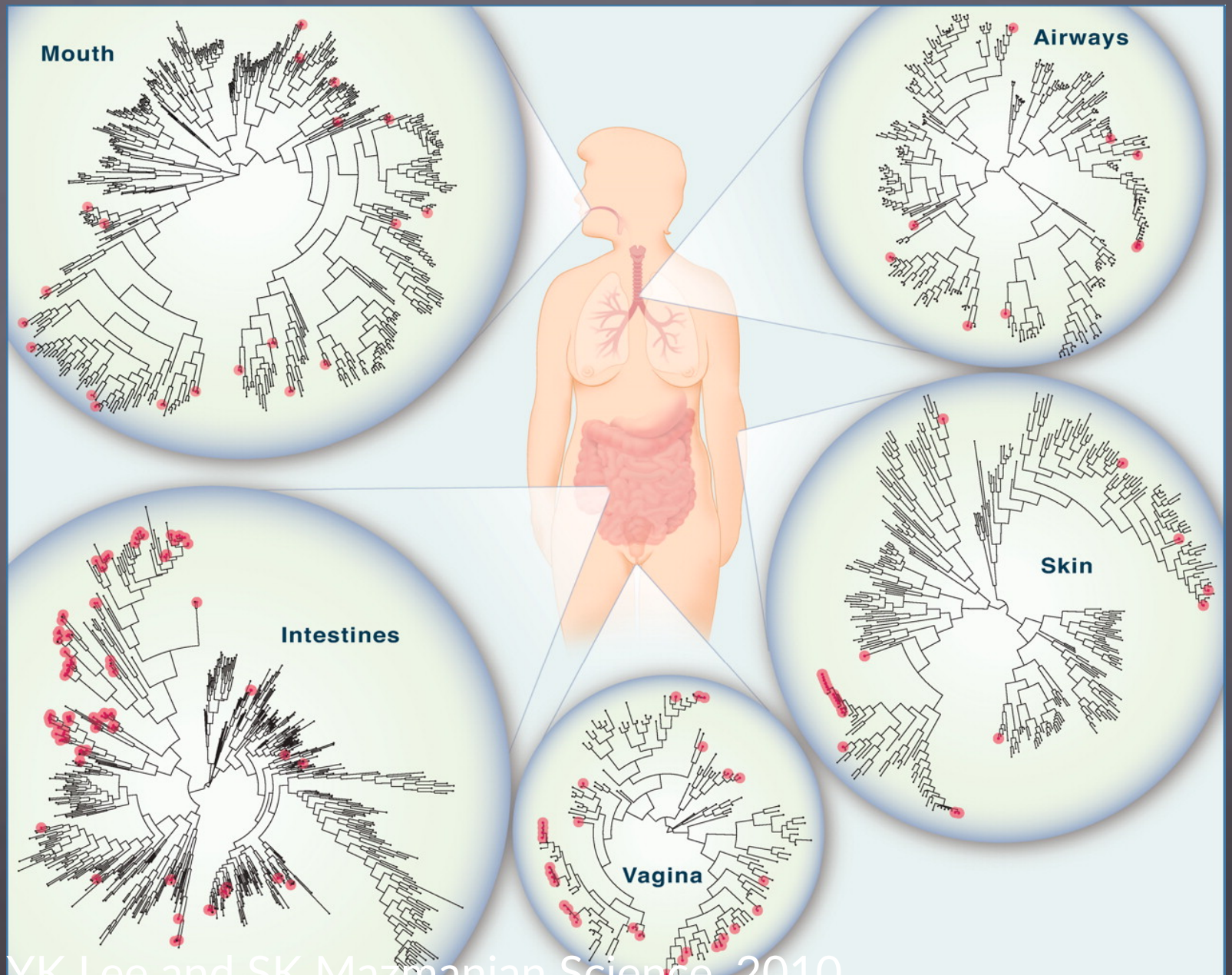
Défis des données hétérogènes du microbiome avec R dada2 et phyloseq

Susan Holmes
@SherlockpHolmes
http://www-stat.stanford.edu/~susan/

Bio-X and Statistics, Stanford University

Montpellier, 17 Juin, 2019

The messes we deal with



YK Lee and SK Mazmanian Science 2010

Défis

- Heterogenieté.
- Incorporation d’information disponible en forme de Graphes ou Arbres.
- Graphiques de haute qualité.
- Robustesse.
- Reproduction des résultats.

Part I

Heterogeneity

`Homogeneous data are all alike;

all heterogeneous data are heterogeneous

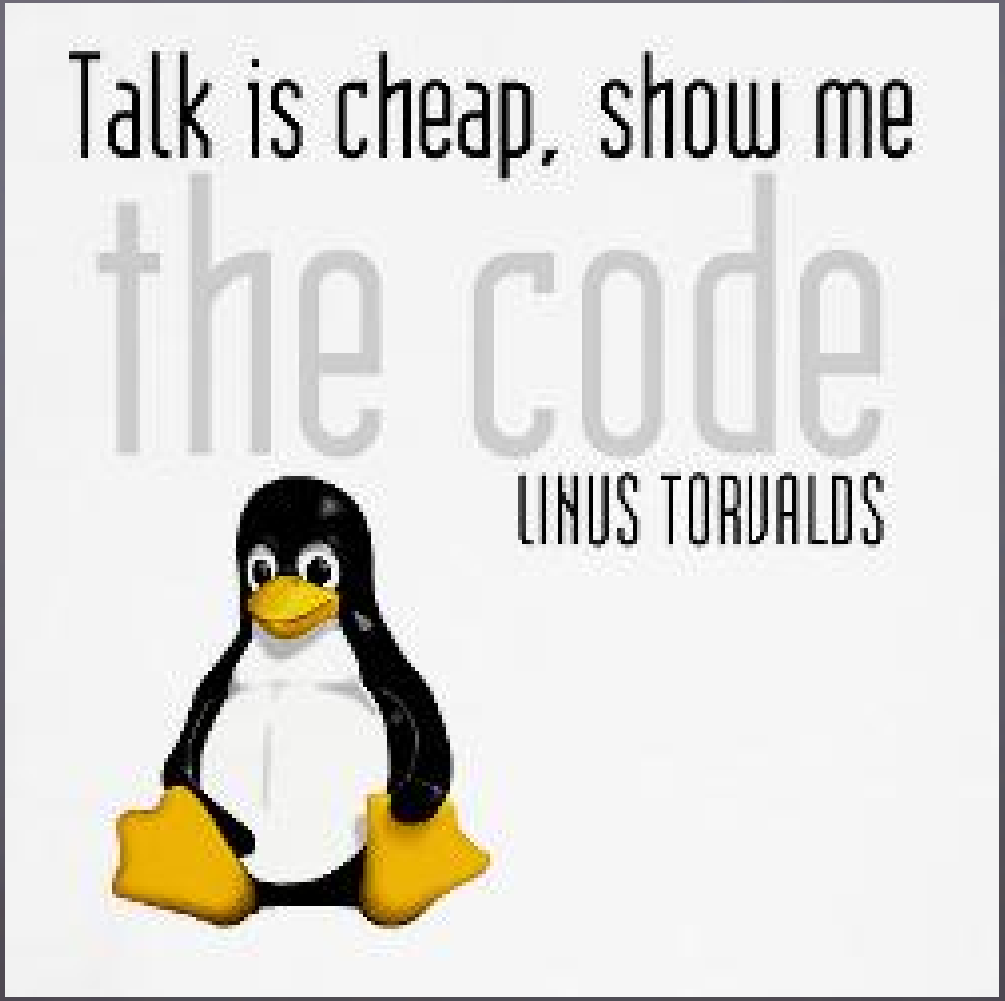
in their own way.'

