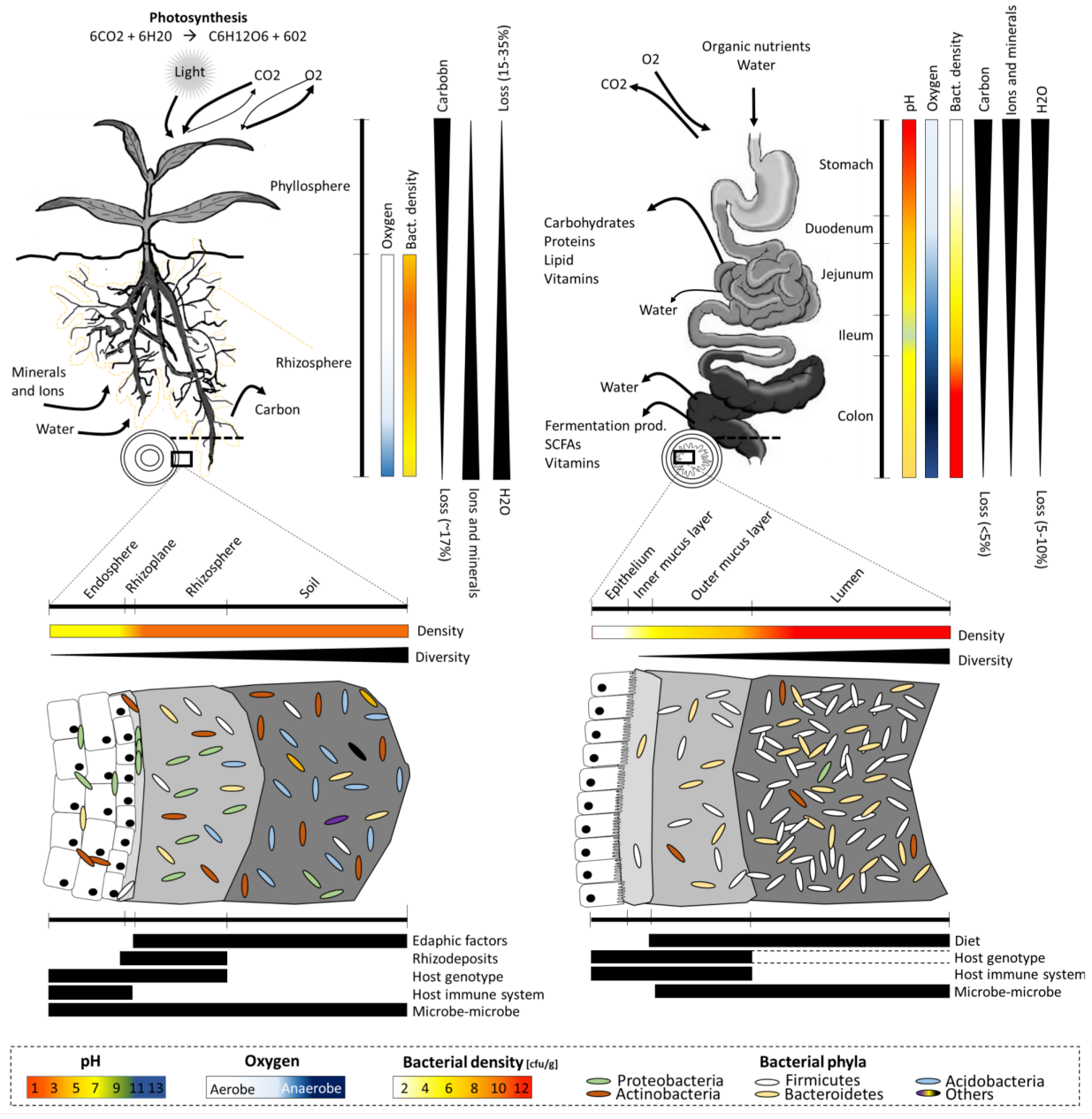
DECRyPT bioinformatics workshop - Introductory talk

Ruben Garrido-Oter
Max Planck Institute for Plant Breeding Research

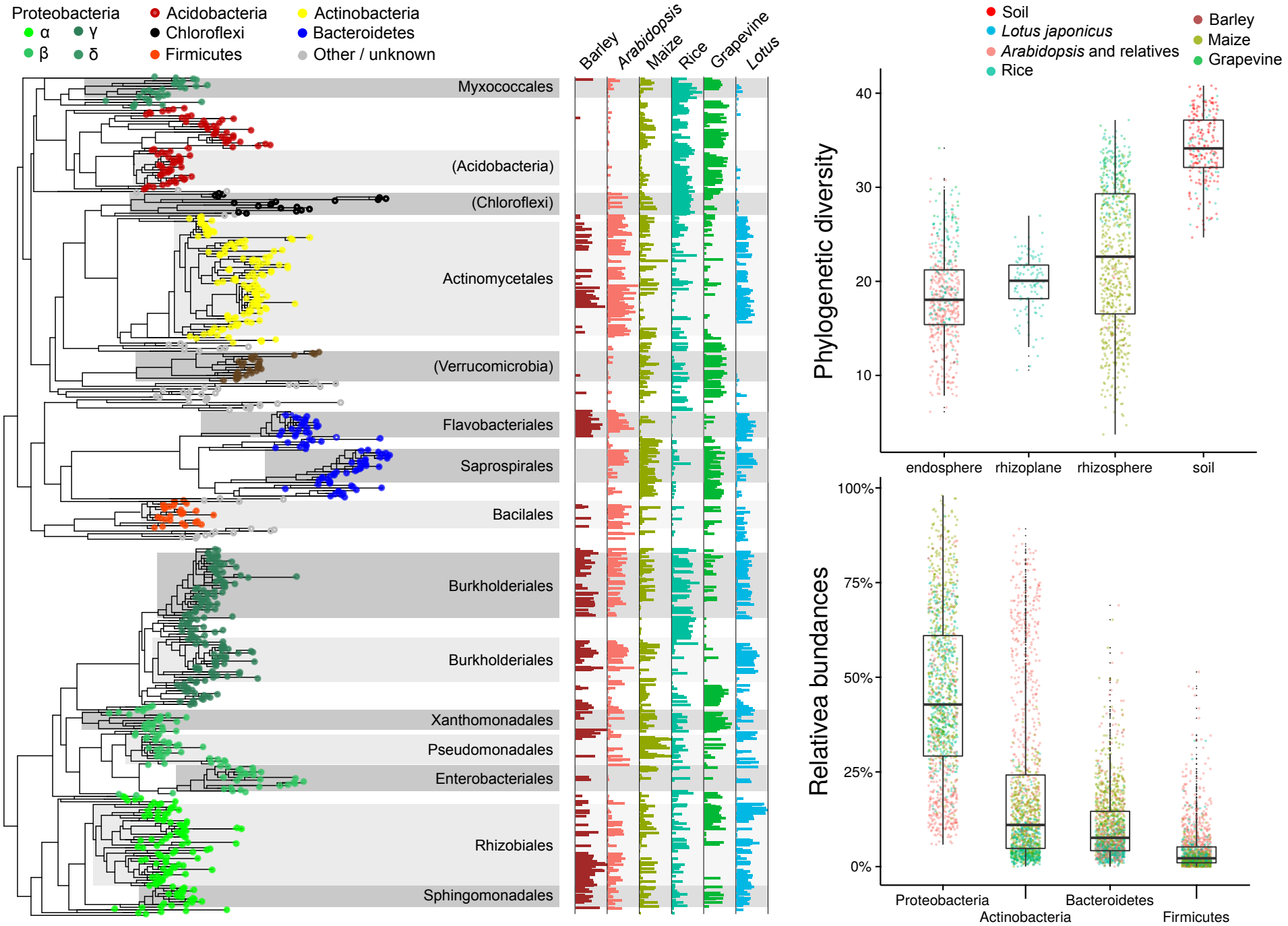
DECRyPT bioinformatics workshop - October 2019

The root microbiota

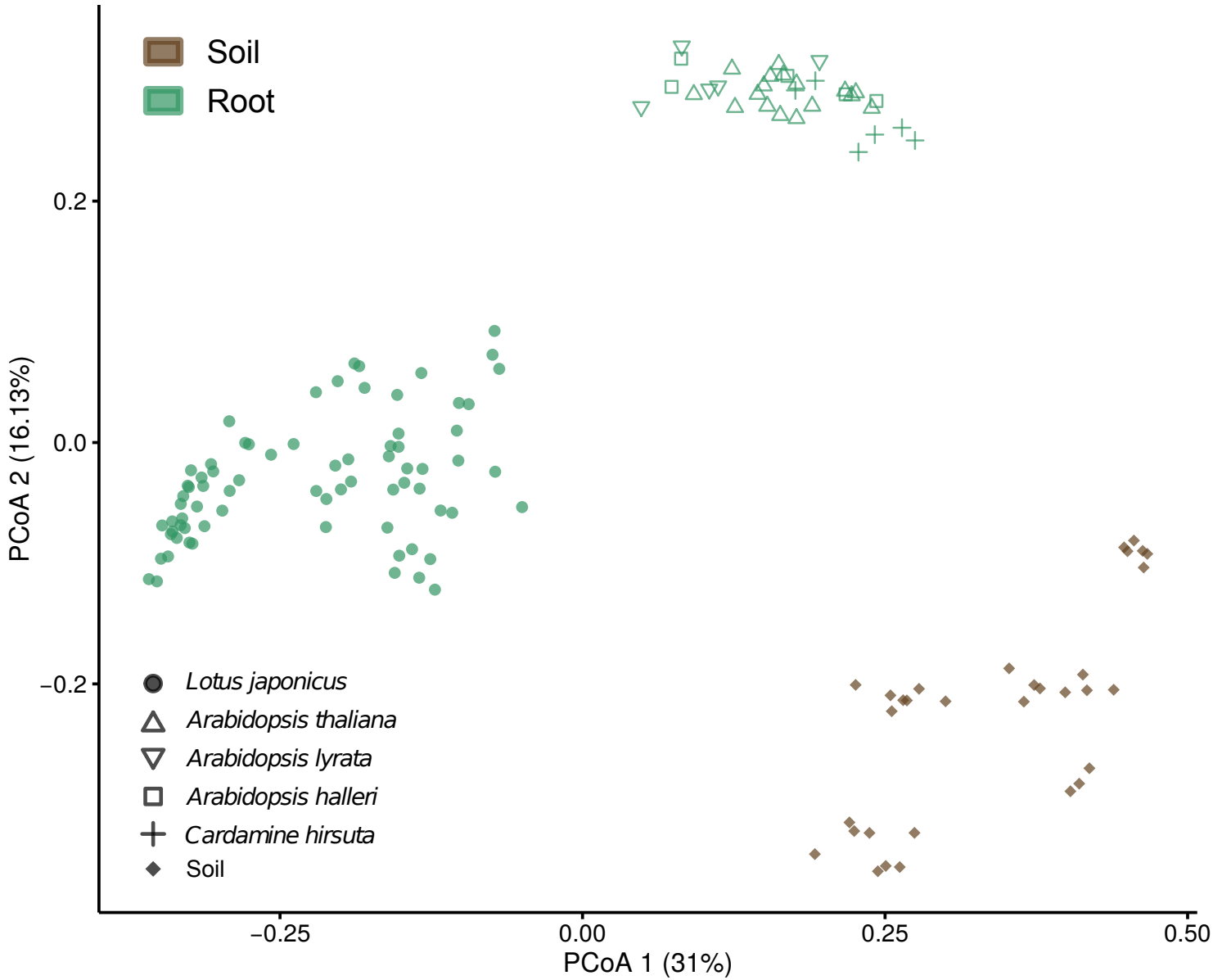


Hacquard, *et al. Cell Host & Microbe*, 2015.

