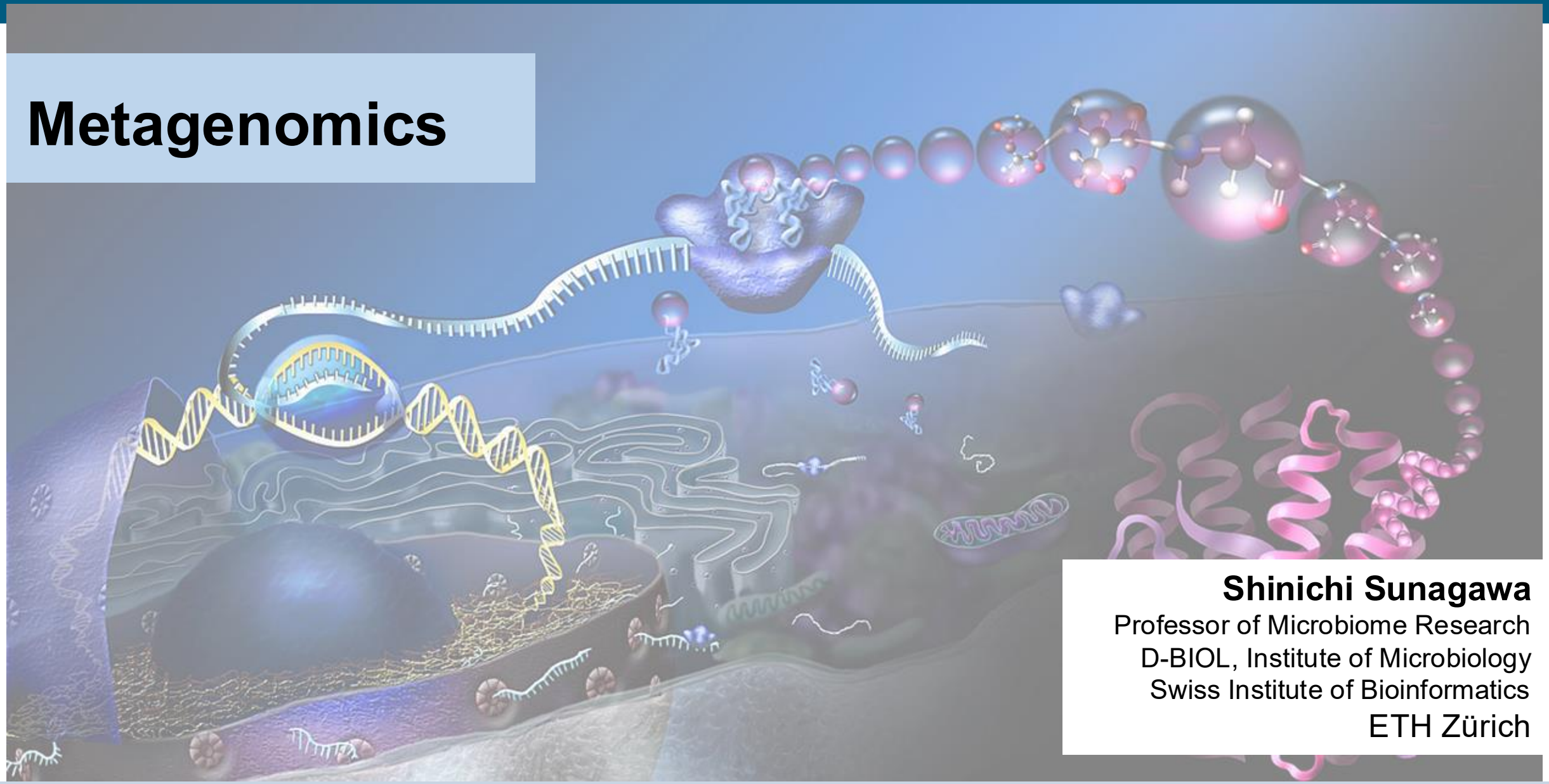
Metagenomics

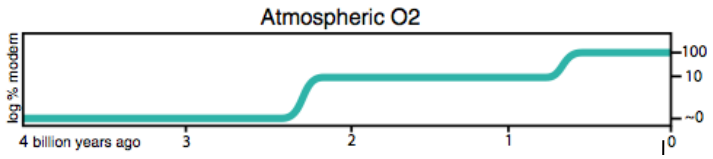


Shinichi Sunagawa
Professor of Microbiome Research
D-BIOL, Institute of Microbiology
Swiss Institute of Bioinformatics
ETH Zürich

Evolution and significance of microbiomes

From the origin of life to today

A



Microorganisms

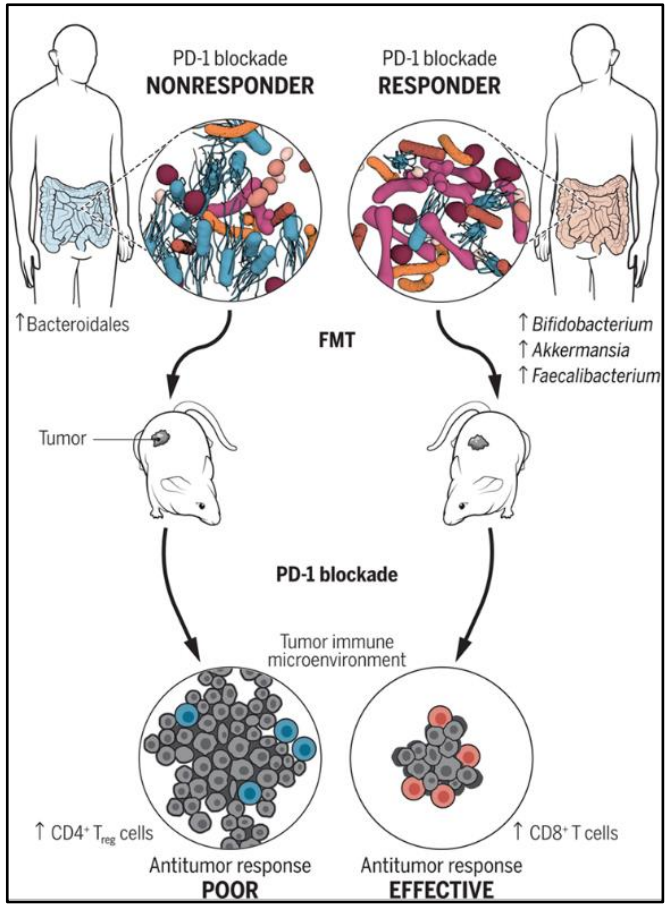
- originated some 3.8 billion years ago
- drive biogeochemical cycles of elements (C, N, P, S, etc.)
- transform energy and biomass

Significance (examples):

