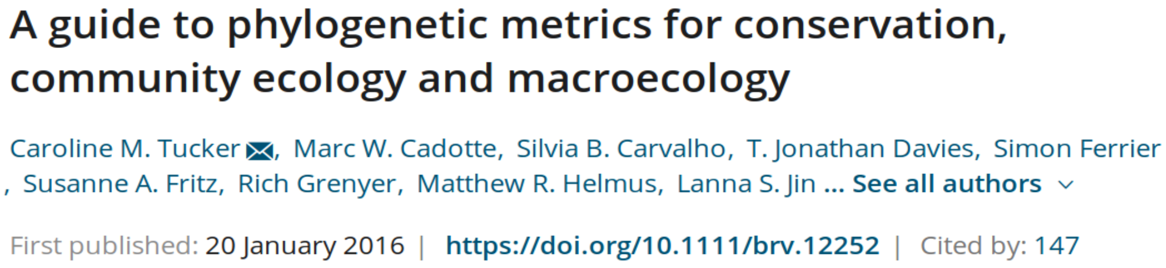
- Richness (observed, Chao1, ACE)
- Evenness (Pielou's evenness)
- Diversity (inverse Simpson, Shannon)

Phylogeny-aware diversities:

- Faith diversity index

Phylogenetic diversity indices



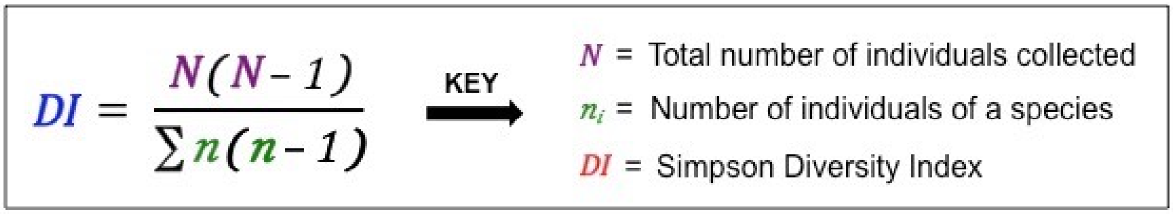


A guide to phylogenetic metrics for conservation, community ecology and macroecology

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



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

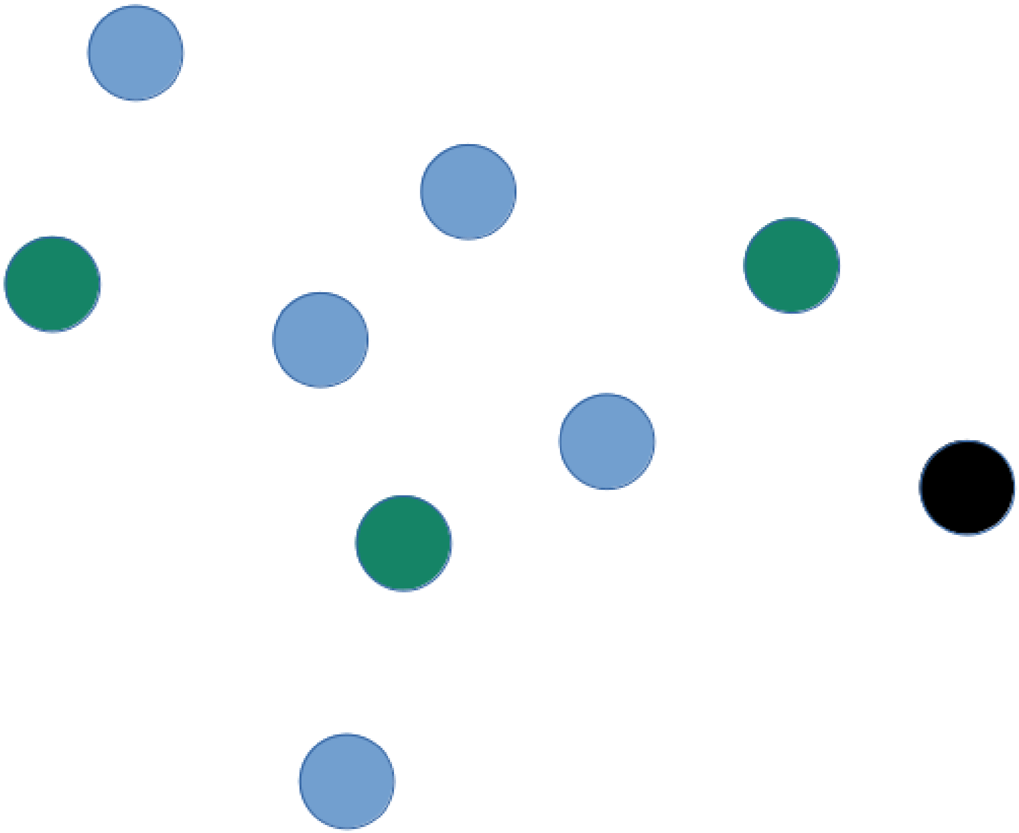
KEY

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:

$$exp(\mathrm{H})$$

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

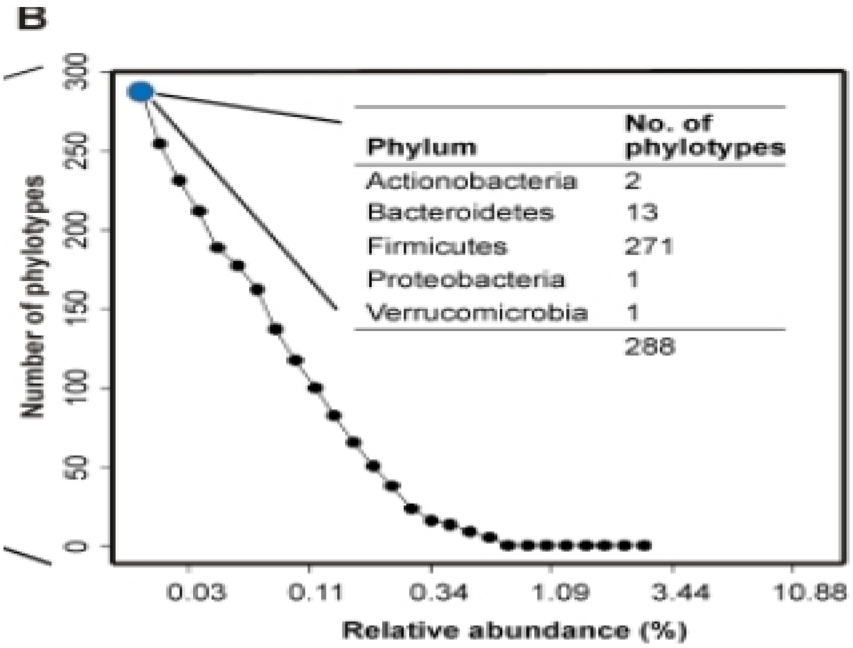
R: richness (number of distinct types)

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



Phylum	No. of phylotypes
Actionobacteria	2
Bacteroidetes	13
Firmicutes	271
Proteobacteria	1
Verrucomicrobia	1
	288

Hill’s alpha diversities

- Richness
- inverse Simpson
- Shannon

Data wrangling

Basic data operations

- Transform

- Subset
- Merge
- **Aggregate**
- Split

Subsetting

Load example data set:

```r
library(mia)
```

```
Loading required package: SummarizedExperiment
```

```
Loading required package: MatrixGenerics
```

```
Loading required package: matrixStats
```

```

Attaching package: 'MatrixGenerics'
```

```
The following objects are masked from 'package:matrixStats':

    colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
    colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
    colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
    colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
    colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
    colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
    colWeightedMeans, colWeightedMedians, colWeightedSds,
    colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
    rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
    rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
    rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
    rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
    rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
    rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
    rowWeightedSds, rowWeightedVars
```

```
Loading required package: GenomicRanges

Loading required package: stats4

Loading required package: BiocGenerics


Attaching package: 'BiocGenerics'


The following objects are masked from 'package:stats':

    IQR, mad, sd, var, xtabs


The following objects are masked from 'package:base':

    anyDuplicated, aperm, append, as.data.frame, basename, cbind,
    colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
    get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
    match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
    Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
    table, tapply, union, unique, unsplit, which.max, which.min


Loading required package: S4Vectors


Attaching package: 'S4Vectors'


The following objects are masked from 'package:base':

    expand.grid, I, unname


Loading required package: IRanges

Loading required package: GenomeInfoDb

Loading required package: Biobase
```

```
Welcome to Bioconductor

    Vignettes contain introductory material; view with
    'browseVignettes()'. To cite Bioconductor, see
    'citation("Biobase")', and for packages 'citation("pkgname")'.


Attaching package: 'Biobase'

The following object is masked from 'package:MatrixGenerics':

    rowMedians

The following objects are masked from 'package:matrixStats':

    anyMissing, rowMedians

Loading required package: SingleCellExperiment

Loading required package: TreeSummarizedExperiment

Loading required package: Biostrings

Loading required package: XVector


Attaching package: 'Biostrings'

The following object is masked from 'package:base':

    strsplit

Loading required package: MultiAssayExperiment
```

```r
data(GlobalPatterns)
tse <- GlobalPatterns
```

Check dimension:

```
dim(tse)
```

```
[1] 19216    26
```

Check dimension for a subset:

```
dim(tse[1:10, 1:3])
```

```
[1] 10  3
```

Transformations

- Presence/absence
- Compositional (percentages)
- Log_{10}
- CLR and other *Aitchison* transformations
- Phylogenetic transformations (e.g. philr)
- Custom transformations

Transformations

Task: Alternative assays

- visualize transformed data; histograms, boxplots
- compare different transformations (scatterplot?)