The plant root microbiota



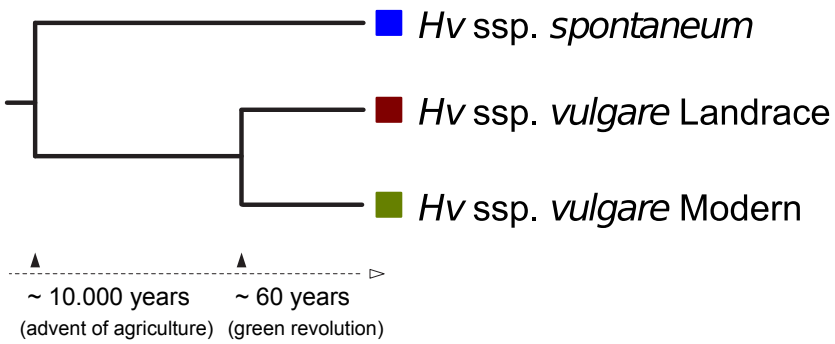
Root microbiota composition in the same soil is plant species-specific

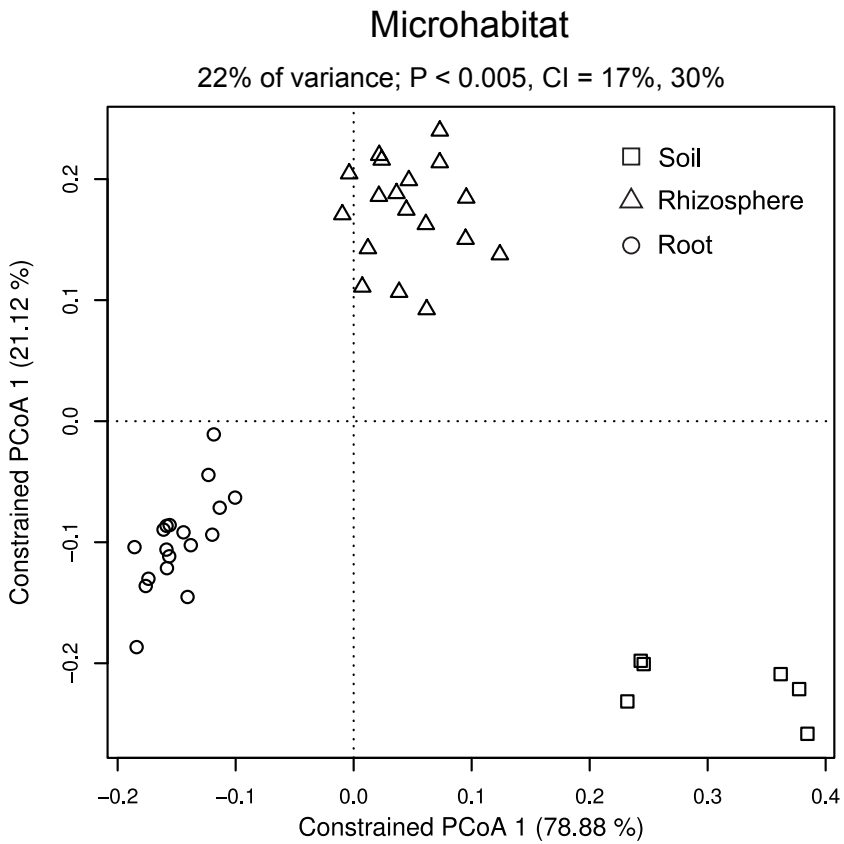


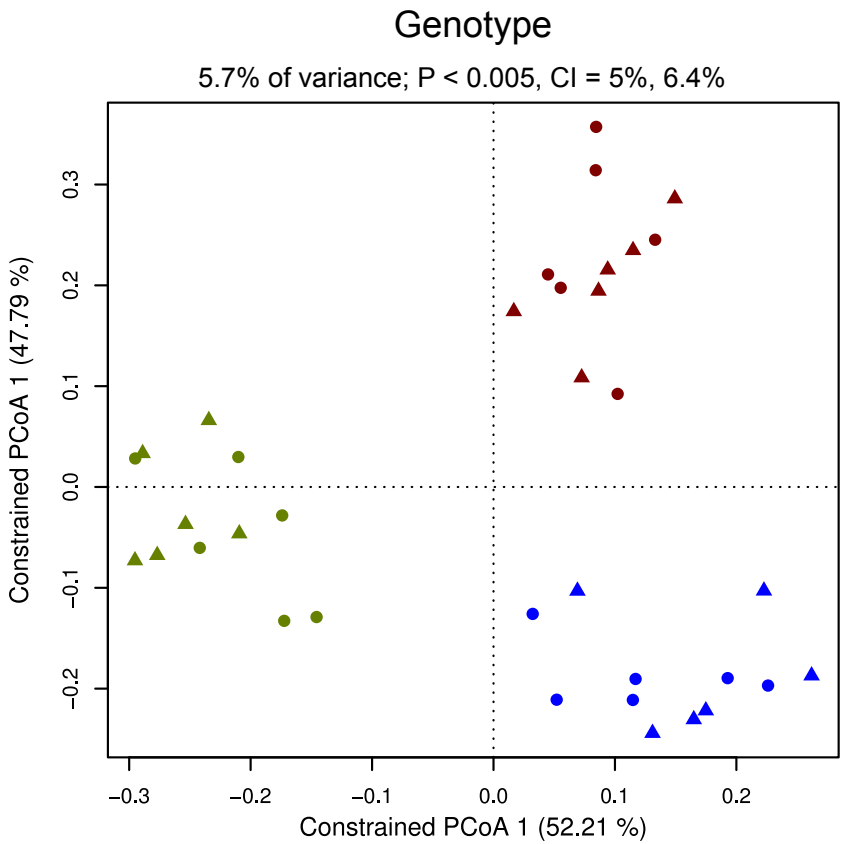
Community profiles of *Lotus* and *Arabidopsis* (and relative species) grown in Cologne Agricultural Soil (CAS)

What is the effect of the host genotype on the microbiota?

Structure of the root microbiota in wild and domesticated barley







Bulgarelli*, Garrido-Oter*, *et al. Cell Host & Microbe*, 2015.

Small genotype effect and low heritability of the plant root and leaf microbiota at the whole community level

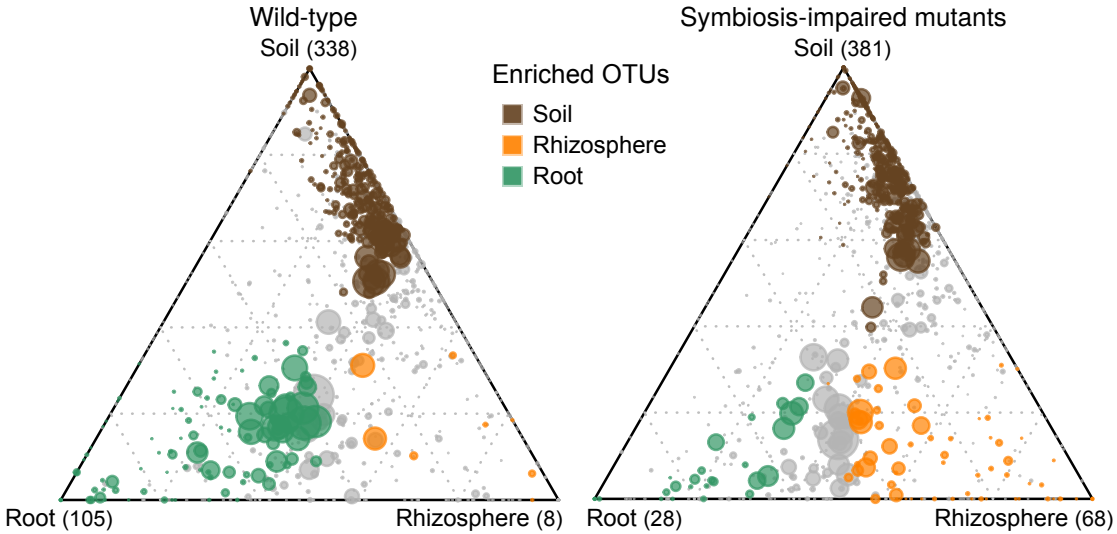
(Pfeiffer *et al.*, 2013; Schlaeppi *et al.*, 2014; Bulgarelli *et al.*; 2015, Wagner *et al.*; 2016)

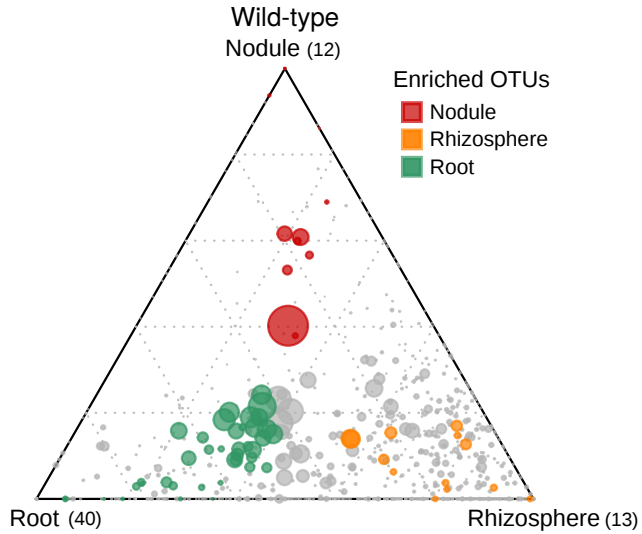
Small genotype effect and low heritability of the plant root and leaf microbiota at the whole community level

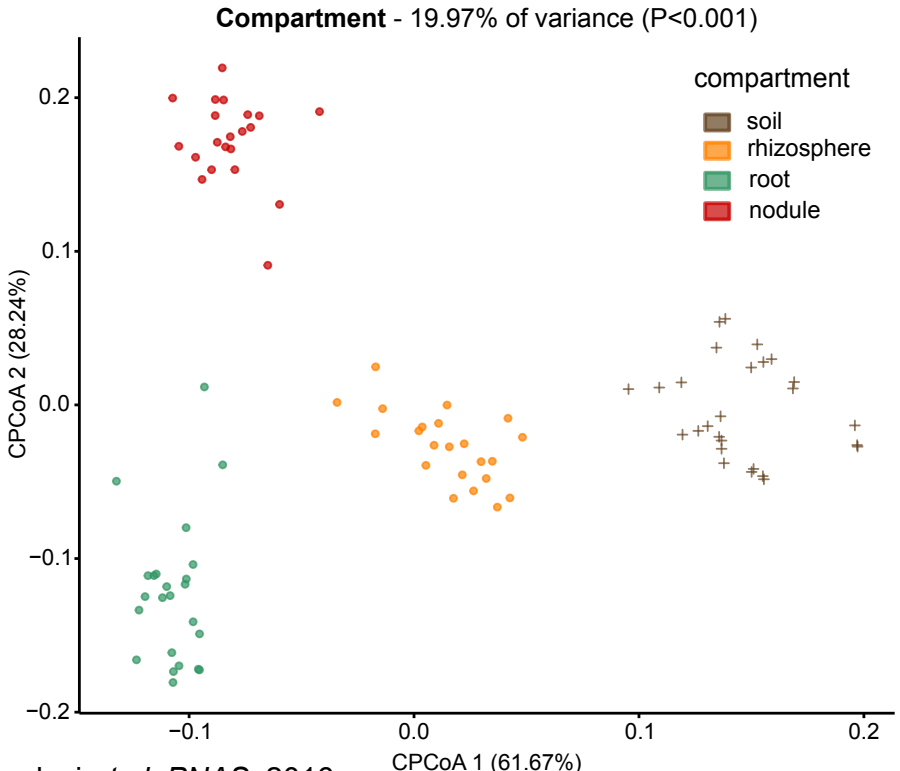
(Pfeiffer *et al.*, 2013; Schlaeppi *et al.*, 2014; Bulgarelli *et al.*; 2015, Wagner *et al.*; 2016)

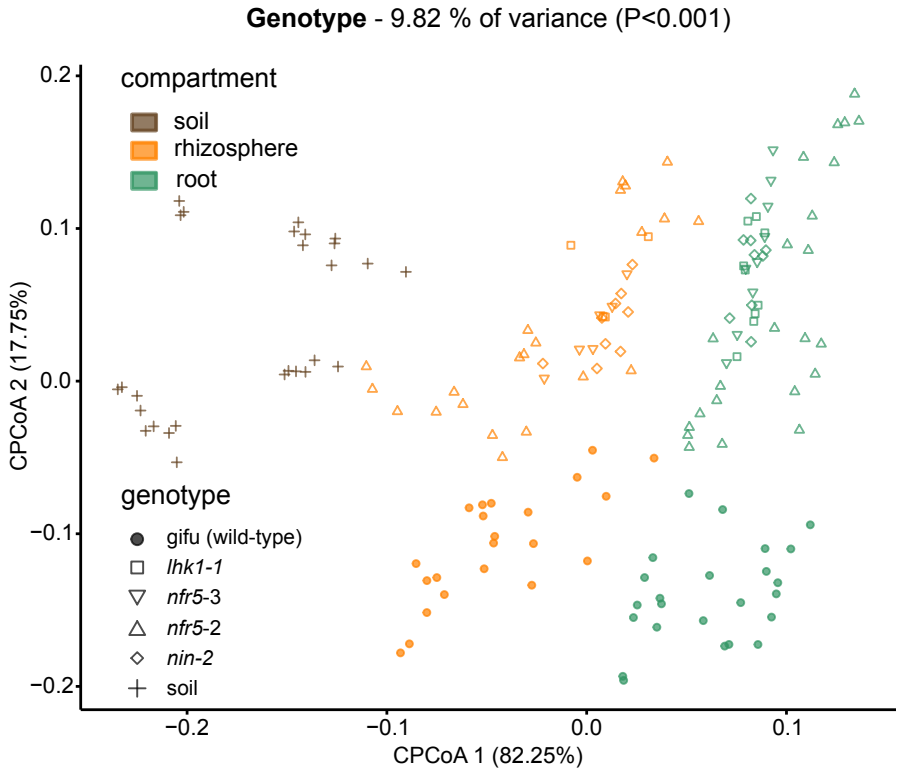
Is this also the case for taxa that are known to interact with the host (e.g. pathogens or symbionts)?