Heterogeneity of Data

- Statut des variables : réponse/ explicatives.
- Cachée (latent)/ ou measurée.
- Types :
  - Continu
  - Binaires, qualitatives.
  - Graphes/ Arbres.
  - Images
  - Information spatiales.
  - Rankings/ rangements.
- Dependences: independent/time series/spatial/mesures repetées.
- Technologies différentes (454, Illumina, MassSpec, NMR, RNA-seq).

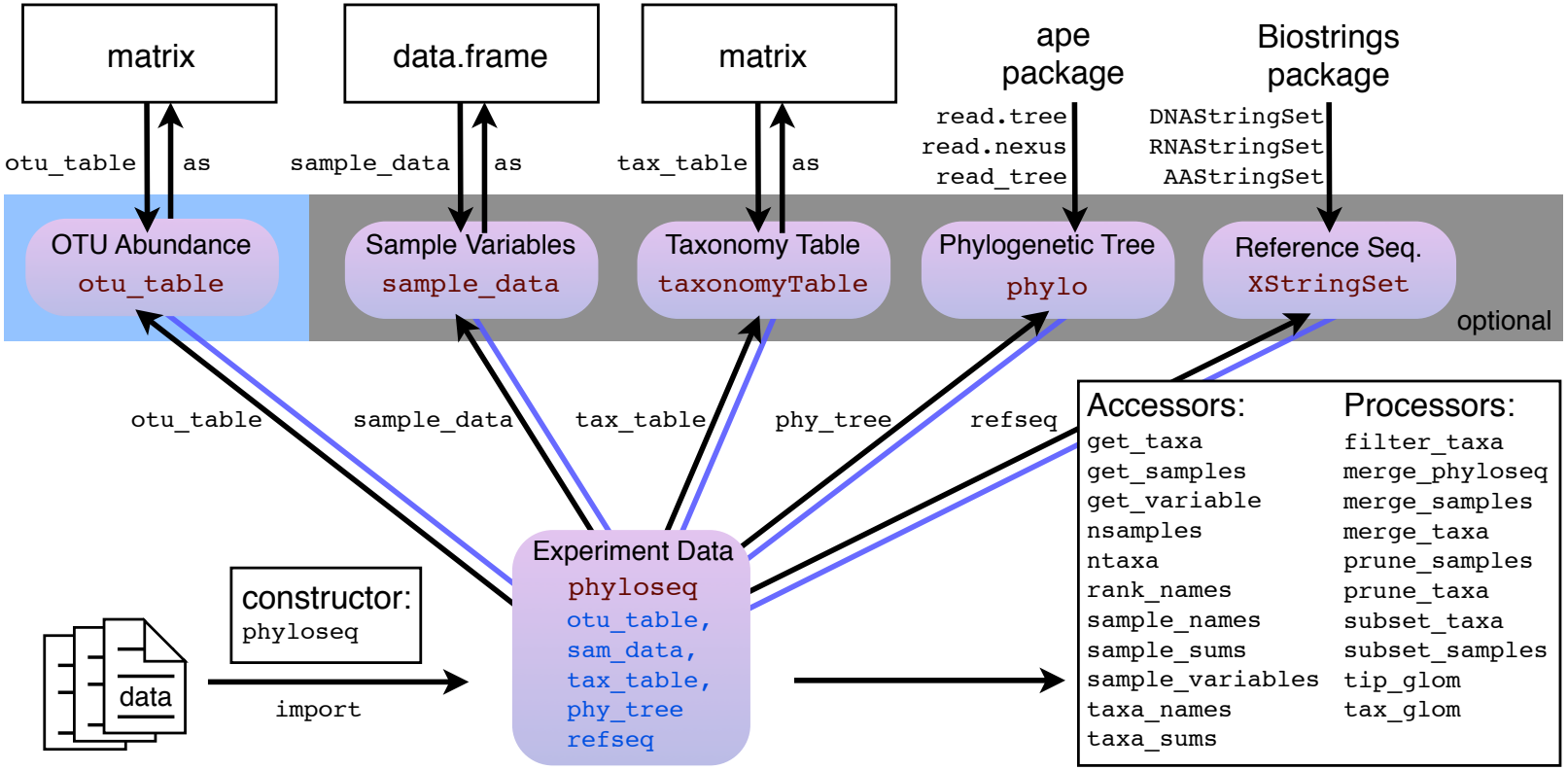
Part II

Implementation



phyloseq

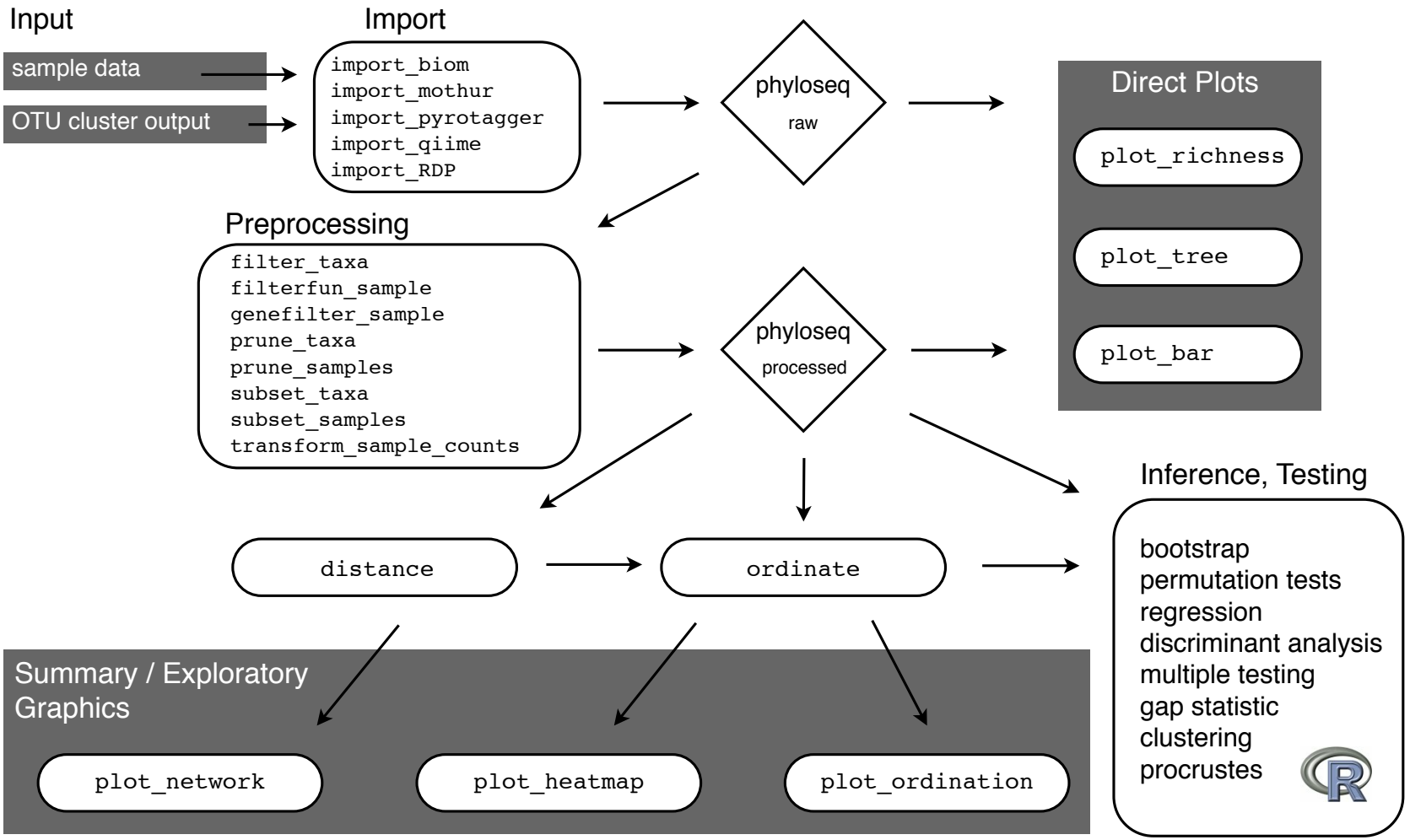
