INSELGRUPPE

# Microbiota-Analysis

Jacqueline Wyss, 21.06.2023

# Overview

1. Definitions
2. Sample processing
3. Data analysis
    - Alpha diversity
    - Beta diversity
    - Taxonomy
    - Differential abundance
    - Correlation
    - Models/classification

# Introduction:
# Microbiology vs Microbiota-Studies

Culturing based

Sequencing based

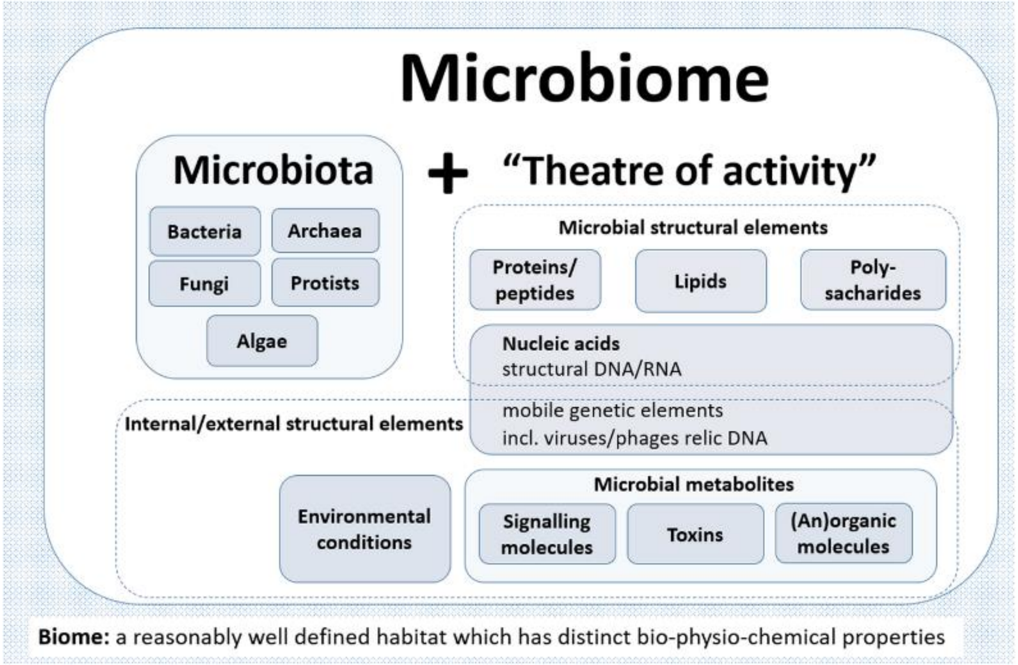
# Introduction: Definitions

Microbiota:

Microbiome:

Meta

- Genomics
- Transcriptomics
- Metabolomics



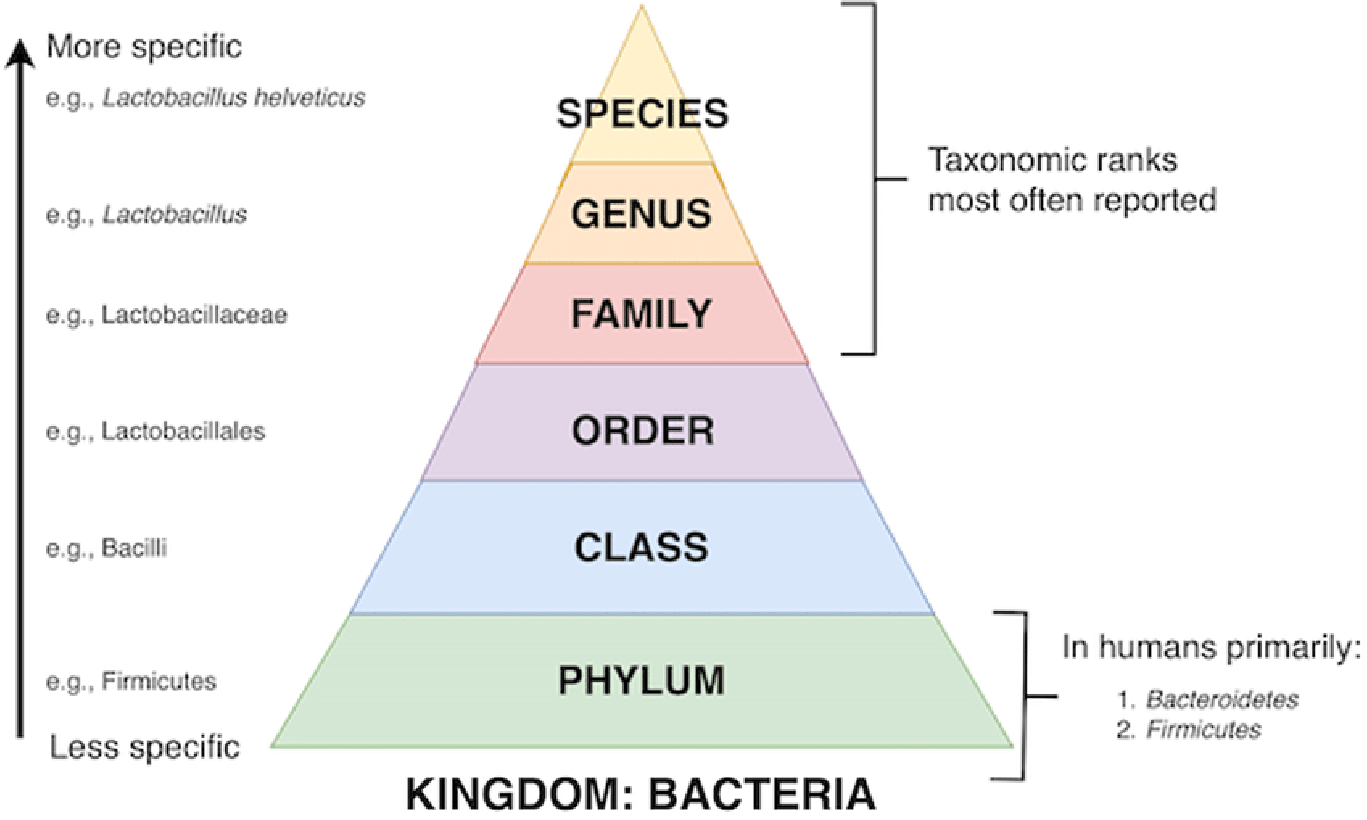
Berg et al. Microbiome 2020

# Introduction: Definitions

Taxonomy:

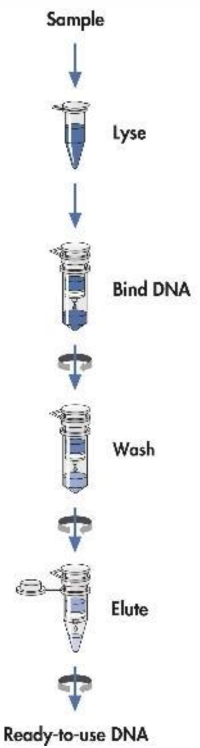
- Species: Strains
- Substrains

Taxa: no definition of the taxonomic level



Ameringen et al. Depression and Anxiety 2019

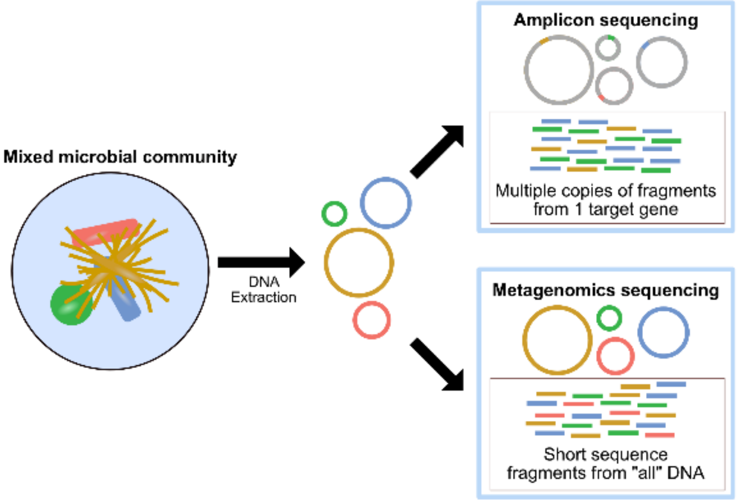
# Workflow:
# Sample preparation



DNA is extracted from the samples

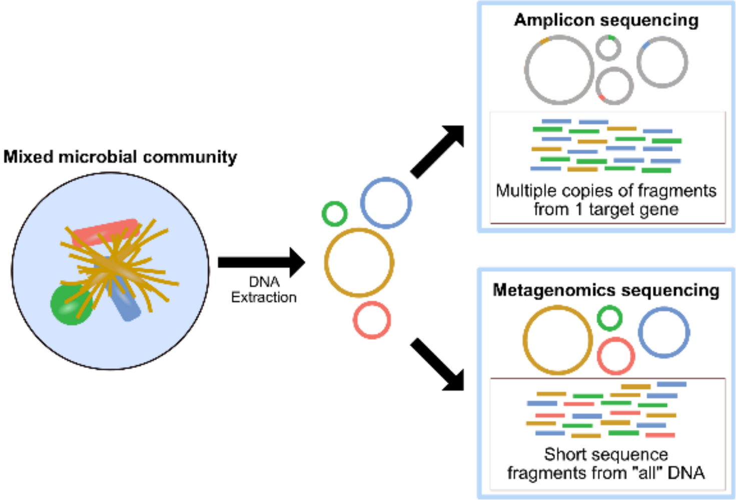
DNA can then be used for sequencing

# Workflow:
# Amplicon vs full metagenomic sequencing


Mixed microbial community
DNA Extraction
Amplicon sequencing
Multiple copies of fragments from 1 target gene
Metagenomics sequencing
Short sequence fragments from "all" DNA

https://astrobiomike.github.io

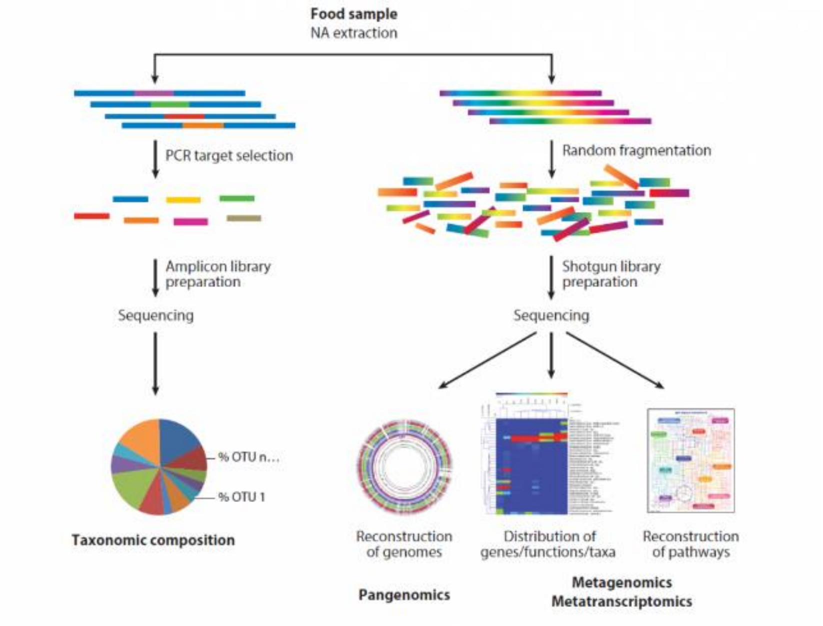
# Workflow:
# Amplicon vs full metagenomic sequencing