Nodule symbiosis impacts *Lotus* root community structure









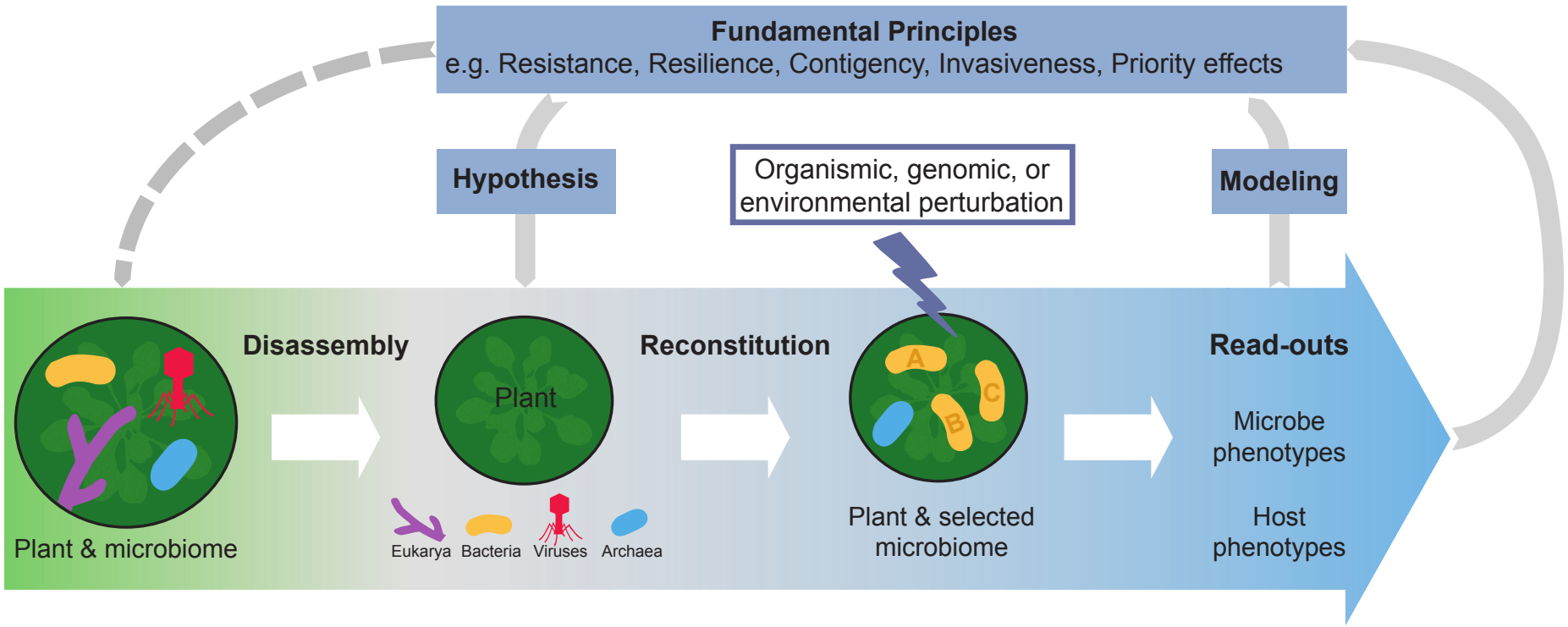
Zgadzaj *et al. PNAS*, 2016.

Community profiling by sequencing of natural communities remains descriptive - hypotheses can be made only based in correlations

Community profiling by sequencing of natural communities remains descriptive - hypotheses can be made only based in correlations

Is it possible to deconstruct the plant microbiota and design microbial communities of reduced complexity?

Synthetic community approach for plant microbiota research

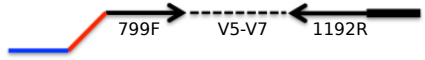


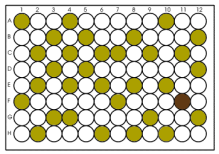
Adapted from Vorholt *et al.*, 2017

Arabidopsis thaliana leaf and root culture collections

Sequence-indexed Rhizobacterial Libraries (IRLs)

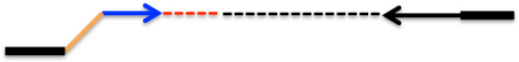
1. 1st PRC step - 96 well barcodes for each plate

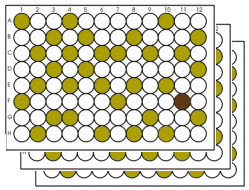




2. Pool PCR products for each plate

3. 2nd PRC step - one barcode for each plate

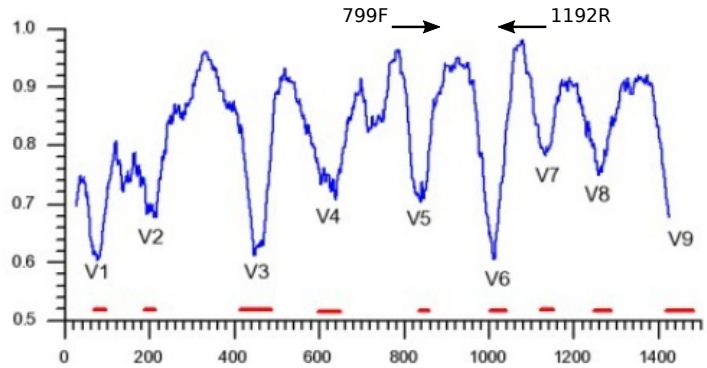




5. Pool and purify

6. Final PCR products for sequencing

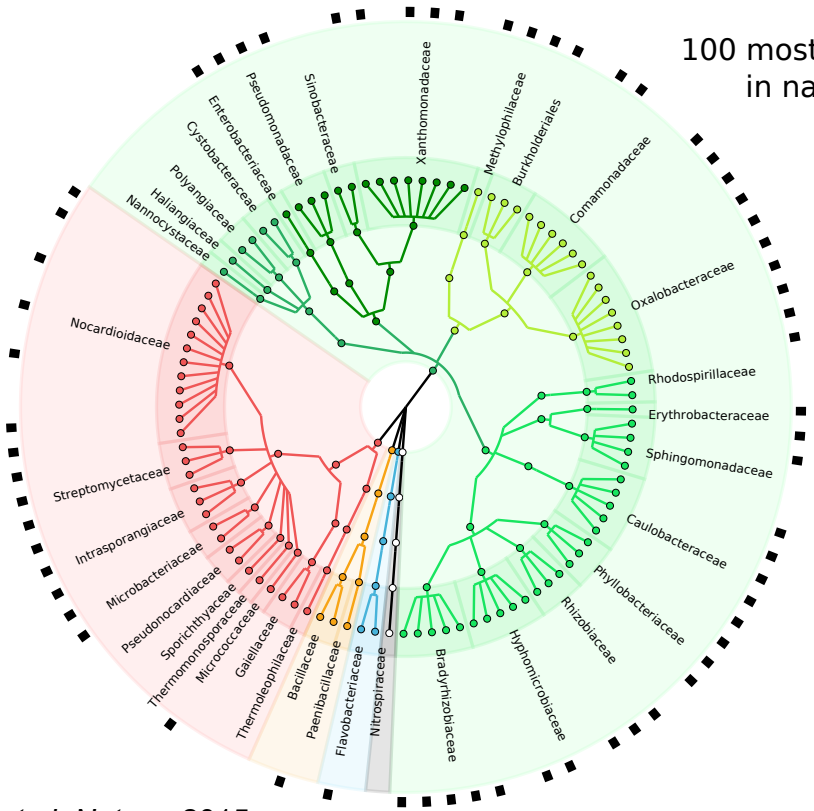
Conservation of the 16S rRNA gene

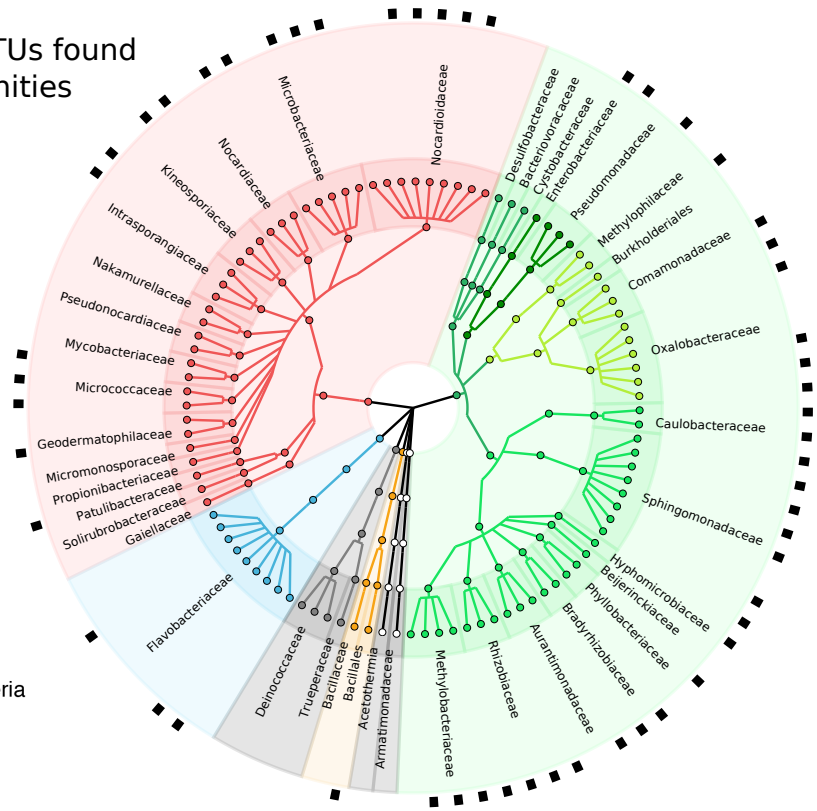


At-**R**SPHERE - 64% recovery rate - 5,812 isolates

At-**L**SPHERE - 54% recovery rate - 2,131 isolates

100 most abundant OTUs found in natural communities





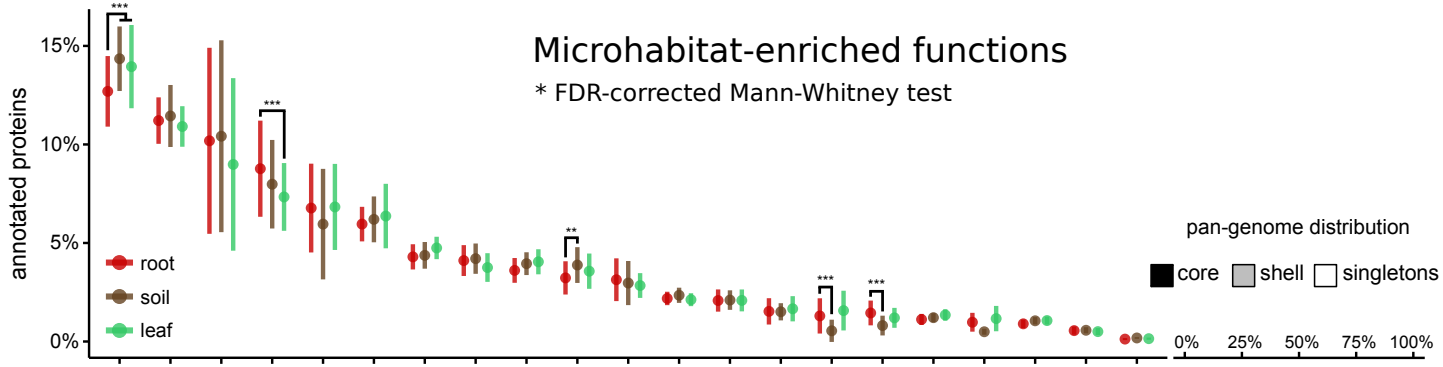
Actinobacteria
Bacteroidetes
Firmicutes
α
β
γ
Proteobacteria
others

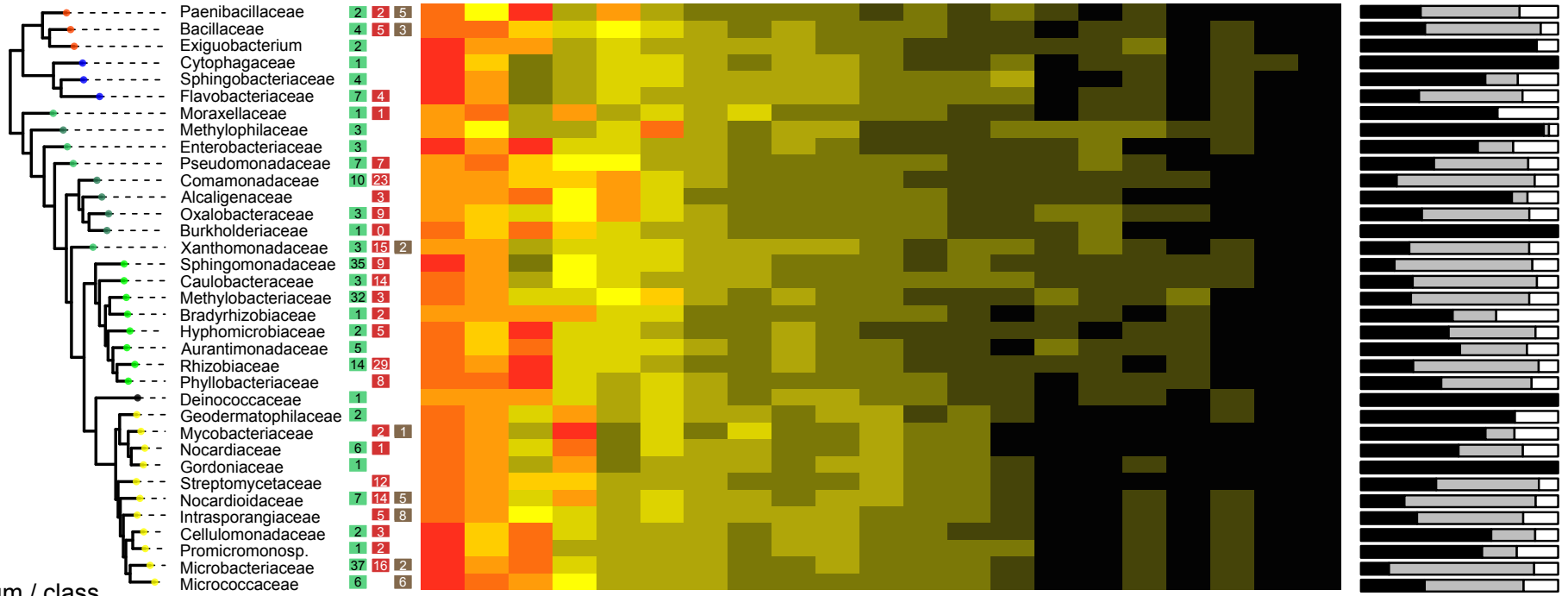
Bai *et al. Nature*, 2015.