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

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

Describing microbial communities – Example 1



GRAPHIC: V. ALTOUNIAN/*SCIENCE*

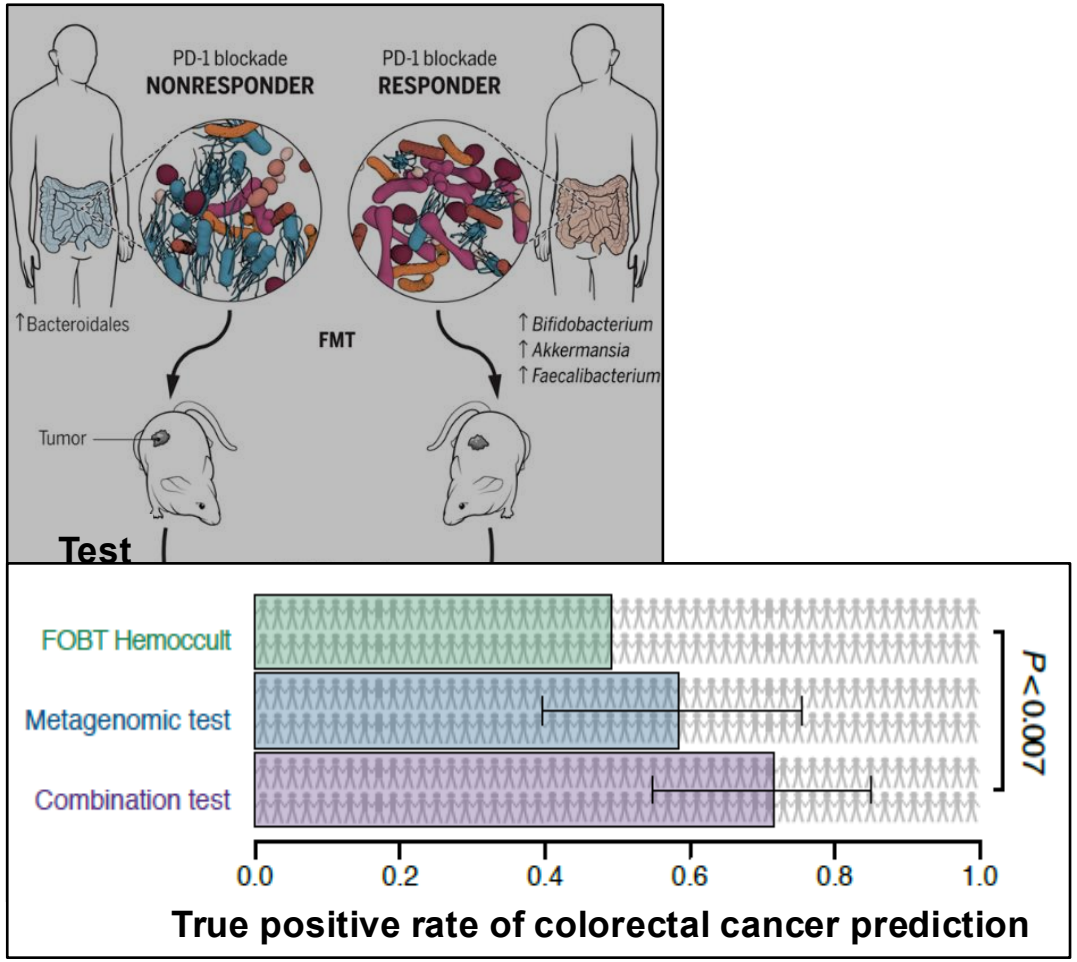
Gut microbial community compositions

- can alter efficacy of treatments

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

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

Describing microbial communities – Example 1



Gut microbial community compositions

- can alter efficacy of treatments

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

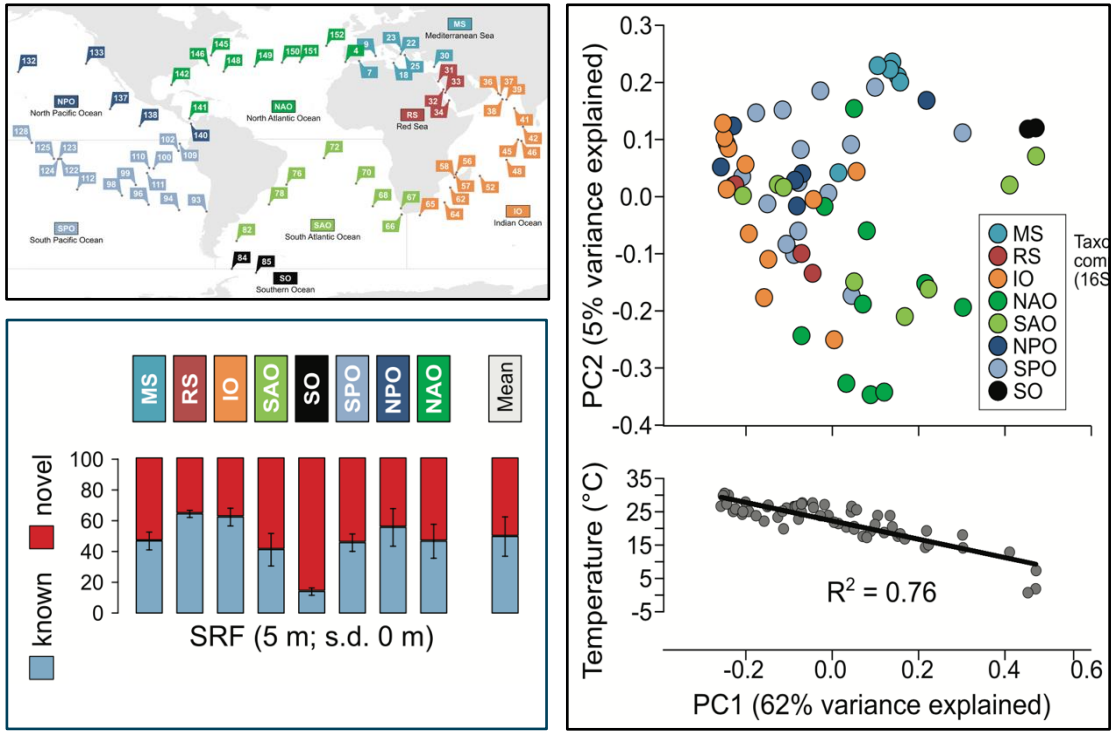
Routy et al., Gopalakrishnan et al., and Matson et al., Science 2018

- can be indicative for diseases

→ Statistical models of fecal microbiota composition can predict colorectal cancer

Zeller et al., 2014; Wirbel et al., 2019

Describing microbial communities – Example 2



Ocean microbial community compositions

- reveal previously unknown organisms and genes (left bottom)

→ implying novel taxa, enzymes and functions

- similarities between communities not determined by geography (right top)

→ but strongly driven by temperature (bottom right)

Sunagawa et al., 2015

Overview

Microbial community structure

- microbial taxonomy and operational taxonomic units
- quantification of microbial community members
- diversity within a microbial community

Differences between microbial communities

- taxonomic differences between microbial communities
- differentially abundant features (e.g., taxa, genes, functions)

Working with microbial community genes and genomes

- reconstruction of microbial community genomes
- gene functional differences between microbial communities

Review: microbial taxonomy

- Microbiologist have adopted the concept of hierarchical taxonomic ranks:

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

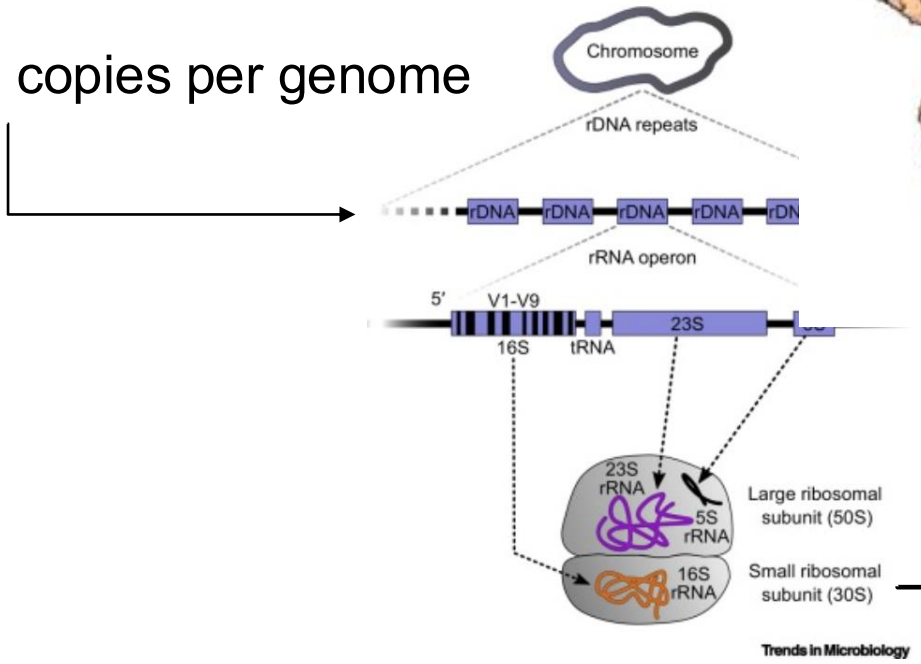
Rank	NCBI
Domain	Bacteria
Phylum	Pseudomonadota
Class	Gammaproteobacteria
Order	Enterobacterales
Family	Enterobacteriaceae
Genus	*Escherichia*
Species	*Escherichia coli*

