Microbiome data analysis - Lecture

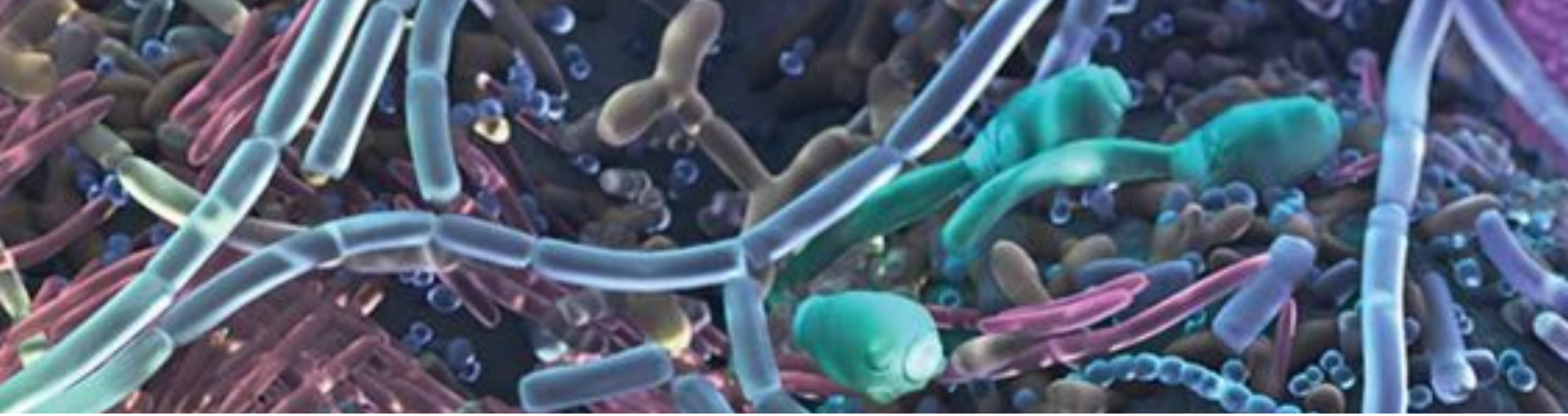
Anouk Zancarini

To cite this version:

Anouk Zancarini. Microbiome data analysis - Lecture. Master. Tools in molecular data analysis - module Microbiome, Netherlands. 2021. hal-04286457

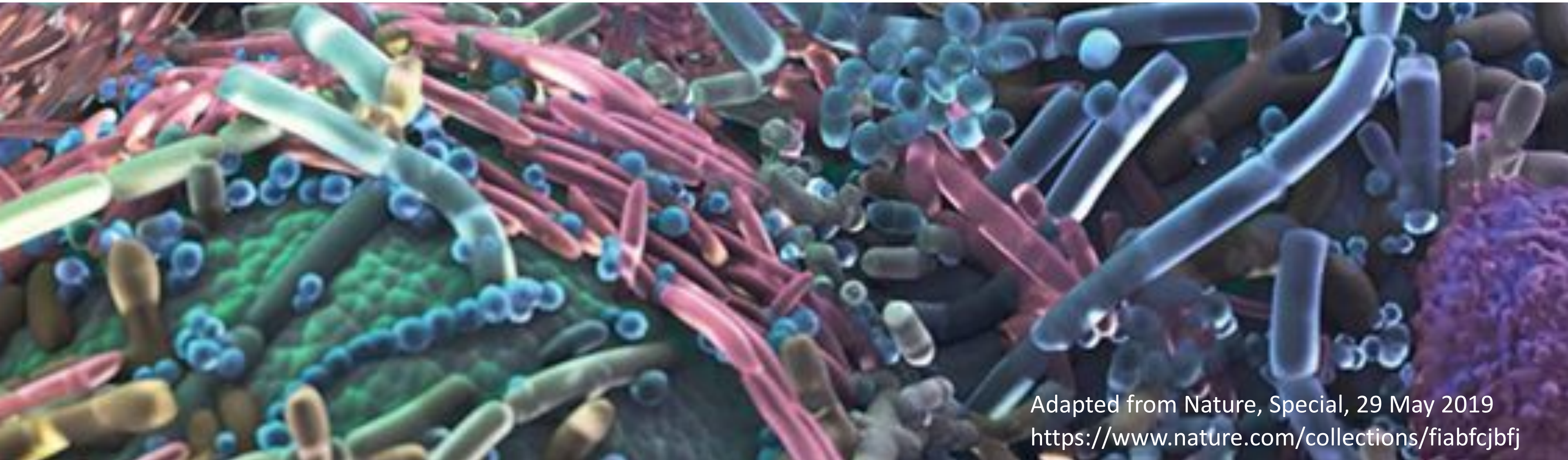
HAL Id: hal-04286457
https://hal.inrae.fr/hal-04286457v1

Submitted on 15 Nov 2023



Microbiome data analysis

Anouk ZANCARINI



Adapted from Nature, Special, 29 May 2019
https://www.nature.com/collections/fiabfcjbfj

Content

What is microbiome?

Part 1

- Definitions
- Microbiome importance
- Scientific questions
- Differences between metagenomics and metabarcoding

How microbiota data are generated?

Part 2

- From samples to sequences
- From sequences to data sets

How microbiota data are analysed?

Part 3

- Alpha-diversity
- Data properties
- Data filtering
- Data normalisation
- Beta-diversity
- Microbial composition

Learning objectives

- ❑ Define microbiome and state microbiome importance
- ❑ Identify differences between metabarcoding and metagenomics
- ❑ Explain how microbiota data are generated (including bias)
- ❑ Explain and preform data pre-processing
- ❑ Explain how microbiota data are analysed
- ❑ Define, perform and interpret alpha-diversity
- ❑ Address sparsity, under-sampling and uneven sampling depth using data filtering and normalization
- ❑ Define, perform and interpret beta-diversity
- ❑ Generate and interpret multivariate data analyses
- ❑ Perform and interpret appropriate statistical tests
- ❑ Visualize and interpret microbial community composition

What is microbiome?

What is microbiome?

Part 1

- Definitions
- Microbiome importance
- Scientific questions
- Differences between metagenomics and metabarcoding

How microbiota data are generated?

Part 2

- From samples to sequences
- From sequences to data sets

How microbiota data are analysed?

Part 3

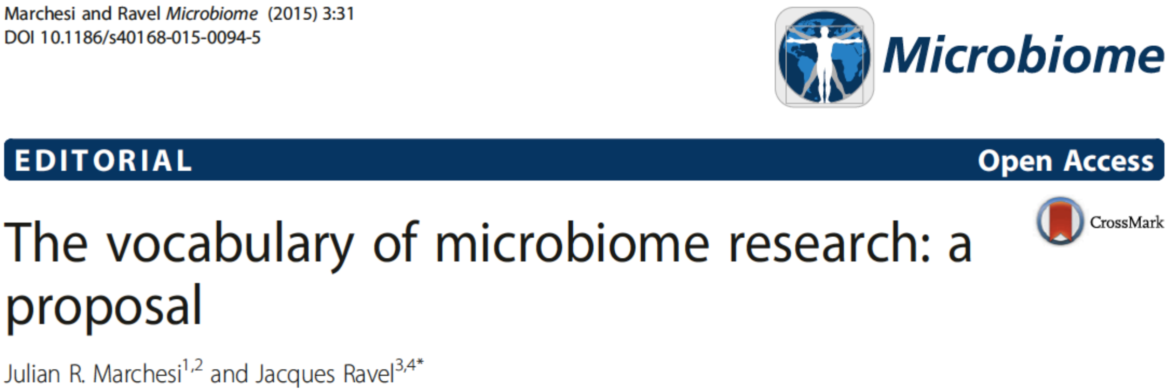
- Alpha-diversity
- Data properties
- Data filtering
- Data normalisation
- Beta-diversity
- Microbial composition

Definitions

Microbiota is the **assemblage of microorganisms** present in a defined environment. Microbiota includes archaea, bacteria, fungi, protists and viruses.

Metagenome is the **collection of genomes** and genes from the members of a microbiota.

Microbiome refers to the **entire habitat**, including the microorganisms (bacteria, archaea, lower and higher eurkaryotes, and viruses), their genomes (*i.e.*, genes), and the surrounding environmental conditions.



Marchesi and Ravel *Microbiome* (2015) 3:31
DOI 10.1186/s40168-015-0094-5

Microbiome

EDITORIAL

Open Access

CrossMark

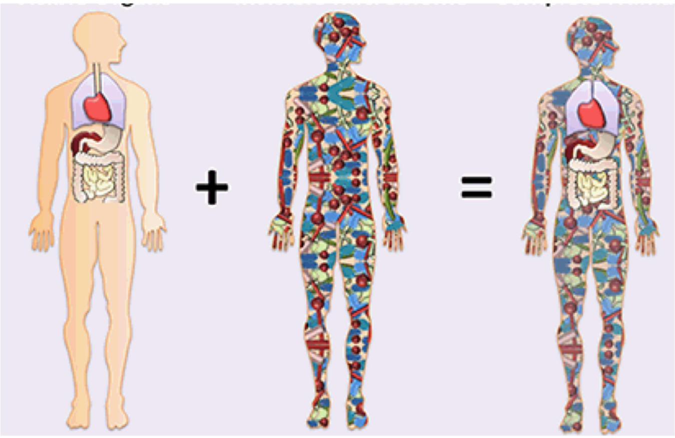
The vocabulary of microbiome research: a proposal

Julian R. Marchesi[1,2] and Jacques Ravel[3,4*]

Microbiome importance

Human microbiome: our second genome

- ~10 times more cells than you
- ~100 times more genes than you
- ~1000s different species



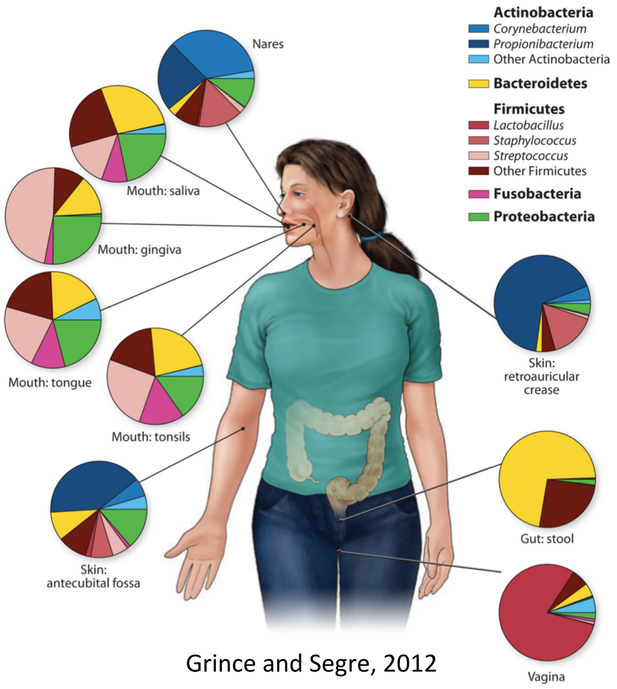
Adapted from Appanna V.D. (2018) The Human Microbiome:
The Origin. In: Human Microbes - The Power Within. Springer, Singapore

The Human Microbiome Project

- Characterize human microbiome
- Analyse its role in human health and disease

Human microbiome links to health

- Influence metabolism
- Modulate drug interaction
- Link to irritable bowel syndrome, cancer, mental health, obesity, diabetes, asthma, etc.



Grince and Segre, 2012

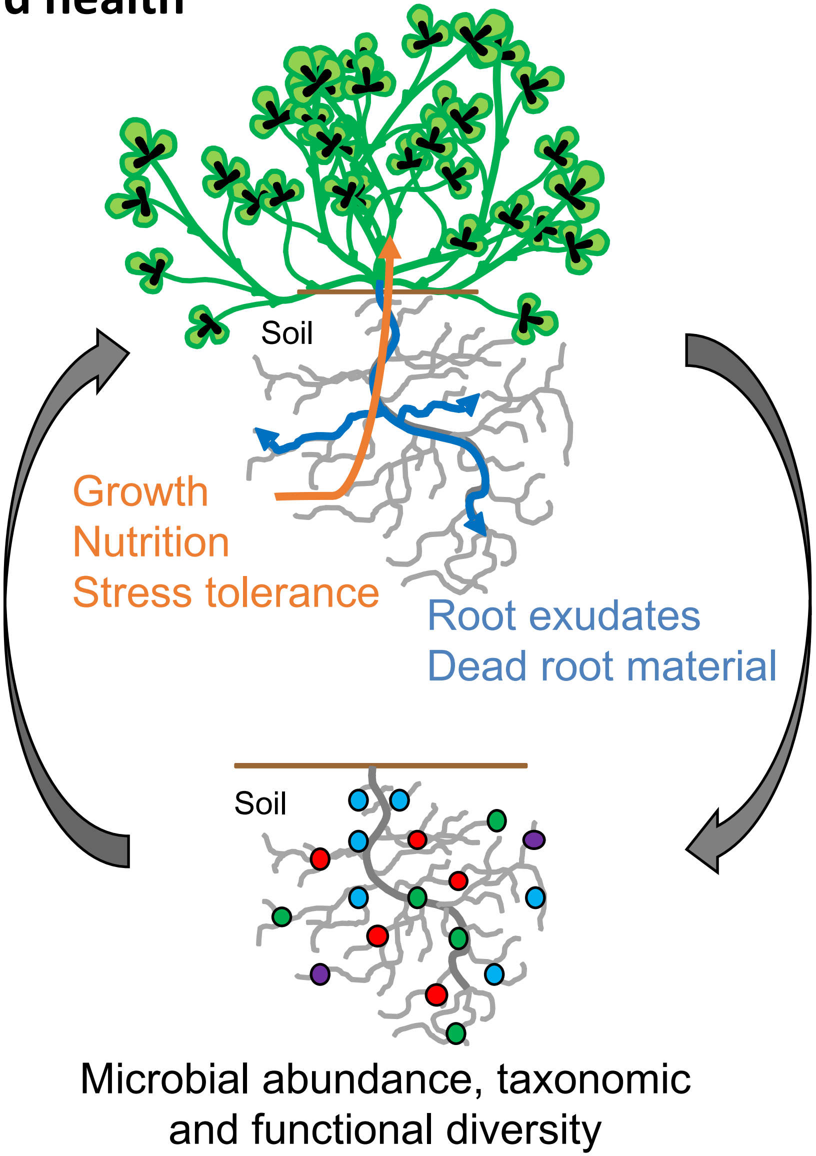
Microbiome importance

Plant microbiome can improve plant growth and health

- Biofertilisation
- Phytostimulation
- Rhizoremediation
- Improve stress tolerance

Plant drives its microbiome

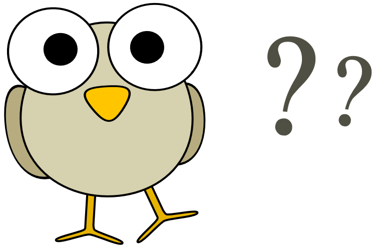
- Root exudates
  (nutrients and signalling molecules)



What is microbiome and its importance?

Test your knowledge…

- Please answer the 3 questions in the following quiz https://bigdata_microbiome.presenterswall.nl/



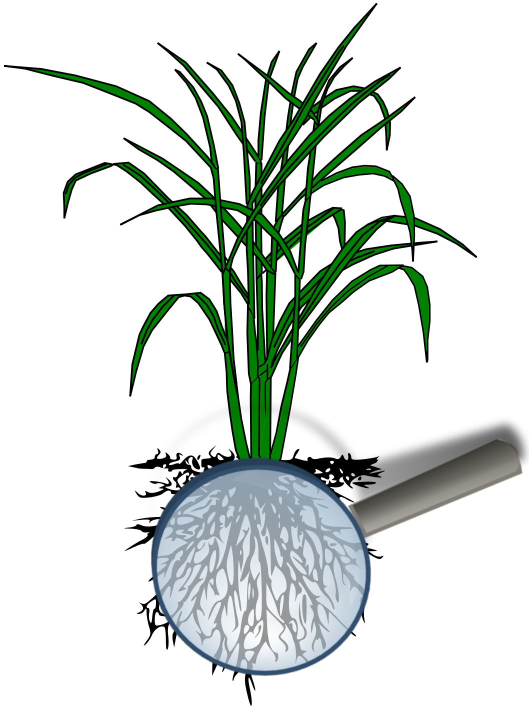
What is microbiome and its importance?

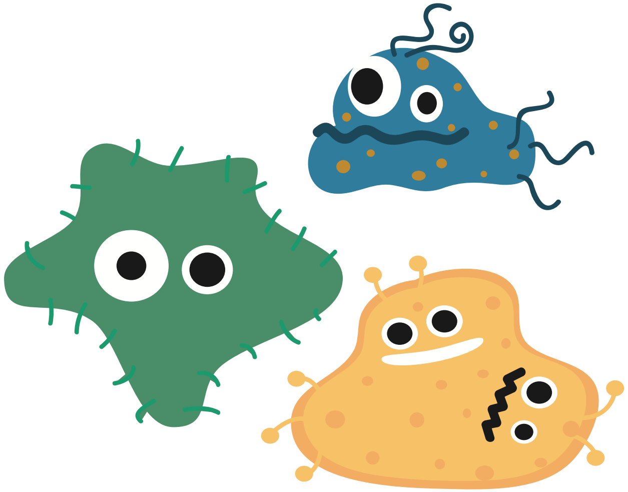
- Microbiota = assemblage of microorganisms
- Metagenome = collection of genomes
- Microbiome refers to the entire habitat
- Microbiome is important in:
  - ecosystem functioning
  - plant growth and health

Learning objectives

- ☑ Define microbiome and state microbiome importance
- □ Identify differences between metabarcoding and metagenomics
- □ Explain how microbiota data are generated (including bias)
- □ Explain and preform data pre-processing
- □ Explain how microbiota data are analysed
- □ Define, perform and interpret alpha-diversity
- □ Address sparsity, under-sampling and uneven sampling depth using data filtering and normalization
- □ Define, perform and interpret beta-diversity
- □ Generate and interpret multivariate data analyses
- □ Perform and interpret appropriate statistical tests
- □ Visualize and interpret microbial community composition

Main biological questions and methods

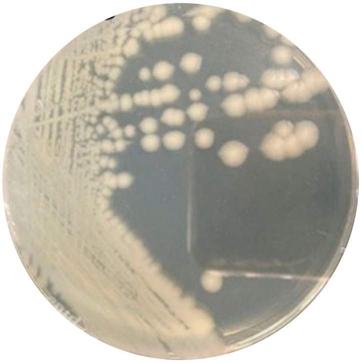




Main biological questions

- ➢ **Who is there?**
- ➢ **What are they doing?**

Challenge:
Most of the microbes are not cultivable

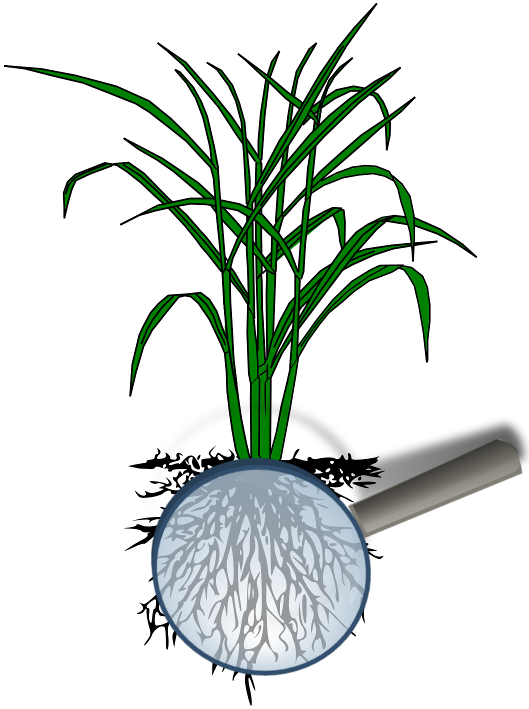


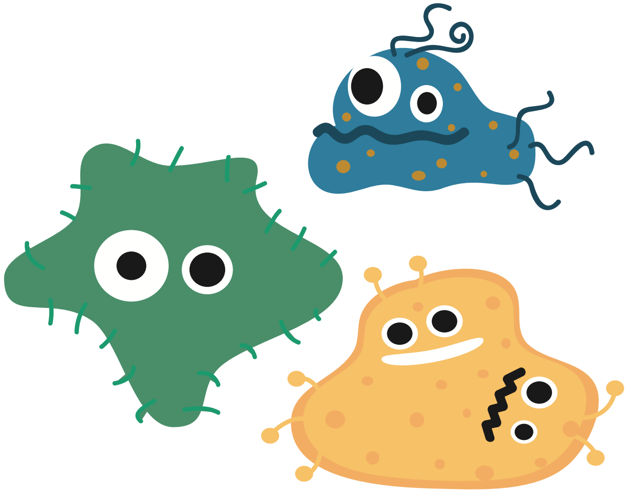
From culturing area

```
CGGATTTATTGGGTTTAAAGGG
CGGATTTATTGGGTTTAAAGGG
CGGATTTATTGGGTTTAAAGGG
CGGATTTACTGGGTGTAAAGGG
CGGATTTACTGGGTGTAAAGGG
CGGATTTATTGGGTTTAAAGGG
CGGATTTATTGGGTTTAAAGGG
```

to sequencing area

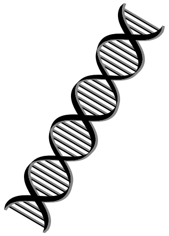
Main biological questions and methods

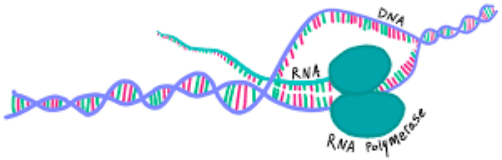


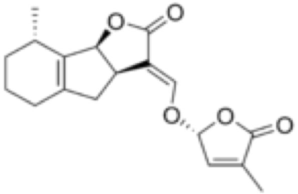


Main biological questions

- ➢ **Who is there?**
- ➢ **What are they doing?**







Metabarcoding
Metagenomics

Metatranscriptomics

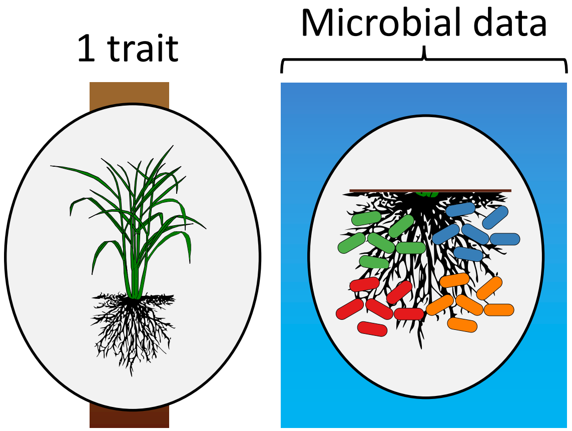
Metaproteomics
Metabolomics

Main biological questions and methods

<u>**Main biological questions**</u>

- **Who is there?**
- **What are they doing?**
- **Which microbe is associated with a specific phenotype?** (*i.e.* feature selection)

Statistical approaches & machine learning

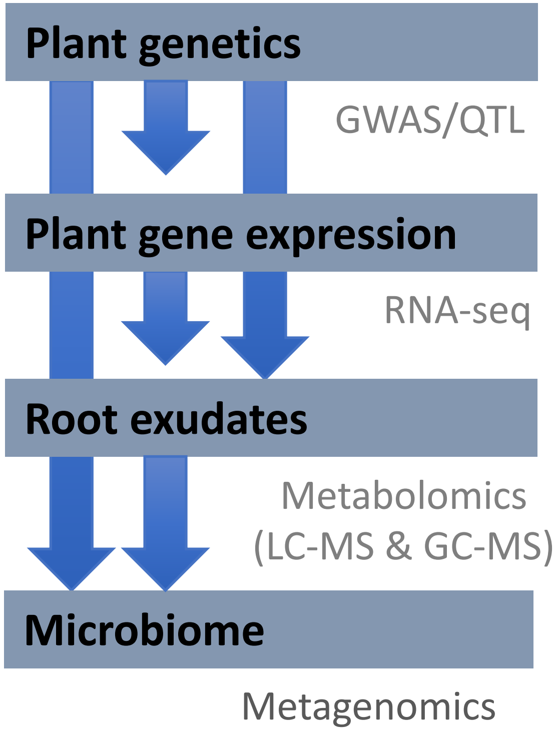


Main biological questions and methods

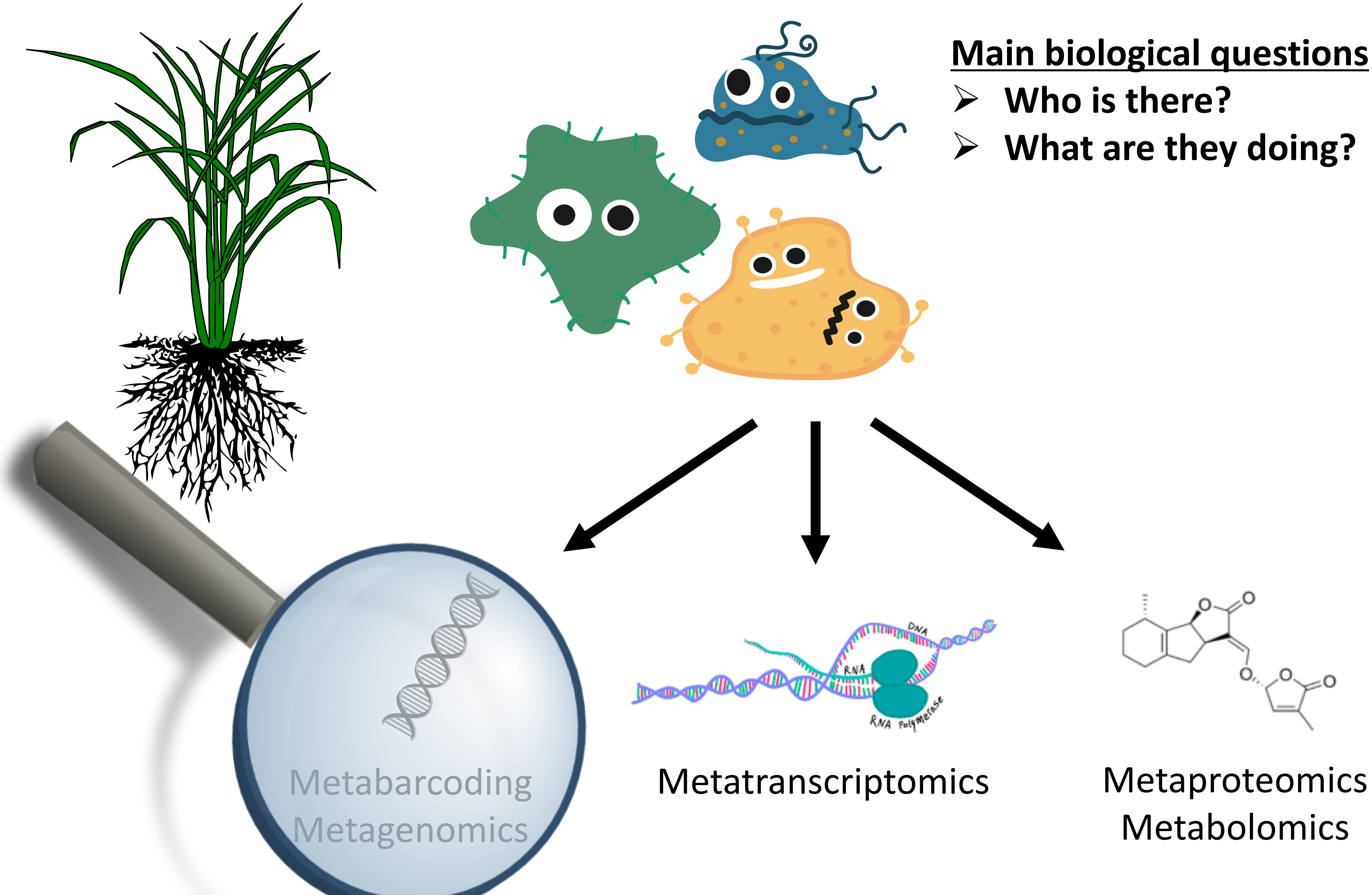
Main biological questions

- **Who is there?**
- **What are they doing?**
- **Which microbe is associated with a specific phenotype?** (*i.e.* feature selection)
- **Unravel how microbiome is recruited?**

Multi-omics approach and data integration



Main biological questions and methods



Main biological questions

- ➢ **Who is there?**
- ➢ **What are they doing?**

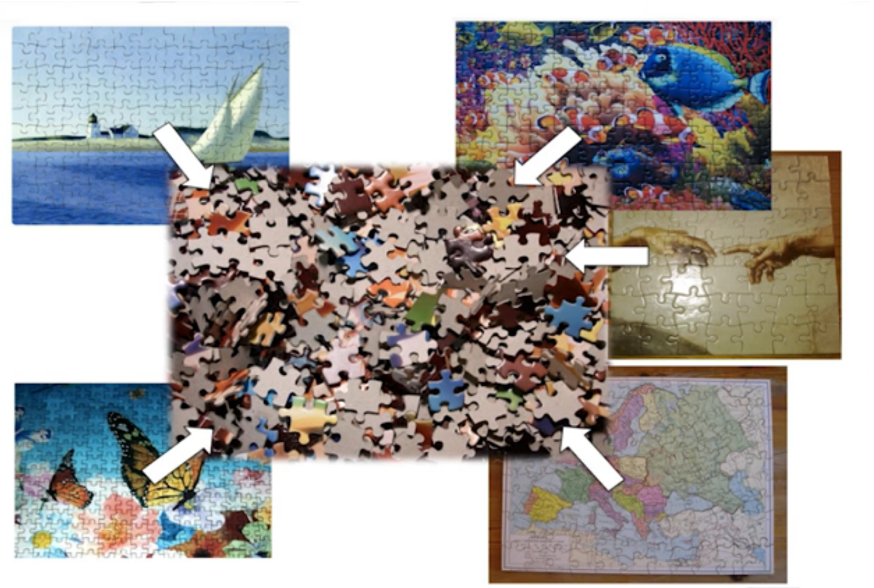
Metabarcoding
Metagenomics

Metatranscriptomics

Metaproteomics
Metabolomics

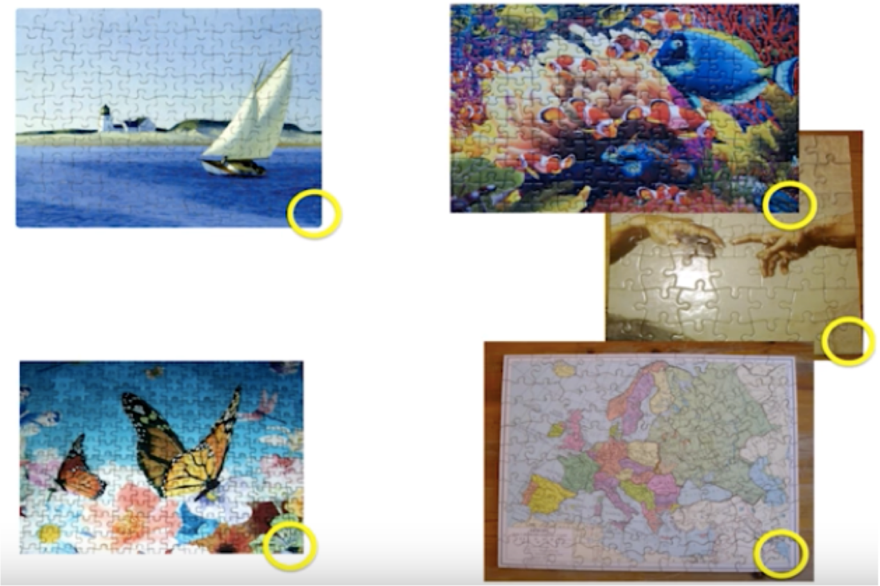
Methods to assess microbial composition and diversity

