Bioinformatics Concept Course: Metagenomics I

Lecture 5:

October 25th, 2021

Shinichi Sunagawa

Microbiome Research Group

Institute of Microbiology, D-BIOL, ETH Zürich

Course schedule

Lecturer	Topic
Sunagawa	Introduction
Kahles	Genomics I
Kahles	Genomics II
Kahles	Genomics III
Sunagawa	Metagenomics I
Sunagawa	Metagenomics II
Sunagawa	Metagenomics III
von Mering	Network bioinformatics I
Gstaiger	Network bioinformatics II
Zamboni	Network bioinformatics III
Snijder	Imaging
Baudis	Ethics
Sunagawa, Rätsch, TBD	Application Symposium

Shinichi Sunagawa
Institute of Microbiology
D-BIOL, ETHZ

Nicola Zamboni
Institute of Molecular Systems Biology
D-BIOL, ETHZ

Christian von Mering
Institute of Molecular Life Sciences
University of Zürich

Andre Kahles
Biomedical Informatics
D-INFK, ETHZ

Berend Snijder
Institute of Microbiology
D-BIOL, ETHZ

Matthias Gstaiger
Institute of Molecular Systems Biology
D-BIOL, ETHZ

Alessandro Blasimme
Institute of Translational Medicine
D-HEST, ETHZ

Course overview – from genomics to metagenomics

In previous lectures/practical courses, you have already learned about:

- Polymerase Chain Reaction (PCR), taxonomy, phylogenetics, multidimensional data

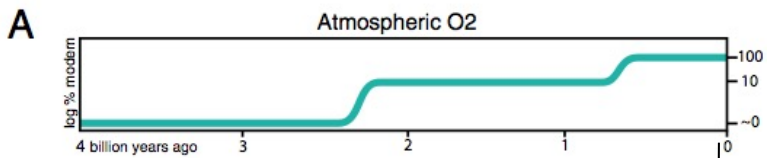
In the genomics section, you have:

- learned how DNA/RNA sequences are generated
- discussed the differences between sequencing technologies
- learned how reads are aligned to a reference sequence
- studied how read counts are used to quantify genomic features (e.g., transcripts)
- understood the need for multiple test corrections

→ The metagenomics section builds upon these concepts and extends them as they are pertinent to microbial communities

Evolution and significance of microbiomes

From the origin of life to today



Microorganisms

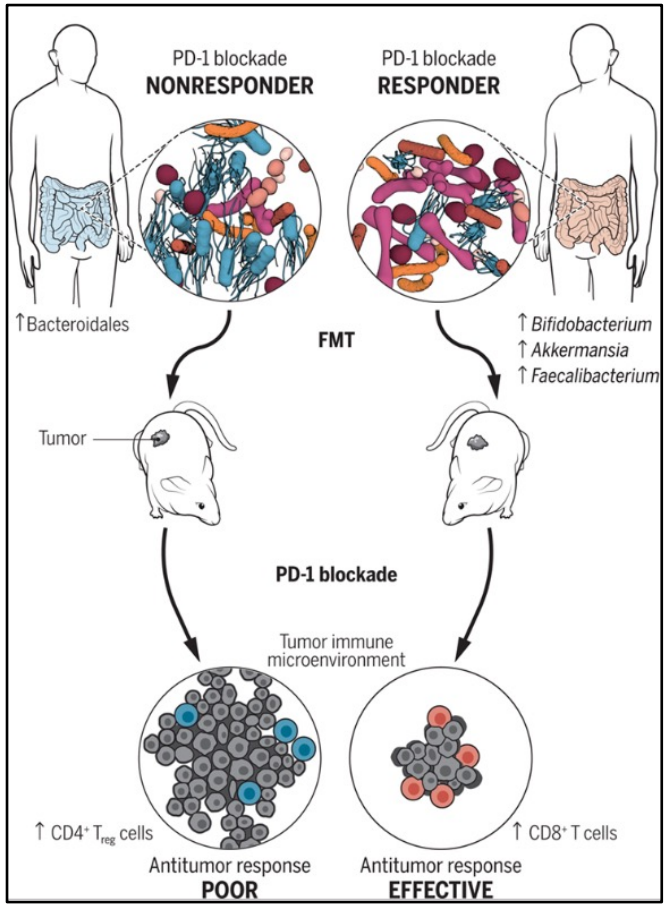
- originated some 3.8 billion years ago
- drive biogeochemical cycles of elements (C, N, P, S, etc.)
- transform energy and biomass

Significance (examples):

- biogeochemistry: e.g., photosynthesis by microbes, carbon fixation/export, nitrogen fixation
- health: help us digest food, provide essential vitamins, train the immune system

McFall-Ngai et al., PNAS, 2013; Venn et al., Science, 2014

Describing microbial communities – Example 1



GRAPHIC: V. ALTOUNIAN/*SCIENCE*

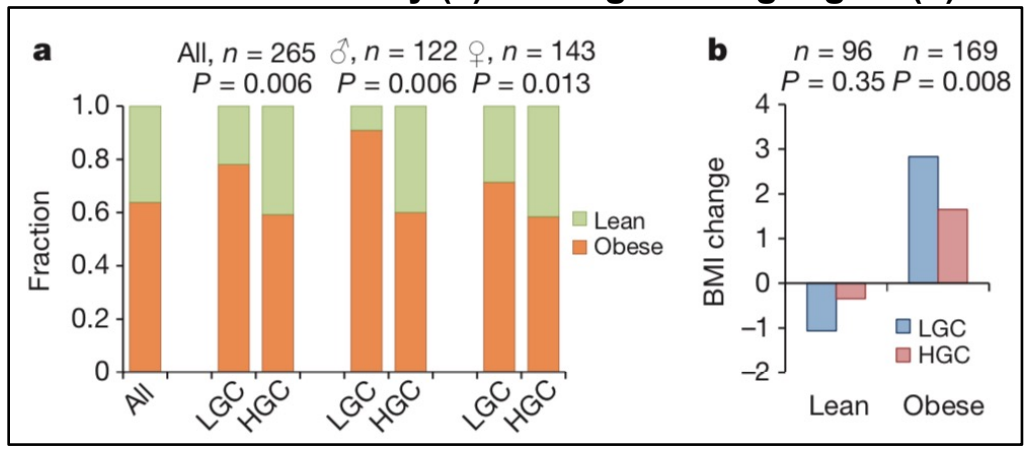
- Compositions of microbial communities are important to characterize, because many host-associated microbes are increasingly implicated in diseases (and personalized medicine)

→ Enrichment of specific microbial taxa may influence the response to cancer immunotherapy

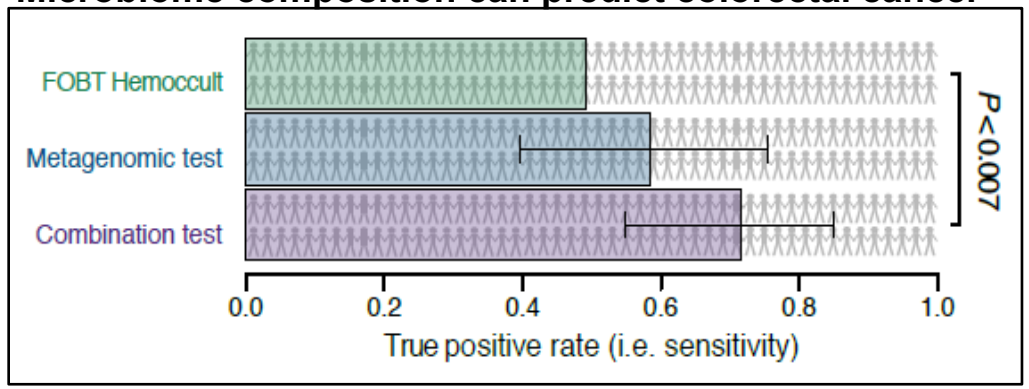
Routy et al., Gopalakrishnan et al., and Matson et al. Science 2018