Mixed microbial community
DNA Extraction
Amplicon sequencing
Multiple copies of fragments from 1 target gene
Metagenomics sequencing
Short sequence fragments from "all" DNA

https://astrobiomike.github.io

In microbiota studies most often a hypervariable region of the 16S subunit is used as a target gene for amplification

# Workflow:
# Amplicon vs full metagenomic sequencing



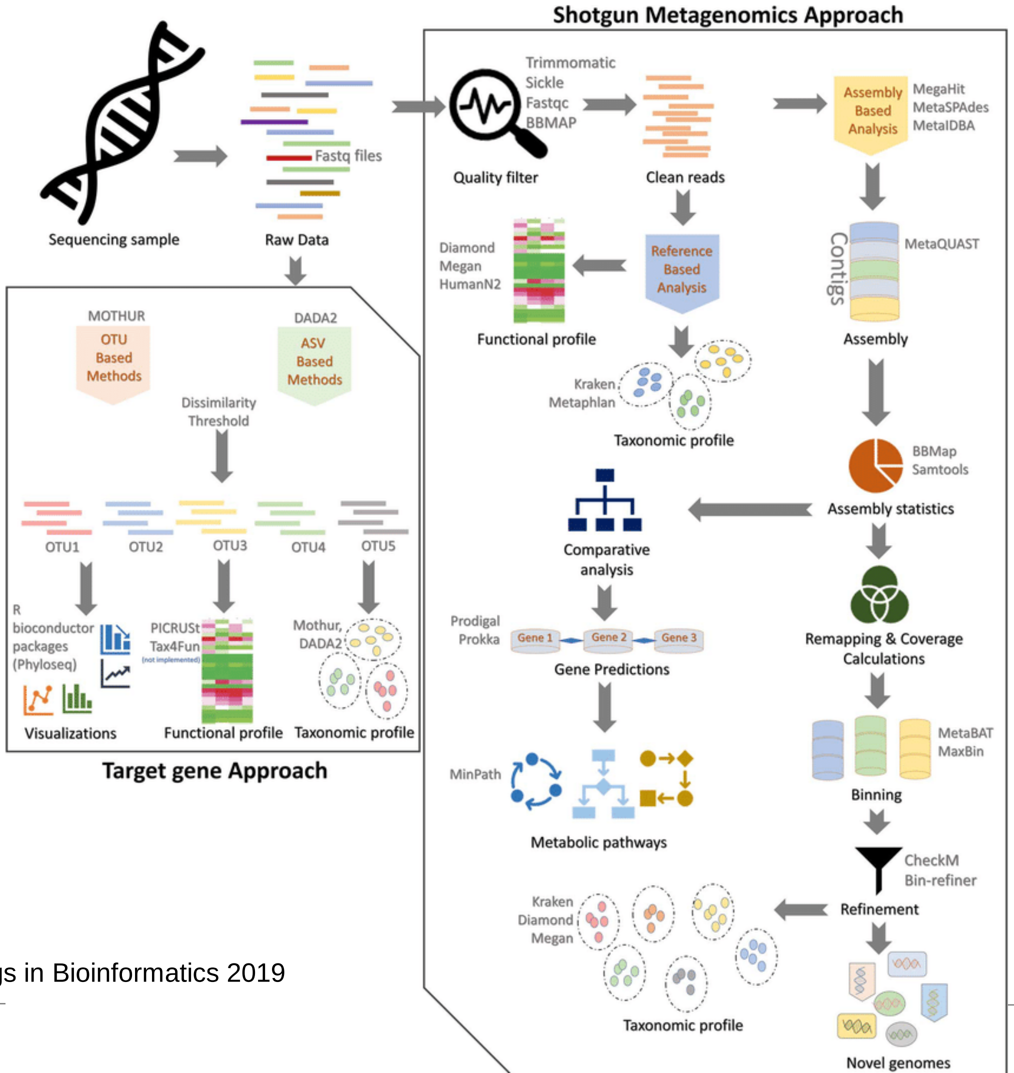
16S sequencing = amplicon sequencing

We mainly get information about the taxa composition

De Filippis F., et al, 2018

# Workflow: Sequencing analysis

Raw sequencing data is analysed by bioinformatic pipelines and packages



Bharti et al. Briefings in Bioinformatics 2019

# Raw sequencing data: FASTQ files

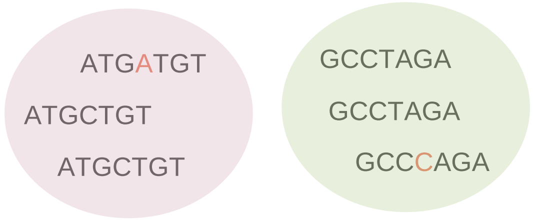
```
@DMJVU:00004:00006
TTCAATTGGCATTAGATACCCTGGTAGTCCACGCCGTAAACGATGTCGAC
TTGGAGGTTGTGCCCTTG
+
;75505057;66ACDCCCCC?DC6<;;;;666;606666,6666666666606066060666--
)-)-
@DMJVU:00004:00009
CTAGGAACCGCATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGT
GCTAGGTGTTAGACCCTTTCCGGGGTTTAGTGCCGCAGCTAACGCATTA
AGCACTCCGCCTGGGGAGTACGACCGCAAGGTTGAAACTCAAAGGAATT
GACGG
+
;;5D0;7606=66;@CD???B:;;7;B@@B7<<<A>A;;;;;6@@;6>=;4==@0;;;05;49C
*44*4.4==*44*4444EAC?5:8//)--
//:3@39@<;;;7;C6;;@B7@CBACC::4:@@<@=B5:::/::::2*2.239.-----
```

Sample identifier

Sequencing result

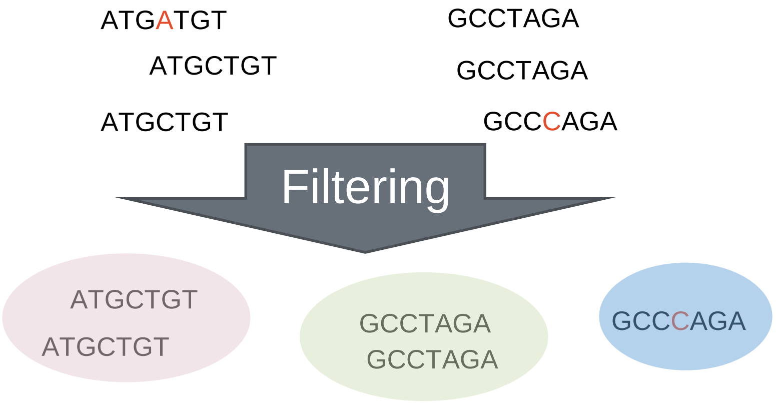
Quality score

# Workflow:
# Grouping reads



## OTU: operational taxonomic units

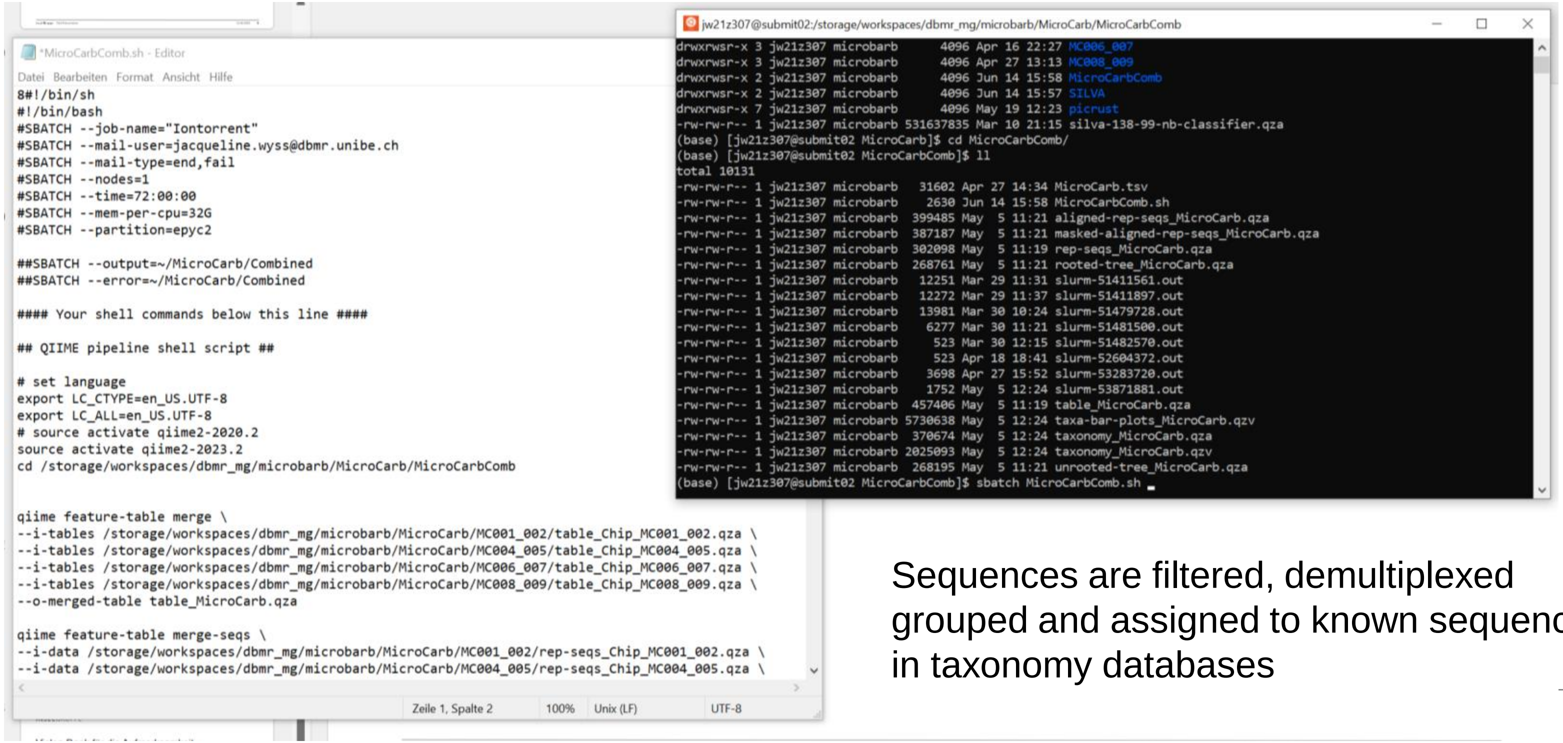
Grouping by consensus (e.g. 97%)



