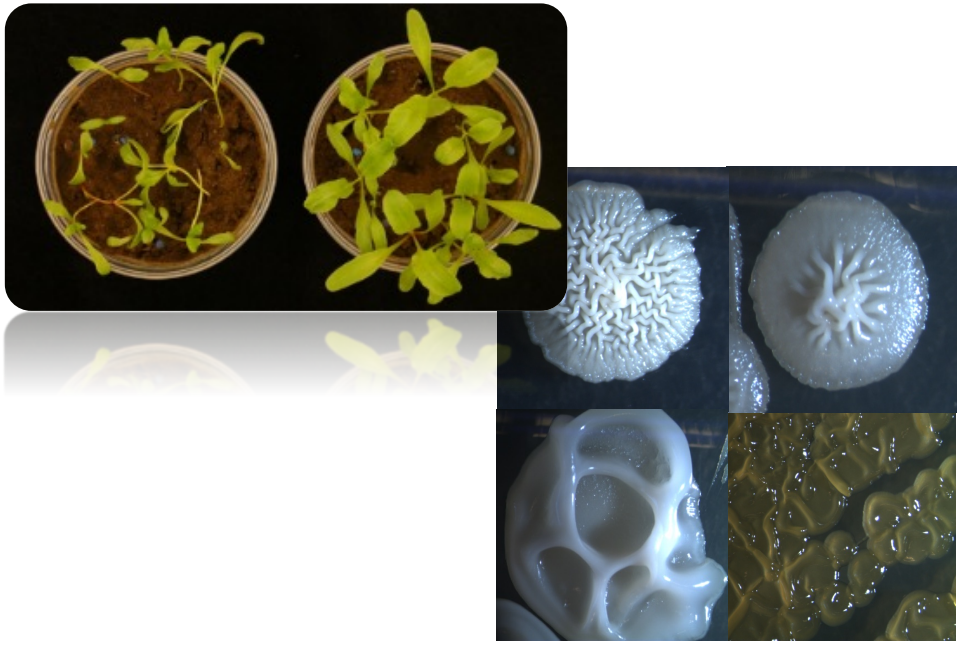
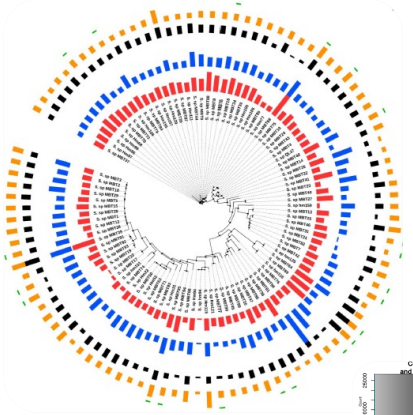
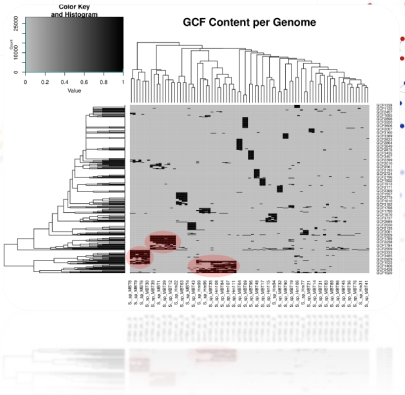
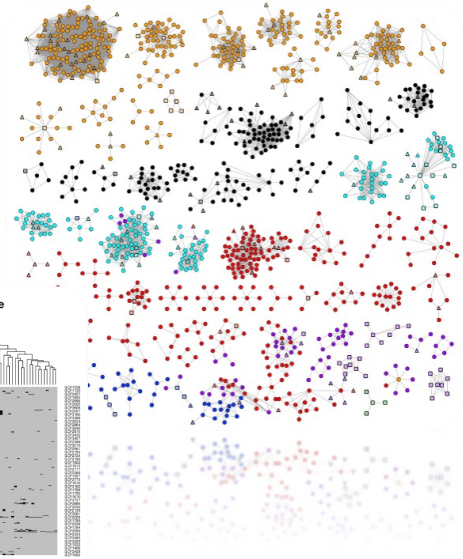
AMPLICON SEQUENCING TECHNIQUES









Victor J **Carrión**

v.j.carrion.bravo@biology.leidenuniv.nl

@VCarryOn1

CARRION LAB

Universiteit Leiden

NIOO

KEY QUESTIONS IN MICROBIOME RESEARCH

Who?
What?
How?

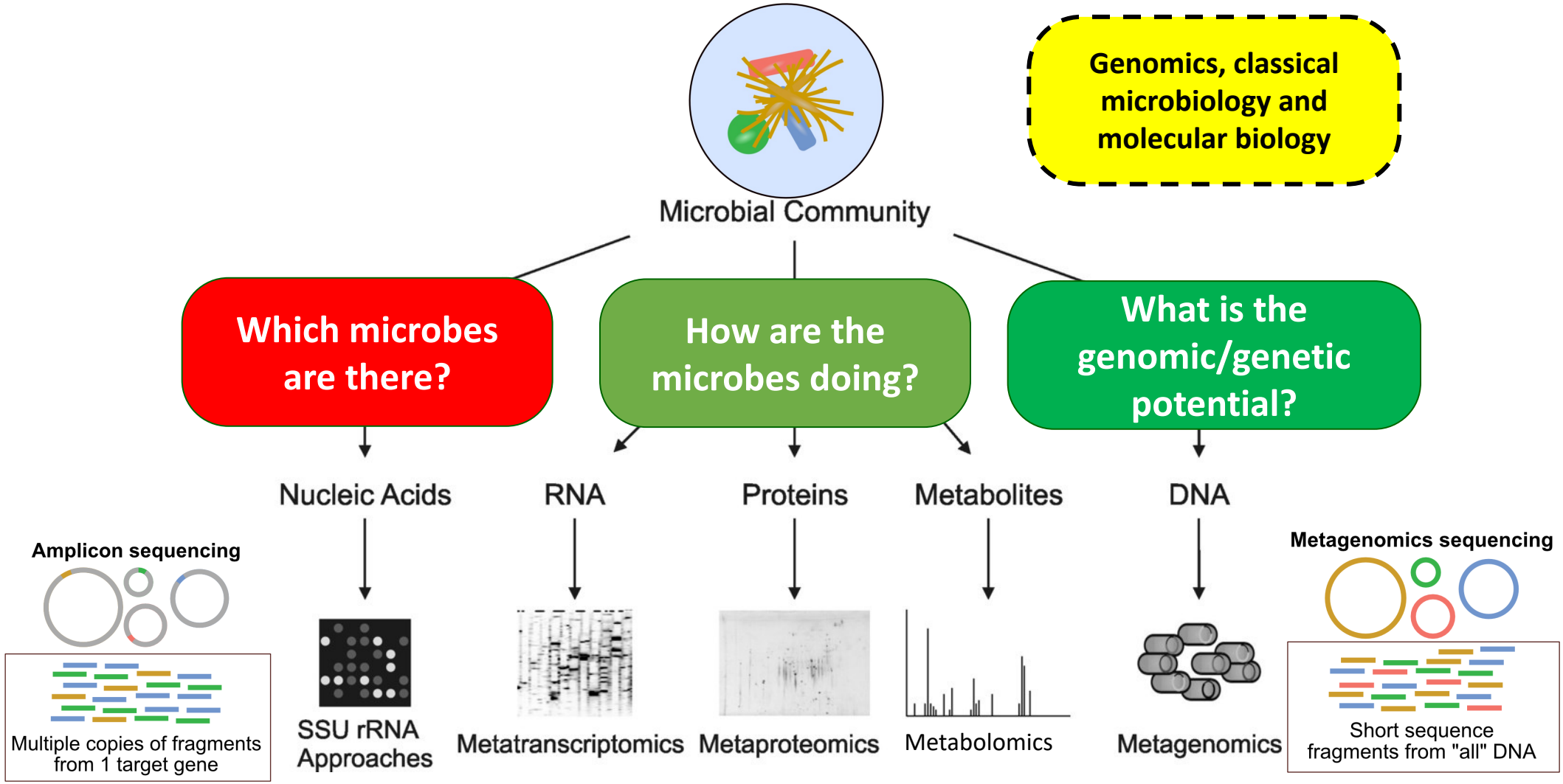
- to what extent do microbiomes influence (eco)system functions?
- what factors, mechanisms drive microbiome assembly & activity?
- are there general patterns in microbiome dynamics & functioning?

→ **systems approach** to study & engineer microbiomes

LEARNING GOALS

- The tool box
- Deciding what to do …
- 16/ITS/18S rRNA
- Clustering vs ASVs
- Microbial diversity measures (alpha)
- Beta diversity
- Differential abundance

THE TOOLBOX …


Genomics, classical microbiology and molecular biology
Microbial Community
Which microbes are there?
How are the microbes doing?
What is the genomic/genetic potential?
Nucleic Acids
RNA
Proteins
Metabolites
DNA
Amplicon sequencing
Multiple copies of fragments from 1 target gene
SSU rRNA Approaches
Metatranscriptomics
Metaproteomics
Metabolomics
Metagenomics
Metagenomics sequencing
Short sequence fragments from "all" DNA

QUESTIONS BEFORE CHOOSING A TOOL

1. What is the research question or hypothesis that the study aims to address?

- Taxonomy, functions or both …

2. What is the biological sample that will be analyzed (e.g., fecal samples, soil samples, water samples, etc.)?

- Possible contaminants, sampling bias …

3. What is the expected complexity of the microbial community in the sample?

- Soil > Rhizosphere > Endosphere … sequencing depth

4. What are the limitations and potential biases of the sequencing technology and analysis methods used?

- Illumina, nanopore, pacbio

5. What are the appropriate controls to ensure the accuracy and reliability of the results?

6. How will the data be analyzed and interpreted to answer the research question or hypothesis?

- MetagenomeSeq/Deseq2/EdgeR

WHAT ABOUT THE METADATA?

- Metadata is information about your samples other than the primary ‘omics’ data. It is data in itself.
- Examples:
  - Date and location where sample was collected
  - Location of raw sample
  - Experimental metadata: controls, replicates, etc.
  - Physical and chemical properties of the environment
  - Ontology designations (ENVO, EMPO)
  - Taxonomy of sample and host

WHY METADATA IS SO IMPORTANT?

- Data are meaningless if you don't know where they came from.

- Microbial communities are highly adapted to their environments; metadata are required to make sense of these patterns.

- Primary data (e.g., sequences and metabolite profiles) can often be regenerated (and may be if technologies improve), but metadata doesn't change and often must be collected at time of sampling.

METADATA AND SEQUENCE DATABASES

Qiita

Database of sequences, observation tables, and metadata

Analysis tools (QIIME)

qiita.ucsd.edu

EBI

European Bioinformatics Institute

European Nucleotide Archive

Central sequence and metadata repository

www.ebi.ac.uk/ena

MIxS

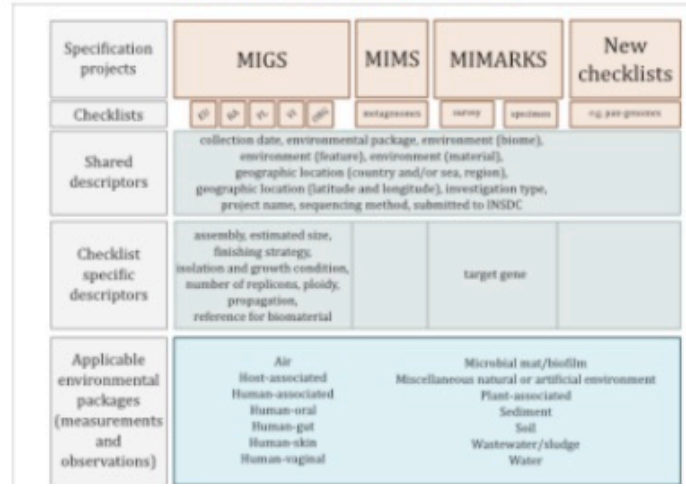
Minimum Information about Any (x) Sequence

x = Genome (G)

x = Metagenome (M)

x = Marker gene (MARK)

gensc.org/mixs



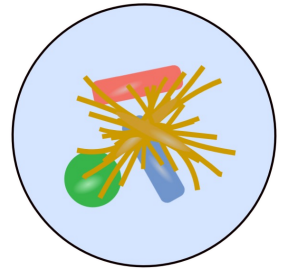
QIITA workshop, L. Thompson

TWO COMMON OPTIONS FOR MICROBIAL DNA

- **Amplicon sequencing** (16S ribosomal RNA (rRNA), 18S rRNA, ITS) Sequence a small section of taxonomically informative target DNA to study microbial composition and diversity

- **Shotgun metagenomic sequencing**
Randomly break up the DNA, sequence all of the fragments to study potential gene function and assemble genomes/partial genomes

THREE MAIN QUESTIONS FOR MICROBIOME RESEARCH



Which microbes are there?