data structure & API



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

phyloseq

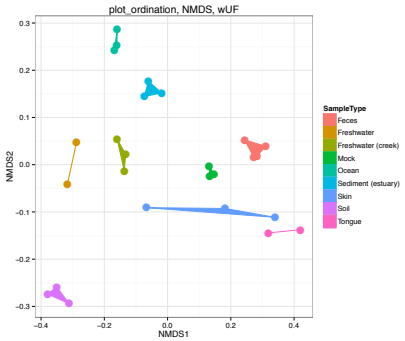
work flow

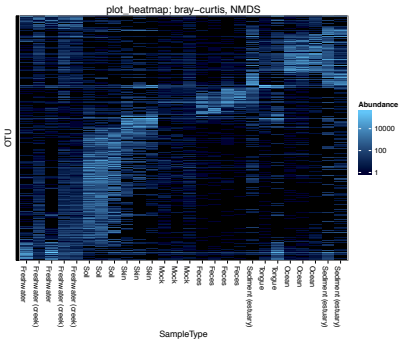


phyloseq

graphics

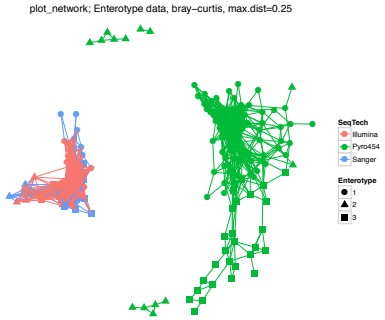
plot_ordination()

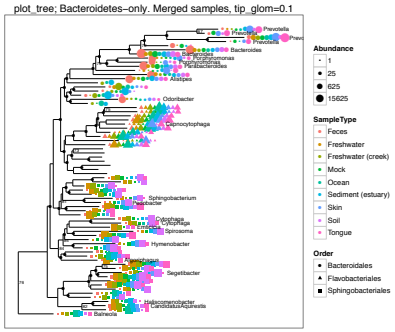




plot_heatmap()

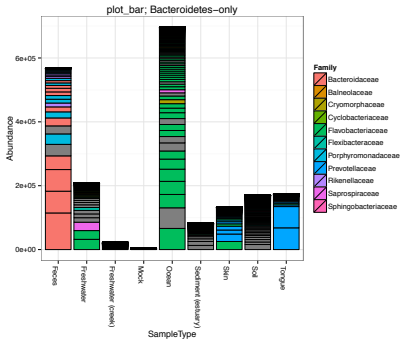
plot_network()

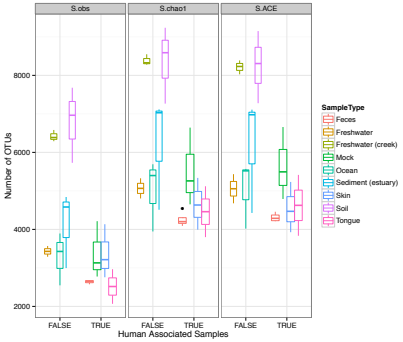




plot_tree()

plot_bar()