Describing microbial communities – Example 2

Less diverse microbiomes (LGC=low gene count) are associated with obesity (a) and higher weight gain (b)



Microbiome composition can predict colorectal cancer



- Many diseases are associated with low microbial diversity (e.g., obesity)
- Microbial community compositions can be indicative for disease (e.g., colorectal cancer)

Le Chatellier et al., Nature, 2014; Zeller et al., MSB, 2014

→ **Today, you will learn how to formally describe the composition and diversity of, and differences between microbial communities**

Overview of the Metagenomics block

Part I - Microbial community structure

- microbial taxonomy and operational taxonomic units (OTUs)
- quantification of diversity within a microbial community (alpha diversity)
- comparison between microbial communities (beta diversity)

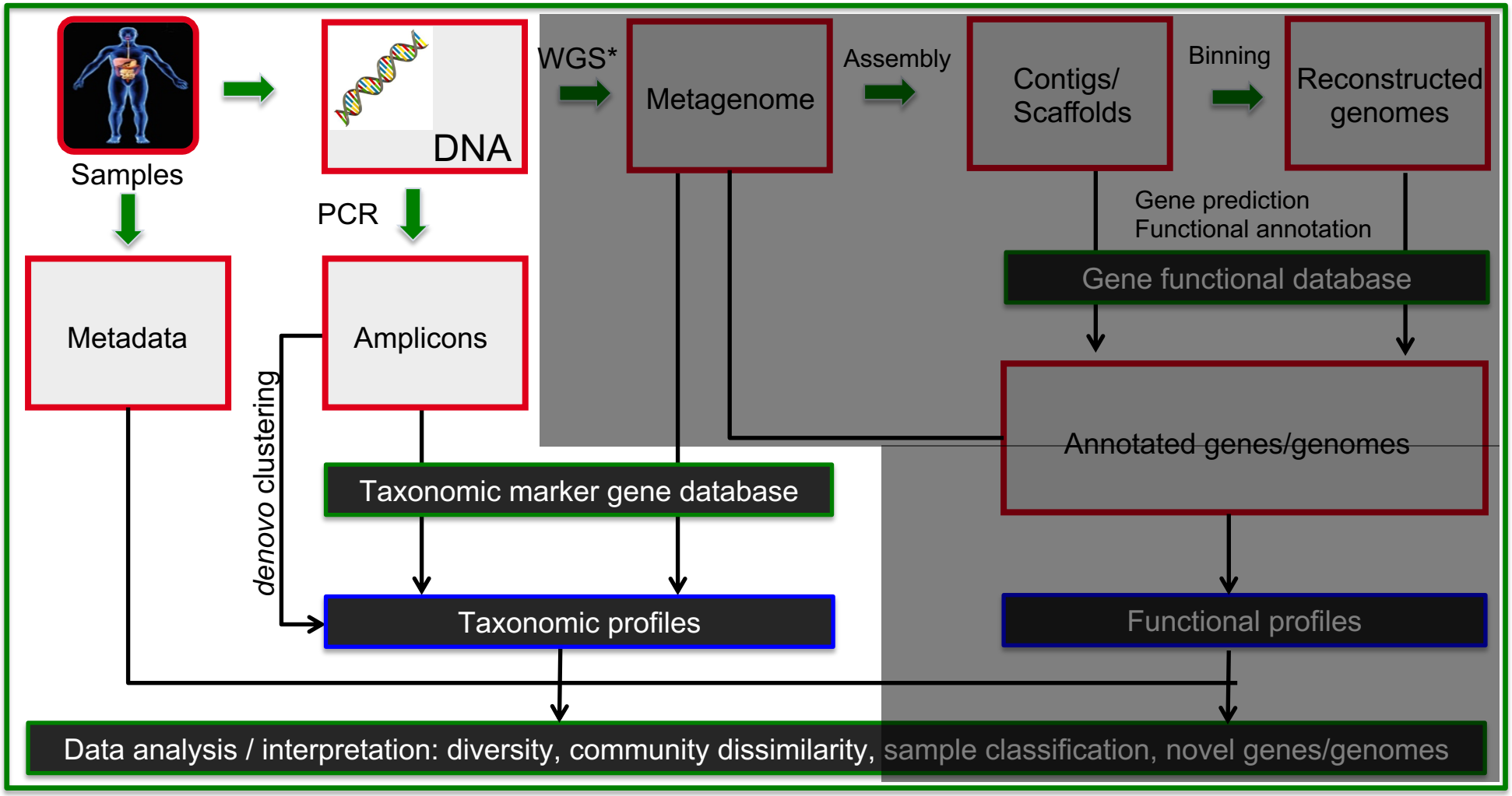
Part II – Reconstruction and annotation of microbial community genomes

- assembly of individual genomes and metagenomes
- binning of metagenomic assemblies into metagenome-assembled genomes
- taxonomic and functional annotation of metagenomes

Part III – Quantitative metagenomics

- generation of *denovo* metagenomic resources
- taxonomic and functional profiling of microbiomes
- classification/regression by supervised machine learning

Today: overview



Learning Objectives

- You can explain the need for and the concept of an operational taxonomic unit (OTUs)
- You can formally describe the diversity of a microbial community as a function of richness and evenness
- You can calculate the difference between two or more microbial community compositions using at least one dissimilarity index

→ During the exercise session following the lecture, you will solve tasks to familiarize yourself with new concepts (and practice the use of R)

Review: microbial taxonomy

- Taxonomy = science of biological classification:
  - Classification is the arrangement of organisms into groups (taxa)
  - Nomenclature refers to the assignment of names to taxonomic groups
  - Identification (also "Annotation", "Phylotyping", or "Classification") refers to the determination of the particular taxon to which a particular isolate belongs
- Based on classical characteristics:
  - morphology, physiology/metabolism, ecology, genetic exchange (transformation, transduction, conjugation)
- Based on molecular characteristics
  - DNA-DNA hybridization (70% same species, 25% same genus)
  - DNA (or protein) sequences of individual genes (e.g., 16S rRNA gene) or complete genomes

TABLE 3.2. Some taxonomically useful morphological characteristics and their variations

Characteristics	*Variations*
Cell morphology	
Unicellular	Cocci, bacilli, vibrios, spirilli, spirochaetes, prosthecate, stalked, sheathed.
Multicellular	Mycelial, filamentous.
Cell arrangement	Single, pairs, chains, bunches, packets.
Staining property	
Gram staining	Gram-positive, gram-negative.
Acid fast staining	Acid-fast, non-acid fast.
Flagellation	Monotrichous, lophotrichous, amphitrichous, peritrichous, endoflagellate or non-flagellate.
Motility	Non-motile, flagellar locomotion, gliding movement, motility due to endoflagella.
Glycocalyx	Capsule present or absent, slime layer.
Spores	Non-sporing, endospore, exospore, conidia, myxospores.
Sporangium	Shape, location of spore.
Cell inclusions	Poly β-hydroxybutyrate, volutin, polysaccharides, sulfur droplets, parasporal protein crystals.
Ultrastructural features	Surface structures of cells —flagella, pili, fimbriae, texture of slime layer.