How are the microbes doing?

What is the genomic/genetic potential?

Who is there?

16S/18S/ITS amplicon sequencing

SO WHAT CHARACTERISTICS OF A GENE MAKE IT A GOOD MARKER?

- **Genes that are ubiquitous** (e.g. important to the function of all living organisms)

- **Genes that contains both:**

  > Conserved region – common between all microbes of interest e.g. a gene region present in all bacteria and archaea (so universal primers can find it)

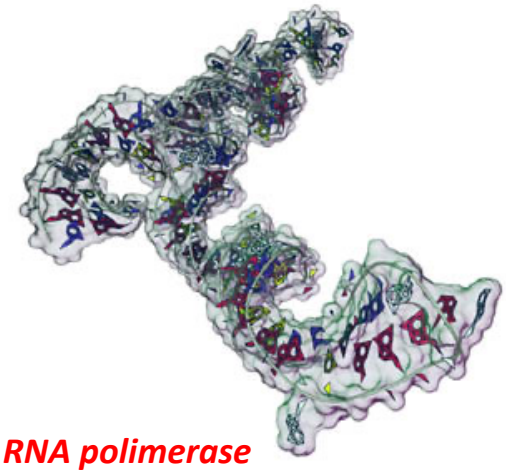
  > Variable region – different between taxa contained within your microbial group of interest e.g. a region within a bacterial marker gene that differentiates *E. coli* or *P. aeruginosa*

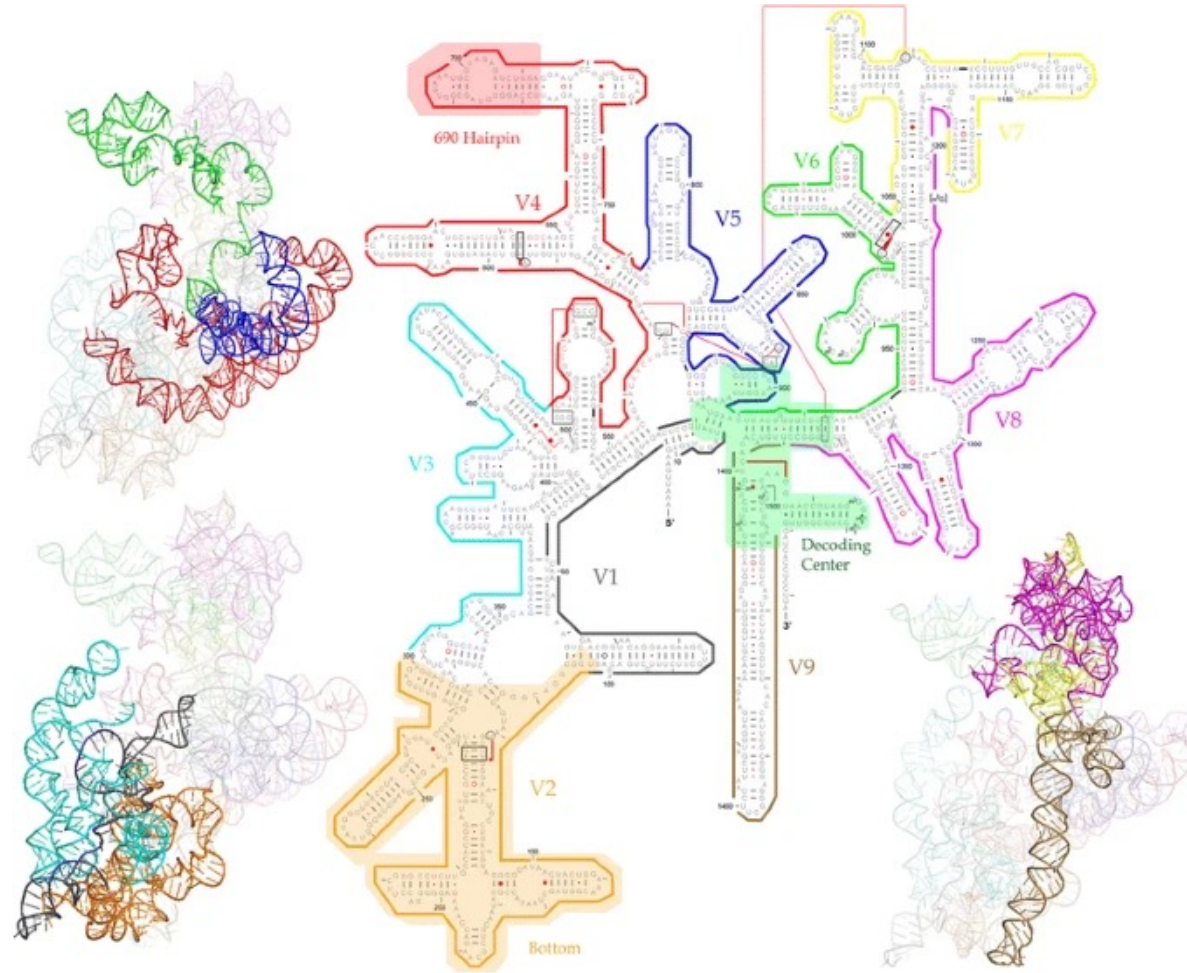
LOOKING FOR A MARKER GENE FOR TAXONOMIC AND PHYLOGENETIC INFORMATION IN MICROBIAL ECOLOGY

✓ Present in all species (ancient gene)

✓ Variable and conserved regions (alternated)

✓ Evolutionary chronometer





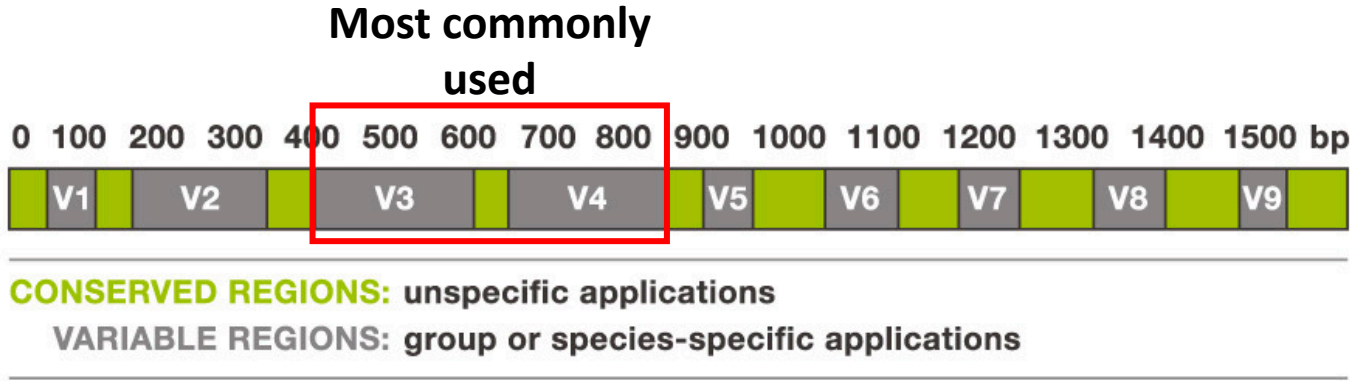
Other genes:

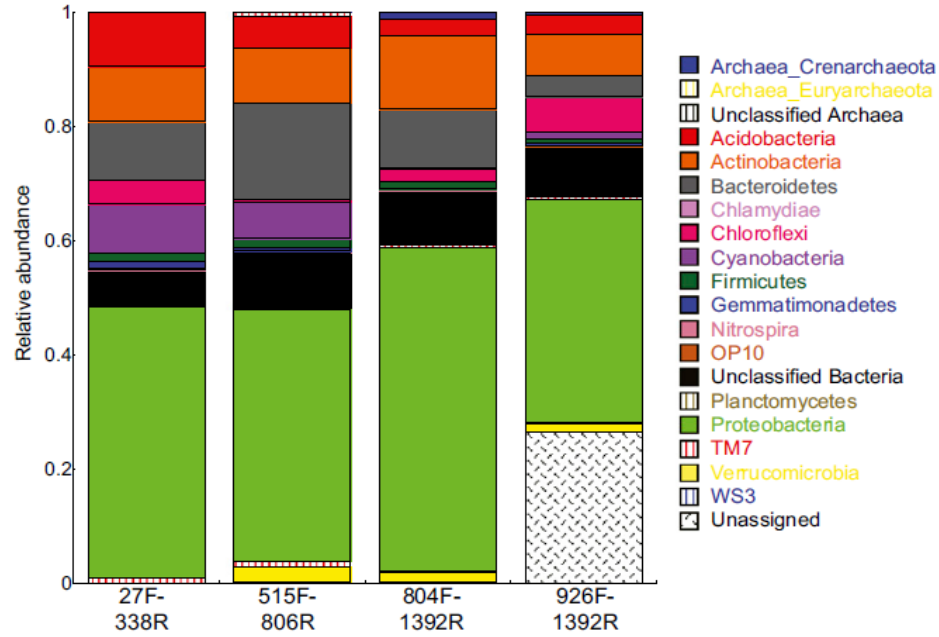
rpoA – codified for RNA polimerase

gyrB – codifies for gyrase protein

(Yang et al., *BMC Bioinformatics,* 2016).

16S rRNA AMPLICON: MICROBIAL COMMUNITY ANALYSIS USING ONLY ONE GENE





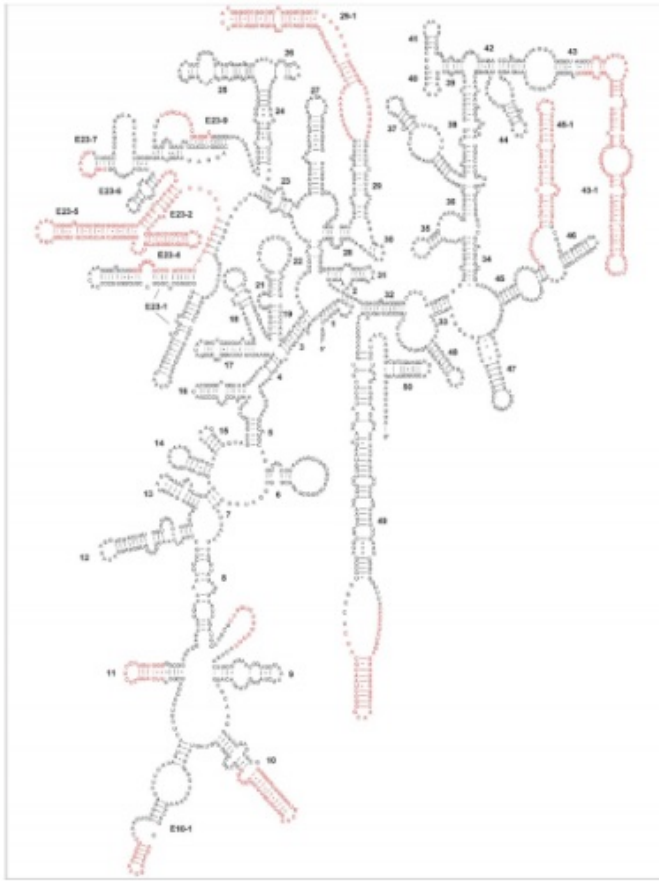
http://www.earthmicrobiome.org/protocols-and-standards

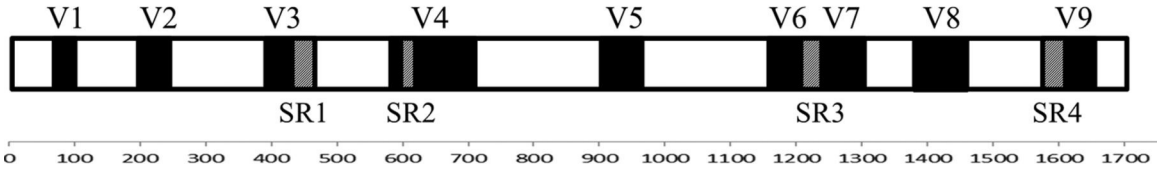
(modified Peiffer et al., *PNAS,* 2013).