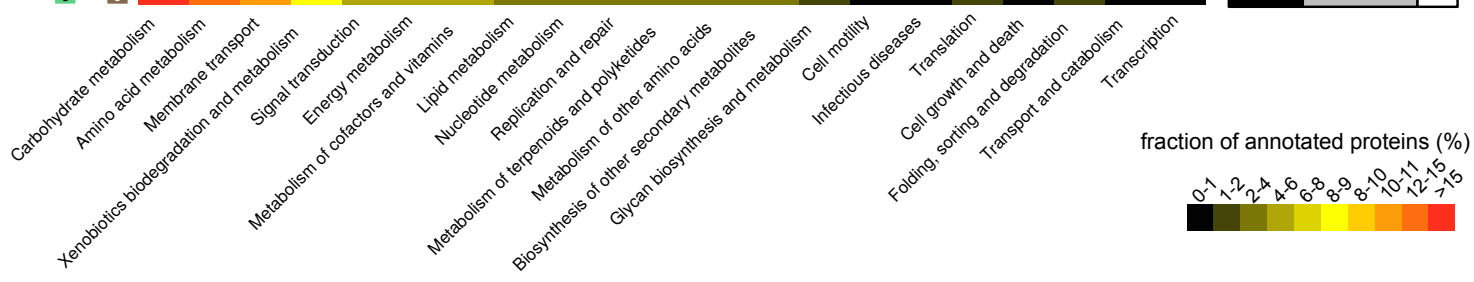
Comparative genomics of *Arabidopsis* sequenced isolates

At-SPHERE core collection

leaf - 206 isolates
root - 194 isolates
soil - 32 isolates







phylum / class

- Deinococcus-Thermus
- Actinobacteria
- Bacteroidetes
- Firmicutes
- Gammaproteobacteria
- Betaproteobacteria
- Alphaproteobacteria

Bai *et al*. *Nature*, 2015.

Culture collections of bacterial communities

Community-level collections

At-SPHERE

Arabidopsis thaliana

432 isolates

35 Families

5 Phyla

99,371 sequenced genes

54-64% natural community coverage

Lj-SPHERE

Lotus japonicus

308 isolates

21 Families

4 Phyla

77,400 sequenced genes

57% natural community coverage

Cr-SPHERE

C. rheinhardtii

184 isolates

18 Families

4 Phyla

52,829 sequenced genes

96% natural community coverage

Population-level collections

Rhizobiales

Multiple hosts

1,314 isolates

5 Families

1 Phylum

~100,000 sequenced genes

Pseudomonas

Multiple hosts

3,913 isolates

31 species

1 Phylum

~187,000 sequenced genes

Bacillus

Multiple hosts

7,463 isolates

19 species

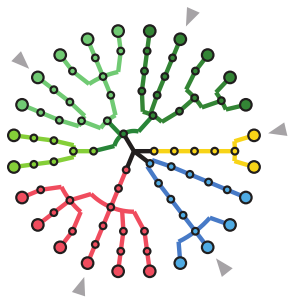
1 Phylum

~158,000 sequenced genes

Computational aspects of SynCom experiment design

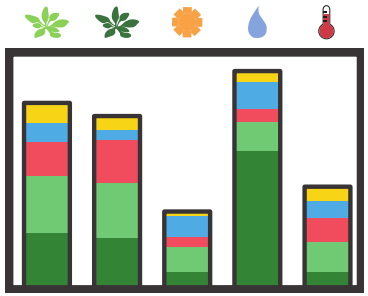
Synthetic community design

Selection based on phylogeny

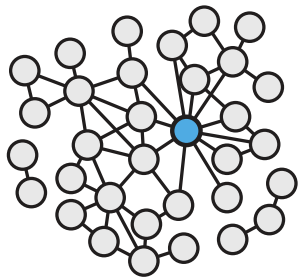


Example

Are communities stable across varying conditions or host genotypes?

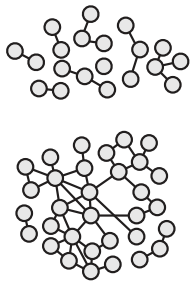


Selection based on interaction networks



Example

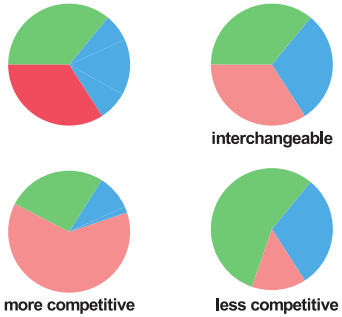
Are highly connected nodes essential for network stability?



Selection based on OTU clustering

Example

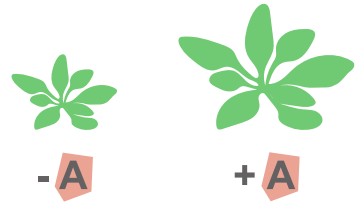
Are OTU exemplars interchangable?



Selection based on function

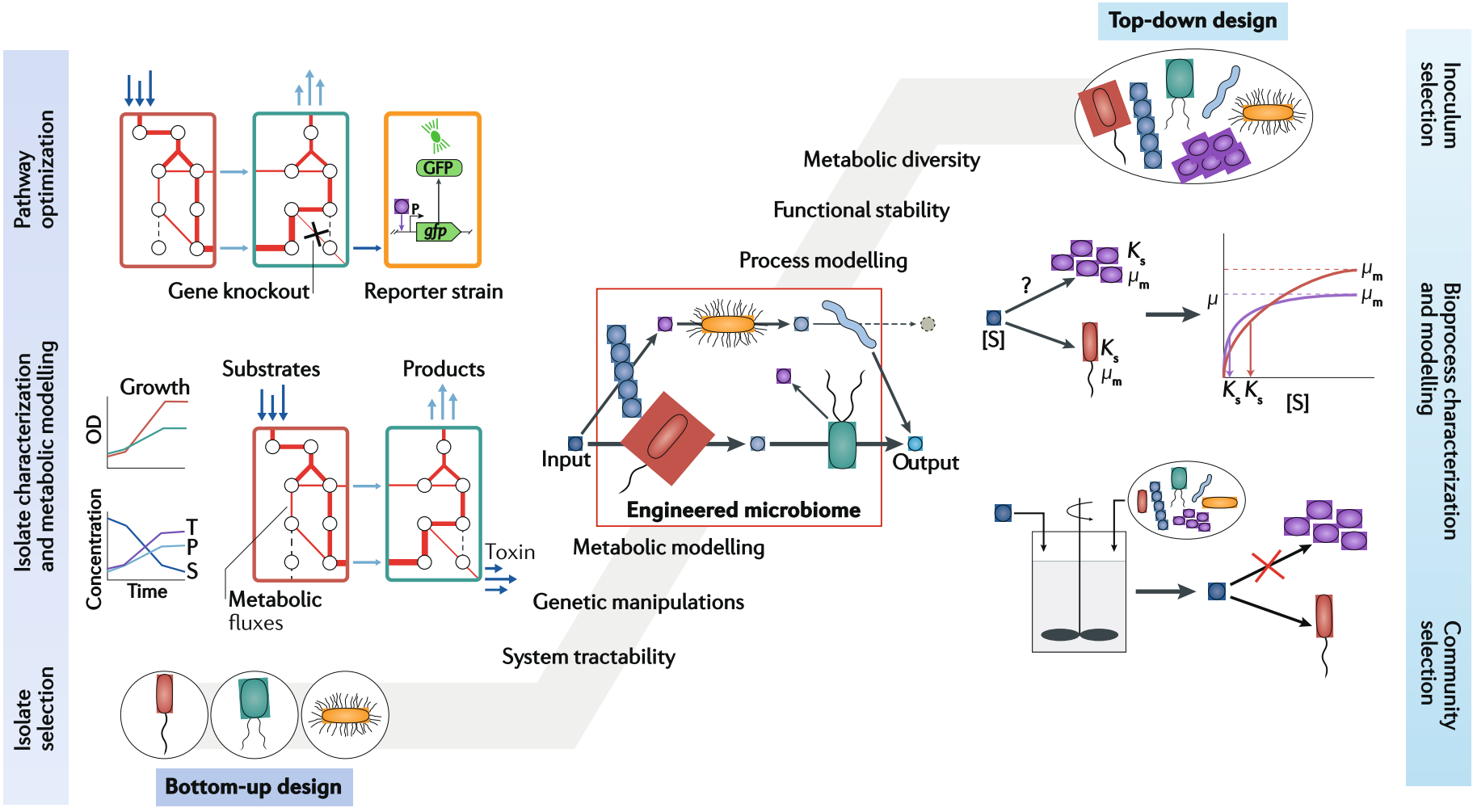
Example

What functions are beneficial for the host?



Adapted from Vorholt *et al.*, 2017

Designing synthetic microbiomes



Lawson *et al.*, 2019

Synthetic community design

Sample each independent SynCom input / start inoculum - relative abundances vs fold changes

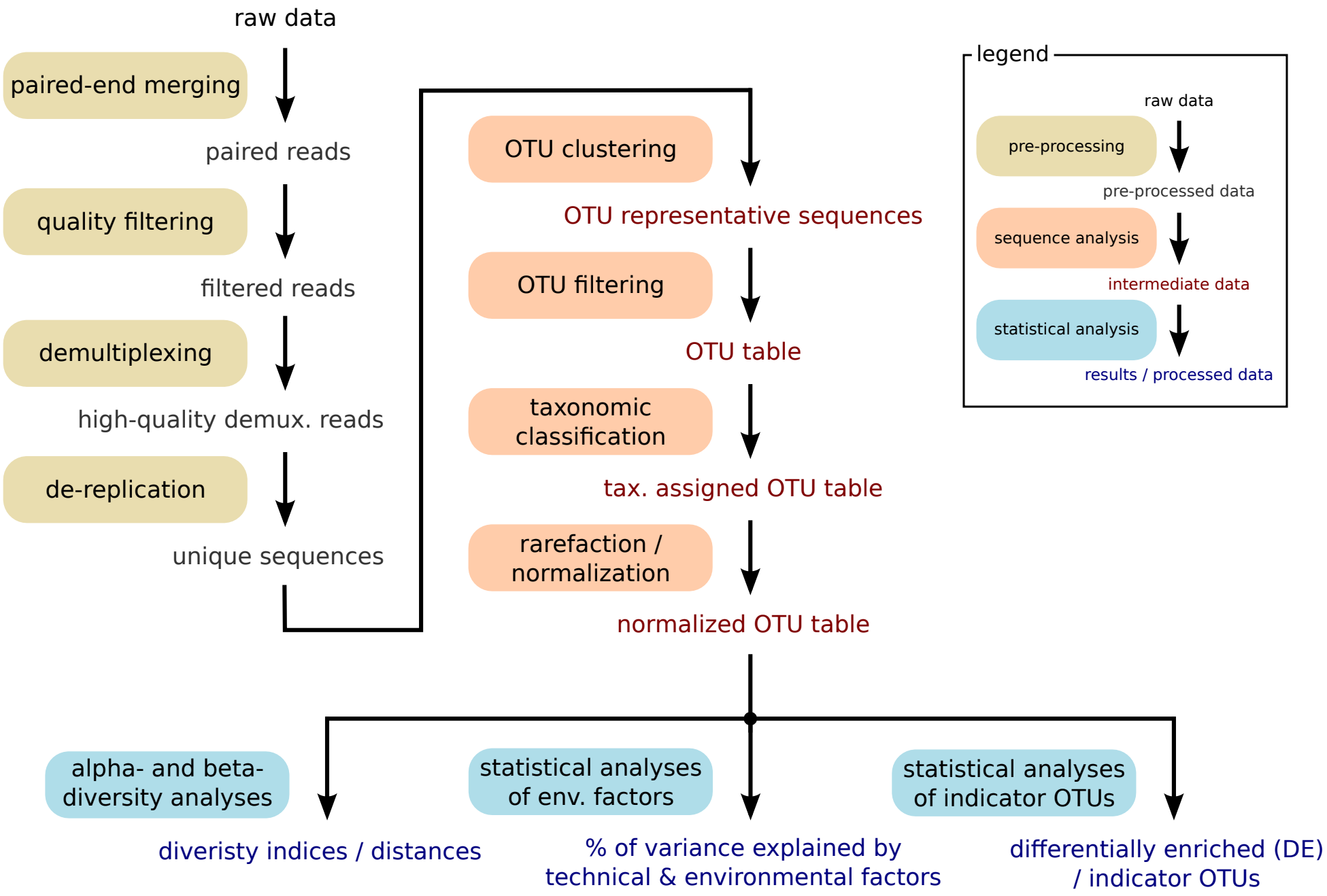
Multiple exemplars within the same species (16S variant) vs single representative strain (fully distinguishable)