TABLE 3.3. Some taxonomically useful physiological and metabolic characteristics and their variations

Character	*Variations*
Nutritional type	Photolithotrophs, chemolithotrophs, photoorganotrophs, chemoorganotrophs
Cell wall components	Peptidoglycans, teichoic and teichuronic acids, protein, polysaccharides, pseudomurein, etc.
Carbon sources	CO_2, sugars, sugar acids, sugar alcohols, polysaccharides, organic acids
Nitrogen sources	Molecular nitrogen, ammonium salts, nitrate, organic nitrogenous compounds
Energy metabolism	Photosynthesis, respiration, fermentation, inorganic substrate oxidation, nitrate and sulphate oxidation
Oxygen relationship	Aerobic, microaerophilic, facultatively anaerobic, obligately anaerobic
Temperature relationships	Mesophilic, facultatively thermophilic, obligately thermophilic, hyperthermophilic, psychrophilic

Review: microbial taxonomy

- Microbiologist have adopted the concept of taxonomic ranks:

  Domain/**K**ingdom, **P**hylum, **C**lass, **O**rder, **F**amily, **G**enus, **S**pecies

TABLE 3.1. Taxonomic ranks or levels in ascending order

Rank or level	*Example*
Species	*E. coli*
Genus	*Escherichia*
Family	Enterobacteriaceae
Order	Enterobacteriales
Class	γ-Proteobacteria
Phylum	Proteobacteria
Domain	Bacteria

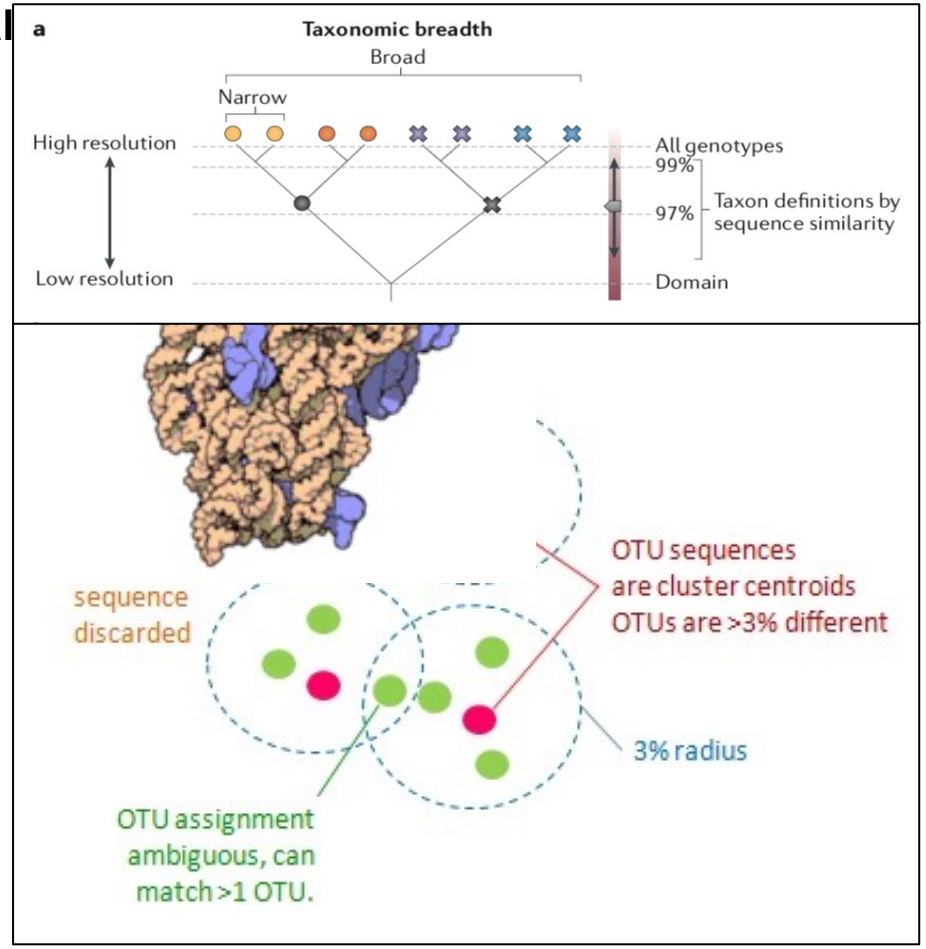
- Phenotypic characteristics:
  - morphology, physiology/metabolism, ecology, exchange of genetic material
- Molecular characteristics
  - DNA-DNA hybridization
  - DNA sequences of individual genes (e.g., 16S rRNA gene) or complete genomes

→ Today, DNA sequencing and computational comparison is the method of choice to determine genetic relatedness

Review: 16S rRNA-based Operational Taxonomic Units (OTUs)

- 16S rRNA
  - present in all prokaryotes
  - conserved function as integral part of the protein synthesis machinery
  - similar mutation rate: → molecular clock
- Proxy for phylogenetic relatedness of organisms
- Owing to lack of prokaryotic species definition, 97% sequence similarity is often used to define ‘species’-like:
  “**Operational Taxonomic Units” (OTUs)**

30S small



Hanson et al., Nat Rev Microbiol, 2012

Microbial community compositions

- Goal: determine ‘who’ is there at what abundance in one or more samples

OTU count table

	OTU1	OTU2	OTU3	OTU4	OTU5	…	Sum
S1	68	38	84	60	60		
S2	9	92	24	0	93		
S3	14	0	21	90	80		
S4	41	34	78	65	29		
S5	3	70	74	63	0		
…							

Rows: S1 to Sn = samples
Columns: OTUs

Step 1: Generation of 16S rRNA amplicon reads by PCR

Community DNA extract

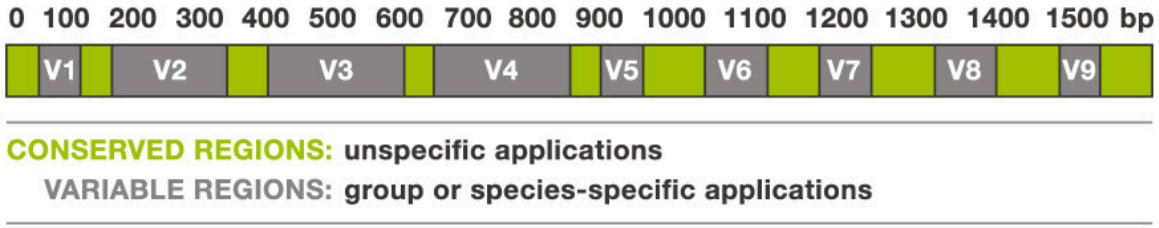
PCR

agtctcgctatgacgtcgtcgtcagactac

gtcgtacgtcgatatttctcgcgccggagc
gtcgtacgtcgatattctcgcggccggagc

agcctacgtcgtcgatagtgcgctagtgtc

- **Primers bind to conserved regions of constant regions.**
- **Variable regions are amplified by PCR**