18S RIBOSOMAL RNA

- Part of the small subunit in **eukaryotic** ribosomes

  - Basic components of all eukaryotic cells.
  - Structural RNA for the small component of eukaryotic cytoplasmic ribosomes
  - Important for maintaining structure of the small subunit
  - Active center of protein synthesis





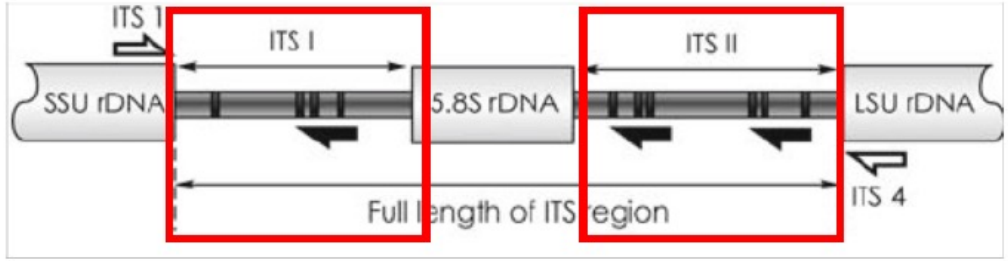
Modified from Ishaq and Wright, 2014

http://www.earthmicrobiome.org/protocols-and-standards

INTERNAL TRANSCRIBED SPACER (ITS)

> Spacer DNA located between the small and large rRNA subunits genes (in the transcribed region)
> Most often used to identify fungi

Eukaryotes



http://www.earthmicrobiome.org/protocols-and-standards

STRENGTHS AND WEAKNESSES OF AMPLICON SEQUENCING

Pros

- ✓ provides a snapshot of the taxonomic diversity
- ✓ inexpensive, can process a lot of samples cheaply
- ✓ works well with low biomass samples and samples with high amounts of host DNA

Cons

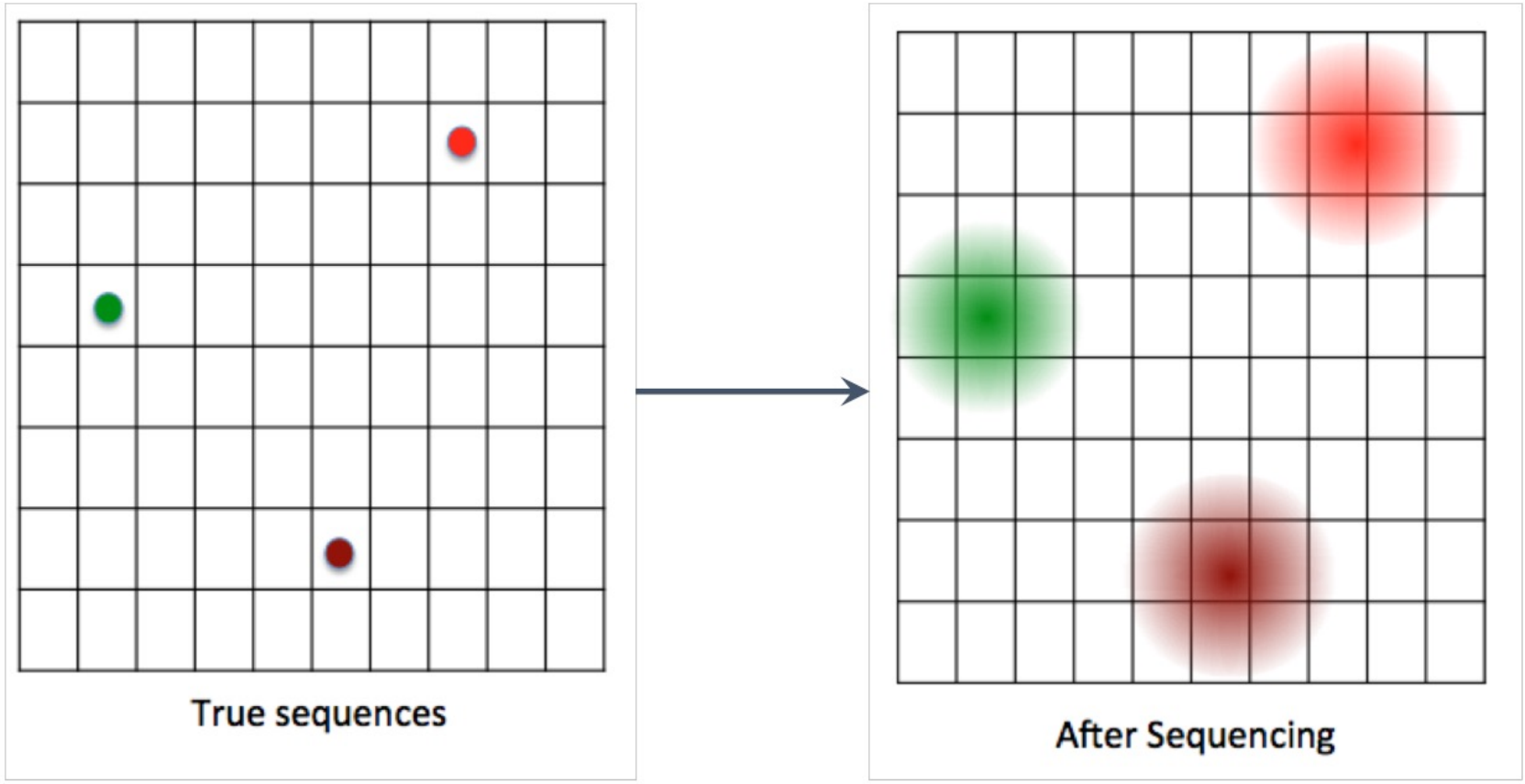
- ✓ not good for strain level identification
- ✓ can be biased based on primer choice, sample preservation methods, and other technical artifacts

LEARNING GOALS

- The tool box
- Deciding what to do …
- 16/ITS/18S rRNA
- Clustering vs ASVs
- Microbial diversity measures (alpha)
- Beta diversity
- Microbiome data analysis
- Differential abundance

WHAT NORMALLY HAPPENS DURING SEQUENCING?

WHAT NORMALLY HAPPENS DURING SEQUENCING?



CLEANING AND MANIPULATING RAW SEQUENCES

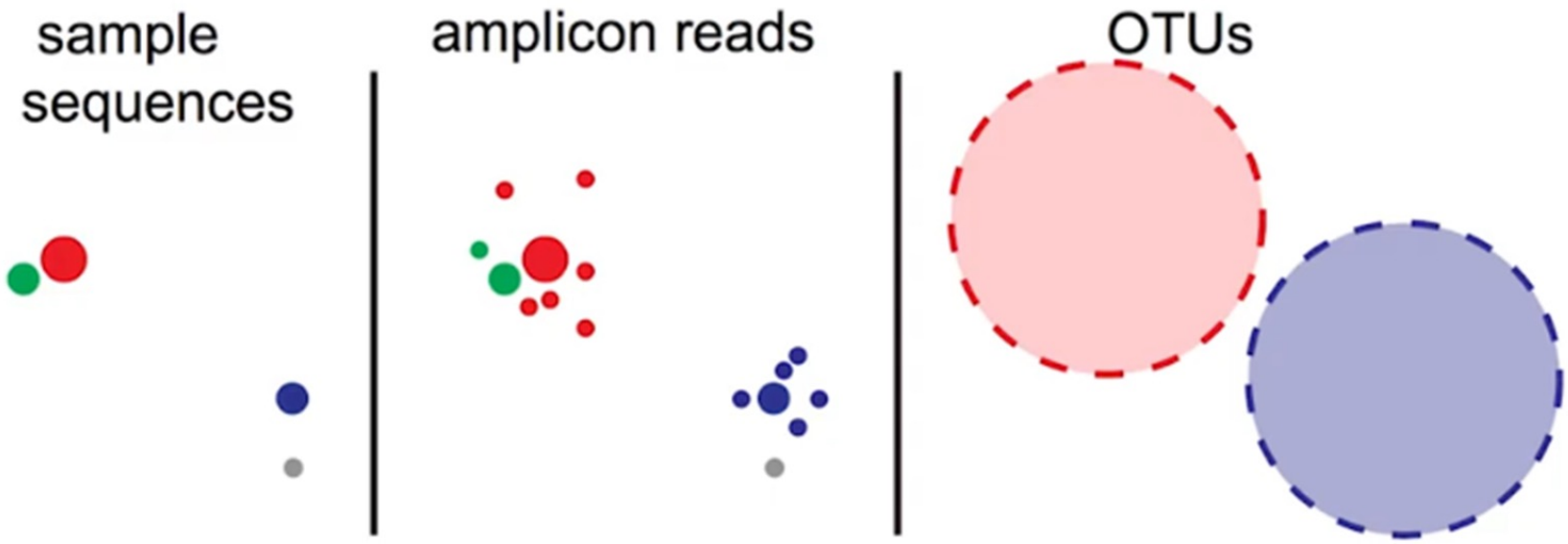
Clustering (OTUs)

- ✓ remove noisy sequences and reduce the amount of sequences to process
- ✓ works based on a given threshold, i.e. 97% similarity
- ✓ There are different methods (1. closed or 2. open reference) and algorithms (UPARSE, uclust, CD-HIT)

Remove noise (ASVs)

- ✓ Find the cleanest sequence
- ✓ Correct and/or discard super noisy sequences
- ✓ Examples are: DADA2 and Deblur

CLUSTERING (Operational Taxonomic Unit, OTU)



97% 16S rRNA sequence identity

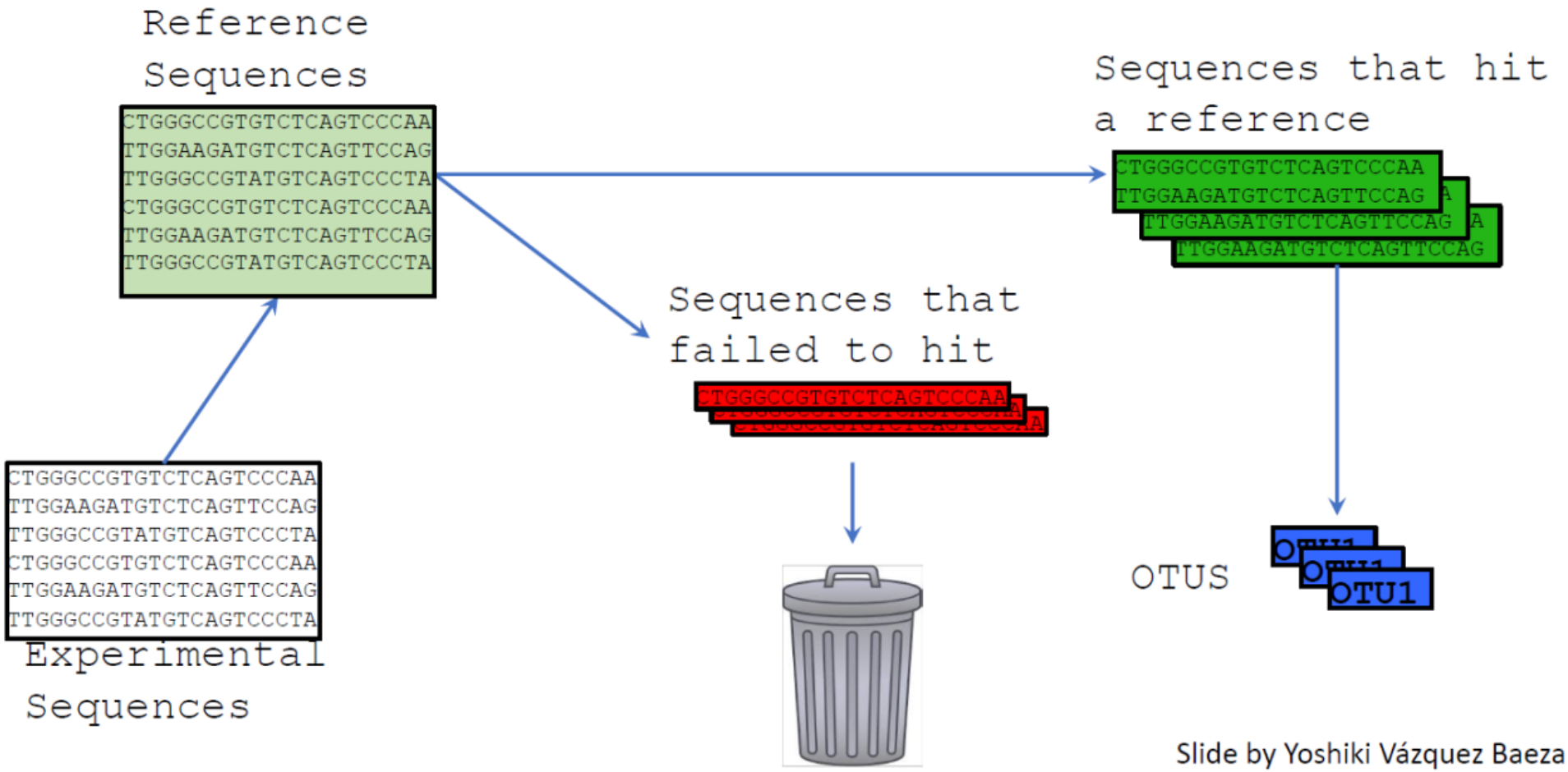
Zheng et al., 2018

CLUSTERING

1. **Closed reference** (must cluster with a database sequence)