- 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**

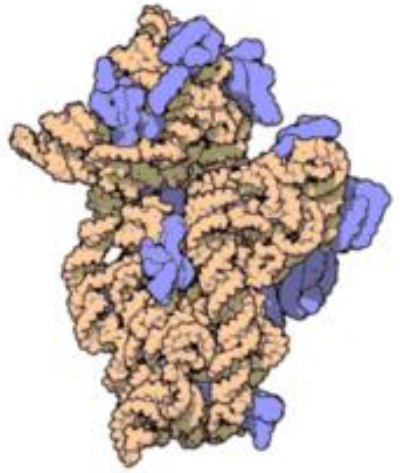
→ Today, DNA sequencing and computational comparison is the method of choice to classify microbial organisms and to study their evolutionary relatedness

16S rRNA-based Operational Taxonomic Units (OTUs)

- 16S rRNA
  - encoded in genomes of all bacteria and archaea conserved function as integral part of the protein synthesis machinery
  - can be present in multiple copies per genome



30S small subunit of ribosomes in prokaryotes

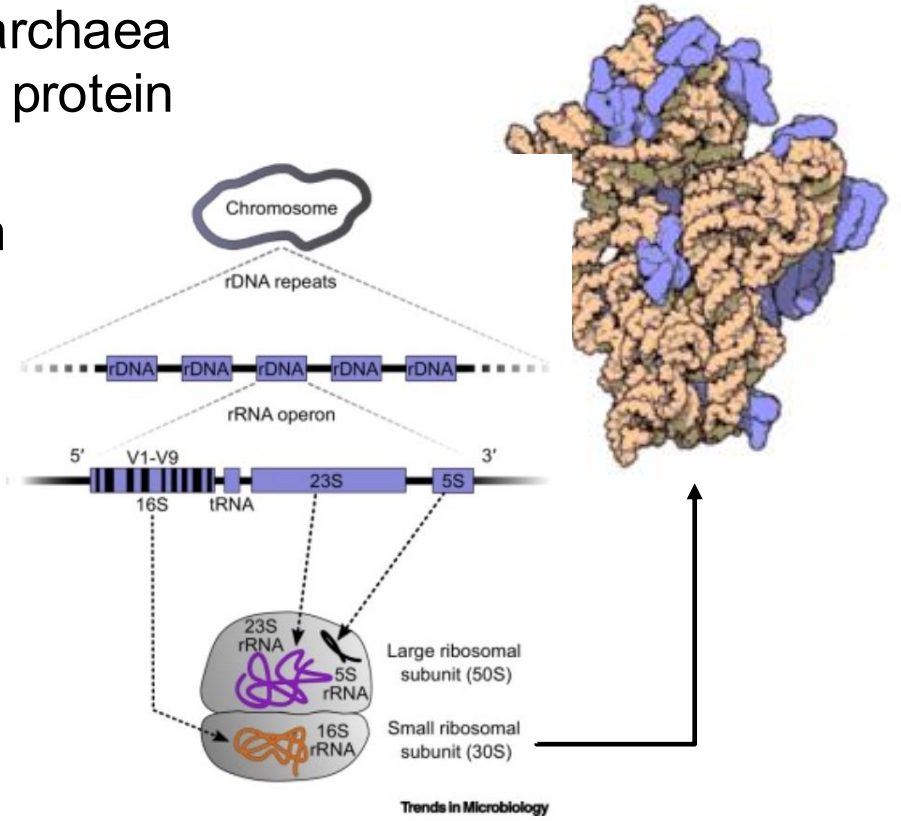


Protein (blue)
rRNA (orange)

16S rRNA-based Operational Taxonomic Units (OTUs)

- 16S rRNA
  - encoded in genomes of all bacteria and archaea conserved function as integral part of the protein synthesis machinery
  - can be present in multiple copies per gen
  - similar mutation rate: → molecular clock

30S small subunit of ribosomes in prokaryotes



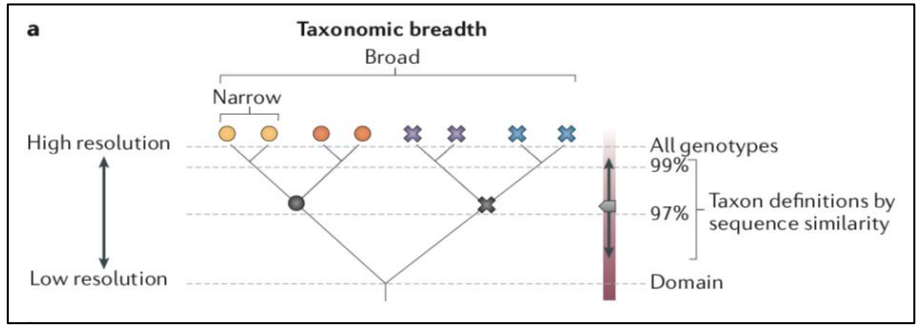
Protein (blue)
rRNA (orange)

Lavrinienko et al., Trends Microbiol, 2021; Wikipedia

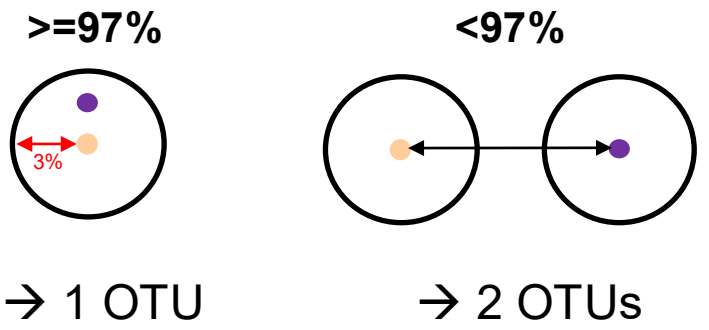
16S rRNA-based Operational Taxonomic Units (OTUs)

- 16S rRNA
  - encoded in genomes of all bacteria and archaea conserved function as integral part of the protein synthesis machinery
  - similar mutation rate: → molecular clock
- Used as proxy for phylogenetic relatedness

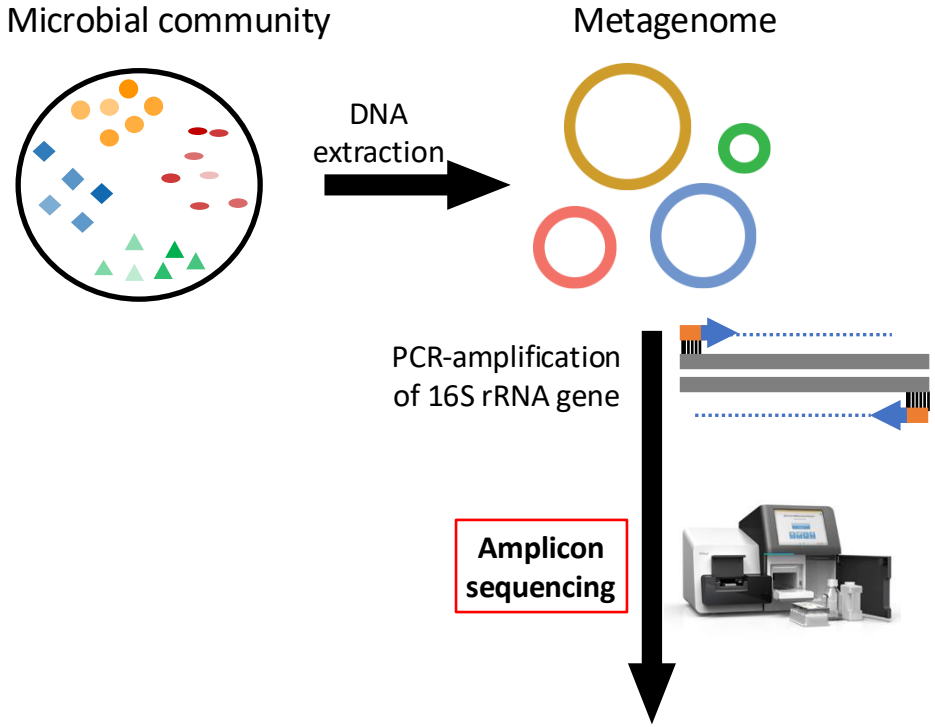
- Owing to lack of prokaryotic species definition, 97% sequence similarity is often used to define 'species'-like:
  "**Operational Taxonomic Units" (OTUs)**