Metagenomics
(shotgun sequencing)



- Sequence all DNA
- Higher cost
- Higher complexity
- Environmental contamination
- Functional information

Metabarcoding
(amplicon sequencing)



- Sequence only specific gene
- Cheaper
- Less complex to analyse
- Primer amplification bias
- No functional information
- Difficult to identify species

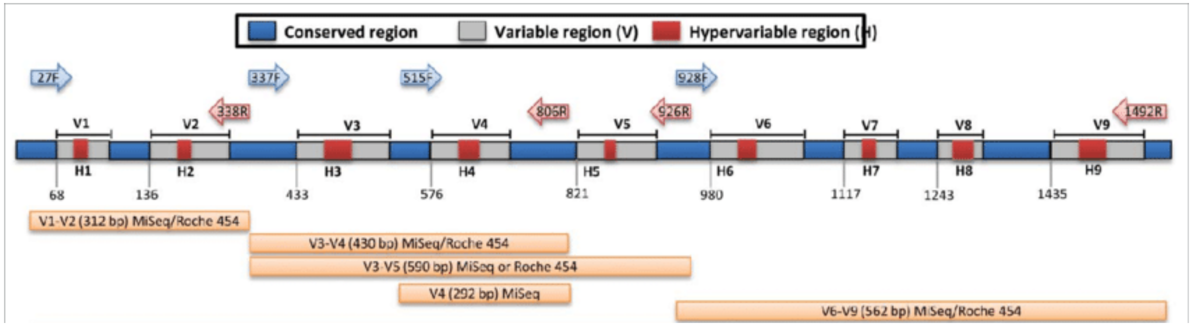
A targeted approach: metabarcoding/amplicon sequencing

Requirements

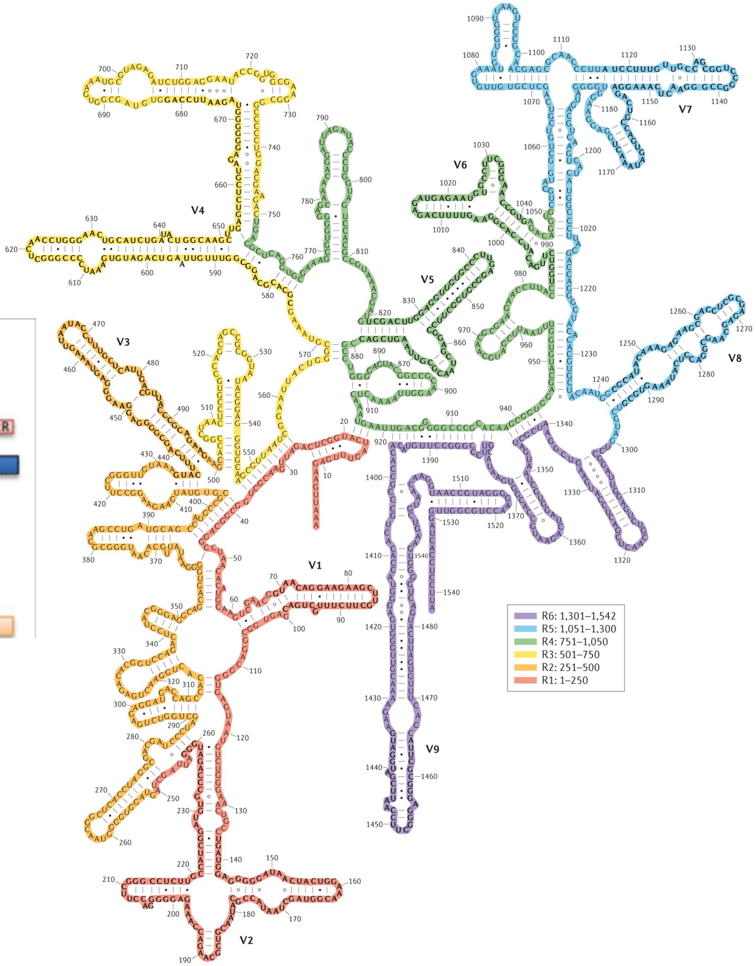
- Gene ubiquitous
- With conserve and variable regions

For Bacteria analysis: 16S rRNA gene

- Gene code for a RNA part of the ribosome



Adapted from Shahi et al 2017



Yarza et al. 2014

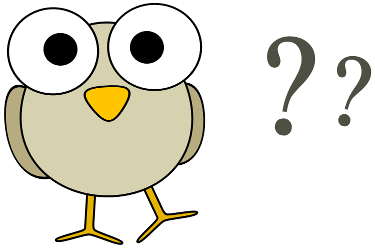
Nature Reviews | Microbiology

For Fungi analysis: 18S rRNA gene or ITS

Main biological questions and methods

Test your knowledge…

- Please answer the 2 questions in the following quiz https://bigdata_microbiome.presenterswall.nl/



Main biological questions and methods

- Who is there? What are they doing?
- Different approaches based on DNA
  - Metagenomics = all DNA
  - Metabarcoding = one specific ubiquitous gene with conserved and variable regions (16S rRNA, 18S rRNA or ITS)

Learning objectives

- ☑ Define microbiome and state microbiome importance
- ☑ Identify differences between metabarcoding and metagenomics
- □ Explain how microbiota data are generated (including bias)
- □ Explain and preform data pre-processing
- □ Explain how microbiota data are analysed
- □ Define, perform and interpret alpha-diversity
- □ Address sparsity, under-sampling and uneven sampling depth using data filtering and normalization
- □ Define, perform and interpret beta-diversity
- □ Generate and interpret multivariate data analyses
- □ Perform and interpret appropriate statistical tests
- □ Visualize and interpret microbial community composition

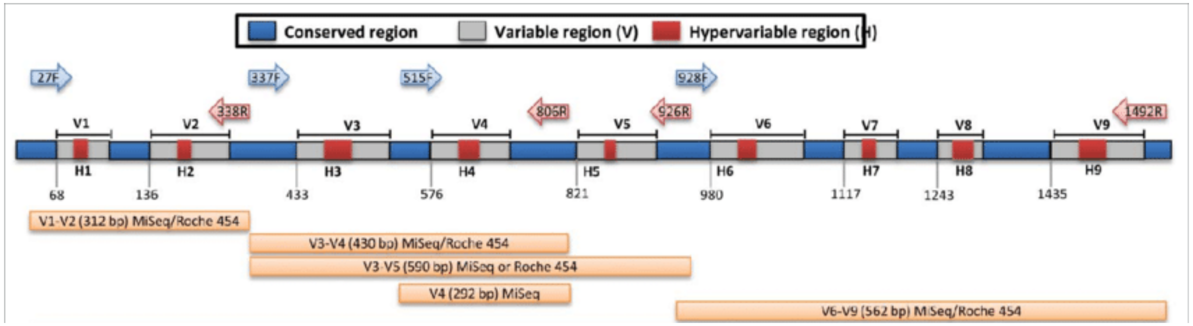
A targeted approach: metabarcoding/amplicon sequencing

Requirements

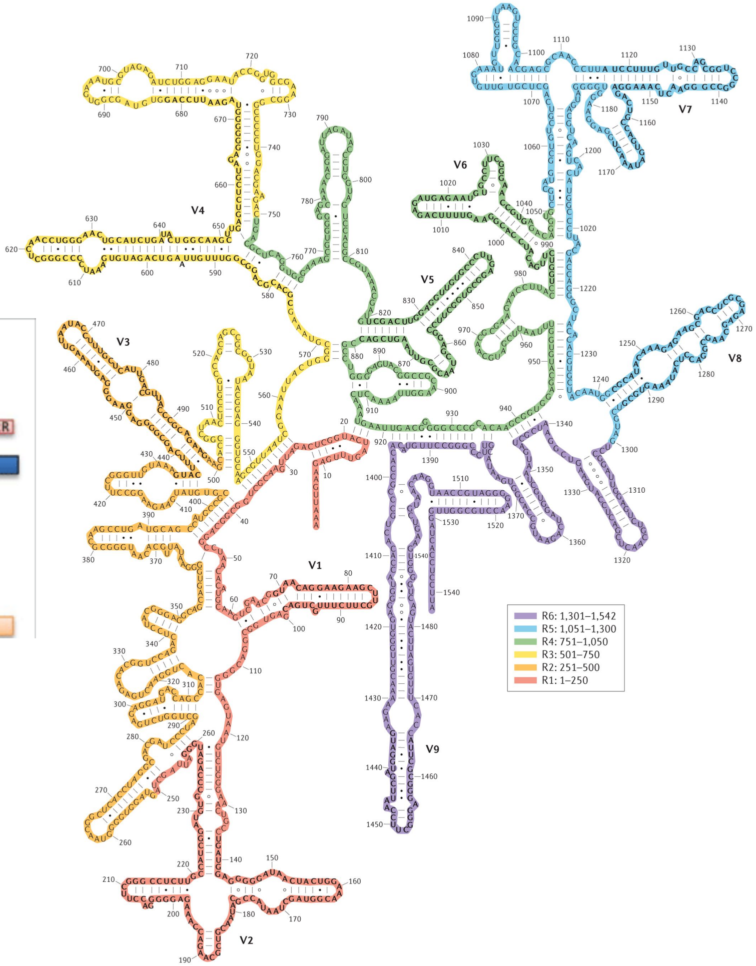
- Gene ubiquitous
- With conserve and variable regions

For Bacteria analysis: 16S rRNA gene

- Gene code for a RNA part of the ribosome



Adapted from Shahi et al 2017



Yarza et al. 2014

Nature Reviews | Microbiology

For Fungi analysis: 18S rRNA gene or ITS

How microbiota data are generated?

What is microbiome?

How microbiota data are generated?

How microbiota data are analysed?

Part 1

- Definitions
- Microbiome importance
- Scientific questions
- Differences between metagenomics and metabarcoding

Part 2

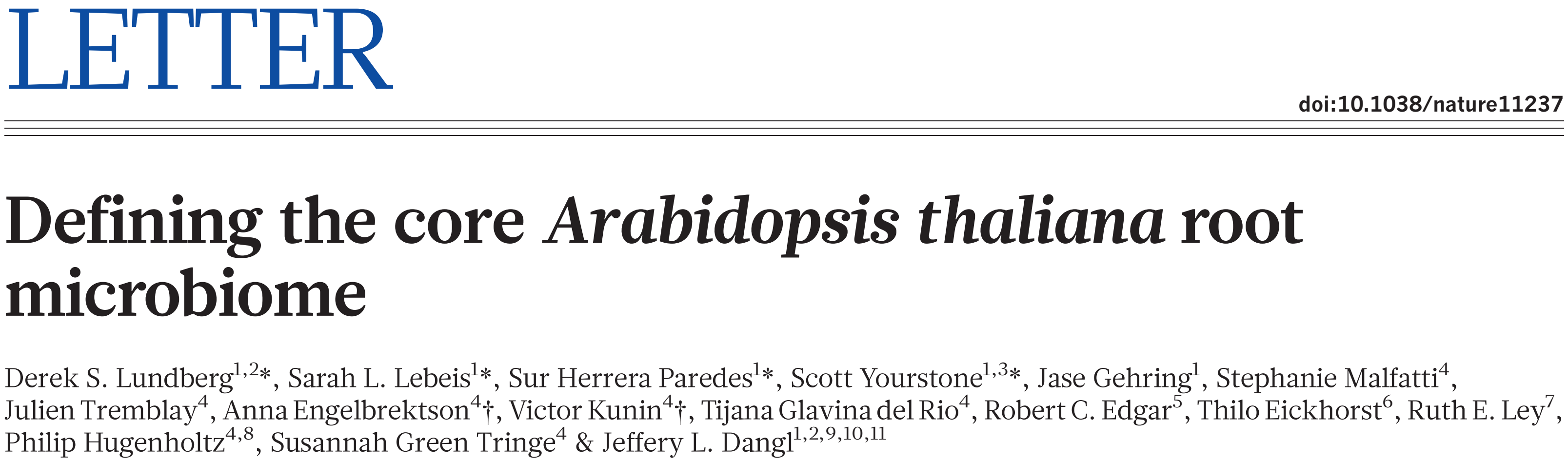
- From samples to sequences
- From sequences to data sets

Part 3

- Alpha-diversity
- Data properties
- Data filtering
- Data normalisation
- Beta-diversity
- Microbial composition

A research example: plant root microbiome

Objective: illustration through a concrete case



LETTER

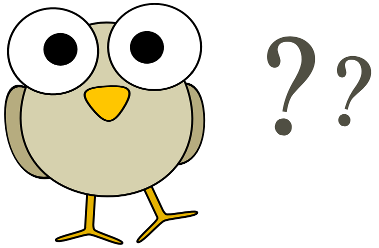
doi:10.1038/nature11237

Defining the core *Arabidopsis thaliana* root microbiome

Derek S. Lundberg[1,2]*, Sarah L. Lebeis[1]*, Sur Herrera Paredes[1]*, Scott Yourstone[1,3]*, Jase Gehring[1], Stephanie Malfatti[4], Julien Tremblay[4], Anna Engelbrektson[4]†, Victor Kunin[4]†, Tijana Glavina del Rio[4], Robert C. Edgar[5], Thilo Eickhorst[6], Ruth E. Ley[7], Philip Hugenholtz[4,8], Susannah Green Tringe[4] & Jeffery L. Dangl[1,2,9,10,11]

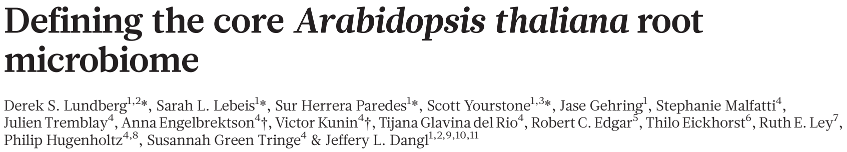
Please answer two quiz questions…

https://bigdata_microbiome.presenterswall.nl/



Step 1: From sample to sequences

Process overview

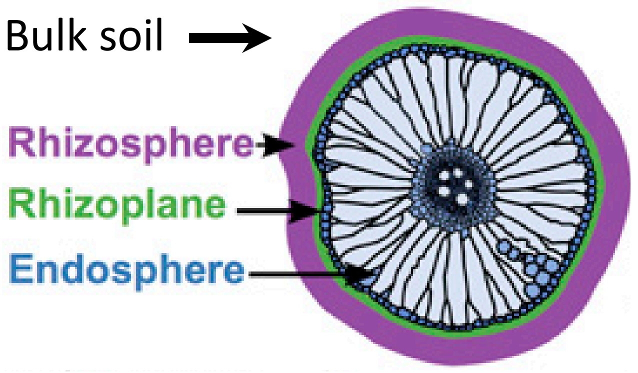


Defining the core *Arabidopsis thaliana* root microbiome

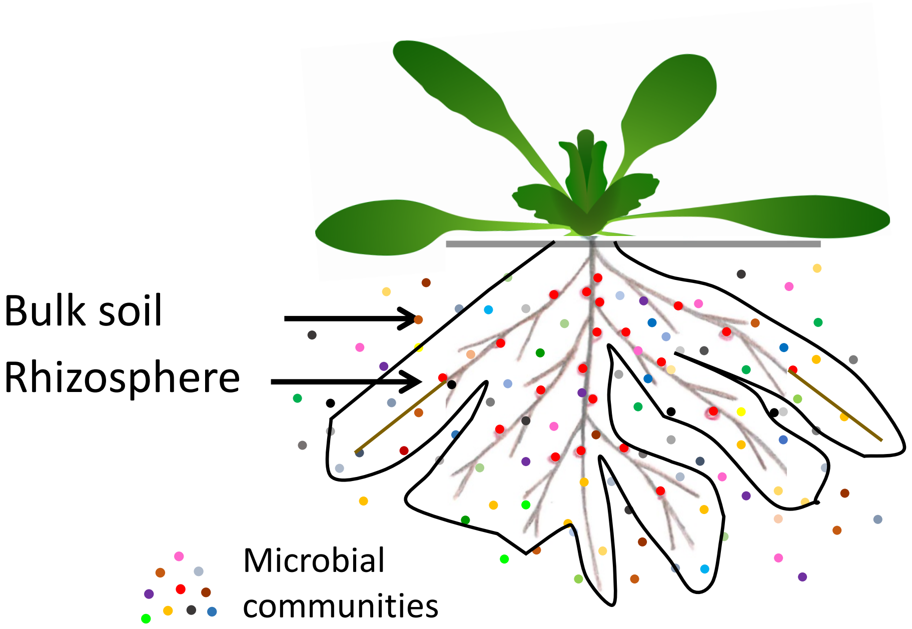
Derek S. Lundberg[1,2]*, Sarah L. Lebeis[1]*, Sur Herrera Paredes[1]*, Scott Yourstone[1,3]*, Jase Gehring[1], Stephanie Malfatti[4], Julien Tremblay[4], Anna Engelbrektson[4]†, Victor Kunin[4]†, Tijana Glavina del Rio[4], Robert C. Edgar[5], Thilo Eickhorst[6], Ruth E. Ley[7], Philip Hugenholtz[4,8], Susannah Green Tringe[4] & Jeffery L. Dangl[1,2,9,10,11]

Sampling

- Three compartments
  - ❑ Bulk soil
  - ❑ Rhizosphere soil
  - ❑ Endosphere



Adapted from Edwards et al. 2014

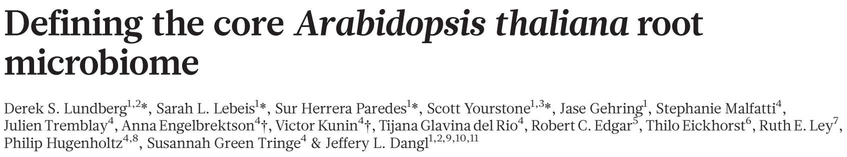


Step 1: From sample to sequences

Process overview

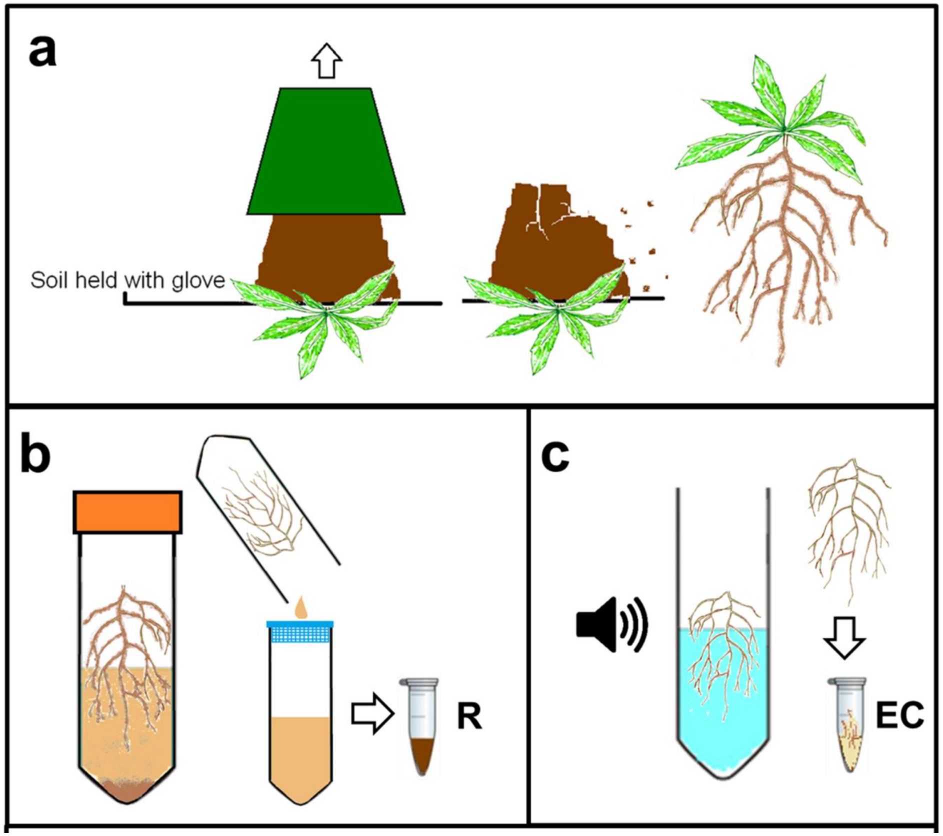
Sampling

- Three compartments
  - ❑ Bulk soil
  - ❑ Rhizosphere soil
  - ❑ Endosphere

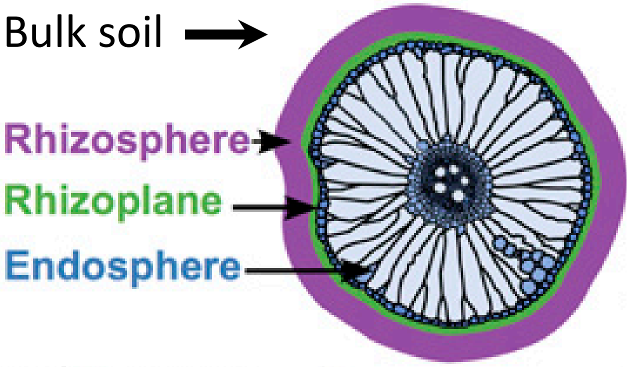


Defining the core *Arabidopsis thaliana* root microbiome

Derek S. Lundberg[1,2]*, Sarah L. Lebeis[1]*, Sur Herrera Paredes[1]*, Scott Yourstone[1,3]*, Jase Gehring[1], Stephanie Malfatti[4], Julien Tremblay[4], Anna Engelbrektson[4]†, Victor Kunin[4]†, Tijana Glavina del Rio[4], Robert C. Edgar[5], Thilo Eickhorst[6], Ruth E. Ley[7], Philip Hugenholtz[4,8], Susannah Green Tringe[4] & Jeffery L. Dangl[1,2,9,10,11]



Lundberg et al. 2012

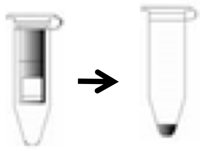


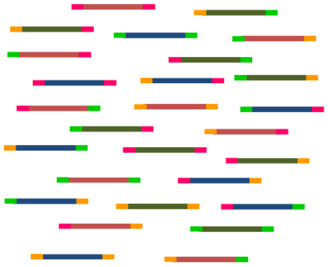
Adapted from Edwards et al. 2014

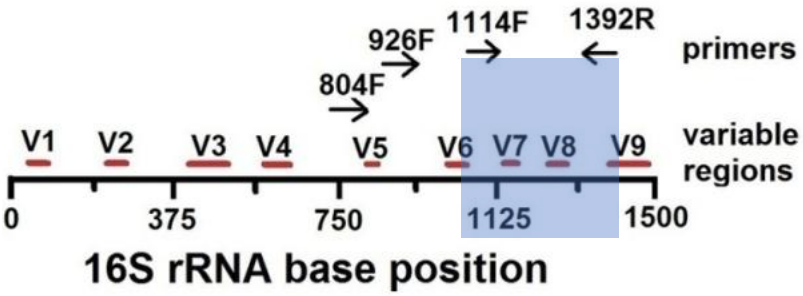
Step 1: From sample to sequences

Process overview

Sampling → DNA extraction → Amplification



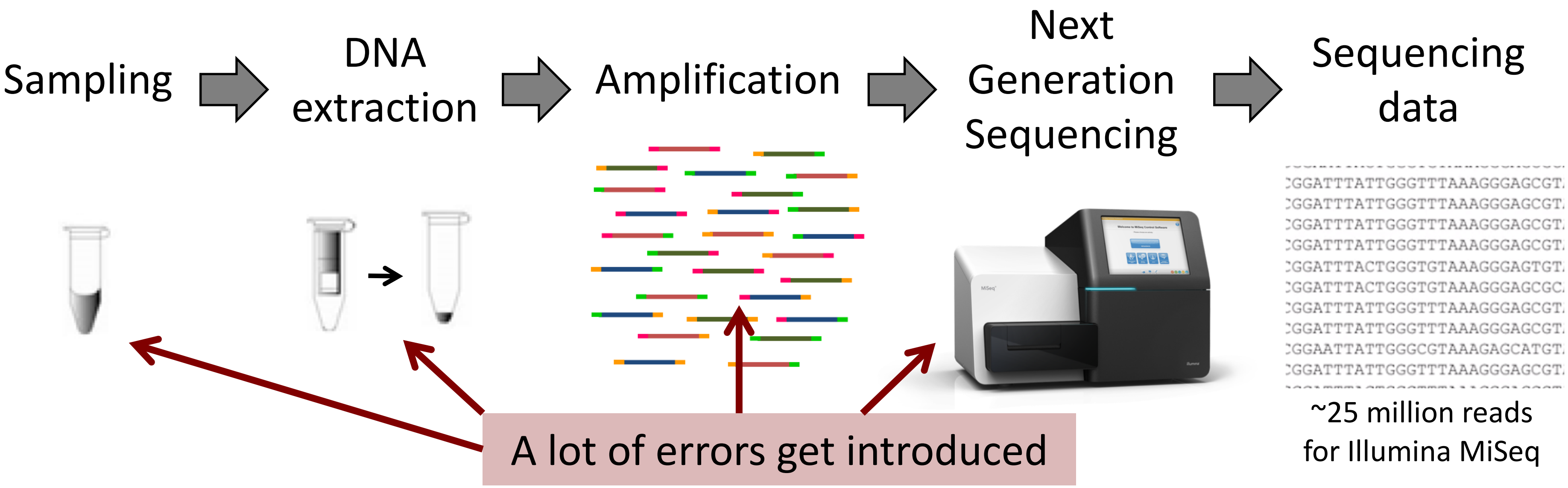




Adapted from Lundberg et al. 2012

Step 1: From sample to sequences

Process overview

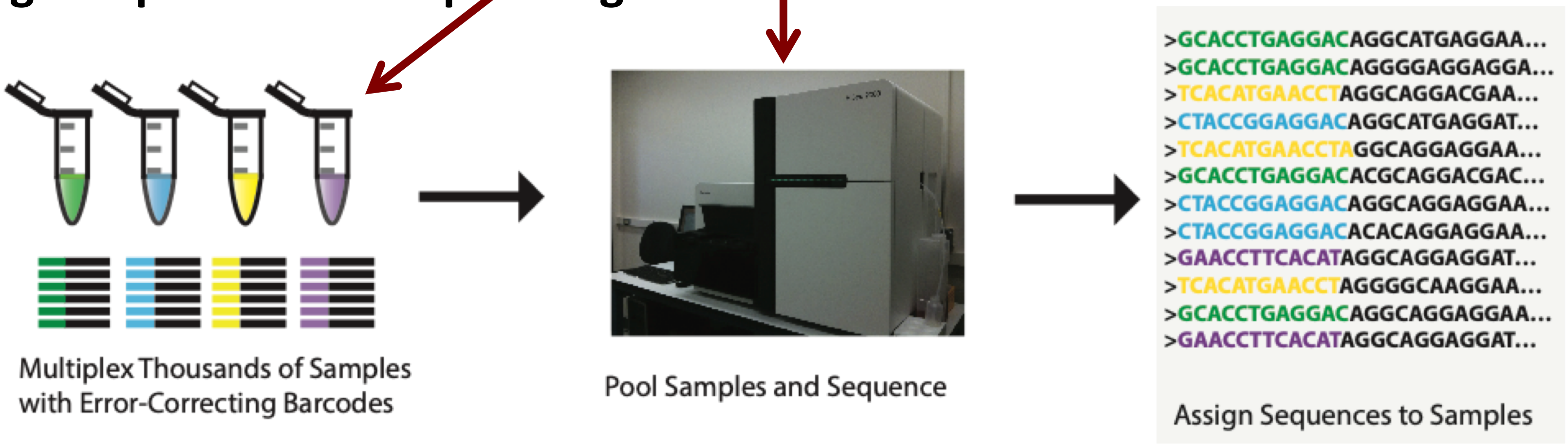


```
CGGATTTATTGGGTTTAAAGGGAGCGT
CGGATTTATTGGGTTTAAAGGGAGCGT
CGGATTTATTGGGTTTAAAGGGAGCGT
CGGATTTATTGGGTTTAAAGGGAGCGT
CGGATTTACTGGGTGTAAAGGGAGTGT
CGGATTTACTGGGTGTAAAGGGAGCGC
CGGATTTATTGGGTTTAAAGGGAGCGT
CGGATTTATTGGGTTTAAAGGGAGCGT
CGGAATTATTGGGCGTAAAGAGCATGT
CGGATTTATTGGGTTTAAAGGGAGCGT
```

