# Heterogeneous Data

Susan Holmes,@SherlockpHolmes

STAMPS, Woods Hole, July 25, 2019

# (Big) Challenges for those of us working from the ground up

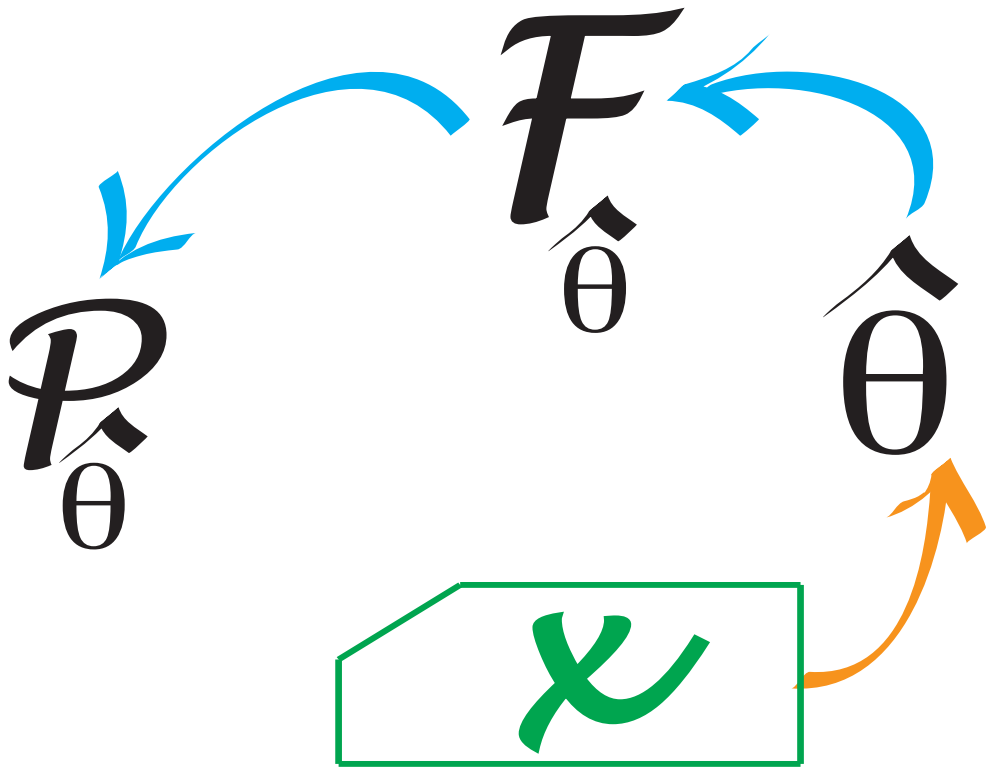
- Heterogeneity.
- Heteroscedasticity.
- Information Leaks.
- Multiplicity of Choices.
- Reproducibility.

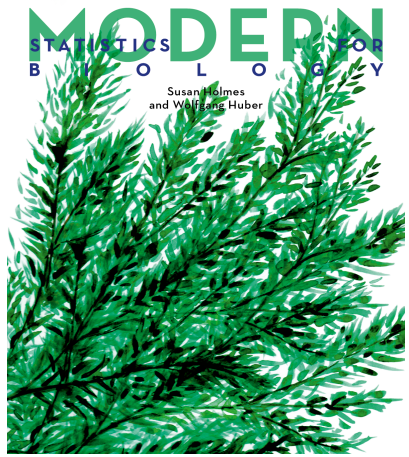
# Heterogeneous data

# Heterogeneity of Data

- Status : response/ explanatory.
- Hidden (latent)/measured.
- Types :
  - Continuous
  - Binary, categorical
  - Graphs/ Trees
  - Images
  - Spatial Information
  - Rankings
- Amounts of dependency: independent/time series/spatial.
- Different technologies used ( Illumina, MassSpec, NMR, RNA-seq).

# Statistics: separate the model from the data




See a complete book:

`http://bios221.stanford.edu/book/`

# Read data are counts, the **data** are not compositional.... *the parameters are!*

- After perturbations amounts of bacteria go up & down.
- Remove contaminants using read numbers (`decontam` and `BARBIE`).
- Estimating depth bias requires read numbers.
- Some bacteria live in symbiosis with others.
- We need the read depths for variability/standard error estimation and uncertainty quantification.
- Data transformations can be used to remove ”multiplicative error” and equalize the variance.

# Glossary (statistical)

- Probabilistic.
- Statistical.
- Estimation.
- Parameter.
- Parametric.
- Nonparametric.
- Independent / dependent.
- Bias.

# Glossary (computational)

**vector** $v[1] \quad v[2] \quad v[3] \quad .....v[n]$

**matrix** The dimension of A here is 4 by 3.

$$\begin{array}{lll} A[1,1] & A[1,2] & A[1,3] \\ A[2,1] & A[2,2] & A[2,3] \\ A[3,1] & A[3,2] & A[3,3] \\ A[4,1] & A[4,2] & A[4,3] \end{array}$$

**factor** A categorical variable with levels: "O", "AB", "B" etc. See here `http://web.stanford.edu/class/bios221/book/Chap-Generative.html`

**list** A container for different objects, usually of different type and dimensions.

**data.frame** A list whose components are of the variables, usually of different types.

**status** response/explanatory: the "formula" `y` $\sim$ `x`.

# Mix and Match data types

- Matrix : Has to be either all numerics or all characters.
- Mixed type: `data.frame`, a special type of list, the name of each component is the variable name.
- Complex list: S4 object with special components called ”slots”.

# Resources for dealing with different data types and structures in R

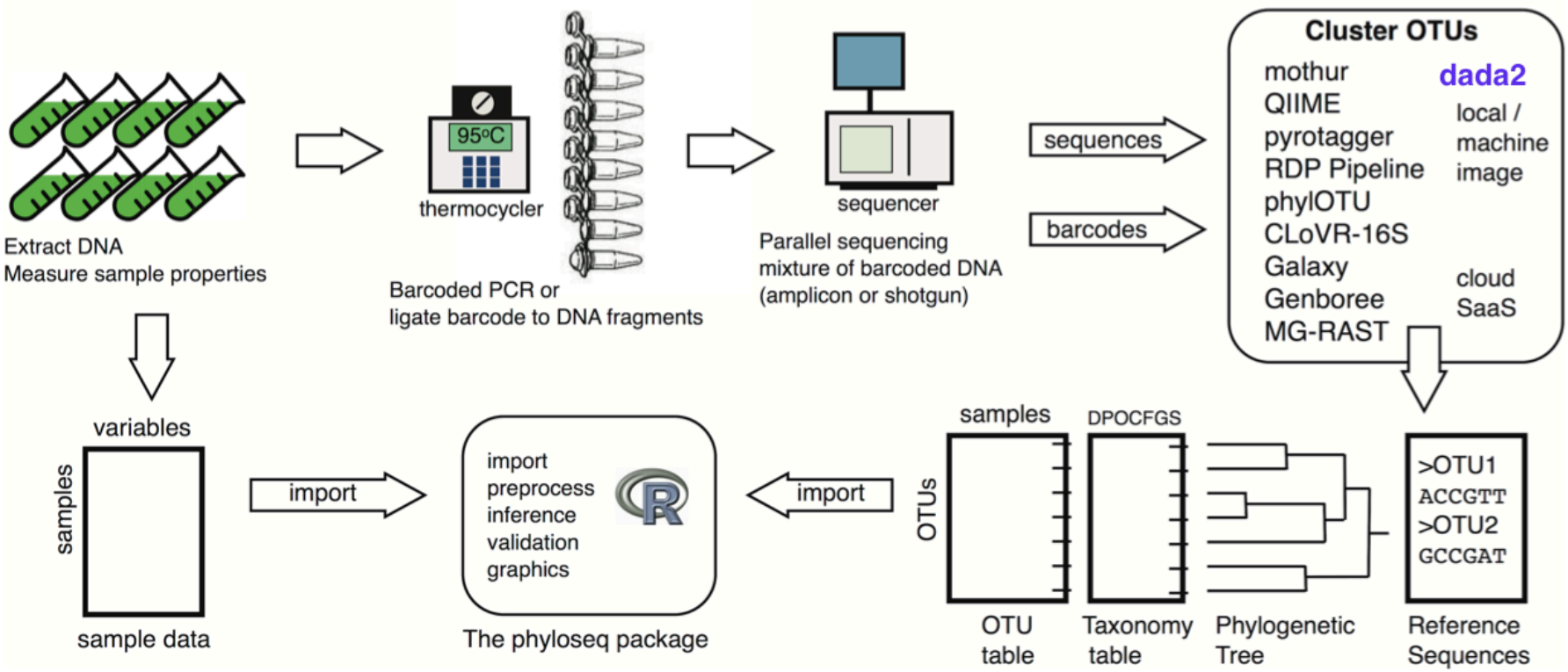
- If you've never used R:
  (a lecture 0 link here: `Lab0-DirectoryFiles.html`).
- About data types: `LabDataTypes.html`
- Hadley's chapter on data structures:
  `http://adv-r.had.co.nz/Data-structures.html`
- Advanced resource: the special S4 data types
  `https://stuartlee.org/post/content/post/2019-07-09-s4-a-short-guide-for-perplexed/`.