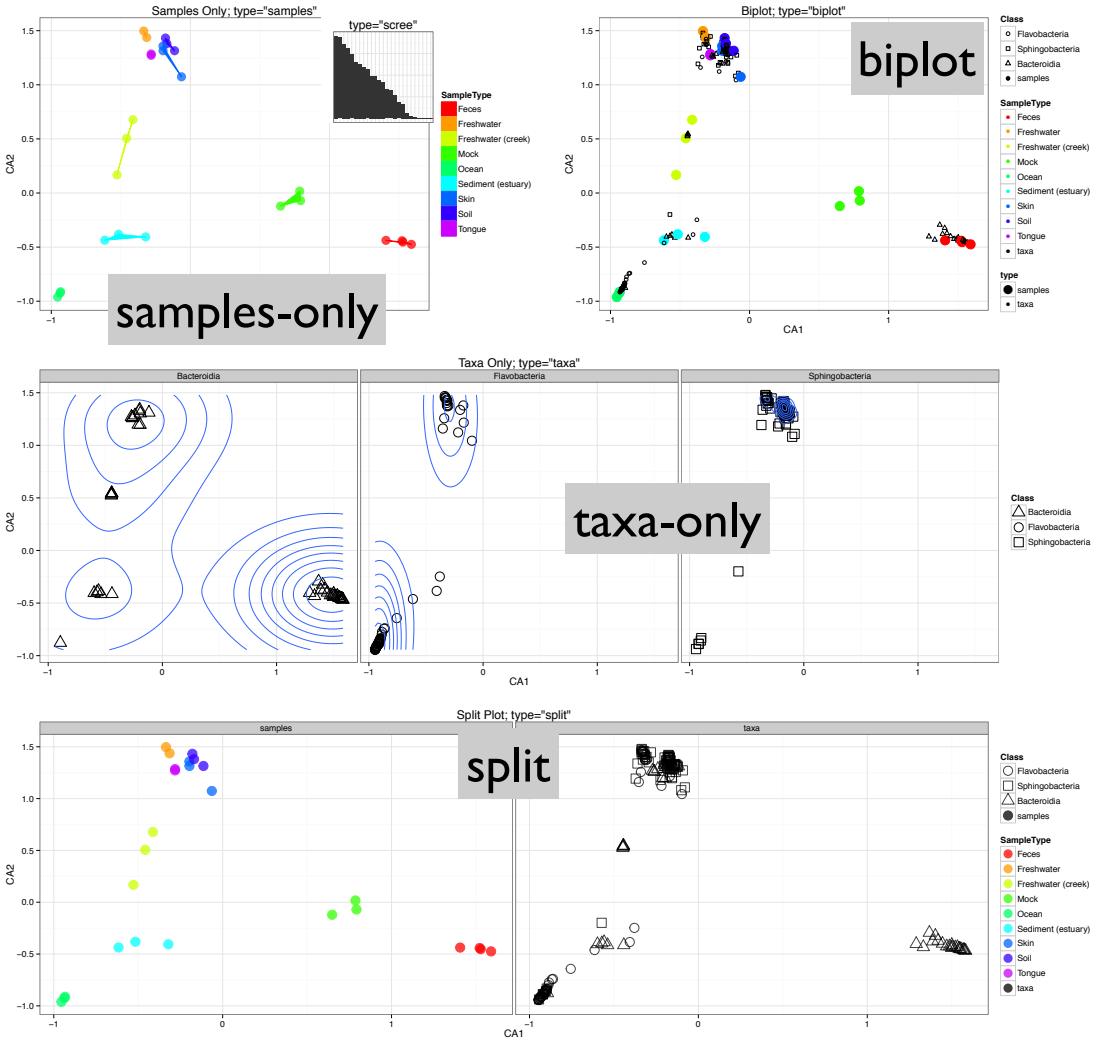


plot_richness()

phyloseq

graphics

plot_ordination()



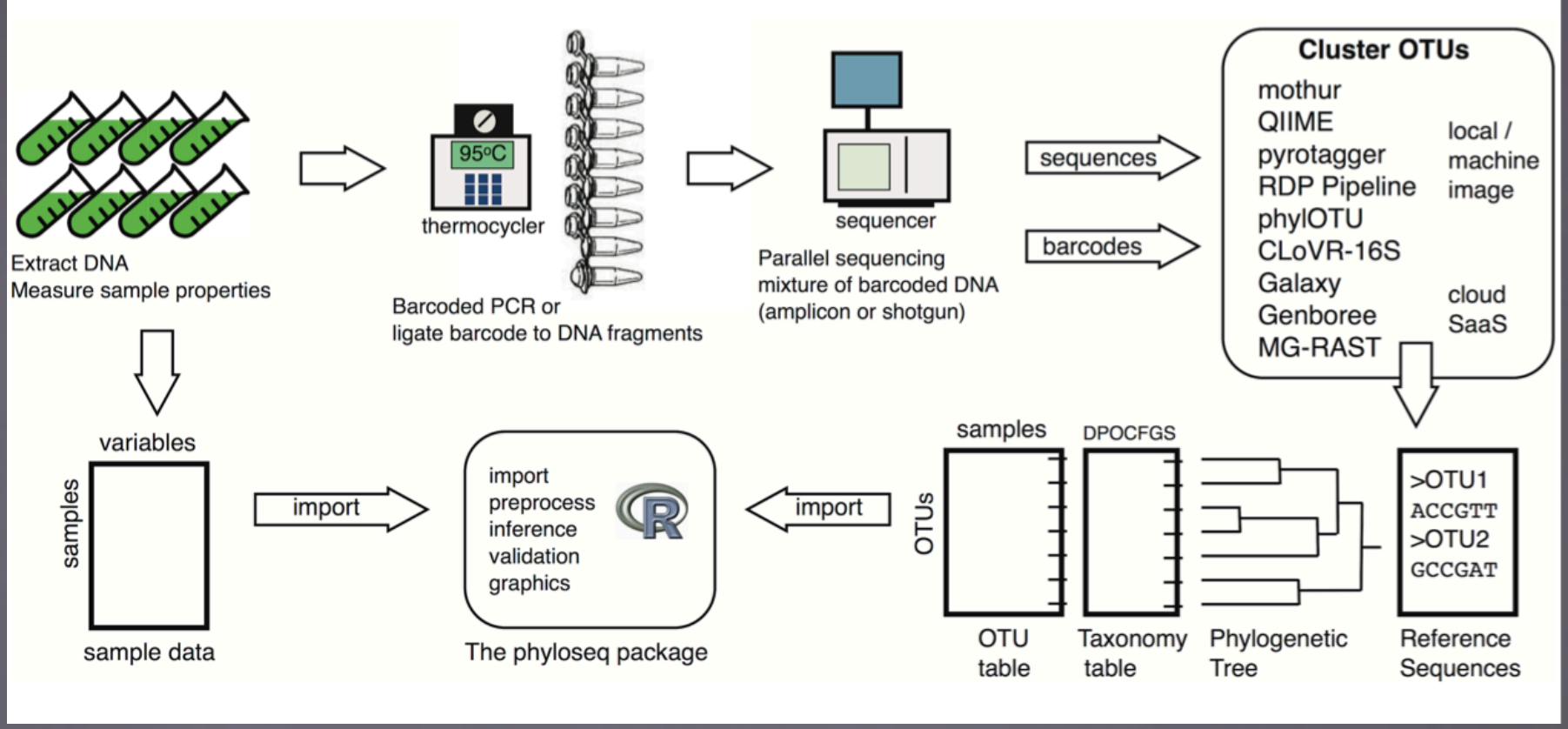
Mapping Variables onto Phylogenies

Example using phyloseq package

```
> data(esophagus)
> es1                <- esophagus
> sn                 <- sample_names(esophagus)
> sample_data(es1) <- sample_data(data.frame(sample=sn, row
```

We create the tree graphic, grouping/coloring by our dummy sample-name variable, and also labeling the number of individuals observed in each sample (if at all). The symbols are slightly enlarged as the number of individuals increases.

```
> plot_tree_phyloseq(es1, color_factor="sample",
     type_abundance_value=TRUE, treeTitle="Kostic example c
```