~25 million reads for Illumina MiSeq

A lot of errors get introduced

Mixing samples in one sequencing run



```
>GCACCTGAGGACAGGCATGAGGAA...
>GCACCTGAGGACAGGGGAGGAGGA...
>TCACATGAACCTAGGCAGGACGAA...
>CTACCGGAGGACAGGCATGAGGAT...
>TCACATGAACCTAGGCAGGAGGAA...
>GCACCTGAGGACACGCAGGACGAC...
>CTACCGGAGGACAGGCAGGAGGAA...
>CTACCGGAGGACACACAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
>TCACATGAACCTAGGGGCAAGGAA...
>GCACCTGAGGACAGGCAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
```

Multiplex Thousands of Samples with Error-Correcting Barcodes

Pool Samples and Sequence

Assign Sequences to Samples

Adapted from Metcalf, Jessica (2014): Overview of data generation, processing and analysis using QIIME. Figshare. https://doi.org/10.6084/m9.figshare.902219.v1

Step 1: From sample to sequences

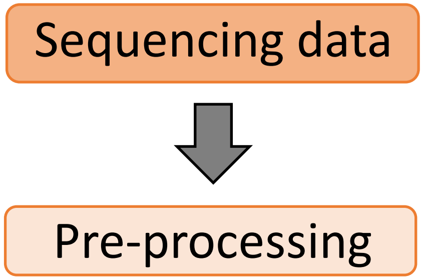
- Dont forget that there are bias
  It will be difficult to
  - Assess the entire microbial community
  - Obtain same amount of sequences per sample

Learning objectives

- ☑ Define microbiome and state microbiome importance
- ☑ Identify differences between metabarcoding and metagenomics
- ☑ Explain how microbiota data are generated (including bias)
- □ Explain and preform data pre-processing
- □ Explain how microbiota data are analysed
- □ Define, perform and interpret alpha-diversity
- □ Address sparsity, under-sampling and uneven sampling depth using data filtering and normalization
- □ Define, perform and interpret beta-diversity
- □ Generate and interpret multivariate data analyses
- □ Perform and interpret appropriate statistical tests
- □ Visualize and interpret microbial community composition

Step 2: From sequences to microbiota data sets

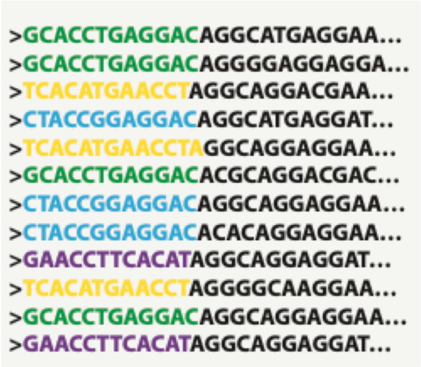
Process overview



Sequencing data

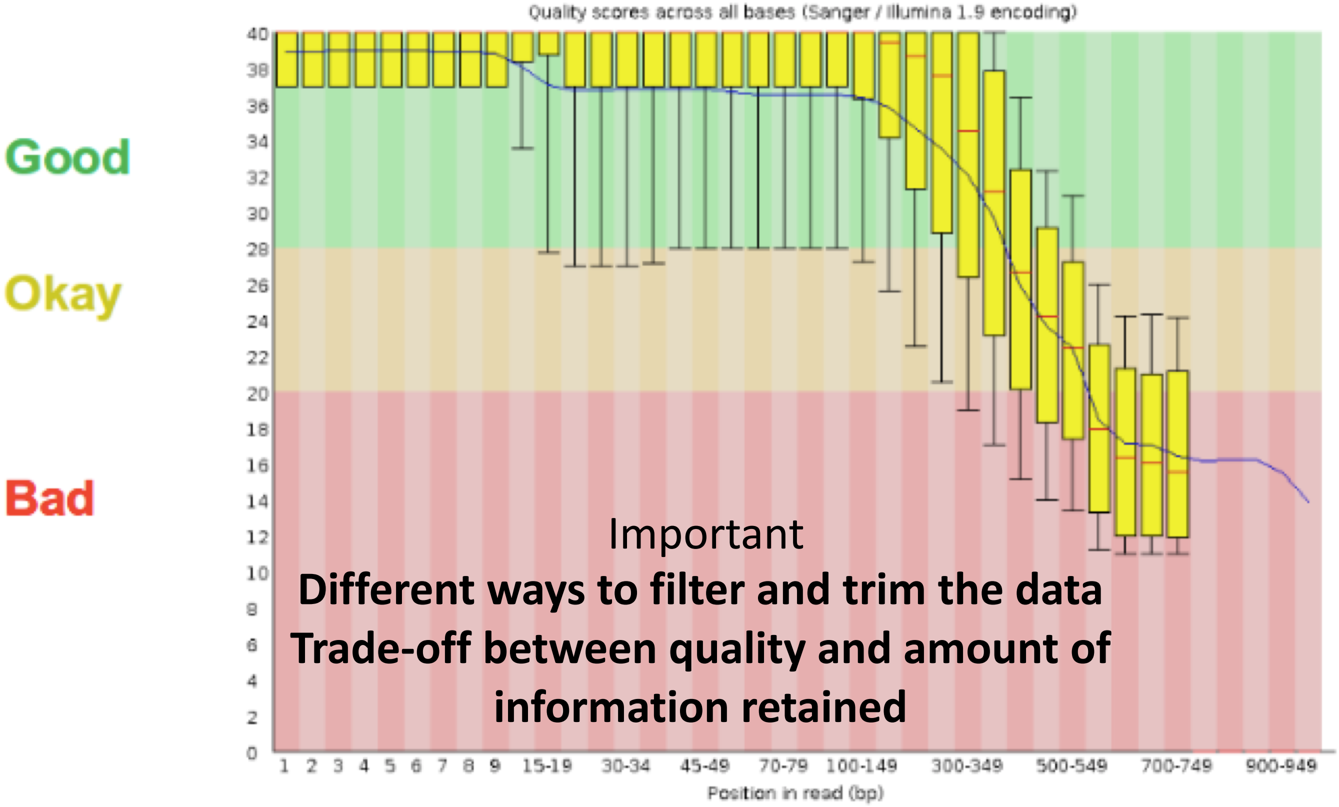
Pre-processing

- De-multiplex (*i.e.* assign a sequence to a sample)
- Remove adaptor and barcode
- Remove low quality reads (*i.e.* filtering step)



```
>GCACCTGAGGACAGGCATGAGGAA...
>GCACCTGAGGACAGGGGAGGAGGA...
>TCACATGAACCTAGGCAGGACGAA...
>CTACCGGAGGACAGGCATGAGGAT...
>TCACATGAACCTAGGCAGGAGGAA...
>GCACCTGAGGACACGCAGGACGAC...
>CTACCGGAGGACAGGCAGGAGGAA...
>CTACCGGAGGACACACAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
>TCACATGAACCTAGGGGCAAGGAA...
>GCACCTGAGGACAGGCAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
```

Metcalf 2014



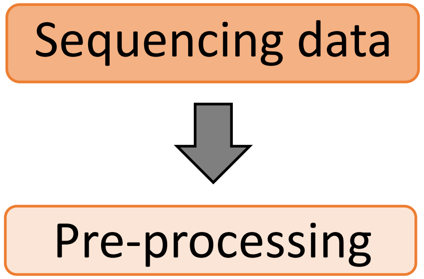
Important

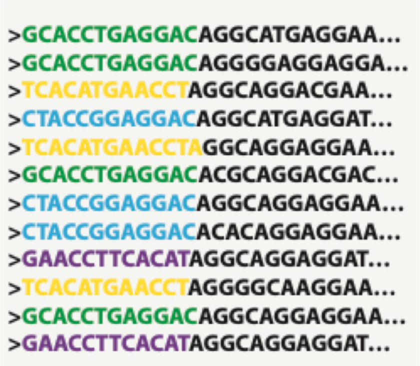
Different ways to filter and trim the data

Trade-off between quality and amount of information retained

Step 2: From sequences to microbiota data sets

Process overview

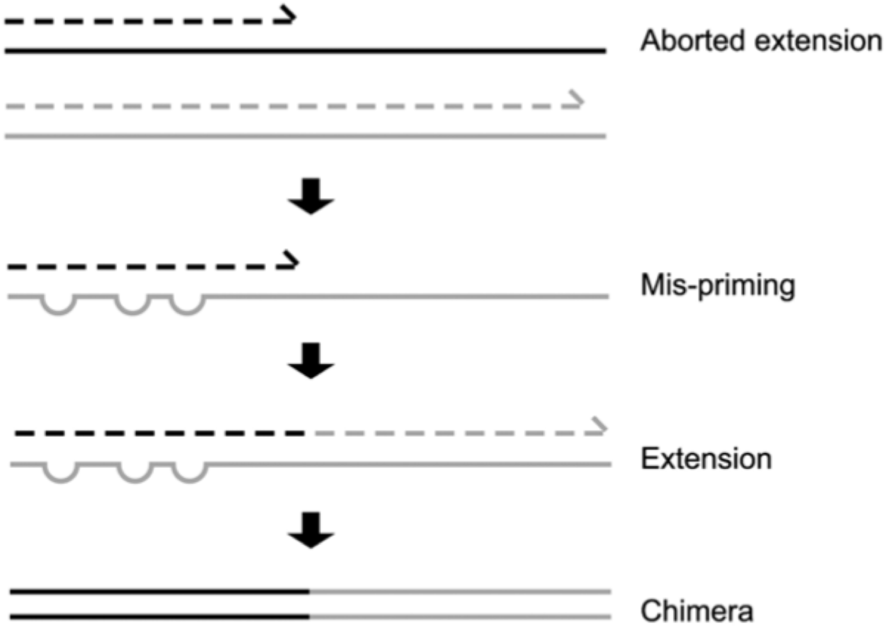




```
>GCACCTGAGGACAGGCATGAGGAA...
>GCACCTGAGGACAGGGGAGGAGGA...
>TCACATGAACCTAGGCAGGACGAA...
>CTACCGGAGGACAGGCATGAGGAT...
>TCACATGAACCTAGGCAGGAGGAA...
>GCACCTGAGGACACGCAGGACGAC...
>CTACCGGAGGACAGGCAGGAGGAA...
>CTACCGGAGGACACACAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
>TCACATGAACCTAGGGGCAAGGAA...
>GCACCTGAGGACAGGCAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
```

Metcalf 2014

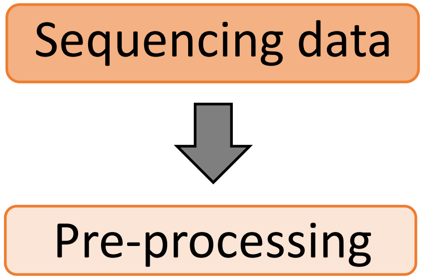
- De-multiplex (*i.e.* assign a sequence to a sample)
- Remove adaptor and barcode
- Remove low quality reads (*i.e.* filtering step)
- Remove chimeras

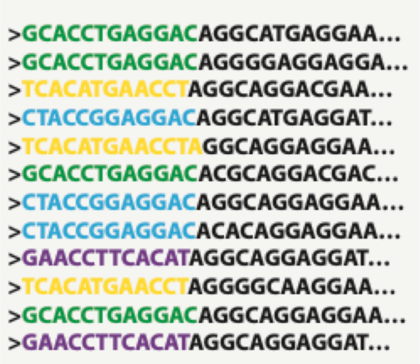


During PCR multiple sequence can combine to form a hybrid
Chimeras must be removed

Step 2: From sequences to microbiota data sets

Process overview

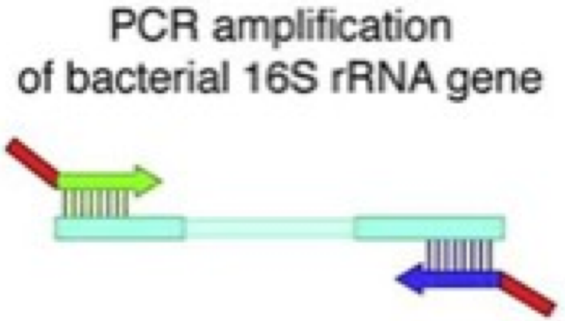




```
>GCACCTGAGGACAGGCATGAGGAA...
>GCACCTGAGGACAGGGGAGGAGGA...
>TCACATGAACCTAGGCAGGACGAA...
>CTACCGGAGGACAGGCATGAGGAT...
>TCACATGAACCTAGGCAGGAGGAA...
>GCACCTGAGGACACGCAGGACGAC...
>CTACCGGAGGACAGGCAGGAGGAA...
>CTACCGGAGGACACACAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
>TCACATGAACCTAGGGGCAAGGAA...
>GCACCTGAGGACAGGCAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
```

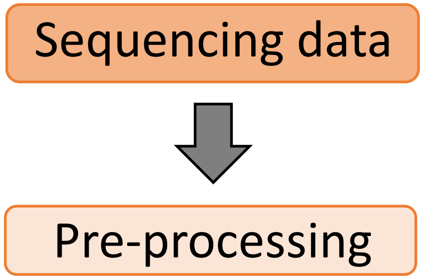
Metcalf 2014

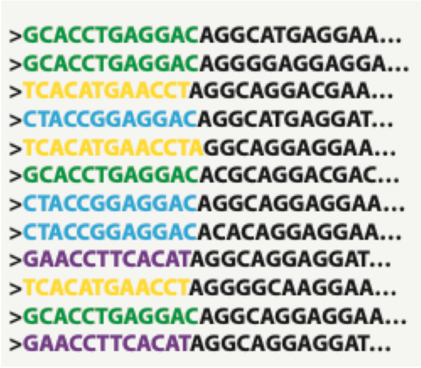
- De-multiplex (*i.e.* assign a sequence to a sample)
- Remove adaptor and barcode
- Remove low quality reads (*i.e.* filtering step)
- Remove chimeras
- Merged pair-end reads



Step 2: From sequences to microbiota data sets

Process overview

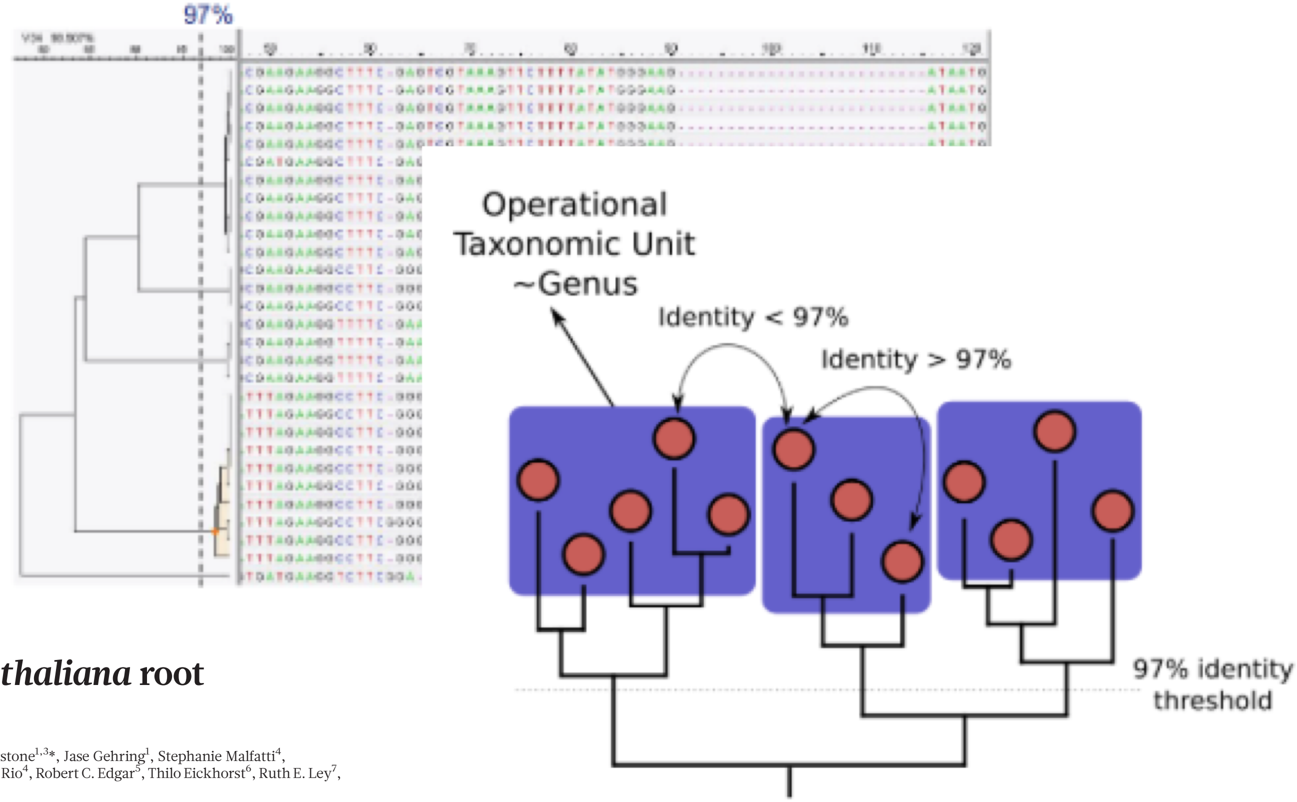




>GCACCTGAGGACAGGCATGAGGAA...
>GCACCTGAGGACAGGGGAGGAGGA...
>TCACATGAACCTAGGCAGGACGAA...
>CTACCGGAGGACAGGCATGAGGAT...
>TCACATGAACCTAGGCAGGAGGAA...
>GCACCTGAGGACACGCAGGACGAC...
>CTACCGGAGGACAGGCAGGAGGAA...
>CTACCGGAGGACACACAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
>TCACATGAACCTAGGGGCAAGGAA...
>GCACCTGAGGACAGGCAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...

Metcalf 2014

- De-multiplex (*i.e.* assign a sequence to a sample)
- Remove adaptor and barcode
- Remove low quality reads (*i.e.* filtering step)
- Remove chimeras
- Merged pair-end reads
- Sequence clustering in OTU

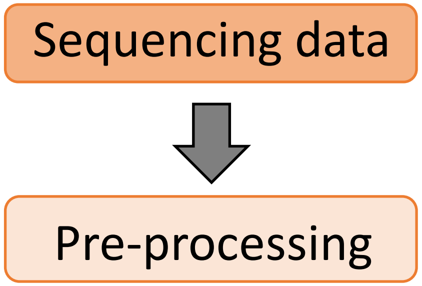


Defining the core *Arabidopsis thaliana* root microbiome

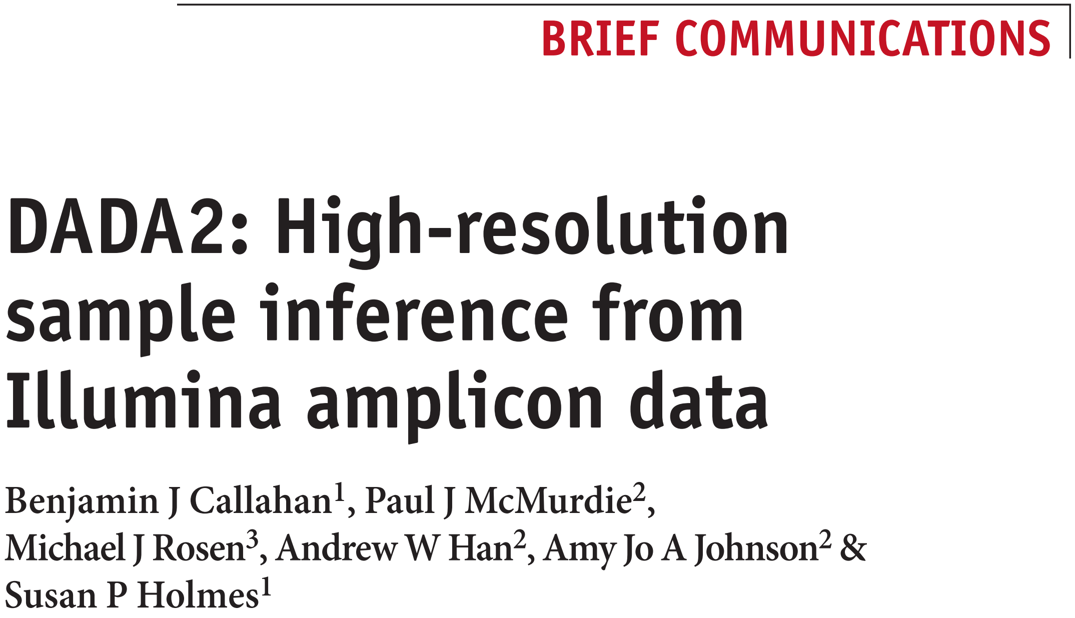
Derek S. Lundberg[1,2]*, Sarah L. Lebeis[1]*, Sur Herrera Paredes[1]*, Scott Yourstone[1,3]*, Jase Gehring[1], Stephanie Malfatti[4], Julien Tremblay[4], Anna Engelbrektson[4]†, Victor Kunin[4]†, Tijana Glavina del Rio[4], Robert C. Edgar[5], Thilo Eickhorst[6], Ruth E. Ley[7], Philip Hugenholtz[4,8], Susannah Green Tringe[4] & Jeffery L. Dangl[1,2,9,10,11]

Step 2: From sequences to microbiota data sets

Process overview



- A new pre-processing pipeline DADA2
- Using Divisive Amplicon Denoising Algorithm (DADA) to correct amplicon errors without constructing OTU (*i.e.* Amplicon Sequence Variants or ASV)



BRIEF COMMUNICATIONS

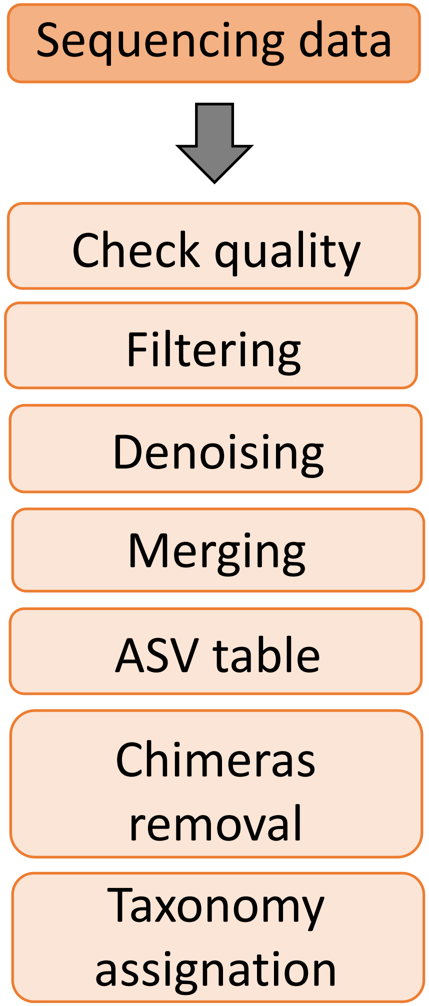
DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan[1], Paul J McMurdie[2], Michael J Rosen[3], Andrew W Han[2], Amy Jo A Johnson[2] & Susan P Holmes[1]

Step 2: From sequences to microbiota data sets

- A new pre-processing pipeline DADA2
- Using Divisive Amplicon Denoising Algorithm (DADA) to correct amplicon errors without constructing OTU (*i.e.* Amplicon Sequence Variants or ASV)

Process overview



Sequencing data

Check quality → *plotQualityProfile()* visualize the quality profile

Filtering → *filterAndTrim()* trims sequences to a specific length and filters based on quality

Denoising → *learnErrors()* learn the error rates & *dada()* implements DADA

Merging → *mergePairs()* merges forward and reverse if they exactly overlap

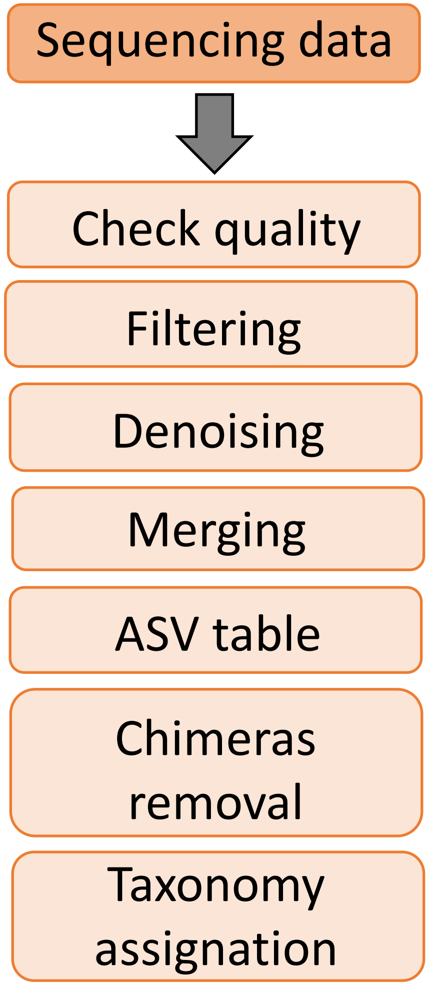
ASV table → *makeSequenceTable()* construct the amplicon sequence variant table

Chimeras removal → *removeBimeraDenovo()* identifies sequences that are exact bimeras (two-parent chimeras) of more abundant sequences

Taxonomy assignation → *assignTaxonomy()* assign taxonomy to the ASV

Step 2: From sequences to microbiota data sets

Process overview



Sequencing data

Check quality

Filtering

Denoising

Merging

ASV table

Chimeras removal

Taxonomy assignment

- Looking for sequence homology with ref databases
- Accuracy depends on quality and completeness of the database

silva

high quality ribosomal RNA databases

SILVA database

Ribosomal Data Project database

rdp

Greengenes database

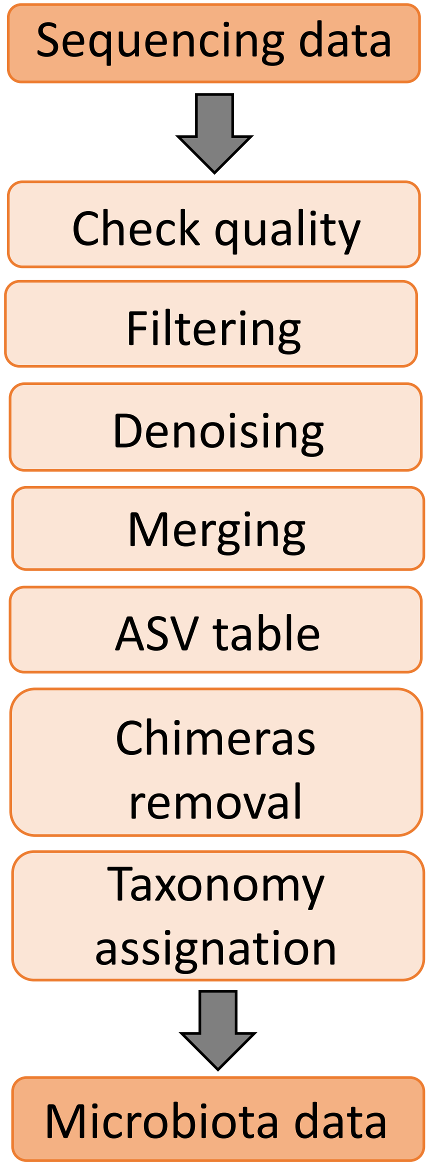
GREENGENES The 16S rRNA Gene Database and Tools

Defining the core *Arabidopsis thaliana* root microbiome

Derek S. Lundberg[1,2]*, Sarah L. Lebeis[1]*, Sur Herrera Paredes[1]*, Scott Yourstone[1,3]*, Jase Gehring[1], Stephanie Malfatti[4], Julien Tremblay[4], Anna Engelbrektson[4]†, Victor Kunin[4]†, Tijana Glavina del Rio[4], Robert C. Edgar[5], Thilo Eickhorst[6], Ruth E. Ley[7], Philip Hugenholtz[4,8], Susannah Green Tringe[4] & Jeffery L. Dangl[1,2,9,10,11]

Step 2: From sequences to microbiota data sets

Process overview



Sequencing data

Check quality

Filtering

Denoising

Merging

ASV table

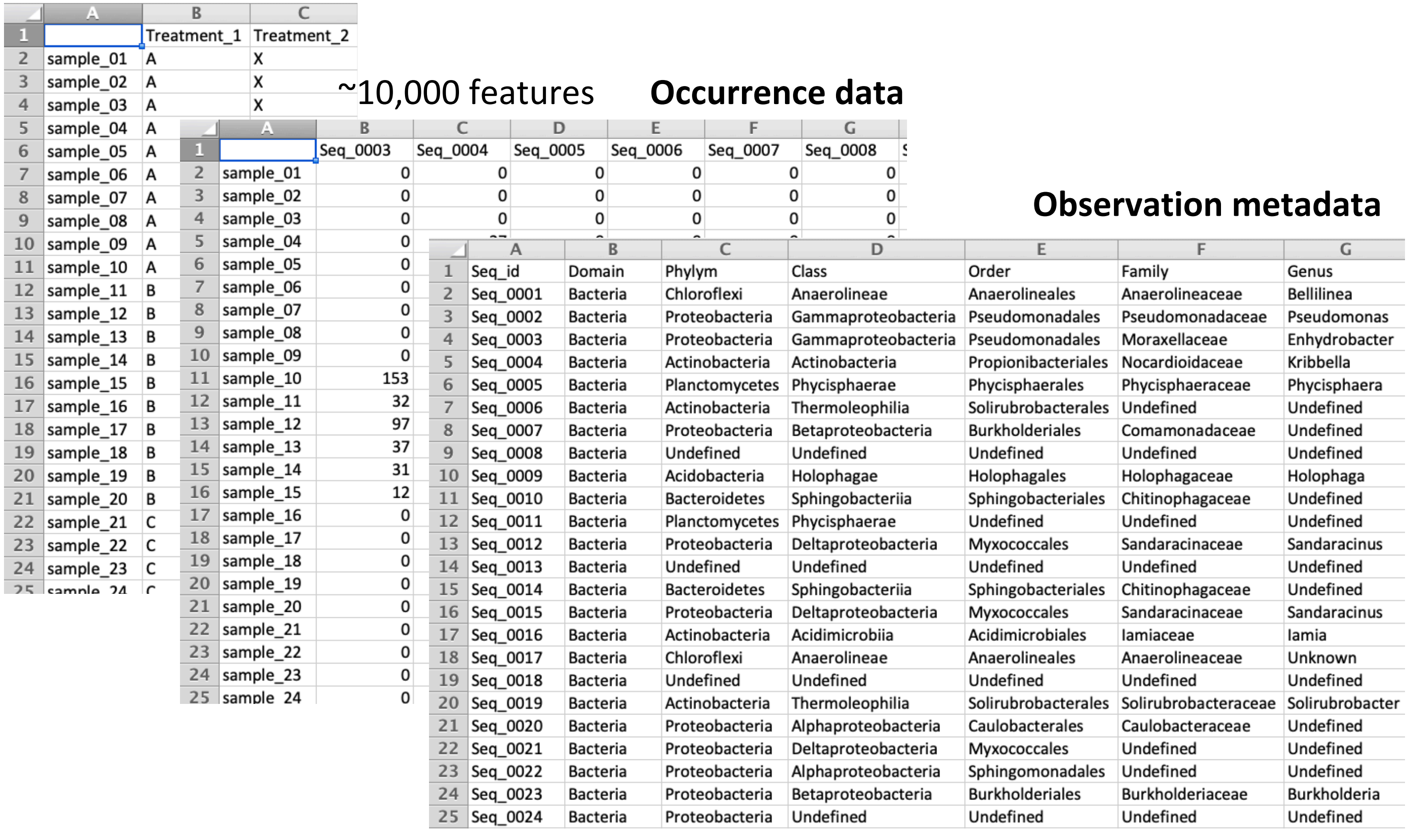
Chimeras removal

Taxonomy assignation

Microbiota data

Data sets output

- Sample metadata
- Occurrence data
- Observation metadata (taxonomic assignation)



