3000788 Intro to Comp Molec Biol

Lecture 8: Metagenomics

Fall 2025

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Today’s agenda

- Microbiome and metagenomics
- Relevant bioinformatics techniques

Metagenomics: analysis of a collection of organisms

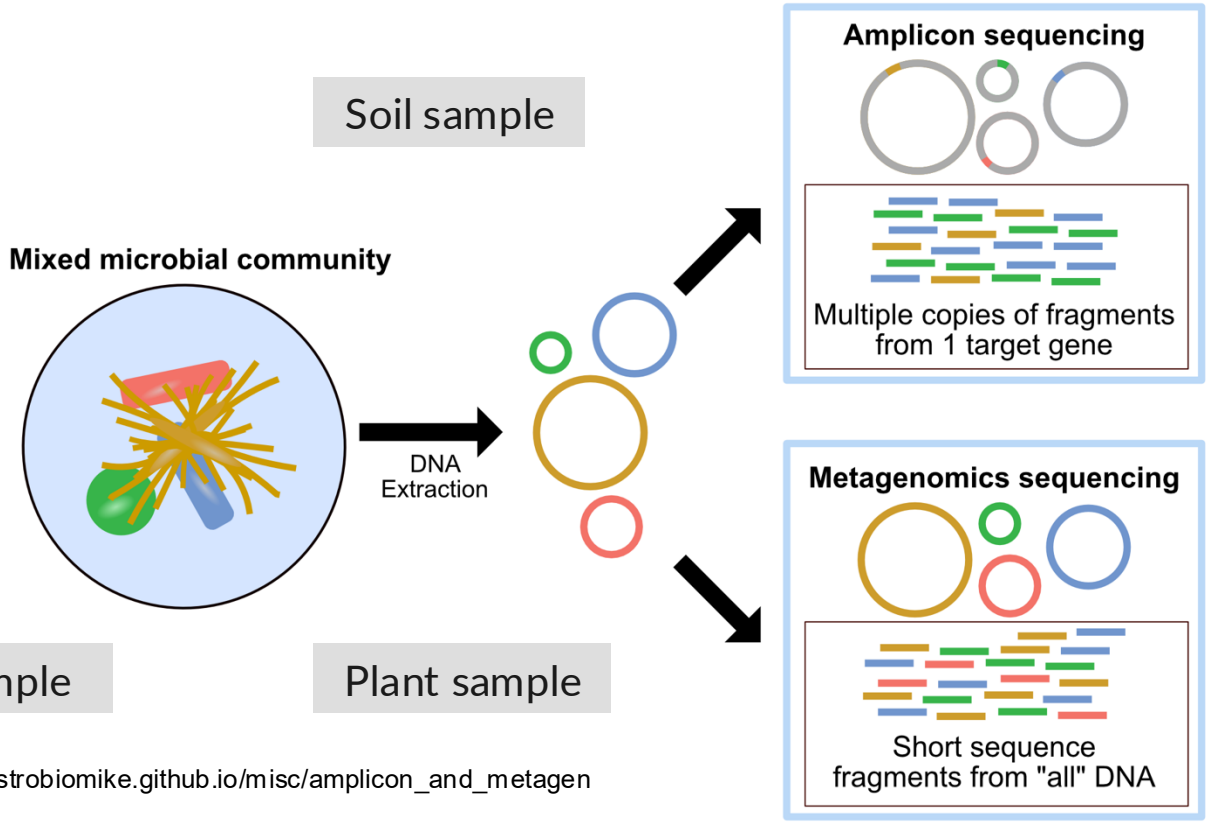
Tumor sample

Soil sample

Stool sample

Factory waste sample

Plant sample



https://astrobiomike.github.io/misc/amplicon_and_metagen

Challenges in metagenomics

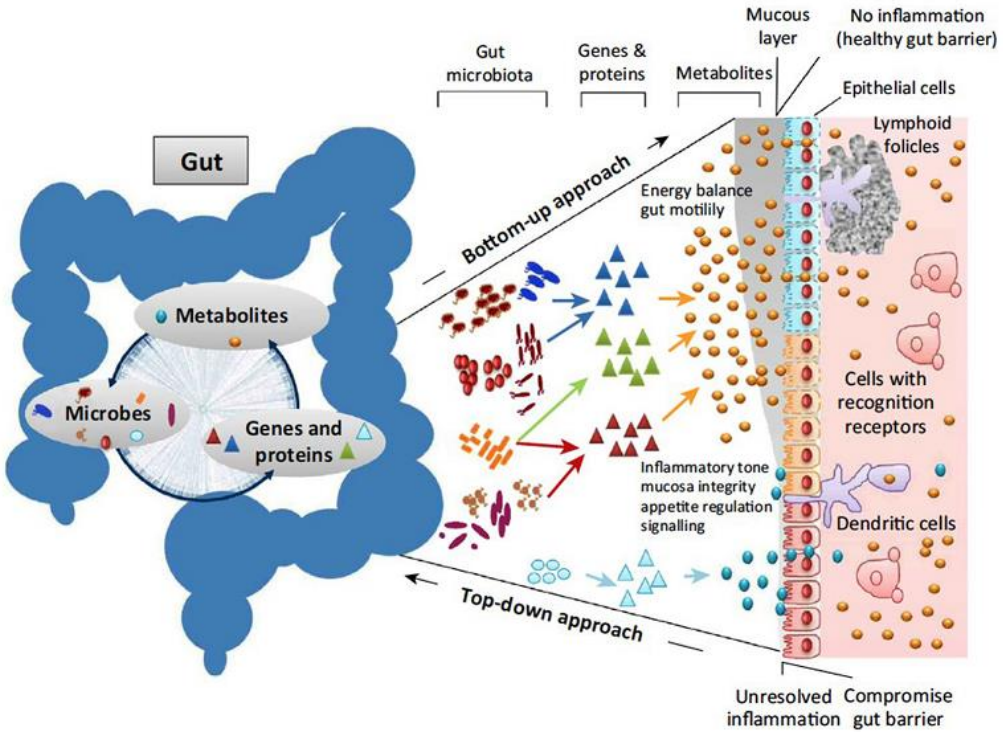
- Reads come from different genomes, how to distinguish?
  - **Clinical samples**: Host versus pathogens
  - **Environmental samples**: Mixed of animals and microorganisms

- **Genome assembly** >> alignment
  - Reference genomes are unavailable
  - Need to know gene sequences to infer functions

- Especially difficult when there are **multiple strains** of the same organisms!

Examples of microbiomes

Gut microbiomes



Moya and Ferrer, Trends in Microbiology 24: 402-413 (2016)

- Microbes digest food and synthesize nutrients and metabolites
- As a community, protect against harmful pathogens
- Reflect host health status

Human Microbiome Project

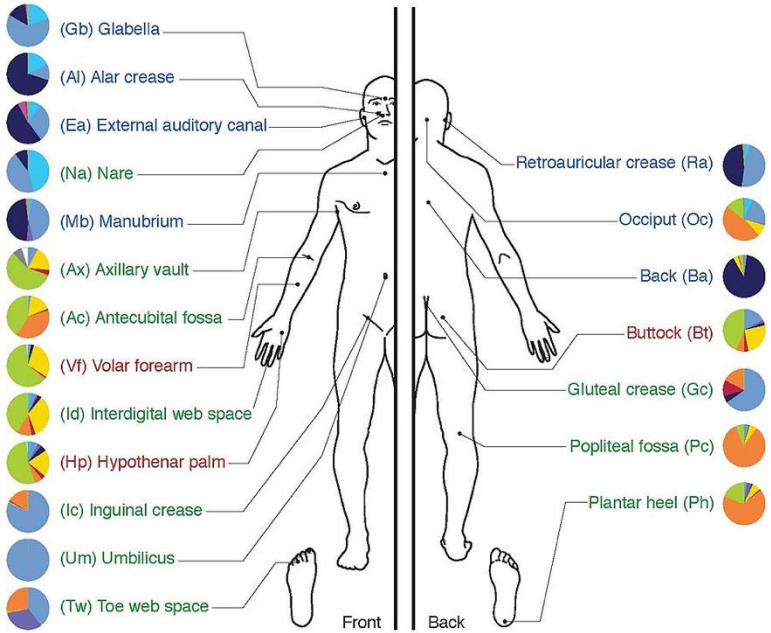
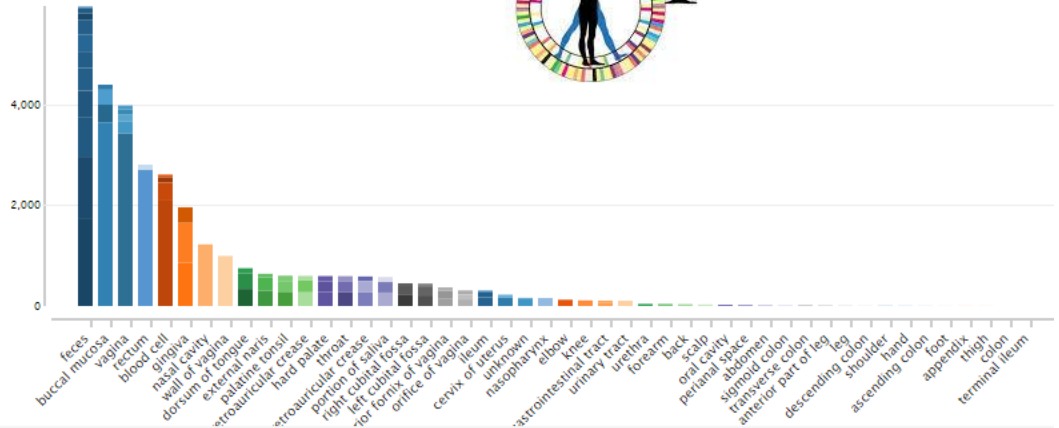