Minimal (~6 members) vs complex SynComs (> 100 members)

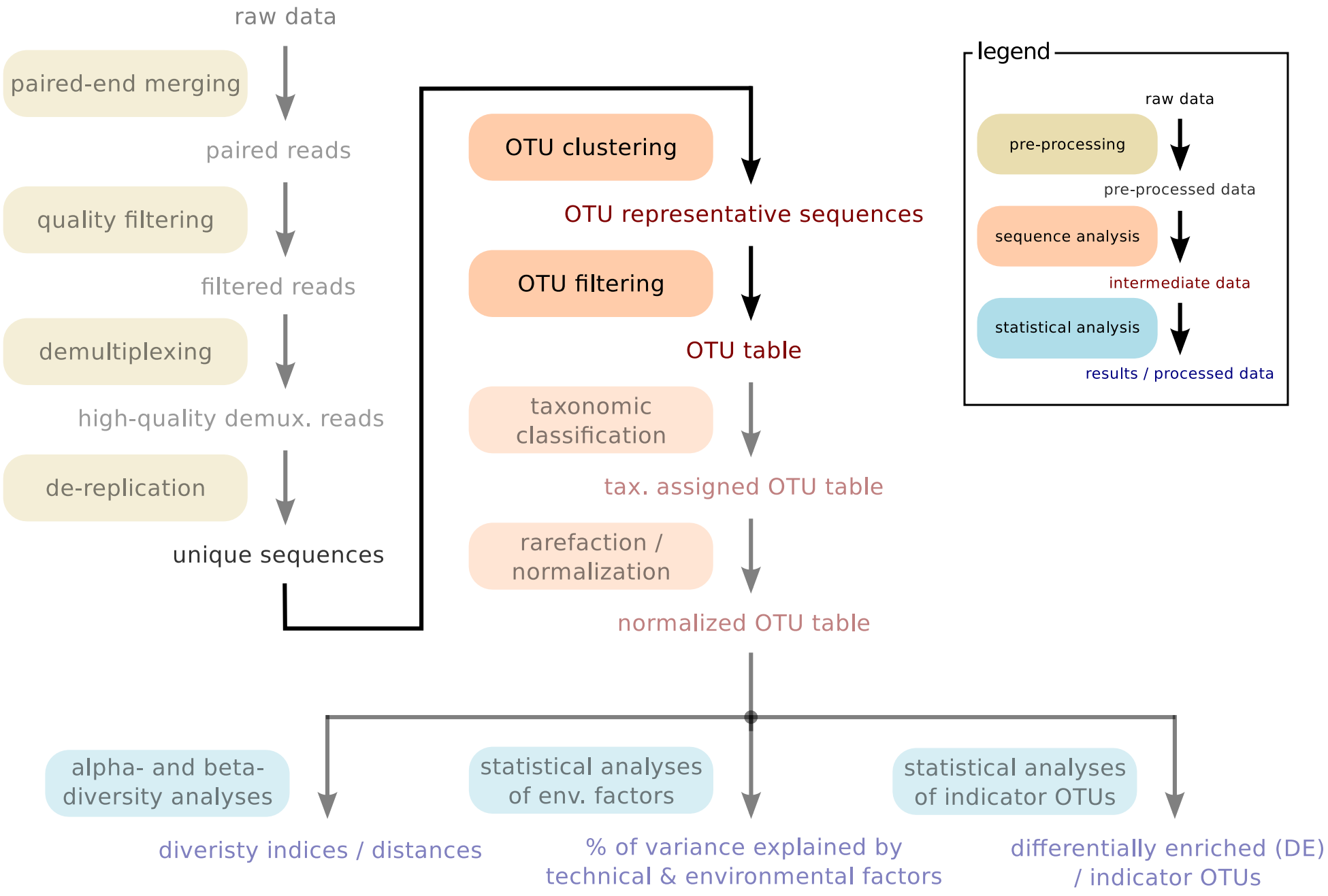
Further considereations: ecological context, origin of isolation, genome quality, purity, etc.

Computational tools for the analysis of SynCom data

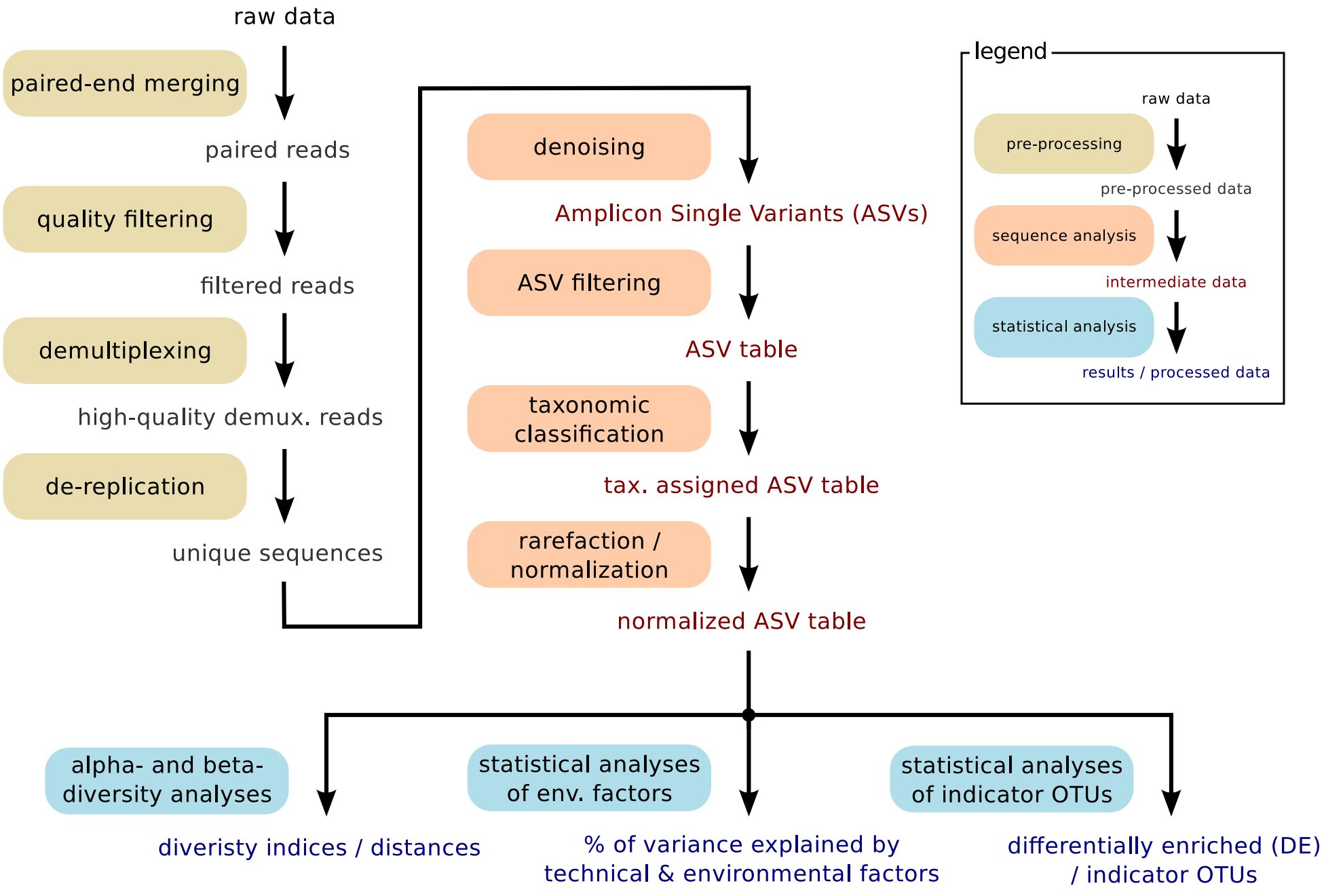
Workflow amplicon data analysis (natural community)



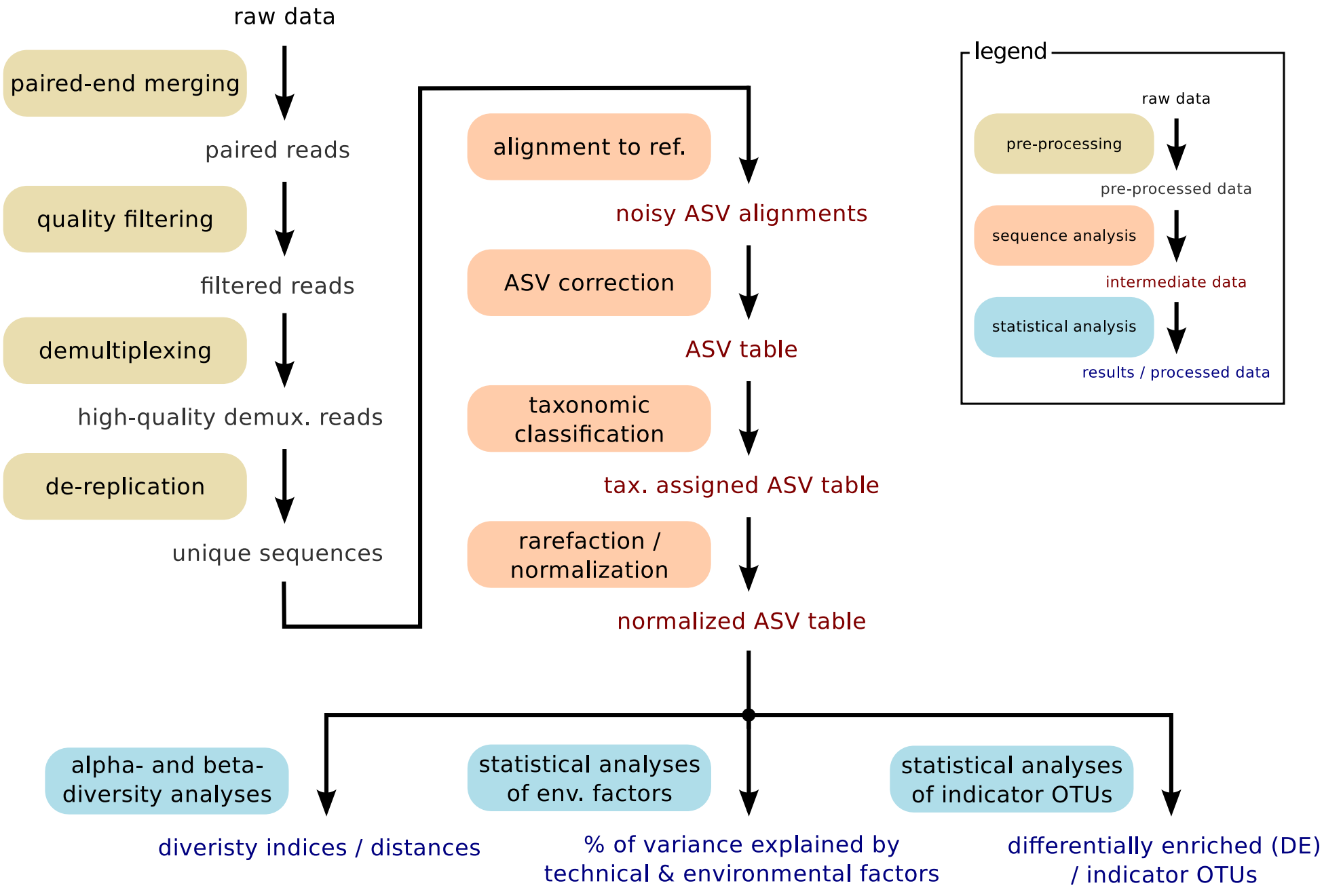
Workflow amplicon data analysis (natural community)