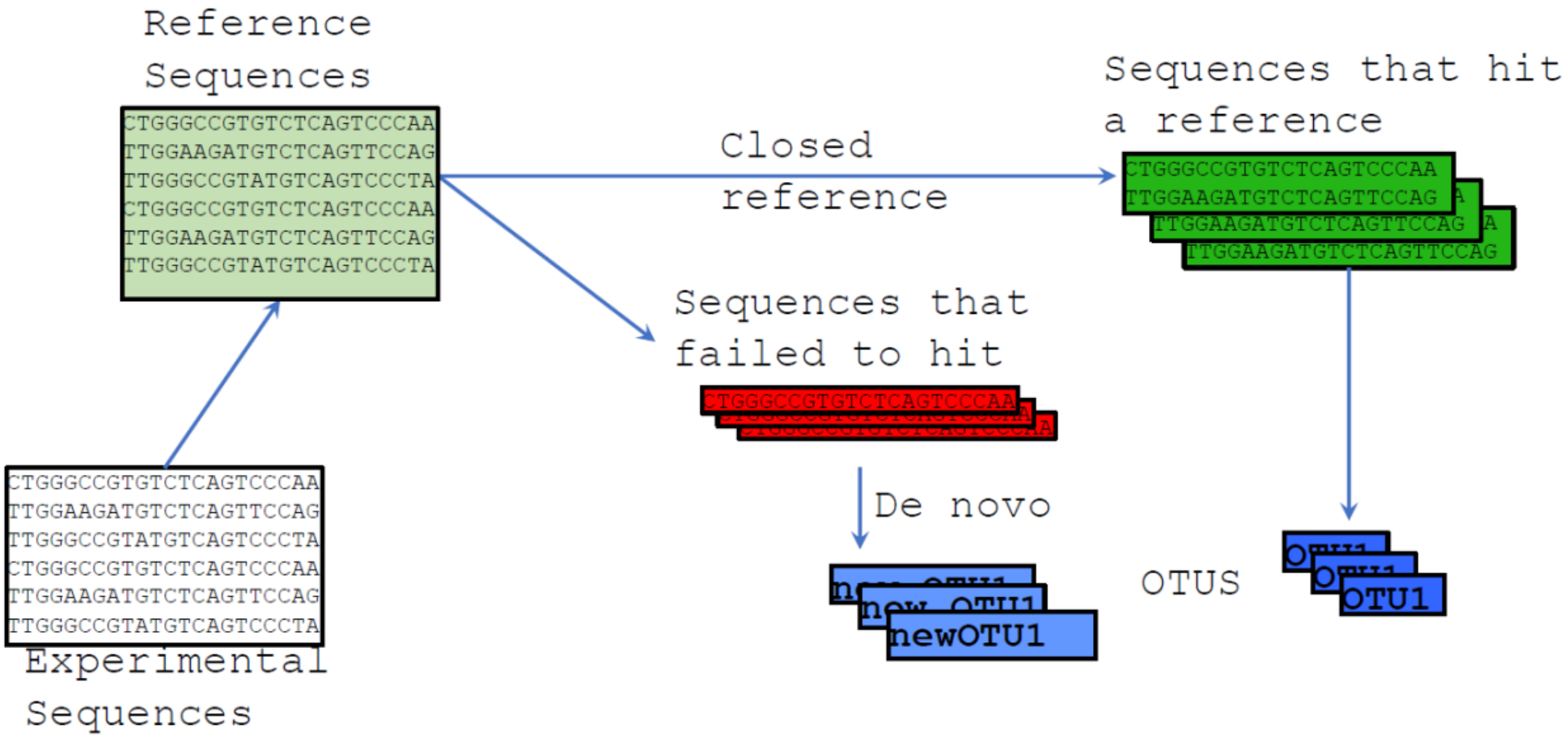
2. Open reference (use database, then *de novo* for sequences not hitting database)

1. CLOSED REFERENCE



Slide by Yoshiki Vázquez Baeza

2. OPEN REFERENCE

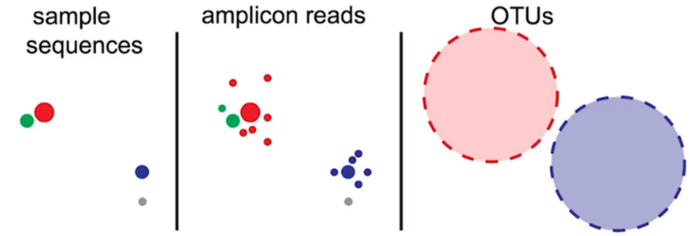


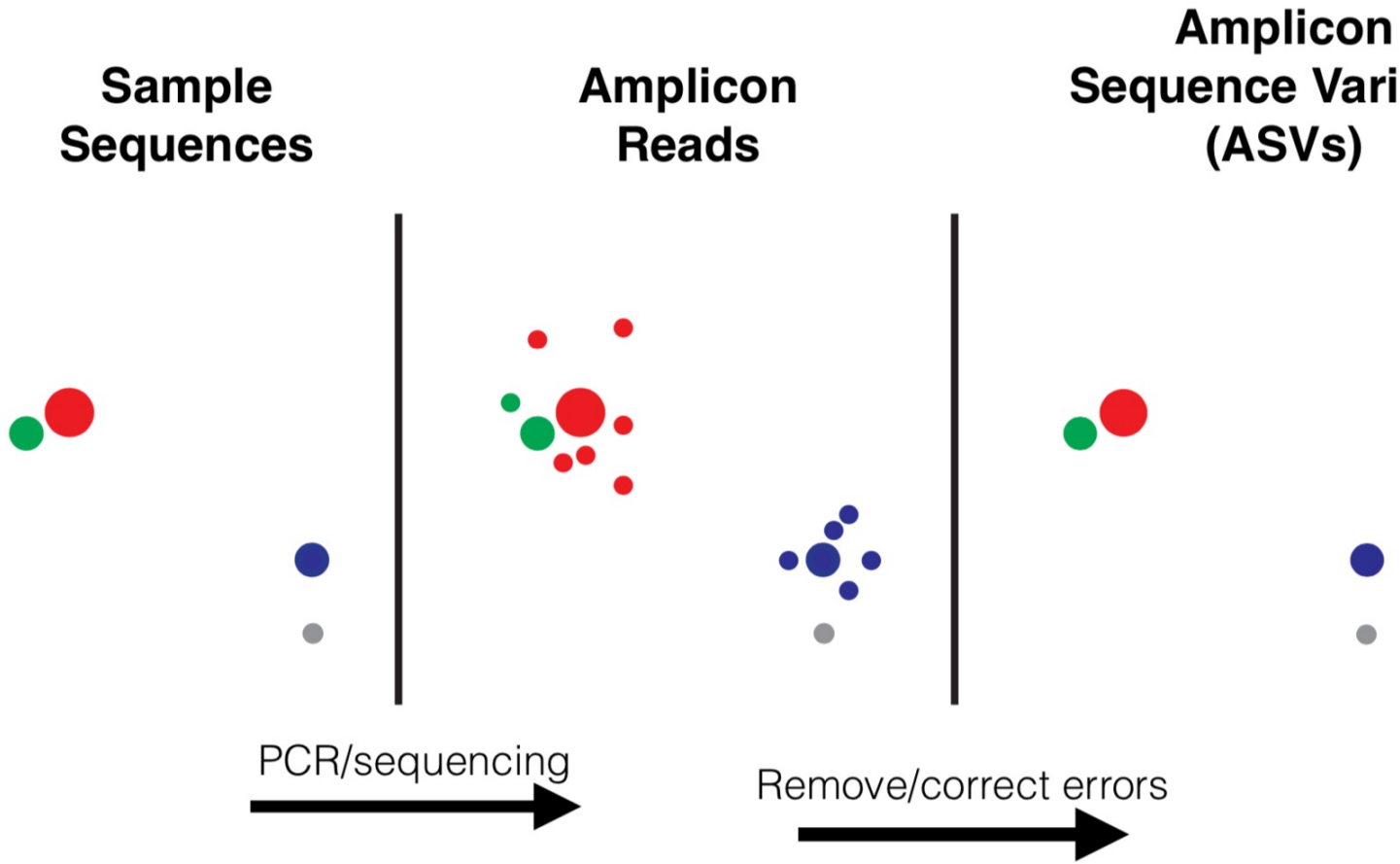
Slide by Yoshiki Vázquez Baeza

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]

AMPLICON SEQUENCE VARIANT (ASV)





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]

AMPLICON SEQUENCE VARIANT (ASV)

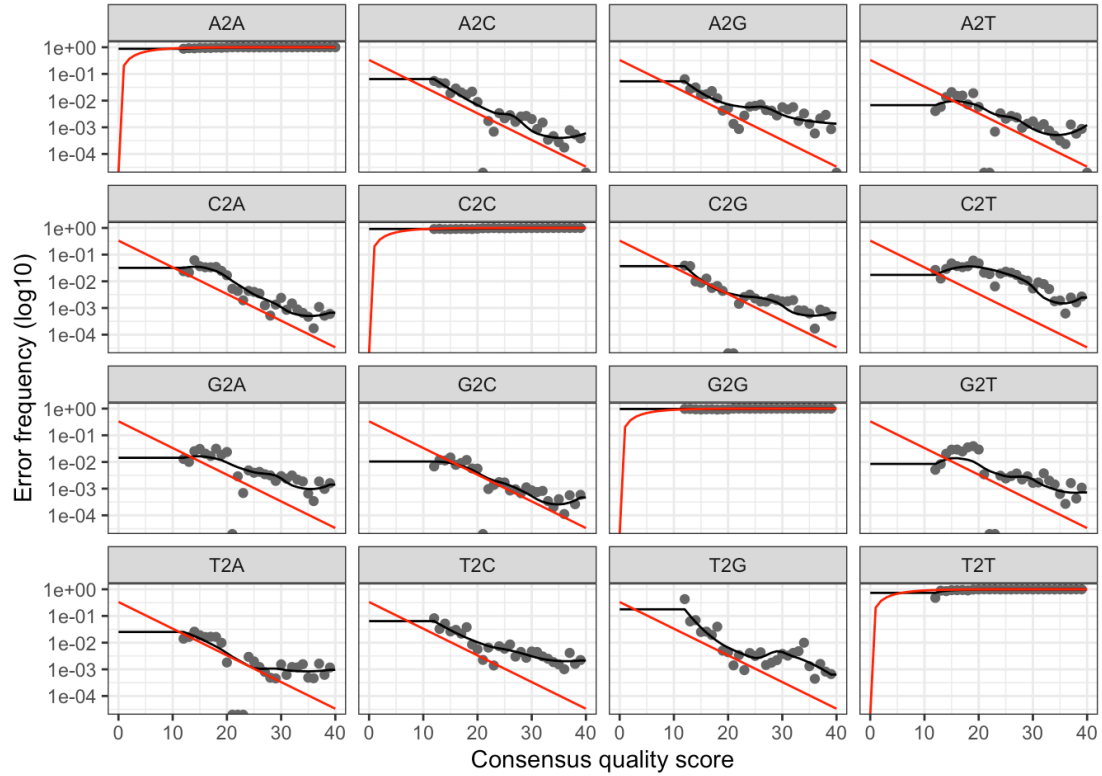
"ASVs are inferred by a de novo process in which biological sequences are discriminated from errors on the basis of, in part, the expectation that biological sequences are more likely to be repeatedly observed than are error-containing sequences."