Image from Wikipedia.com

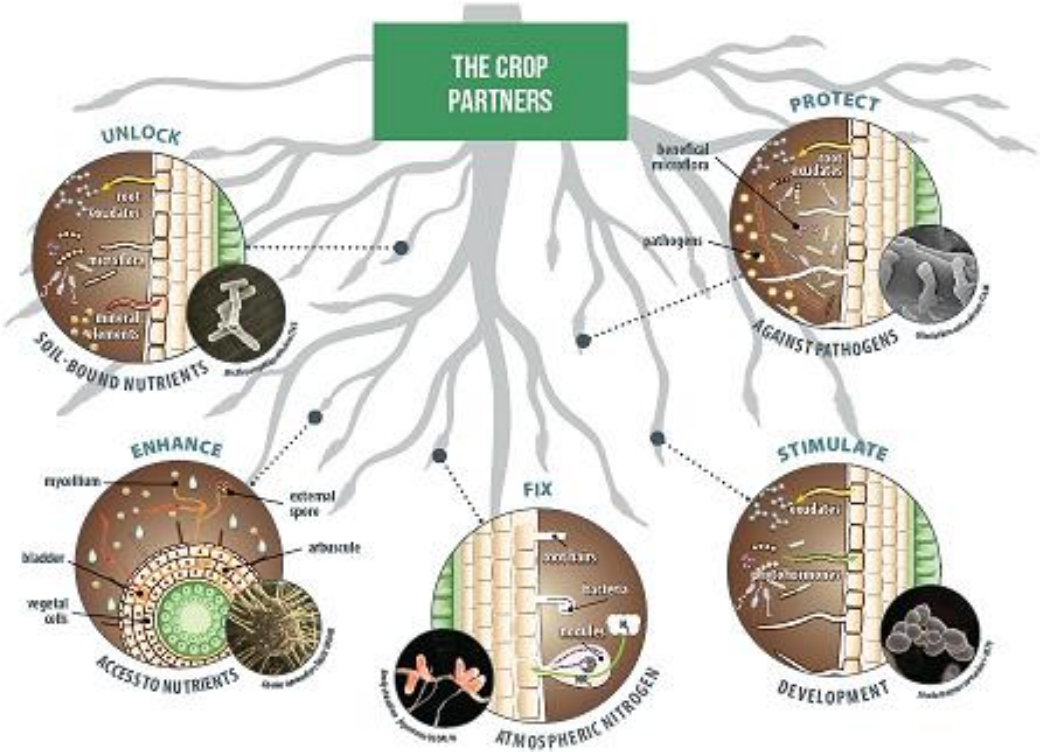


DATA PORTAL SUMMARY

Data Release 1.1 - September 26, 2017

STUDIES	PRIMARY SITE	SAMPLES	FILES
21	48	32,181	118,597

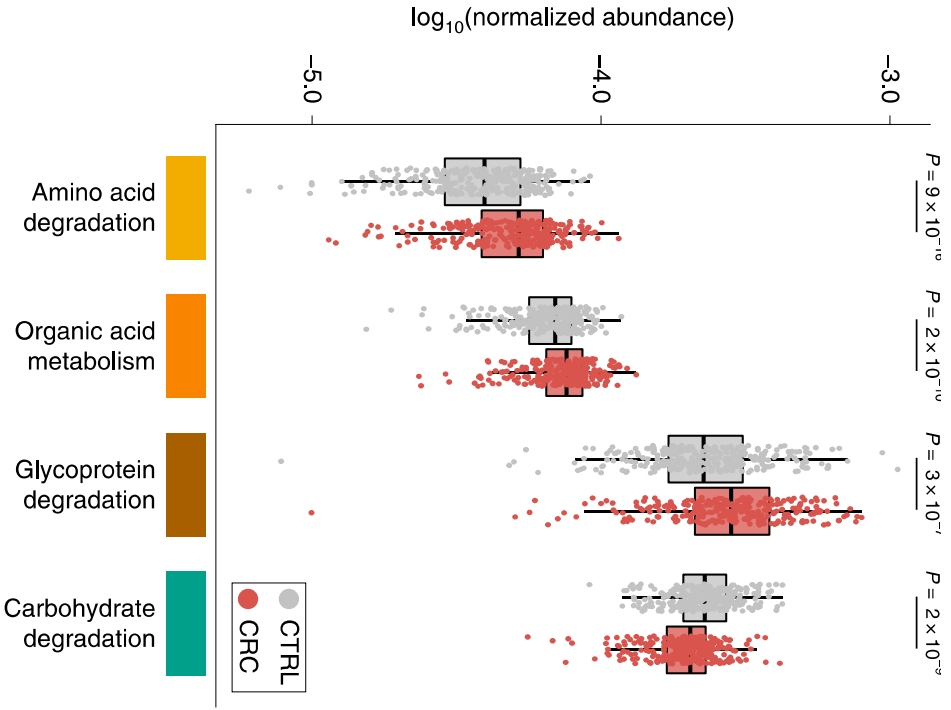
Plant and soil microbiomes



- Decompose and synthesize nutrients for plants
- As a community, protect against harmful pathogens
- Stimulate plant growth

http://www.lallemandplantcare.com/en/our-solutions/rhizosphere-inoculants/

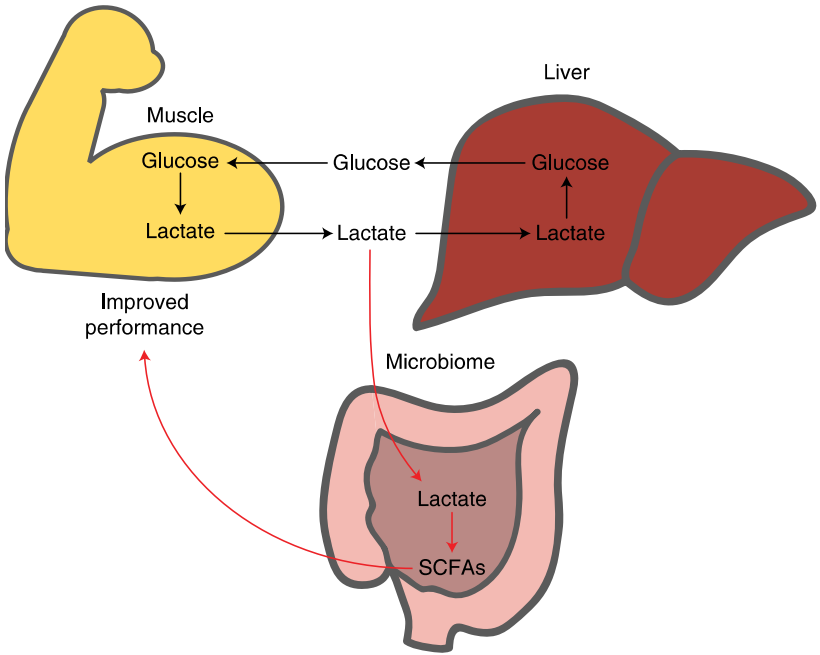
Tumor microbiome



Wirbel *et al*. Nat Med 25:679-689 (2019)

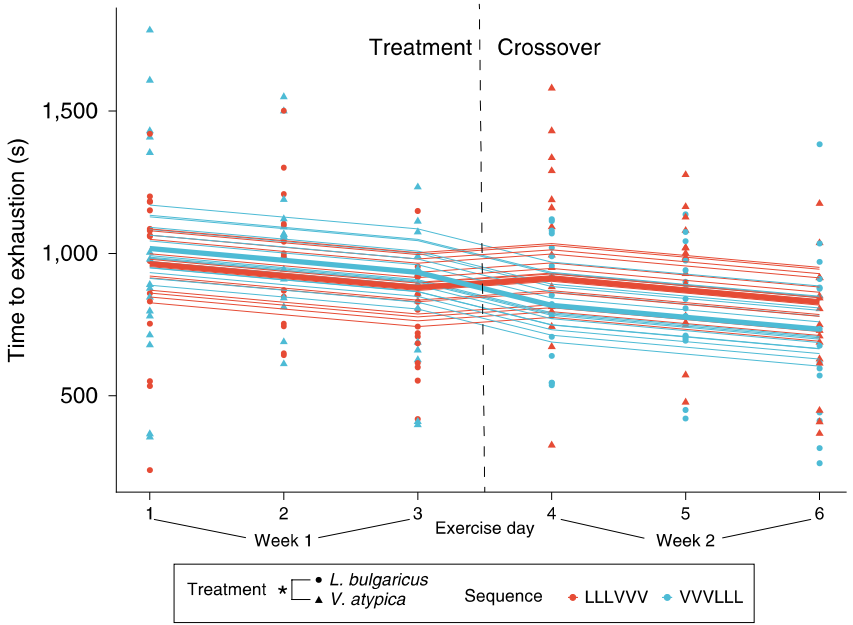
- Exchange nutrients with cancer cells in tumor microenvironment to sustain tumor growth
- Influence host immune response
- Can serve as biomarkers and targets for cancer treatment
  - Disrupt microbiome to disrupt cancer growth

Lactate-utilizing bacteria in athelete guts



Scheiman *et al*. Nat Med 25:1104-1109 (2019)

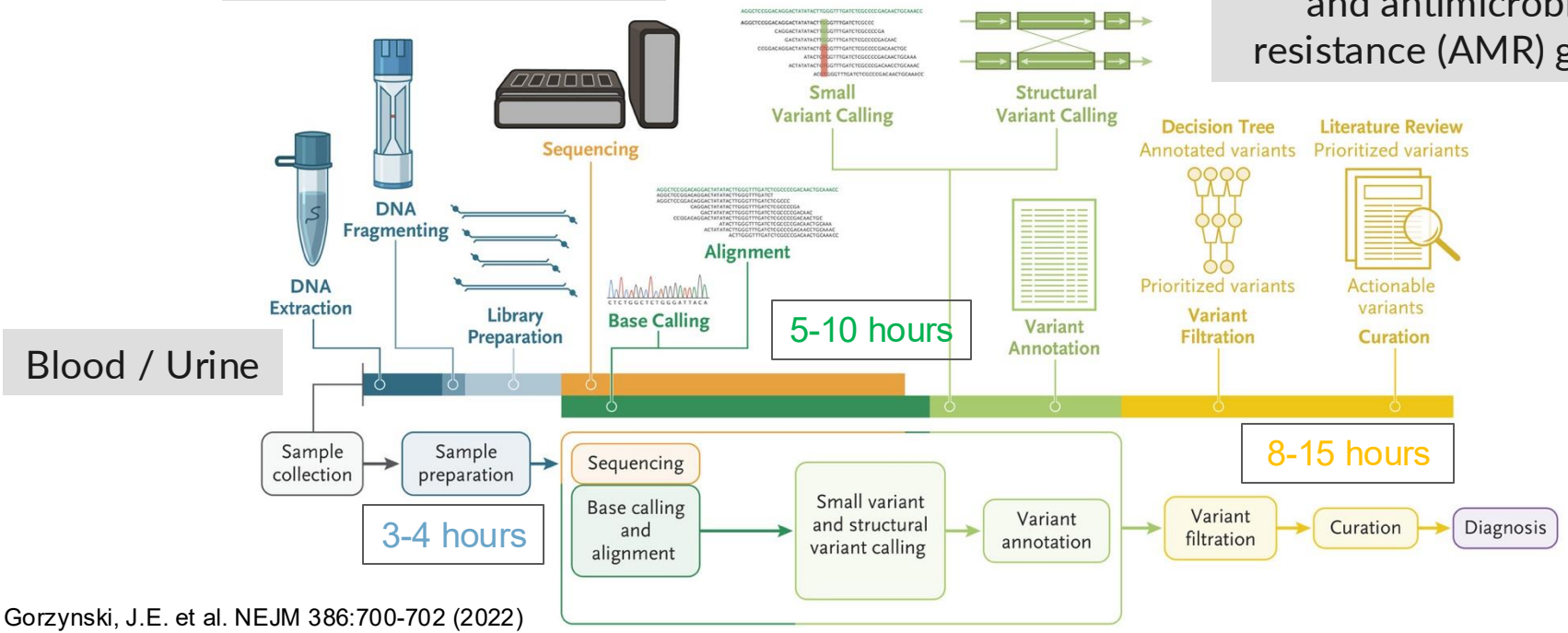
Increased athletic ability in mice with transplanted microbiome



Rapid pathogen detection for clinical decision

Nanopore sequencing

Identify pathogen species and antimicrobial resistance (AMR) genes



Gorzynski, J.E. et al. NEJM 386:700-702 (2022)

Wildlife conservation via metagenomics monitoring

Père David's deer and their gut microbiome

Different dietary plants (nutrition, isotopic evidences)

Next-generation sequencing bioinformatics' analysis

Dissimilarity

Gut microbial composition

Gut microbial function (cellulose digestion, salt-related metabolism)

Conservation

Reflecting increased evolutionary potential and resilience in response to environment changes

Helping us select a putative translocation region

Yao, R. et al. Evol Bioinfo 15:1-3 (2019)

Metagenomics for One Health

https://www.isglobal.org/en/healthisglobal

- Monitoring of pathogens in environmental, wildlife, livestock, and agricultural microbiomes
- Monitoring of pathogen evolution during an active outbreak
- **Resistome**: Targeted sequencing of known AMR genes

Some research questions in metagenomics

- **Health**: microbiome functional markers for diseases and health statuses

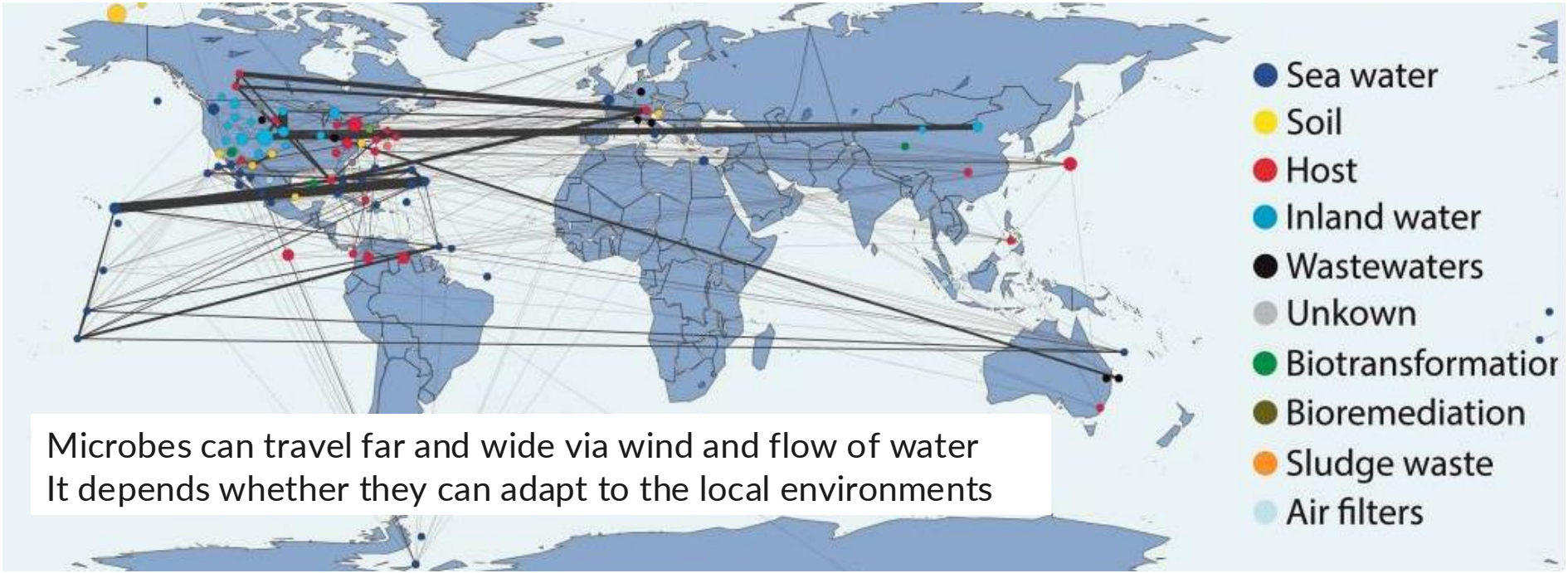
- **Ecology**:
  - Change in microbiome due to human actions
    - Factory and hospital wastes
    - Global warming
  - Microbiome of extreme conditions → new biotechnology tools