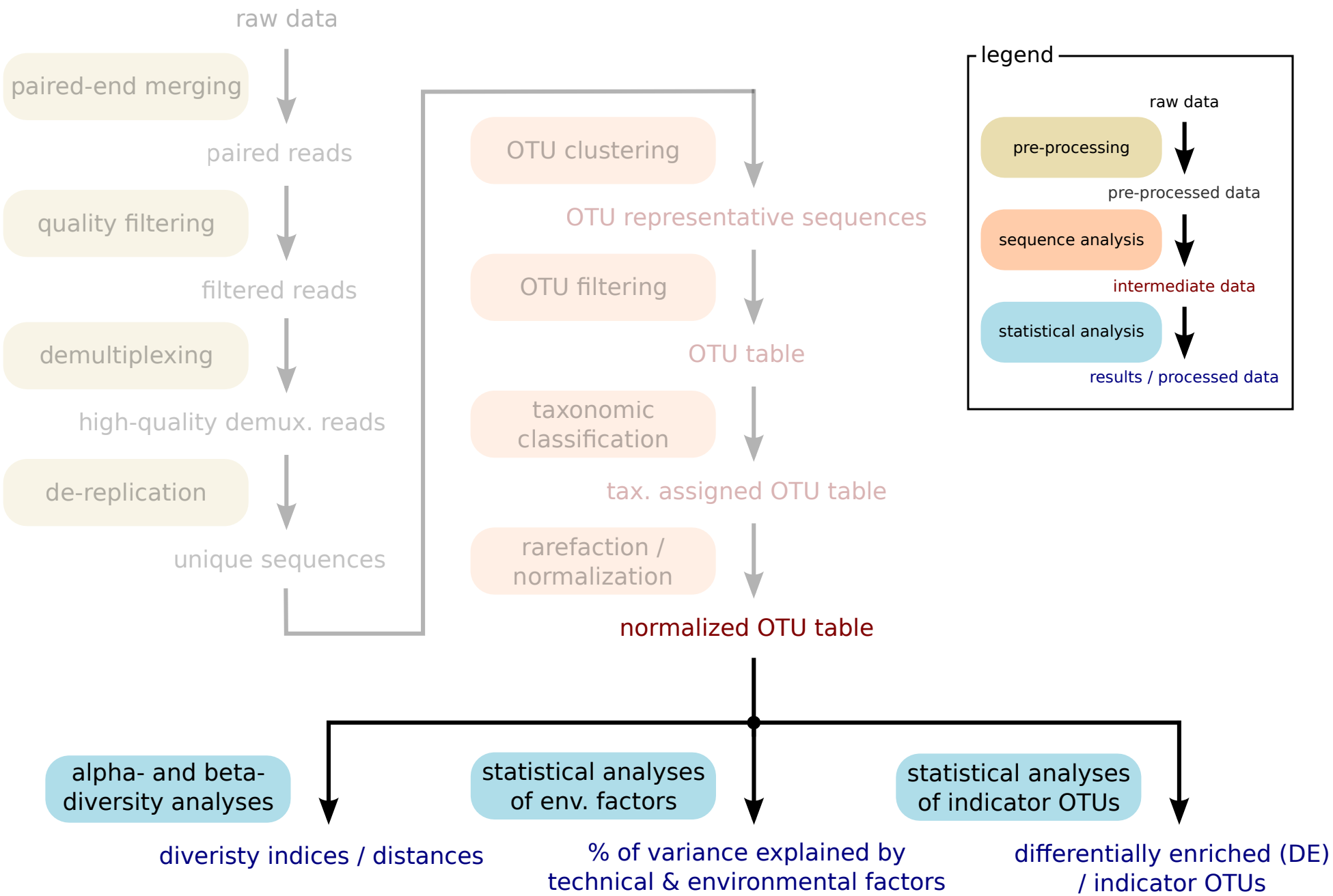
Workflow amplicon data analysis (natural community)



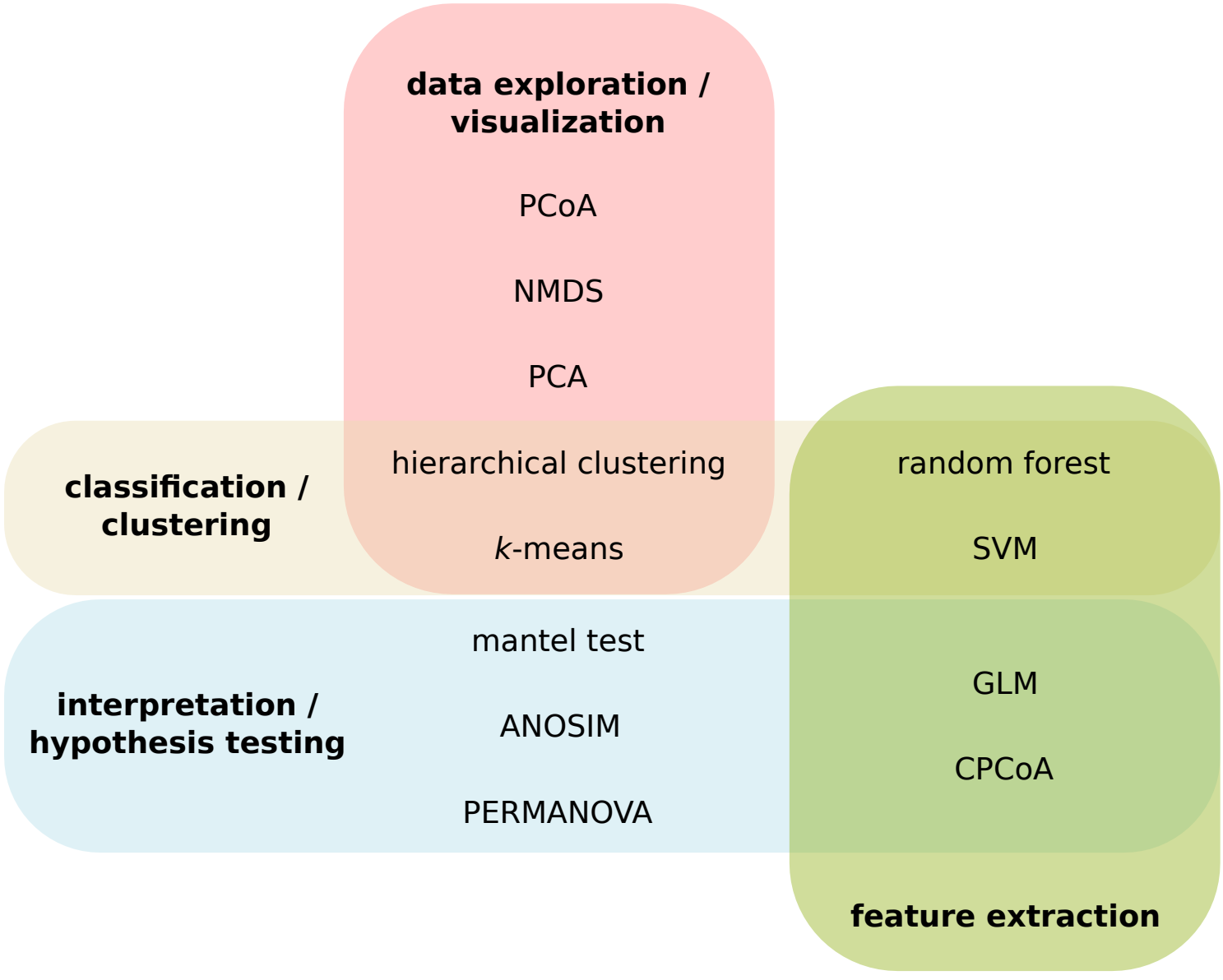
Workflow amplicon data analysis (SynCom)



Workflow amplicon data analysis (natural community)



Overview of common multivariate analysis methods in microbial ecology



Ordination methods in microbial ecology

PCA (Principal Component Analysis; Pearson, 1901)

consists on rotating the original system of coordinates to maximize dispersion
input are coordinates of datapoints in a high-dimensional space
most widely used and simple (fast) ordination method
R function: prcomp (stats)

PCoA (Principal Coordinate Analysis; Gower, 1966)

similar to PCA but first transforms distances into coordinates in a new space
input are pairwise distances between datapoints
popular in microbial ecology because it allows employing various distances
R function: cmdscale (stats)

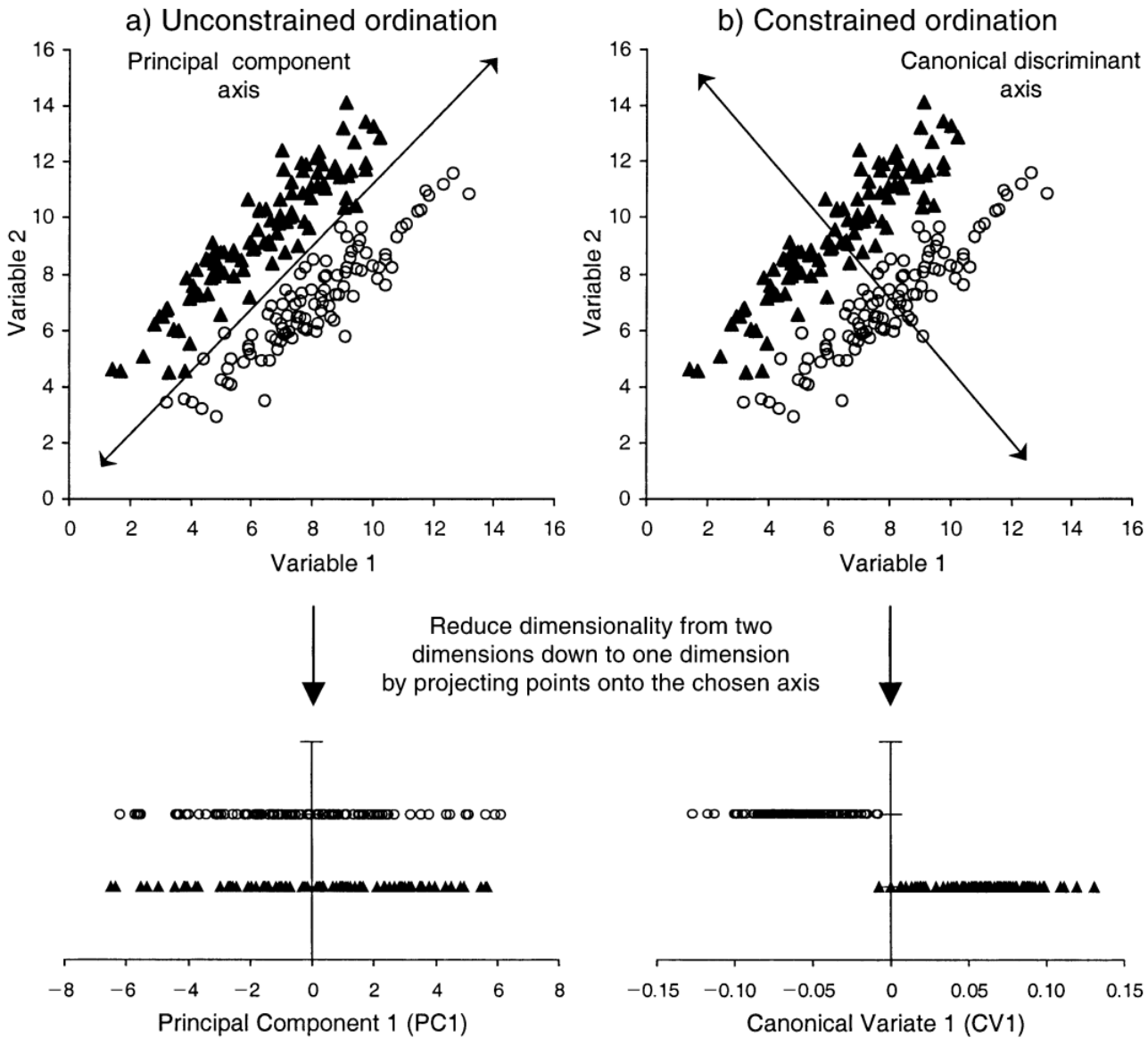
NMDS (Non-metric Multidimensional Scaling; Kruskal et al., 1964)

numerical rather than analytical method (slow(er), non-deterministic)
number of dimensions *k* are chosen *a priori*
all variance of the data is used to distribute points in a *k*-dimensional space
(Euclidean) distances in the new space are monotonically related to original distances
R function: isoMDS (MASS)

CPCoA (Constrained Principal Coordinate Analysis; Legendre and Legendre, 1998)