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]

AMPLICON SEQUENCE VARIANT (ASV)

Accommodate	Acommodate, accomodate
Definitely	Definately, definatly
Embarrass	Embarass, embaras
Occurrence	Occurance, occurence
Separate	Seperate, seperete
Weird	Wierd, weired



OPEN

The ISME Journal (2017) 11, 2639–2643
www.nature.com/ismej

PERSPECTIVE

Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

Benjamin J Callahan[1], Paul J McMurdie[2] and Susan P Holmes[3]

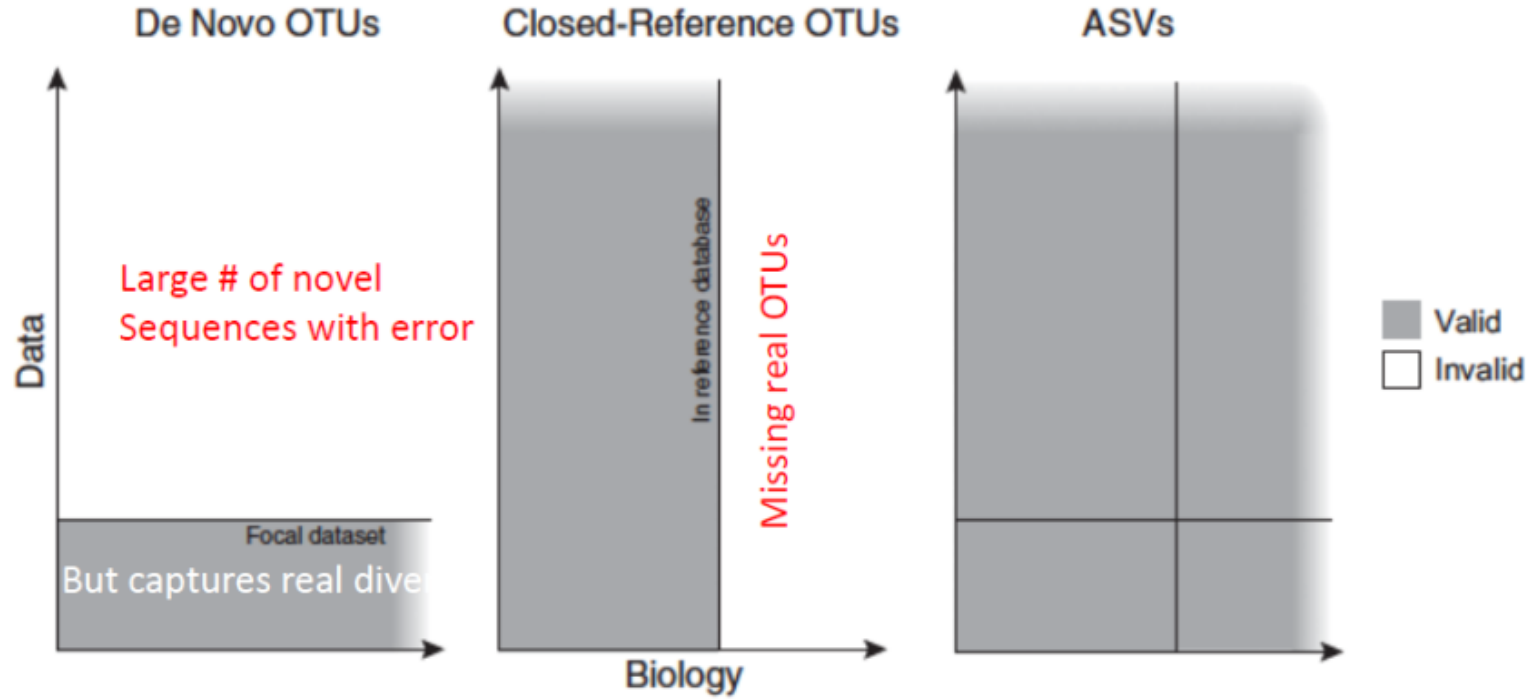


Figure 1 The extent of the validity of *de novo* OTUs, closed-reference OTUs and ASVs determined from a focal data set.

TAXONOMIC CLASSIFICATION

silva

A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.
SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (***Bacteria***, ***Archaea*** **and** ***Eukarya***).

GREENGENES
The 16S rRNA Gene Database and Tools

rdp

NCBI
National Center for
Biotechnology Information

RESEARCH Open Access

SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare?

Monika Balvočiūtė* and Daniel H. Huson

From The Fifteenth Asia Pacific Bioinformatics Conference
Shenzhen, China. 16–18 January 2017

Abstract

Background: A key step in microbiome sequencing analysis is read assignment to taxonomic units. This is often performed using one of four taxonomic classifications, namely SILVA, RDP, Greengenes or NCBI. It is unclear how similar these are and how to compare analysis results that are based on different taxonomies.

Results: We provide a method and software for mapping taxonomic entities from one taxonomy onto another. We use it to compare the four taxonomies and the Open Tree of life Taxonomy (OTT).

Conclusions: While we find that SILVA, RDP and Greengenes map well into NCBI, and all four map well into the OTT, mapping the two larger taxonomies on to the smaller ones is problematic.

Keywords: Metagenomics, Taxonomic classification, OTU assignment, NCBI, Silva, RDP, Greengenes, Open tree of life

Background

Microbiome sequencing analysis is concerned with sequencing DNA from microorganisms living in certain environments without cultivating them in laboratory. In a typical taxonomy guided approach [1], sequencing reads are first binned into taxonomic units and then the microbial composition of samples is analyzed and compared in detail (see Fig. 1).

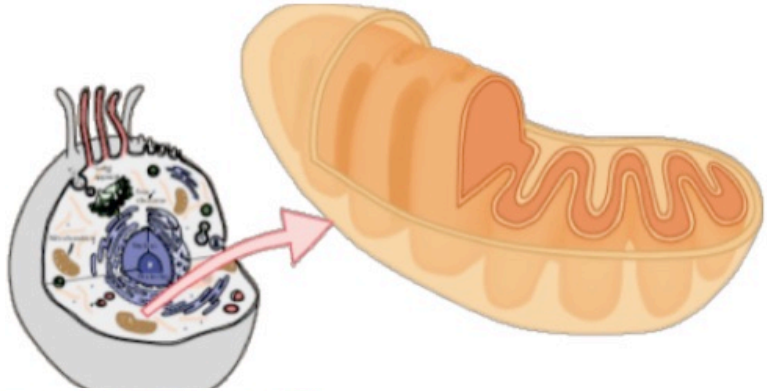
The two main technical ingredients of taxonomic analysis are the reference taxonomy used and the binning approach employed. Binning is usually performed either by aligning reads against reference sequences (e.g. [2]) or using k-mer based techniques (e.g. [3]). Taxonomic binning of 16S reads is usually based on one of these four taxonomies: SILVA [4], RDP [5], Greengenes [6] or NCBI [7].

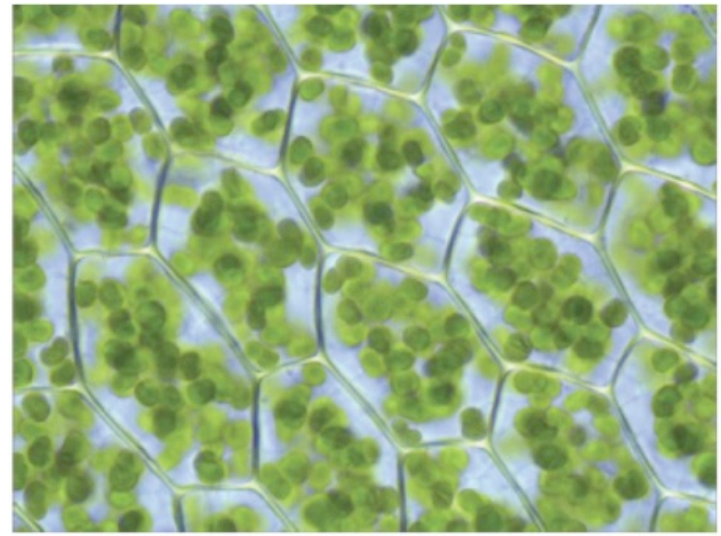
whether results obtained using one classification can easily be carried over to another.

We define and explore an algorithm for mapping one taxonomy into another. This method allows us to compare taxonomies and is the basis for a tool that makes analyses on different classifications comparable to each other by mapping them onto a common taxonomy. While our main focus is on the four most popular taxonomic trees, we also consider the recently published Open Tree of life Taxonomy (OTT) [9].

We found that SILVA, RDP and Greengenes can be mapped into NCBI and OTT with few conflicts, but not vice versa. There is a great deal of difference between taxonomies that arise because of the differences in size and structure.

ARE ALL THE SEQUENCES IDENTIFIED PART OF THE MICROBIOME?





Chloroplast 16S rRNA sequences:
k__Bacteria;p__Cyanobacteria;c__Chloroplast;o__Chlorophyta;f__;g__
k__Bacteria;p__Cyanobacteria;c__Chloroplast;o__Chlorophyta;f__Trebouxiophyceae;g__
k__Bacteria;p__Cyanobacteria;c__Chloroplast;o__Streptophyta;f__;g__

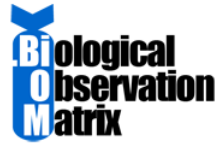
Mitochondrial 16S sequences:
k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rickettsiales;f__mitochondria

OTU/ASV TABLES

biom-format.org

THE BIOLOGICAL OBSERVATION MATRIX (BIOM) FORMAT

Contents :: BIOM Documentation »



Y67 | fx | 15

	A	B	C	D	E	F	G	H	I	J	BP	BQ	BR	BS	BT	BU	BV	BW	BX
1	#OTU ID	Wild_C_a	Variety_l	Wild_A_a	Wild_A_a	Wild_C_a	Wild_A_a	Wild_A_a	Variety_l	Wild_C_n	Domain	Phylum	Class	Order	Family	Genus			
2	OTU_5	325	0	587	485	760	122	344	594	0	Bacteria	Proteoba	Alphaprc	Rhizobia	Rhizobia	Rhizobium			
3	OTU_6528	3	0	3	0	0	0	0	2	0	Bacteria	Proteoba	Alphaprc	Rhizobia	Rhizobia	Rhizobium			
4	OTU_16182	4	0	1	0	2	0	5	0	0	Bacteria	Verrucom	Subdivisi	unclassit	unclassit	Subdivision3_genera_incertae_sedis			
5	OTU_9	425	705	173	446	588	278	918	315	506	Bacteria	Proteoba	Alphaprc	Sphingor	Sphingor	unclassified_Sphingomonadaceae			
6	OTU_3449	4	0	1	1	1	0	4	1	0	Bacteria	Candidat	unclassit	unclassit	unclassit	Saccharibacteria_genera_incertae_sedis			
7	OTU_67	100	4	196	133	98	44	97	144	4	Bacteria	Proteoba	Betaprot	Burkhold	Burkhold	Cupriavidus			
8	OTU_14268	52	104	12	45	85	43	39	25	81	Bacteria	Acidobac	Acidobac	unclassit	unclassit	unclassified_Acidobacteria_Gp3			
9	OTU_15790	60	0	108	86	88	77	172	73	1	Bacteria	Proteoba	Betaprot	Burkhold	unclassit	unclassified_Burkholderiales			
10	OTU_1	5130	520	21658	6767	6185	7335	5342	18475	425	Bacteria	Proteoba	Alphaprc	Rhizobia	Rhizobia	Rhizobium			
11	OTU_1430	11	0	2	1	6	0	2	5	0	Bacteria	unclassit	unclassit	unclassit	unclassit	unclassified_Bacteria			
12	OTU_14794	117	149	25	35	119	32	133	48	131	Bacteria	Acidobac	Acidobac	unclassit	unclassit	Gp1			
13	OTU_642	18	0	1	18	18	6	20	13	0	Bacteria	unclassit	unclassit	unclassit	unclassit	unclassified_Bacteria			
14	OTU_15481	79	0	21	8	82	0	29	82	0	Bacteria	Proteoba	Betaprot	Burkhold	unclassit	unclassified_Burkholderiales			
15	OTU_10734	12	0	3	6	15	8	19	9	0	Bacteria	Acidobac	Acidobac	unclassit	unclassit	Gp6			
16	OTU_777	25	0	9	4	9	3	0	15	0	Bacteria	Bacteroic	unclassit	unclassit	unclassit	unclassified_Bacteroidetes			
17	OTU_2977	2	0	0	3	0	0	3	1	0	Bacteria	Acidobac	Acidobac	unclassit	unclassit	Gp4			
18	OTU_943	9	0	3	13	11	7	27	5	0	Bacteria	unclassit	unclassit	unclassit	unclassit	unclassified_Bacteria			
19	OTU_1450	81	45	27	84	76	36	114	26	35	Bacteria	Acidobac	Acidobac	unclassit	unclassit	Gp1			
20	OTU_13439	623	41	1347	398	582	470	532	1284	19	Bacteria	Proteoba	Alphaprc	Rhizobia	Rhizobia	Rhizobium			
21	OTU_53	105	0	188	124	102	141	156	118	0	Bacteria	Actinoba	Actinoba	Actinomy	Nocardio	Aeromicrobium			
22	OTU_2472	10	0	1	1	7	3	6	2	0	Bacteria	Acidobac	Acidobac	unclassit	unclassit	Gp6			
23	OTU_9620	1	0	0	0	0	0	0	0	0	Bacteria	unclassit	unclassit	unclassit	unclassit	unclassified_Bacteria			
24	OTU_15861	1	0	0	0	0	0	0	0	0	Bacteria	unclassit	unclassit	unclassit	unclassit	unclassified_Bacteria			
25	OTU_10113	1	0	0	0	4	1	1	4	0	Bacteria	Acidobac	Acidobac	unclassit	unclassit	Gp6			
26	OTU_7	595	42	349	294	1124	255	511	1860	69	Bacteria	Actinoba	Actinoba	Actinomy	Streptom	Streptomyces			
27	OTU_13170	445	4	360	226	410	160	668	511	1	Bacteria	Proteoba	Alphaprc	Rhizobia	Rhizobia	Rhizobium			
28	OTU_387	12	42	0	0	5	0	4	1	51	Bacteria	Proteoba	Gammap	unclassit	unclassit	unclassified_Gammaproteobacteria			
29	OTU_9047	979	1	568	196	1124	165	705	1945	0	Bacteria	Proteoba	Alphaprc	Rhizobia	Rhizobia	Rhizobium			
30	OTU_15070	54	22	15	30	66	18	63	49	13	Bacteria	Proteoba	Betaprot	unclassit	unclassit	unclassified_Betaproteobacteria			

All samples | workotutable

LEARNING GOALS

- The tool box
- Deciding what to do …
- 16/ITS/18S rRNA
- Clustering vs ASVs
- Microbial diversity measures (alpha)
- Beta diversity
- Microbiome data analysis
- Differential abundance

MICROBIAL DIVERSITY

Comparing microbial communities

DIVERSITY MEASURES

- What is there? How much is there?