Identity of 16S rRNA gene sequences



Amplification of 16S rRNA gene fragments by PCR



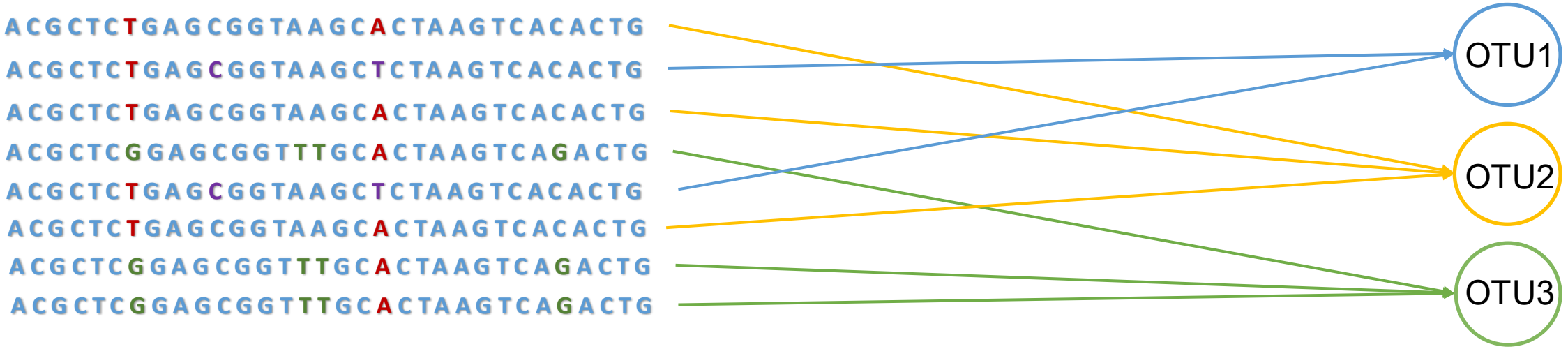
16S rRNA amplicons

ACGCTCTGAGCGGTAAGCACTAAGTCACACTG
ACGCTCTGAGCGGTAAGCTCTAAGTCACACTG
ACGCTCTGAGCGGTAAGCACTAAGTCACACTG
ACGCTCGGAGCGGTTTGCACTAAGTCAGACTG
ACGCTCTGAGCGGTAAGCTCTAAGTCACACTG
ACGCTCTGAGCGGTAAGCACTAAGTCACACTG
ACGCTCGGAGCGGTTTGCACTAAGTCAGACTG
ACGCTCGGAGCGGTTTGCACTAAGTCAGACTG

Who is there?

Quantification of OTU abundances

All amplicons are aligned to best matching OTU 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

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 pairs, please discuss:

Q1: What are the factors that influence the differences between samples?
How could the differences be formally described (i.e., measured in quantitative terms)?

Q2: How may the number of reads per sample impact the results?
What measures can be taken to account for this effect?

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,

where n*i* = the number individuals of the *i*-th OTU
and n = total number of individuals, that is:
p*i* = n*i* / n

Pielou's evenness (J')

$$J' = \frac{H'}{H'_{\max}}$$

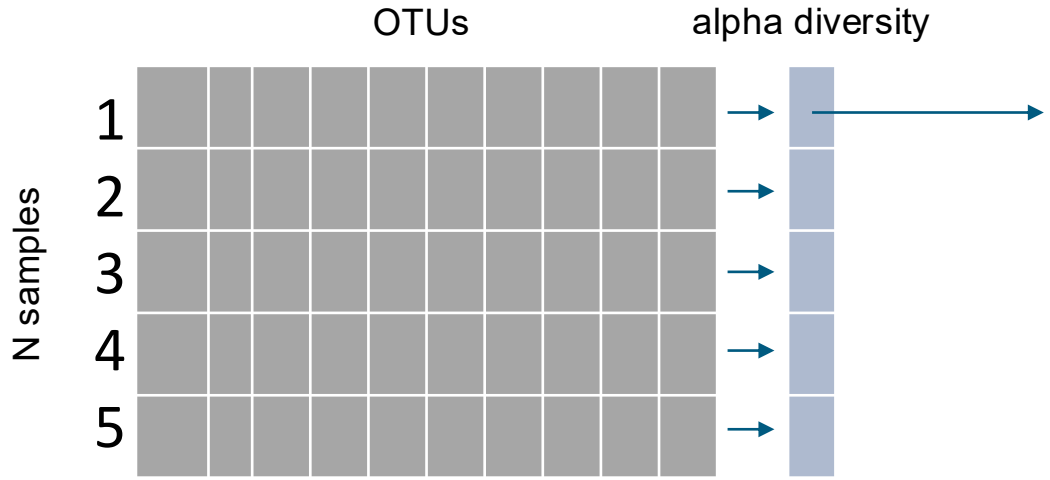
where $$H'_{max} = -\sum_{i=1}^{R} \frac{1}{R} ln \frac{1}{R} = \ln R$$

that is, every species is equally present

Summary

Microbial community composition

- Microbial taxonomy, ASVs and operational taxonomic units (OTUs): **definitions and clustering**
- Counting OTUs: **taxonomic profiling**
- Diversity <u>within</u> a microbial community: **alpha diversity**



<u>Shannon diversity index</u>
Function of:
- Richness (number of detected OTUs)
- Evenness (frequency distribution of detected OTUs)

Normalization/rarefaction

Overview of the Metagenomics part

Microbial community structure

- microbial taxonomy and operational taxonomic units
- quantification of microbial community members
- diversity within a microbial community

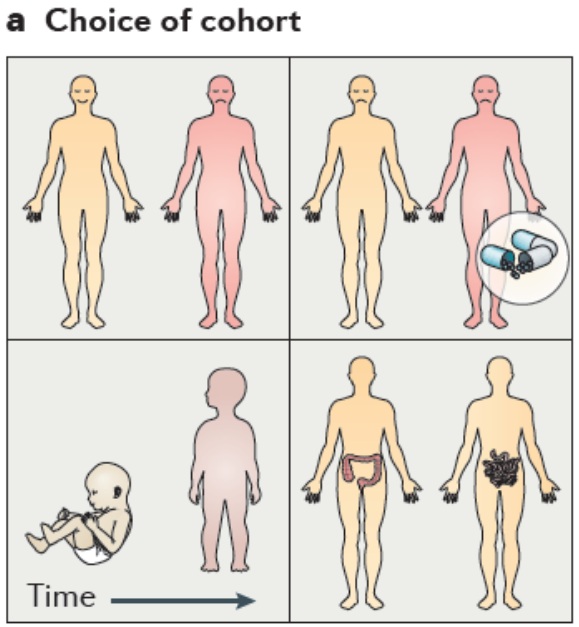
Differences between microbial communities

- taxonomic differences between microbial communities
- differentially abundant features (e.g., taxa, genes, functions)

Working with microbial community genes and genomes

- reconstruction of microbial community genomes
- gene functional differences between microbial communities

Microbiome-wide association studies are analogous to GWAS



Analogous to GWAS, the microbiome can be linked to:

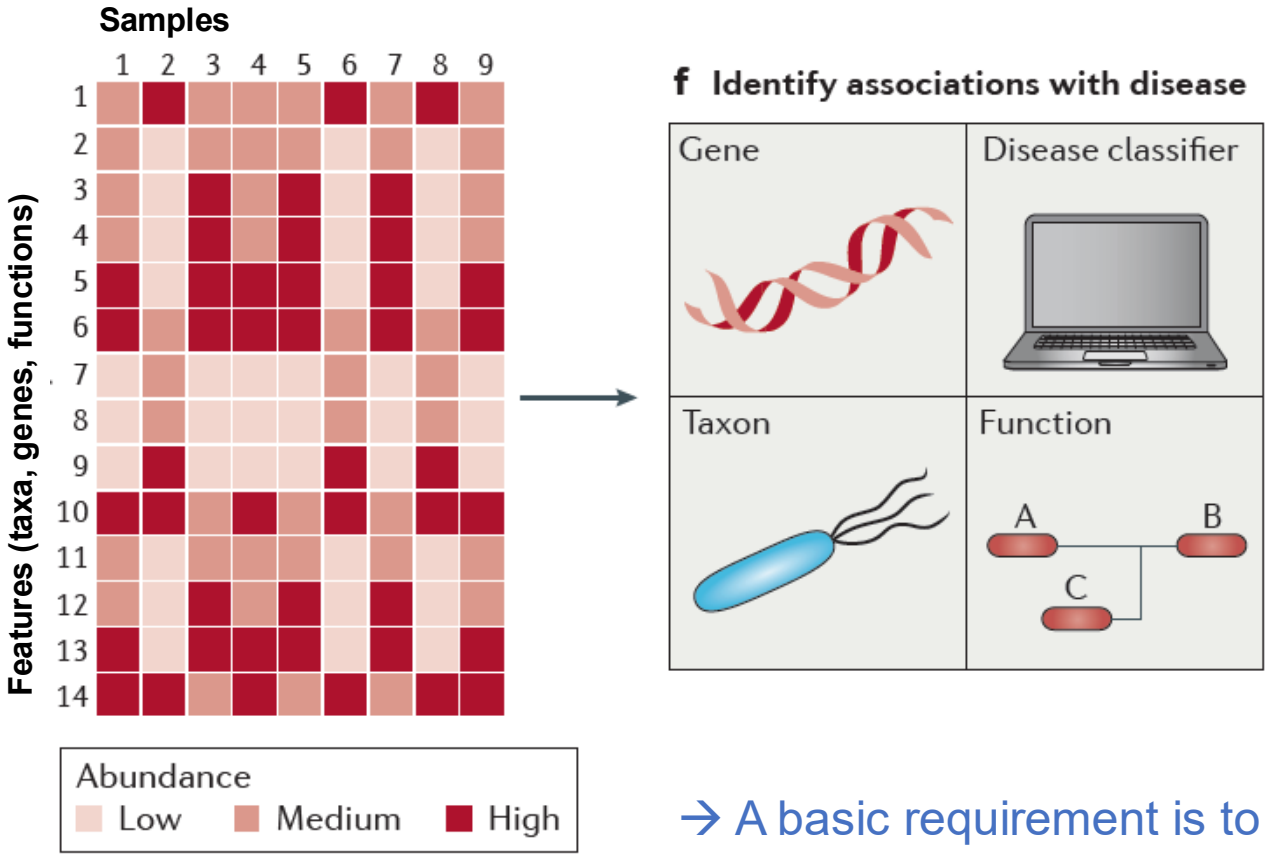
- groups of individuals and/or health states
- differential response to drugs (or nutrition)
- organismal development (or disease progression)
- differences between body sites

Examples:

- asymptomatic individuals vs colorectal cancer patients
- cardiatic drug digoxin inactivation by *Eggerthella lenta*
- *Bifidobacterium* spp. decrease with age
- body-site specific taxa

Wang and Jia, NRM, 2016

Microbiome-wide association studies are analogous to GWAS

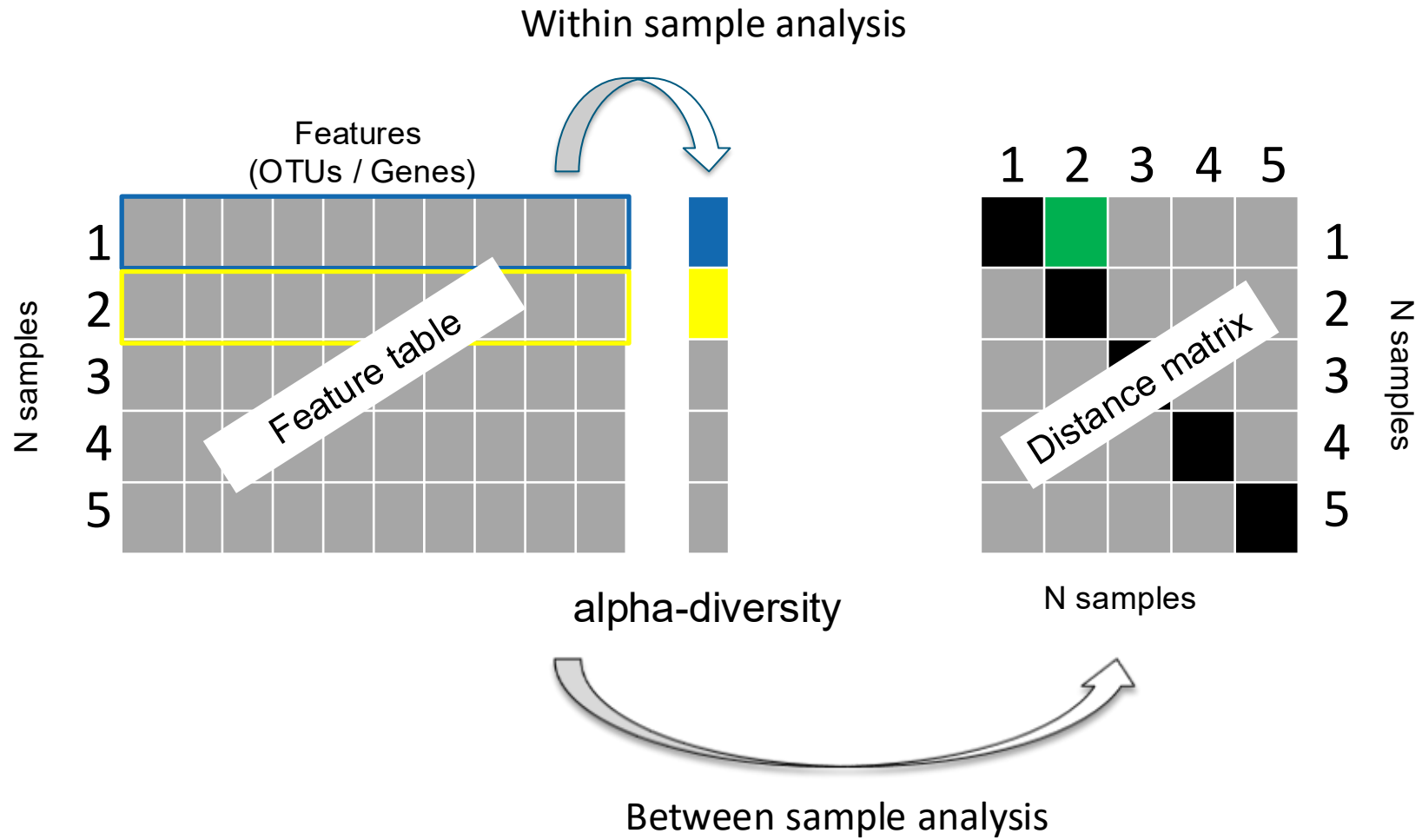


- DNA sequencing reads are analyzed to quantify the abundance of taxa, genes or functions (or to generalize: “features”)
- Abundance tables are analyzed to determine differentially abundant features, e.g., between groups of samples, to identify biomarkers
- Machine learning is used to classify samples and/or to identify relationships between the microbiome and clinical/environmental phenotypes

→ A basic requirement is to quantify the differences between samples

Wang and Jia, NRM, 2016

How many of which OTUs/genes are found in a sample?
How similar are the OTUs/gene compositions between samples?