Agglomeration

- taxonomic units
- TreeSE objects

Agglomeration

Agglomerate microbiota data to higher taxonomic levels:

- chapter 6.3
- agglomerateByRank
- compare diversity or prevalent features between levels

Alternative experiments

Alternative assays vs. alternative experiments?

- Store agglomerated data: *altExp*
- Do all levels at once: *splitByRanks*

Splits

Splitting by:

- taxonomic units
- sample or feature groups

Taxonomic ranks & *altExp*

The alternative experiments (*altExp*) mechanism allows us to include multiple abundance tables at different taxonomic levels.

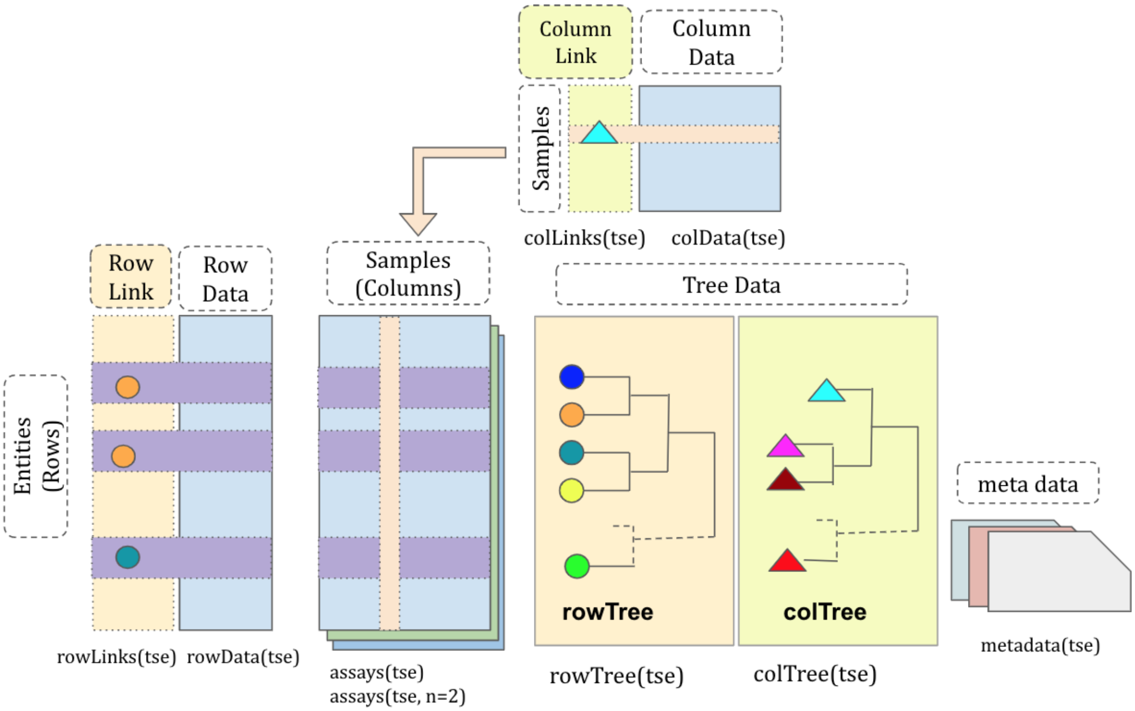
Option	Rows (features)	Cols (samples)	Recommendation
assays	match	match	Data transformations
altExp	free	match	Alternative experiments
MultiAssay	free	free (mapping)	Multi-omic experiments

Alternative experiments and assays?

- Pick clr assay from Genus-level data table?
- Compare Shannon diversity from Genus and Species levels?

TreeSummarizedExperiment

Huang et al. F1000, 2021



Visualization

Ordination

- Visualize example data with PCoA using Bray-Curtis dissimilarity
- Visualize example data with PCA using Aitchison distance (CLR + Euclid)

Heatmaps

- Visualize abundance variation for selected taxa on a heatmap

Trees

- Visualize phylogenetic tree using the examples