# Human Microbiome: What are the data?

| | |
|---|---|
| DNA | A biomarker DNA count (16sRNA-gene especially). |
| DNA | All the Genomic material present (shotgun metagenomics). |
| RNA | What genes are being turned on (gene expression), transcriptomics. |
| Mass Spec | Specific signatures of chemical compounds present. |
| Clinical | Multivariate information about patients' clinical status, medication, weight. |
| Environmental | Location, nutrition, time. |
| Domain Knowledge | Metabolic networks, phylogenetic trees, gene ontologies. |

# phyloseq

Joey McMurdie (joey711 on github).
Available in Bioconductor.

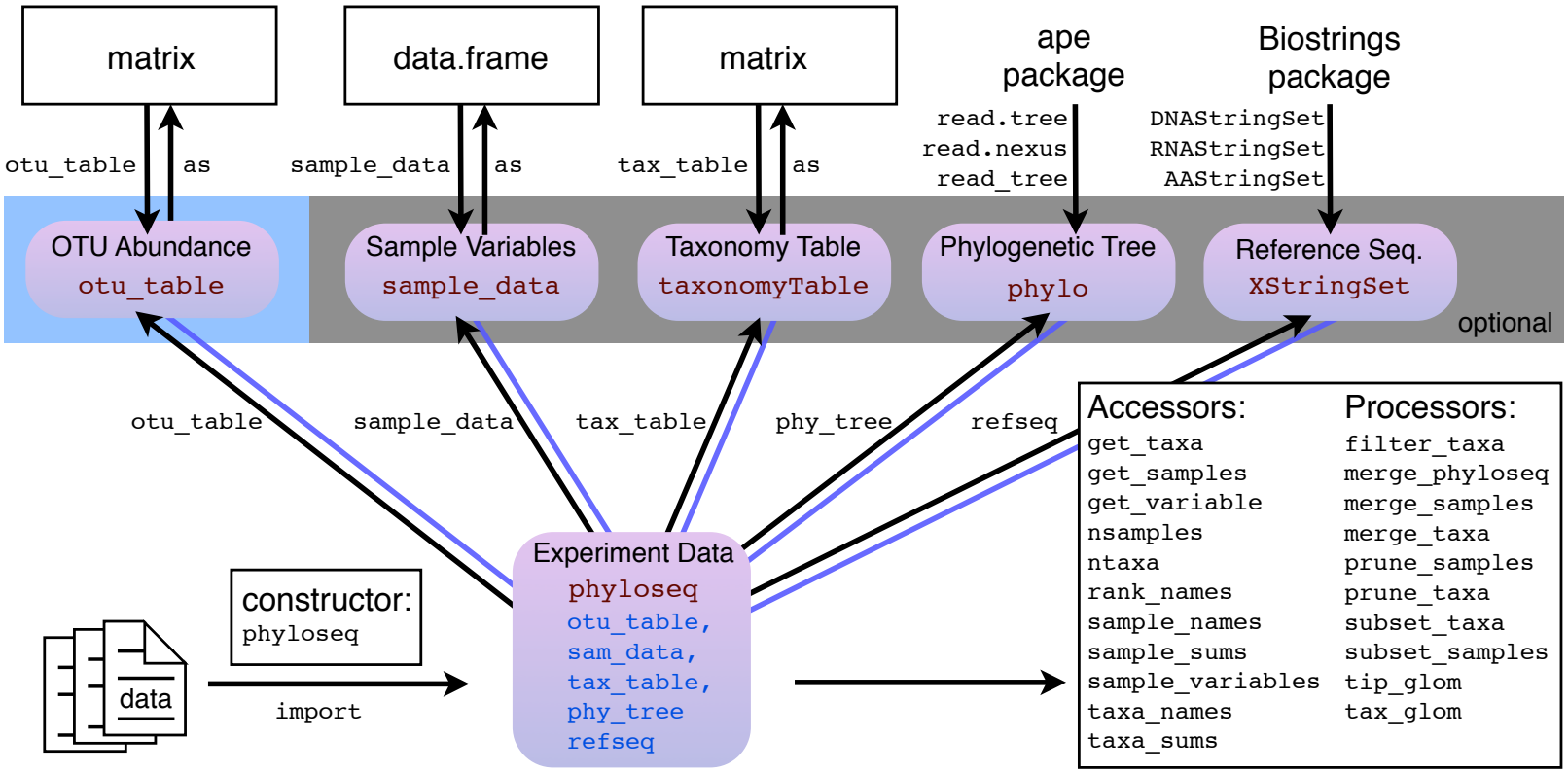



## Heterogeneous Data Objects

Input and data manipulation with `phyloseq`
(McMurdie and Holmes, 2013, Plos ONE)
As always in R: object oriented data.