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	0
2	1
3	1
4	0
5	4

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

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

Mini-quiz

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

a) Differences in the evenness between two samples are not accounted for

b) Differences in the abundance of OTUs shared between samples are not accounted for

c) Differences in the abundance of OTUs not shared between two samples are not accounted for

d) All of the above

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

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

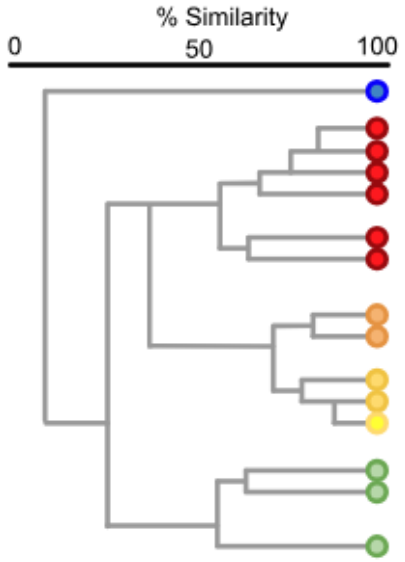
Hellinger: $$D = \sqrt{\sum_{i=1}^{S}\left(\sqrt{\frac{a_i}{a_+}} - \sqrt{\frac{c_i}{c_+}}\right)^2} \text{ with } a_+ = \sum_{i=1}^{S} a_i$$

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

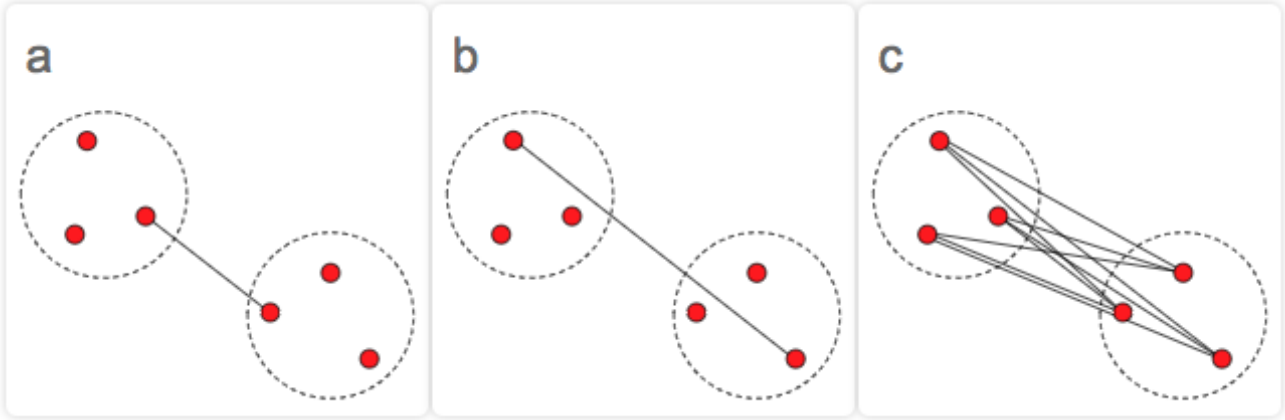
Visualize dissimilarities between microbial communities

- For 2 (xy) or 3 (xyz) variables, data can be easily visualized in two or three dimensional space
- For multi (n>3) dimensional data, distances can be ‘projected’ into lower dimensional space

Hierarchical clustering



Linkage algorithms

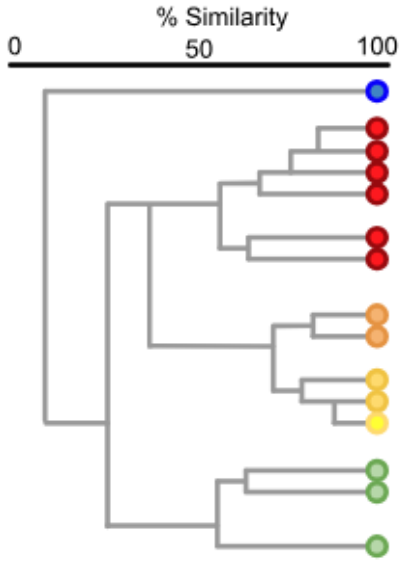


single complete average

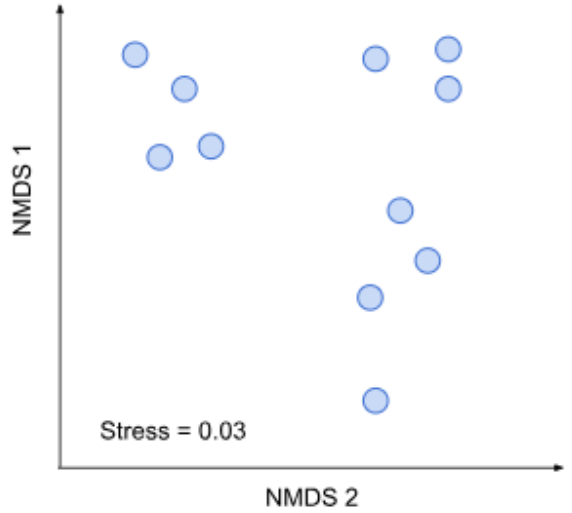
Visualize dissimilarities between microbial communities

- For 2 (xy) or 3 (xyz) variables, data can be easily visualized in two or three dimensional space
- For multi (n>3) dimensional data, distances can be ‘projected’ into lower dimensional space

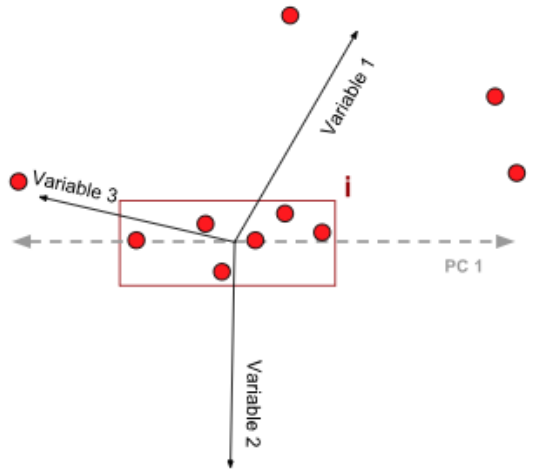
Hierarchical clustering



Non-metric dimensional scaling (NMDS)

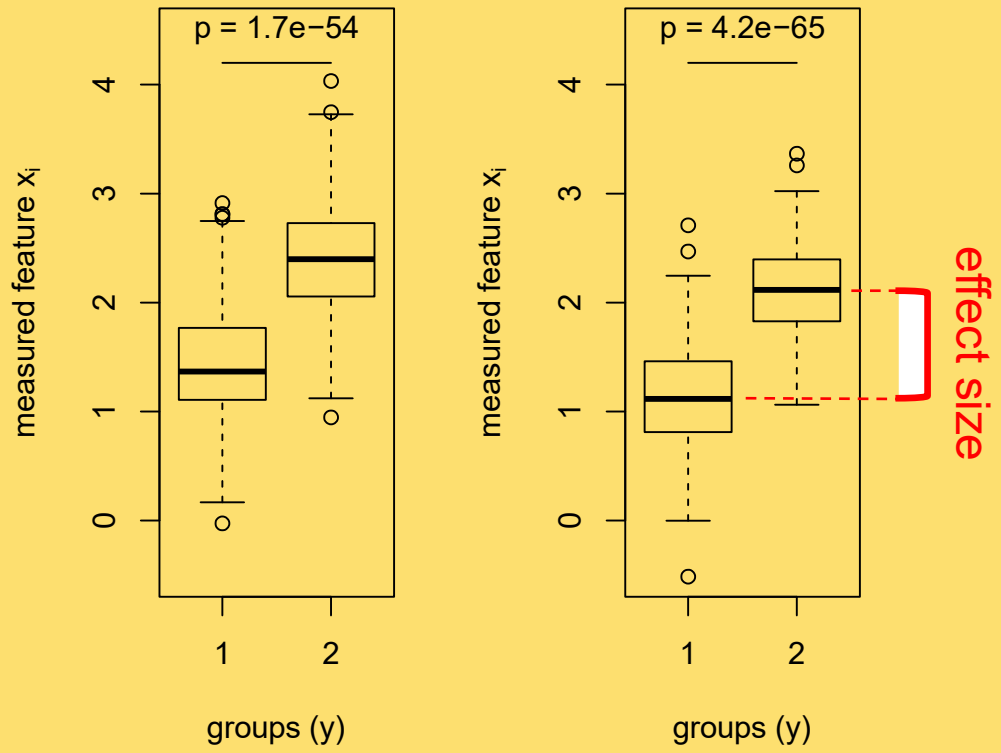


Principal component or coordinate analysis (PCA or PCoA)



Determine differentially abundant features

Hypothesis testing: could an observed difference also be observed by chance?