## ASV: amplicon sequencing variants

Denoising and quality filtering before grouping exact matches

# Workflow:
# Scripts and Server



*MicroCarbComb.sh - Editor
Datei Bearbeiten Format Ansicht Hilfe
Zeile 1, Spalte 2
100%
Unix (LF)
UTF-8

```
8#!/bin/sh
#!/bin/bash
#SBATCH --job-name="Iontorrent"
#SBATCH --mail-user=jacqueline.wyss@dbmr.unibe.ch
#SBATCH --mail-type=end,fail
#SBATCH --nodes=1
#SBATCH --time=72:00:00
#SBATCH --mem-per-cpu=32G
#SBATCH --partition=epyc2

##SBATCH --output=~/MicroCarb/Combined
##SBATCH --error=~/MicroCarb/Combined

#### Your shell commands below this line ####

## QIIME pipeline shell script ##

# set language
export LC_CTYPE=en_US.UTF-8
export LC_ALL=en_US.UTF-8
# source activate qiime2-2020.2
source activate qiime2-2023.2
cd /storage/workspaces/dbmr_mg/microbarb/MicroCarb/MicroCarbComb


qiime feature-table merge \
--i-tables /storage/workspaces/dbmr_mg/microbarb/MicroCarb/MC001_002/table_Chip_MC001_002.qza \
--i-tables /storage/workspaces/dbmr_mg/microbarb/MicroCarb/MC004_005/table_Chip_MC004_005.qza \
--i-tables /storage/workspaces/dbmr_mg/microbarb/MicroCarb/MC006_007/table_Chip_MC006_007.qza \
--i-tables /storage/workspaces/dbmr_mg/microbarb/MicroCarb/MC008_009/table_Chip_MC008_009.qza \
--o-merged-table table_MicroCarb.qza

qiime feature-table merge-seqs \
--i-data /storage/workspaces/dbmr_mg/microbarb/MicroCarb/MC001_002/rep-seqs_Chip_MC001_002.qza \
--i-data /storage/workspaces/dbmr_mg/microbarb/MicroCarb/MC004_005/rep-seqs_Chip_MC004_005.qza \
```

jw21z307@submit02:/storage/workspaces/dbmr_mg/microbarb/MicroCarb/MicroCarbComb

```
drwxrwsr-x 3 jw21z307 microbarb      4096 Apr 16 22:27 MC006_007
drwxrwsr-x 3 jw21z307 microbarb      4096 Apr 27 13:13 MC008_009
drwxrwsr-x 2 jw21z307 microbarb      4096 Jun 14 15:58 MicroCarbComb
drwxrwsr-x 2 jw21z307 microbarb      4096 Jun 14 15:57 SILVA
drwxrwsr-x 7 jw21z307 microbarb      4096 May 19 12:23 picrust
-rw-rw-r-- 1 jw21z307 microbarb 531637835 Mar 10 21:15 silva-138-99-nb-classifier.qza
(base) [jw21z307@submit02 MicroCarb]$ cd MicroCarbComb/
(base) [jw21z307@submit02 MicroCarbComb]$ ll
total 10131
-rw-rw-r-- 1 jw21z307 microbarb   31602 Apr 27 14:34 MicroCarb.tsv
-rw-rw-r-- 1 jw21z307 microbarb    2630 Jun 14 15:58 MicroCarbComb.sh
-rw-rw-r-- 1 jw21z307 microbarb  399485 May  5 11:21 aligned-rep-seqs_MicroCarb.qza
-rw-rw-r-- 1 jw21z307 microbarb  387187 May  5 11:21 masked-aligned-rep-seqs_MicroCarb.qza
-rw-rw-r-- 1 jw21z307 microbarb  302098 May  5 11:19 rep-seqs_MicroCarb.qza
-rw-rw-r-- 1 jw21z307 microbarb  268761 May  5 11:21 rooted-tree_MicroCarb.qza
-rw-rw-r-- 1 jw21z307 microbarb   12251 Mar 29 11:31 slurm-51411561.out
-rw-rw-r-- 1 jw21z307 microbarb   12272 Mar 29 11:37 slurm-51411897.out
-rw-rw-r-- 1 jw21z307 microbarb   13981 Mar 30 10:24 slurm-51479728.out
-rw-rw-r-- 1 jw21z307 microbarb    6277 Mar 30 11:21 slurm-51481500.out
-rw-rw-r-- 1 jw21z307 microbarb     523 Mar 30 12:15 slurm-51482570.out
-rw-rw-r-- 1 jw21z307 microbarb     523 Apr 18 18:41 slurm-52604372.out
-rw-rw-r-- 1 jw21z307 microbarb    3698 Apr 27 15:52 slurm-53283720.out
-rw-rw-r-- 1 jw21z307 microbarb    1752 May  5 12:24 slurm-53871881.out
-rw-rw-r-- 1 jw21z307 microbarb  457406 May  5 11:19 table_MicroCarb.qza
-rw-rw-r-- 1 jw21z307 microbarb 5730638 May  5 12:24 taxa-bar-plots_MicroCarb.qzv
-rw-rw-r-- 1 jw21z307 microbarb  370674 May  5 12:24 taxonomy_MicroCarb.qza
-rw-rw-r-- 1 jw21z307 microbarb 2025093 May  5 12:24 taxonomy_MicroCarb.qzv
-rw-rw-r-- 1 jw21z307 microbarb  268195 May  5 11:21 unrooted-tree_MicroCarb.qza
(base) [jw21z307@submit02 MicroCarbComb]$ sbatch MicroCarbComb.sh
```

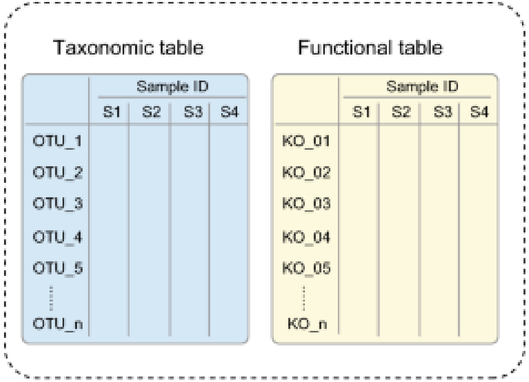
Sequences are filtered, demultiplexed grouped and assigned to known sequences in taxonomy databases

# Workflow:
# Feature table

| | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| Taxa 1 | 0 | 0 | 2 |
| Taxa 2 | 1 | 0 | 0 |
| Taxa 3 | 10 | 2 | 15 |
| Taxa 4 | 0 | 1 | 0 |

Feature tables are sometimes filtered (rare taxa, low abundance, rarefaction)

# Microbiota data: analytic challenges



Taxonomic table

|  | Sample ID |  |  |  |
|---|---|---|---|---|
|  | S1 | S2 | S3 | S4 |
| OTU_1 |  |  |  |  |
| OTU_2 |  |  |  |  |
| OTU_3 |  |  |  |  |
| OTU_4 |  |  |  |  |
| OTU_5 |  |  |  |  |
| ⋮ |  |  |  |  |
| OTU_n |  |  |  |  |

Functional table

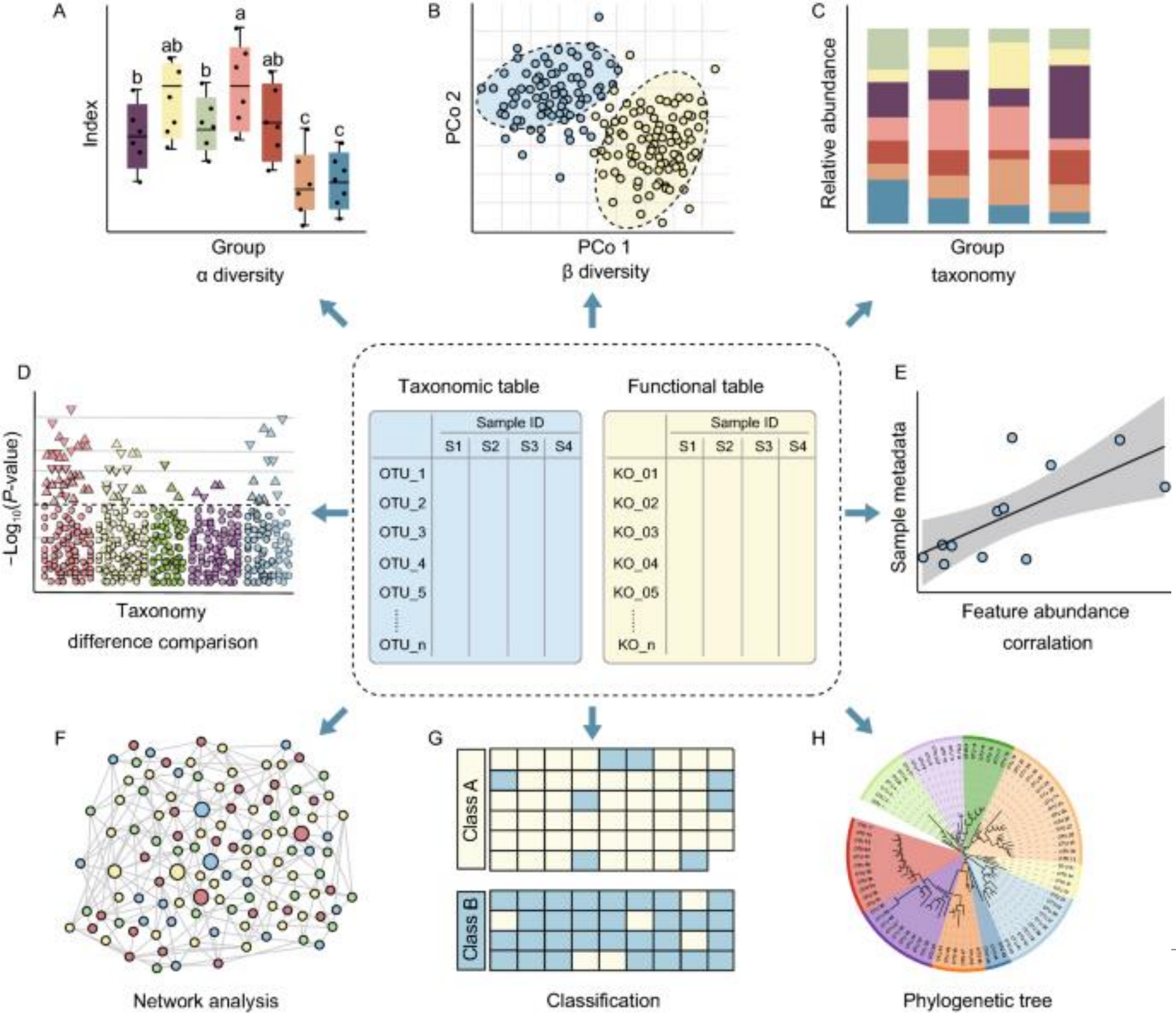
|  | Sample ID |  |  |  |
|---|---|---|---|---|
|  | S1 | S2 | S3 | S4 |
| KO_01 |  |  |  |  |
| KO_02 |  |  |  |  |
| KO_03 |  |  |  |  |
| KO_04 |  |  |  |  |
| KO_05 |  |  |  |  |
| ⋮ |  |  |  |  |
| KO_n |  |  |  |  |