Heterogeneous Data Objects

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

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

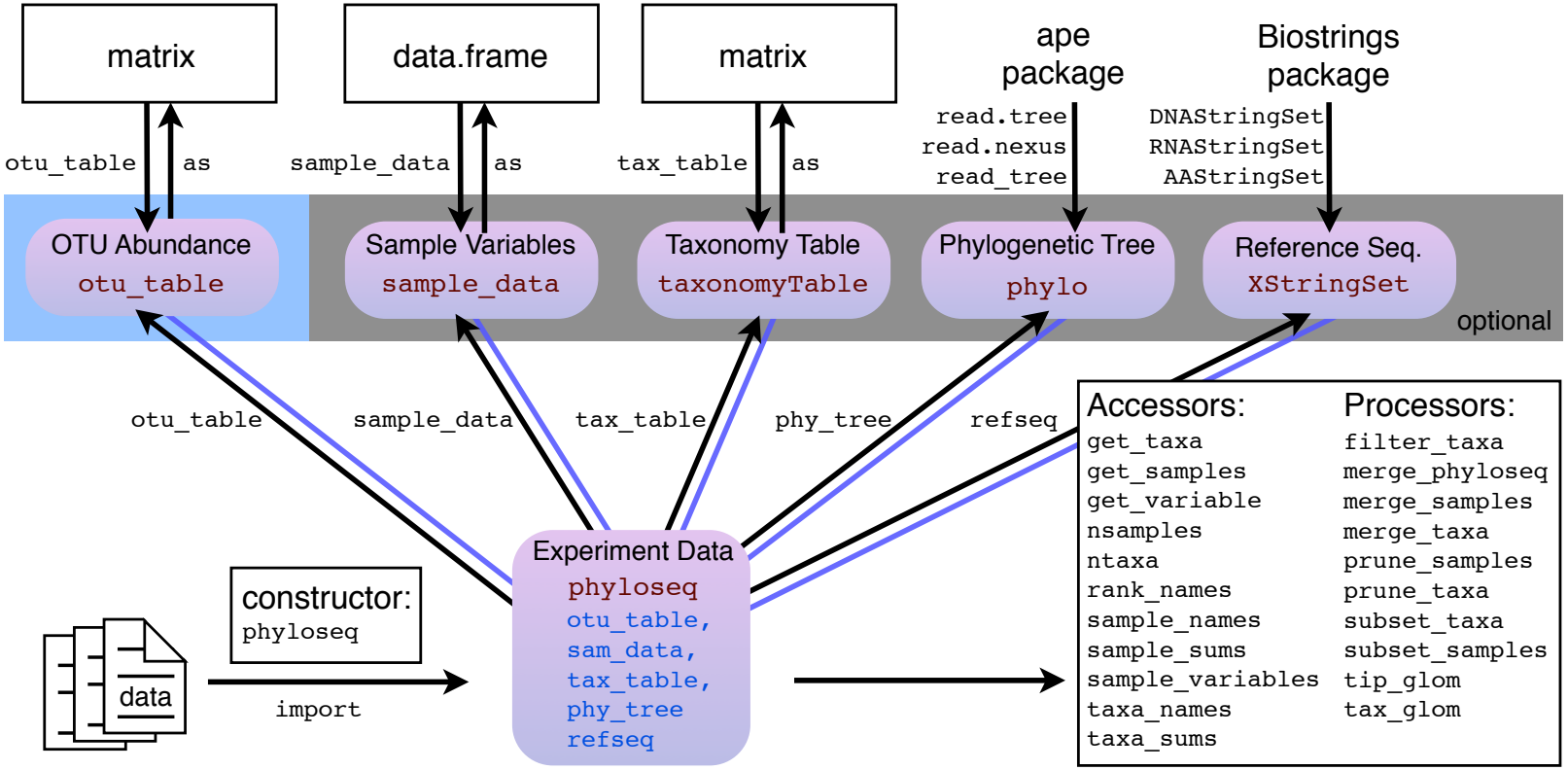
phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data

Paul J. McMurdie, Susan Holmes

Published: April 22, 2013 • https://doi.org/10.1371/journal.pone.0061217

phyloseq

data structure & API



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

Representation utiles: Plusieurs Matrices

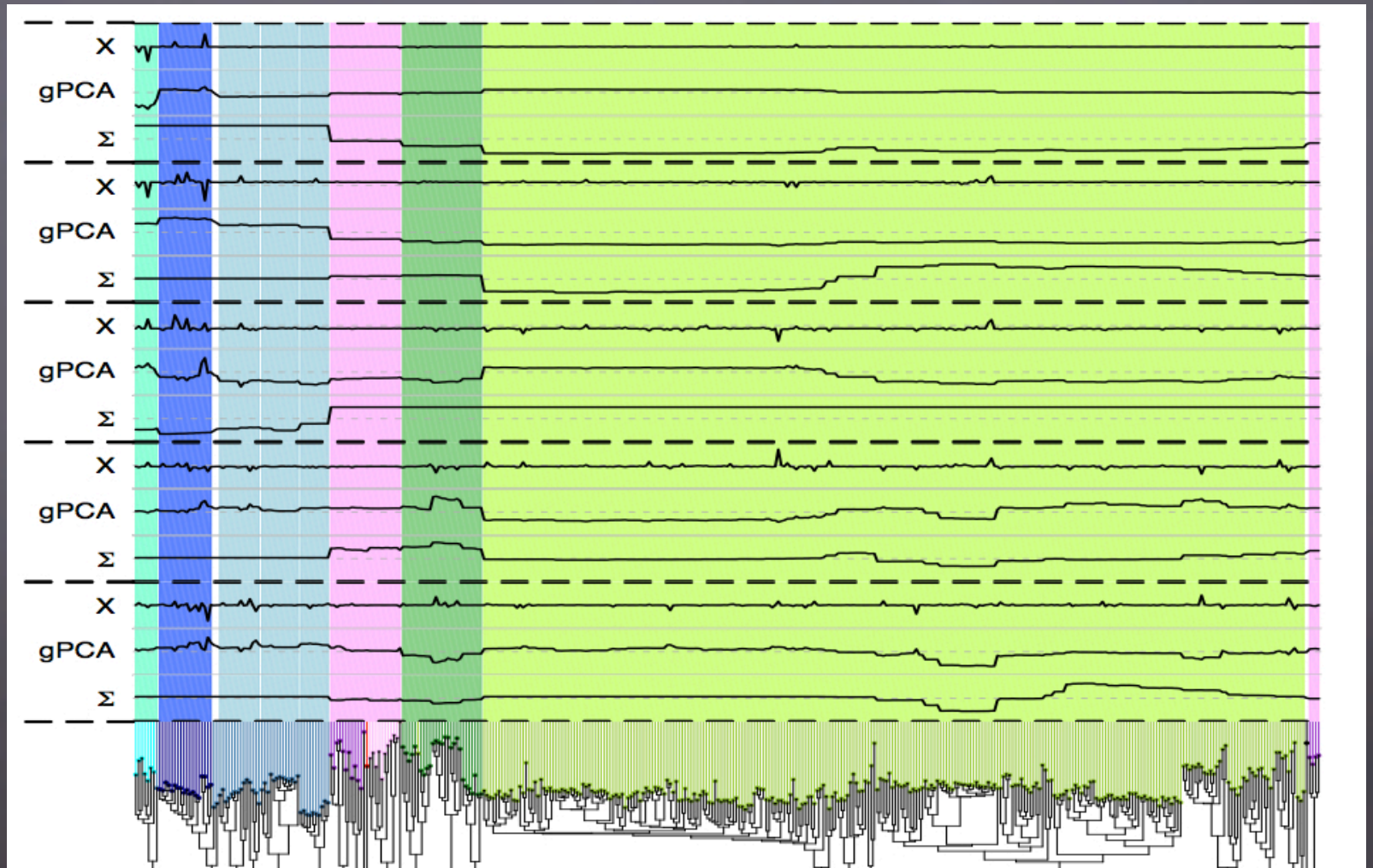
with metrics..