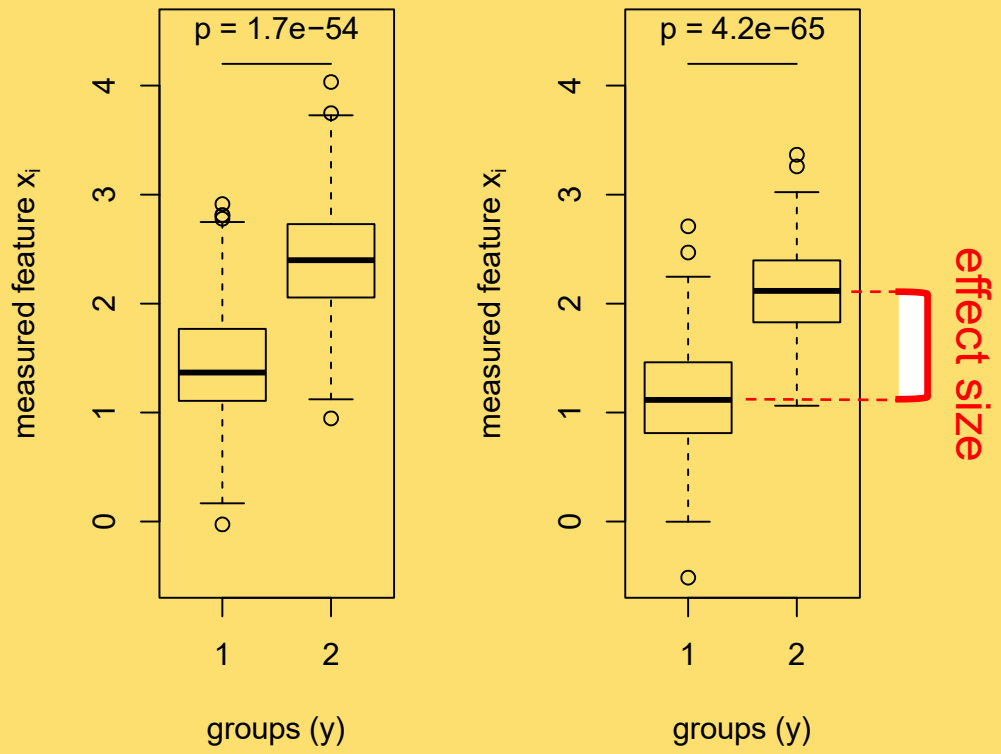
Question 1: in a clinical trial, you observe differences in the taxonomic composition of stool samples from healthy vs. diseased individuals. Assuming it to be a true effect, what do you expect from sampling additional individuals?

a) The fold change (effect size) of differentially abundant taxa to become larger
b) The p-value associated with these changes to decrease
c) The confidence interval around the fold change to increase

Sullivan and Feinn, J Grad Med Educ., 2012

Determine differentially abundant features

Hypothesis testing: could an observed difference also be observed by chance?



Question 2: the likelihood of observing significantly different features between samples by chance increases with the number of features for which a test is performed. What measures can be taken to correct for errors introduced by such multiple comparisons?

a) Correct the p-value according to the number of tests performed

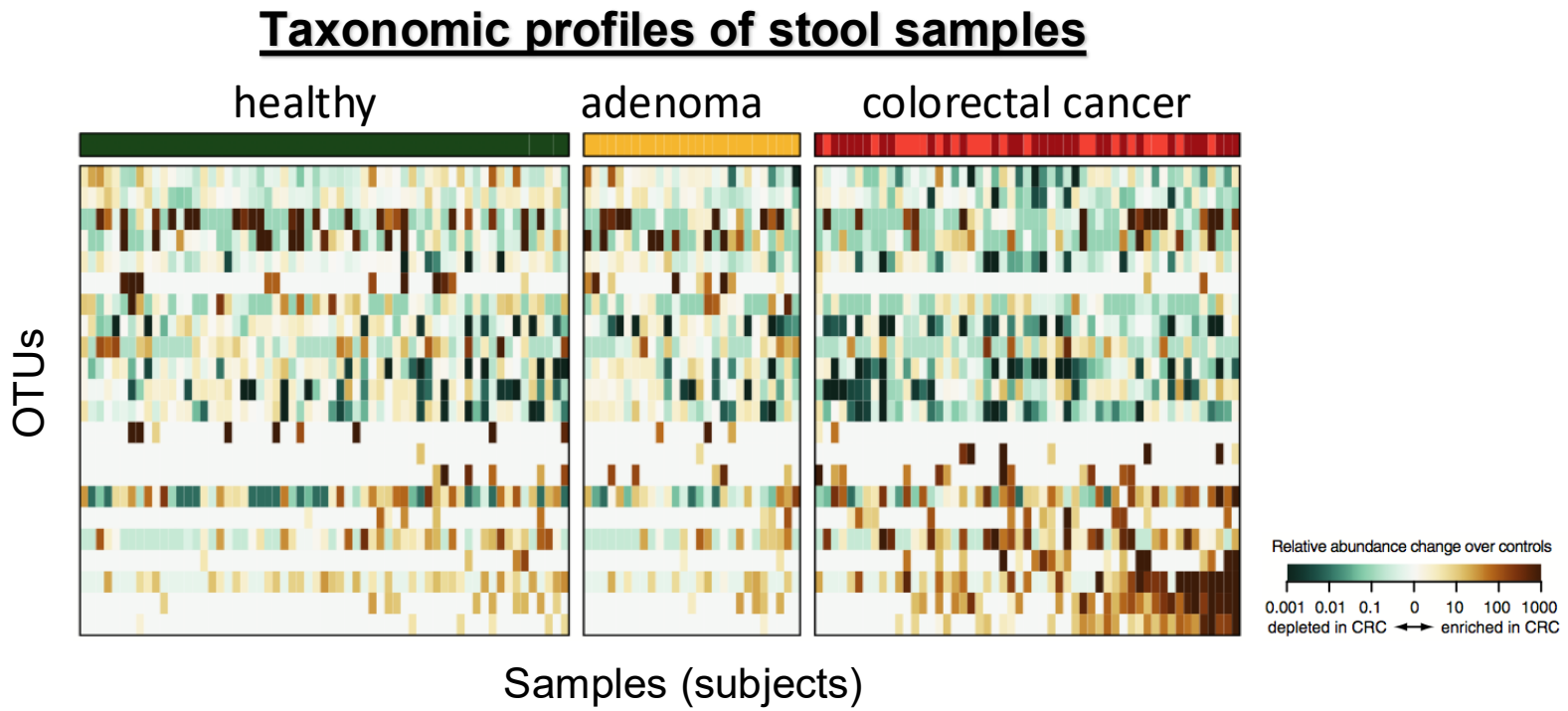
b) Repeat the test multiple times to reduce the error

c) Reduce the number of features that are tested

→ label-agnostic modifications to matrix

Quantitative differences between microbial communities

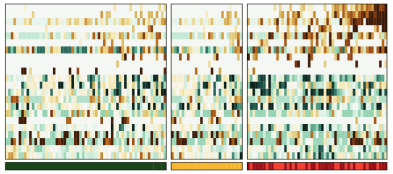
- Describing quantitative differences between microbial community compositions:
  - can identify taxa (or other features) as disease markers