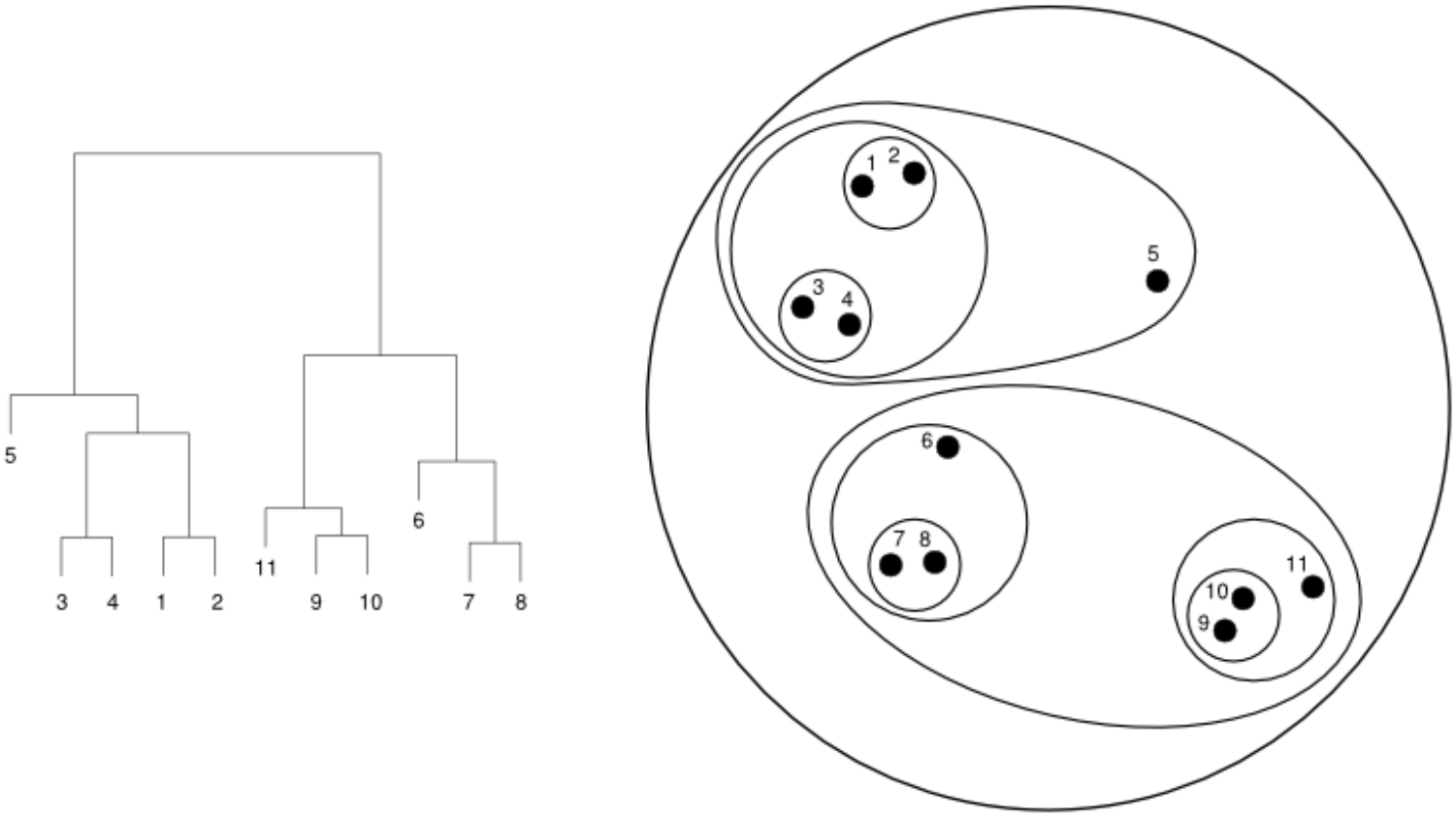
similar to PCoA but attempts to maximize separation between groups (env. variables)
used to address specific hypotheses (e.g. significant differences among groups)
statistical test of hypothesis by permutation procedures
R function: capscale (vegan)

Constrained vs. unconstrained ordination



Anderson and Willis, 2003

Hierarchical clustering



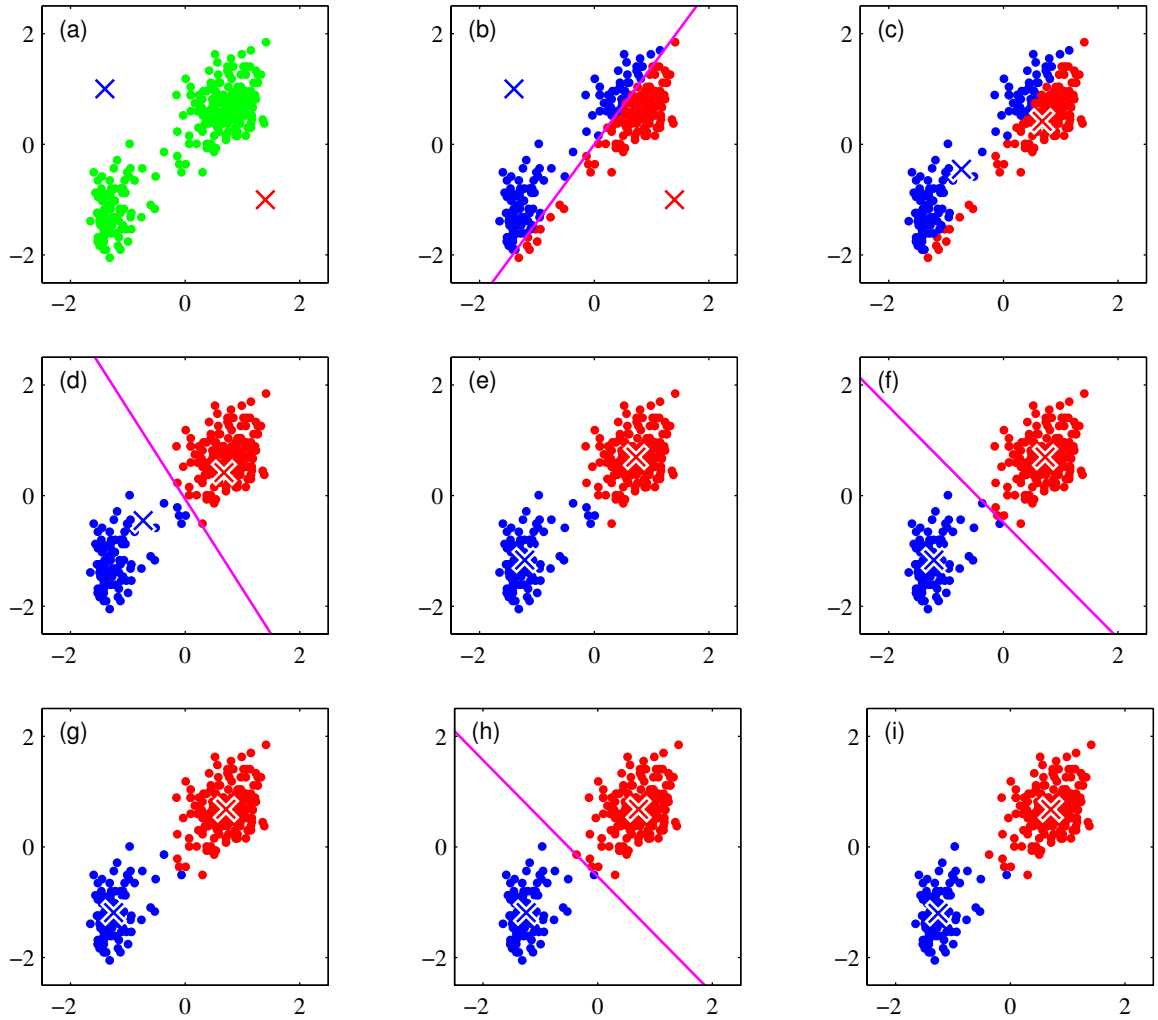
unsupervised clustering algorithm that can be used for data visualization

there are many variants (single linkage, average linkage, UPGMA, etc.)

fast and robust, can capture non-linear groups of datapoints

difficult to choose the number of clusters *k*

k-means clustering



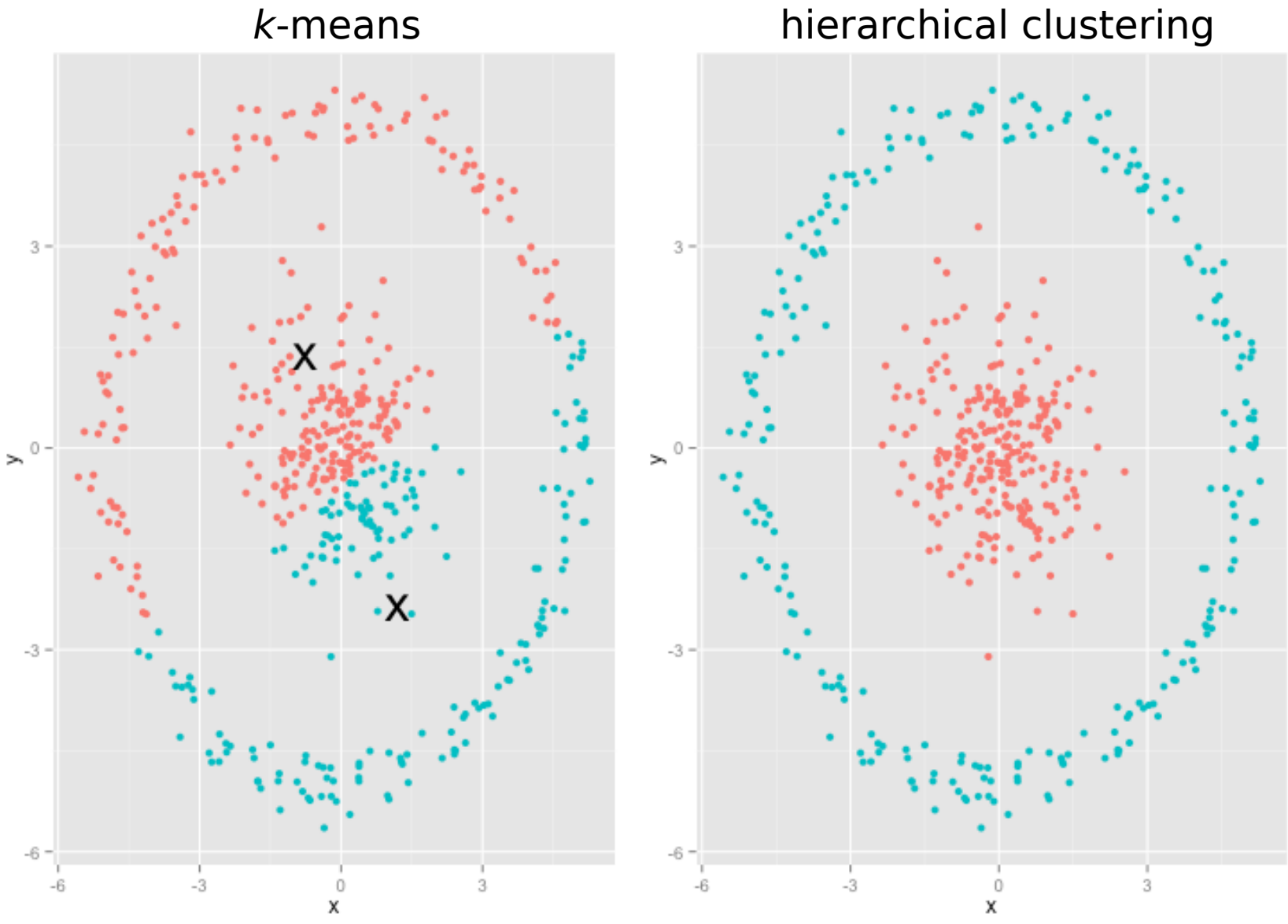
widely used, fast and robust unsupervised clustering method

depends on initialization of centroids

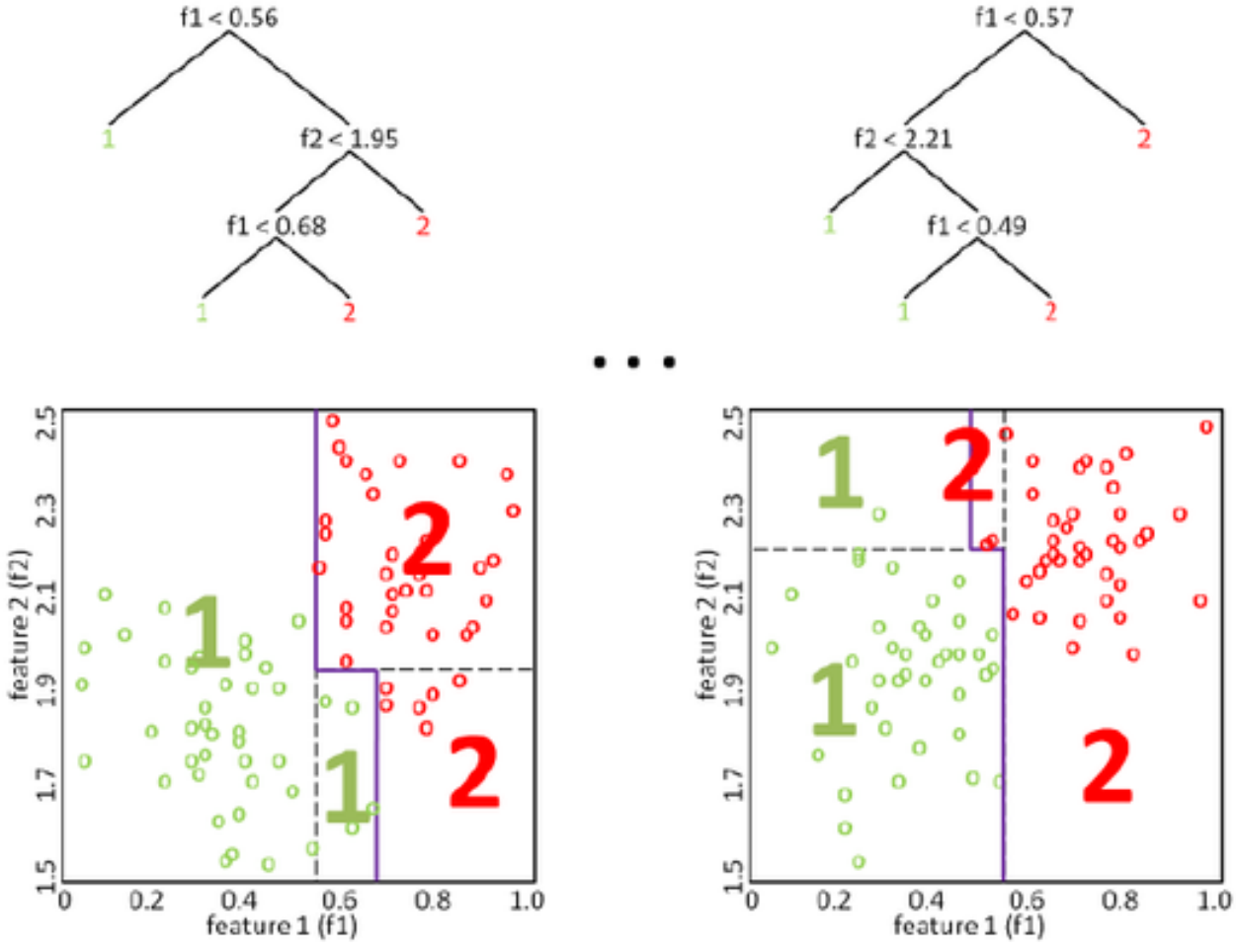
difficult to choose the number of clusters k