# phyloseq

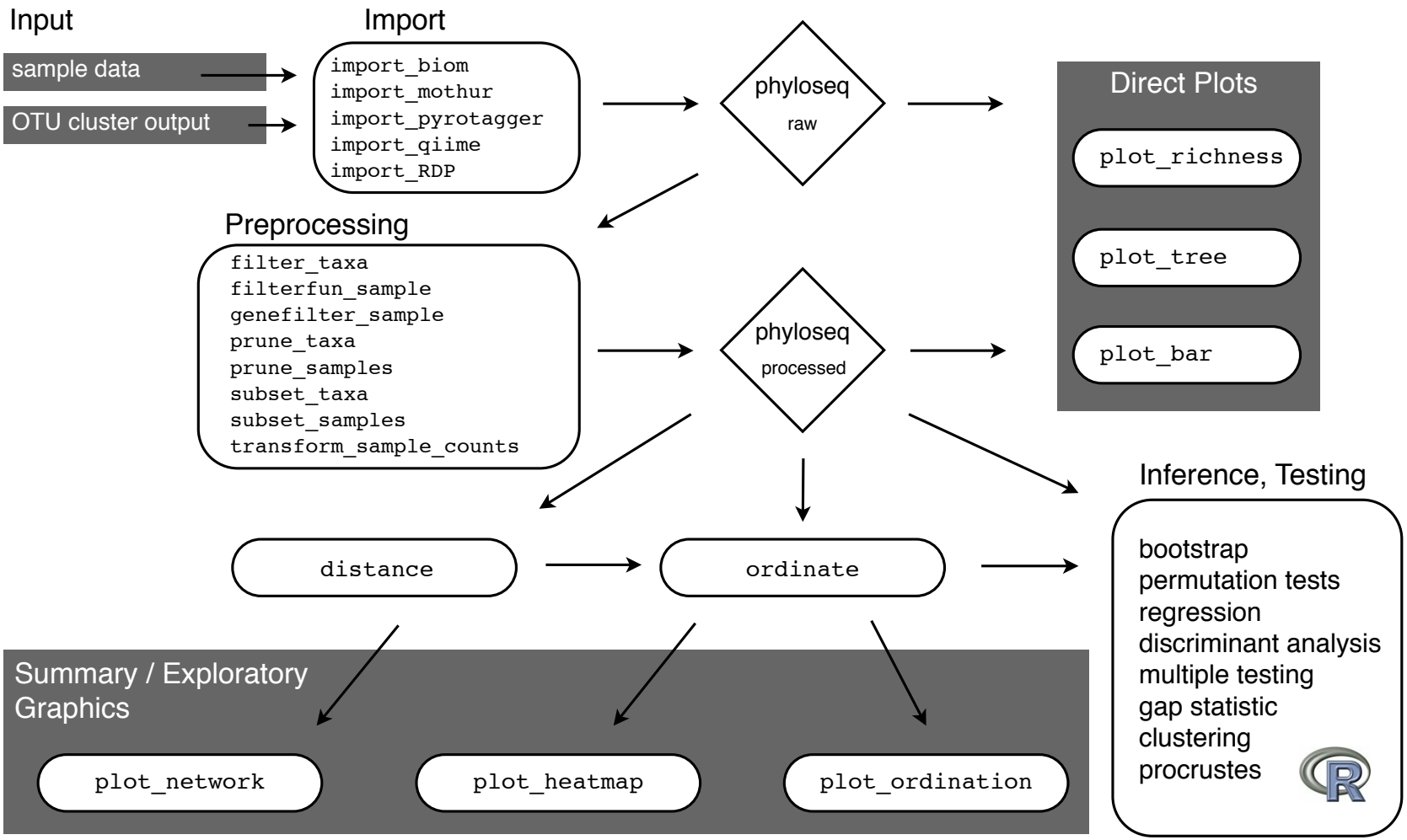
data structure & API




http://joey711.github.io/phyloseq/

# phyloseq

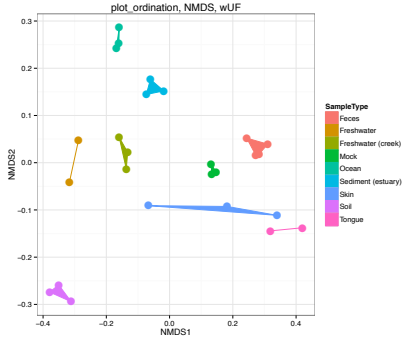
work flow

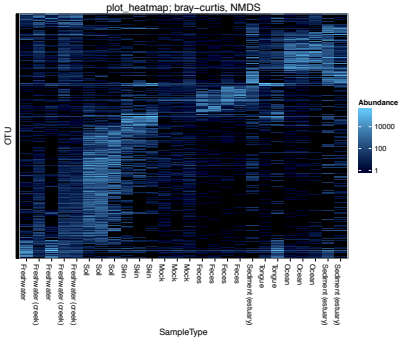


# phyloseq

graphics

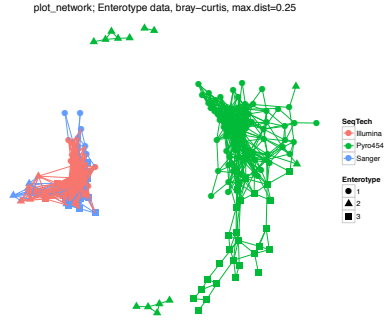
plot_ordination()

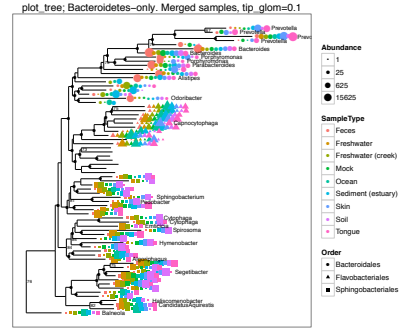



plot_heatmap()

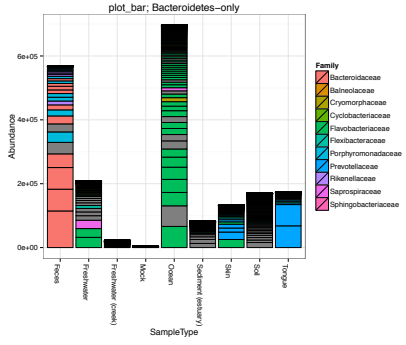
plot_network()

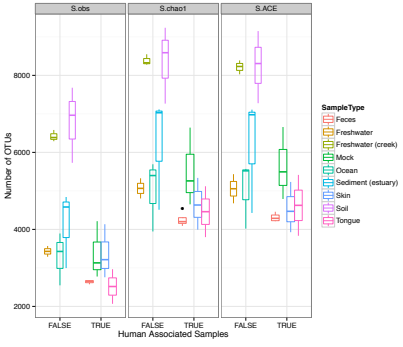



plot_tree()

plot_bar()