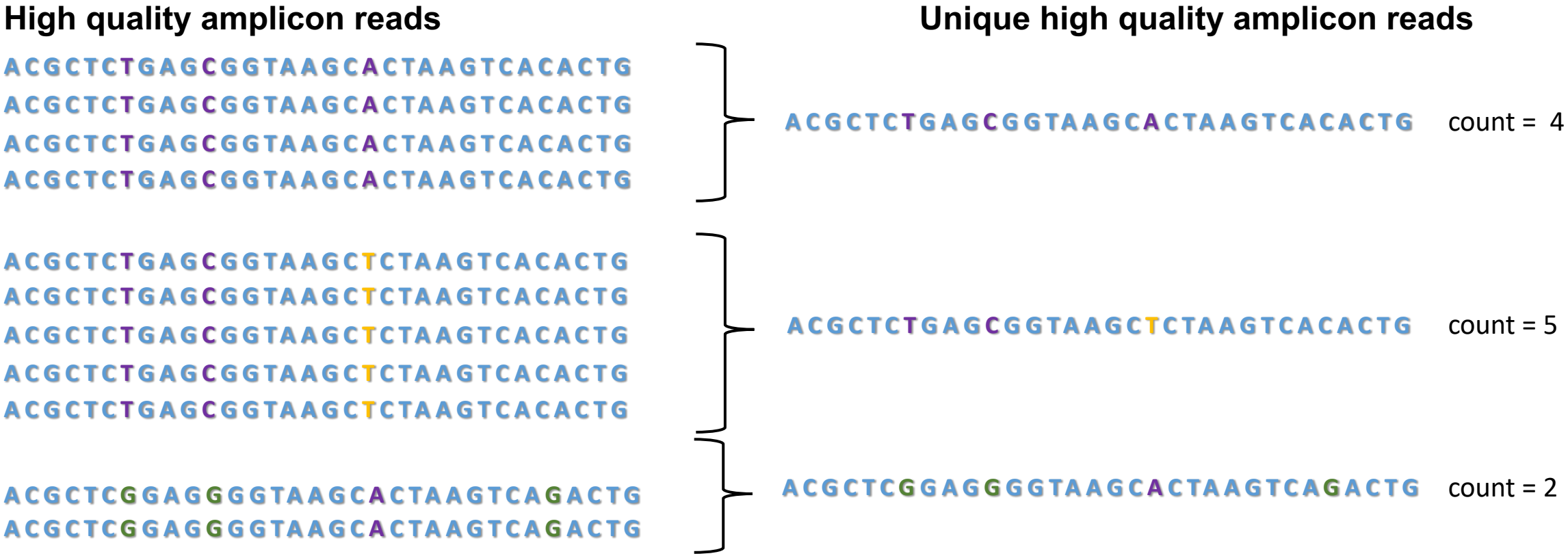
Example

- "V4 primers" yield ca. 250 bp long amplicon reads

- **After sequencing, amplicon reads are quality controlled, yielding high quality amplicon reads**

→ Number of reads are proportional to number of gene copies in the community

Step 2: De-replication of identical sequences



- All reads are aligned to each other to identify identical sequences
- Unique sequences are kept and the number of identical sequences is counted
- Output are unique sequences with records of identical sequences

Step 3: Heuristic clustering of sequences into OTUs

Deterministic approach: calculate all pairwise similarities

→ too "expensive" (resource and time consuming)

Heuristic approach:

1) Unique high quality reads are sorted by counts (high to low)
2) Read with highest count is centroid of a new OTU (N=1)
3) Next read is compared to all OTU centroids

   2 different possibilities:

   a) Centroid sequence and new read are >= 97% identical
      - read becomes new member of the OTU (N=1)

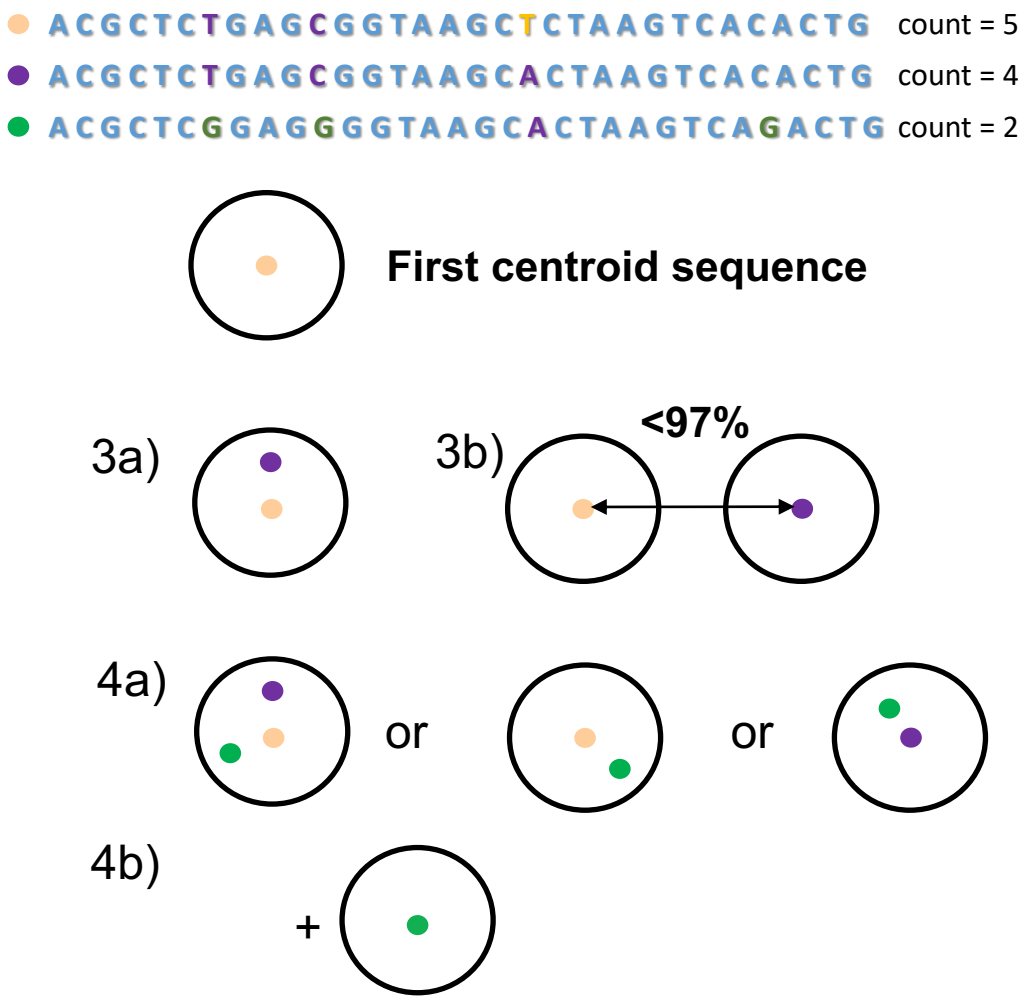
   b) Centroid sequence and new read are < 97% identical
      - read becomes centroid of a new OTU (N=2)