- Time series of abundance matrices.
- Different types of data on same samples (taxa counts, clinical variates, spatial location).
- Networks and trees over time.
- Explanatory (environmental) variables, Response variables.

Holmes (2005), Duality Diagrams, matrices with metrics.

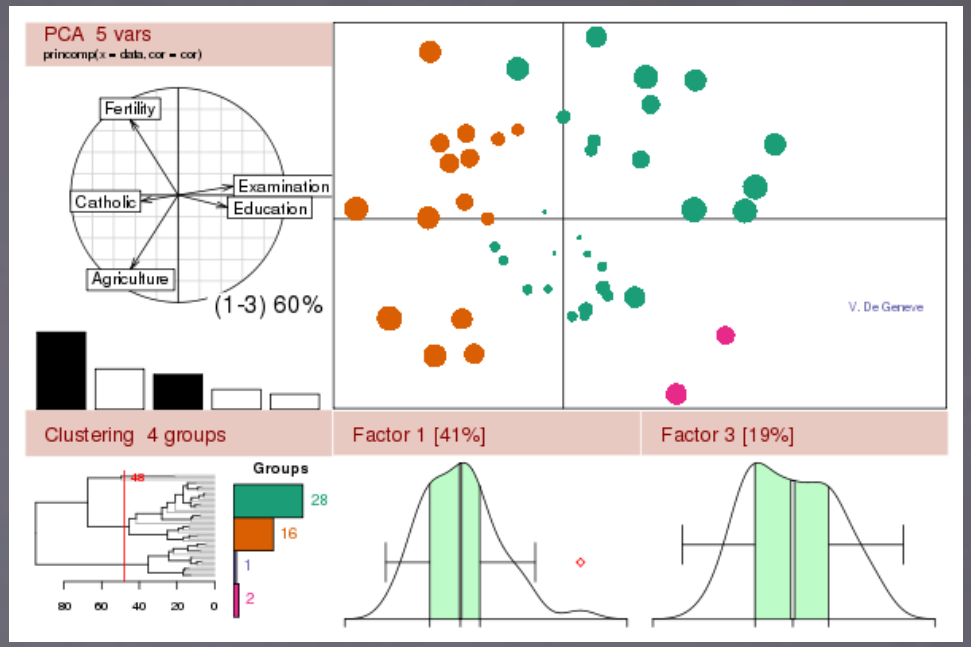
Incorporer les abundances (comptages) des espèces avec l'arbres de phylogénie

Bik et al, 2006, Science.
Purdom, 2011, Annals of Applied Statistics.



Graphiques

.



High powered graphics packages with layered functionalities:

- `ade4` .
- Philosophy: Leland Wilkinson's Grammar of Graphics.
- Implementation :`ggplot2`
  (Hadley Wickham youtube video)
- Animated gifs, interactive graphics.

Part III

Robustesse

High Breakdown (median,)

Sparse (L1 minimizing)

Nonparameteric (ranks, nmMDS, ..)

Part IV

Confirmatory Analysis: Nonparametric tests

Nonparametric Tests

- Canonical Correspondence Analysis tests on a factor.
- Mantel's Test between distance matrices
- Multiple testing correction.
- Bootstrap tests.

Everything is done by shuffling labels

Example:

Is there a shedding effect in Relman/Hoy Mice data?

```
> shed = scan("shed.txt")
> shedf = as.factor(shed)[-(70:71)]
> resca.shed = cca(t(pib.nz) ~ shedf)
> anova(resca.shed)
Permutation test for cca under reduced model

Model: cca(formula = t(pib.nz) ~ shedf)
         Df  Chisq      F N.Perm Pr(>F)
Model     6 0.3825 2.0627    199  0.005 **
Residual 87 2.6886
---
```