- **Agriculture**: soil-plant microbiome interactions

- Pathogen surveillance

Dynamics of microbiomes

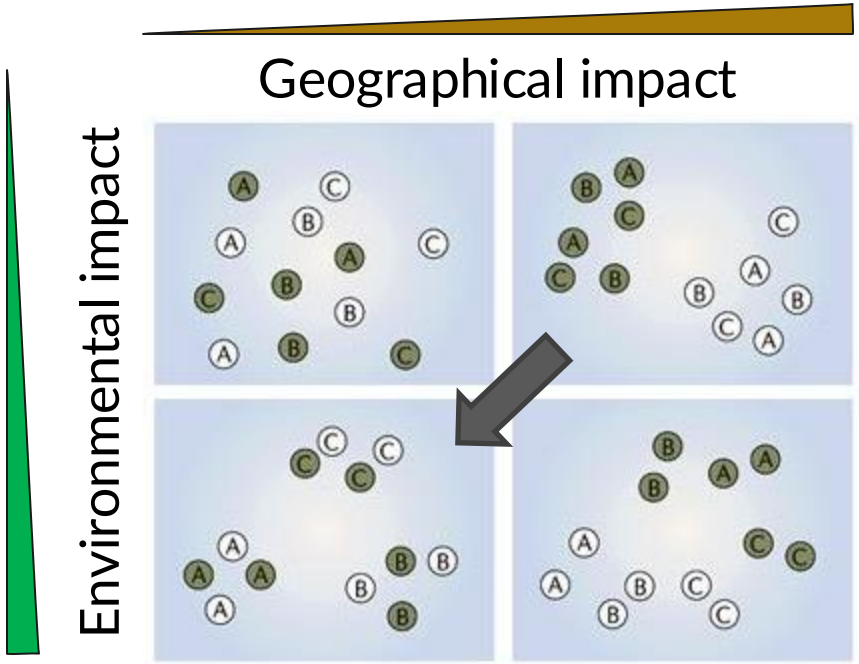
Everything is everywhere, but the nature selects



Microbes can travel far and wide via wind and flow of water
It depends whether they can adapt to the local environments

Fondi *et al.*, Genome Biol Evol 8: 1388-1400 (2016)

Environment defines microbiome more than geography



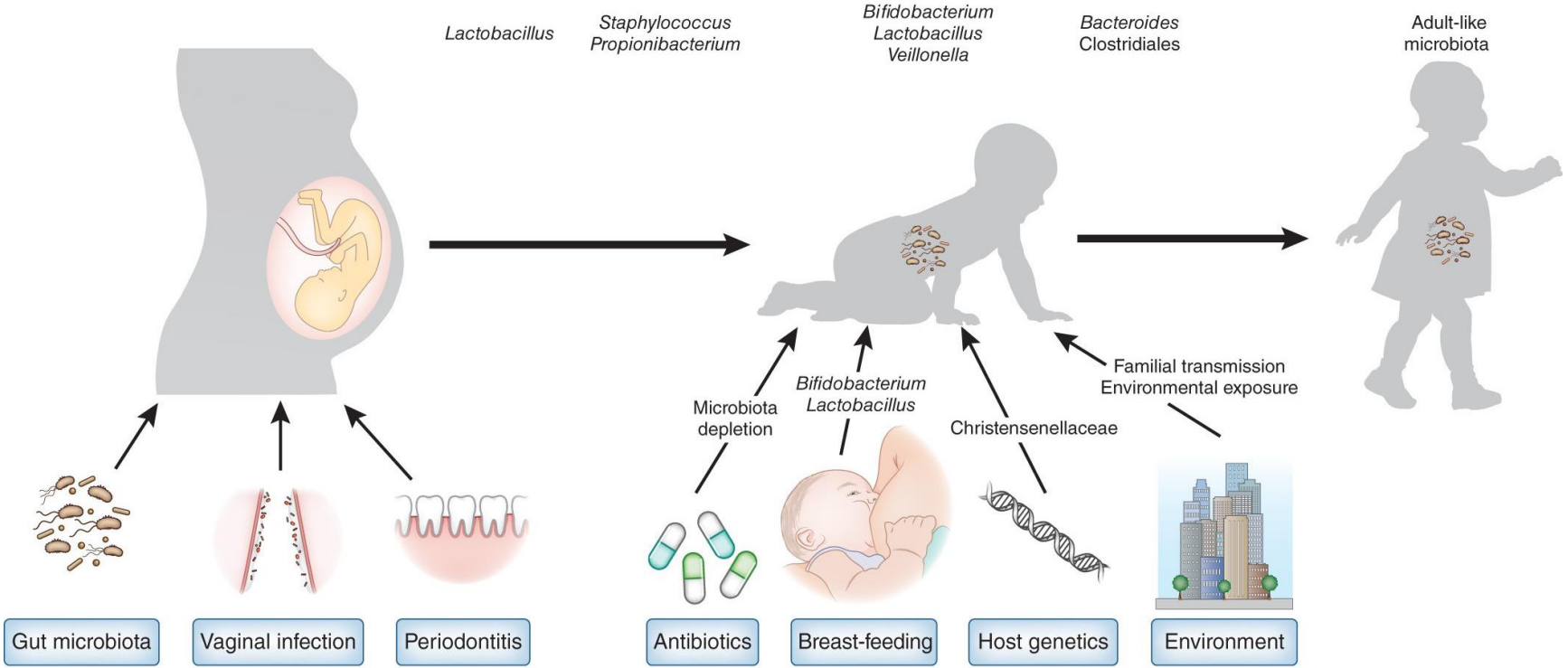
indicate same geography

A, B, C indicate same environment

- **Large animals** can withstand changing environments but cannot travel world-wide
- **Microbes** can travel world-wide but cannot withstand changing environments

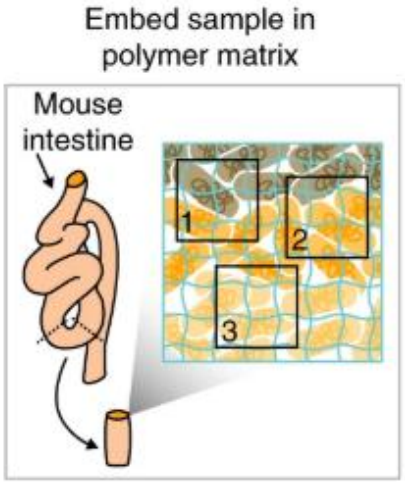
Adapted from Martiny *et al.*, Nature Review Microbiology 4: 101-112 (2006)

Microbiomes are extremely dynamics

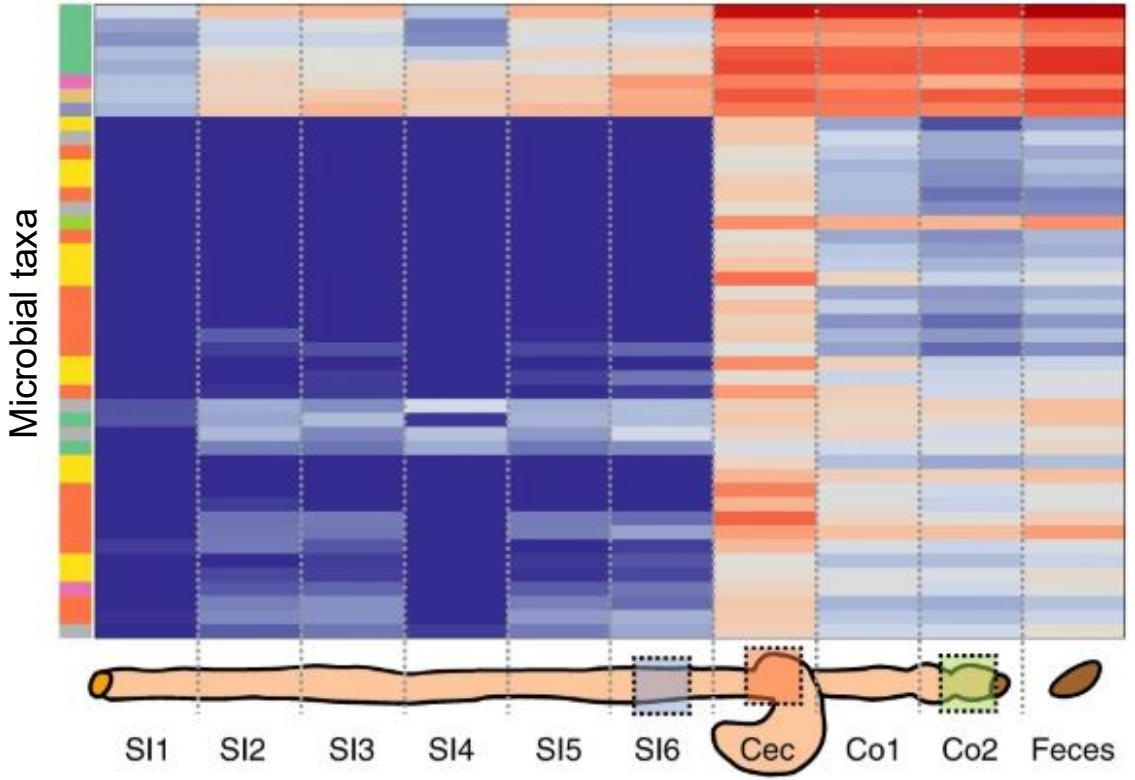


Tamburini, S. et al. Nat Med 22:713-722 (2016)

Spatial variability of gut microbiome



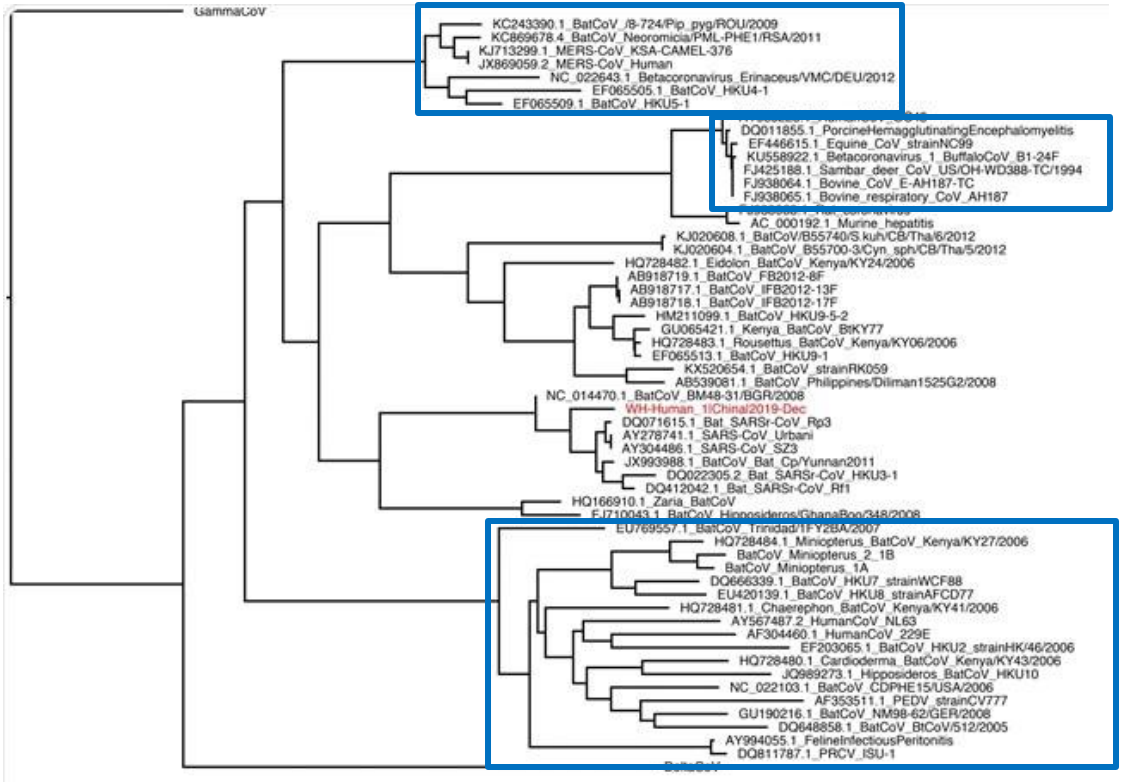
Sheth *et al*., Nature Biotech 37: 877-883 (2019)