4) Next read is compared to all OTU centroids

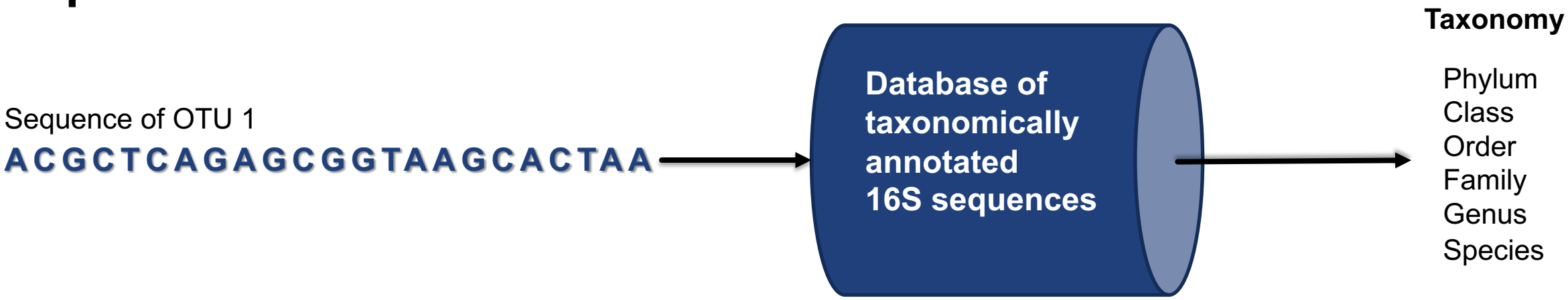
   2 different possibilities:

   a) Any centroid sequence and new read are >= 97% identical
      - read becomes new member of the OTU (N=N)

   b) Any centroid sequence and new read are < 97% identical
      - read becomes centroid of a new OTU (N=N+1)



Step 4: Taxonomic annotation of OTUs

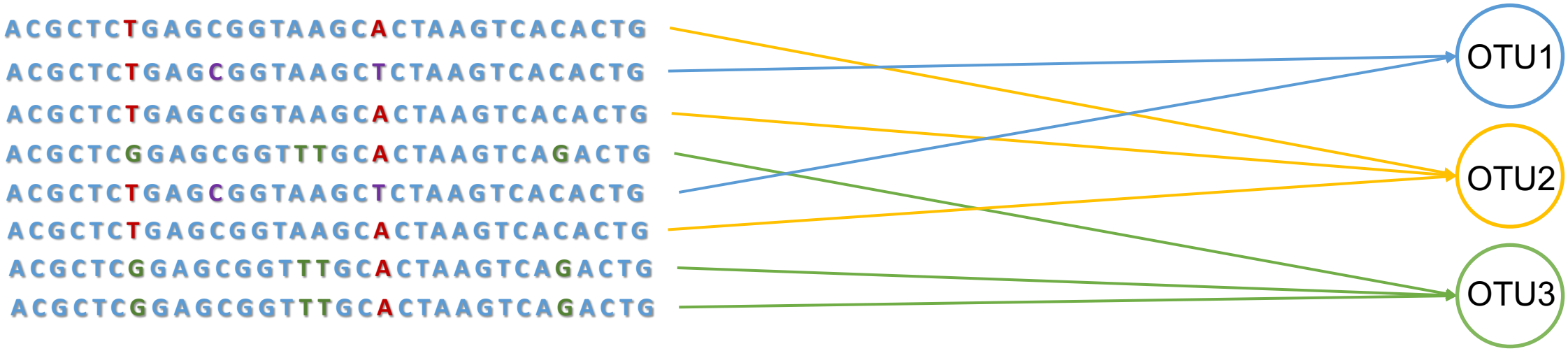


- Identification of taxon to which an OTU belongs
  - The centroid sequence of each OTU is compared to a database of annotated 16S rRNA gene sequences

→ sequences are assigned to taxonomic ranks: phylum, class, family etc.

Step 5: Quantification of OTU abundances

All reads are aligned to best matching OTU centroid sequence (and counted)

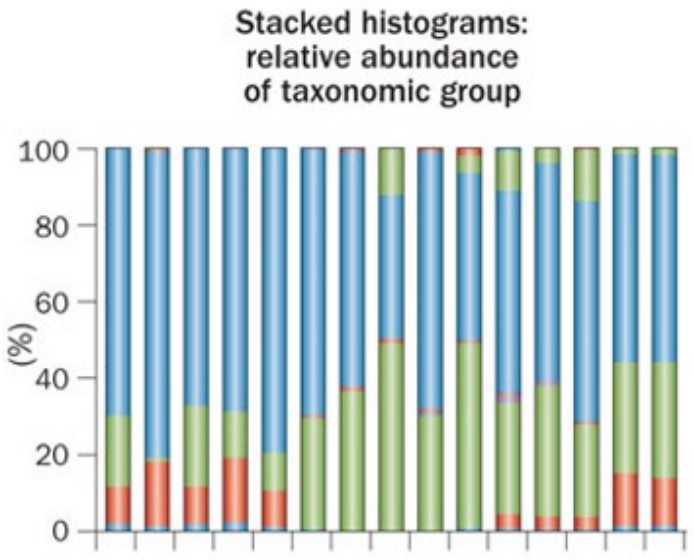


The result is an OTU count table, summarizing read counts for each OTU for each sample:

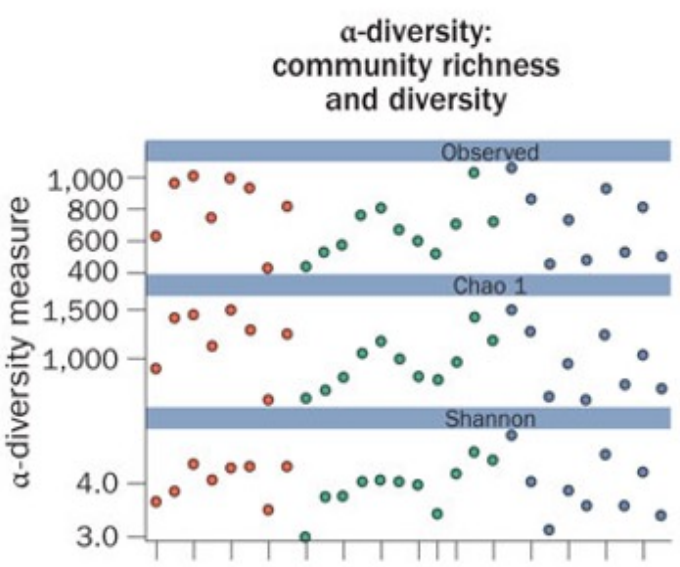
OTU	S1	S2	S3
OTU1	234	87	166
OTU2	23	0	93
OTU3	2	137	191
OTU4	455	0	112
OTU5	23	229	66

Data analysis / interpretation: diversity, community dissimilarity, sample classification