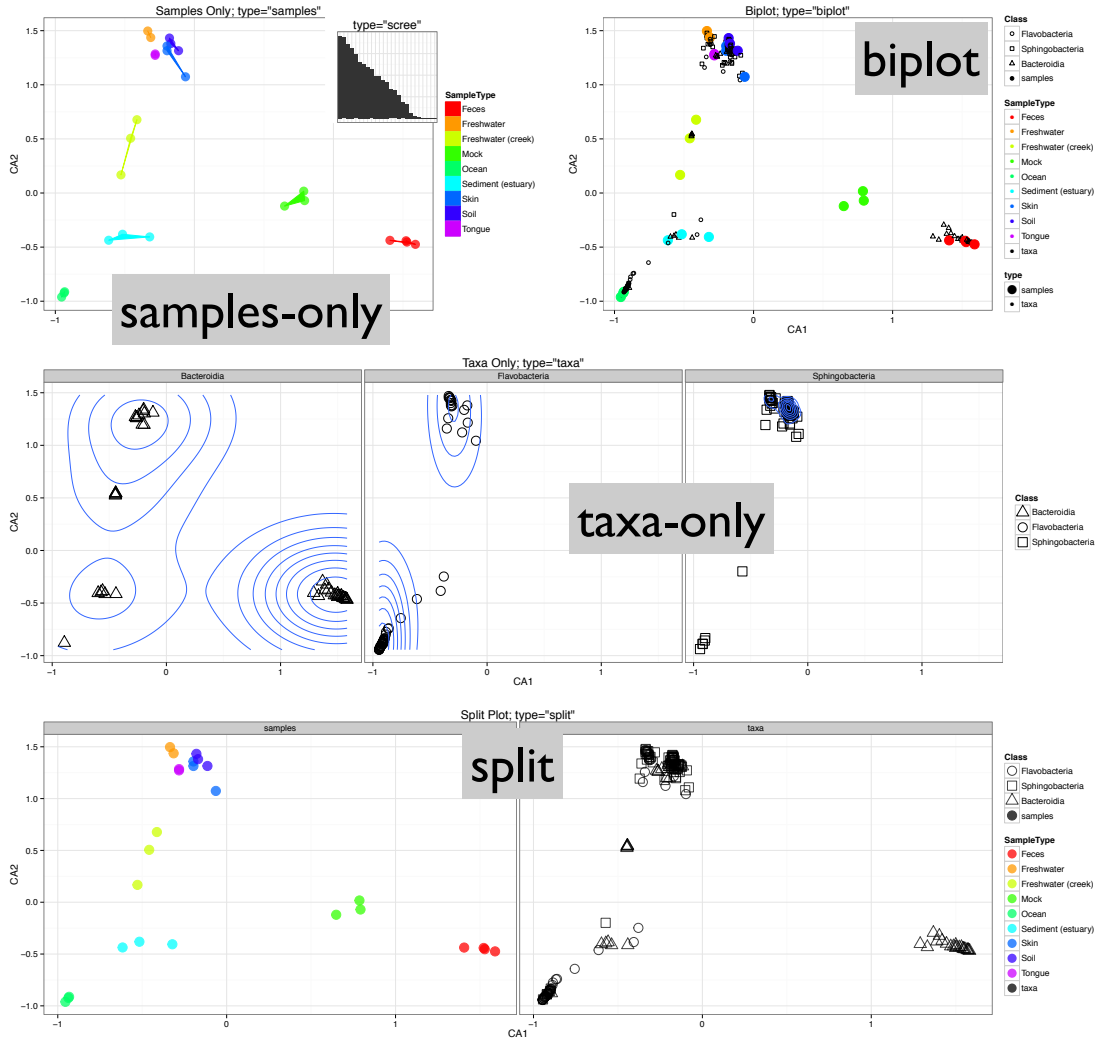




plot_richness()

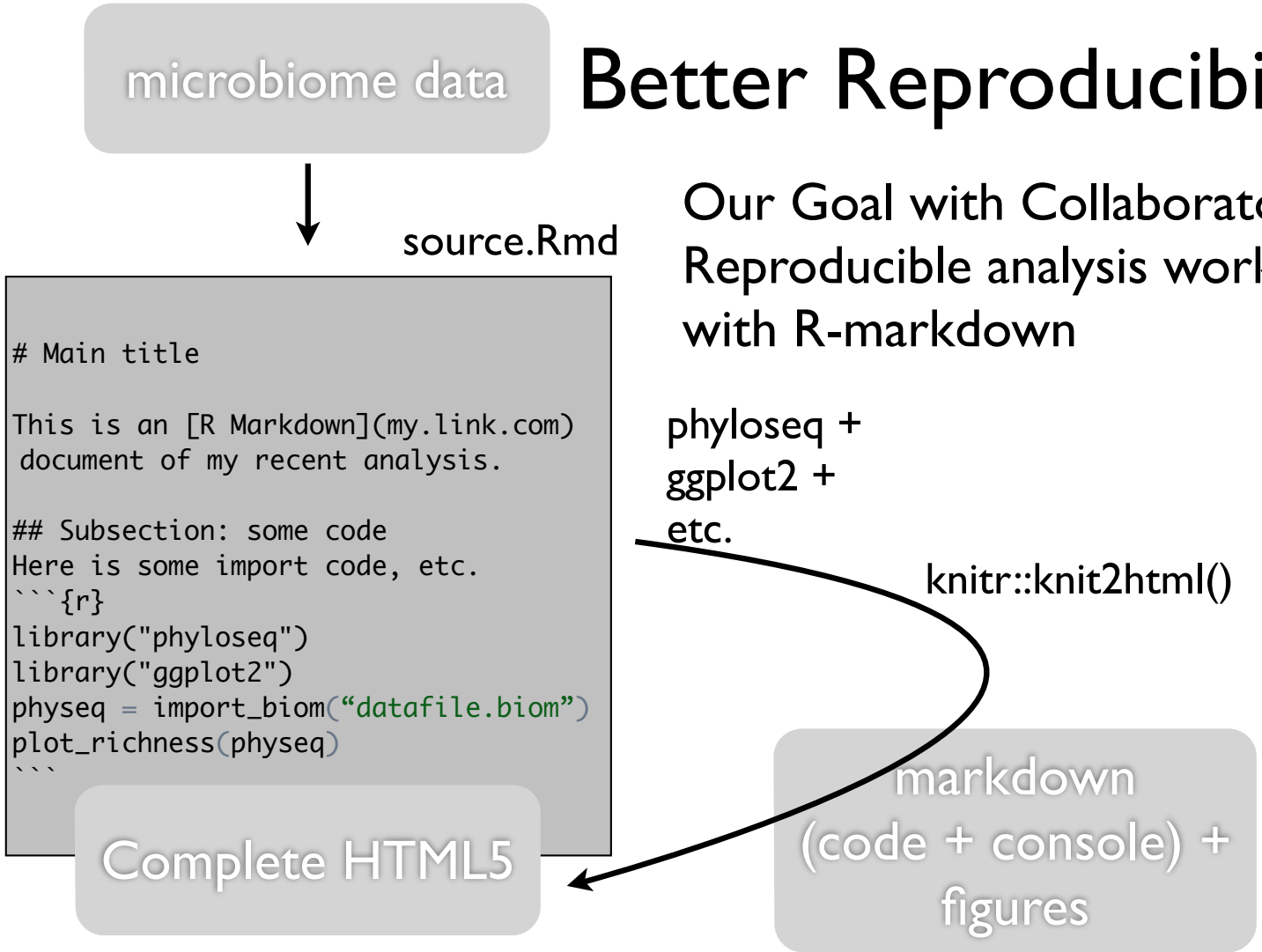
# phyloseq

graphics

plot_ordination()