Microbiota data is:

**multivariable** (= because every taxon is a variable)

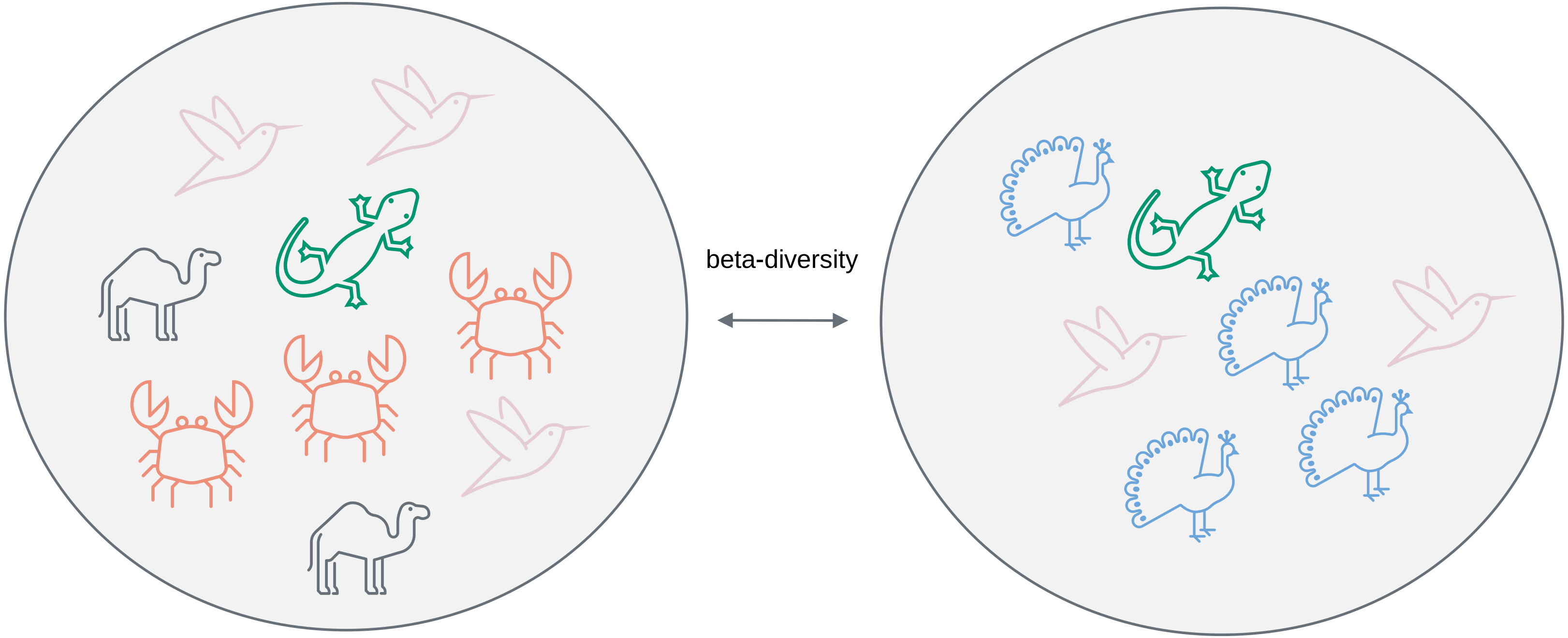
**sparse** (= have many zeros, because taxa are often only present in few samples)

**compositional** (= are not absolute measurements but are proportions)



Liu et al. Protein & Cell 2021

# Microbiota as an ecosystem



beta-diversity

alpha-diversity

alpha-diversity

# Diversity measures
# alpha diversity

Most simple: Species richness= total number of species at a site

Most popular alpha diversity measures:

- Simpson: between 0 and 1

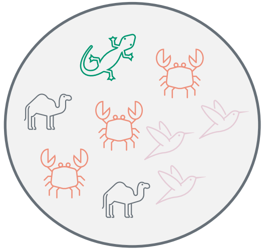
$$D = 1 - \sum_{i=1}^{S} \frac{n_i(n_i - 1)}{n(n-1)}$$

- Shannon: max= ln S

$$H' = -\sum_i p_i \cdot \ln p_i \quad \text{mit } p_i = \frac{n_i}{N}$$

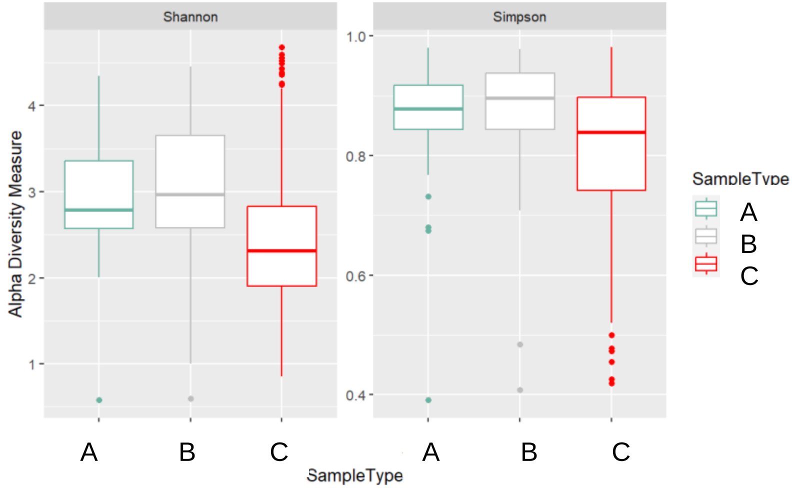
- Others: Chao1, ACE, ..

Simpson diversity example calculation:
1*0/9*8 = 0
2*1/9*8 = 0.027
3*2/9*8 = 0.083
3*2/9*8 = 0.083
1-(2*0.083+0.027) = 0.8



# Alpha diversity Visualisation

Often visualised as boxplots per groups



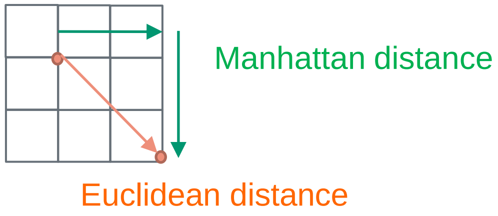
Alpha diversity differences statistically evaluated like any other measurement:

- Wilcox-test, ANOVA

# Diversity measures
# beta diversity

Popular beta-diversity measures:

- Euclidean distance
- Aitchinson distance: central log transform + euclidean distance
- Bray-curtis dissimilarity
- UniFrac + weighted UniFrac: includes phylogenetic information
- Jaccard, Chao, Mahalobani, etc.



$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

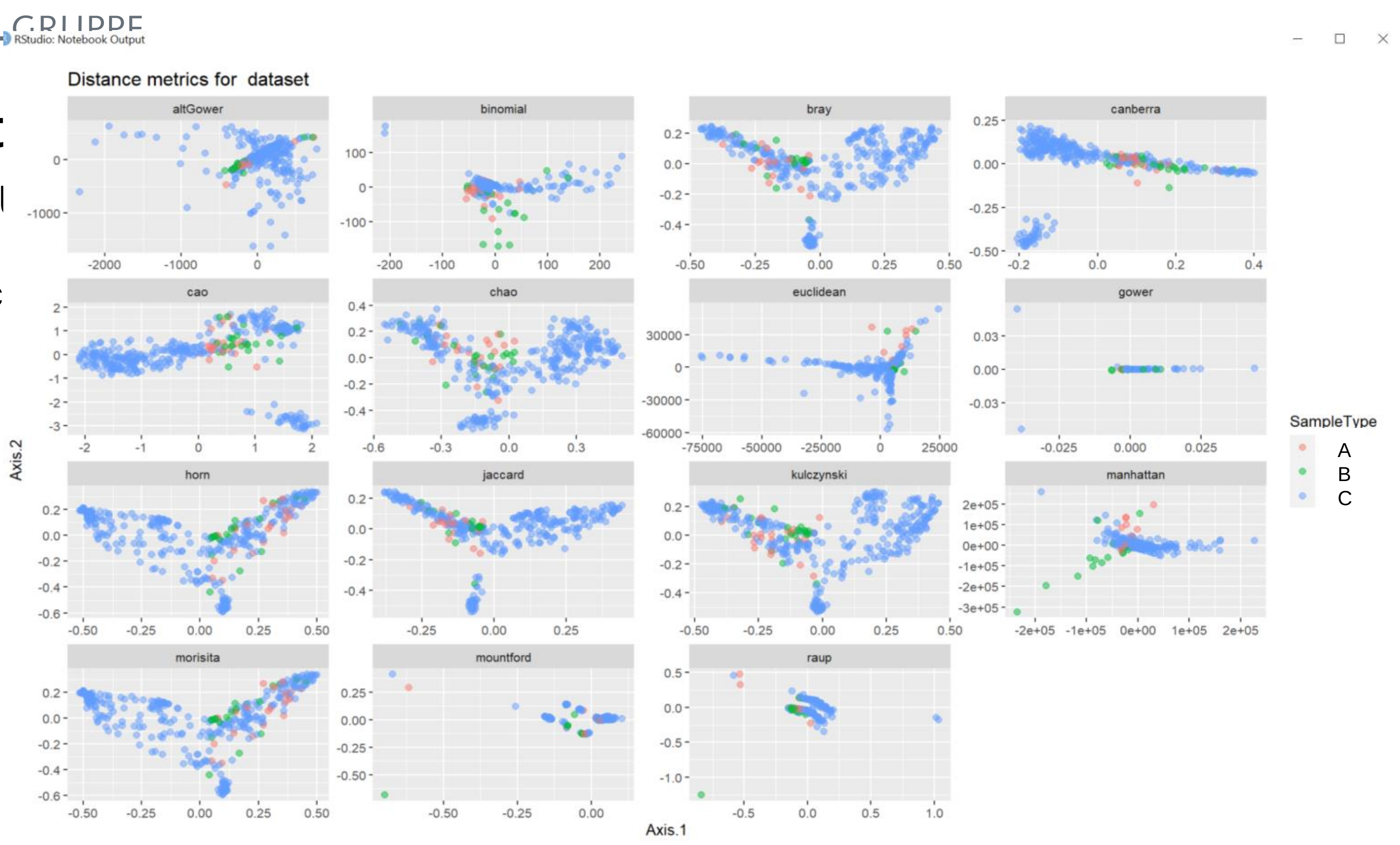
C= common species
S= total species at a site

$$\left( \frac{sum\ of\ unshared\ branch\ lengths}{sum\ of\ all\ tree\ branch\ lengths} \right) = fraction\ of\ total\ unshared\ branch\ lengths$$