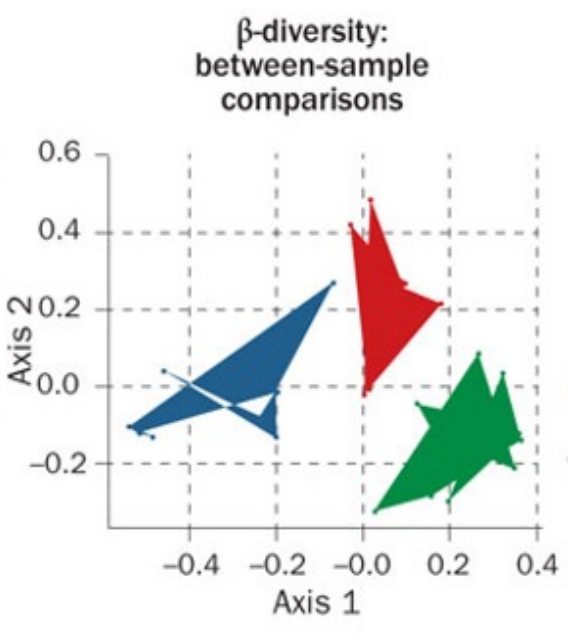
Typical analyses using OTU count tables



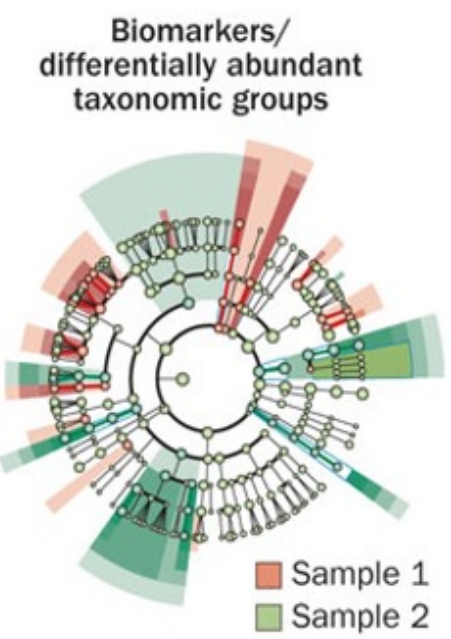
Determine the **relative abundance** of different taxa in microbial communities



Determine the **taxonomic diversity** of microbial communities
→ alpha diversity



Compare **differences between** microbial community compositions
→ beta diversity



Determine **differential abundance** of taxa between samples

In-class task 1: alpha diversity

- Assume 4 different samples (A-D), each with 100 reads sequenced

OTUs	Sample A	Sample B	Sample C	Sample D
1	20	1	25	0
2	20	10	25	0
3	20	20	0	0
4	20	30	25	0
5	20	39	25	100
Sum	100	100	100	100

- In groups of 2, discuss how the diversity of one sample could be formally described (i.e., measured in quantitative terms):
  - How are the 100 reads distributed among the 5 OTUs?
  - Are all OTUs present in a given sample?

→ What are the factors that influence the differences between samples?

In-class task 1: alpha diversity

Shannon's diversity index (*H'*)

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

R = richness
p*i* = the proportion of the *i*-th OTU
n*i* = the number individuals of the *i*-th OTU
n = total number of individuals

Pielou's evenness (J*'*)

$$J' = \frac{H'}{\ln R}$$

→ Please download the spread sheet named "Exercise – Diversity" from Moodle, and calculate *H' and J'* for the example data.

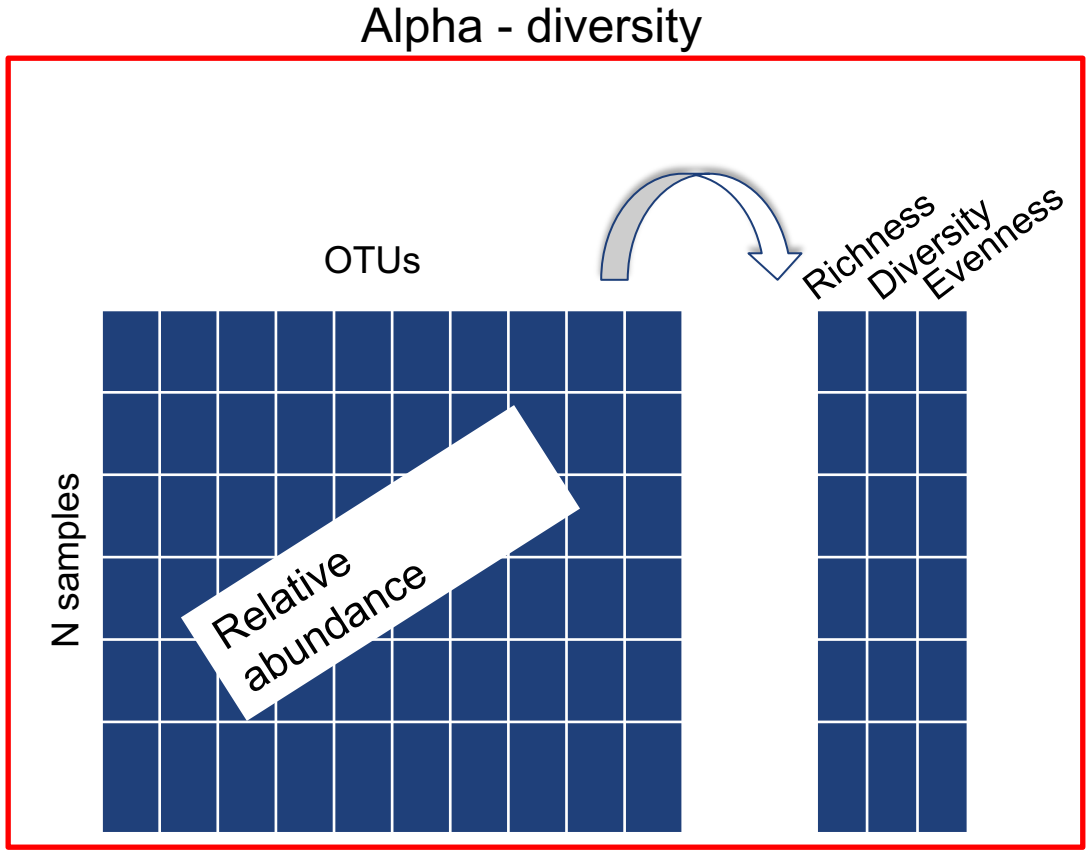
Test your newly acquired knowledge

OTUs	Sample A	Sample B	Sample C	Sample D
1	1	1	4	1
2	1	2	2	2
3	1	3	1	3
4	0	4	0	4
5	0	5	0	0

Which of the samples on the left (A-D) is:

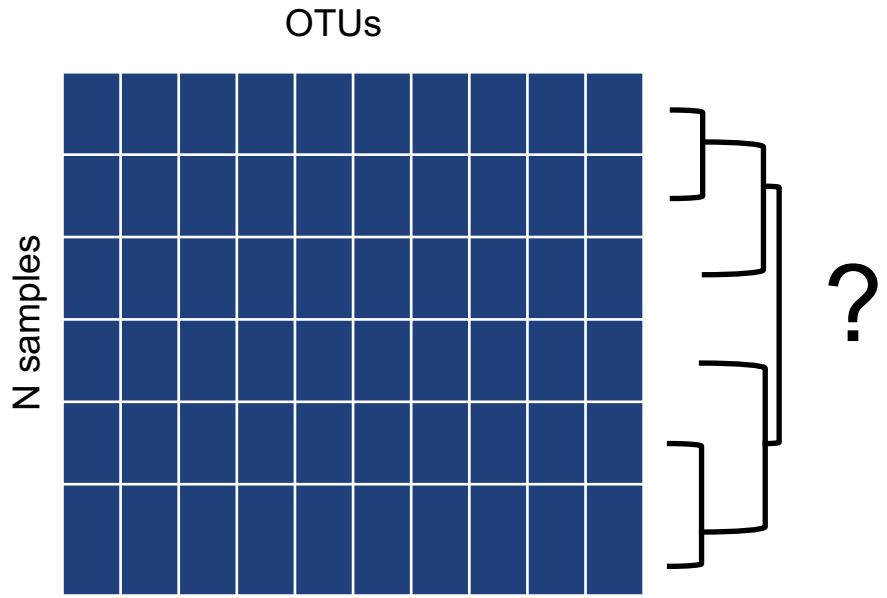
- **the richest?**
- **the most even?**
- **the most diverse?**

Concept of alpha diversity - summary



Beta diversity: between sample dissimilarity

- Now that we learned how to describe the diversity of an individual sample, how do we compare different communities to each other?