Sample metadata

~100 samples

	A	B	C
1		Treatment_1	Treatment_2
2	sample_01	A	X
3	sample_02	A	X
4	sample_03	A	X
5	sample_04	A	
6	sample_05	A	
7	sample_06	A	
8	sample_07	A	
9	sample_08	A	
10	sample_09	A	
11	sample_10	A	
12	sample_11	B	
13	sample_12	B	
14	sample_13	B	
15	sample_14	B	
16	sample_15	B	
17	sample_16	B	
18	sample_17	B	
19	sample_18	B	
20	sample_19	B	
21	sample_20	B	
22	sample_21	C	
23	sample_22	C	
24	sample_23	C	
25	sample_24	C	

~10,000 features **Occurrence data**

	A	B	C	D	E	F	G
1		Seq_0003	Seq_0004	Seq_0005	Seq_0006	Seq_0007	Seq_0008
2	sample_01	0	0	0	0	0	0
3	sample_02	0	0	0	0	0	0
4	sample_03	0	0	0	0	0	0
5	sample_04	0					
6	sample_05	0					
7	sample_06	0					
8	sample_07	0					
9	sample_08	0					
10	sample_09	0					
11	sample_10	153					
12	sample_11	32					
13	sample_12	97					
14	sample_13	37					
15	sample_14	31					
16	sample_15	12					
17	sample_16	0					
18	sample_17	0					
19	sample_18	0					
20	sample_19	0					
21	sample_20	0					
22	sample_21	0					
23	sample_22	0					
24	sample_23	0					
25	sample_24	0					

Observation metadata

	A	B	C	D	E	F	G
1	Seq_id	Domain	Phylym	Class	Order	Family	Genus
2	Seq_0001	Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	Anaerolineaceae	Bellilinea
3	Seq_0002	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas
4	Seq_0003	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Enhydrobacter
5	Seq_0004	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Nocardioidaceae	Kribbella
6	Seq_0005	Bacteria	Planctomycetes	Phycisphaerae	Phycisphaerales	Phycisphaeraceae	Phycisphaera
7	Seq_0006	Bacteria	Actinobacteria	Thermoleophilia	Solirubrobacterales	Undefined	Undefined
8	Seq_0007	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Undefined
9	Seq_0008	Bacteria	Undefined	Undefined	Undefined	Undefined	Undefined
10	Seq_0009	Bacteria	Acidobacteria	Holophagae	Holophagales	Holophagaceae	Holophaga
11	Seq_0010	Bacteria	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Chitinophagaceae	Undefined
12	Seq_0011	Bacteria	Planctomycetes	Phycisphaerae	Undefined	Undefined	Undefined
13	Seq_0012	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	Sandaracinaceae	Sandaracinus
14	Seq_0013	Bacteria	Undefined	Undefined	Undefined	Undefined	Undefined
15	Seq_0014	Bacteria	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Chitinophagaceae	Undefined
16	Seq_0015	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	Sandaracinaceae	Sandaracinus
17	Seq_0016	Bacteria	Actinobacteria	Acidimicrobiia	Acidimicrobiales	Iamiaceae	Iamia
18	Seq_0017	Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	Anaerolineaceae	Unknown
19	Seq_0018	Bacteria	Undefined	Undefined	Undefined	Undefined	Undefined
20	Seq_0019	Bacteria	Actinobacteria	Thermoleophilia	Solirubrobacterales	Solirubrobacteraceae	Solirubrobacter
21	Seq_0020	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Undefined
22	Seq_0021	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	Undefined	Undefined
23	Seq_0022	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Undefined	Undefined
24	Seq_0023	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia
25	Seq_0024	Bacteria	Proteobacteria	Undefined	Undefined	Undefined	Undefined

Taxonomic assignment

Example of the bacteria *Escherichia coli* O157:H7

Domain	Bacteria
Kingdom	Eubacteria
Phylum	Proteobacteria
Class	Gammaproteobacteria
Order	Enterobacterales
Family	Enterobacteriaceae
Genus	Escherichia-Shigella
Species	*Escherichia coli*
Strain	O157:H7

Taxonomic assignment

Example of the bacteria *Escherichia coli* O157:H7 -> ASV_6287

Domain	Bacteria
Kingdom	Eubacteria
Phylum	Proteobacteria
Class	Gammaproteobacteria
Order	Enterobacterales
Family	Enterobacteriaceae
Genus	Undefined
Species	Undefined
Strain	-

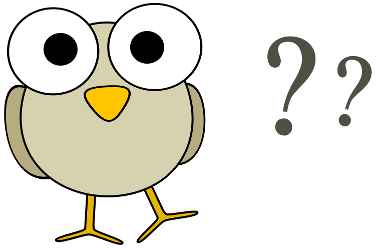
Step 2: From sequences to microbiota data sets

- Data pre-processing: always a trade-off between quality and quantity
- OTU Operational Taxonomic Units ≠ ASV Amplicon Sequence Variants
- Go from fasta files to three tables
  - occurrence table
  - taxonomic assignation
  - sample metadata

How microbiota data are generated?

Test your knowledge…

- Please answer the 3 questions in the following quiz https://bigdata_microbiome.presenterswall.nl/



Practice time: from sequences to microbiota data sets

In the tutorial, look at:

- Getting ready
- Inspect read quality profiles
- Filter and trim
- Learn the error rates
- Sample inference
- Merge paired reads
- Construct sequence table
- Remove chimeras
- Track reads through the pipeline
- Assign taxonomy

Tutorial link:

http://benjjneb.github.io/dada2/tutorial.html

Script on Canvas or link:

https://scienceparkstudygroup.github.io/microbiome-lesson/02-data-preprocess-fastq-to-asv/index.html

Learning objectives

- ☑ Define microbiome and state microbiome importance
- ☑ Identify differences between metabarcoding and metagenomics
- ☑ Explain how microbiota data are generated (including bias)
- ☑ Explain and preform data pre-processing
- □ Explain how microbiota data are analysed
- □ Define, perform and interpret alpha-diversity
- □ Address sparsity, under-sampling and uneven sampling depth using data filtering and normalization
- □ Define, perform and interpret beta-diversity
- □ Generate and interpret multivariate data analyses
- □ Perform and interpret appropriate statistical tests
- □ Visualize and interpret microbial community composition

How microbiota data are analysed?

What is microbiome?

Part 1

- Definitions
- Microbiome importance
- Scientific questions
- Differences between metagenomics and metabarcoding

How microbiota data are generated?

Part 2

- From samples to sequences
- From sequences to data sets

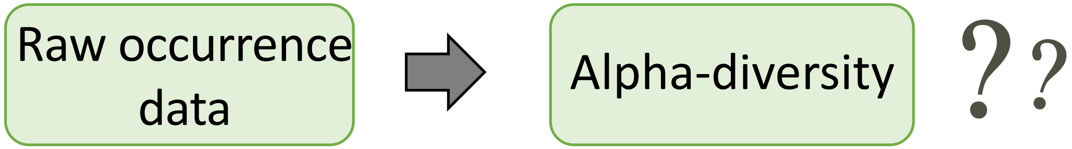
How microbiota data are analysed?

Part 3

- Alpha-diversity
- Data properties
- Data filtering
- Data normalisation
- Beta-diversity
- Microbial composition

Step 3: From microbiota data sets to data visualisation

Process overview


Raw occurrence data

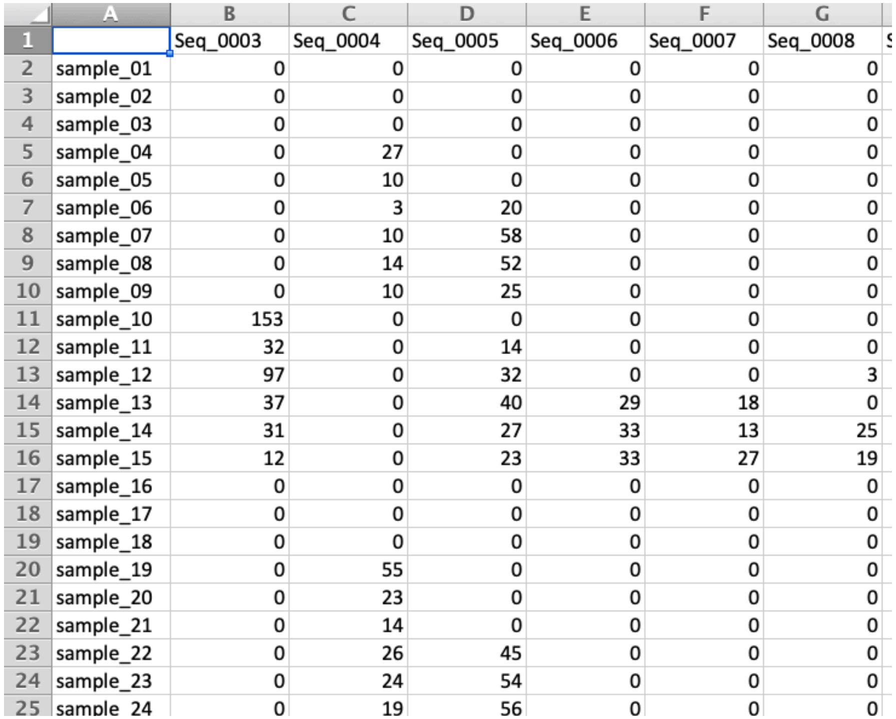
Alpha-diversity

??

~10,000 features

Occurrence data

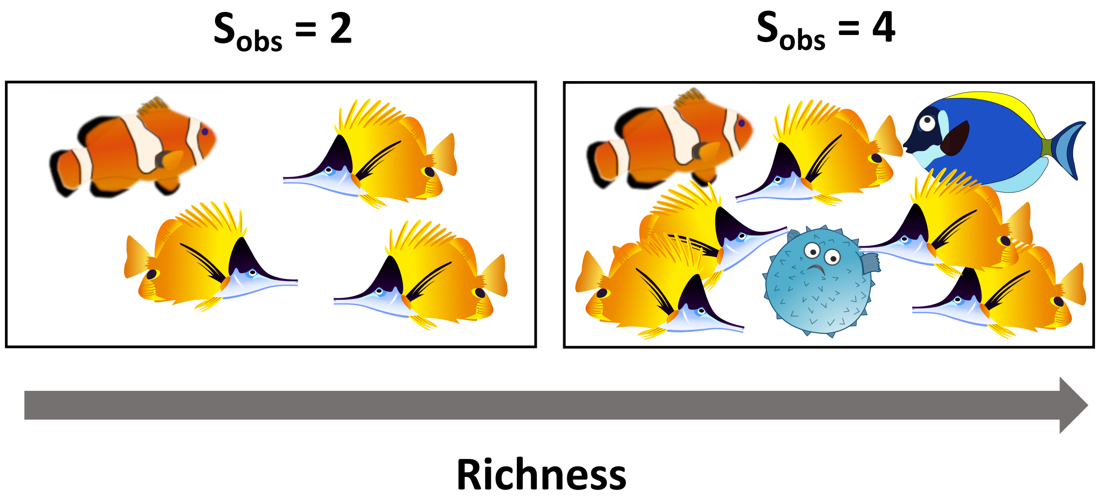
~100 samples



	A	B	C	D	E	F	G
1		Seq_0003	Seq_0004	Seq_0005	Seq_0006	Seq_0007	Seq_0008
2	sample_01	0	0	0	0	0	0
3	sample_02	0	0	0	0	0	0
4	sample_03	0	0	0	0	0	0
5	sample_04	0	27	0	0	0	0
6	sample_05	0	10	0	0	0	0
7	sample_06	0	3	20	0	0	0
8	sample_07	0	10	58	0	0	0
9	sample_08	0	14	52	0	0	0
10	sample_09	0	10	25	0	0	0
11	sample_10	153	0	0	0	0	0
12	sample_11	32	0	14	0	0	0
13	sample_12	97	0	32	0	0	3
14	sample_13	37	0	40	29	18	0
15	sample_14	31	0	27	33	13	25
16	sample_15	12	0	23	33	27	19
17	sample_16	0	0	0	0	0	0
18	sample_17	0	0	0	0	0	0
19	sample_18	0	0	0	0	0	0
20	sample_19	0	55	0	0	0	0
21	sample_20	0	23	0	0	0	0
22	sample_21	0	14	0	0	0	0
23	sample_22	0	26	45	0	0	0
24	sample_23	0	24	54	0	0	0
25	sample_24	0	19	56	0	0	0

Alpha-diversity

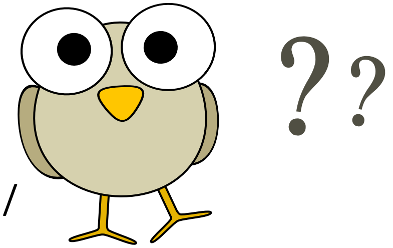
- Diversity **within one sample**/ecosystem (usually calculated at feature level)
- Alpha-diversity indices
  - ❑ Richness represents the number of species observed (S_{obs})



S_{obs} = 2

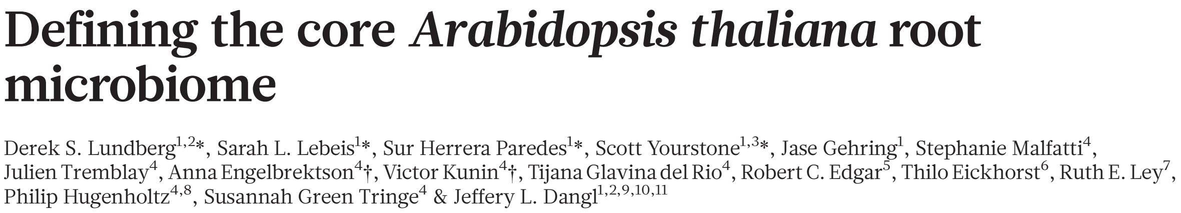
S_{obs} = 4

Richness



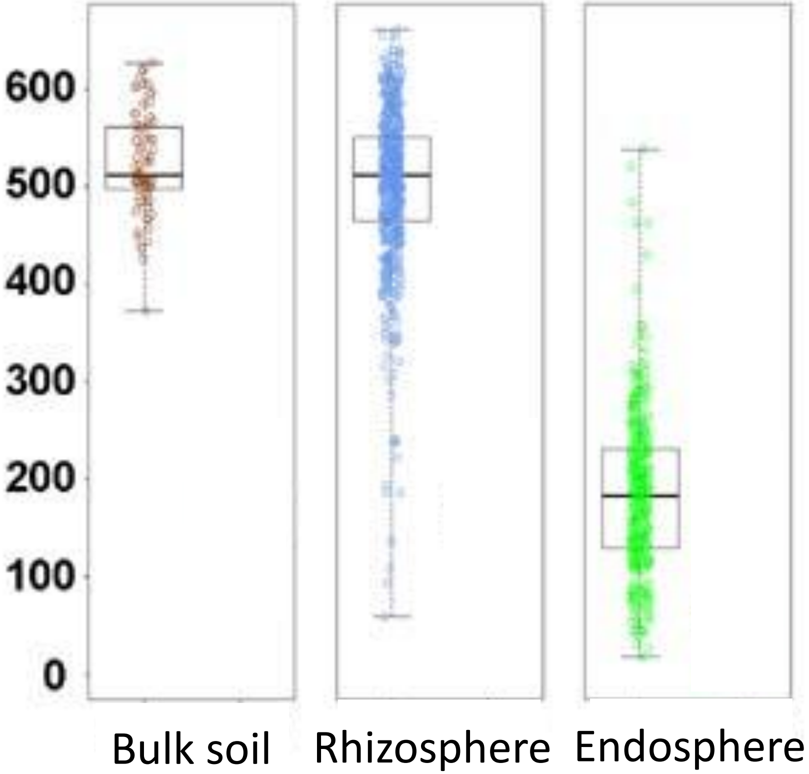
https://bigdata_microbiome.presenterswall.nl/

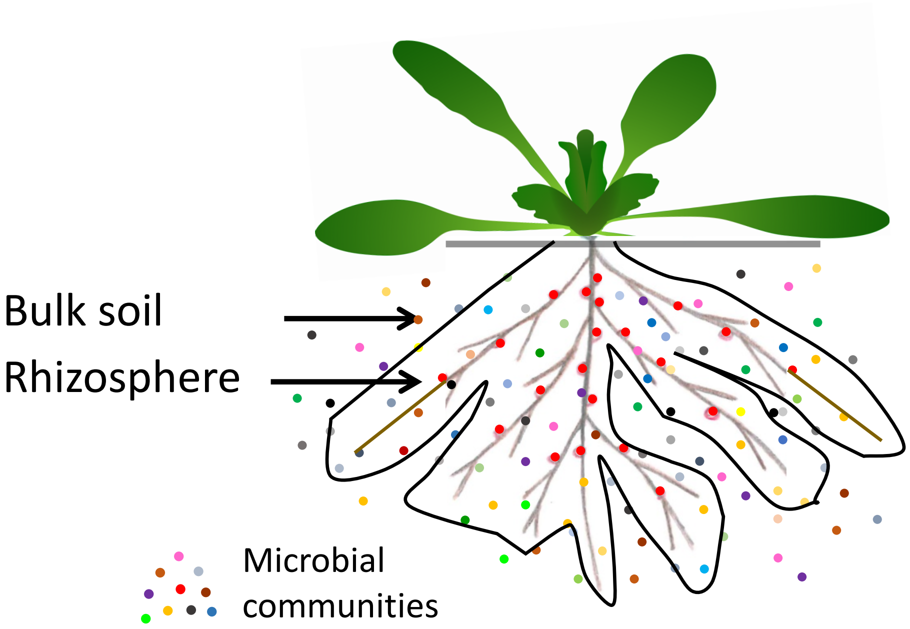
Alpha-diversity



Defining the core *Arabidopsis thaliana* root microbiome

Derek S. Lundberg[1,2]*, Sarah L. Lebeis[1]*, Sur Herrera Paredes[1]*, Scott Yourstone[1,3]*, Jase Gehring[1], Stephanie Malfatti[4], Julien Tremblay[4], Anna Engelbrektson[4]†, Victor Kunin[4]†, Tijana Glavina del Rio[4], Robert C. Edgar[5], Thilo Eickhorst[6], Ruth E. Ley[7], Philip Hugenholtz[4,8], Susannah Green Tringe[4] & Jeffery L. Dangl[1,2,9,10,11]





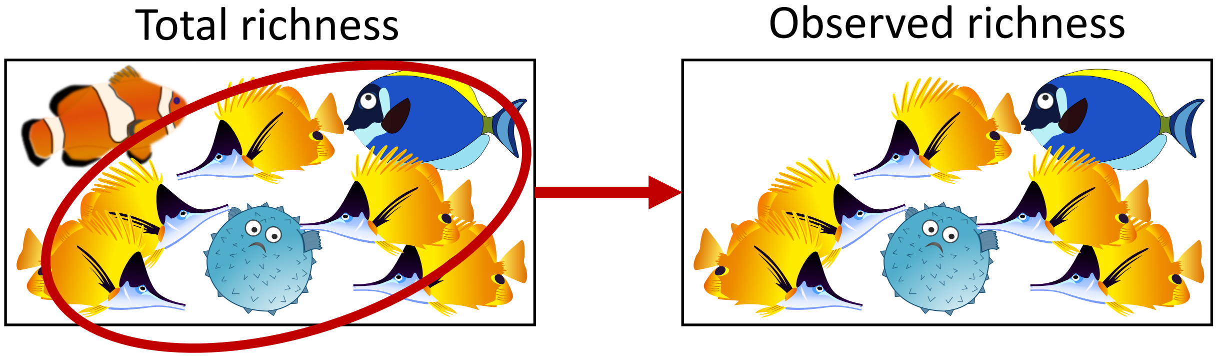
Alpha-diversity

- Diversity **within one sample**/ecosystem (usually calculated at feature level)
- Alpha-diversity indices
  - ❑ Richness represents the number of species observed (S_{obs})
  - ❑ Chao1 estimates total richness (S_1)

$$S_1 = S_{obs} + \frac{F_1^2}{2F_2}$$

S_{obs} Number of species
F_1 Number of singletons
F_2 Number of doubletons

REMARK: **Chao1 can only be calculated on raw data**

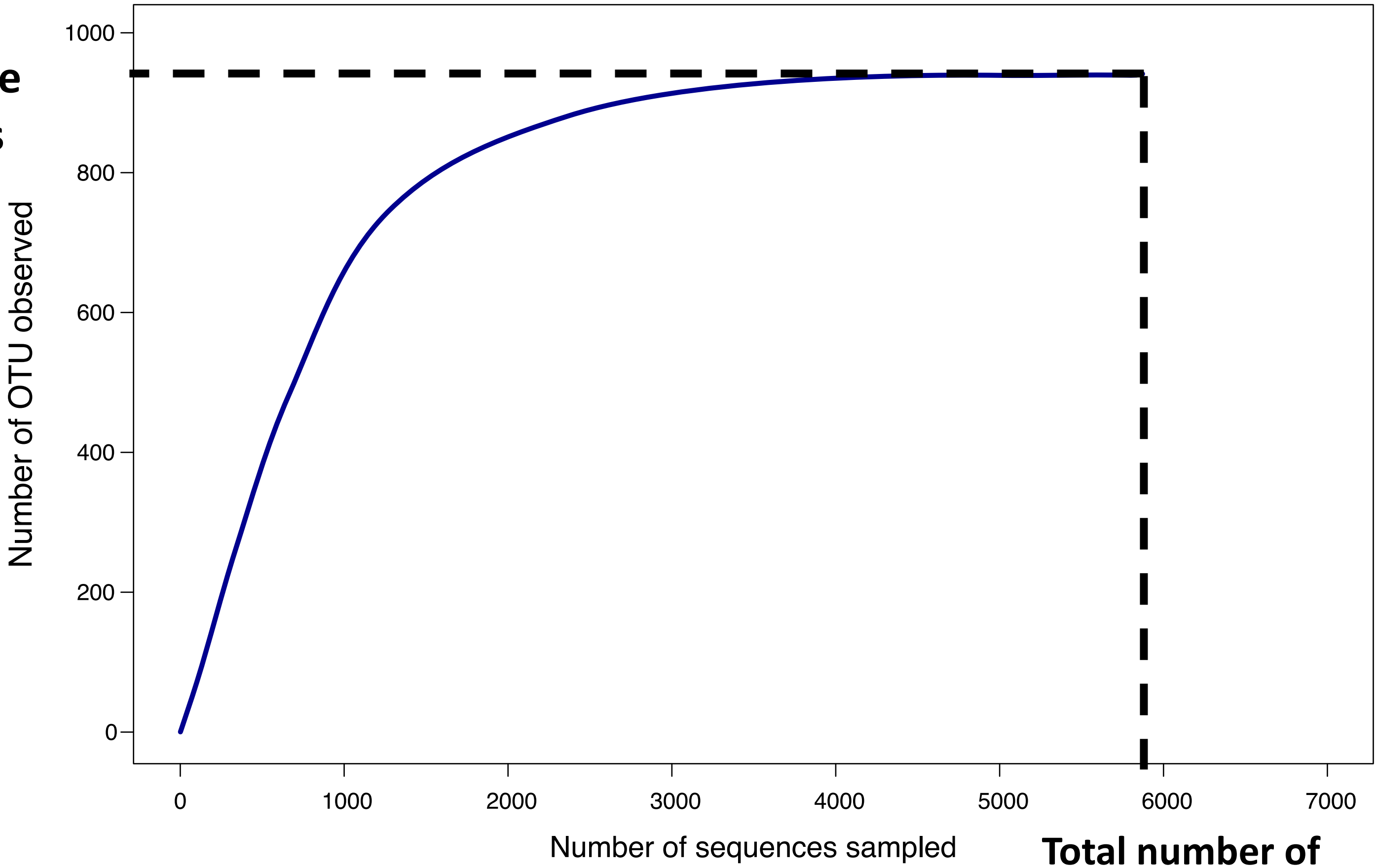


REMARK: **Difference between observed richness and Chao1 give you information about the sequencing depth (enough if Richness = Chao1; not enough if Richness << Chao1)**

Sequencing depth

- Rarefaction curve

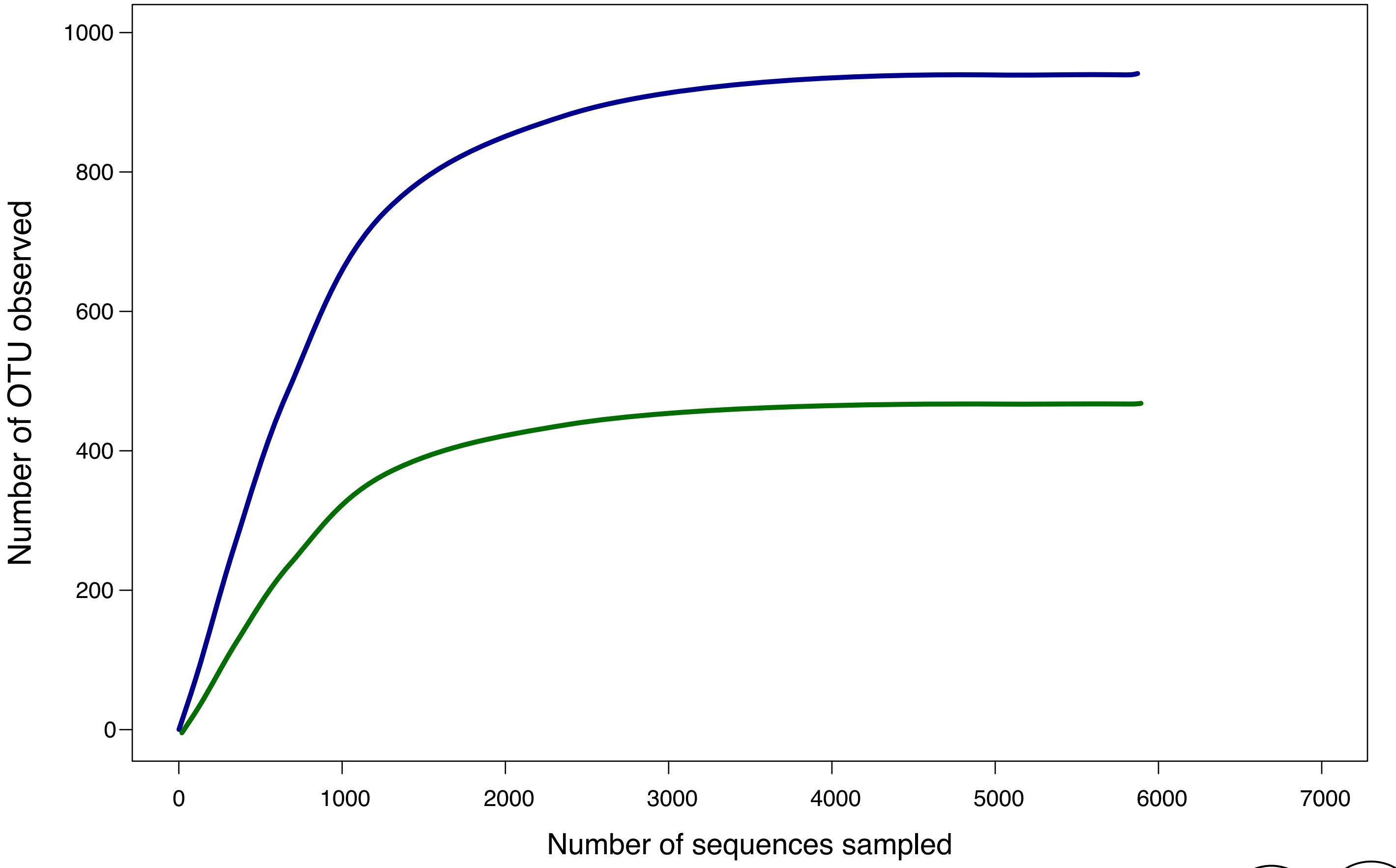
Richness (S_{obs}) Number of feature observed for this sample



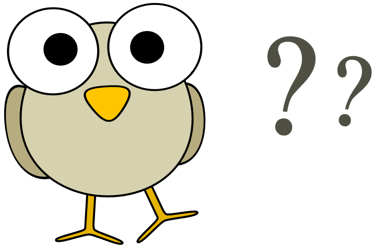
Total number of sequences/reads for this sample