# Better Reproducibility



# Part I

## An Example

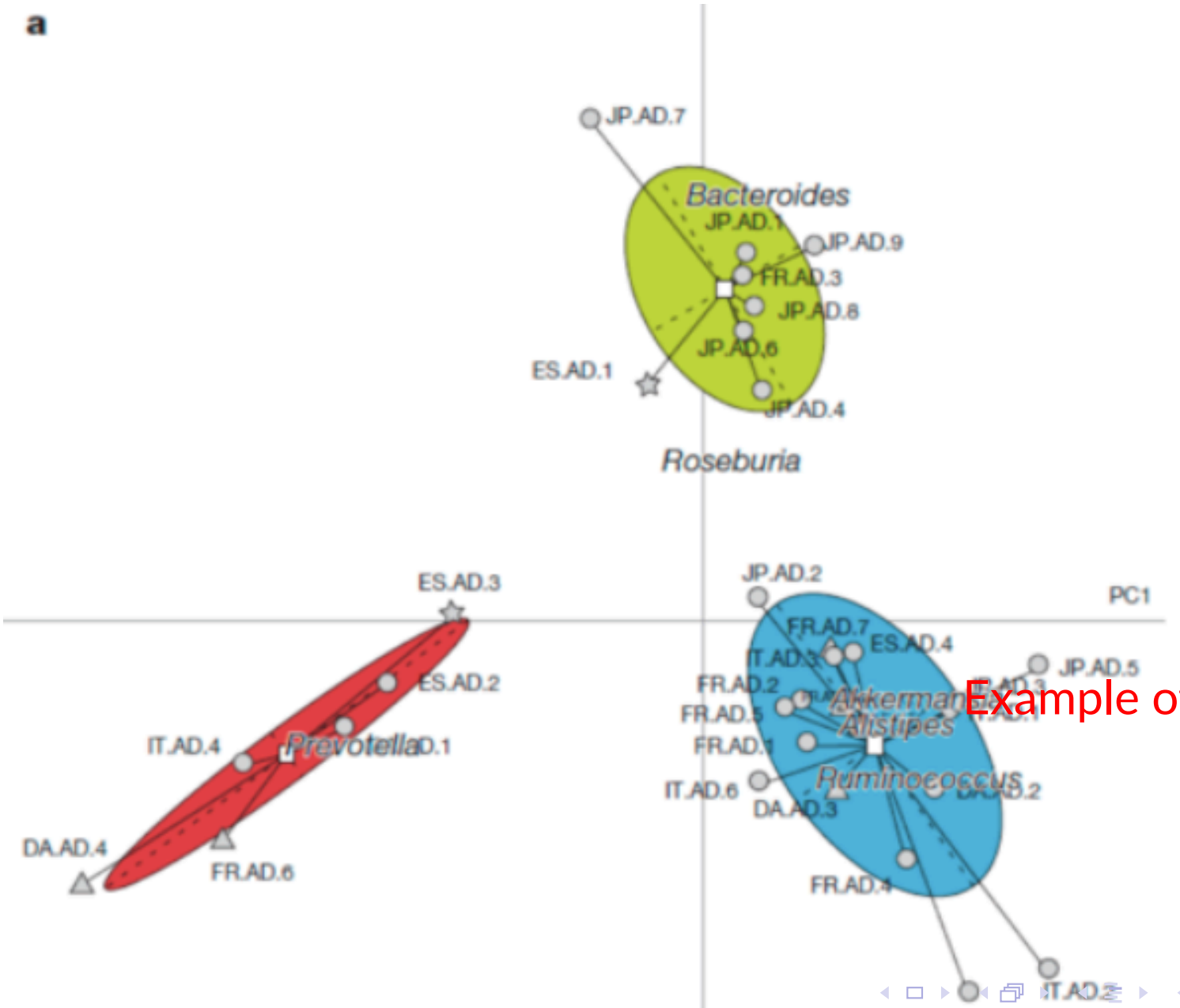
ARTICLE

doi:10.1038/nature09944

# Enterotypes of the human gut microbiome

Manimozhiyan Arumugam[1]*, Jeroen Raes[1,2]*, Eric Pelletier[3,4,5], Denis Le Paslier[3,4,5], Takuji Yamada[1], Daniel R. Mende[1], Gabriel R. Fernandes[1,6], Julien Tap[1,7], Thomas Bruls[3,4,5], Jean-Michel Batto[7], Marcelo Bertalan[8], Natalia Borruel[9], Francesc Casellas[9], Leyden Fernandez[10], Laurent Gautier[8], Torben Hansen[11,12], Masahira Hattori[13], Tetsuya Hayashi[14], Michiel Kleerebezem[15], Ken Kurokawa[16], Marion Leclerc[7], Florence Levenez[7], Chaysavanh Manichanh[9], H. Bjørn Nielsen[8], Trine Nielsen[11], Nicolas Pons[7], Julie Poulain[3], Junjie Qin[17], Thomas Sicheritz-Ponten[8,18], Sebastian Tims[15], David Torrents[10,19], Edgardo Ugarte[3], Erwin G. Zoetendal[15], Jun Wang[17,20], Francisco Guarner[9], Oluf Pedersen[11,21,22,23], Willem M. de Vos[15,24], Søren Brunak[8], Joel Doré[7], MetaHIT Consortium†, Jean Weissenbach[3,4,5], S. Dusko Ehrlich[7] & Peer Bork[1,25]

**Our knowledge of species and functional composition of the human gut microbiome is rapidly increasing, but it is still based on very few cohorts and little is known about variation across the world. By combining 22 newly sequenced faecal metagenomes of individuals from four countries with previously published data sets, here we identify three robust clusters (referred to as enterotypes hereafter) that are not nation or continent specific. We also confirmed the enterotypes in two published, larger cohorts, indicating that intestinal microbiota variation is generally stratified, not continuous. This indicates further the existence of a limited number of well-balanced host–microbial symbiotic states that might respond differently to diet and drug intake. The enterotypes are mostly driven by species composition, but abundant molecular functions are not necessarily provided by abundant species, highlighting the importance of a functional analysis to understand microbial communities. Although individual host properties such as body mass index, age, or gender cannot explain the observed enterotypes, data-driven marker genes or functional modules can**




Example of Study

# Summary of the study

- Choose the data transformation (here proportions replaced the original counts).
  ... log, rlog, subsample, prop, orig.
- Take a subset of the data, some samples declared as outliers. ... leave out 0, 1, 2 ,..,9, + criteria (10).....
- Filter out certain taxa (unknown labels, rare, etc...)
  ... remove rare taxa (threshold at 0.01%, 1%, 2%,...)
- Choose a distance.
  ... 40 choices in vegan/phyloseq.
- Choose an ordination method and number of coordinates.
  ... MDS, NMDS, k=2,3,4,5..
- Choose a clustering method, choose a number of clusters.
  ... PAM, KNN, density based, hclust ...
- Choose an underlying continuous variable (gradient or group of variables: manifold).
- Choose a graphical representation.

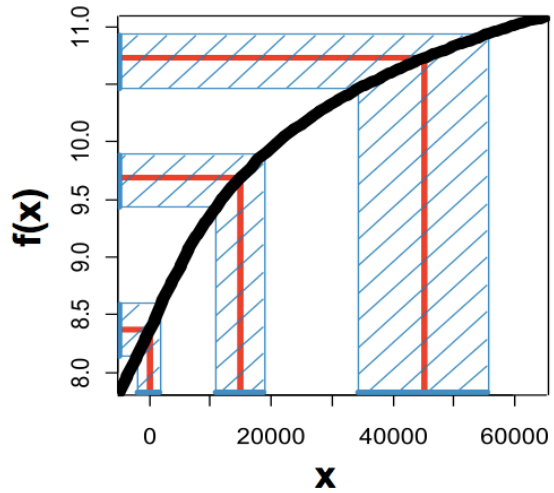
There are thus more than 200 million possible ways of analyzing this data:

$$5 \times 100 \times 10 \times 40 \times 8 \times 16 \times 2 \times 4 = 204800000$$

# Heteroscedasticity

**Different variances** across runs and samples and bacteria.

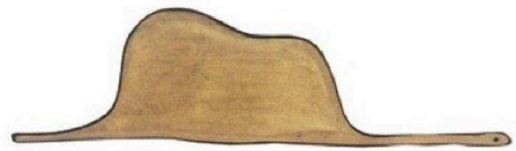
- variance stabilizing transformation




- Transformations (variance stabilizing): `log`, `asinh`, etc.
  For more information: Waste Not, want not video lecture
  Tutorial on DESeq2 and phyloseq
- Rank-transformations (see `LabRobustness.html`).

# Paths in thinking about these heterogeneous systems

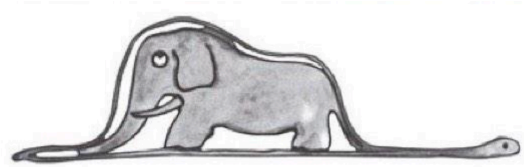
- Think in layers: latent variables or factors enable interpretation.



hidden variables.