In-class task 2: beta diversity

OTUs	Sample A
1	1
2	1
3	1
4	0
5	0

OTUs	Sample B
1	1
2	1
3	1
4	1
5	1

OTUs	Sample C
1	4
2	1
3	1
4	0
5	0

OTUs	Sample D
1	2
2	2
3	2
4	0
5	0

→ In pairs, please discuss how pairwise similarities of samples A, B, C, and D could be quantified?

→ Both qualitative differences vs quantitative differences can be taken into account.

In-class task 2: beta diversity

OTUs	Sample A
1	1
2	1
3	1
4	0
5	0

OTUs	Sample B
1	1
2	1
3	1
4	1
5	1

OTUs	Sample C
1	4
2	1
3	1
4	0
5	0

OTUs	Sample D
1	2
2	2
3	2
4	0
5	0

Example: Jaccard index/dissimilarity

Jaccard index: **J = a / (a + b + c)**

where
a = # of species shared
b= # of species unique to sample 1
c= # of species unique to sample 2

Jaccard distance / dissimilarity: **D = 1 - J**

→ **Note: For Jaccard distance, only presence/absence of species are considered!**

Mini-quiz

What is / are limitation(s) of the Jaccard index?

a) Differences in evenness between two samples are not accounted for
b) Differences in the abundance of OTUs that are shared between two samples are not accounted for
c) Differences in the abundance of OTUs that are not shared between two samples are not accounted for
d) All of the above

Other distance (dissimilarity) measures

The formulae for calculating the ecological distances are:

Bray-Curtis: $$D = 1 - 2\frac{\sum_{i=1}^{S}\min(a_i, c_i)}{\sum_{i=1}^{S}(a_i + c_i)}$$

Kulczynski: $$D = 1 - \frac{1}{2}\left(\frac{\sum_{i=1}^{S}\min(a_i, c_i)}{\sum_{i=1}^{S} a_i} + \frac{\sum_{i=1}^{S}\min(a_i, c_i)}{\sum_{i=1}^{S} c_i}\right)$$

Euclidean: $$D = \sqrt{\sum_{i=1}^{S}(a_i - c_i)^2}$$

Chi-square: $$D = \sqrt{\sum_{i=1}^{S}\frac{(a_+ + c_+)}{(a_i + c_i)}\left(\frac{a_i}{a_+} - \frac{c_i}{c_+}\right)^2} \quad \text{with } a_+ = \sum_{i=1}^{S} a_i$$

Hellinger: $$D = \sqrt{\sum_{i=1}^{S}\left(\sqrt{\frac{a_i}{a_+}} - \sqrt{\frac{c_i}{c_+}}\right)^2} \quad \text{with } a_+ = \sum_{i=1}^{S} a_i$$

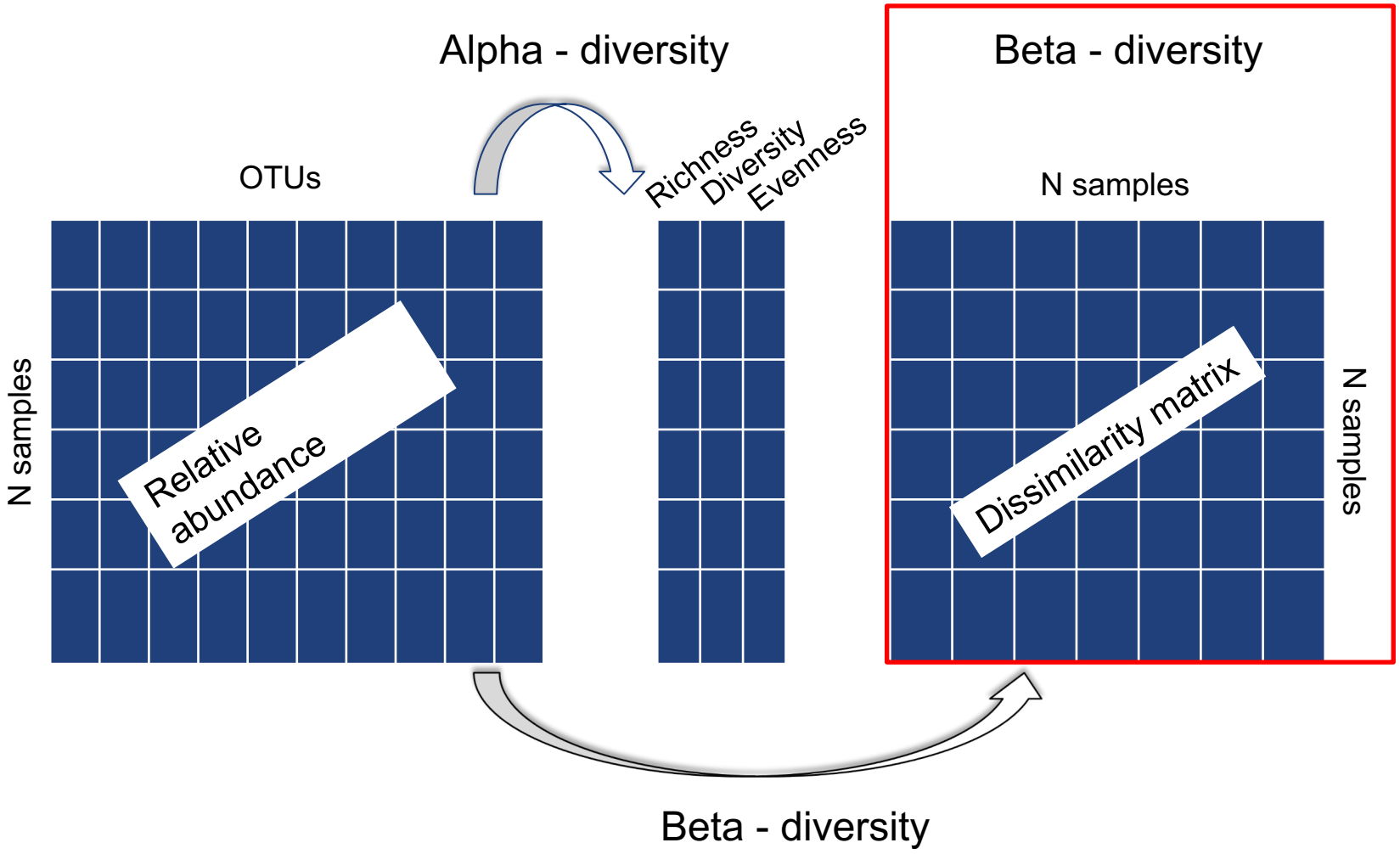
UniFrac distance = phylogenetically weighted distance

Lozupone et al., 2005

→ Using the example spread sheet, calculate all pairwise Jaccard and Bray-Curtis distances for the example data.

a_i = abundance of taxon *i* in sample *a*, and
c_i = abundance of taxon *i* in sample *c*