Sequencing depth

- Rarefaction curve

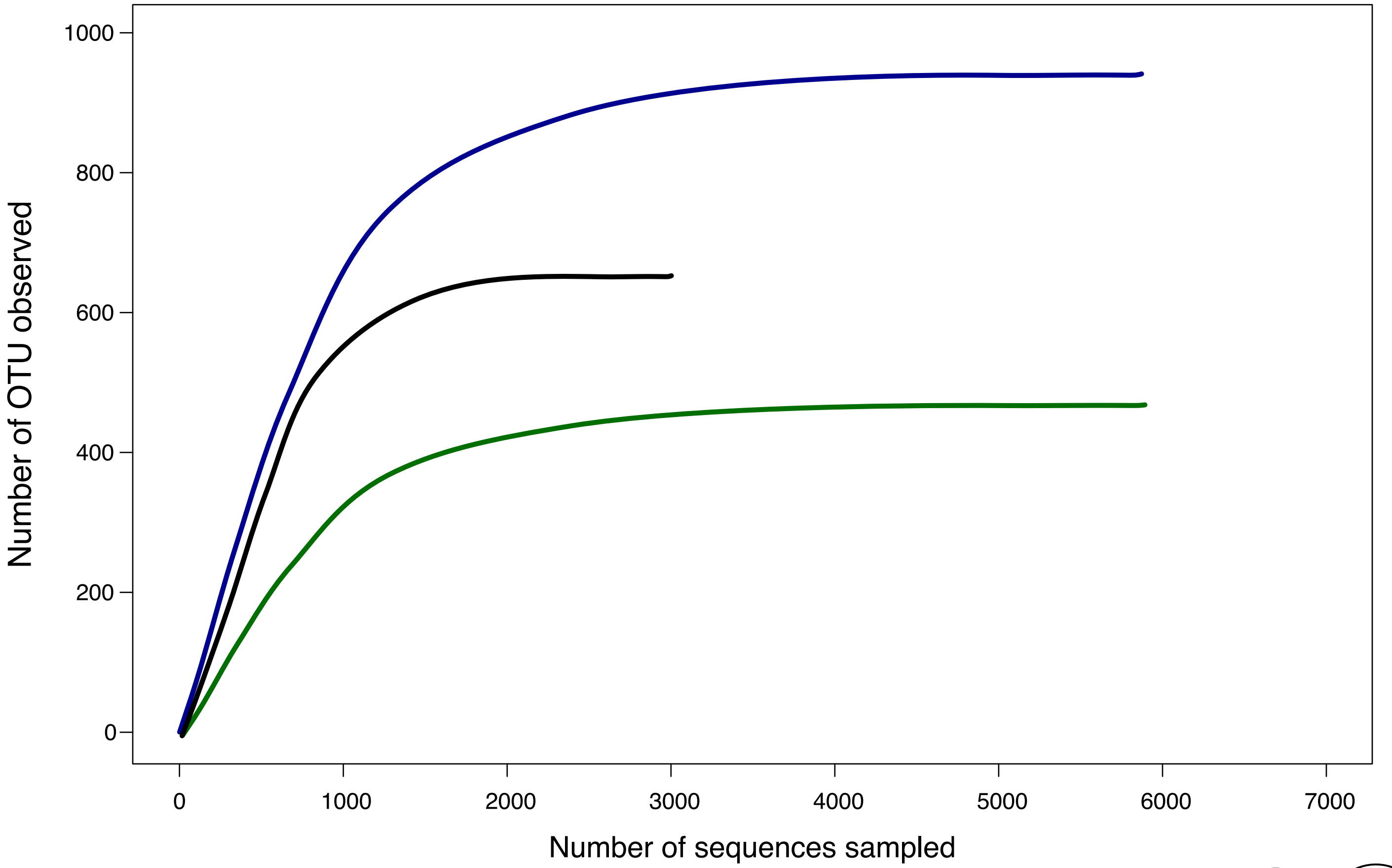


https://bigdata_microbiome.presenterswall.nl/

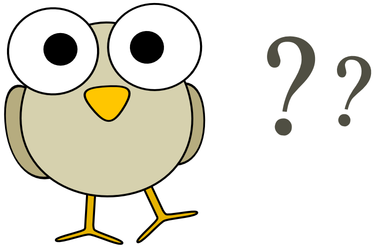


Sequencing depth

- Rarefaction curve

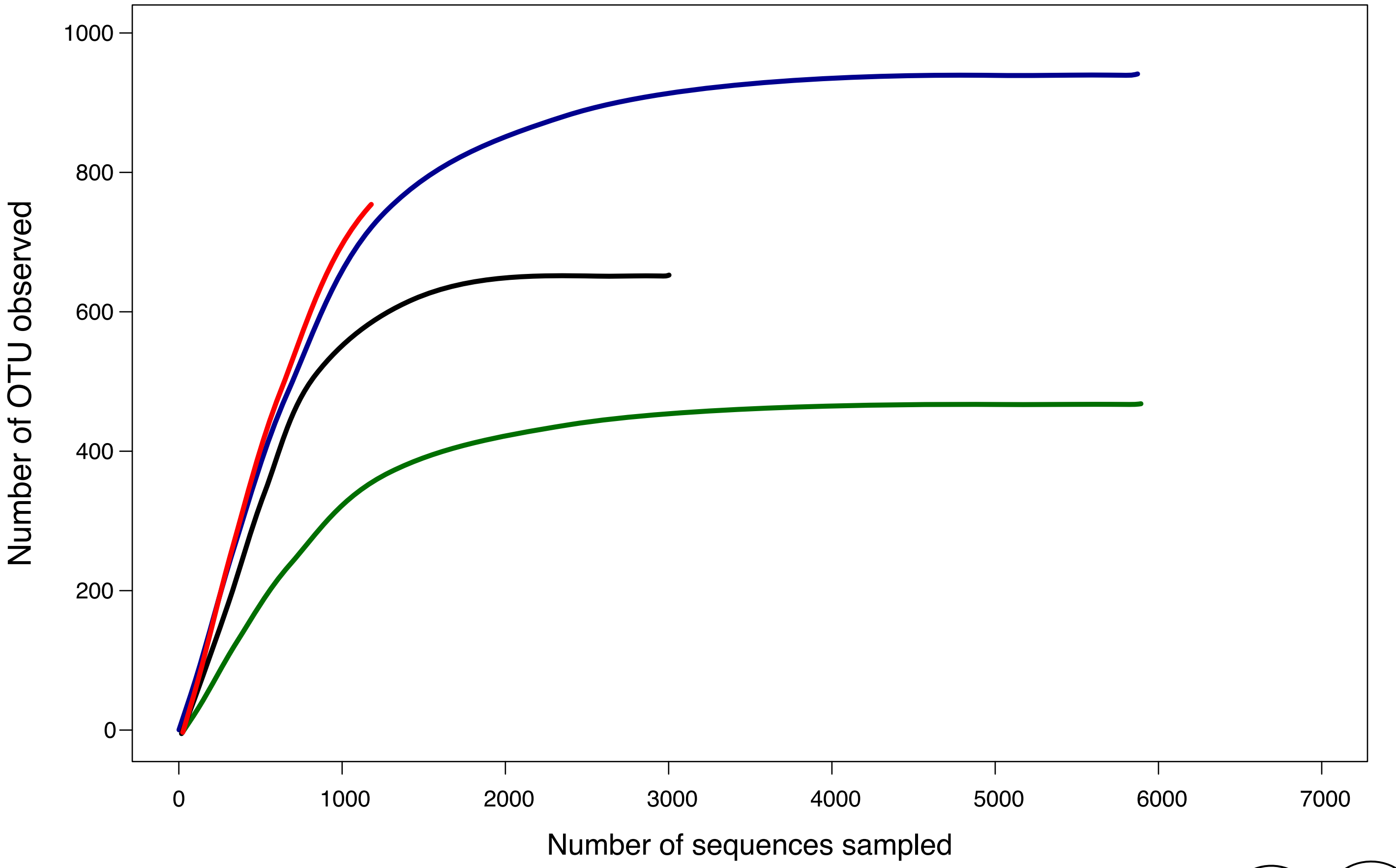


https://bigdata_microbiome.presenterswall.nl/

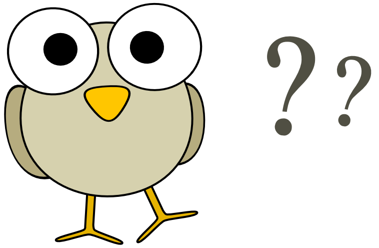


Sequencing depth

- Rarefaction curve



https://bigdata_microbiome.presenterswall.nl/

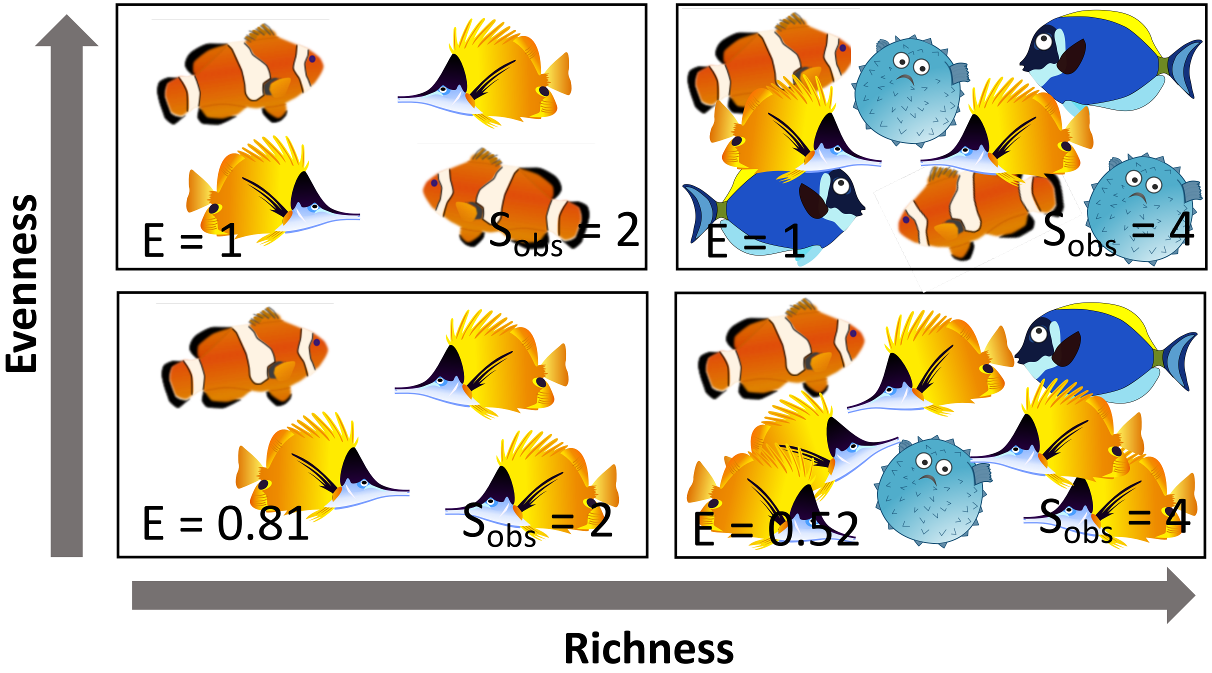


Alpha-diversity

- Diversity **within one sample**/ecosystem (usually calculated at feature level)
- Alpha-diversity indices
  - ❑ Richness represents the number of species observed (S_{obs})
  - ❑ Chao1 estimates total richness (S_1)
  - ❑ Pielou's evenness provide information about equity in species abundance

$$\mathbf{E = -\sum_{i=1}^{Sobs} p_i \ln p_i \;\; / \;\; \ln(S_{obs})}$$

p_i proportion of individuals belonging to the i^{th} species



Alpha-diversity

- Diversity **within one sample**/ecosystem (usually calculated at feature level)
- Alpha-diversity indices
  - ❑ Richness represents the number of species observed (S_{obs})
  - ❑ Chao1 estimates total richness (S_1)
  - ❑ Pielou’s evenness provide information about equity in species abundance
  - ❑ Shannon provides information about both richness and evenness (H’)

$$H' = -\sum_{i=1}^{Sobs} p_i \ln p_i$$

p_i proportion of individuals belonging to the i^{th} species

Alpha-diversity

- Diversity **within one sample**/ecosystem (usually calculated at feature level)
- Alpha-diversity indices
  - ❑ Richness represents the number of species observed (S_{obs})
  - ❑ Chao1 estimates total richness (S_1)
  - ❑ Pielou's evenness provide information about equity in species abundance
  - ❑ Shannon provides information about both richness and evenness (H')
- Statistical tests
  - ❑ Normal distribution: t-test or ANOVA
  - ❑ No normal distribution: Mann Whitney or Kruskal Wallis

Alpha-diversity

- Diversity within one sample/ecosystem
- Should be calculated on raw data
- Observed richness = number of features observed
- Chao1 = total richness
- Evenness = equity in feature abundance
- Shannon <= richness and evenness
- Sequencing depth => did I catch all the diversity?

Practice time: alpha-diversity

In the tutorial, look at:

- Home page
- 1. Introduction
- 4. Alpha-diversity

Tutorial link:

https://scienceparkstudygroup.github.io/microbiome-lesson/index.html

Learning objectives

☑ Define microbiome and state microbiome importance

☑ Identify differences between metabarcoding and metagenomics

☑ Explain how microbiota data are generated (including bias)

☑ Explain and preform data pre-processing

□ Explain how microbiota data are analysed

☑ Define, perform and interpret alpha-diversity

□ Address sparsity, under-sampling and uneven sampling depth using data filtering and normalization

□ Define, perform and interpret beta-diversity

□ Generate and interpret multivariate data analyses

□ Perform and interpret appropriate statistical tests

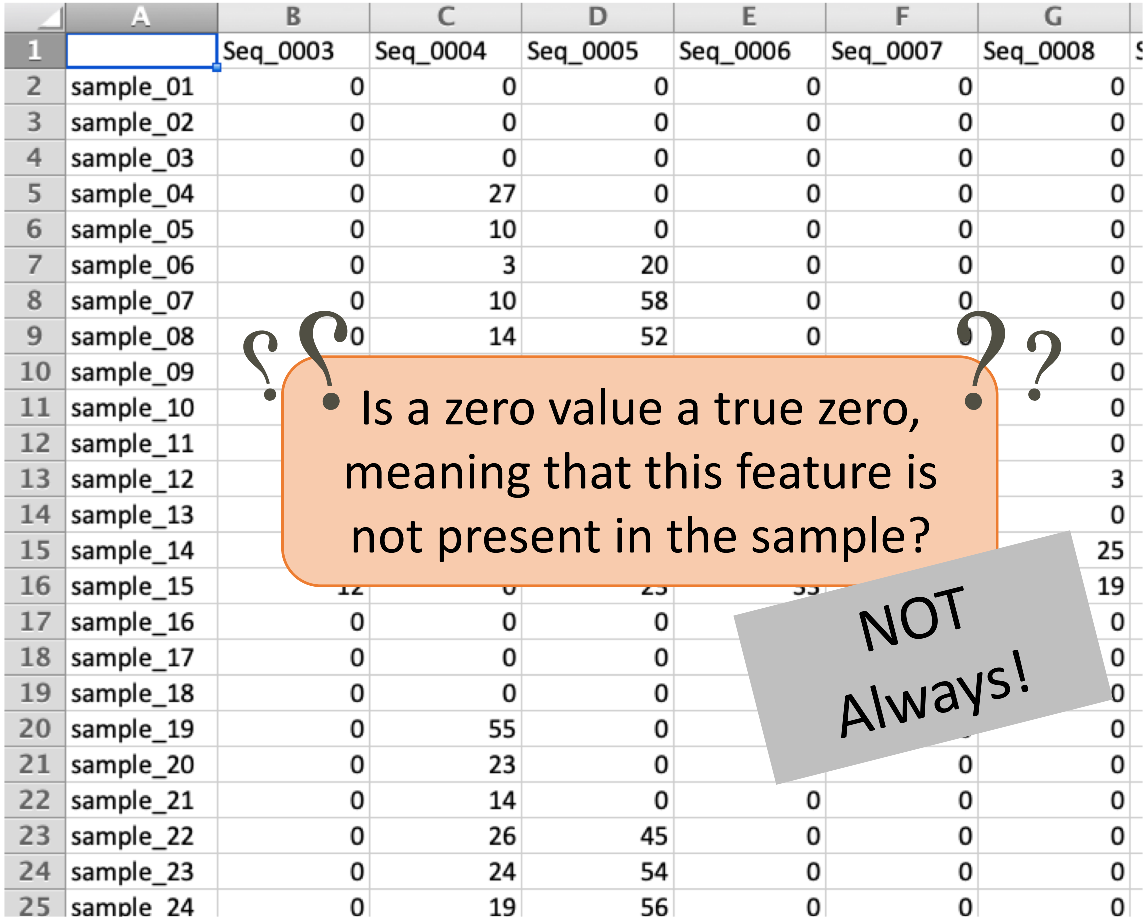
□ Visualize and interpret microbial community composition

Microbiota data properties

Occurrence table

~10,000 features

~100 samples



	Seq_0003	Seq_0004	Seq_0005	Seq_0006	Seq_0007	Seq_0008
sample_01	0	0	0	0	0	0
sample_02	0	0	0	0	0	0
sample_03	0	0	0	0	0	0
sample_04	0	27	0	0	0	0
sample_05	0	10	0	0	0	0
sample_06	0	3	20	0	0	0
sample_07	0	10	58	0	0	0
sample_08	0	14	52	0	0	0
sample_09						0
sample_10						0
sample_11						0
sample_12						3
sample_13						0
sample_14						25
sample_15	12	0	23	33		19
sample_16	0	0	0			0
sample_17	0	0	0			0
sample_18	0	0	0			0
sample_19	0	55	0			0
sample_20	0	23	0		0	0
sample_21	0	14	0	0	0	0
sample_22	0	26	45	0	0	0
sample_23	0	24	54	0	0	0
sample_24	0	19	56	0	0	0

Is a zero value a true zero, meaning that this feature is not present in the sample?

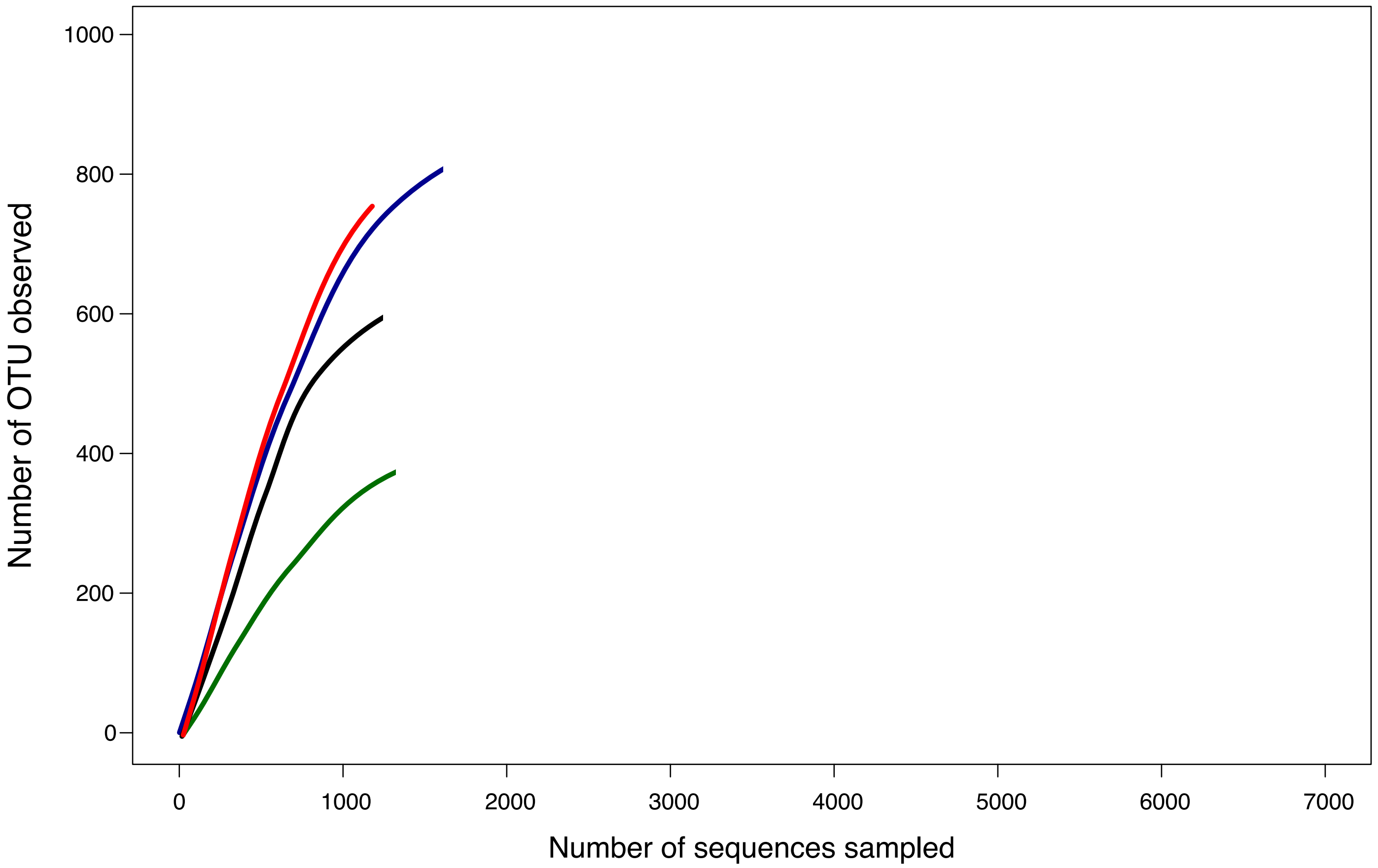
NOT Always!

- n << p
- Sparse data (~80% of 0)

Filter the data in order to decrease low quality or uninformative features

Sequencing depth

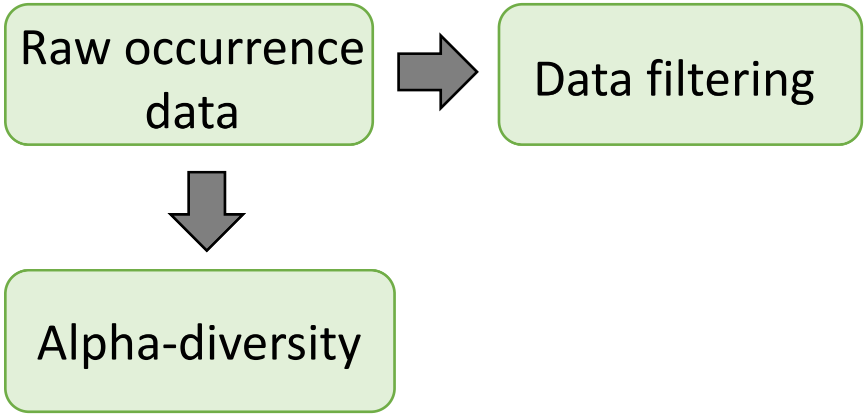
- Rarefaction curve



REMARK: **If the sequencing depth is not enough, it will be difficult to compare difference between samples for low counts.**
Therefore, it will be better to remove features that have only low counts.

Step 3: From microbiota data sets to data visualisation

Process overview



Raw occurrence data

Data filtering

Alpha-diversity

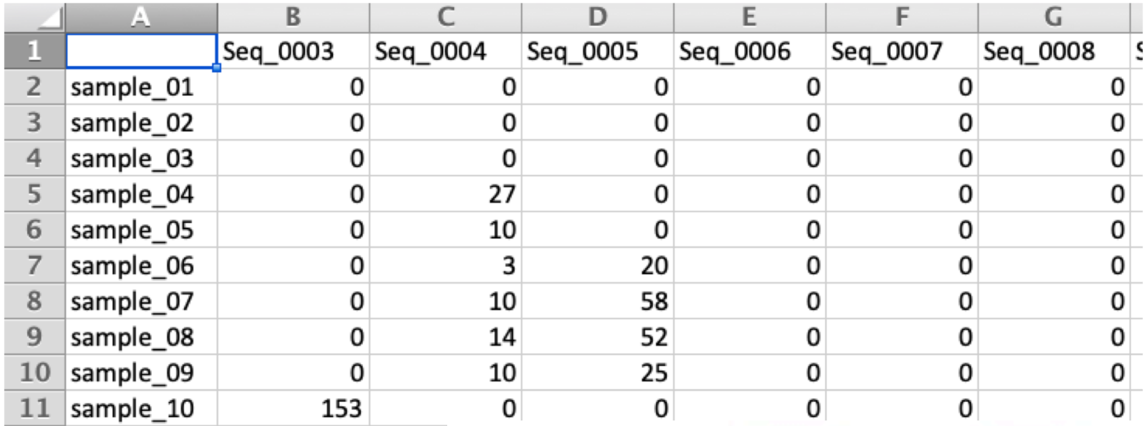
Challenge:
Remove uninformative & low quality reads
Trade-off between quantity and quality

Microbiota data properties

Occurrence table

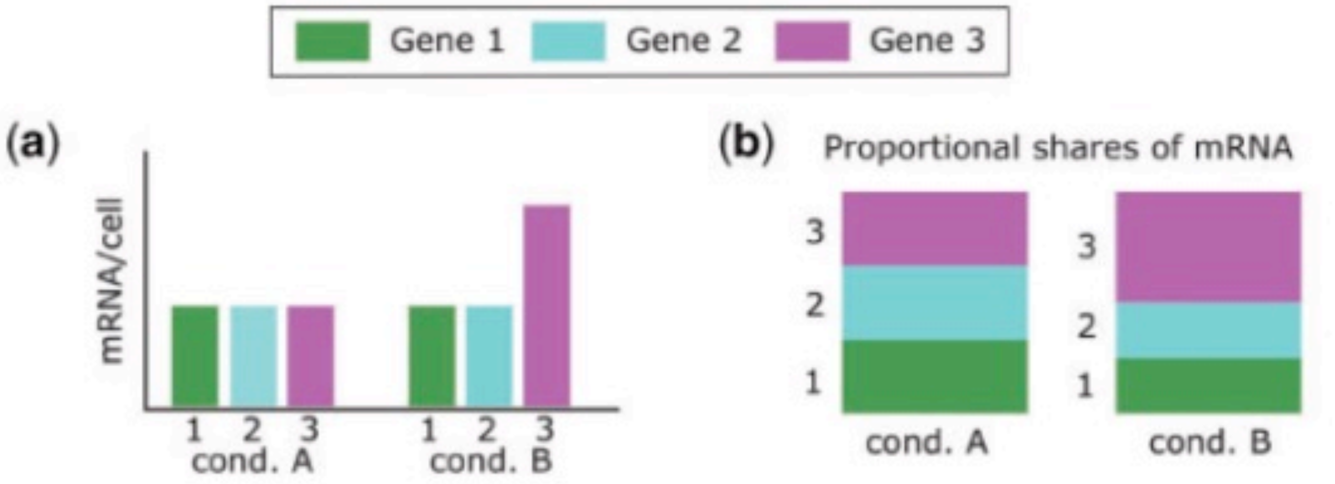
~10,000 features

~100 samples



	A	B	C	D	E	F	G
1		Seq_0003	Seq_0004	Seq_0005	Seq_0006	Seq_0007	Seq_0008
2	sample_01	0	0	0	0	0	0
3	sample_02	0	0	0	0	0	0
4	sample_03	0	0	0	0	0	0
5	sample_04	0	27	0	0	0	0
6	sample_05	0	10	0	0	0	0
7	sample_06	0	3	20	0	0	0
8	sample_07	0	10	58	0	0	0
9	sample_08	0	14	52	0	0	0
10	sample_09	0	10	25	0	0	0
11	sample_10	153	0	0	0	0	0
12	sample_11	32					
13	sample_12	97					
14	sample_13	37					
15	sample_14	31					
16	sample_15	12					
17	sample_16	0					
18	sample_17	0					
19	sample_18	0					
20	sample_19	0					
21	sample_20	0					
22	sample_21	0					
23	sample_22	0					
24	sample_23	0					
25	sample_24	0					

- n << p
- Sparse data (~80% of 0)
- Compositional data



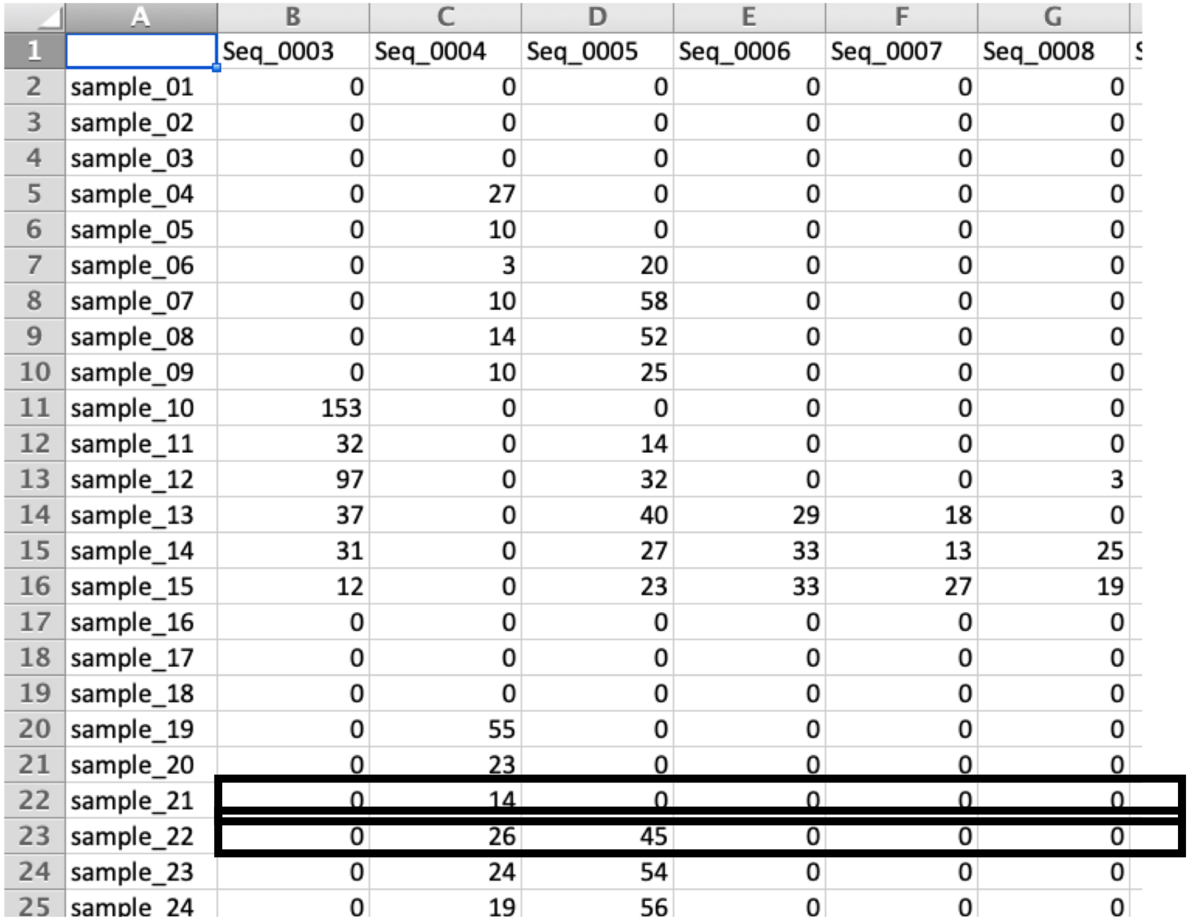
REMARK: **We describe relative abundances**