Operational taxonomic unit (OTU)

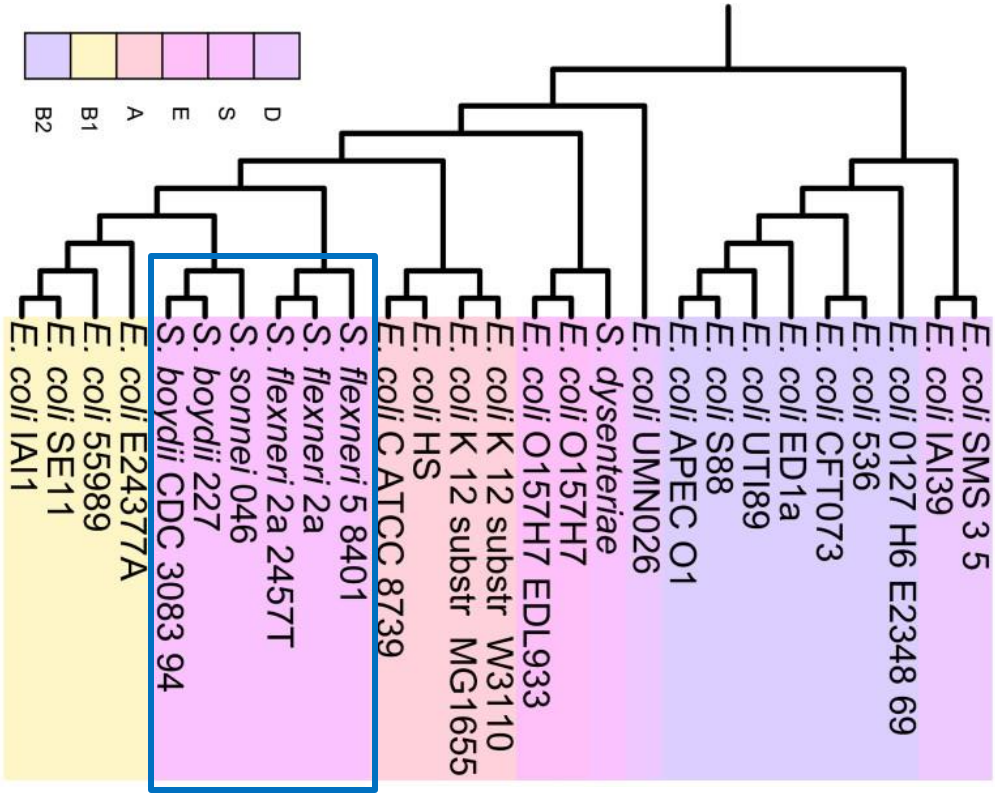
Traditional phylogeny: Cluster of similar sequences

- Is sequence similarity alone sensitive enough to distinguish **closely-related microbes**?
- If sequences are all very similar, **where to draw the cutoff**?



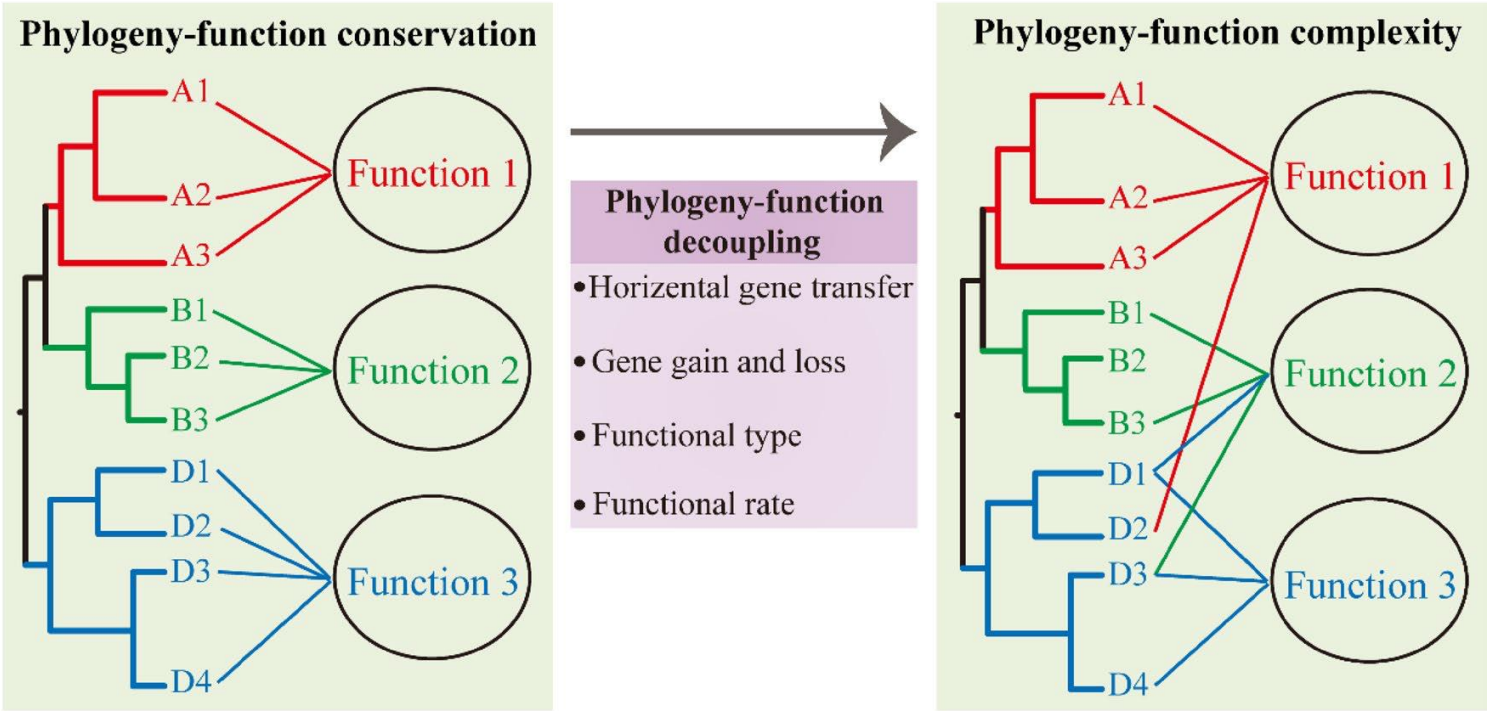
When sequence similarity alone is not enough

- Genus *Shigella* are pathogens that evolved from an *E. coli* ancestor
- 80-90% similarity to some *E. coli* clades
- Definition of taxonomy may **require both genotypes and phenotypes**



Skippington and Ragan, BMC Genomics 12:532 (2011)

Phylogeny based on function (inferred from genes)



Liu *et al.*, Marine Life Science & Technology 1: 112-128 (2019)

A rough OTU similarity cutoff for microbes

nature reviews microbiology

Published: 14 August 2014

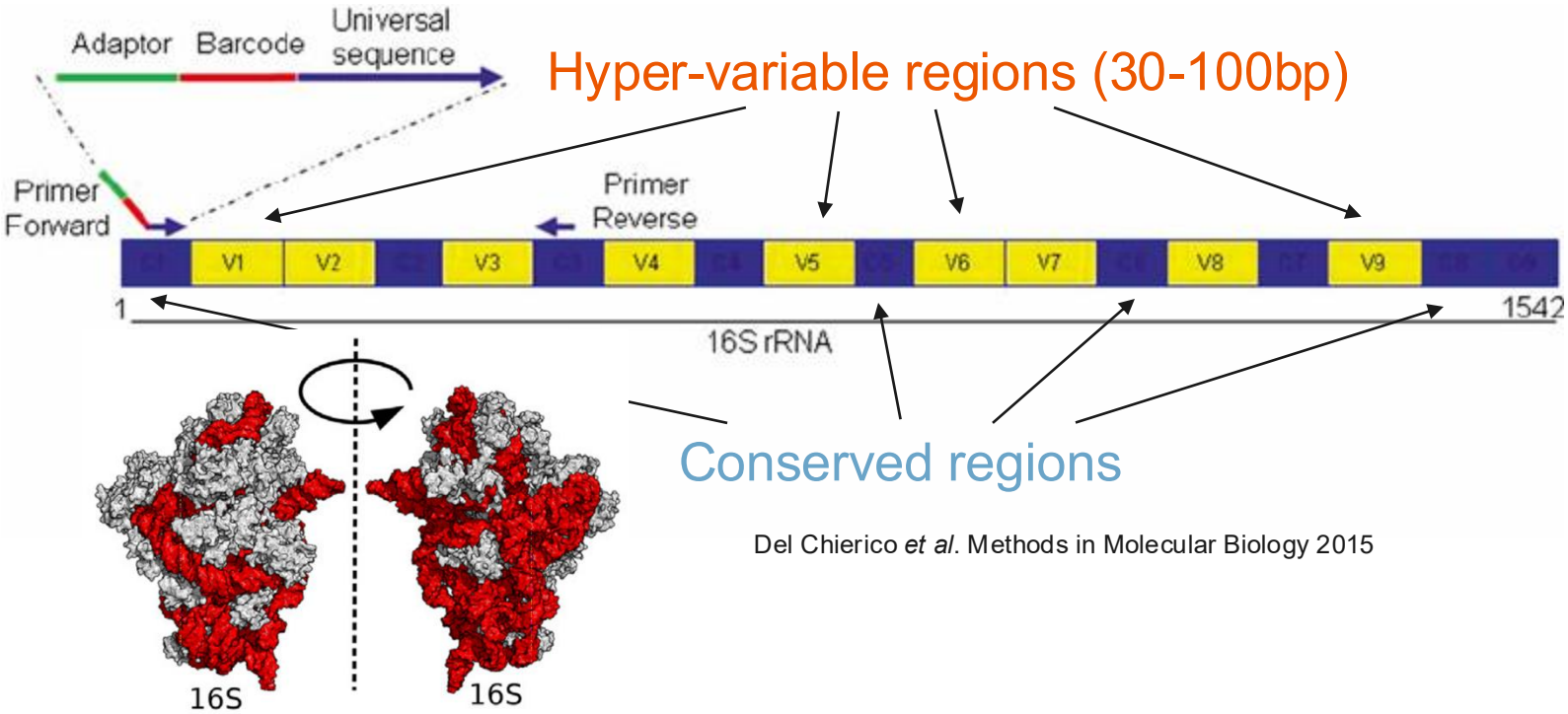
Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences

Pablo Yarza, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B. Whitman, Jean Euzéby, Rudolf Amann & Ramon Rosselló-Móra

- Based on annotated **16S rRNA genes**
  - 94.5% similarity for genus
  - 86.5% similarity for family
  - 75% similarity for phylum
- Near-complete sequences are required to get accurate classification