# Paths in thinking about these heterogeneous systems

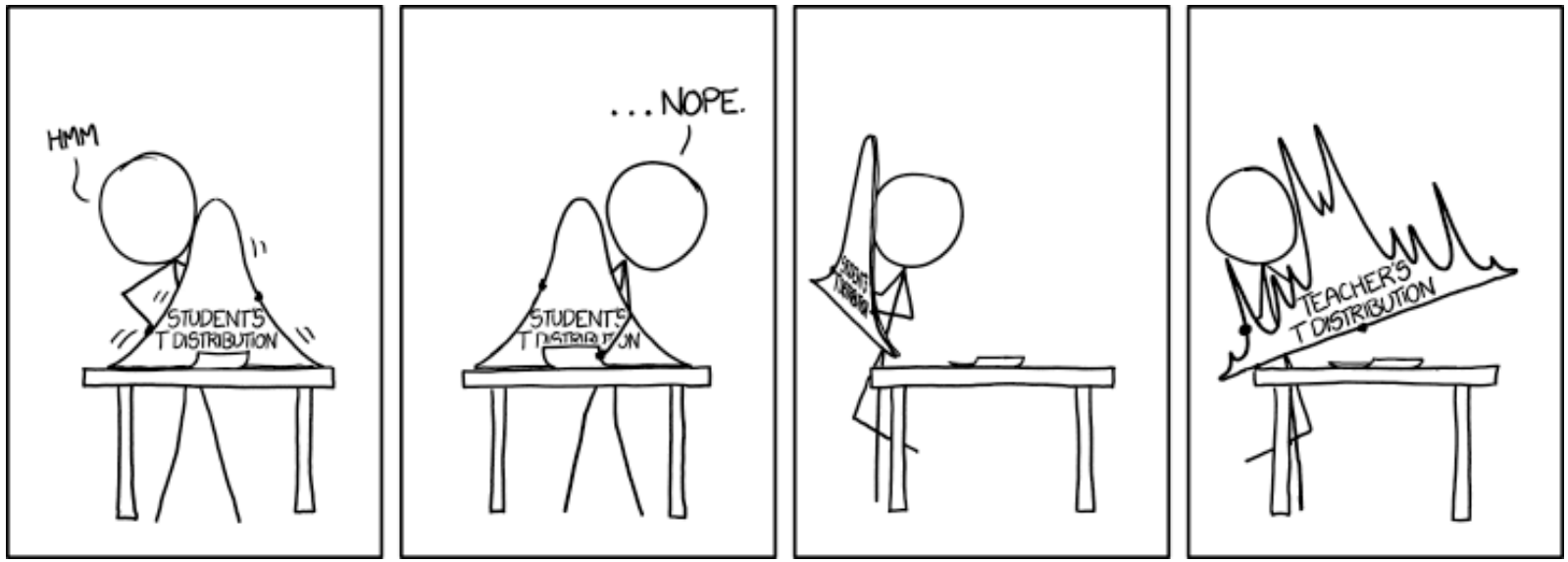
- Think in layers: latent variables or factors enable interpretation.



hidden variables.

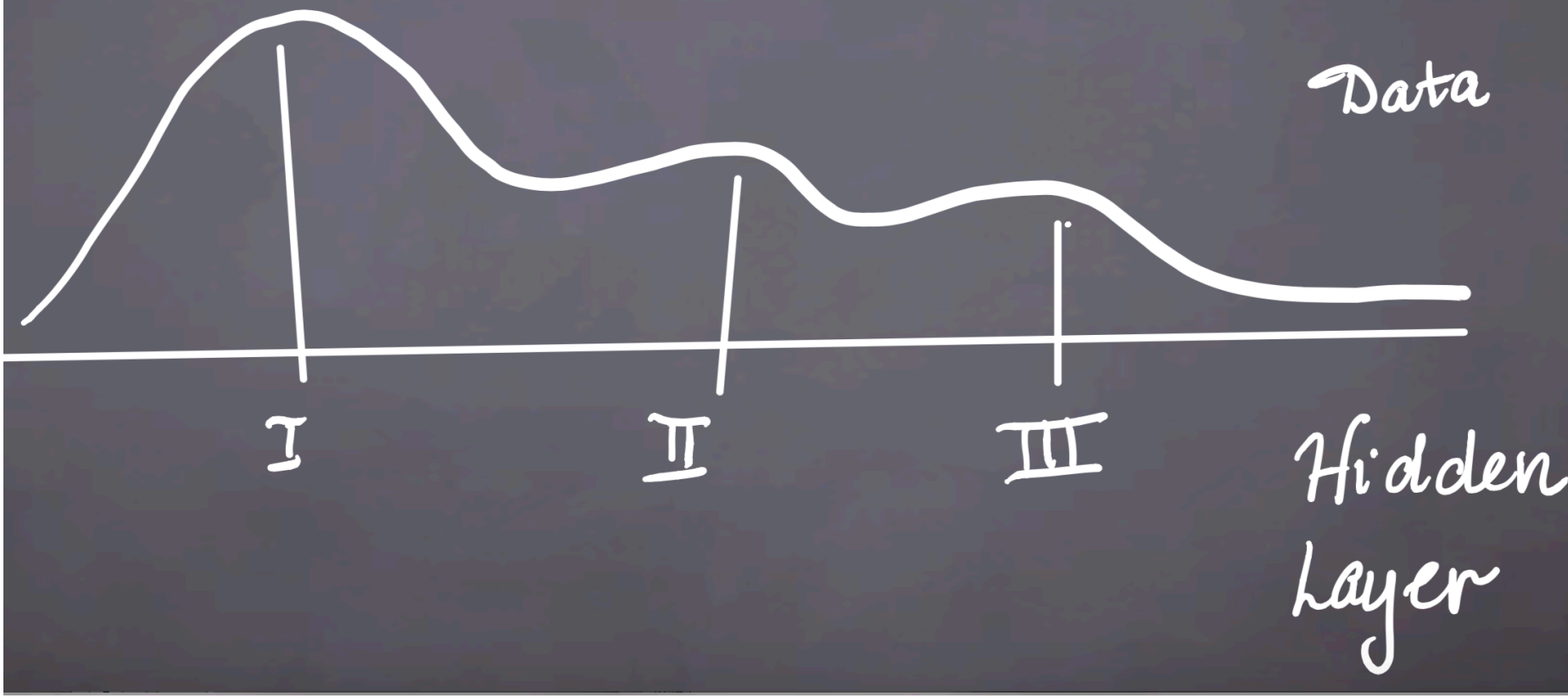
# Paths in thinking about these heterogeneous systems

- Think in terms of mixtures (not one parametric population).