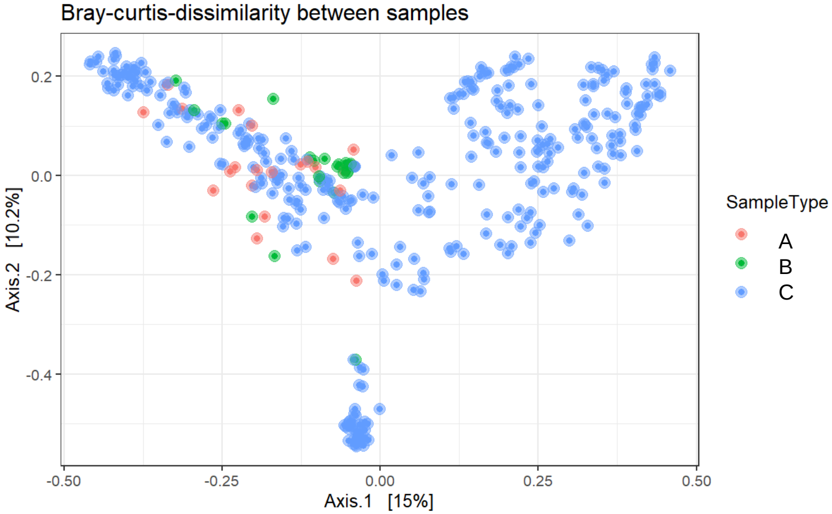
# Bet

# Vis

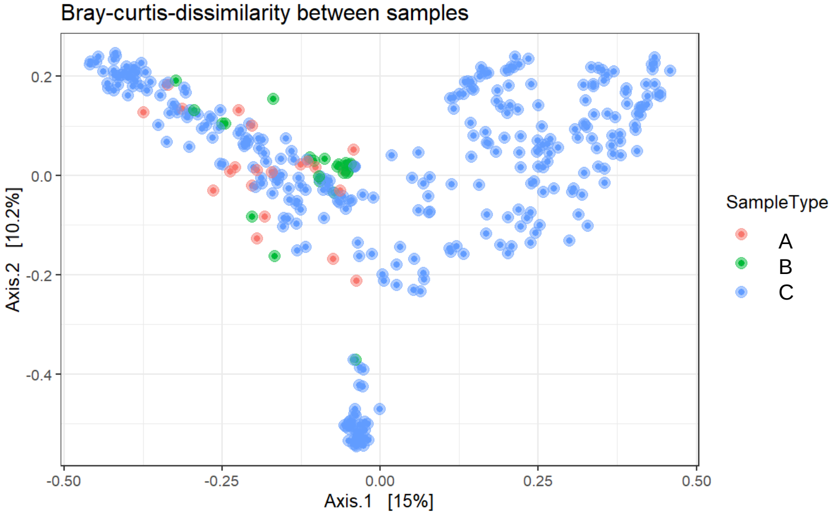
Choic

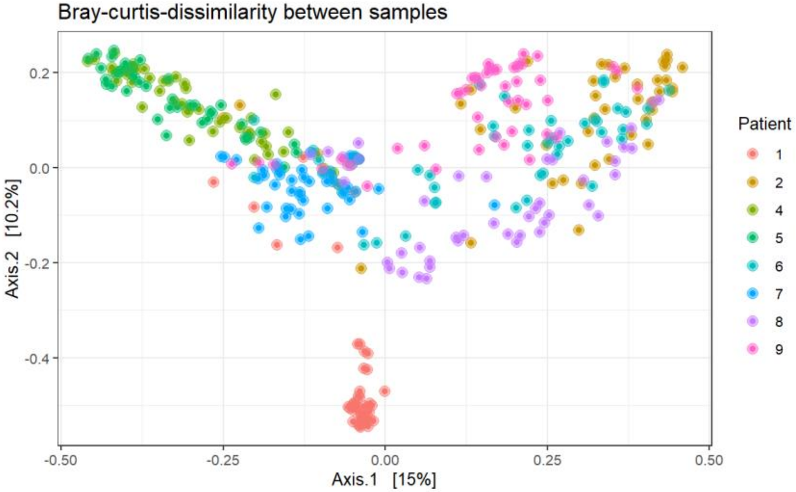


# Beta diversity
# Pitfall: Batch effects



# Beta diversity
# Pitfall: Batch effects



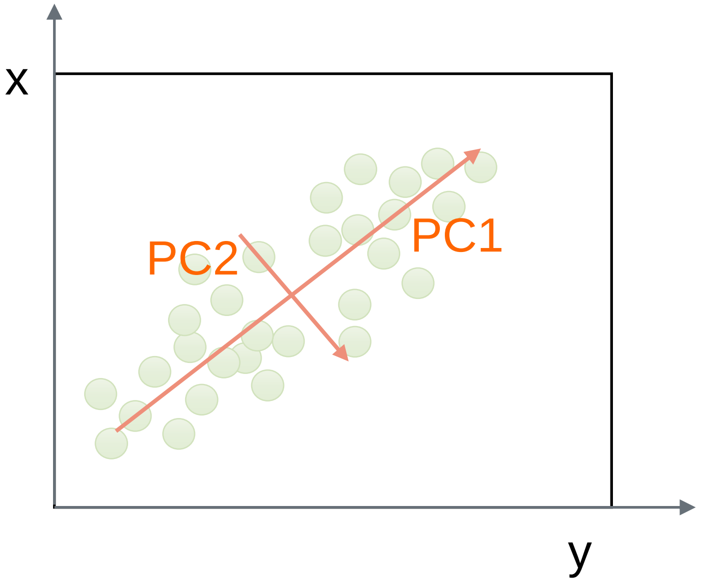


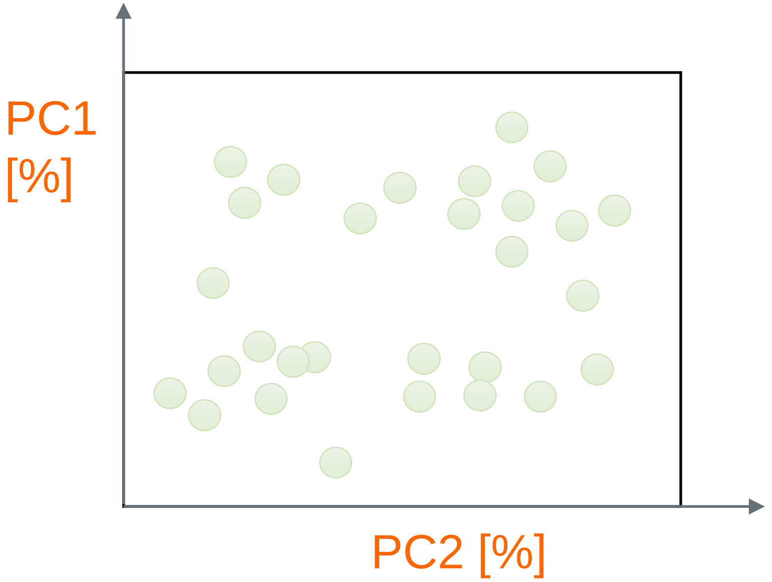
# Beta diversity
# Ordination

Multivariate data can be plotted in a 2D way using dimensionality reduction techniques

- PCA:  for any multivariate data set, characterizing the main axes of variance in the data
- PCoA: main axes of variance for distance matrices

# Principal components
# Idea





# Beta diversity
# Ordination

Multivariate data can be plotted in a 2D way using dimensionality reduction techniques

- PCA: for any multivariate data set, characterizing the main axes of variance in the data
- PCoA: main axes of variance for distance matrices

- NMDS (Non-Metric Multidimensional Scaling): iterative process, axes can not be interpreted

- Other: correspondance analysis (CA), redundancy analysis (RDA): uses metadata variables for scaling

# Beta diversity
# Visualisation

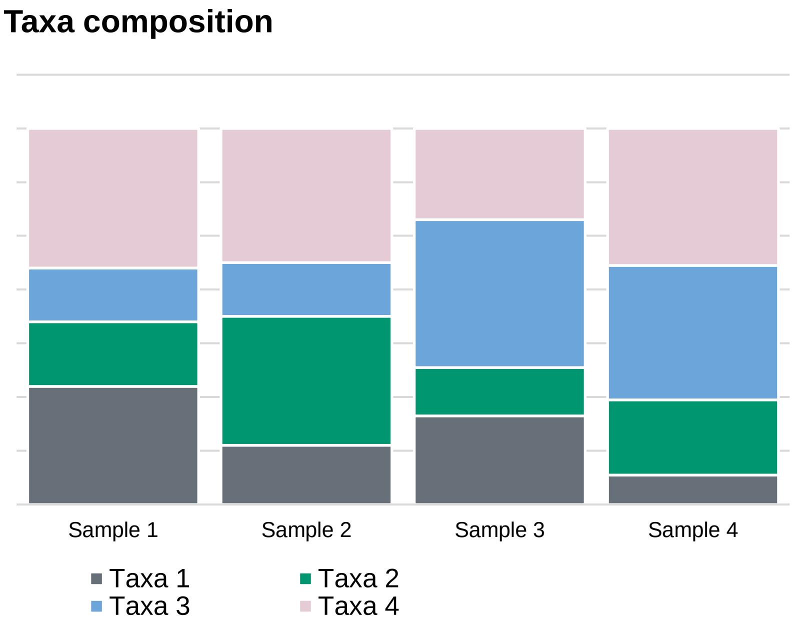
Choice of distance method is variable, ultimate goal is to see separation of groups

# Beta diversity
# Visualisation

Significant differences between groups need to tested by methods assessing multivariable data

- ANOSIM, ADONIS, PERMANOVA, MRPP

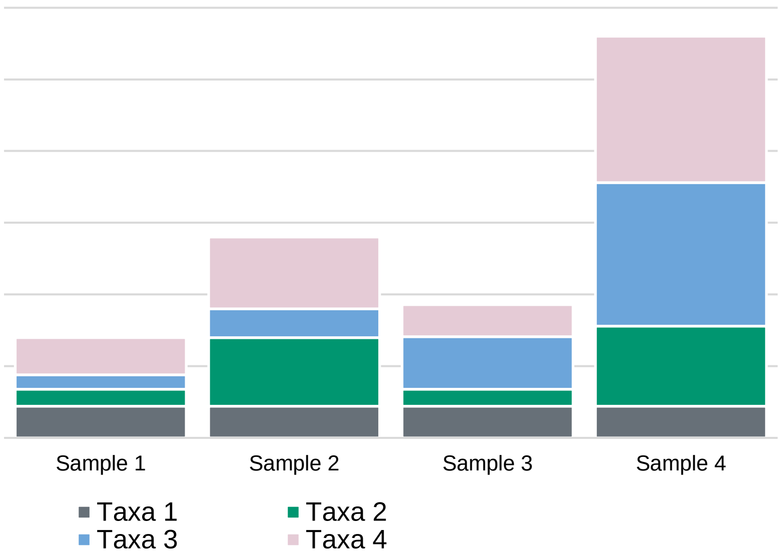
# Taxonomy



Taxonomy results are not absolute, taxa composition are proportions

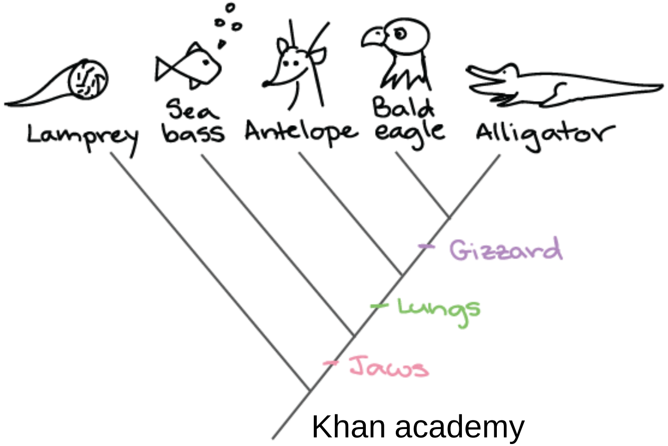
# Taxonomy

**Taxa absolute numbers**



This can lead to distortions if the overall bacterial count fluctuates between samples