Ribosomal RNA analysis

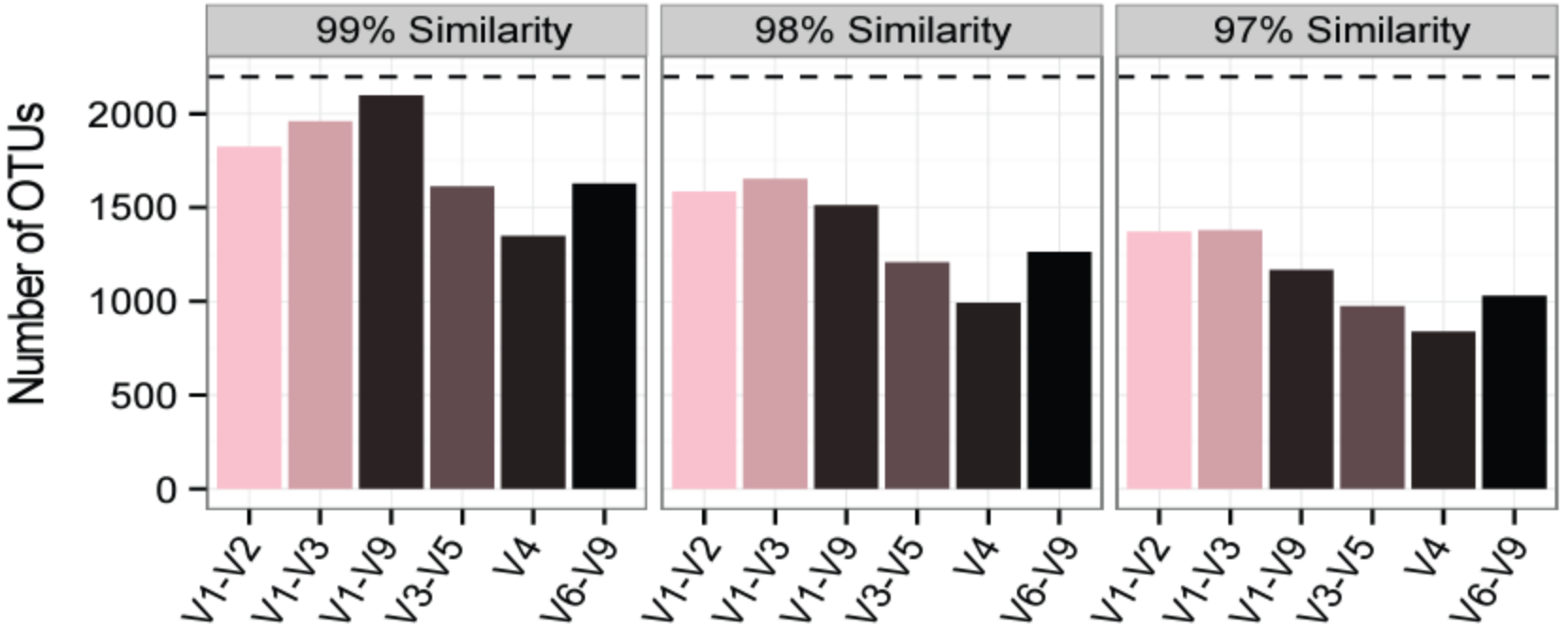
16S rRNA in prokaryotes



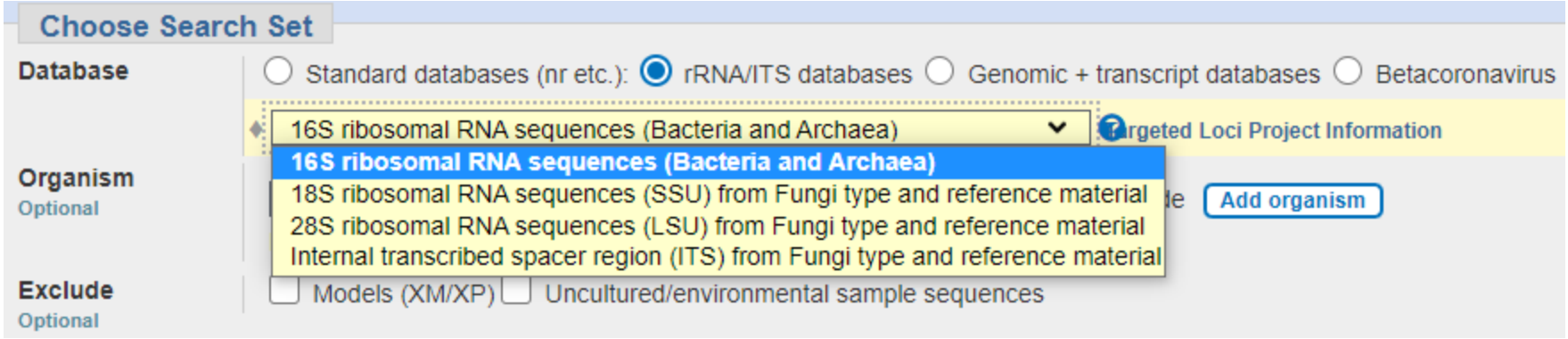
Del Chierico *et al*. Methods in Molecular Biology 2015

Ramazzotti and Bacci, Metagenomics: Perspectives, Methods, and Applications,103-119 (2018)

Different variations across hypervariable regions



rRNA sequence alignment



- Endpoint of rRNA sequence analysis is taxonomic assignment
- **Short read**: accurate only up to family and genus
- **Long read**: species and sub-species level classification is possible

rRNA databases

silva
high quality ribosomal RNA databases

Home SILVAngs Browser Search ACT Download Doc

SILVA

Welcome to the SILVA rRNA database project

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***).

SILVA are the official databases of the software package **ARB**.

rdp

ANNOUNCEMENTS

RDP News

12/12/2018 RDP and Fungene Pipelines are back

RDP Release 11, Update 5 :: September 30, 2016

3,356,809 16S rRNAs :: 125,525 Fungal 28S rRNAs
Find out what's new in RDP Release 11.5 ***here***.

GREENGENES

The Greengenes Database

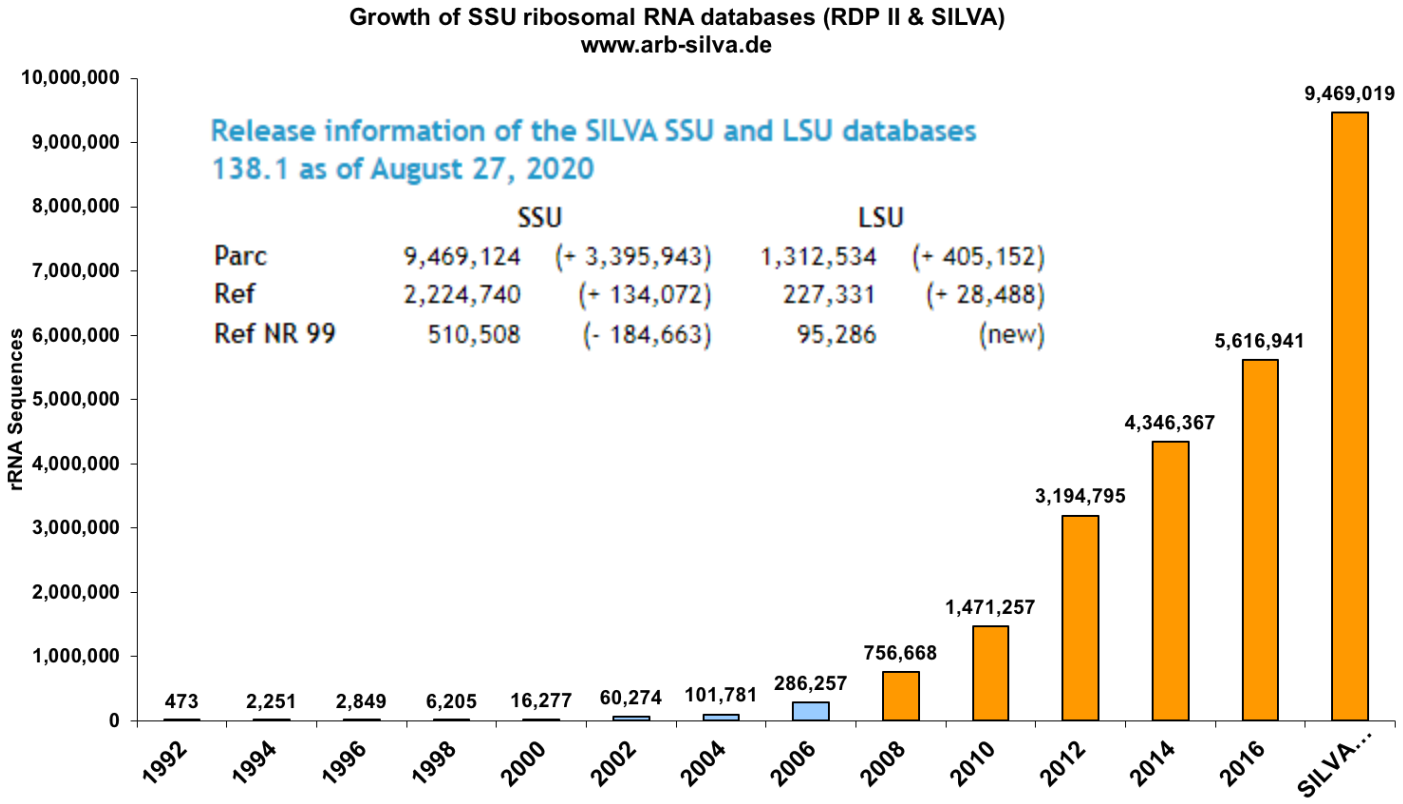
Browse links below to download versions of the Greengenes 16S rRNA gene database or experimental datasets created with the PhyloChip 16S rRNA microarray. Beware that these publicly available versions of the Greengenes database utilize taxonomic terms proposed from phylogenetic methods applied **years ago between 2012 and 2013**. Since then, a variety of novel phylogenetic methods have been proposed for Archaea and Bacteria. For a recent example see Yokono 2018.

unite
community

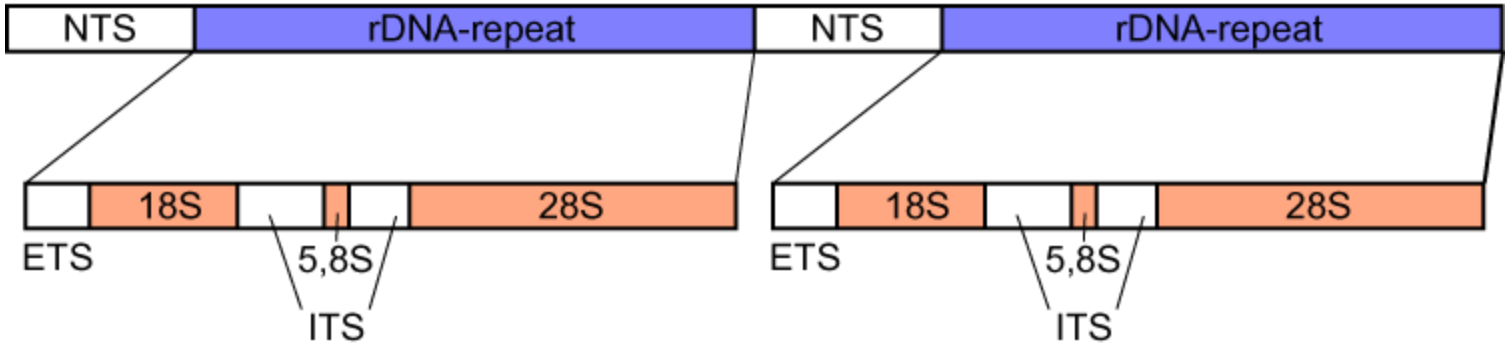
rDNA ITS based identifica

Current version: **8.2**; Last updated: 2020-01-15 (**read more**)
Number of ITS sequences (UNITE+INSD): **2 480 043**; Number of UNITE fungal Sp

Explosion of known rRNA sequences



Internal transcribed spacer (ITS) in fungi and algae

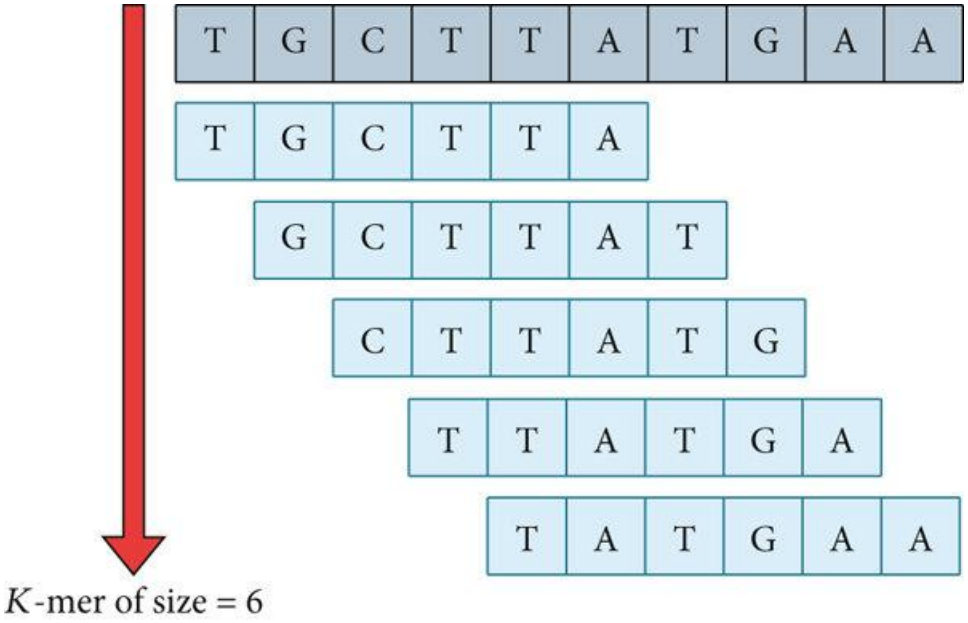


wikipedia.com

- Located between rRNA repeats
- Typically study ITS1 and ITS2 regions
- 400-1000 bp