Example: Is there a subject effect in Katie's data?

```
 subjcc = vegan::cca(tnorepnz ~ subject)
 anova(subjcc)

Permutation test for cca under reduced model

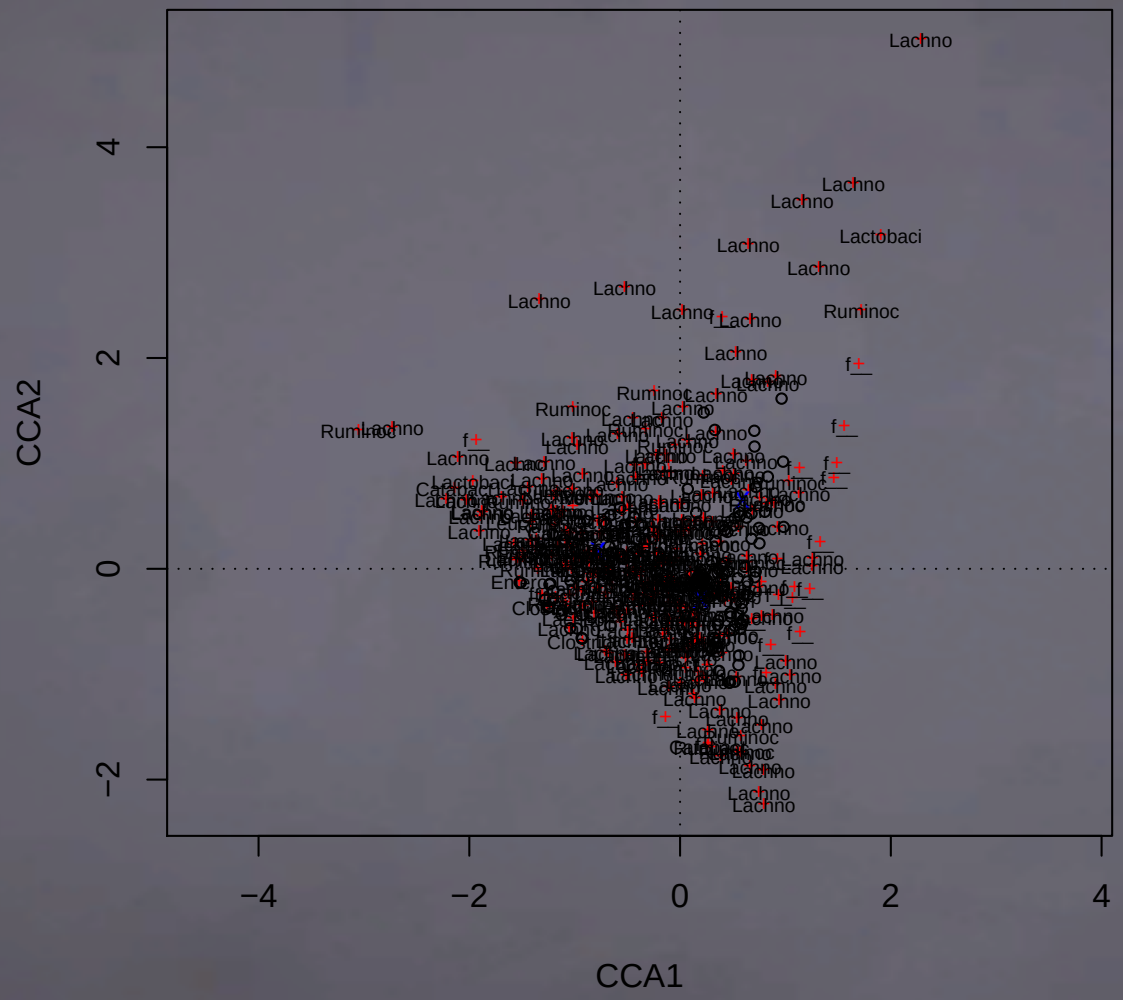
Model: cca(formula = tnorepnz ~ subject)
          Df   Chisq       F N.Perm Pr(>F)
Model      7  1.2177 10.7999    199  0.005 **
Residual 472  7.6027
---
Signif. codes:  0 ,***, 0.001 **, 0.01 ,* 0.05 . 0.1
```

```
> cca.cage = vegan::cca(t(tcmall) ~ cagef)
> plot(cca.cage)
> text(cca.cage, choices = c(1, 2), label = cagef, display
> anova(cca.cage)
Permutation test for cca under reduced model

Model: cca(formula = t(tcmall) ~ cagef)
          Df  Chisq      F N.Perm Pr(>F)
Model      2 0.3722 7.6501    199  0.005 **
Residual 178 4.3305
---
```

```
> plot(cca.cage, scaling = 1)
> text(cca.cage, scaling = 1, choices = c(1, 2), display =
+     cex = 0.6)
> title("Species, Cage effect, 1,2")
```

Species, Cage effect, 1,2



Over-representation of certain phyla

Set	Over-represented	Universe	
Microbiome	Families/Phyla	Species Present	Test
Gene Expression	Ontological groups	Filtered Genes	

in both cases: hypergeometric / Fisher's exact test.
We define the set of prefiltered species (**species universe**) as those that passed the threshold test of being present (> 6000) in at least 31 of the arrays.
This method is especially relevant here as the tree does not show equal representation of different families and phyla.

Hypergeometric Tests for over representation of certain phyla)

- IBS higher group had significantly more Bacteriodetes
- overrepresentation of Firmicutes in the healthy controls.
- At the family level, the results showed that the families of Oxalobacteraceae, Prevotellaceae, Burkholderiaceae, Sphingobacteriaceae were significantly overrepresented in IBS.
- Conversely, the most significantly enriched family in control rats were Lachnospiraceae, including Ruminococcus sp., followed by Erysipelotrichaeceae and Clostridiaceae.

Example of Shiny-Phyloseq

Better Reproducibility

microbiome data

source.Rmd

```
# Main title

This is an [R Markdown](my.link.com)
 document of my recent analysis.

## Subsection: some code
Here is some import code, etc.
```{r}
library("phyloseq")
library("ggplot2")
physeq = import_biom(“datafile.biom”)
plot_richness(physeq)
```
```

Our Goal with Collaborators:
Reproducible analysis workflow
with R-markdown

phyloseq +
ggplot2 +
etc.

knitr::knit2html()