Microbiota data properties

Occurrence table

~10,000 features

~100 samples

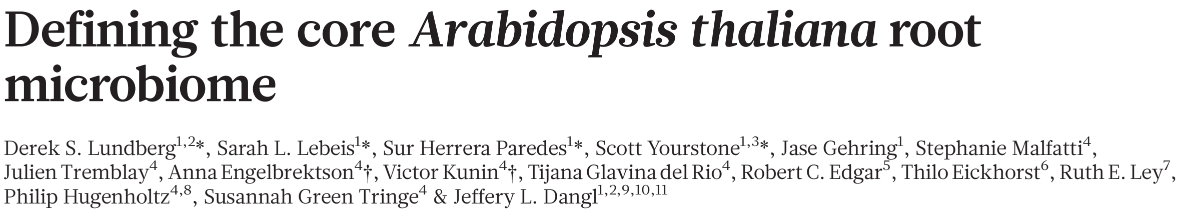


	A	B	C	D	E	F	G
1		Seq_0003	Seq_0004	Seq_0005	Seq_0006	Seq_0007	Seq_0008
2	sample_01	0	0	0	0	0	0
3	sample_02	0	0	0	0	0	0
4	sample_03	0	0	0	0	0	0
5	sample_04	0	27	0	0	0	0
6	sample_05	0	10	0	0	0	0
7	sample_06	0	3	20	0	0	0
8	sample_07	0	10	58	0	0	0
9	sample_08	0	14	52	0	0	0
10	sample_09	0	10	25	0	0	0
11	sample_10	153	0	0	0	0	0
12	sample_11	32	0	14	0	0	0
13	sample_12	97	0	32	0	0	3
14	sample_13	37	0	40	29	18	0
15	sample_14	31	0	27	33	13	25
16	sample_15	12	0	23	33	27	19
17	sample_16	0	0	0	0	0	0
18	sample_17	0	0	0	0	0	0
19	sample_18	0	0	0	0	0	0
20	sample_19	0	55	0	0	0	0
21	sample_20	0	23	0	0	0	0
22	sample_21	0	14	0	0	0	0
23	sample_22	0	26	45	0	0	0
24	sample_23	0	24	54	0	0	0
25	sample_24	0	19	56	0	0	0

Sum = 14
Sum = 71

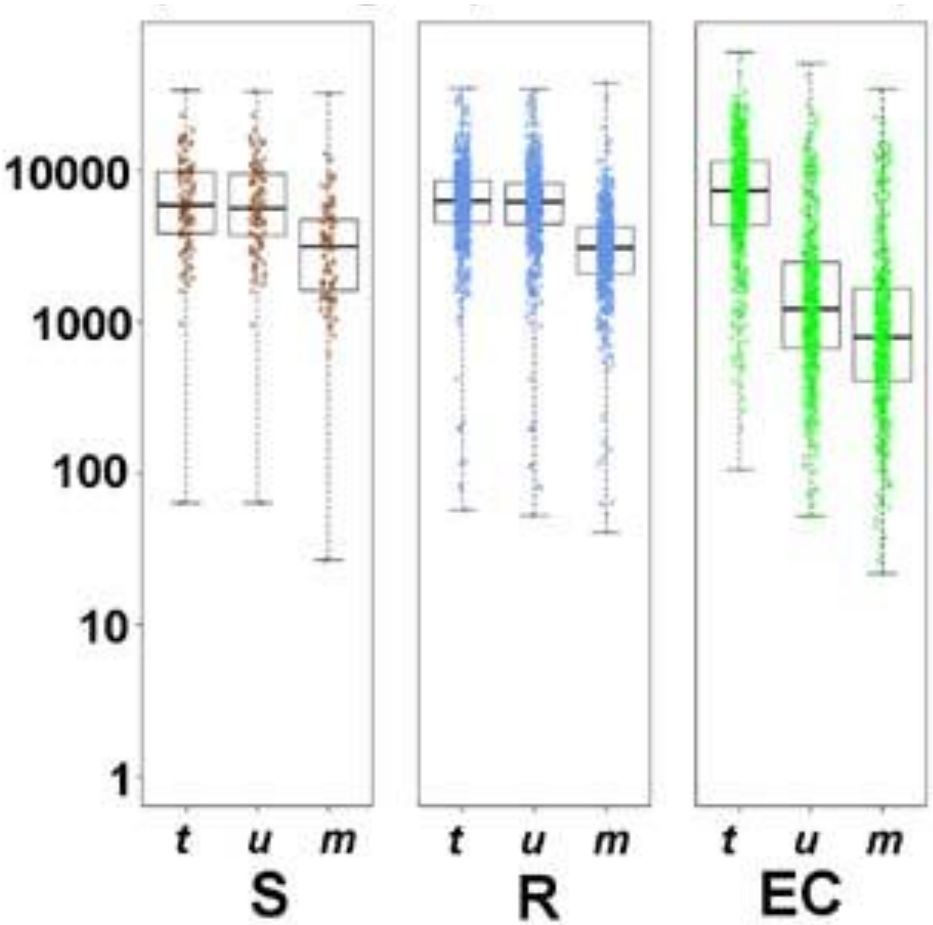
- n << p
- Sparse data (~80% of 0)
- Compositional data
- Different library sizes (total number of reads/ sequences per sample)

Microbiota data properties: library size per sample



Defining the core *Arabidopsis thaliana* root microbiome

Derek S. Lundberg[1,2]*, Sarah L. Lebeis[1]*, Sur Herrera Paredes[1]*, Scott Yourstone[1,3]*, Jase Gehring[1], Stephanie Malfatti[4], Julien Tremblay[4], Anna Engelbrektson[4]†, Victor Kunin[4]†, Tijana Glavina del Rio[4], Robert C. Edgar[5], Thilo Eickhorst[6], Ruth E. Ley[7], Philip Hugenholtz[4,8], Susannah Green Tringe[4] & Jeffery L. Dangl[1,2,9,10,11]



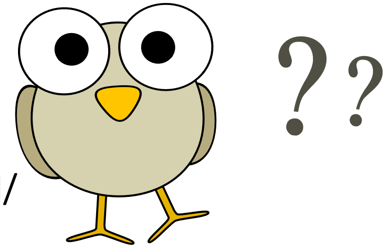
Lundberg et al. 2012

Microbiota data properties: library size per sample

- Library size is the **total number of reads per sample**

	seq_1	seq_2	seq_3	(…)	seq_p	total_reads
sample_1	500	80	20		5	10,000
sample_2	500	80	20		5	1,000
sample_3	50	8	2		0	1,000
(…)						
sample_n	2000	0	2		0	10,000

https://bigdata_microbiome.presenterswall.nl/



Microbiota data properties

- ➢ Microbiota data usually sparse => need filtering especially when sequencing depth was not enough
- ➢ Uneven library size => need normalisation for sample comparison

Practice time: microbiota data properties

In the tutorial, look at:

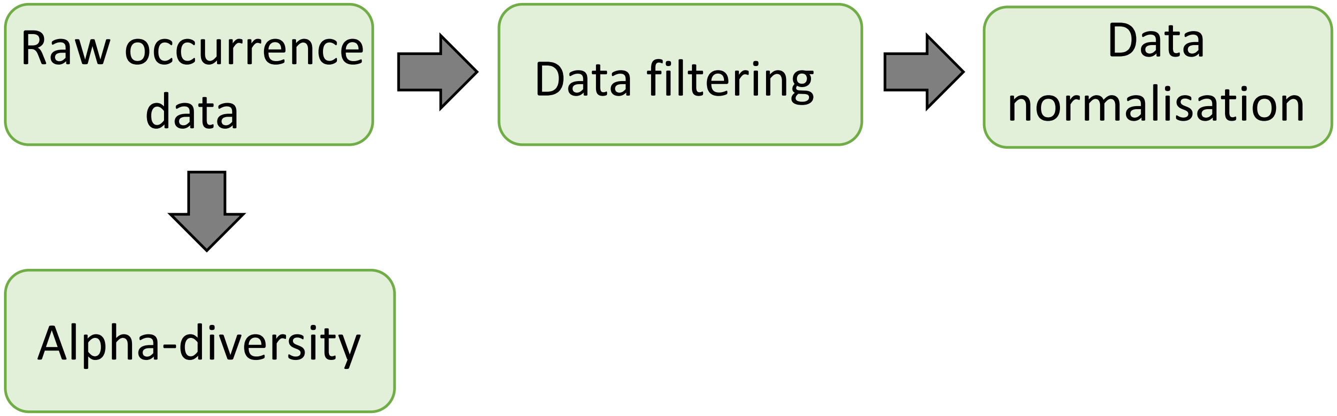
- 3. Data exploration and properties

Tutorial link:

https://scienceparkstudygroup.github.io/microbiome-lesson/03-data-exploration-and-properties/index.html

Step 3: From microbiota data sets to data visualisation

Process overview



Raw occurrence data

Data filtering

Data normalisation

Alpha-diversity

Microbiota data normalisation

- Different normalisation methods available (depend on your downstream analysis)
  - **Total Sum Normalisation**: dividing the reads for each OTU in a sample by the total number of reads in that sample and multiplying by 100

	seq_1	seq_2	seq_3	(…)	seq_p	total_reads
sample_1	500	80	20		5	10,000
sample_2	500	80	20		5	1,000
sample_3	50	8	2		0	1,000

⇩

	seq_1	seq_2	seq_3	(…)	seq_p	total_reads
sample_1						
sample_2						
sample_3						

Microbiota data normalisation

- Different normalisation methods available
  - ❑ **Total Sum Normalisation**: dividing the reads for each OTU in a sample by the total number of reads in that sample and multiplying by 100

	seq_1	seq_2	seq_3	(…)	seq_p	total_reads
sample_1	500	80	20		5	10,000
sample_2	500	80	20		5	1,000
sample_3	50	8	2		0	1,000

⇩

	seq_1	seq_2	seq_3	(…)	seq_p	total_reads
sample_1						100
sample_2						100
sample_3						100

Microbiota data normalisation

- Different normalisation methods available
  - **Total Sum Normalisation**: dividing the reads for each OTU in a sample by the total number of reads in that sample and multiplying by 100

	seq_1	seq_2	seq_3	(…)	seq_p	total_reads
sample_1	500	80	20		5	10,000
sample_2	500	80	20		5	1,000
sample_3	50	8	2		0	1,000

⇩

	seq_1	seq_2	seq_3	(…)	seq_p	total_reads
sample_1	0.05	0.008	0.002		0.0005	100
sample_2	0.5	0.08	0.02		0.005	100
sample_3	0.05	0.008	0.002		0	100

Microbiota data normalisation

- Different normalisation methods available
  - **Total Sum Normalisation**: dividing the reads for each OTU in a sample by the total number of reads in that sample and multiplying by 100
  - **Rarefy**: randomly subsampling each sample to the lowest read depth of any sample

	seq_1	seq_2	seq_3	(…)	seq_p	total_reads
sample_1	500	80	20		5	10,000
sample_2	500	80	20		5	1,000
sample_3	50	8	2		0	1,000

	seq_1	seq_2	seq_3	(…)	seq_p	total_reads
sample_1						1,000
sample_2						1,000
sample_3						1,000

Microbiota data normalisation

- Different normalisation methods available
  - ❑ **Total Sum Normalisation**: dividing the reads for each OTU in a sample by the total number of reads in that sample and multiplying by 100
  - ❑ **Rarefy**: randomly subsampling each sample to the lowest read depth of any sample

	seq_1	seq_2	seq_3	(…)	seq_p	total_reads
sample_1	500	80	20		5	10,000
sample_2	500	80	20		5	1,000
sample_3	50	8	2		0	1,000

⇩

	seq_1	seq_2	seq_3	(…)	seq_p	total_reads
sample_1						1,000
sample_2	500	80	20		5	1,000
sample_3	50	8	2		0	1,000

Microbiota data normalisation

- Different normalisation methods available
  - ❑ **Total Sum Normalisation**: dividing the reads for each OTU in a sample by the total number of reads in that sample and multiplying by 100
  - ❑ **Rarefy**: randomly subsampling each sample to the lowest read depth of any sample

	seq_1	seq_2	seq_3	(…)	seq_p	total_reads
sample_1	500	80	20		5	10,000
sample_2	500	80	20		5	1,000
sample_3	50	8	2		0	1,000

⬇

	seq_1	seq_2	seq_3	(…)	seq_p	total_reads
sample_1	52	8	1		0	1,000
sample_2	500	80	20		5	1,000
sample_3	50	8	2		0	1,000

Microbiota data normalisation

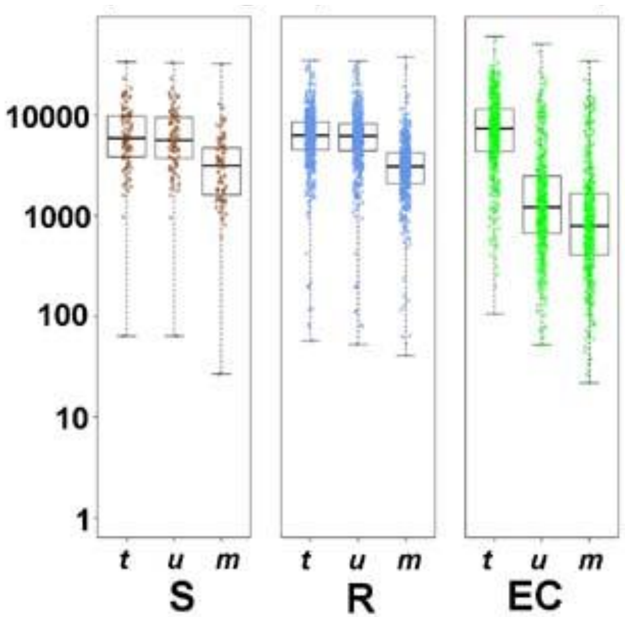
- Different normalisation methods available
  - **Total Sum Normalisation**: dividing the reads for each OTU in a sample by the total number of reads in that sample and multiplying by 100
  - **Rarefy**: randomly subsampling each sample to the lowest read depth of any sample

REMARK: **When the sequencing depth is not enough and you have big differences in library sizes (~x10), it is better to rarefy your data than calculate percentage**

Defining the core *Arabidopsis thaliana* root microbiome

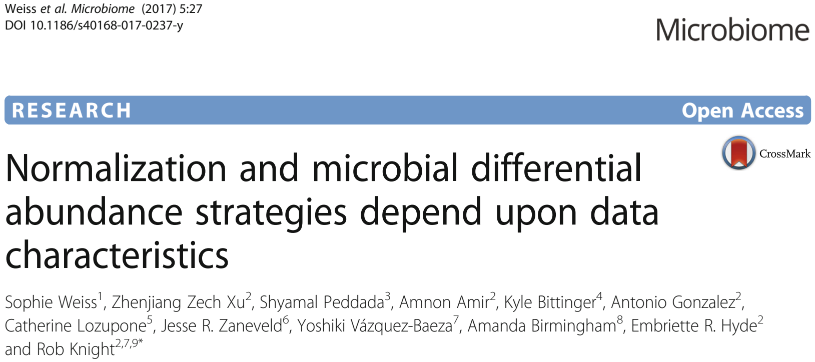
Derek S. Lundberg[1,2]*, Sarah L. Lebeis[1]*, Sur Herrera Paredes[1]*, Scott Yourstone[1,3]*, Jase Gehring[1], Stephanie Malfatti[4], Julien Tremblay[4], Anna Engelbrektson[4]†, Victor Kunin[4]†, Tijana Glavina del Rio[4], Robert C. Edgar[5], Thilo Eickhorst[6], Ruth E. Ley[7], Philip Hugenholtz[4,8], Susannah Green Tringe[4] & Jeffery L. Dangl[1,2,9,10,11]

- ➢ Rarefied at 1000 reads per sample



Microbiota data normalisation

- Different normalisation methods available
  - ❑ **Total Sum Normalisation**: dividing the reads for each OTU in a sample by the total number of reads in that sample and multiplying by 100
  - ❑ **Rarefy**: randomly subsampling each sample to the lowest read depth of any sample
  - ❑ **DESeq-VS**: a variance stabilizing transformation (used for RNA-seq analysis)
  - ❑ **edgeR-TMM**: a trimmed mean of M-values normalisation



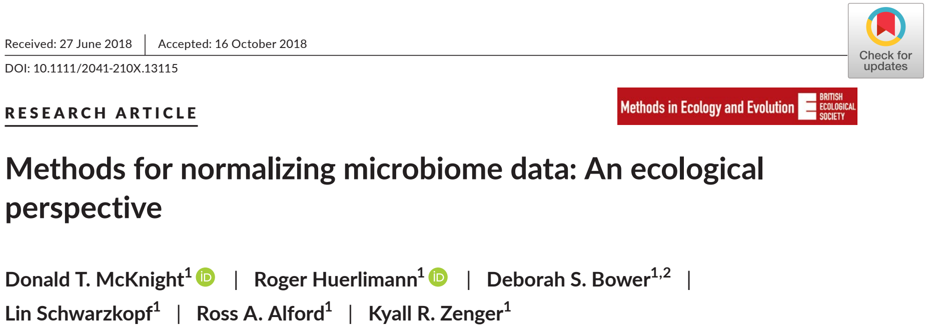
Weiss *et al. Microbiome* (2017) 5:27
DOI 10.1186/s40168-017-0237-y

Microbiome

RESEARCH Open Access

Normalization and microbial differential abundance strategies depend upon data characteristics

Sophie Weiss[1], Zhenjiang Zech Xu[2], Shyamal Peddada[3], Amnon Amir[2], Kyle Bittinger[4], Antonio Gonzalez[2], Catherine Lozupone[5], Jesse R. Zaneveld[6], Yoshiki Vázquez-Baeza[7], Amanda Birmingham[8], Embriette R. Hyde[2] and Rob Knight[2,7,9*]



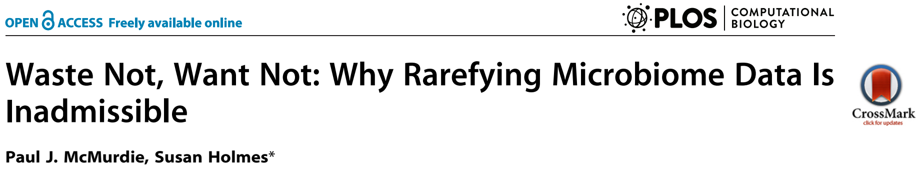
Received: 27 June 2018 | Accepted: 16 October 2018
DOI: 10.1111/2041-210X.13115

RESEARCH ARTICLE

Methods in Ecology and Evolution

Methods for normalizing microbiome data: An ecological perspective

Donald T. McKnight[1] | Roger Huerlimann[1] | Deborah S. Bower[1,2] | Lin Schwarzkopf[1] | Ross A. Alford[1] | Kyall R. Zenger[1]



OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes*

Microbiota data normalisation

- Different normalisation methods for sample comparison
  - For community level analysis (TSN or rarefying)
  - For differential abundance testing (DESeq-VS or edgeR-TMM)
- Better to use rarefying when sequencing depth is not enough and there are big differences in library sizes

Practice time: data filtering and normalisation

In the tutorial, look at:

- 5. Data filtering and normalisation

Tutorial link:

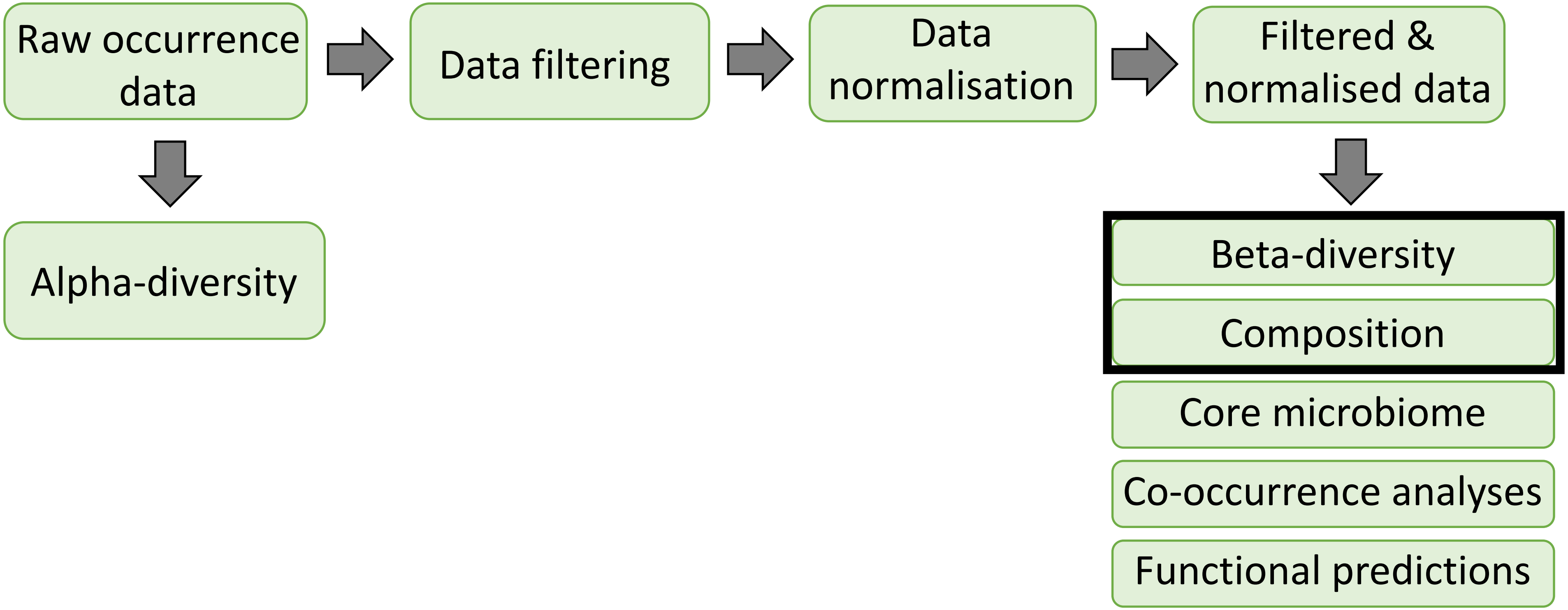
https://scienceparkstudygroup.github.io/microbiome-lesson/05-data-filtering-and-normalisation/index.html

Learning objectives

- ☑ Define microbiome and state microbiome importance
- ☑ Identify differences between metabarcoding and metagenomics
- ☑ Explain how microbiota data are generated (including bias)
- ☑ Explain and preform data pre-processing
- □ Explain how microbiota data are analysed
- ☑ Define, perform and interpret alpha-diversity
- ☑ Address sparsity, under-sampling and uneven sampling depth using data filtering and normalization
- □ Define, perform and interpret beta-diversity
- □ Generate and interpret multivariate data analyses
- □ Perform and interpret appropriate statistical tests
- □ Visualize and interpret microbial community composition

Step 3: From microbiota data sets to data visualisation

Process overview

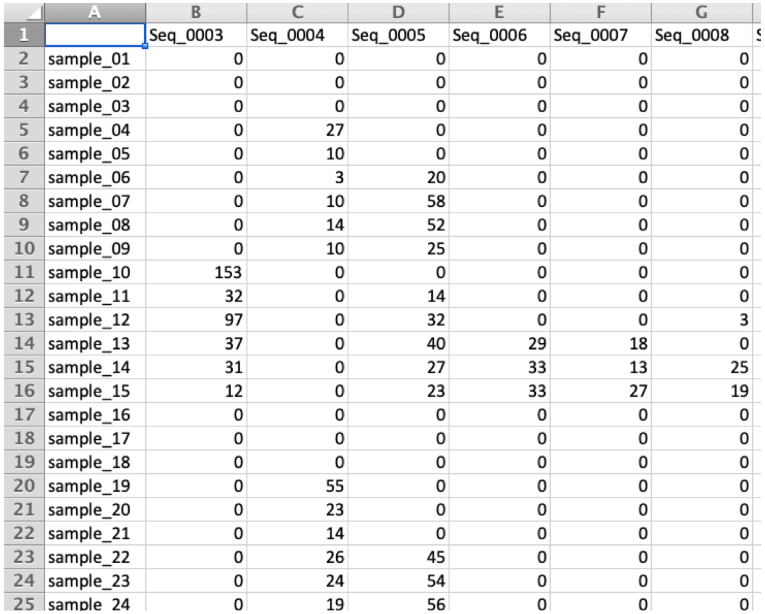


Beta-diversity

- Diversity **between two samples/ecosystems** (feature level)
- Calculate distances between samples

~10,000 features

~100 samples

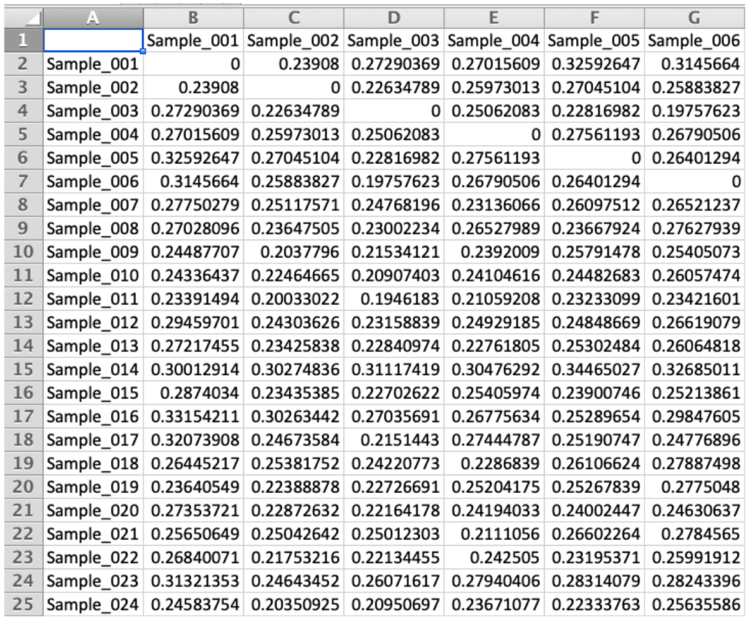


	A	B	C	D	E	F	G
1		Seq_0003	Seq_0004	Seq_0005	Seq_0006	Seq_0007	Seq_0008
2	sample_01	0	0	0	0	0	0
3	sample_02	0	0	0	0	0	0
4	sample_03	0	0	0	0	0	0
5	sample_04	0	27	0	0	0	0
6	sample_05	0	10	0	0	0	0
7	sample_06	0	3	20	0	0	0
8	sample_07	0	10	58	0	0	0
9	sample_08	0	14	52	0	0	0
10	sample_09	0	10	25	0	0	0
11	sample_10	153	0	0	0	0	0
12	sample_11	32	0	14	0	0	0
13	sample_12	97	0	32	0	0	3
14	sample_13	37	0	40	29	18	0
15	sample_14	31	0	27	33	13	25
16	sample_15	12	0	23	33	27	19
17	sample_16	0	0	0	0	0	0
18	sample_17	0	0	0	0	0	0
19	sample_18	0	0	0	0	0	0
20	sample_19	0	55	0	0	0	0
21	sample_20	0	23	0	0	0	0
22	sample_21	0	14	0	0	0	0
23	sample_22	0	26	45	0	0	0
24	sample_23	0	24	54	0	0	0
25	sample_24	0	19	56	0	0	0

Occurrence table

~100 samples

~100 samples



	A	B	C	D	E	F	G
1		Sample_001	Sample_002	Sample_003	Sample_004	Sample_005	Sample_006
2	Sample_001	0	0.23908	0.27290369	0.27015609	0.32592647	0.3145664
3	Sample_002	0.23908	0	0.22634789	0.25973013	0.27045104	0.25883827
4	Sample_003	0.27290369	0.22634789	0	0.25062083	0.22816982	0.19757623
5	Sample_004	0.27015609	0.25973013	0.25062083	0	0.27561193	0.26790506
6	Sample_005	0.32592647	0.27045104	0.22816982	0.27561193	0	0.26401294
7	Sample_006	0.3145664	0.25883827	0.19757623	0.26790506	0.26401294	0
8	Sample_007	0.27750279	0.25117571	0.24768196	0.23136066	0.26097512	0.26521237
9	Sample_008	0.27028096	0.23647505	0.23002234	0.26527989	0.23667924	0.27627939
10	Sample_009	0.24487707	0.2037796	0.21534121	0.2392009	0.25791478	0.25405073
11	Sample_010	0.24336437	0.22464665	0.20907403	0.24104616	0.24482683	0.26057474
12	Sample_011	0.23391494	0.20033022	0.1946183	0.21059208	0.23233099	0.23421601
13	Sample_012	0.29459701	0.24303626	0.23158839	0.24929185	0.24848669	0.26619079
14	Sample_013	0.27217455	0.23425838	0.22840974	0.22761805	0.25302484	0.26064818
15	Sample_014	0.30012914	0.30274836	0.31117419	0.30476292	0.34465027	0.32685011
16	Sample_015	0.2874034	0.23435385	0.22702622	0.25405974	0.23900746	0.25213861
17	Sample_016	0.33154211	0.30263442	0.27035691	0.26775634	0.25289654	0.29847605
18	Sample_017	0.32073908	0.24673584	0.2151443	0.27444787	0.25190747	0.24776896
19	Sample_018	0.26445217	0.25381752	0.24220773	0.2286839	0.26106624	0.27887498
20	Sample_019	0.23640549	0.22388878	0.22726691	0.25204175	0.25267839	0.2775048
21	Sample_020	0.27353721	0.22872632	0.22164178	0.24194033	0.24002447	0.24630637
22	Sample_021	0.25650649	0.25042642	0.25012303	0.2111056	0.26602264	0.2784565
23	Sample_022	0.26840071	0.21753216	0.22134455	0.242505	0.23195371	0.25991912
24	Sample_023	0.31321353	0.24643452	0.26071617	0.27940406	0.28314079	0.28243396
25	Sample_024	0.24583754	0.20350925	0.20950697	0.23671077	0.22333763	0.25635586