linear boundaries for classification

Linear vs. non-linear classification boundaries



Random Forests



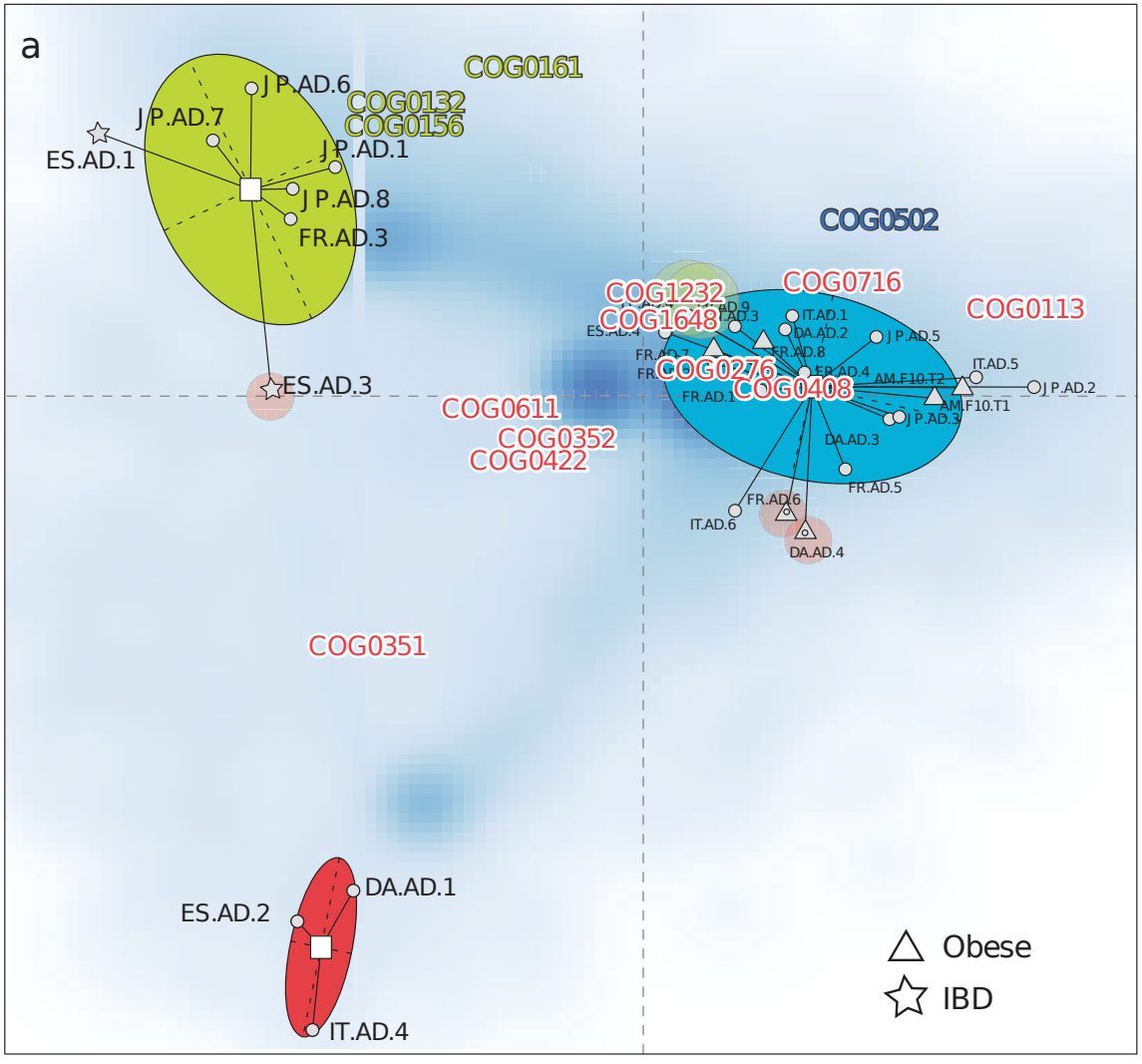
supervised clustering method that learns how to discriminate groups based on decision trees

classify observations into large groups based on predictor values (OTU abundances)

high classification accuracy, can be used for feature extraction (predictor OTUs)

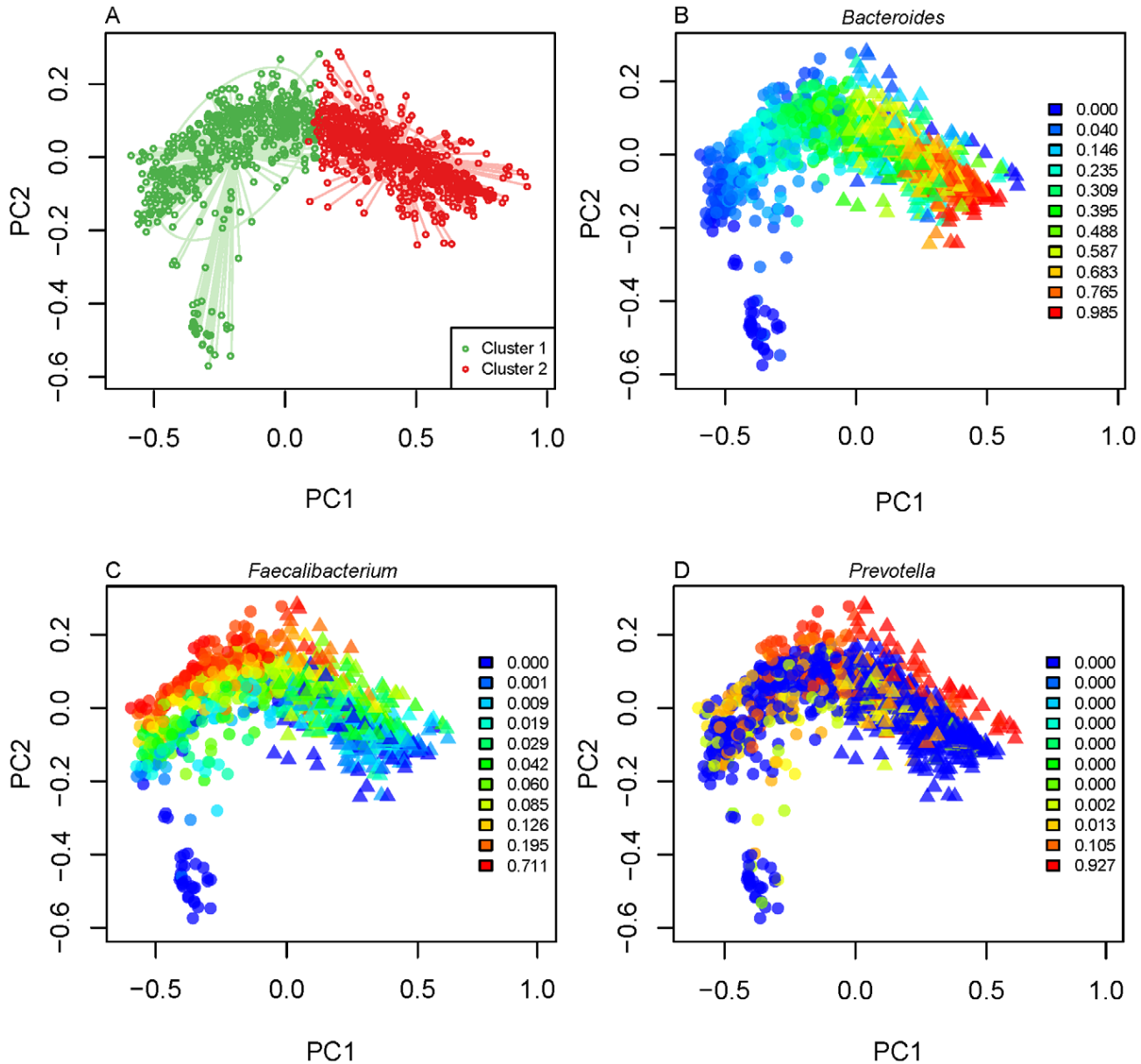
prone to over-fitting, requires cross-validation

Data exploration vs. classification



Enterotypes of the human gut microbiome
(Arumugam *et al.*, 2011)

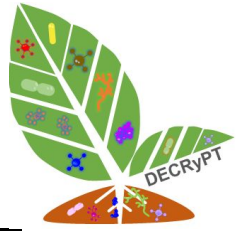
Data exploration vs. classification



Enterotypes of the human gut microbiome
(Koren *et al.*, 2013)

Workshop introduction and content outline

Day 1 - Monday, October 28



16:00 – 17:00	Introductory lecture Ruben Garrido-Oter, MPIPZ Cologne
17:00 – 18:30	Hands-on session: Introduction to bash scripting for sequencing data analysis Eik Dahms, Ruben Garrido-Oter, MPIPZ Cologne
18:30	**Joint Evening, Pizza in the TATA Bar**

Day 2 - Monday, October 29

09:00 – 09:15	Briefing and introduction Ruben Garrido-Oter, MPIPZ Cologne
09:15 – 09:45	Short talk: State-of-the-art approaches for amplicon data analyses of synthetic and natural communities Pengfan Zhang, Ruben Garrido-Oter, MPIPZ Cologne
09:45 – 10:40	Hands-on session: Read quality control, sample demultiplexing, error correction and OTU clustering Pengfan Zhang, Ruben Garrido-Oter, MPIPZ Cologne
10:40 – 10:55	**Coffee Break**

Day 2 - Monday, October 29

10:55 – 11:30	Short talk: Taxonomic characterization and diversity analyses Pengfan Zhang, Ruben Garrido-Oter, MPIPZ Cologne
11:30 – 12:30	Hands-on session: Taxonomic classification and indices of alpha and beta diversity Pengfan Zhang, Ruben Garrido-Oter, MPIPZ Cologne
12.30 – 13:15	**Lunch Break at canteen**

Day 2 - Monday, October 29

13:15 – 13:45	Short talk: Diversity analyses and community data visualization Ruben Garrido-Oter, Pengfan Zhang, MPIPZ
13:45 – 15:15	Hands-on session: Alpha and beta diversity analyses and ordination methods Ruben Garrido-Oter, Pengfan Zhang, MPIPZ Cologne
15:15 – 15:30	**Coffee break**

Day 2 - Monday, October 29

15:30 – 16:30	Hands-on session: Data normalization and enrichment tests Pengfan Zhang, Ruben Garrido-Oter, MPIPZ Cologne
16:30 – 16:45	**Short break**
16:45 – 18:00	Hands-on session: Constrained ordination methods and statistical approaches to assess the effect of environmental variables Ruben Garrido-Oter, Pengfan Zhang, MPIPZ Cologne
19:00	**Joint Evening at “Brauhaus Pütz”, Engelbertstraße 67, 50674 Köln**

Day 3 - Monday, October 30

09:00 – 09:15	Briefing and introduction Ruben Garrido-Oter, MPIPZ Cologne
09:15 – 09:45	Short talk: Overview of RNA-Seq data analysis workflows Yulong Niu, MPIPZ, Cologne
09:45 – 10:40	Hands-on session: Read quality assessment and *de novo* transcriptome assembly Fantin Mesny, Yulong Niu, MPIPZ, Cologne
10:40 – 10:55	**Coffee Break**

Day 3 - Monday, October 30

10:55 – 11:15	Short talk: Overview of RNA-Seq data alignment methods Yulong Niu, Fantin Mesny, MPIPZ, Cologne
11:15 – 12:30	Hands-on session: Read alignment and pseudo-alignment Yulong Niu, Fantin Mesny, MPIPZ, Cologne
12.30 – 13:15	**Lunch Break at canteen**

Day 3 - Monday, October 30

13:15 – 13:45	Short talk: Methods for differential expression analysis of RNA-Seq data Fantin Mesny, Yulong Niu, MPIPZ, Cologne
13:45 – 15:15	Hands-on session: Differential expression analysis in R Fantin Mesny, Yulong Niu, MPIPZ, Cologne
15:15 - 15:30	**Coffee Break**

Day 3 - Monday, October 30

15:30 – 15:45	Short talk: RNA-Seq data visualization Yulong Niu, Fantin Mesny, MPIPZ, Cologne
15:45 – 16:45	Hands-on session: Expression data visualization in R Yulong Niu, Fantin Mesny, MPIPZ, Cologne
16:45 – 17:00	Final remarks and farewell Ruben Garrido-Oter, MPIPZ, Cologne
17:00	**End**

Setting up

Laptop login

user: spp
password: Winter!2

Server logins

user: spp01 - spp30
password: Winter!2

github repository:

https://github.com/YulongNiu/MPIPZ_SPP_workshop/