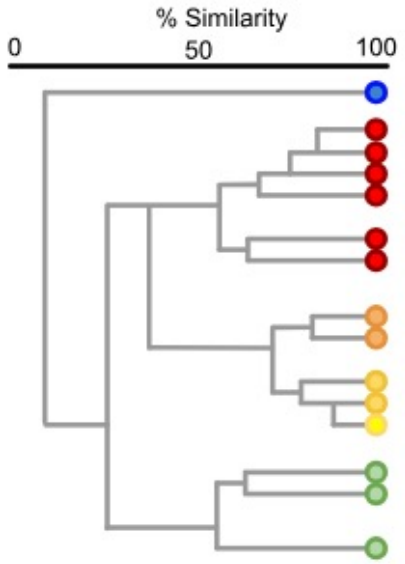
Within sample descriptions → between sample comparisons



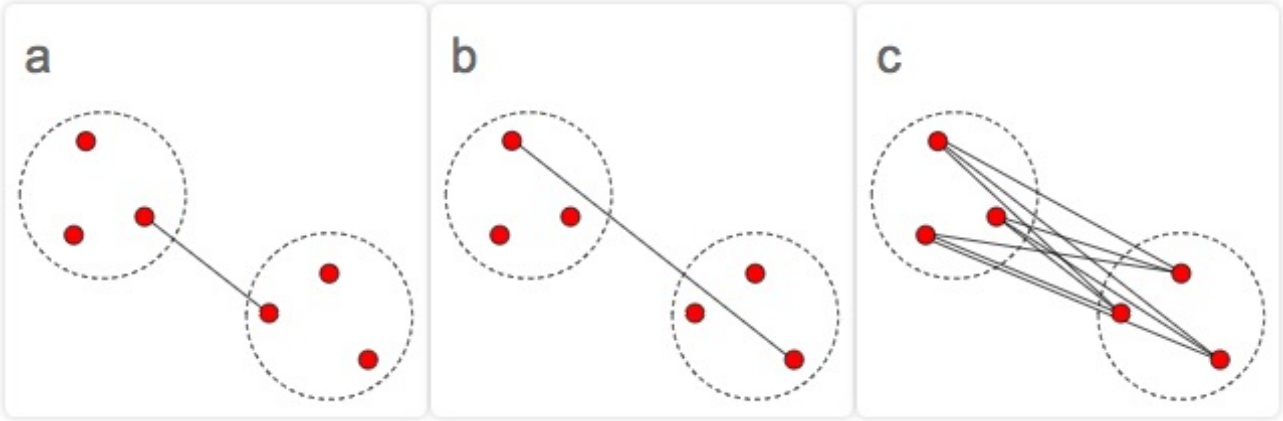
Visualize dissimilarities between microbial communities

- For 2 (xy) or 3 (xyz) variables, data can be easily visualized
- For multi (n>3) dimensional data, calculate distances and ‘project’ into lower dimensional space

Hierarchical clustering



Linkage algorithms

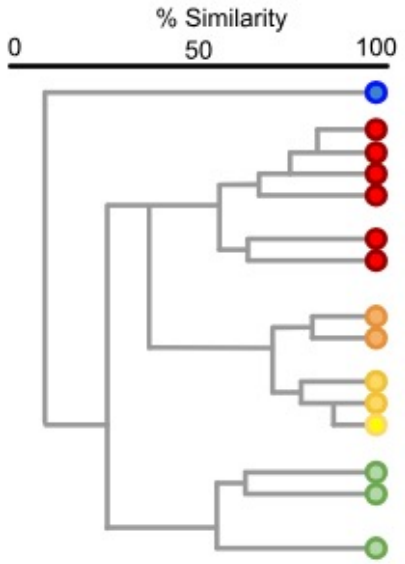


single complete average

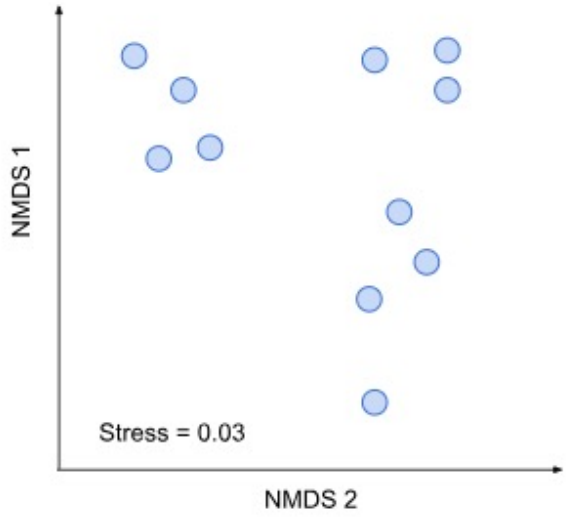
Visualize dissimilarities between microbial communities

- For 2 (xy) or 3 (xyz) variables, data can be easily visualized
- For multi (n>3) dimensional data, calculate distances and 'project' into lower dimensional space

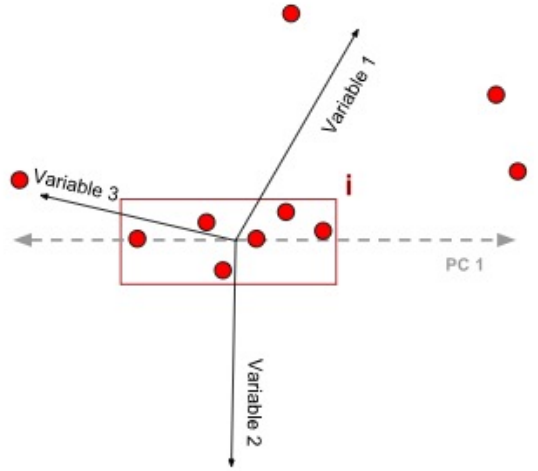
Hierarchical clustering



Non-metric dimensional scaling (NMDS)



Principal component or coordinate analysis (PCA or PCoA)



Summary

- Metagenomics provides information about microorganisms that are often difficult or impossible to be cultivated in their natural environment
- Due to the lack of species concept for prokaryotes, researchers use sequence identity cutoffs of taxonomic markers to define **O**perational **T**axonomic **U**nits
- Microbial community diversity describes the richness (number of species) of taxa and their evenness (the distribution of their abundances)
  - Alpha diversity (within sample diversity) is a function of richness and evenness
  - Alpha diversity can be quantified by different diversity indices (e.g., Shannon)
  - Beta diversity describes differences between microbial communities
  - Beta diversity can be quantified by different dissimilarity indices (e.g., Jaccard, Bray-Curtis)

Example exam questions

If a 99% (rather than 97%) sequence identity cutoff was used to define an OTU, how would this affect the richness of a sample?

a) Richness would increase

b) Richness would decrease

c) Since only evenness would be affected, there is no effect on richness

Assume you have two samples A and B. For sample A, 1,000 amplicons were generated, while for sample B only 500. What measure do you have to take, so that the richness of both samples can be compared?

Literature / web-resources

- The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet
  - https://www.ncbi.nlm.nih.gov/books/NBK54011
- GUide to STatistical Analysis in Microbial Ecology
  - https://sites.google.com/site/mb3gustame/home
- Kindt R and Coe R. 2005. *Tree diversity analysis. A manual and software for common statistical methods for ecological and biodiversity studies.* Nairobi: World Agroforestry Centre (ICRAF).