Alpha diversity: within sample.

How many different OTUs/ASVs/ESVs are there? Alpha diversity

richness (observed OTUs)

evenness (Shannon)

- How similar or different are samples?

Beta diversity: between samples.

DIVERSITY MEASURES

What is there? How much is there?

Alpha diversity: within sample.

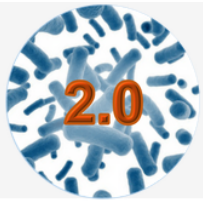
- **Alpha diversity and species richness**: Number of species in a given sample
- Shannon: How even are species abundances distributed?
- **Phylogenetic diversity**: The phylogenetic distance of the observed sequences
- Coverage: The estimated proportion of total diversity observed in a given dataset
- **Functional diversity: In genes or processes**

TOOLS FOR THE ANALYSIS OF MICROBIOMES…

nature protocols PROTOCOL
https://doi.org/10.1038/s41596-019-0264-1

Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data

Jasmine Chong[1], Peng Liu[1], Guangyan Zhou[1] and Jianguo Xia[1,2,3,4*]



MicrobiomeAnalyst -- comprehensive statistical, functional and integrative analysis of microbiome data

Home | Formats | Forum | Updates | Resources | Contact

Marker Data Profiling Analyze marker gene counts data	**Shotgun Data Profiling** Analyze shotgun metagenomics data	**Taxon Set Analysis** Discover enriched microbial signatures
Microbiome Metabolomics Co-analyze microbiome & metabolomics data	**Statistical Meta-analysis** Integrate multiple marker gene data	**Raw Data Processing** Convert raw 16S reads to ASV table

DIVERSITY MEASURES: ALPHA DIVERSITY
METHOD 1 -> SPECIES COUNT

non-phylogenetic, alpha diversity metric measuring richness

Plant A

Pseudomonas aeruginosa
Pseudomonas fluorescens
Pseudomonas putida
Escherichia coli

Plant B

Pseudomonas aeruginosa
Pseudomonas fluorescens

Plant C

Pseudomonas aeruginosa

Observed species

Plant A: 4
Plant B: 2
Plant C: 1

Slide adapted from the Rob knight Lab

SPECIES COUNT FAILS TO CAPTURE RELATEDNESS

Plant A

Pseudomonas aeruginosa
Pseudomonas fluorescens
Pseudomonas putida

Plant B

Pseudomonas aeruginosa
Pseudomonas fluorescens
Escherichia coli

Plant C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

Observed species

Plant A -> 3
Plant B -> 3
Plant C -> 3

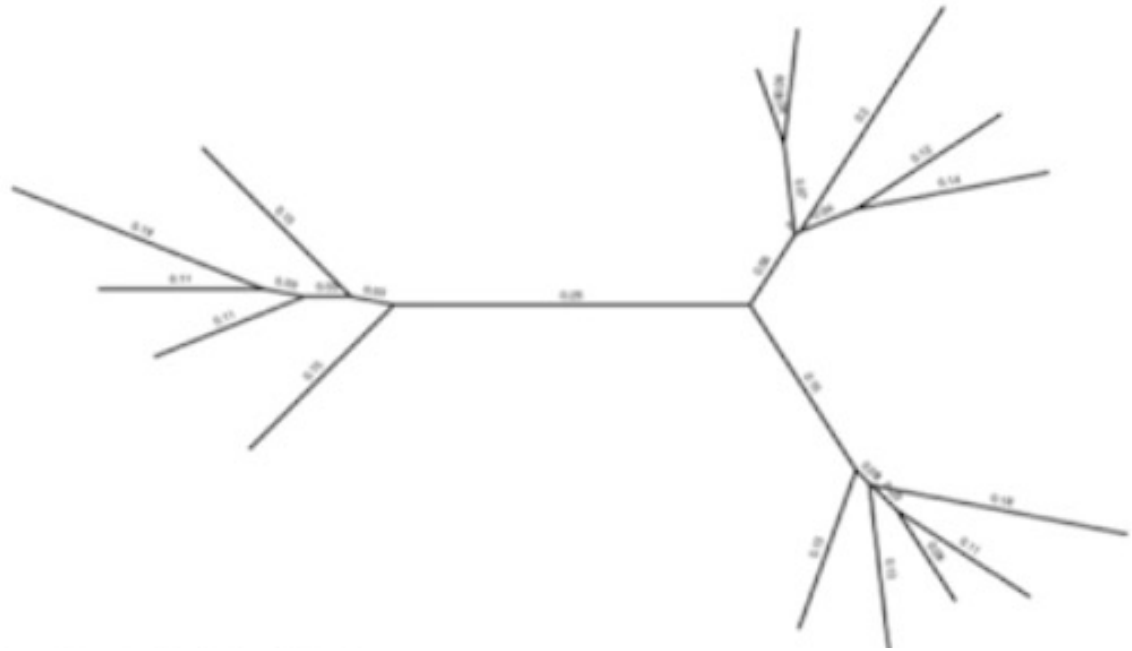
Conclusion
Plant A, B and C are equally diverse

METHOD 2: PHYLOGENETIC DIVERSITY

Plant A
Pseudomonas aeruginosa
Pseudomonas fluorescens
Pseudomonas putida

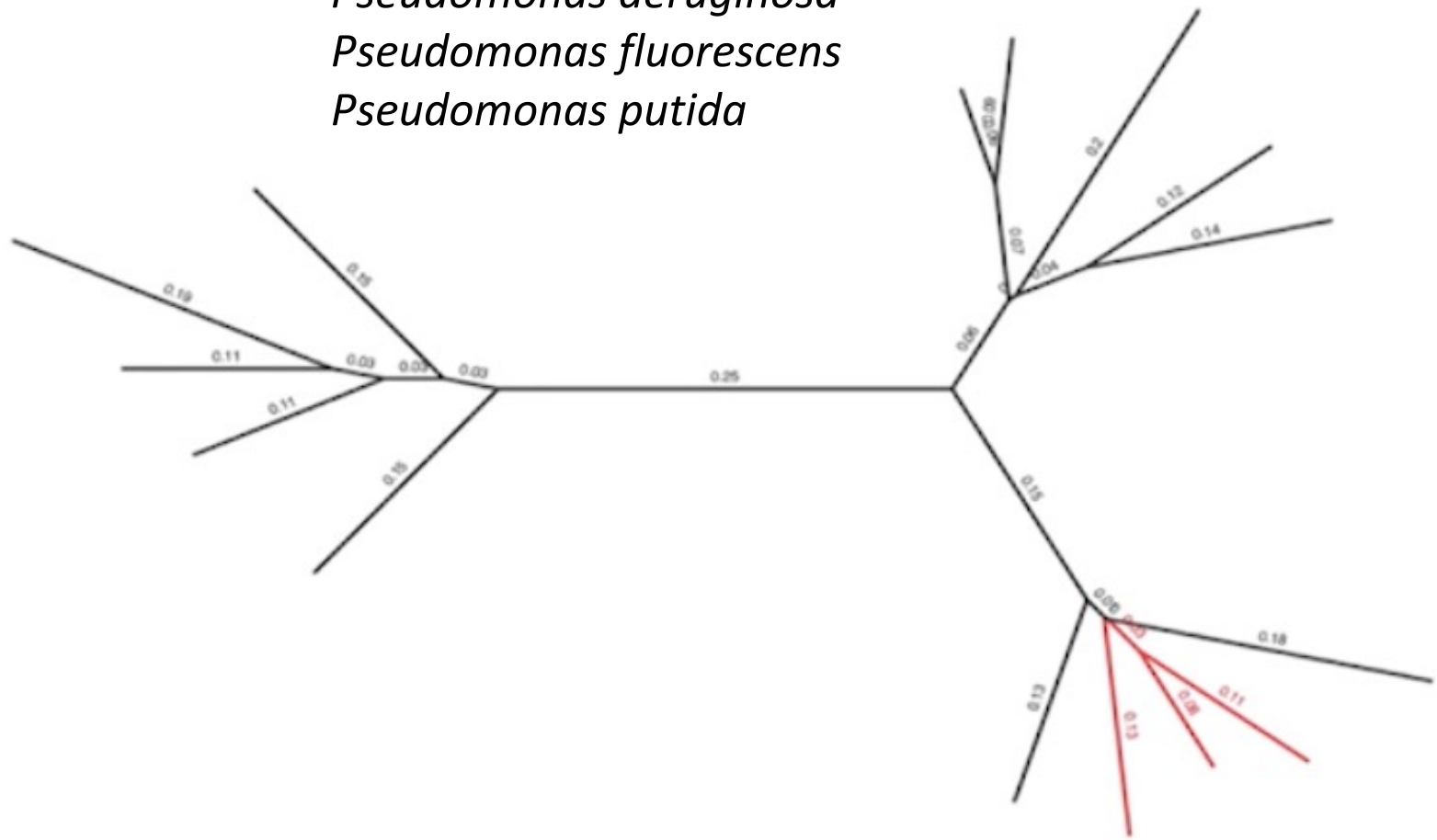
Plant B
Pseudomonas aeruginosa
Pseudomonas fluorescens
Escherichia coli

Plant C
Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii



Plant A

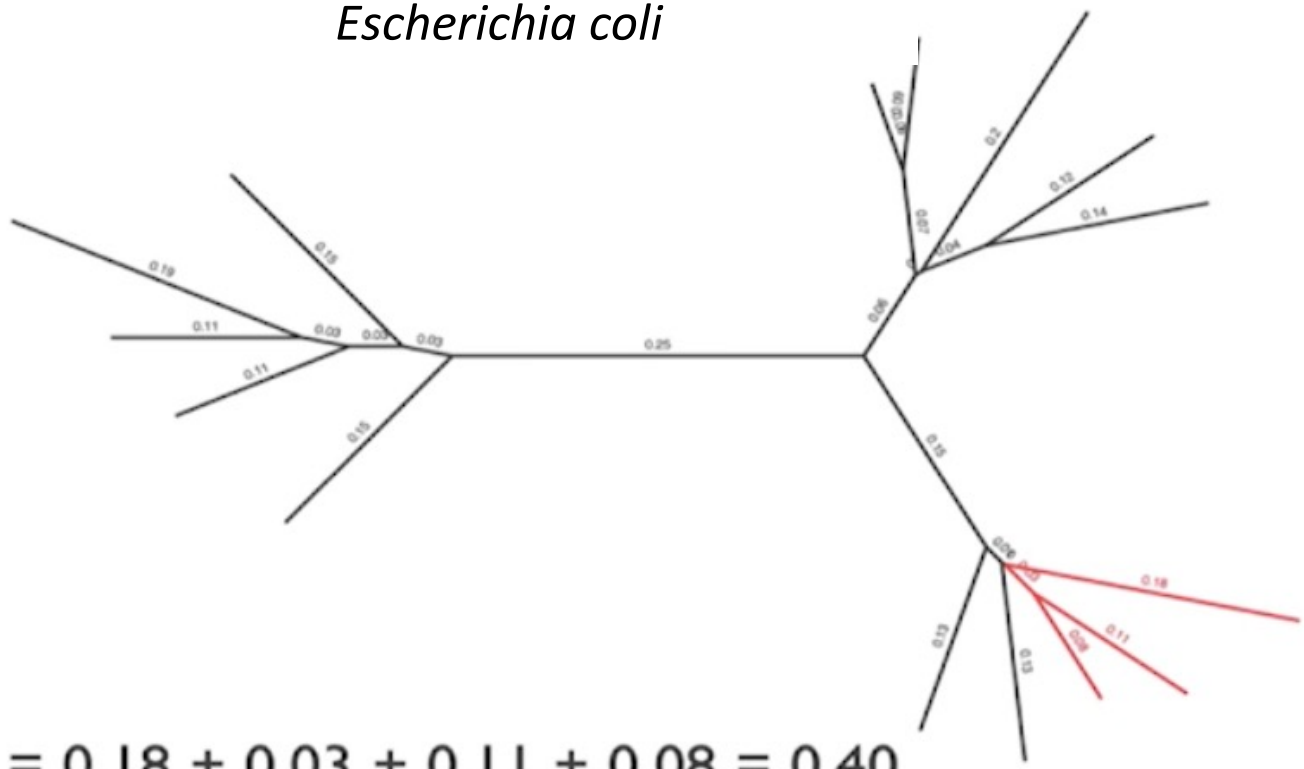
Pseudomonas aeruginosa
Pseudomonas fluorescens
Pseudomonas putida