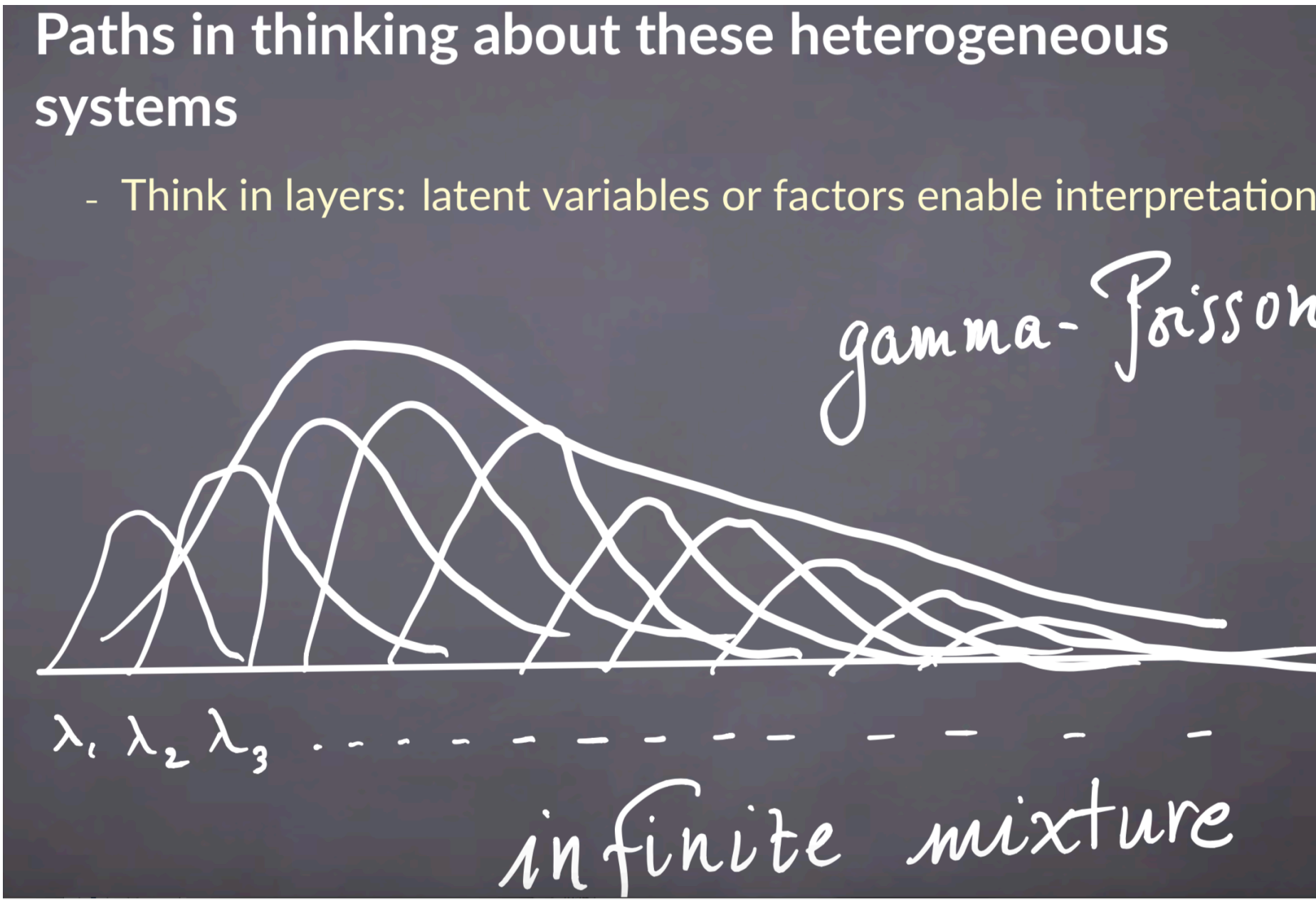
# Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation.



# Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation



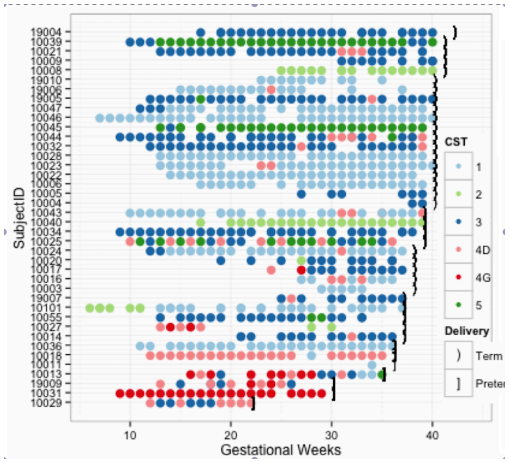
# The Yoda of Silicon Valley

## “premature optimization is the root of all evil in coding”

“premature summarization is the root of all evil in statistics”

# Different Levels of Dependencies

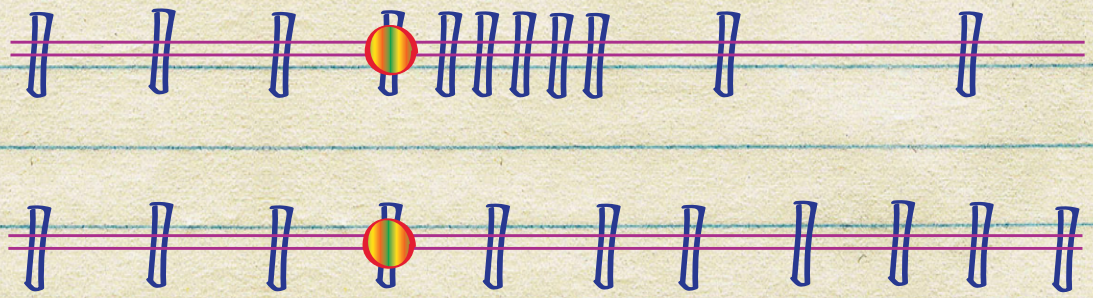
- Spatial (2 or 3 dimensional), oral microbiome study with Diana Proctor (Data and code here: ).
- Longitudinal studies:
  Equally spaced: (pregnancy).




Stability

Perturbation study (multidomain data).

# Not equally distant time points.



Between point **variation** should be as equal as possible.
See Peter Diggle's text : Analysis of Longitudinal Data, 2002.
Chapter on design of high-throughput experiments

# Part II

## Statistics is Hard, we are all Outsiders?

# The Jargon: we need the right words

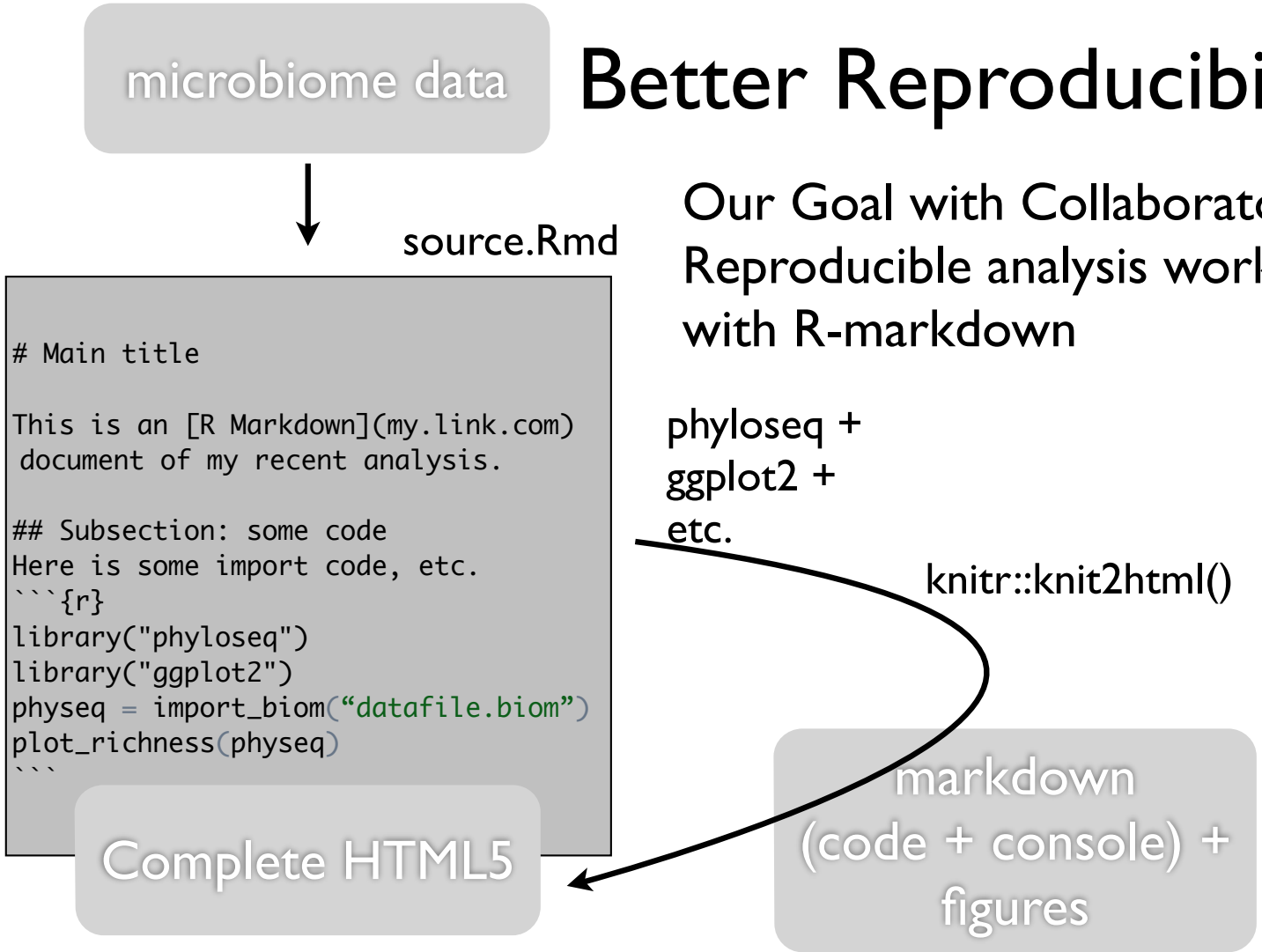
We can often find helpful information about biological phenomena by using Google or Wikipedia, because the words are quite specific:
Lymphocytes, CD45, epigenetics,.....
Statistics often uses common words with hundreds of other meanings:
Normal, Geometric, Independent, Expectation, Process, Mean, Bias, Conditional, Cluster, Random, Variance, generative, parameter ...