DNA *k*-mers

k-mer: substring of length *k*



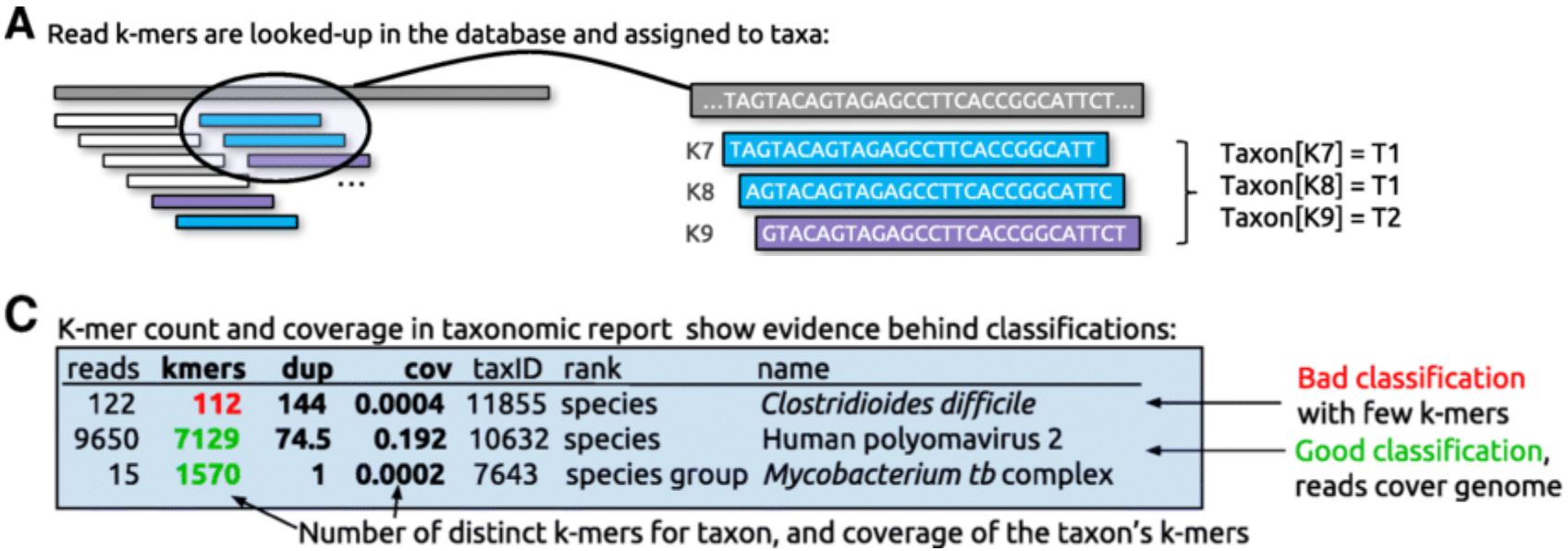
- Similar to the concept of **word** in BLAST
- A long alignment must also contain shorter matches
- **How about searching for short matches instead**?

Gunasekaran, H. et al. Computational and Mathematical Methods in Medicine (2021)

k-mer-based taxonomic classification

- We want to estimate P(genus | observed w_1, w_2, ..., w_n), where w_i's are *k*-mer
- But P(observed w_1, w_2, ..., w_n | genus) is much easier to calculate since we know the rRNA sequences
- **Bayes' rule**: P(genus | words) ∝ P(words | genus) x P(genus), where P(genus) is the prior probability of observing a genus

k-mer matching result



Breitwieser, F. et al. Genome Biology 19:198 (2018)

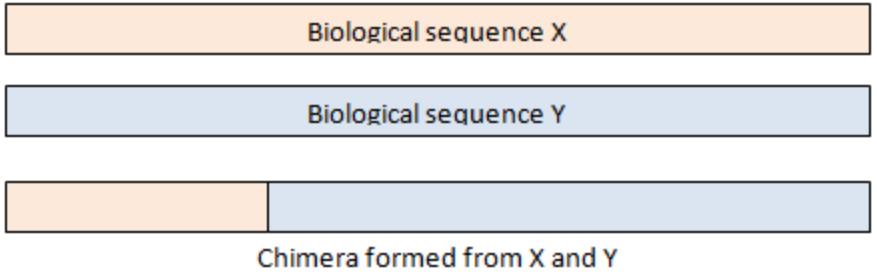
Benefits of using *k*-mer matching

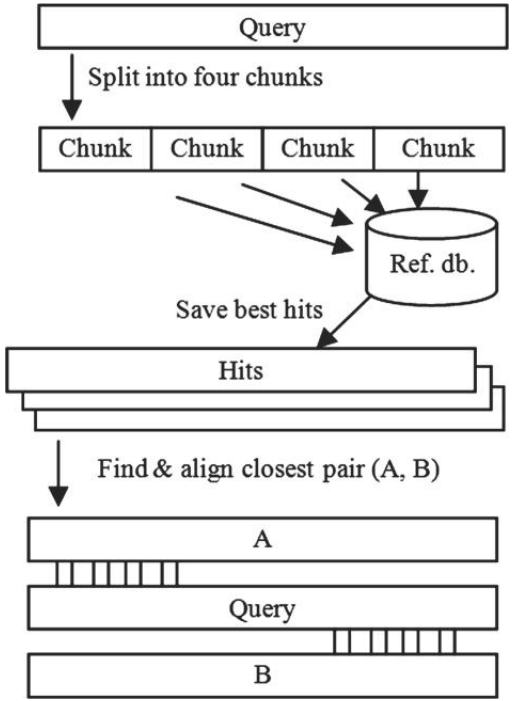
- Much faster than regular alignment
  - Find short perfect matches instead of dynamic programming

- Filter reads for downstream analysis
  - Group reads for assembly
  - Remove host contamination
  - Remove reads mapped to abundant bacteria

Metagenomics read processing

Chimeric reads in amplicon sequencing

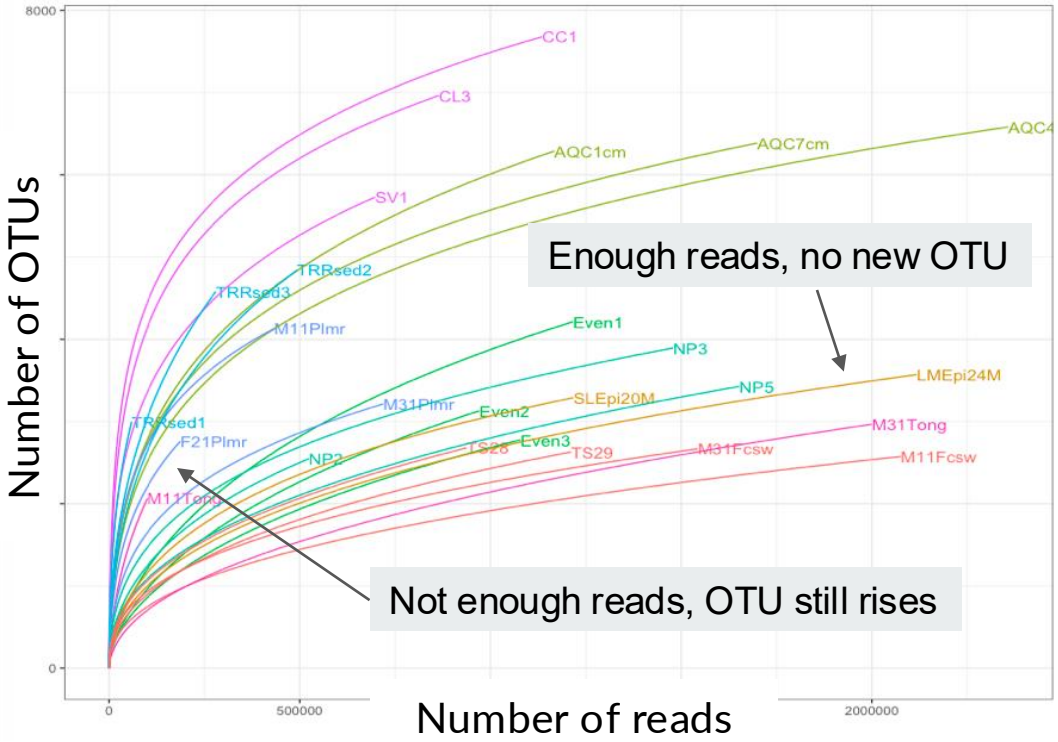




- Produced during PCR amplification
- Detected by alignment different portion of the reads to rRNA databases
  - Same idea as gapped alignment

Edgar *et al.* Bioinformatics 27:2194-2200 (2011)

Rarefaction curve



- One curve = one sample
- Show the gain in unique OTU as more reads are obtained
- **Plateau**: additional reads do not detect more OTU = enough reads

Image from PhyloSeq GitHub

Complexity of microbiome composition

Diversity indices/ Parameters	Description	Formula
Shannon diversity index (H)	Estimator of species richness and species evenness: more weight on species richness	$H = -\sum_{i=1}^{s}(p_i \ln p_i)$ where s is the number of OTUs and p_i is the proportion of the community represented by OTU i.
Simpson's index (D)	Estimator of species richness and species evenness: more weight on species evenness	$D = \frac{1}{\sum_{i=1}^{s} p_i^2}$ where s is the total number of species in the community and p_i is the proportion of community represented by OTU i.

S = number of taxa

Kim, B.-R. et al. J Microbiol Biotechnol 27:2089-2093 (2017)

- **Richness** = number of distinct species
- **Evenness** = lack of dominant species

Behavior of Shannon entropy

Diversity indices/ Parameters	Description	Formula
Shannon diversity index (H)	Estimator of species richness and species evenness: more weight on species richness	$H = -\sum_{i=1}^{s}(p_i \ln p_i)$ where s is the number of OTUs and p_i is the proportion of the community represented by OTU i.
Simpson's index (D)	Estimator of species richness and species evenness: more weight on species evenness	$D = \frac{1}{\sum_{i=1}^{s} p_i^2}$ where s is the total number of species in the community and p_i is the proportion of community represented by OTU i.

S = number of taxa

Kim, B.-R. et al. J Microbiol Biotechnol 27:2089-2093 (2017)

- Entropy is maximized when $p_i = \frac{1}{S} \rightarrow H_{max} = \ln(S)$
- Entropy is minimized when there is one dominant species $H_{min} = 0$

Behavior of Simpson's index

Diversity indices/ Parameters	Description	Formula
Shannon diversity index (H)	Estimator of species richness and species evenness: more weight on species richness	$H = -\sum_{i=1}^{s}(p_i \ln p_i)$ where s is the number of OTUs and p_i is the proportion of the community represented by OTU i.
Simpson's index (D)	Estimator of species richness and species evenness: more weight on species evenness	$D = \frac{1}{\sum_{i=1}^{s} p_i^2}$ where s is the total number of species in the community and p_i is the proportion of community represented by OTU i.