PD = 0.13 + 0.03 + 0.11 + 0.08 = 0.35

Plant B

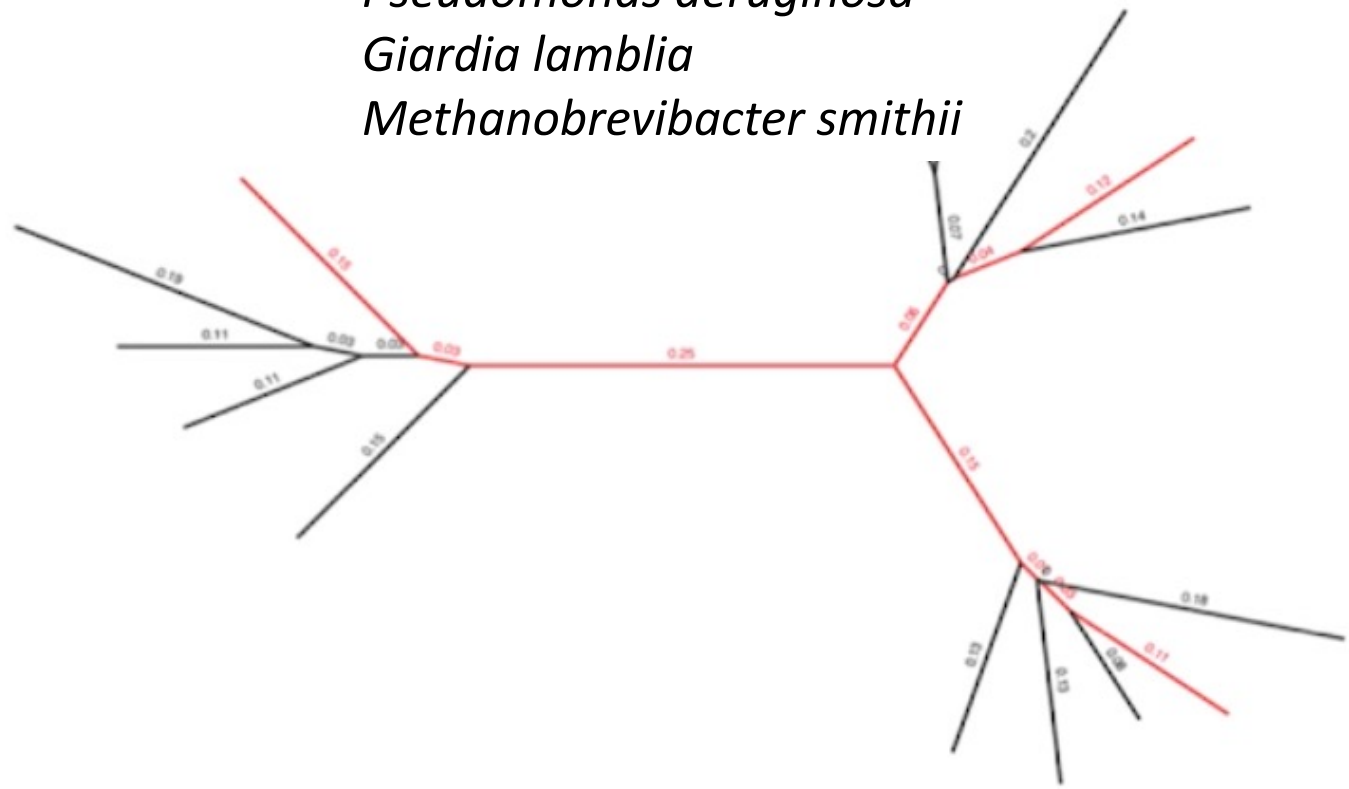
Pseudomonas aeruginosa
Pseudomonas fluorescens
Escherichia coli



PD = 0.18 + 0.03 + 0.11 + 0.08 = 0.40

Plant C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii



PD = 0.15+0.03+0.25+0.06+0.04+0.12+0.15+0.01+0.03+0.11= 0.95

METHOD 2: PHYLOGENETIC DIVERSITY

Plant A
Pseudomonas aeruginosa
Pseudomonas fluorescens
Pseudomonas putida

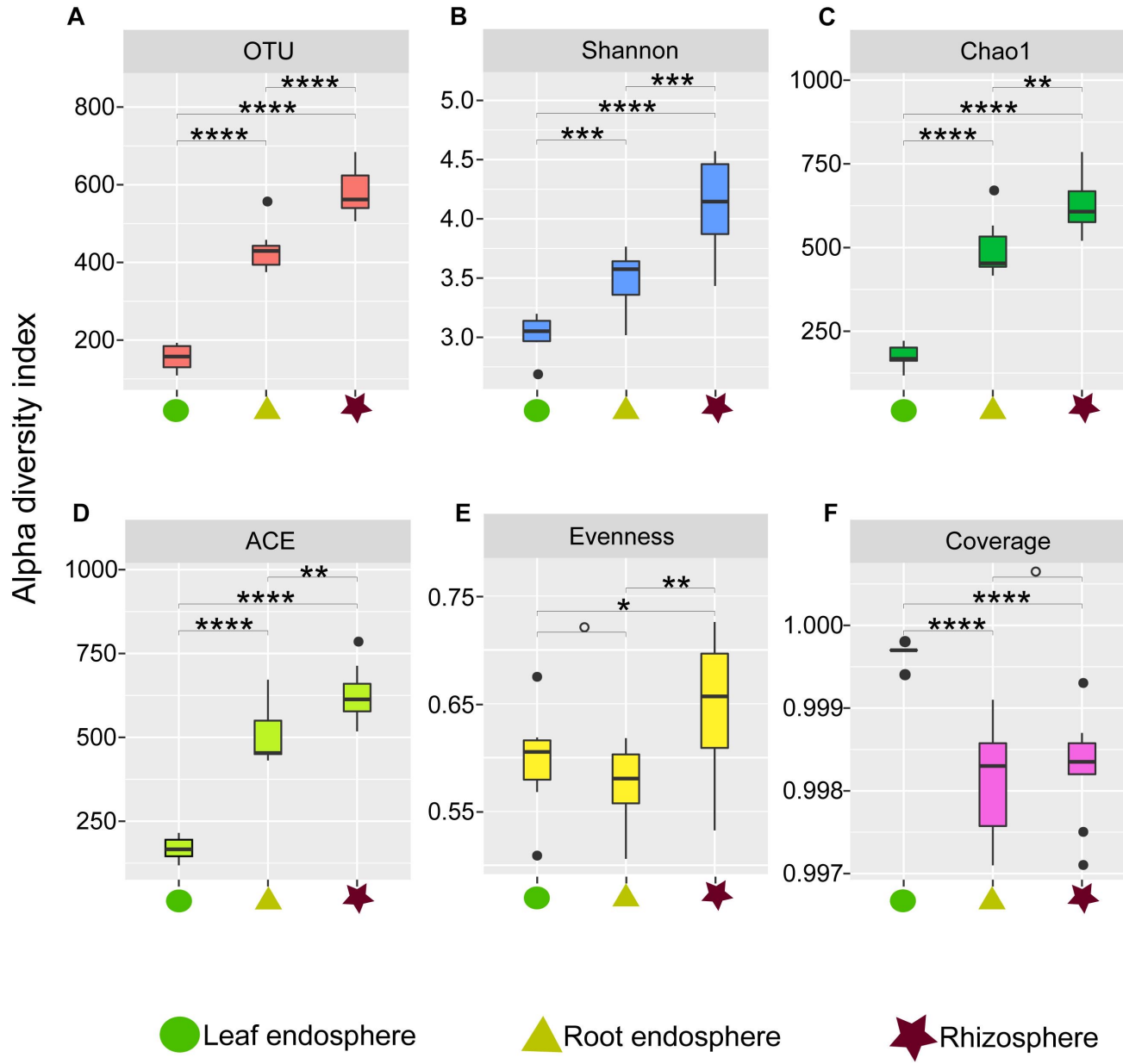
Plant B
Pseudomonas aeruginosa
Pseudomonas fluorescens
Escherichia coli

Plant C
Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

PD = 0.35 < PD = 0.40 < PD = 0.95

Sample C is more diverse than sample B, which is more diverse than sample A

EXAMPLES



Quian et al., 2019

LEARNING GOALS

- The tool box
- Deciding what to do …
- 16/ITS/18S rRNA
- Clustering vs ASVs
- Microbial diversity measures (alpha)
- Beta diversity
- Microbiome data analysis
- Differential abundance
- Visualization

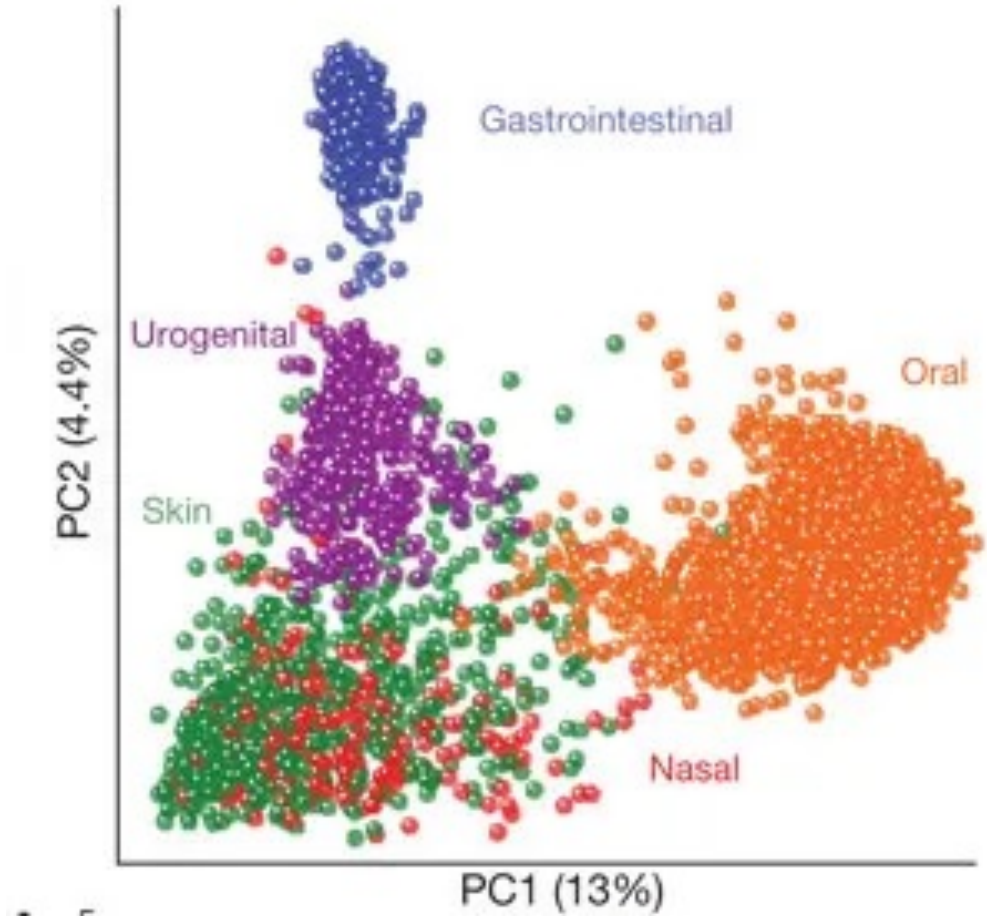
DIVERSITY MEASURES

- How similar or different are samples?