markdown
(code + console) +
figures

Complete HTML5

Part V

An Example

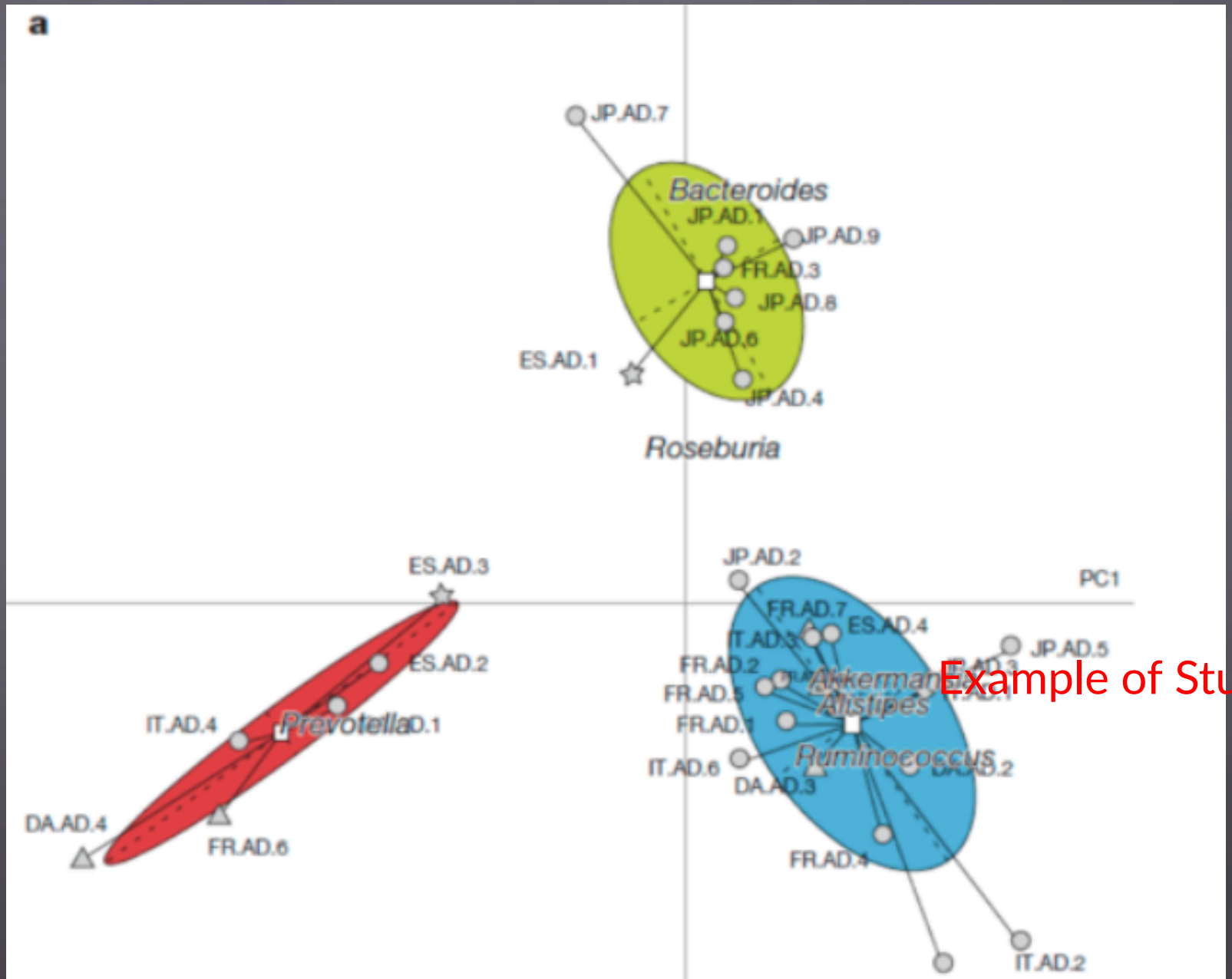
ARTICLE

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

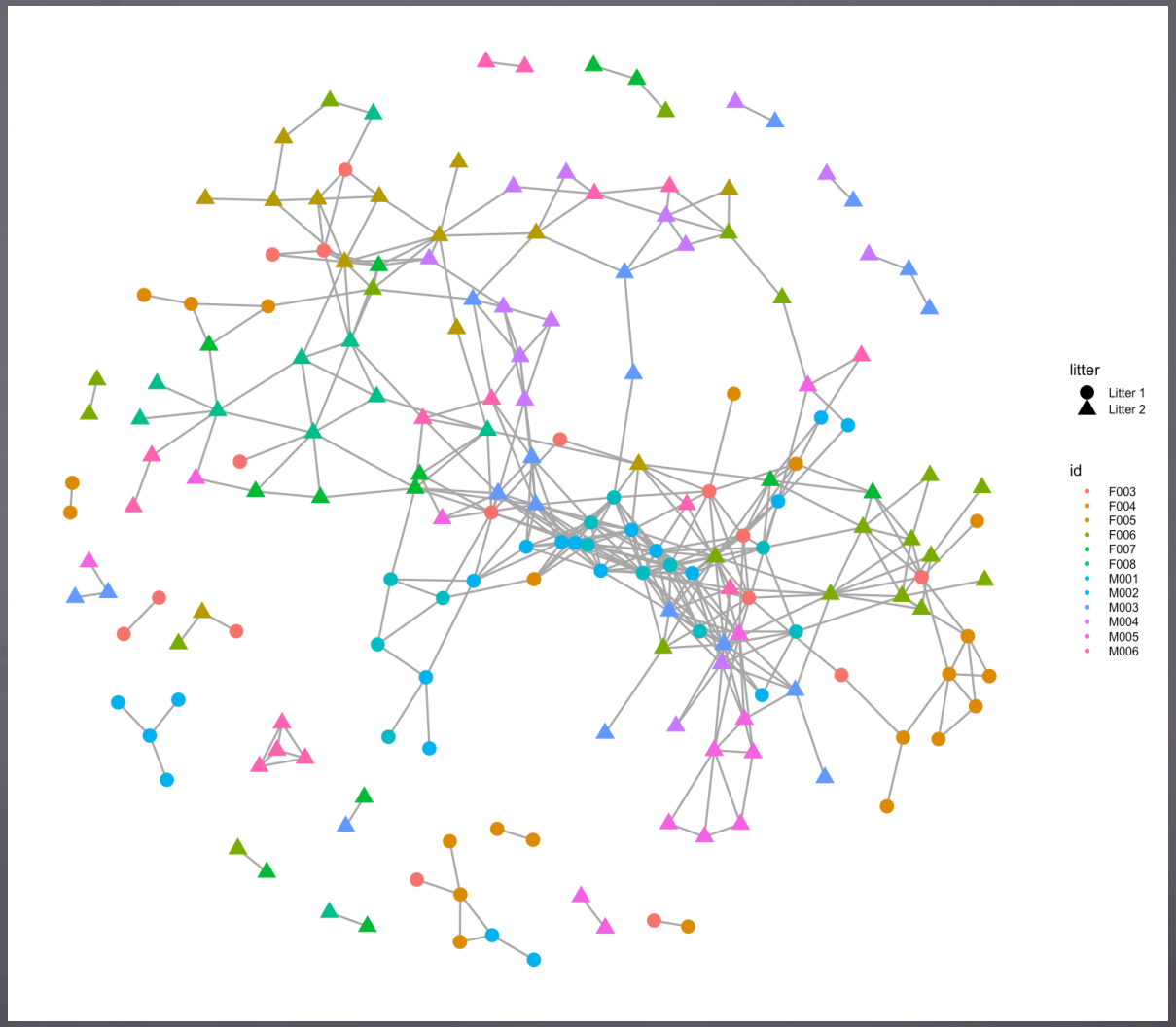
Summary of the study

- 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:

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

Des tests qui utilisent les graphes



https://bioconductor.org/help/course-materials/2017/BioC2017/Day1/Workshops/Microbiome/MicrobiomeWorkflowII.html#graph-based analyses

Nature **473,** 174–180 (2011); doi:10.1038/nature09944 and corrigendum **474,** 666 (2011); doi:10.1038/nature10187

It has been drawn to our attention that the methods described in the main text and the Supplementary Information of this Article have been considered by some researchers to be insufficient to enable them to identify enterotypes in their own data sets. Enterotypes were originally defined in this Article (page 177) as "densely populated areas in a multi-dimensional space of community composition" and should not be seen as discrete clusters, but as a way of stratifying samples to reduce complexity. Additionally, the Fig. 2 legend should not imply that between-class analysis is simply a method of visualizing principal component analysis (PCA); rather, it is a supervised rather than an unsupervised analysis of data because it incorporates the outcome of clustering of data. To simplify enterotype identification in the original and other data sets, we have developed a comprehensive tutorial at http://enterotype.embl.de—which is a website on enterotypes that will be updated as methods improve. We thank Ivica Letunic and Paul Costea from EMBL for setting up the tutorial.

Your turn

Link to `http://bios221.stanford.edu/stamps/Phyloseq_Lab.html`.
Link to `http://bios221.stanford.edu/stamps/Phyloseq_Lab.Rmd`.