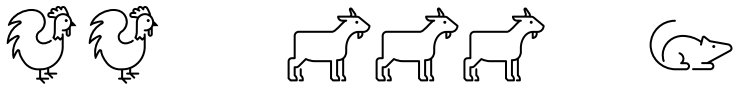
S = number of taxa

Kim, B.-R. et al. J Microbiol Biotechnol 27:2089-2093 (2017)

- D is maximized when $p_i = \frac{1}{S}$ → $D_{max} = S$
- D is minimized when there is only dominant species $D_{min} = 1$

Similarity between microbial compositions

Bray-Curtis dissimilarity (intersection over union)

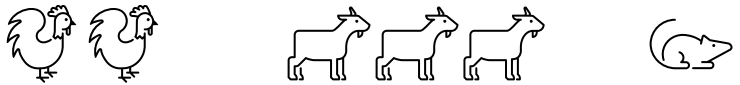
Sample 1 

Sample 2 

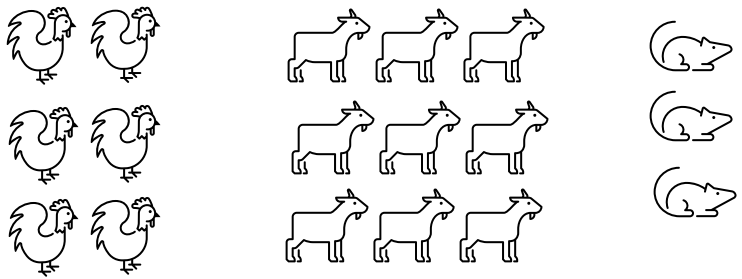
- S_1 = # of individuals in sample 1 = 6
- S_2 = # of individuals in sample 2 = 5
- Overlap = 1 chicken + 2 goat + 1 mouse = 4
- Bray-Curtis dissimilarity = $1 - \frac{2 \times \text{Overlap}}{S_1 + S_2} = 1 - \frac{8}{11} = \frac{3}{11}$

Impact of sequencing depth on Bray-Curtis

Sample 1

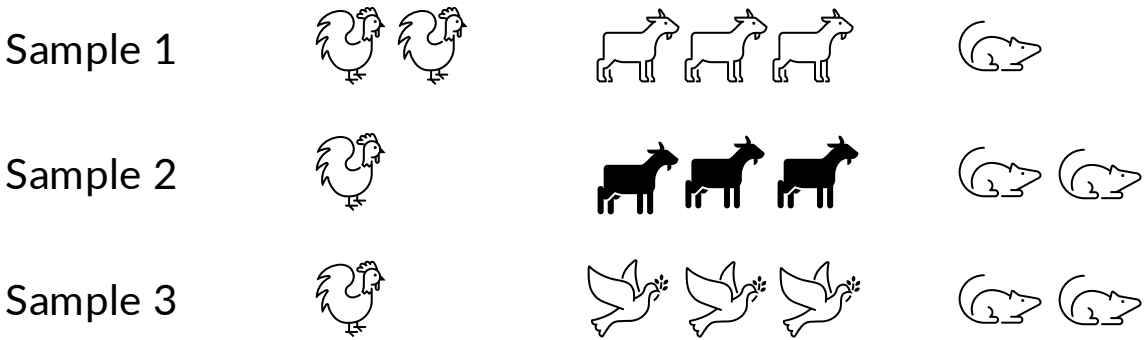


Sample 2



- Bray-Curtis is only suitable between samples with similar depths
- Both samples have 2:3:1 composition
- Bray-Curtis dissimilarity = $1 - \frac{2 \times \text{Overlap}}{S_1 + S_2} = 1 - \frac{12}{24} = \frac{1}{2}$

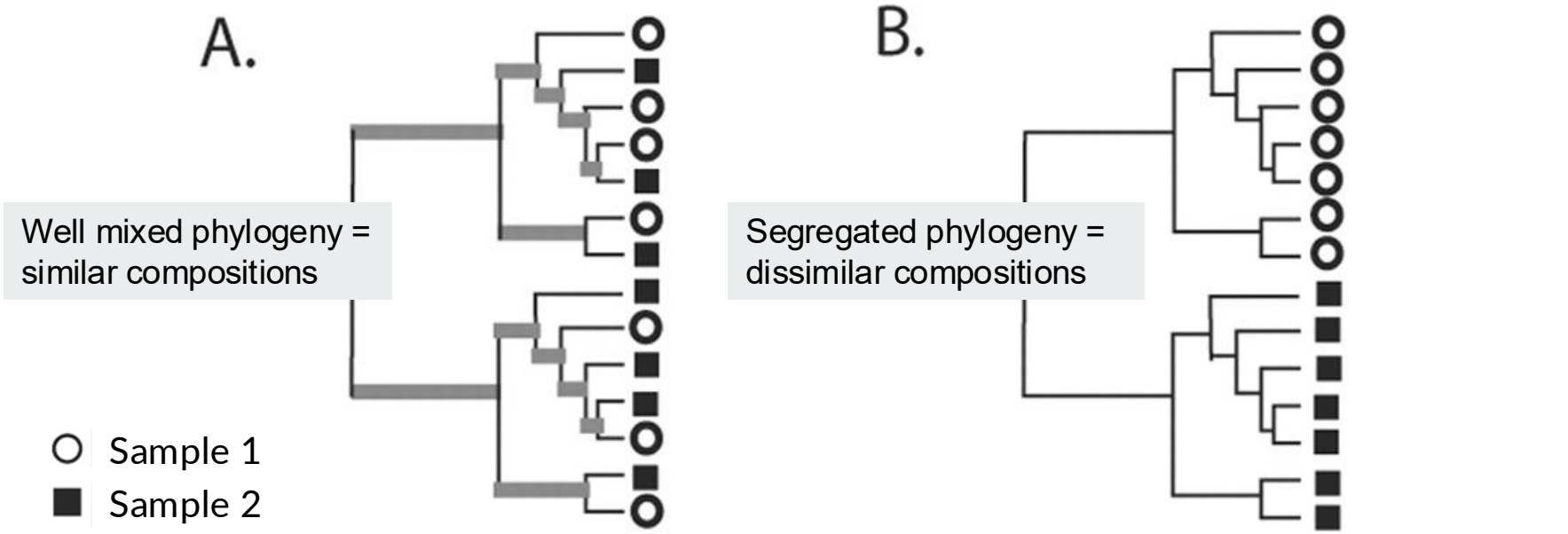
Bray-Curtis doe not consider taxonomic similarity



- Difference between white goats and black goats is treated the same as the difference between goats and birds
- Discordant with the view of functional interpretation of metagenomics

UniFrac distance

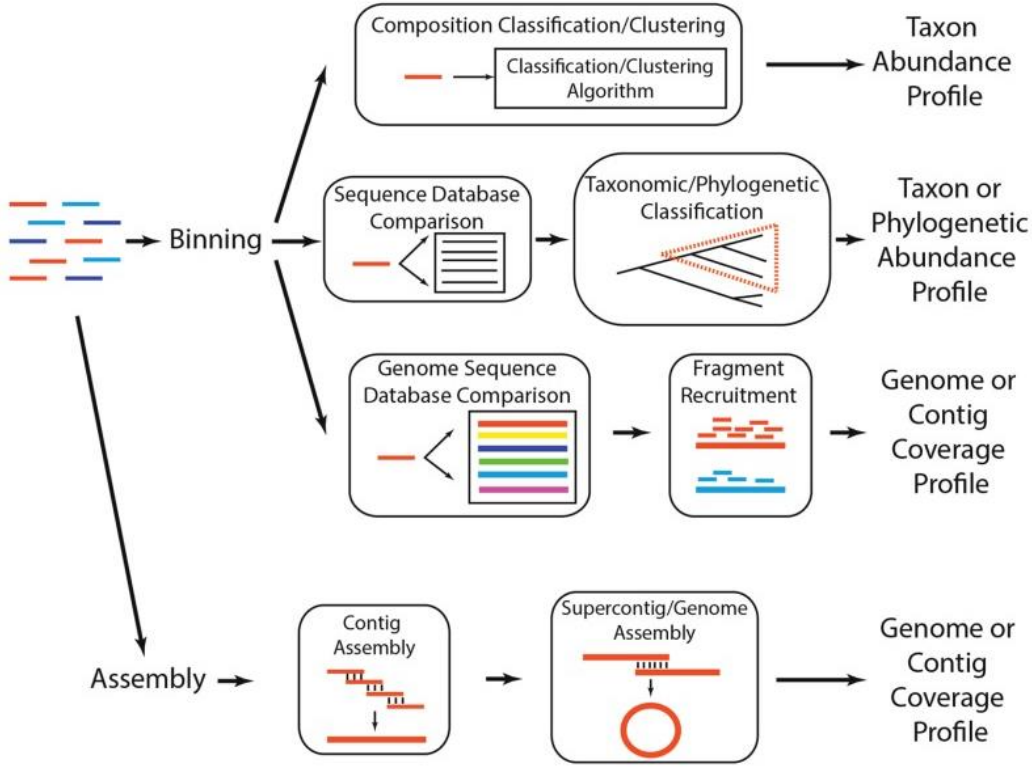
Lozupone, C. and Knight, R. Applied and Environmental Microbiology 71 (2005)



- UniFrac = fraction of shared phylogenetic branches between samples
- Can be weighted or unweighted by taxa abundances

Shotgun metagenomics

Shotgun metagenomics

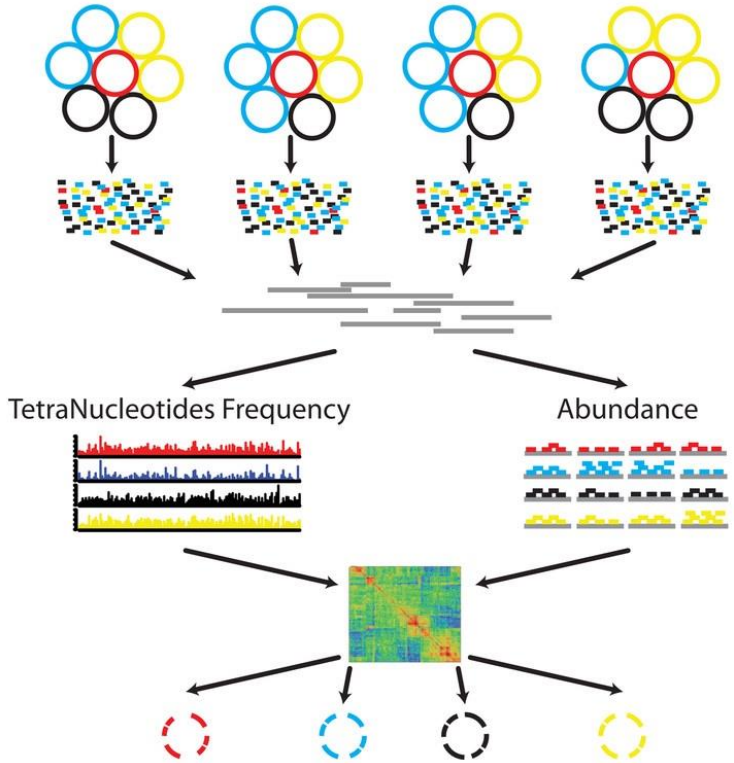


- **Binning**: grouping DNA/RNA from the same host organisms together

- Direct assembly is possible for abundant species and long reads

Sharpton *et al*. Front. Plant Sci. 5:209 (2014)

Read binning strategies

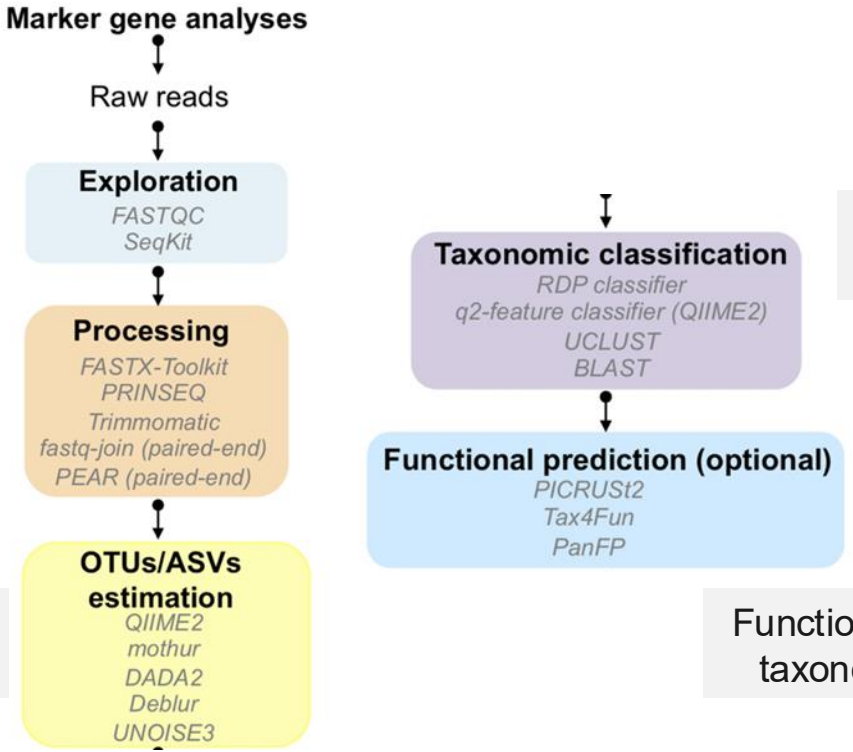


- ***k*-mer**: reads originated from the same species have similar *k*-mer profiles (like GC content)
- **Abundance**: reads originated from the same organism have correlated abundance profiles across samples