# Useful Concepts with (Hard) Buzzwords

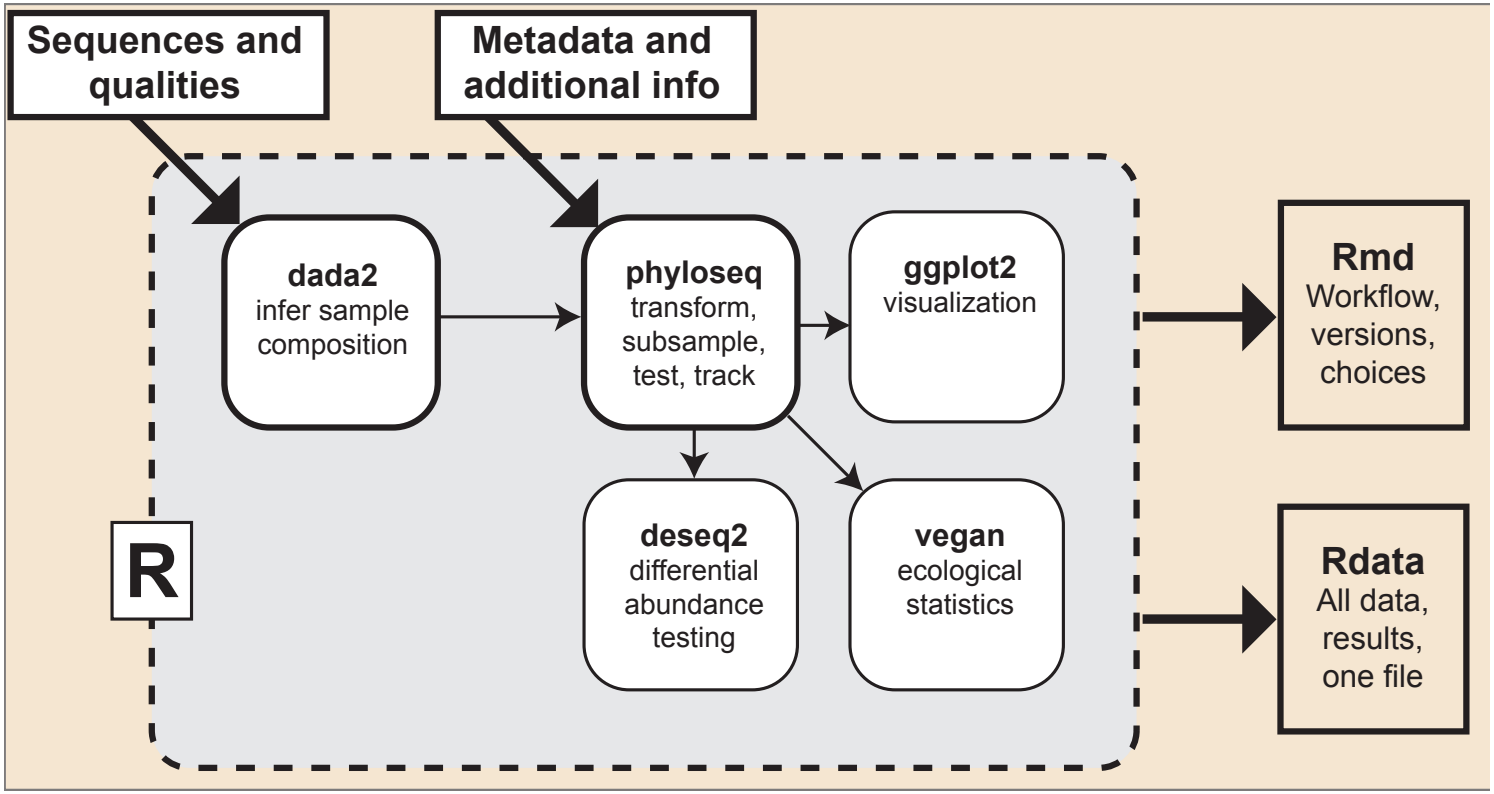
- Dependence.
- Multivariate.
- Effective sample size.
- Latent Variables.
- Noise.
- Robust.
- Spatio-Temporel Data Analysis.
- Degrees of Freedom.
- Heteroscedasticity.

# False Friends

# Better Reproducibility



# Reproducible Research Workflow



See complete workflow on Bioconductor channel of F1000:
http://f1000research.com/articles/5-1492/v1


CrossMark click for updates

RESEARCH ARTICLE

# Bioconductor workflow for microbiome data analysis: from raw reads to community analyses [version 1; referees: awaiting peer review]

Ben J. Callahan[1], Kris Sankaran[1], Julia A. Fukuyama[1], Paul J. McMurdie[2], Susan P. Holmes[1]

Author affiliations

Grant information

This article is included in the Bioconductor channel.

# Reproducible research R markdown code and data

- Waste not, want not paper, Plos Comp Bio.
- Complete workflow from reads to community networks, F1000Research.
  F1000Research paper
  Bioconductor workflow (html).
- Oral Microbiome
- Web page with all code and images as they appear from the code. Complete Analysis of Colonic Cleanout Data
- Enterotypes, oral microbiome PSB 2016.
- Treelapse for antibiotics

# R packages and resources

phyloseq: http://bioconductor.org/packages/stats/bioc/phyloseq/

dada2: http://bioconductor.org/packages/stats/bioc/dada2/

treelapse: https://krisrs1128.github.io/treelapse/

treelapse antibiotics http://statweb.stanford.edu/~kriss1/antibiotic.html

microbiome_pvlm: https://github.com/krisrs1128/microbiome_plvm

decontam: https://github.com/benjjneb/decontam/

adaptiveGPCA: https://cran.r-project.org/web/packages/adaptiveGPCA/index.html

bootLong: https://github.com/PratheepaJ/bootLong/blob/master/vignettes/Workflow.Rmd

Modern Statistics for Modern Biology

http://bios221.stanford.edu/book/

# Solutions for microbiome analyses: respect the data.

- Poor data quality, information $\longrightarrow$ quality scores & probability.
- Maintain all information $\longrightarrow$ sequences are names.
- Reproducibility $\longrightarrow$ complete code source.
- Heterogeneity $\longrightarrow$ multicomponent objects:phyloseq.
- Training and collaboration $\longrightarrow$ Rmd and html.

# More problems: todo next time and further

- Multivariate statistics= analysing matrices.
- Multivariate data combined with trees and networks.
- Multidomain problems (metagenomics, RNA-seq, images, ...).
- Interpretation $\longrightarrow$ latent variables (gradients or clusters).
- Model choices (transformation choices).