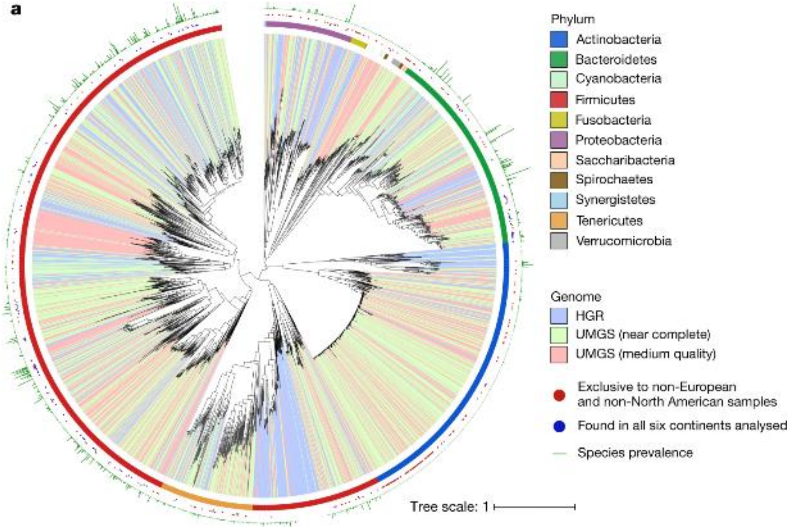
# Taxonomy
# Phylogenetic trees



Khan academy

Phylogenetic trees can be built with reference databases or with information from the sequencing reads

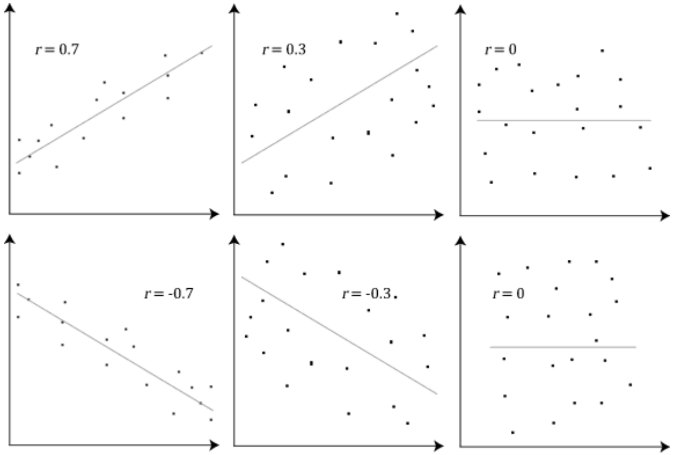
# Taxonomy
# Phylogenetic trees



Phylogenetic trees can be built with reference databases or with information from the sequencing reads

Almeida et al. Nature, 2019

# Analysis:
# Correlation analysis



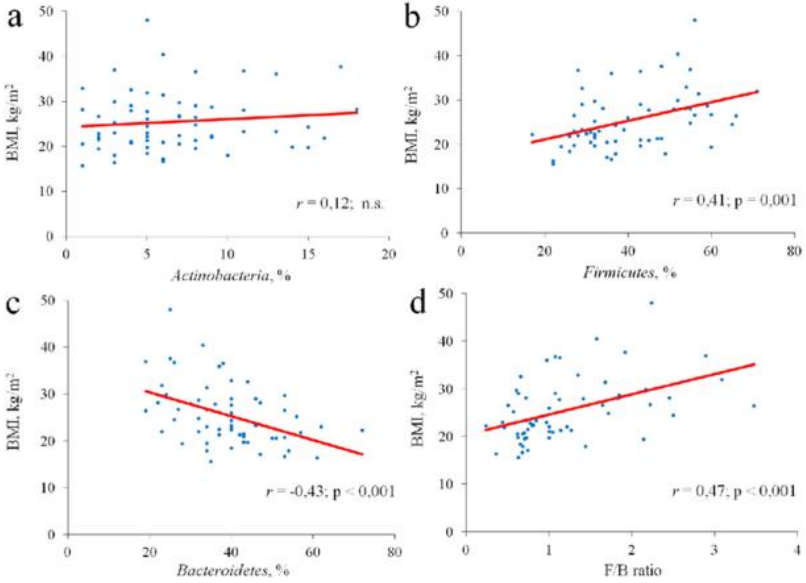
Wikipedia: Correlation

Data can be assessed for correlation structure

- Between microbiota
- Between microbiota and metadata

Because every taxa is tested seperately results have to be adjusted by multiple testing correction

# Correlation analysis: Visualisation

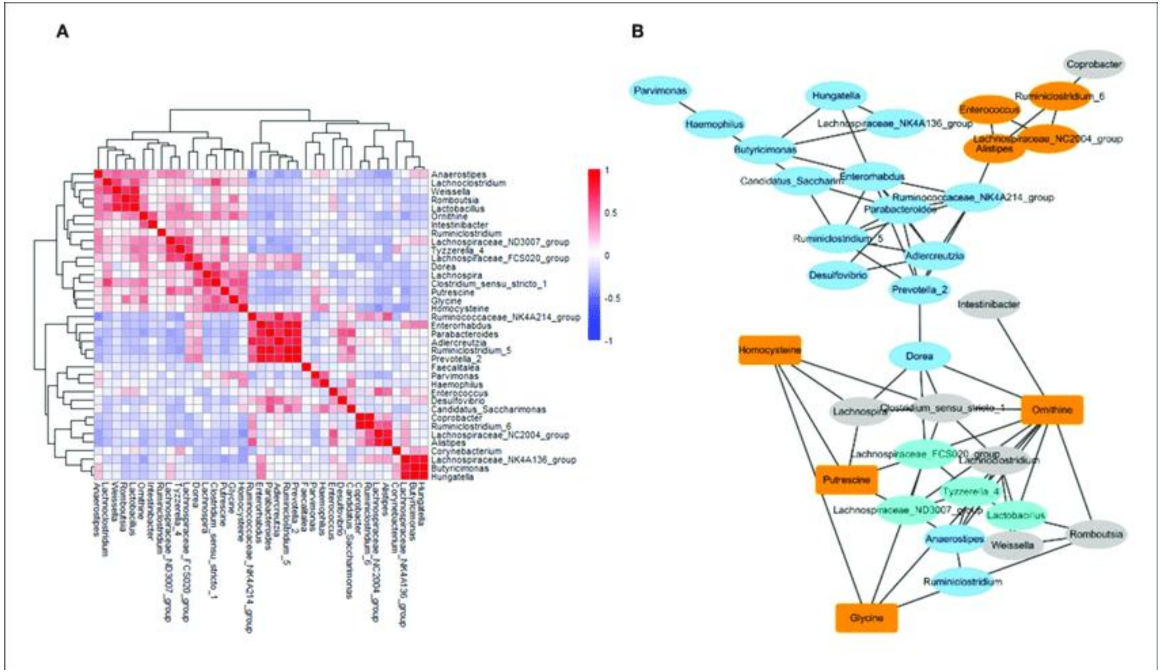


Koliada et al. BMC Microbiology 2017

Correlations with one predictor can be depticted as a regression

Multiple correlations can be depicted as a network analysis or in a correlation plot

# Correlation analysis: Visualisation



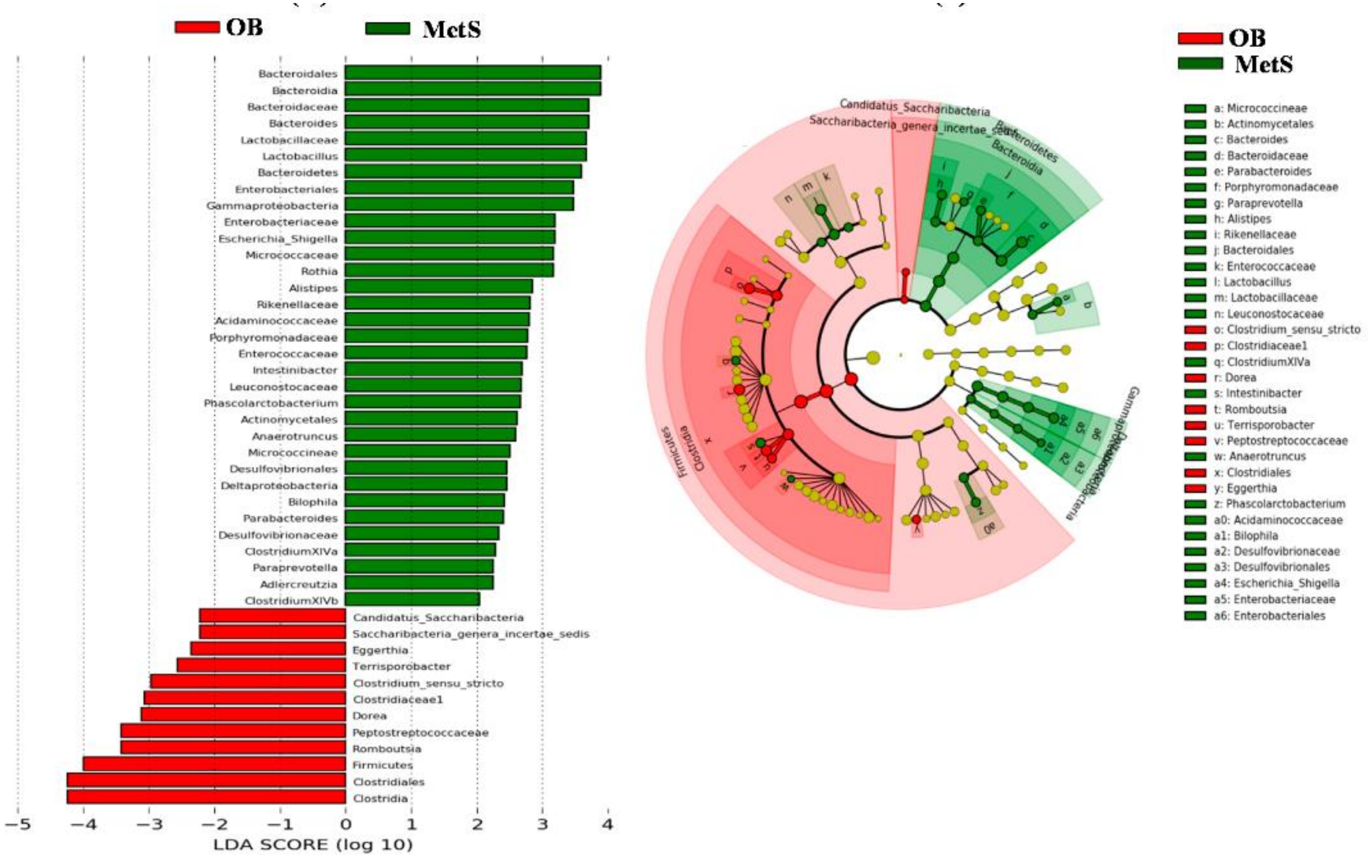
Zhu et al. Frontiers in Cellular and Infection Microbiology, 2019