Distances matrix

Beta-diversity

- Diversity between two samples/ecosystems (feature level)
- Calculate distances between samples
  - ❑ **Jaccard** (presence/absence in occurrence table)

$$\mathbf{J_{AB} = AB\ /\ (AB + A + B)}$$

J_{AB}: Jaccard similarity between samples A and B
AB: species present in A and B
A: species only present in A
B: species only present in B

Beta-diversity

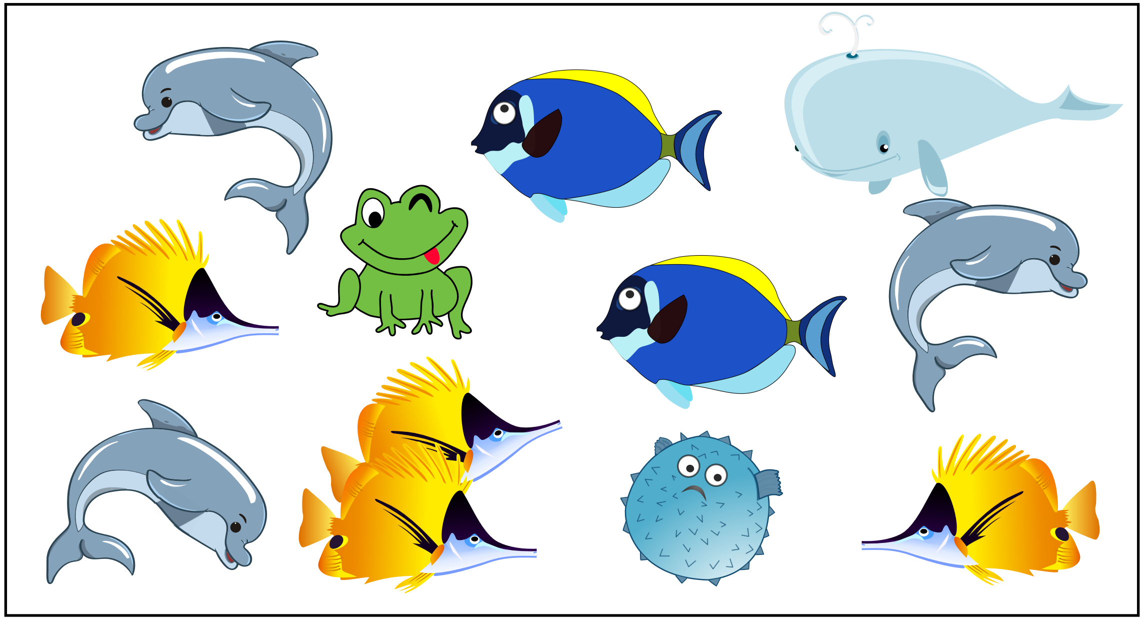
- Diversity between two samples/ecosystems (feature level)
- Calculate distances between samples
  - ❑ Jaccard (presence/absence in occurrence table)
  - ❑ **Bray-Curtis** (occurrence table)

$$\boldsymbol{d\mathrm{BC}_{AB} = \Sigma_{s=1} |A_S - B_S| / (n_A + n_B)}$$

$d\mathrm{BC}_{AB}$: Bray Curtis distance
A_S: number of reads for species S in sample A
B_S: number of reads for species S in sample B
n_A: total number of reads in sample A
n_B: total number of reads in sample B

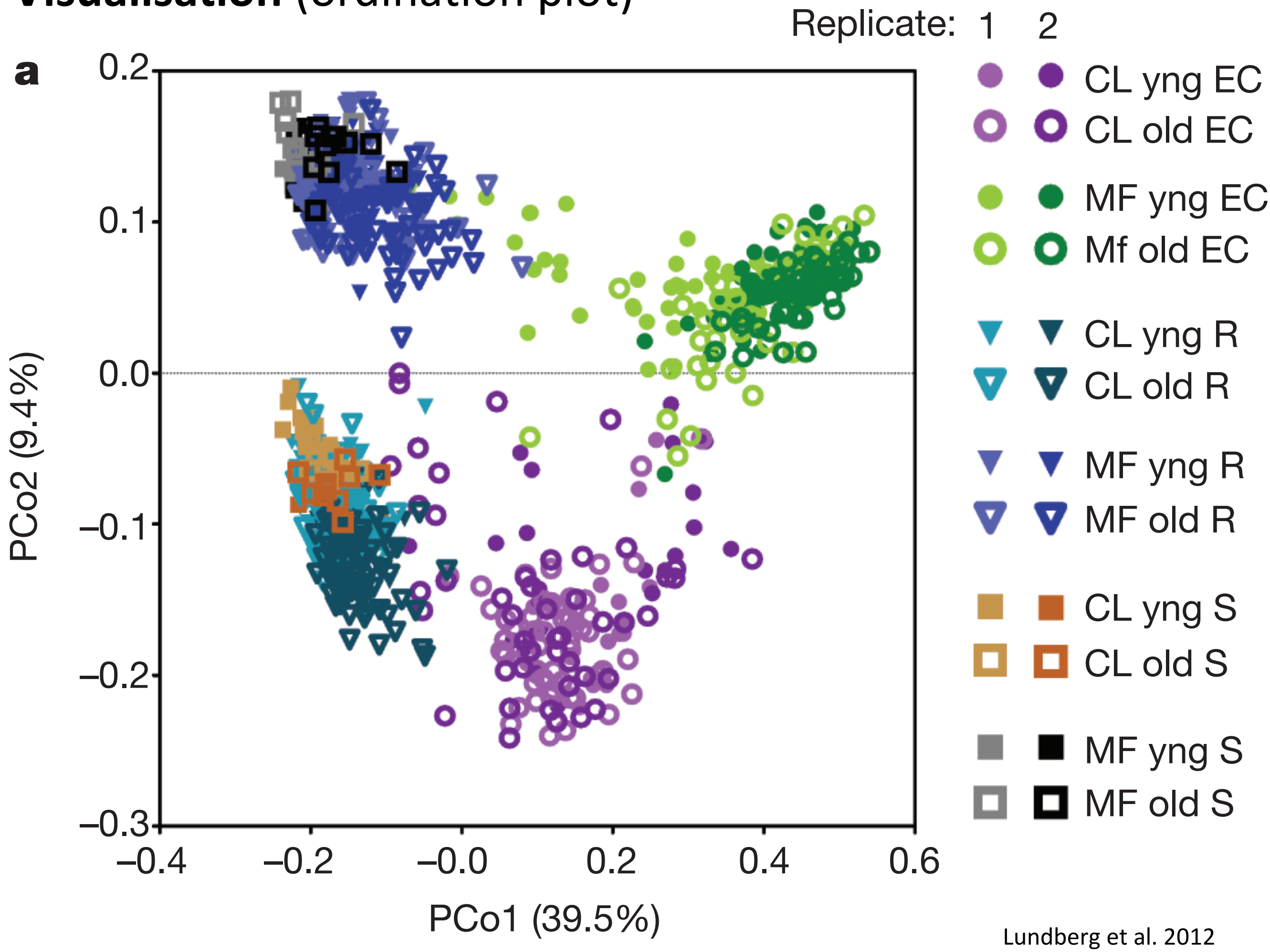
Beta-diversity

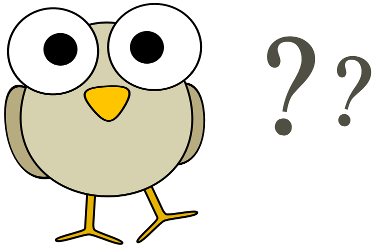
- Diversity between two samples/ecosystems (feature level)
- Calculate distances between samples
  - ❑ Jaccard (presence/absence in occurrence table)
  - ❑ Bray-Curtis (occurrence table)
  - ❑ Unifrac (occurrence table and phylogeny)
    - ❑ Unweighted
    - ❑ Weighted



Beta-diversity

- Diversity between two samples/ecosystems (feature level)
- Calculate distances between samples
- **Visualisation** (ordination plot)

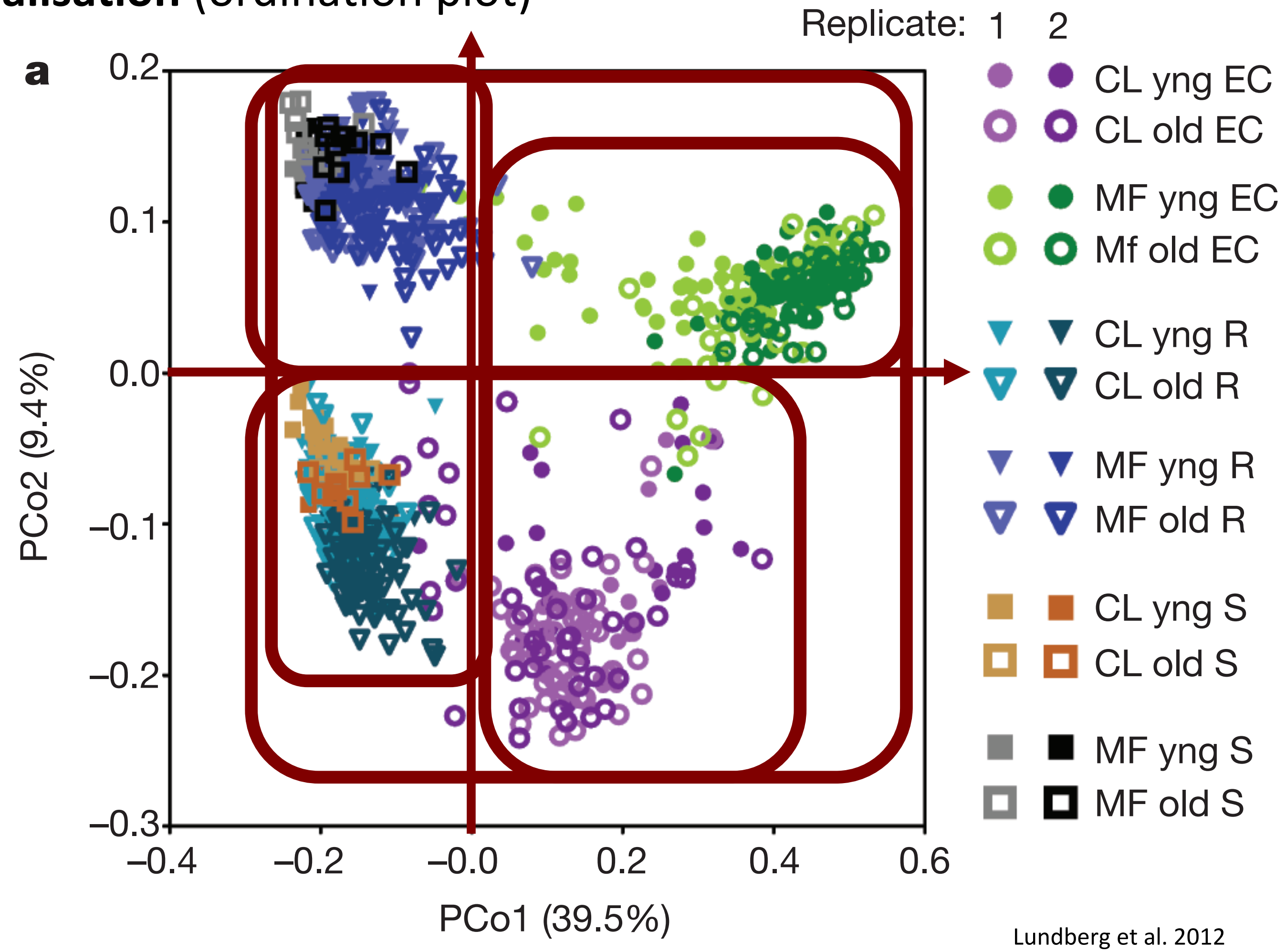




Lundberg et al. 2012

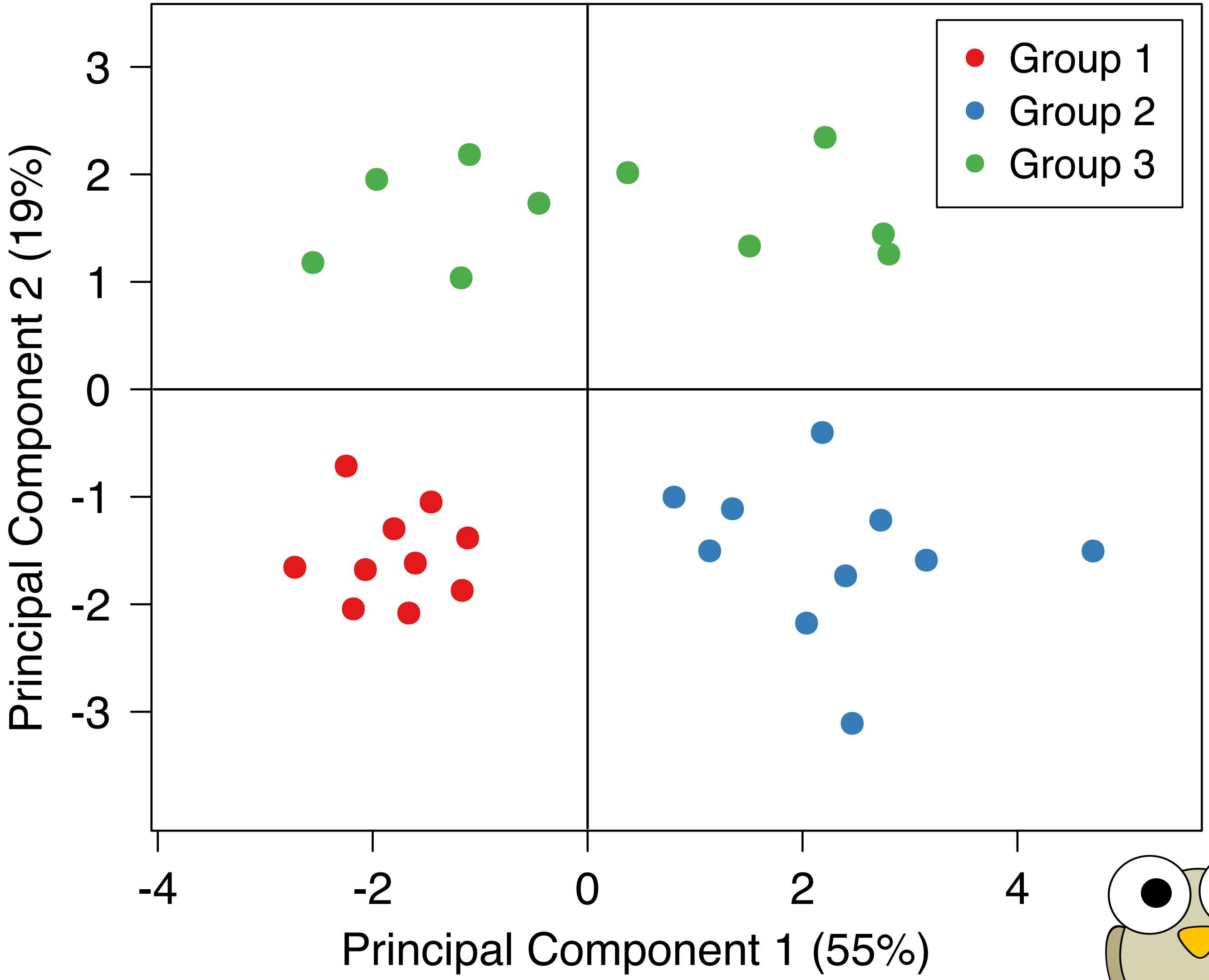
Beta-diversity

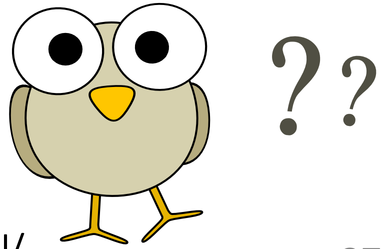
- Diversity between two samples/ecosystems (feature level)
- Calculate distances between samples
- **Visualisation** (ordination plot)



Lundberg et al. 2012

How do we interpret an ordination plot such as PCA?



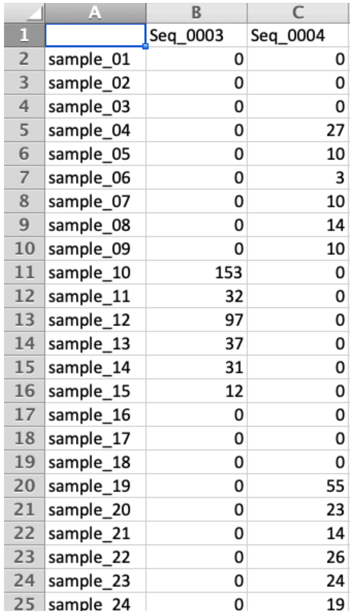


https://bigdata_microbiome.presenterswall.nl/

Why do we use ordination plot such as PCA?

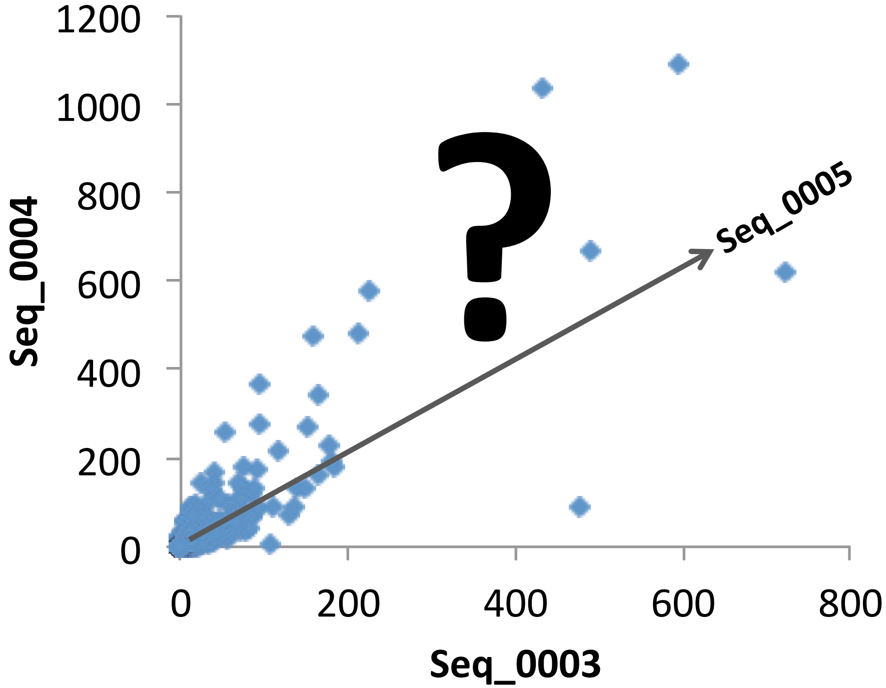
- Visualisation of multivariate data

~100 samples



	A	B	C
1		Seq_0003	Seq_0004
2	sample_01	0	0
3	sample_02	0	0
4	sample_03	0	0
5	sample_04	0	27
6	sample_05	0	10
7	sample_06	0	3
8	sample_07	0	10
9	sample_08	0	14
10	sample_09	0	10
11	sample_10	153	0
12	sample_11	32	0
13	sample_12	97	0
14	sample_13	37	0
15	sample_14	31	0
16	sample_15	12	0
17	sample_16	0	0
18	sample_17	0	0
19	sample_18	0	0
20	sample_19	0	55
21	sample_20	0	23
22	sample_21	0	14
23	sample_22	0	26
24	sample_23	0	24
25	sample_24	0	19

Occurrence table

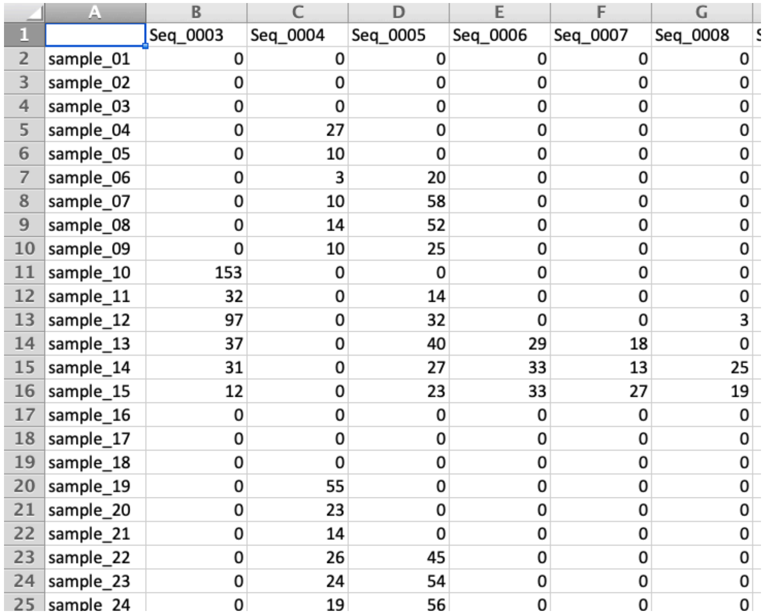


Why do we use ordination plot such as PCA?

- Reduce the dimensionality of a data set

~10,000 features

~100 samples

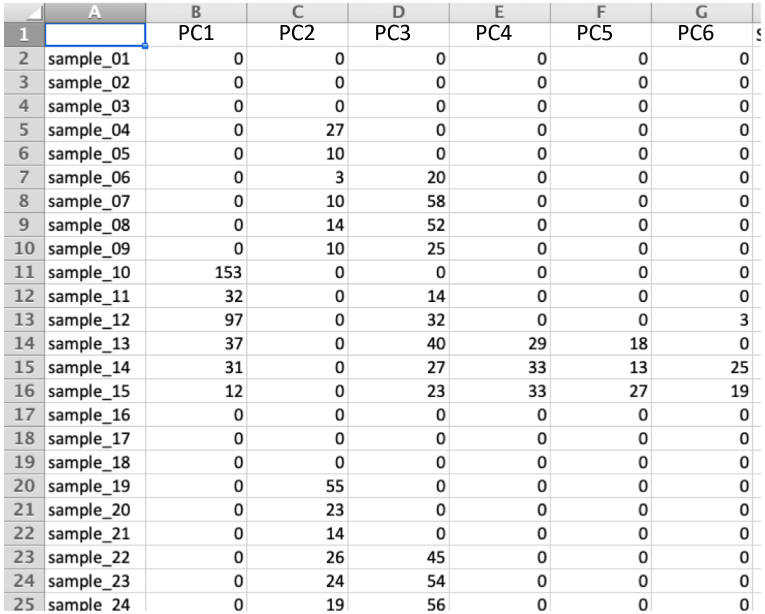


	A	B	C	D	E	F	G
1		Seq_0003	Seq_0004	Seq_0005	Seq_0006	Seq_0007	Seq_0008
2	sample_01	0	0	0	0	0	0
3	sample_02	0	0	0	0	0	0
4	sample_03	0	0	0	0	0	0
5	sample_04	0	27	0	0	0	0
6	sample_05	0	10	0	0	0	0
7	sample_06	0	3	20	0	0	0
8	sample_07	0	10	58	0	0	0
9	sample_08	0	14	52	0	0	0
10	sample_09	0	10	25	0	0	0
11	sample_10	153	0	0	0	0	0
12	sample_11	32	0	14	0	0	0
13	sample_12	97	0	32	0	0	3
14	sample_13	37	0	40	29	18	0
15	sample_14	31	0	27	33	13	25
16	sample_15	12	0	23	33	27	19
17	sample_16	0	0	0	0	0	0
18	sample_17	0	0	0	0	0	0
19	sample_18	0	0	0	0	0	0
20	sample_19	0	55	0	0	0	0
21	sample_20	0	23	0	0	0	0
22	sample_21	0	14	0	0	0	0
23	sample_22	0	26	45	0	0	0
24	sample_23	0	24	54	0	0	0
25	sample_24	0	19	56	0	0	0

Occurrence table

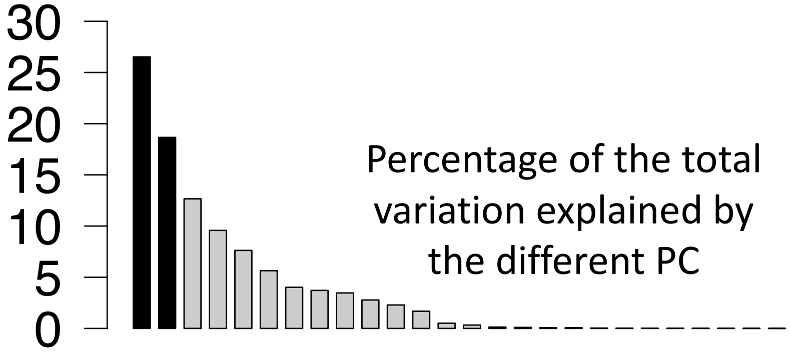
~30 features

~100 samples



	A	B	C	D	E	F	G
1		PC1	PC2	PC3	PC4	PC5	PC6
2	sample_01	0	0	0	0	0	0
3	sample_02	0	0	0	0	0	0
4	sample_03	0	0	0	0	0	0
5	sample_04	0	27	0	0	0	0
6	sample_05	0	10	0	0	0	0
7	sample_06	0	3	20	0	0	0
8	sample_07	0	10	58	0	0	0
9	sample_08	0	14	52	0	0	0
10	sample_09	0	10	25	0	0	0
11	sample_10	153	0	0	0	0	0
12	sample_11	32	0	14	0	0	0
13	sample_12	97	0	32	0	0	3
14	sample_13	37	0	40	29	18	0
15	sample_14	31	0	27	33	13	25
16	sample_15	12	0	23	33	27	19
17	sample_16	0	0	0	0	0	0
18	sample_17	0	0	0	0	0	0
19	sample_18	0	0	0	0	0	0
20	sample_19	0	55	0	0	0	0
21	sample_20	0	23	0	0	0	0
22	sample_21	0	14	0	0	0	0
23	sample_22	0	26	45	0	0	0
24	sample_23	0	24	54	0	0	0
25	sample_24	0	19	56	0	0	0

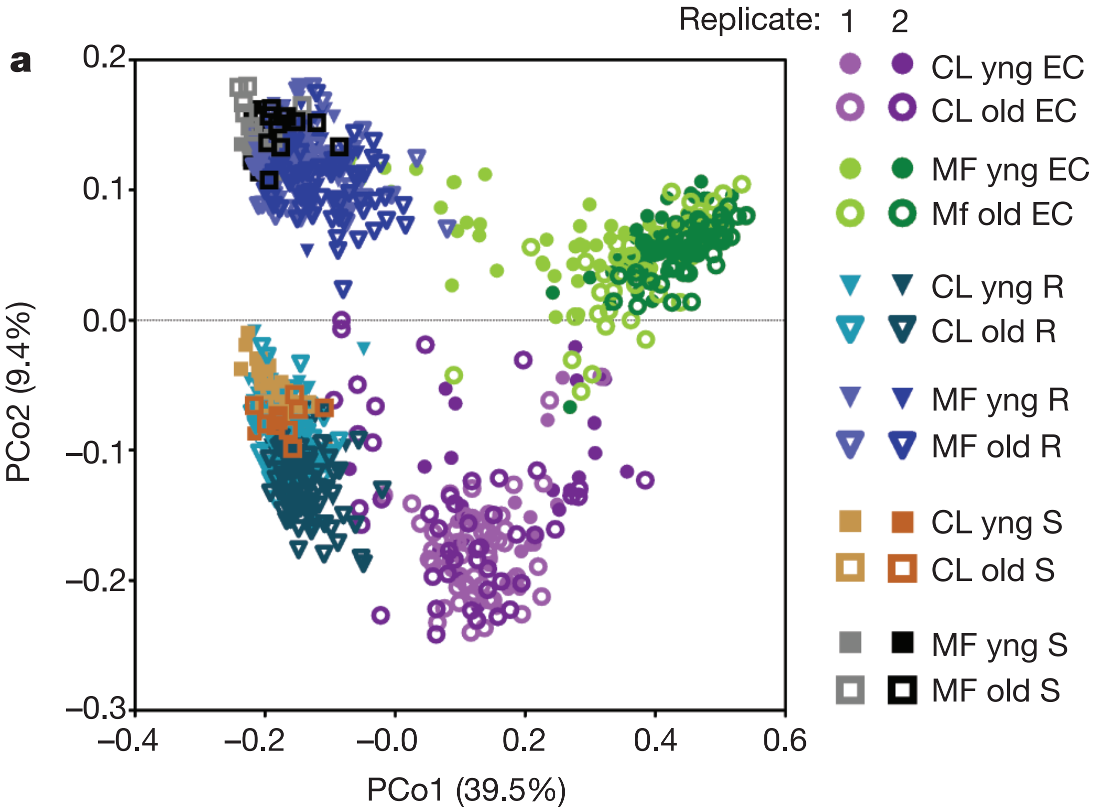
Component table



Percentage of the total variation explained by the different PC

Beta-diversity

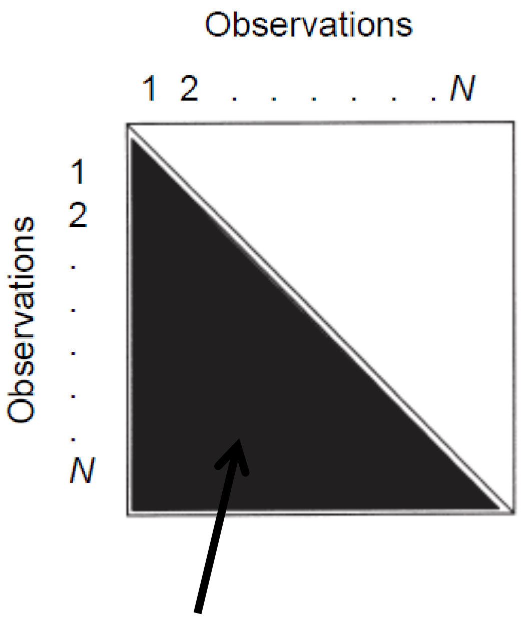
- Diversity between two samples/ecosystems (feature level)
- Calculate distances between samples
- Visualisation (ordination plot)
  - **Principal Coordinate Analysis (PCoA)**
    => can handle different types of distance measurements (such as Bray-Curtis)