Kang *et al.* PeerJ (2015)

Resources for metagenomics

Amplicon analysis pipeline (e.g., 16S rRNA)

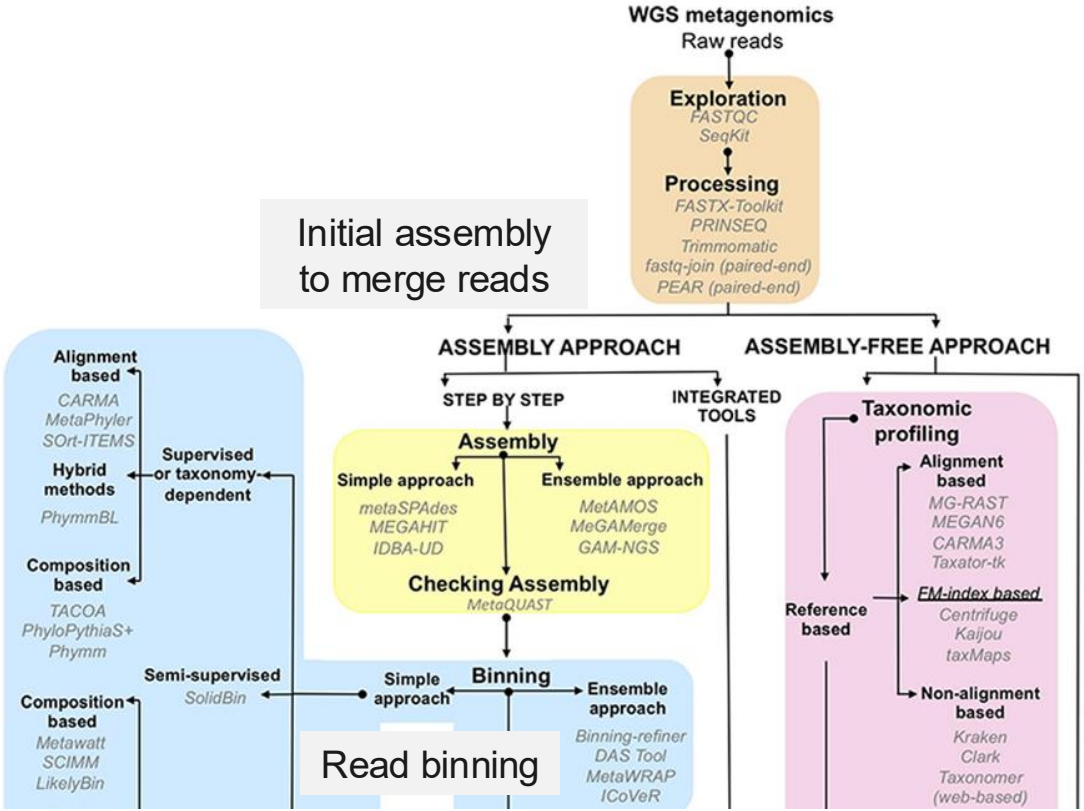


k-mer-based classification or alignment

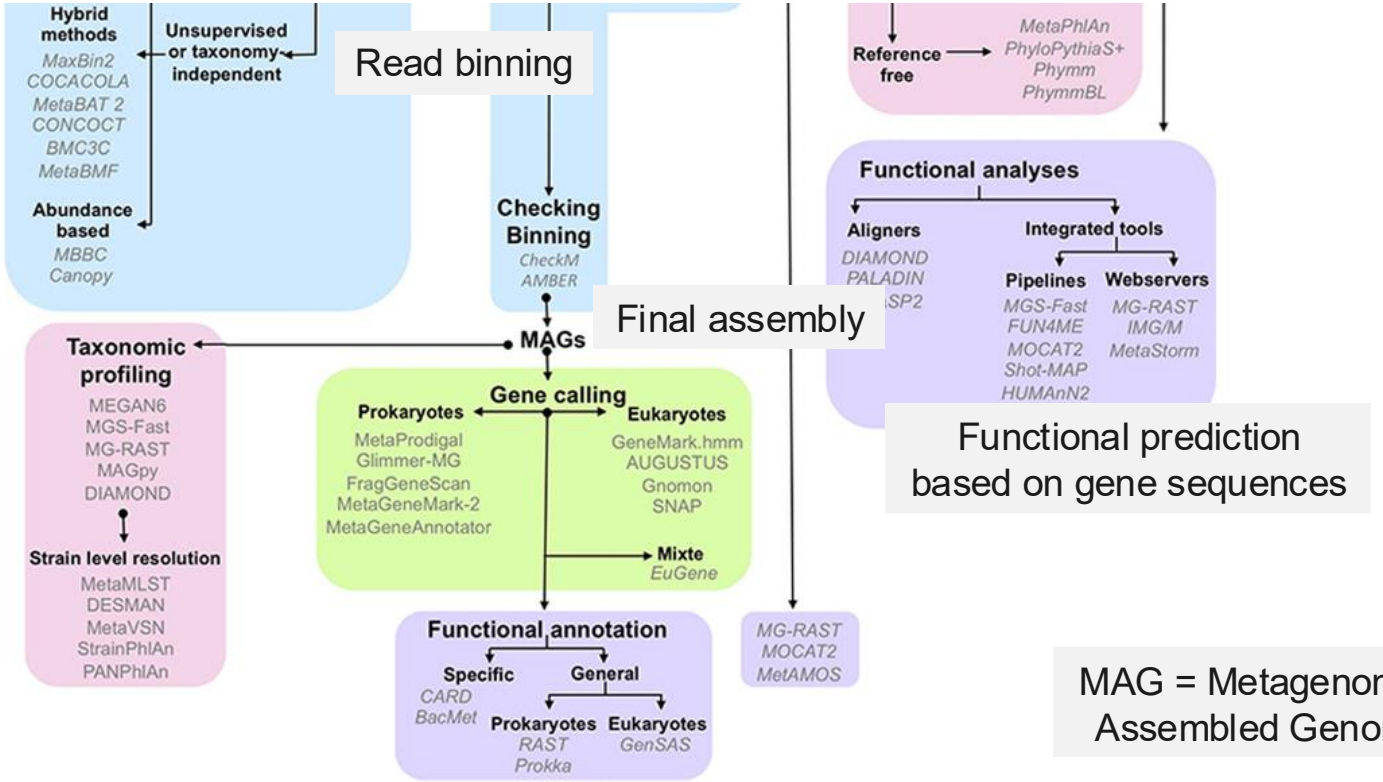
Remove noises, reads with high error, and chimera

Functional prediction based on taxonomy-gene association

Shotgun analysis pipeline



Shotgun analysis pipeline (continued)

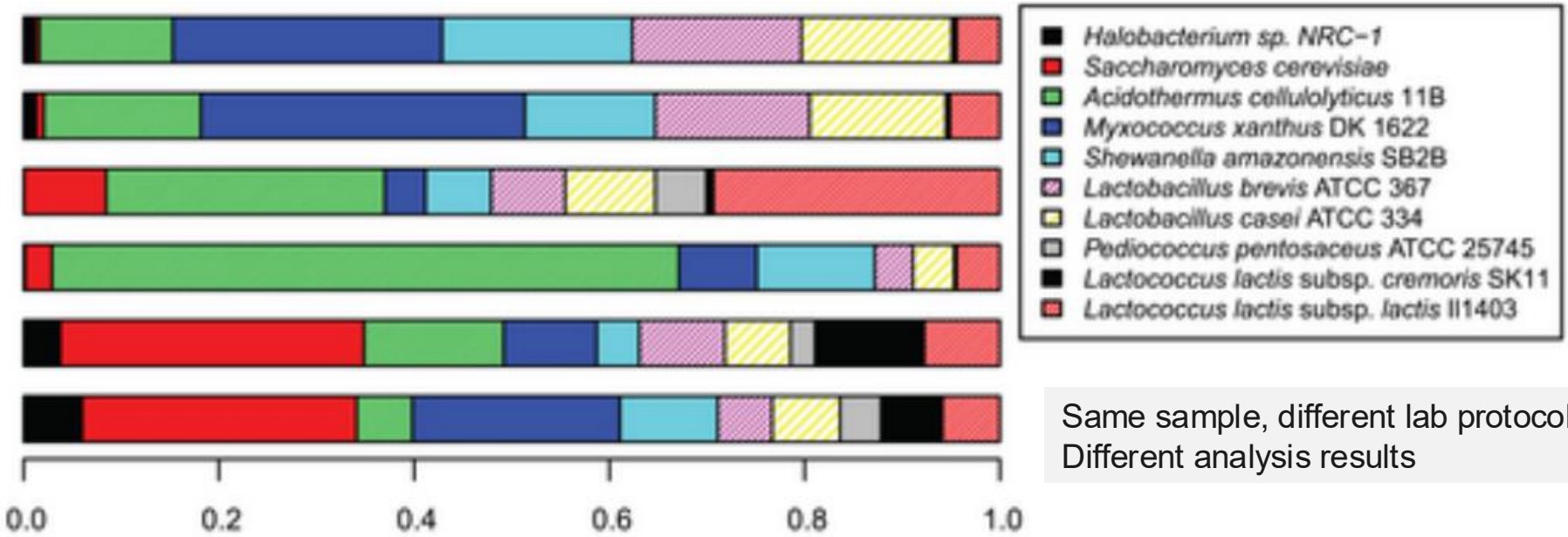


Functional prediction based on gene sequences

MAG = Metagenomics Assembled Genome

Caution: Variability of microbiome

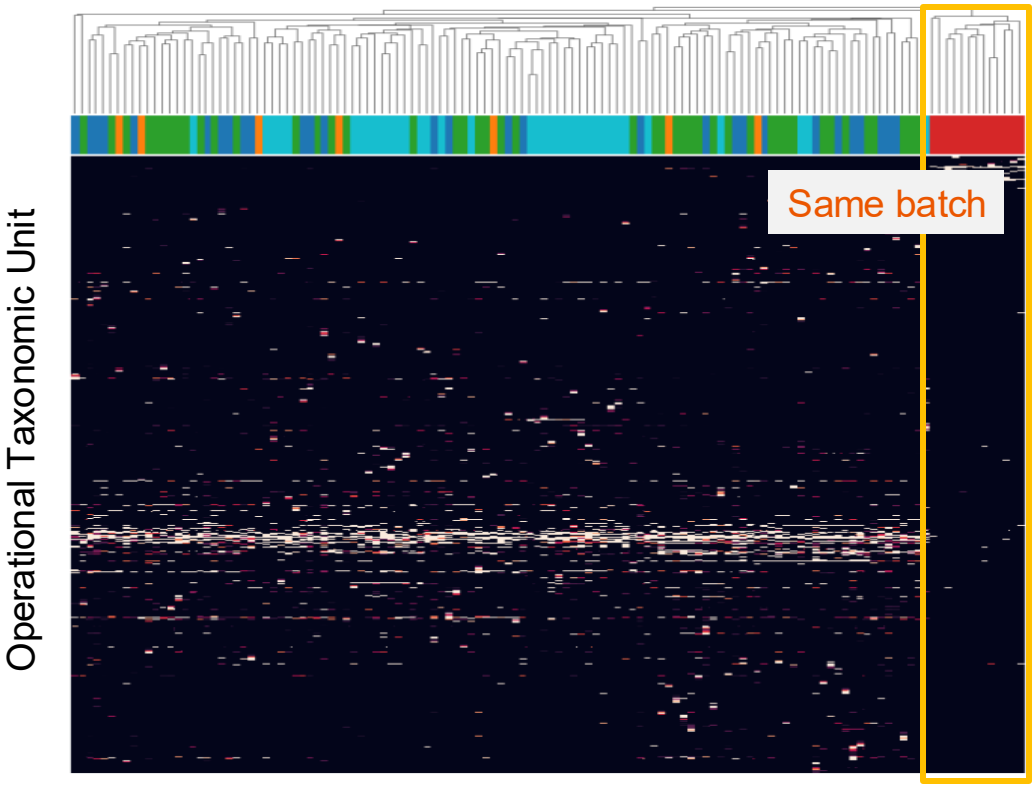
Microbiome profile influenced by lab protocol



Same sample, different lab protocols
Different analysis results

Source: http://www.cbs.dtu.dk/courses/27626

Batch effects



- A common problem
- **Good batch design**: distribute samples of the same class across batches
- Don’t group them in the same batch

Any question?

- See you next time