Beta diversity: between samples.

BETA DIVERSITY: PRINCIPAL COORDINATES ANALYSIS (PCoA)

Dimension reduction plot to map distance metric between samples

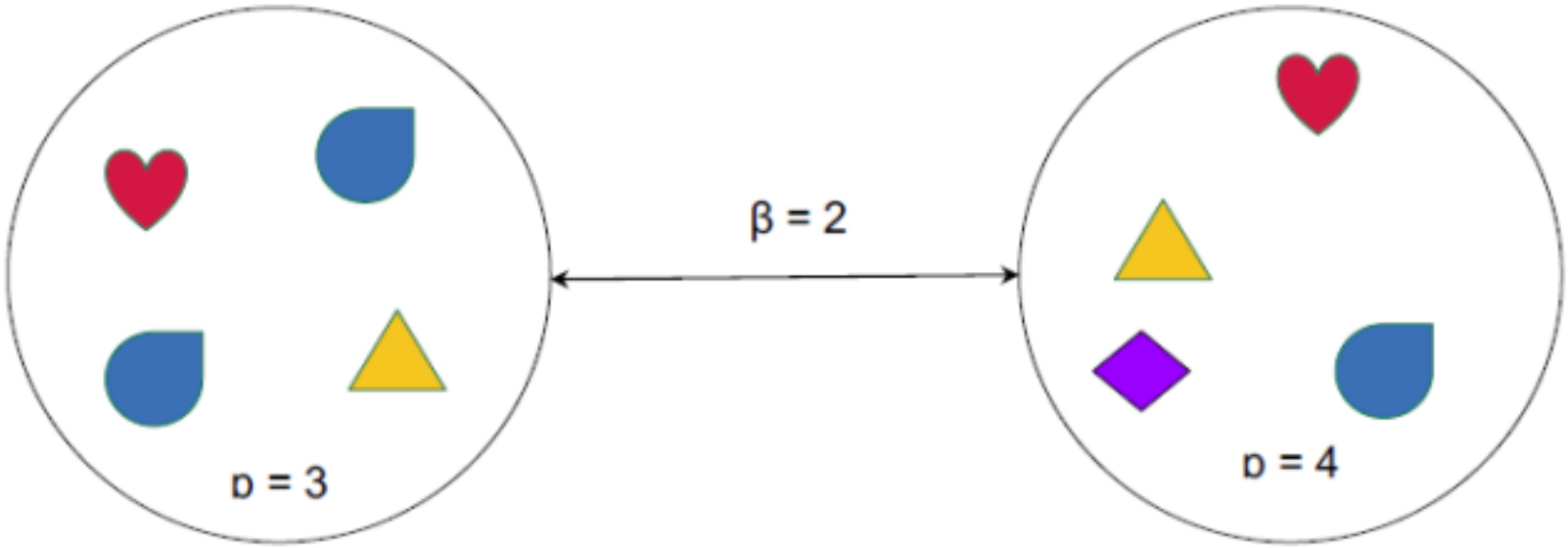


BETA DIVERSITY

Difference in microbiome composition between samples.

> Difference in microbiome composition between samples measured using distance metrics

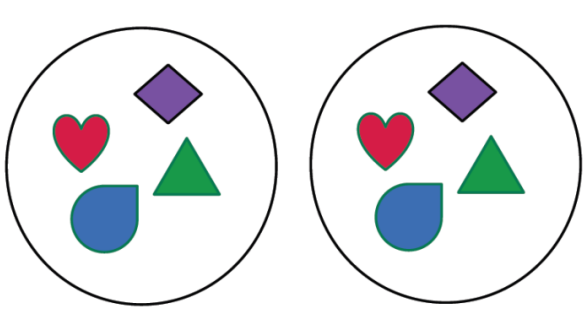
> Dependent on what samples you are comparing



EMBL-EBI

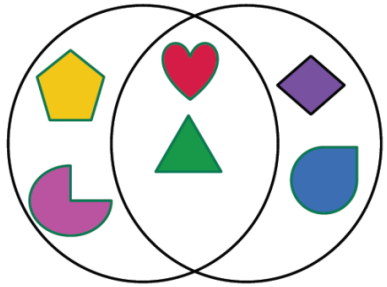
BETA DIVERSITY: JACCARD DISTANCE

Measure of dissimilarity. Does not consider abundance



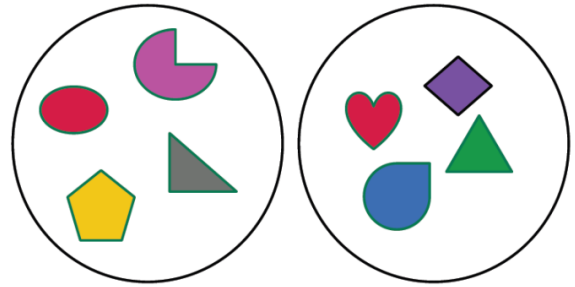
$d_j = 0$

(100 % similarity)



$d_j = 0.5$

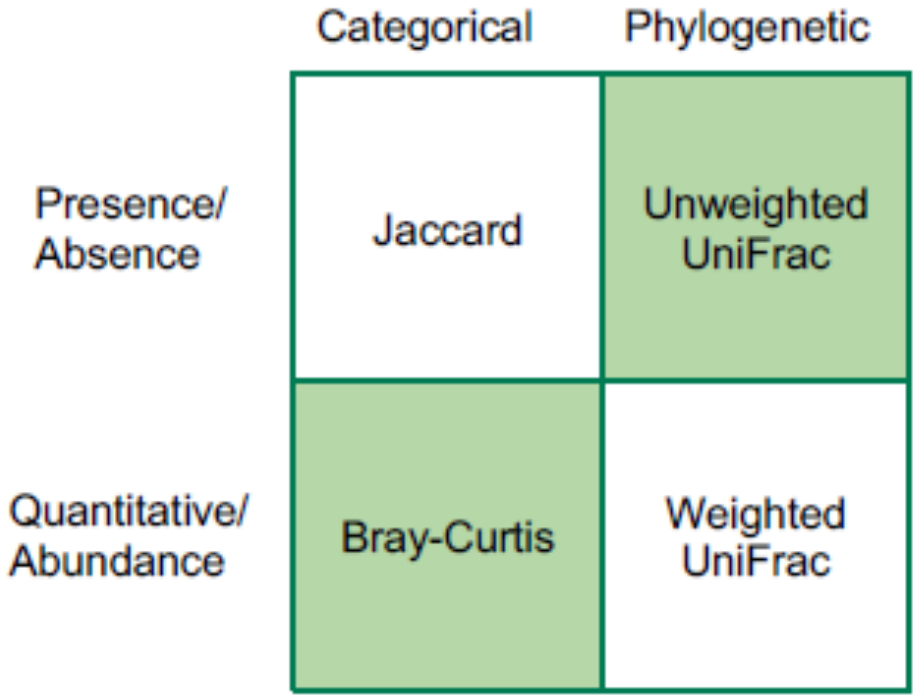
(50 % similarity)



$d_j = 1$

(0 % similarity)

BETA DIVERSITY

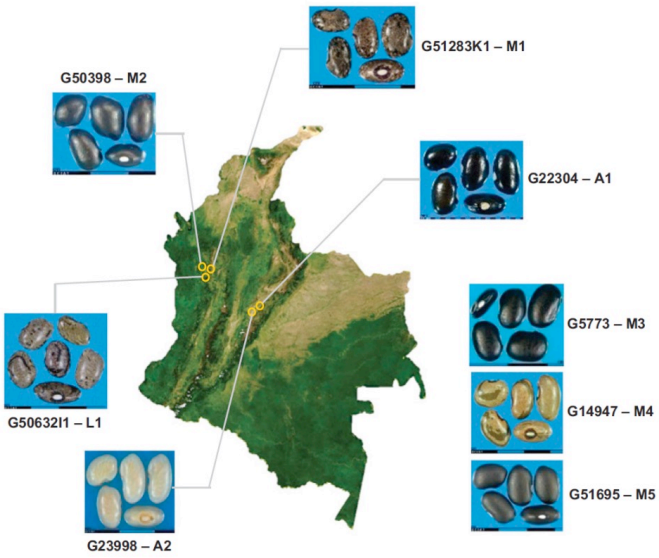


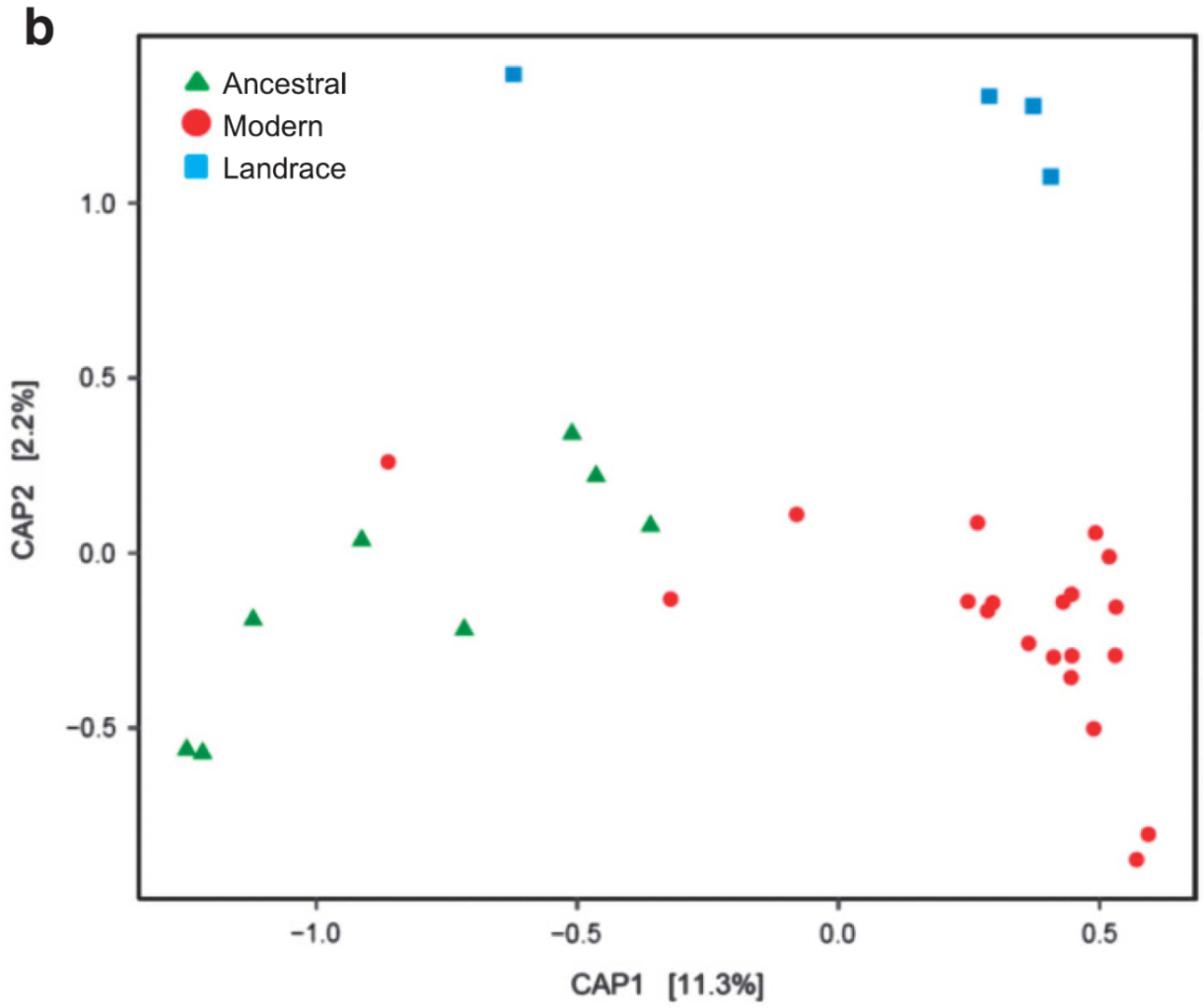
Other Distance metrics

Manhattan
Euclidean
Canberra
Bray
Kulczynski
Gower
mountford

Kowalska-Duplaga (2019)

BETA DIVERSITY: EXAMPLE





Pérez-Jaramillo et al., 2017

LEARNING GOALS

- The tool box
- Deciding what to do …
- 16/ITS/18S rRNA
- Clustering vs ASVs
- Microbial diversity measures (alpha)
- Beta diversity
- Microbiome data analysis
- Differential abundance