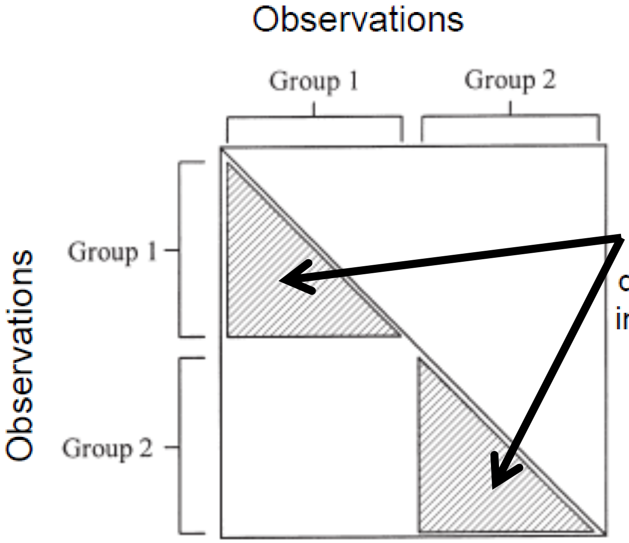
Lundberg et al. 2012

Beta-diversity

- Diversity between two samples/ecosystems (feature level)
- Calculate distances between samples
- Visualisation (ordination plot)
- Statistical comparison among sets of communities
  - ❑ **PERMANOVA**: ANOVA type method based on sample to sample distances to compare within and between group distances & P-value by permutation



SS_T = sum of the squared distances in the half-matrix, divided by the total number of observations (*N*)



SS_W = sum of the squared distances between replicates in the same group, divided by the number of replicates per group (*n*).

$$SS_A = SS_T - SS_W$$

$$F = \frac{SS_A/(a-1)}{SS_W/(N-a)}$$

Adapted from Anderson 2001

Beta-diversity

- Diversity between two samples/ecosystems (feature level)
- Calculate distances between samples
- Visualisation (ordination plot)
- Statistical comparison among sets of communities
  - **PERMANOVA**: ANOVA type method based on sample to sample distances to compare within and between group distances & P-value by permutation
  - **ANOSIM**: Similar to Permanova, but analysis is performed on ranked distances

Beta-diversity

- ➢ Diversity between two samples/ecosystems
- ➢ Different distance measurements:
  - Jaccard (occurrence table: presence/absence)
  - Bray-Curtis (occurrence table: abundance)
  - Unifrac (occurrence table and phylogeny)
- ➢ Visualisation using ordination plot (PCOA)

Practice time: beta-diversity

In the tutorial, look at:

- 6. Beta-diversity

Tutorial link:

https://scienceparkstudygroup.github.io/microbiome-lesson/06-beta-diversity/index.html

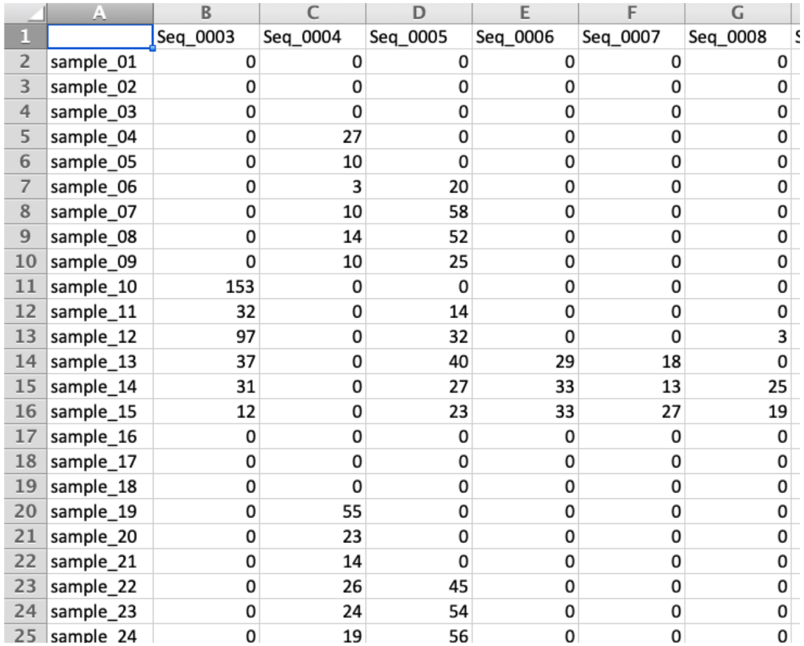
Learning objectives

- ☑ Define microbiome and state microbiome importance
- ☑ Identify differences between metabarcoding and metagenomics
- ☑ Explain how microbiota data are generated (including bias)
- ☑ Explain and preform data pre-processing
- □ Explain how microbiota data are analysed
- ☑ Define, perform and interpret alpha-diversity
- ☑ Address sparsity, under-sampling and uneven sampling depth using data filtering and normalization
- ☑ Define, perform and interpret beta-diversity
- ☑ Generate and interpret multivariate data analyses
- ☑ Perform and interpret appropriate statistical tests
- □ Visualize and interpret microbial community composition

Microbial composition

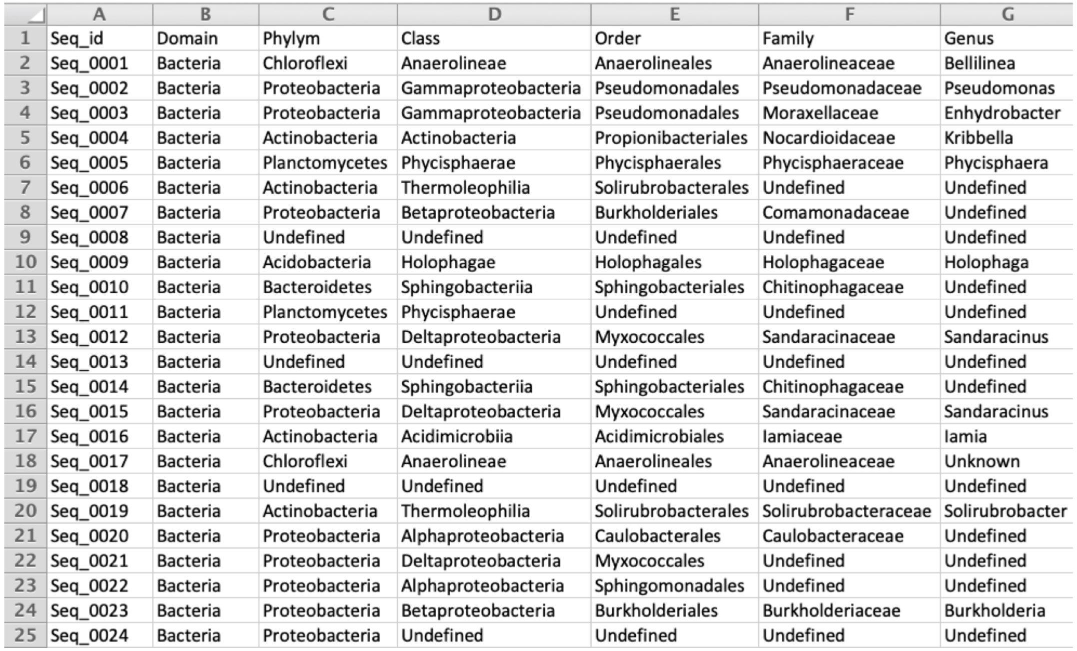
- Aggregate sequences according to their taxonomic assignment

~10,000 features **Occurrence data**



	A	B	C	D	E	F	G
1		Seq_0003	Seq_0004	Seq_0005	Seq_0006	Seq_0007	Seq_0008
2	sample_01	0	0	0	0	0	0
3	sample_02	0	0	0	0	0	0
4	sample_03	0	0	0	0	0	0
5	sample_04	0	27	0	0	0	0
6	sample_05	0	10	0	0	0	0
7	sample_06	0	3	20	0	0	0
8	sample_07	0	10	58	0	0	0
9	sample_08	0	14	52	0	0	0
10	sample_09	0	10	25	0	0	0
11	sample_10	153	0	0	0	0	0
12	sample_11	32	0	14	0	0	0
13	sample_12	97	0	32	0	0	3
14	sample_13	37	0	40	29	18	0
15	sample_14	31	0	27	33	13	25
16	sample_15	12	0	23	33	27	19
17	sample_16	0	0	0	0	0	0
18	sample_17	0	0	0	0	0	0
19	sample_18	0	0	0	0	0	0
20	sample_19	0	55	0	0	0	0
21	sample_20	0	23	0	0	0	0
22	sample_21	0	14	0	0	0	0
23	sample_22	0	26	45	0	0	0
24	sample_23	0	24	54	0	0	0
25	sample_24	0	19	56	0	0	0

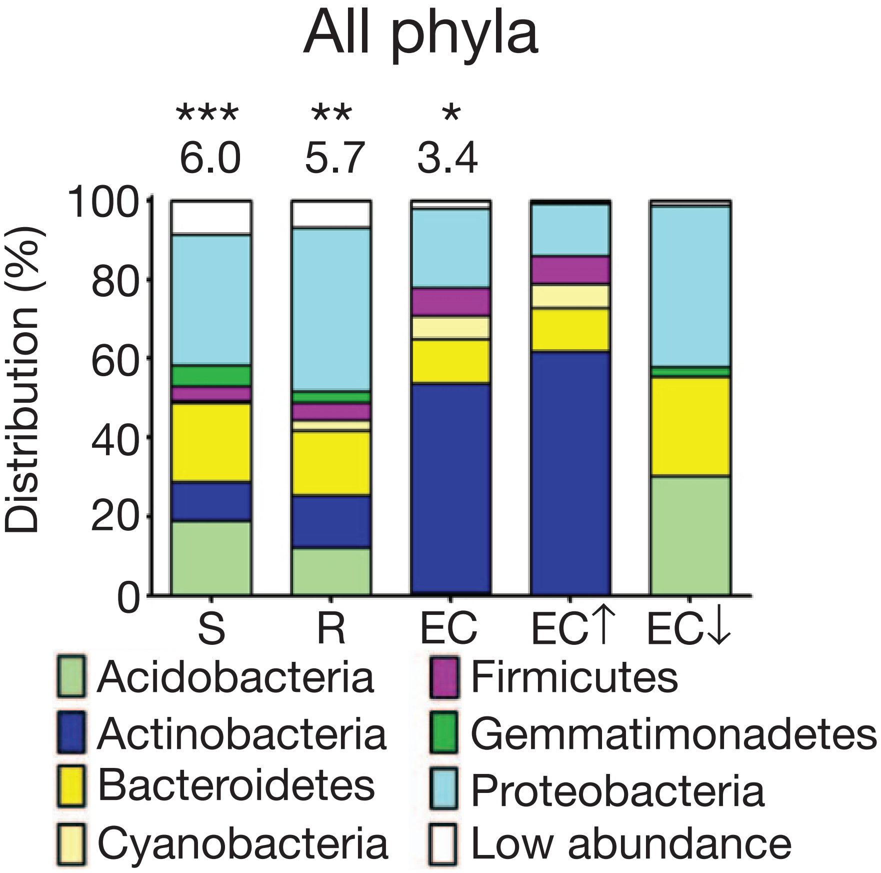
Observation metadata

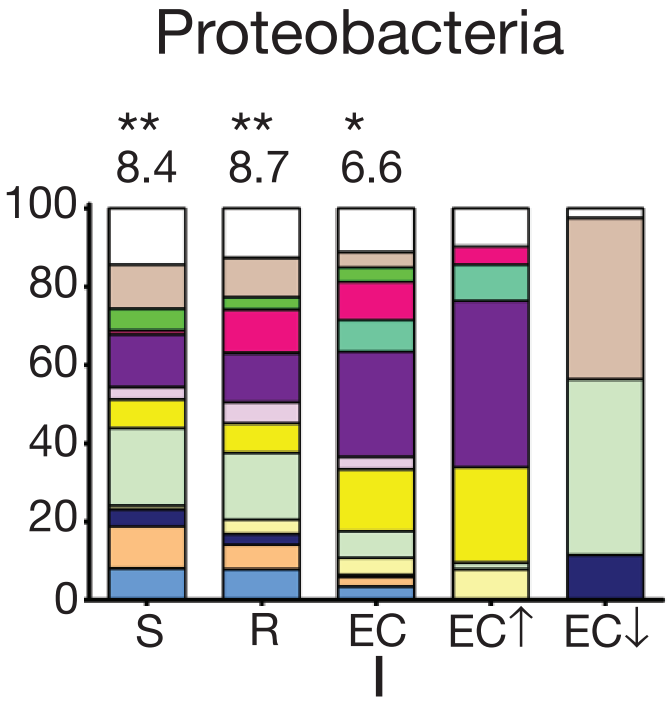


	A	B	C	D	E	F	G
1	Seq_id	Domain	Phylym	Class	Order	Family	Genus
2	Seq_0001	Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	Anaerolineaceae	Bellilinea
3	Seq_0002	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas
4	Seq_0003	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Enhydrobacter
5	Seq_0004	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Nocardioidaceae	Kribbella
6	Seq_0005	Bacteria	Planctomycetes	Phycisphaerae	Phycisphaerales	Phycisphaeraceae	Phycisphaera
7	Seq_0006	Bacteria	Actinobacteria	Thermoleophilia	Solirubrobacterales	Undefined	Undefined
8	Seq_0007	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Undefined
9	Seq_0008	Bacteria	Undefined	Undefined	Undefined	Undefined	Undefined
10	Seq_0009	Bacteria	Acidobacteria	Holophagae	Holophagales	Holophagaceae	Holophaga
11	Seq_0010	Bacteria	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Chitinophagaceae	Undefined
12	Seq_0011	Bacteria	Planctomycetes	Phycisphaerae	Undefined	Undefined	Undefined
13	Seq_0012	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	Sandaracinaceae	Sandaracinus
14	Seq_0013	Bacteria	Undefined	Undefined	Undefined	Undefined	Undefined
15	Seq_0014	Bacteria	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Chitinophagaceae	Undefined
16	Seq_0015	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	Sandaracinaceae	Sandaracinus
17	Seq_0016	Bacteria	Actinobacteria	Acidimicrobiia	Acidimicrobiales	Iamiaceae	Iamia
18	Seq_0017	Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	Anaerolineaceae	Unknown
19	Seq_0018	Bacteria	Undefined	Undefined	Undefined	Undefined	Undefined
20	Seq_0019	Bacteria	Actinobacteria	Thermoleophilia	Solirubrobacterales	Solirubrobacteraceae	Solirubrobacter
21	Seq_0020	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Undefined
22	Seq_0021	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	Undefined	Undefined
23	Seq_0022	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Undefined	Undefined
24	Seq_0023	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia
25	Seq_0024	Bacteria	Proteobacteria	Undefined	Undefined	Undefined	Undefined

Microbial composition

- Aggregate sequences according to their taxonomic assignment
- Plot microbial composition


All phyla


Proteobacteria

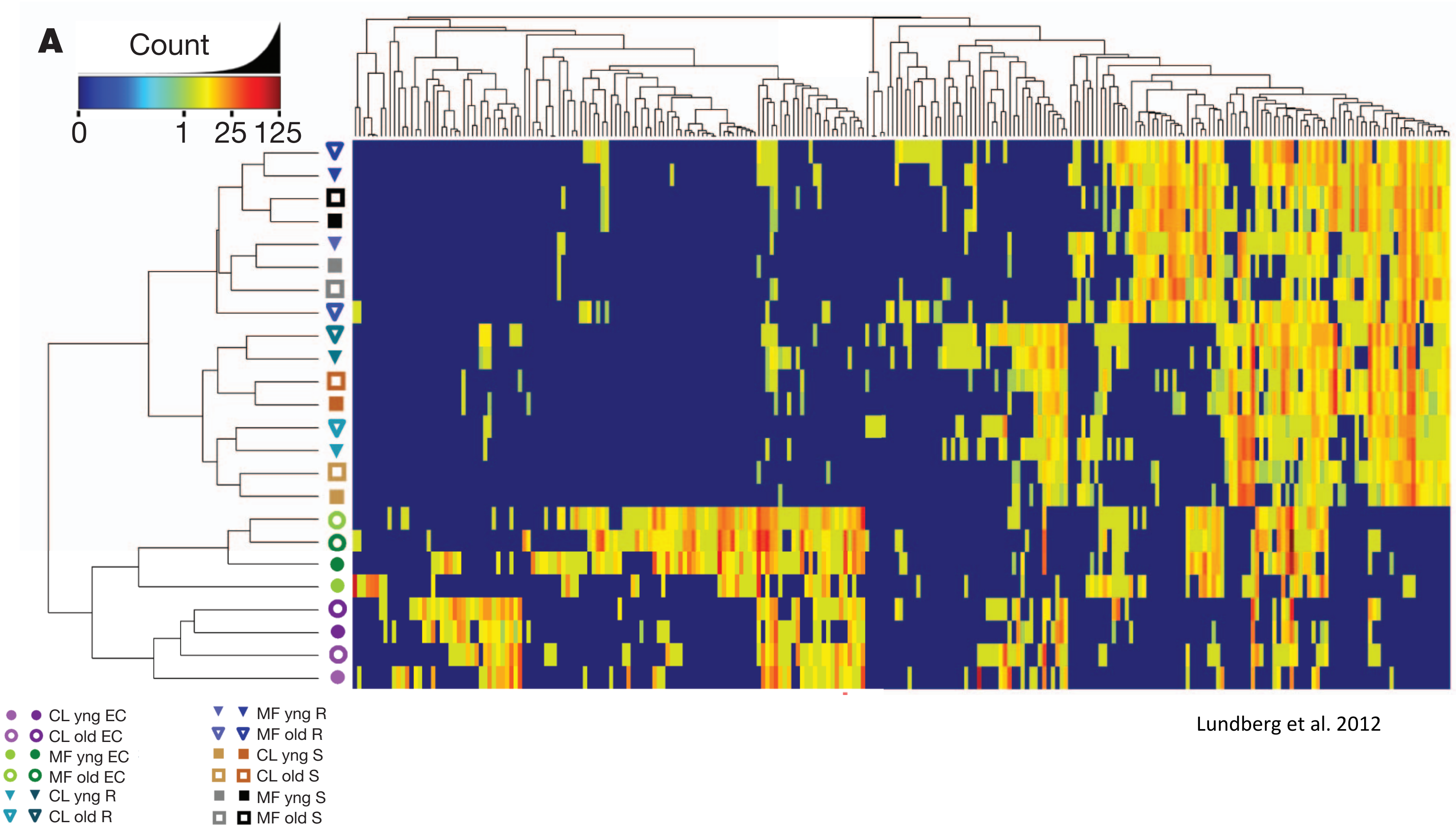
Lundberg et al. 2012

Defining the core *Arabidopsis thaliana* root microbiome

Derek S. Lundberg[1,2]*, Sarah L. Lebeis[1]*, Sur Herrera Paredes[1]*, Scott Yourstone[1,3]*, Jase Gehring[1], Stephanie Malfatti[4], Julien Tremblay[4], Anna Engelbrektson[4]†, Victor Kunin[4]†, Tijana Glavina del Rio[4], Robert C. Edgar[5], Thilo Eickhorst[6], Ruth E. Ley[7], Philip Hugenholtz[4,8], Susannah Green Tringe[4] & Jeffery L. Dangl[1,2,9,10,11]

Microbial composition

- Aggregate sequences according to their taxonomic assignment
- Plot microbial composition



Lundberg et al. 2012

Practice time: microbial composition

In the tutorial, look at:

- 7. Bacterial community composition

Tutorial link:

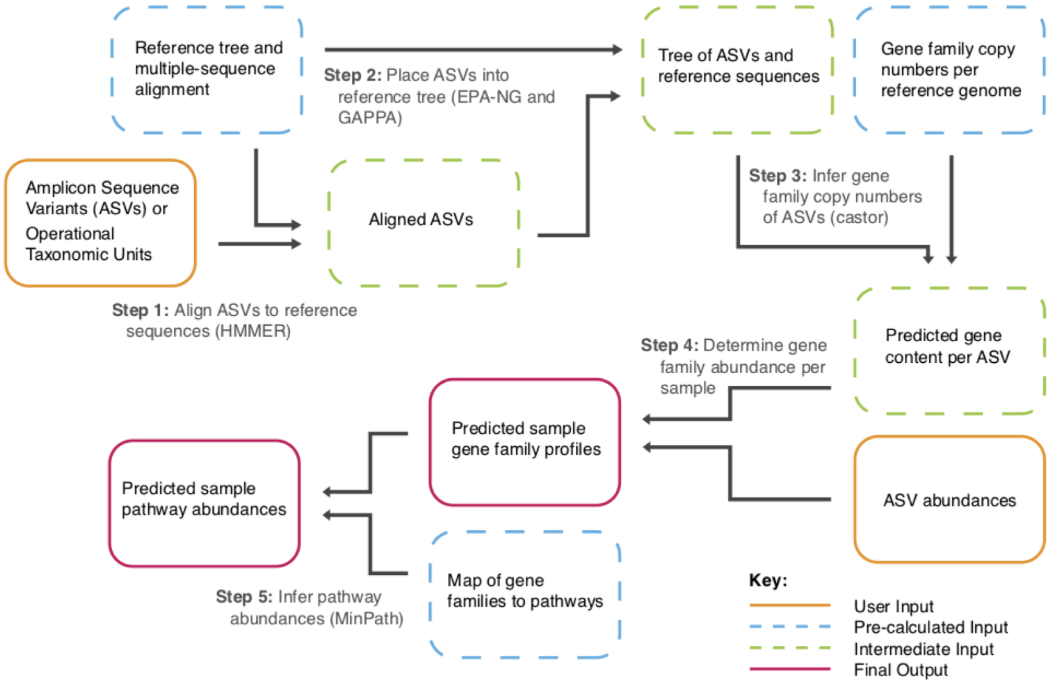
https://scienceparkstudygroup.github.io/microbiome-lesson/07-bacterial-composition/index.html

Other classic microbiota analyses and perspectives

- Co-occurrence analyses
- Functional prediction (*e.g.* PICRUST)
- New sequencing technologies
  - ❑ Long reads for a better identification
  - ❑ No amplification

MinION

PacBio



Reference tree and multiple-sequence alignment

Amplicon Sequence Variants (ASVs) or Operational Taxonomic Units

Step 1: Align ASVs to reference sequences (HMMER)

Aligned ASVs

Step 2: Place ASVs into reference tree (EPA-NG and GAPPA)

Tree of ASVs and reference sequences

Gene family copy numbers per reference genome

Step 3: Infer gene family copy numbers of ASVs (castor)

Predicted gene content per ASV

Step 4: Determine gene family abundance per sample

ASV abundances

Predicted sample gene family profiles

Map of gene families to pathways

Step 5: Infer pathway abundances (MinPath)

Predicted sample pathway abundances

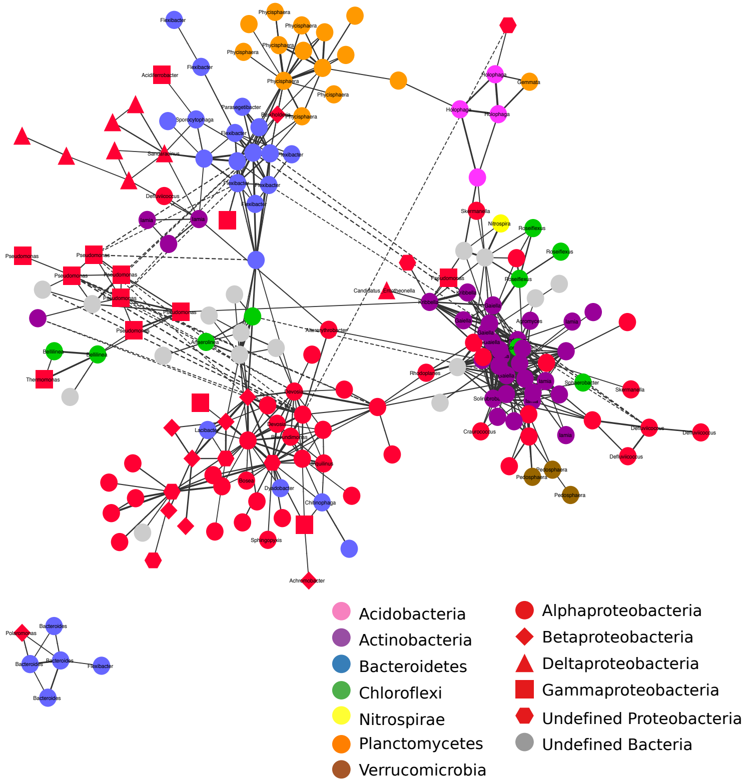
Key:

User Input

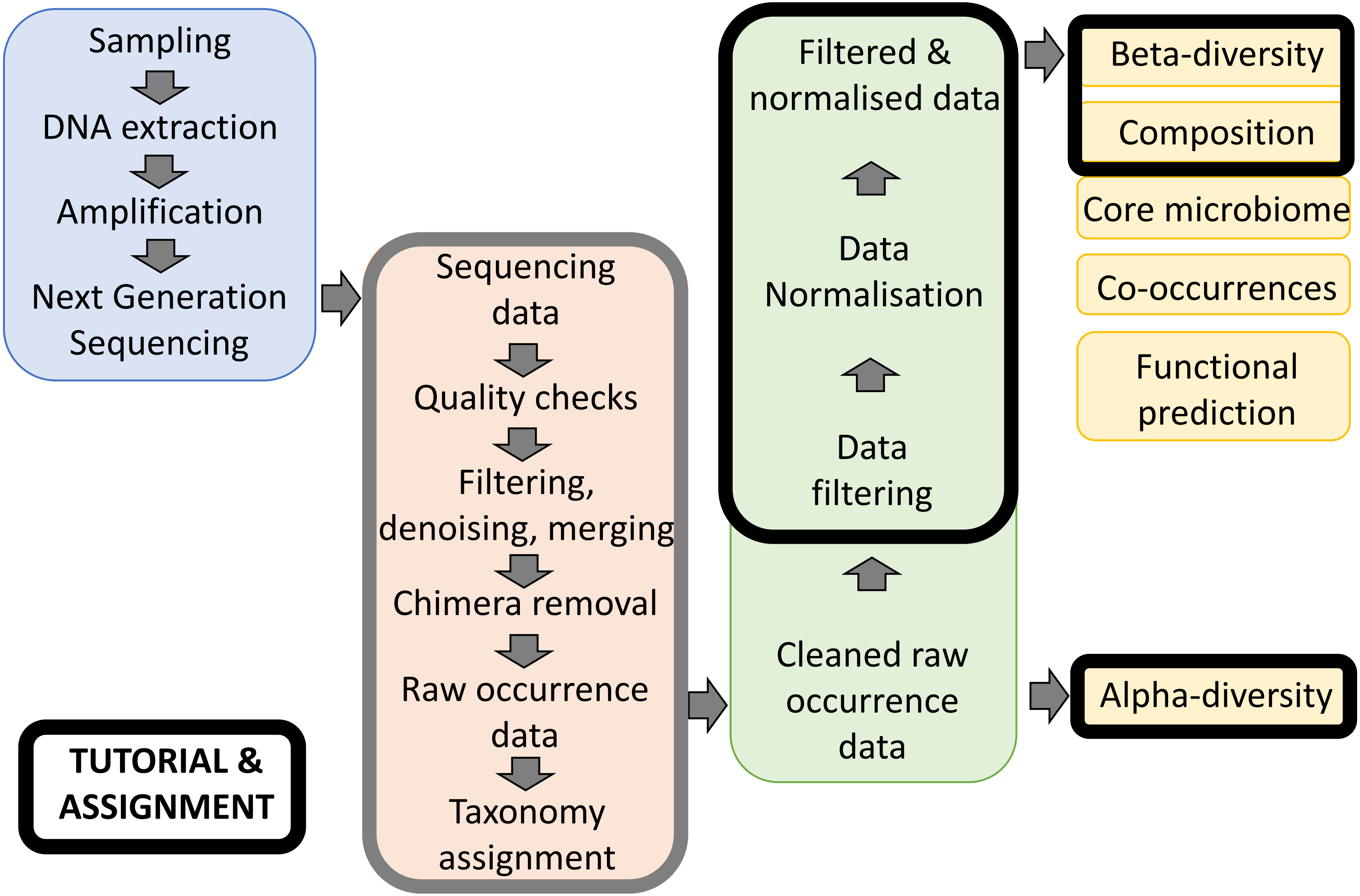
Pre-calculated Input

Intermediate Input

Final Output



Microbiota analysis : data analysis overview



Sampling

DNA extraction

Amplification

Next Generation Sequencing

Sequencing data

Quality checks

Filtering, denoising, merging

Chimera removal

Raw occurrence data

Taxonomy assignment

Cleaned raw occurrence data

Data filtering

Data Normalisation

Filtered & normalised data

Beta-diversity

Composition

Core microbiome

Co-occurrences

Functional prediction

Alpha-diversity

TUTORIAL & ASSIGNMENT

Microbiota analysis : results discussion

- Scientific context, research question and experimental design
- Data properties (*i.e.* sparsity and library size)
- Data filtering and normalisation
- Alpha-diversity
- Beta-diversity
- Microbial composition
- Conclusion

Learning objectives

- ☑ Define microbiome and state microbiome importance
- ☑ Identify differences between metabarcoding and metagenomics
- ☑ Explain how microbiota data are generated (including bias)
- ☑ Explain and preform data pre-processing
- ☑ Explain how microbiota data are analysed
- ☑ Define, perform and interpret alpha-diversity
- ☑ Address sparsity, under-sampling and uneven sampling depth using data filtering and normalization
- ☑ Define, perform and interpret beta-diversity
- ☑ Generate and interpret multivariate data analyses
- ☑ Perform and interpret appropriate statistical tests
- ☑ Visualize and interpret microbial community composition

Microbiota data analysis assignment

- Scientific context, research question and experimental design
- Data properties (*i.e.* sparsity and library size)
- Data filtering and normalisation
- Alpha-diversity
- Beta-diversity
- Microbial composition
- Conclusion

Detailed instructions available on Canvas

- Rmarkdown report in pdf
- Think about reproducibility
  - What have you done?
  - Why?
- Include, describe and interpret your plots & statistical results