Correlations with one predictor can be depticted as a regression

Multiple correlations can be depicted as a network analysis or in a heatmap

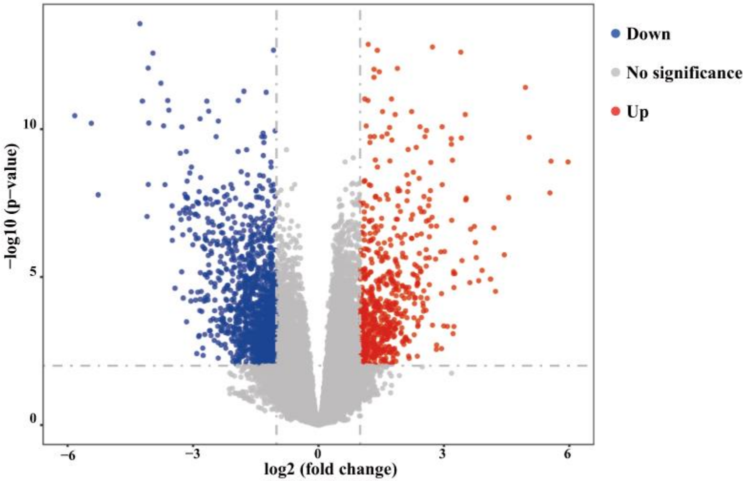
# Analysis
# Differential abundance

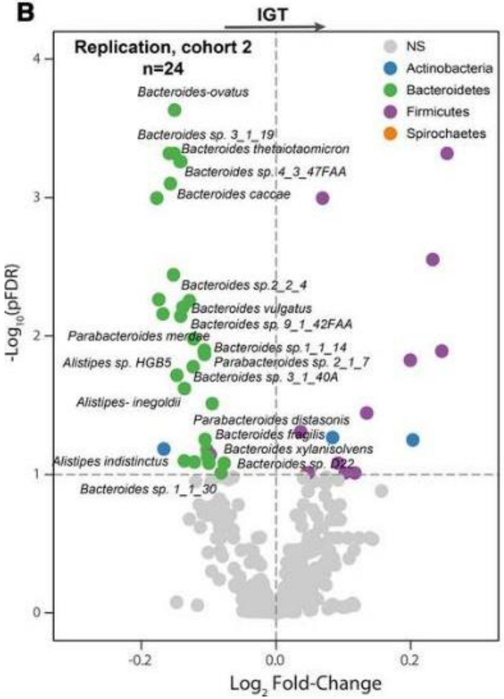
Differental abundance tells us which taxa differ significantly between groups


OB
MetS
Bacteroidales
Bacteroidia
Bacteroidaceae
Bacteroides
Lactobacillaceae
Lactobacillus
Bacteroidetes
Enterobacteriales
Gammaproteobacteria
Enterobacteriaceae
Escherichia_Shigella
Micrococcaceae
Rothia
Alistipes
Rikenellaceae
Acidaminococcaceae
Porphyromonadaceae
Enterococcaceae
Intestinibacter
Leuconostocaceae
Phascolarctobacterium
Actinomycetales
Anaerotruncus
Micrococcineae
Desulfovibrionales
Deltaproteobacteria
Bilophila
Parabacteroides
Desulfovibrionaceae
ClostridiumXIVa
Paraprevotella
Adlercreutzia
ClostridiumXIVb
Candidatus_Saccharibacteria
Saccharibacteria_genera_incertae_sedis
Eggerthia
Terrisporobacter
Clostridium_sensu_stricto
Clostridiaceae1
Dorea
Peptostreptococcaceae
Romboutsia
Firmicutes
Clostridiales
Clostridia
−5
−4
−3
−2
−1
0
1
2
3
4
LDA SCORE (log 10)
OB
MetS
Candidatus_Saccharibacteria
Saccharibacteria_genera_incertae_sedis
Bacteroidetes
Bacteroidia
Firmicutes
Clostridia
Gammaproteobacteria
Deltaproteobacteria
a: Micrococcineae
b: Actinomycetales
c: Bacteroides
d: Bacteroidaceae
e: Parabacteroides
f: Porphyromonadaceae
g: Paraprevotella
h: Alistipes
i: Rikenellaceae
j: Bacteroidales
k: Enterococcaceae
l: Lactobacillus
m: Lactobacillaceae
n: Leuconostocaceae
o: Clostridium_sensu_stricto
p: Clostridiaceae1
q: ClostridiumXIVa
r: Dorea
s: Intestinibacter
t: Romboutsia
u: Terrisporobacter
v: Peptostreptococcaceae
w: Anaerotruncus
x: Clostridiales
y: Eggerthia
z: Phascolarctobacterium
a0: Acidaminococcaceae
a1: Bilophila
a2: Desulfovibrionaceae
a3: Desulfovibrionales
a4: Escherichia_Shigella
a5: Enterobacteriaceae
a6: Enterobacteriales

# Differential abundance Visualisation

## Volcano plot

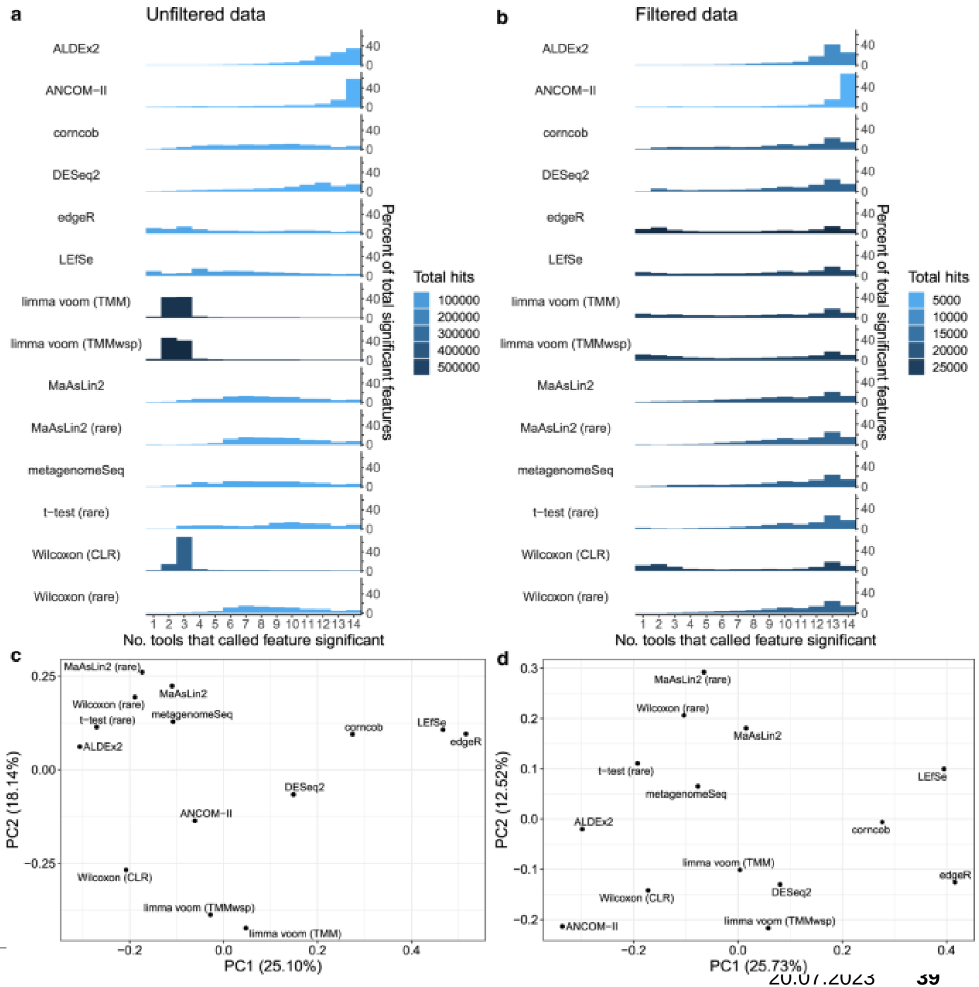




Arnoriaga-Rodríguez et al., Gut, 2021

# Analysis
# Differential abundance

Methods are not standardised
and can output different results



Nearing et al. Nature communications 2022

# Classification: Model fitting

If we have some hypothesis about the underlying structure of the data we can try to fit a model

- e.g. classification into two disease groups
- Model parameters can give information about which features (=e.g. taxa) are important

# Classification:
# Model fitting

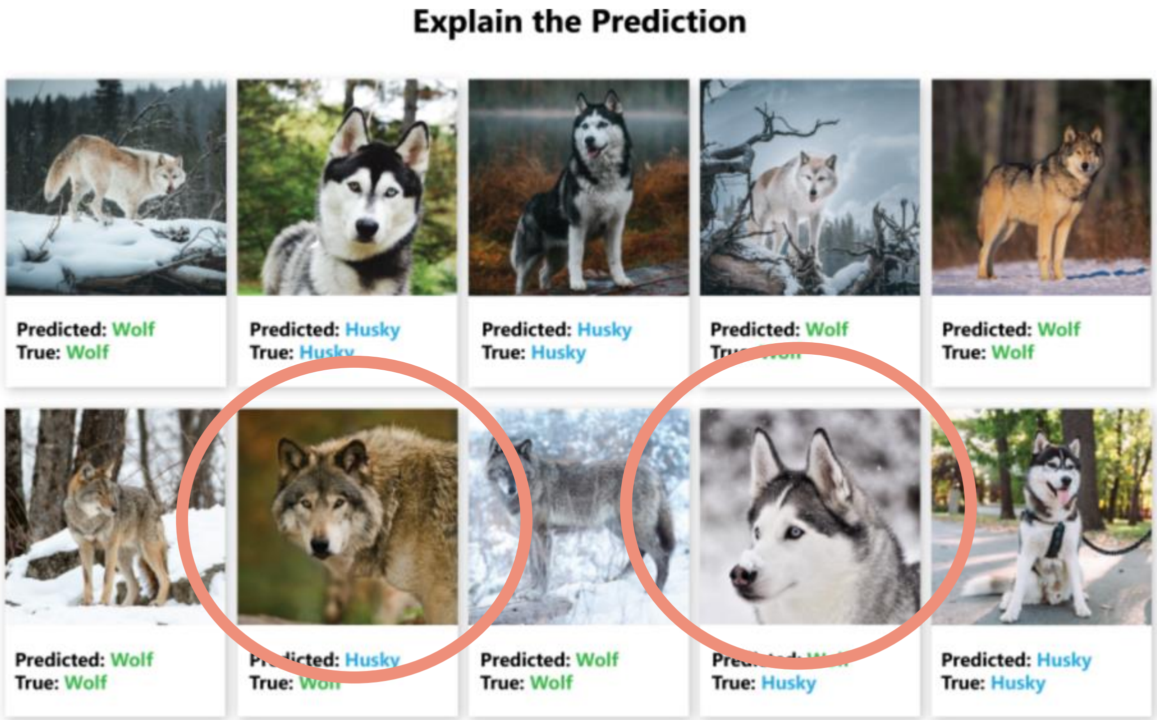
If we have some hypothesis about the underlying structure of the data we can try to fit a model

- e.g. classification into two disease groups
- Model parameters can give information about which features (=e.g. taxa) are important

Examples:

- Multiple linear regression
- Random forest
- Partial Least-Squares Discriminant Analysis (PLS-DA)
- Neural networks

# Model fitting: pitfalls



https://carpentries-incubator.github.io/data-science-ai-senior-researchers/05-Problems-with-AI/index.html

# Overfitting

Overfitting: If a model uses a lot of features it can be fitted very well to a training data set

- Extreme: one variable classifies one samples -> perfect fit
- Problem: **This model will not be informative for a new unrelated data**

Solution: Validation with new data

- Cross-validation
- Splitting of data before analysis
- Second evaluation cohort

Better

# Model evaluation: AUC-ROC-curves

Good AUC results

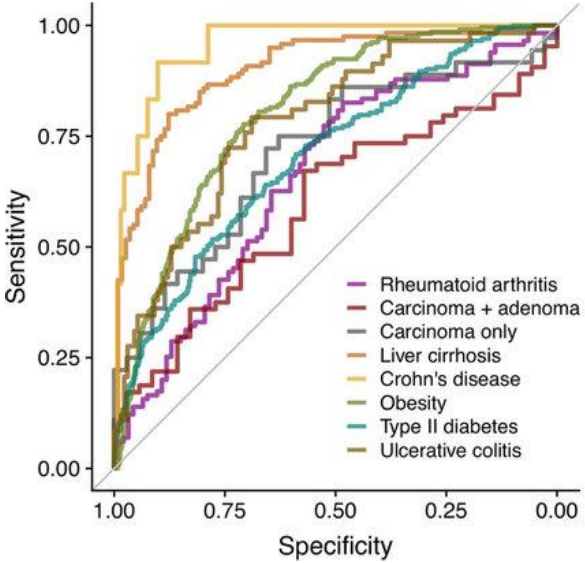
- Over 0.7: acceptable
- Over 0.8: excellent
- Over 0.9: outstanding

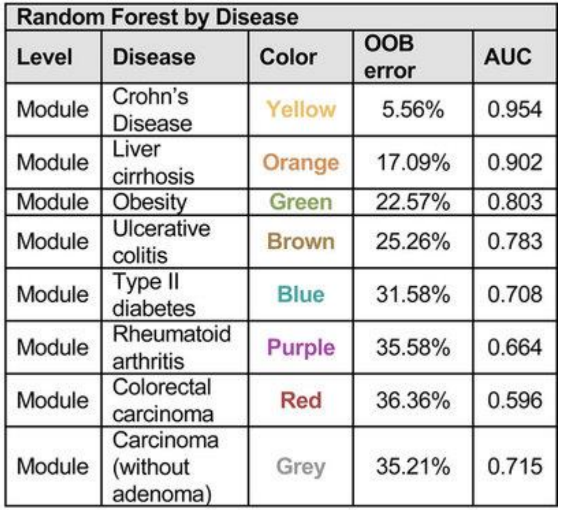
True positive rate (Precision) = TP/(TP+FN)

False positive rate= FP/(FP+TN)

Sensitivity (Recall): 1- TPR

Specificity: 1- FPR





| Random Forest by Disease | | | | |
|---|---|---|---|---|
| **Level** | **Disease** | **Color** | **OOB error** | **AUC** |
| Module | Crohn's Disease | Yellow | 5.56% | 0.954 |
| Module | Liver cirrhosis | Orange | 17.09% | 0.902 |
| Module | Obesity | Green | 22.57% | 0.803 |
| Module | Ulcerative colitis | Brown | 25.26% | 0.783 |
| Module | Type II diabetes | Blue | 31.58% | 0.708 |
| Module | Rheumatoid arthritis | Purple | 35.58% | 0.664 |
| Module | Colorectal carcinoma | Red | 36.36% | 0.596 |
| Module | Carcinoma (without adenoma) | Grey | 35.21% | 0.715 |

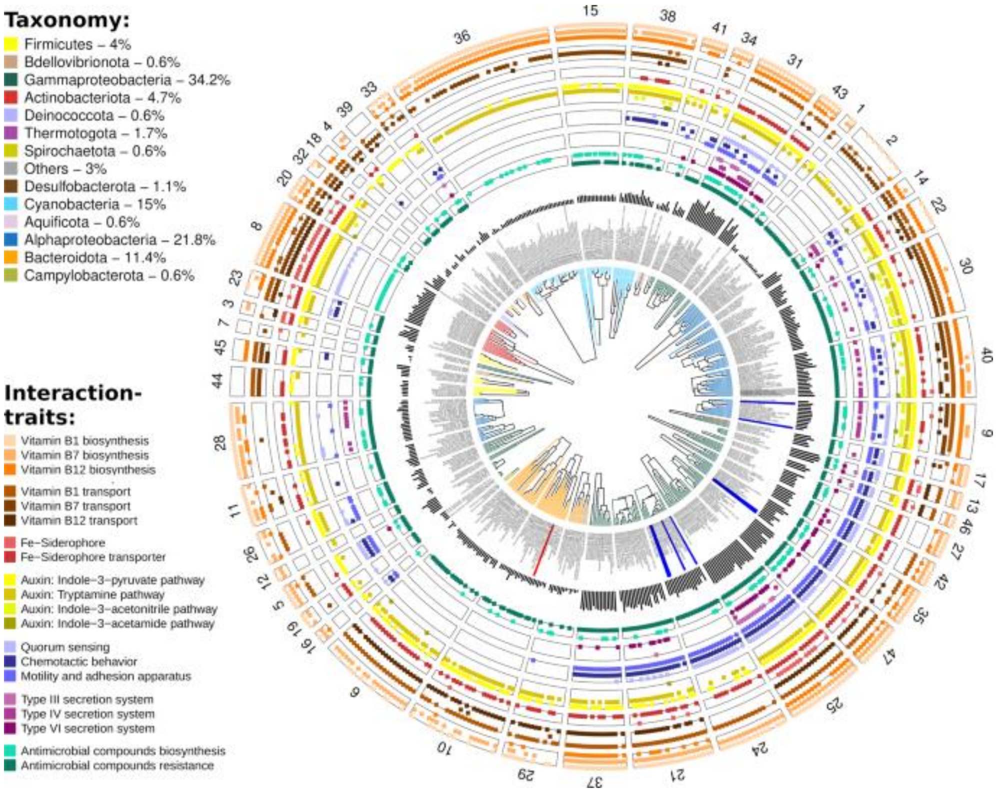
Armour et. al, mSystems, 2019

# Metagenomic analysis: Functional potential

Metagenomic sequencing analyses the whole genome of  bacteria

This includes bacterial enzymes:

- Analysis of functional potential



Zoccarato et. al, communications biology, 2022

# Metagenomic analysis: Functional potential

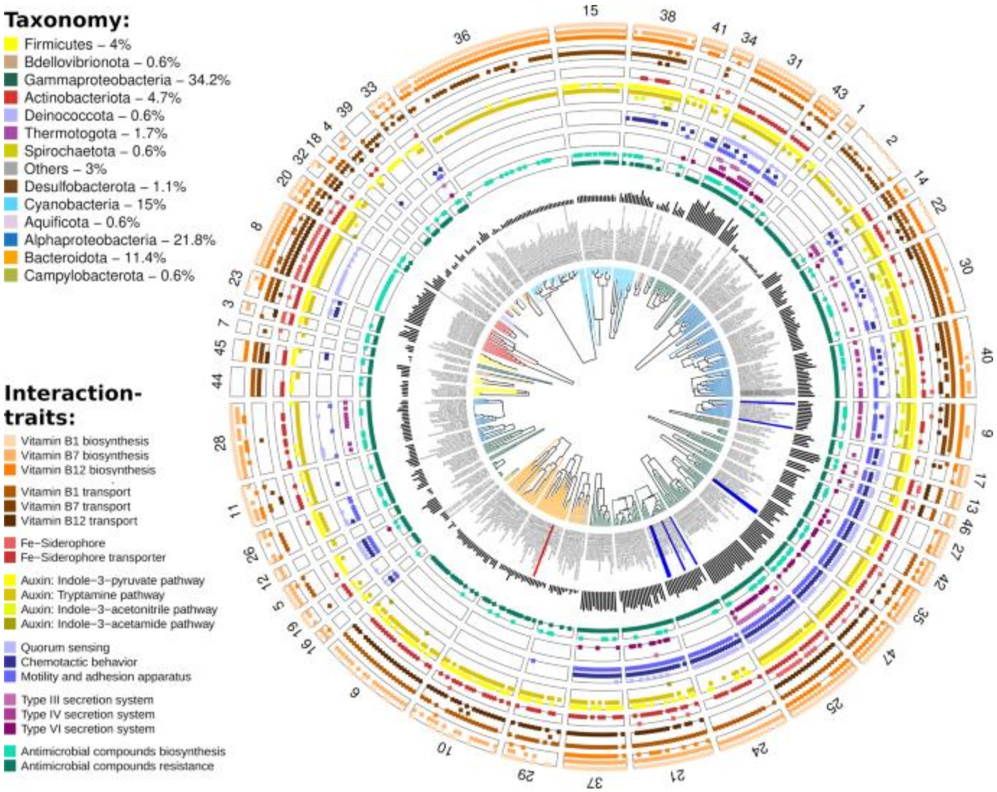
Metagenomic sequencing analyses the whole genome of  bacteria

This includes bacterial enzymes:

- Analysis of functional potential

Because of higher resolution also:

- Species level taxonomic profiling



Zoccarato et. al, communications biology, 2022

# Type of analysis depends on the research question

# What questions can we answer?

- We can identify if specific taxa are present
- We can compare if a sample has reduced alpha diversity (e.g. after antibiotic therapy)
- We can compare samples with others (e.g. is a sample more similar to other from heathly people or from a disease group)

- We can establish taxa signatures that are diagnostic or predicitve for clinical variables
- Metagenomic sequencing can give answers about metabolic functions of taxa: pathways, metabolites, bacterial signalling molecules, ..

INSELGRUPPE

# Questions?