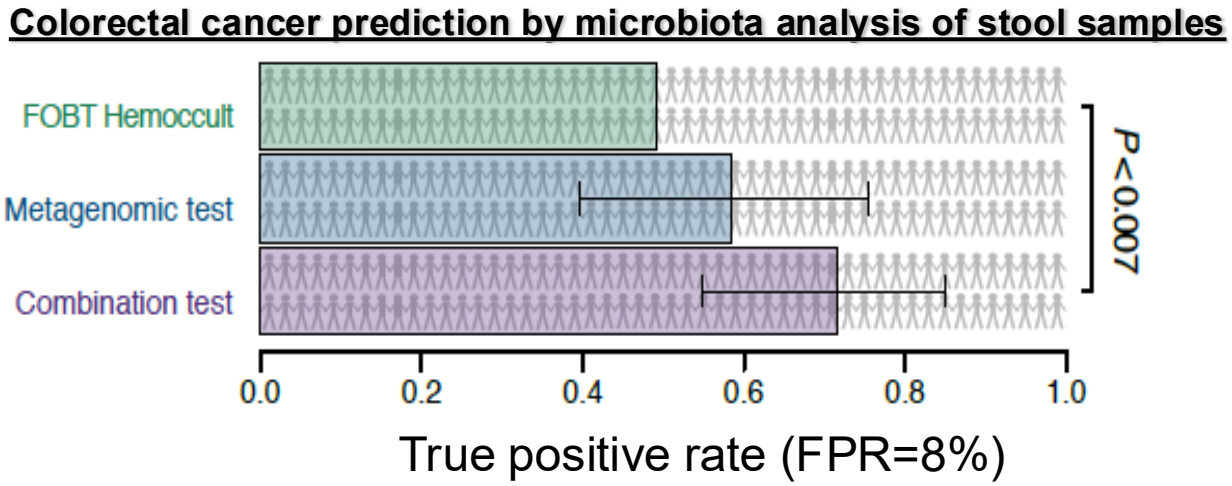


Zeller et al., MSB, 2014

Quantitative differences between microbial communities

- Describing quantitative differences between microbial community compositions:
  - and be used to make predictions (classify) new samples, e.g. by machine learning





Summary – Part II

- Dissimilarities of microbial community compositions (beta diversity) can be quantified by different diversity indices
- Microbiome wide association studies aim at identifying relationships between microbiome features (taxa, genes, functions) and phenotypes
- Statistical testing can reveal differentially abundant features (potential biomarkers) between groups of samples

Overview of the Metagenomics part

Microbial community structure

- microbial taxonomy and operational taxonomic units
- quantification of microbial community members
- diversity within a microbial community

Differences between microbial communities

- taxonomic differences between microbial communities
- differentially abundant features (e.g., taxa, genes, functions)

Working with microbial community genes and genomes

- reconstruction of microbial community genomes
- gene functional differences between microbial communities

Reconstruction and annotation of microbial community genomes

- Most organisms in microbial communities have not been isolated and cultured
  - However, we can sequence microbial community DNA, reconstruct genomes and predict protein sequences / structures
- Genomes reconstructed from natural environments capture microbial diversity on Earth
  - New data challenge long-standing concepts
- Predicted genes inform about functional capabilities and traits of organisms
  - “Who is there?” → “What can they do?” → “Who can do what?”
- Genomic information enables discovery of new enzymes and microbially produced compounds
  - Potential to identify new drug leads or proteins with desired or new functions
- Microbial gene functions may explain differential responses to same treatment
  - Analysis of microbiomes may inform personalized treatments

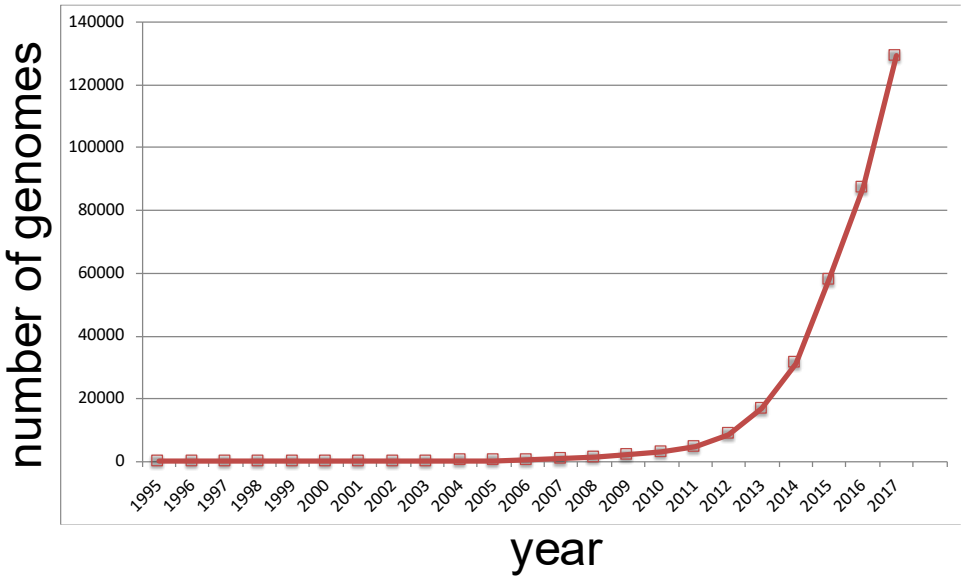
Sequencing of microbial isolate genomes

- First bacterial genome (1995): *Haemophilus influenzae* *Fleischmann et al. 1995*
- Followed by many isolated pathogens of diseases (e.g., plague, anthrax, tuberculosis, Lyme disease)
- Many isolates of important non-pathogenic species: e.g., *Prochlorococcus*, *Lactobacillus*, *Bradyrhizobium*

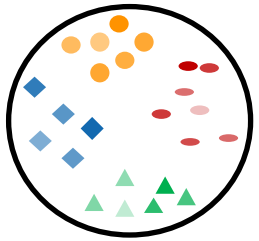
- Bacteria and archaea have ca. 500–10,000 genes, arrayed on usually circular DNA molecules (e.g., chromosomes and plasmids)
- Protein coding genes are on average ca. 1,000 base pairs long
- Their genomes are ca. 600,000–12 million bp in size (human 2 x 3 billion bp)

Background: added value of metagenomics

Microbial isolate genome sequences



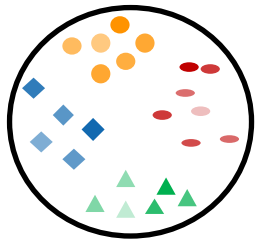
9/2023	10/2025
312,000	3,150,000



Microbial community



→However, most bacteria and archaea have not been isolated and sequenced.
Metagenomics provides access to genomic information within a microbial community. This allows us to ask: “what can they do?”, “who can do what?”



DNA extraction

Library preparation

Microbial community

Overview of generic* metagenomics workflow

*This is generic; specific workflows can vary on the order of steps here and how they are done.

When working with your own data you should never follow any pipeline blindly. There can be critical differences based on your data.

might be done by sequencing facility

sequencing facility → fastq files → demultiplex (split samples by barodes) → fastqc/multiqc → quality filter/trim (remove adapters/**primers**) → fasta/q files

fastq files: @HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G.<G<AGGII...

Some tools:
- sabre
- fastx_demux (usearch/vsearch)
- idemp
- fastx barcode splitter (fastx-toolkit)

Some tools:
- trimmomatic
- bbduk.sh (bbtools suite of tools)

fasta/q files: >HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...

read-based analysis ← no-assembly path

assembly path

consider testing assemblies with and w/o digital normalization

Some tools:
- bbnorm
- diginorm

Count Table

	Sample_A	Sample_B	...
obj_1	0	428	...
obj_2	306	323	...
obj_3	217	1	...
...	...	...	...

MetaQUAST is a great tool for comparing assemblies

contigs

map individual sample reads to (co)-assembly → Generate coverage information (mapping)

Some tools:
- bowtie2
- bwa

Some assemblers and tools:
- Megahit (assembler)
- SPAdes (assembler)
- idba-ud (assembler)
- MetAMOS (assembler and analysis pipeline)
- MetaCompass (reference-guided)
- MetagenomeScope (visualize assembly graphs)

Gene calling Functional/taxonomic profiling

Some tools:
- prodigal (identifies open reading frames)
- prokka (runs prodigal and performs annotations)
- GHOSTKOALA (web-hosted KEGG annotations)
- BLAST (protein nr db/refseq/COGs)

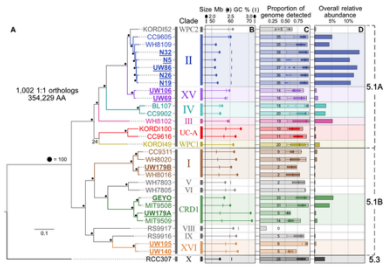
Recovering genomes from metagenomes → **A note on MAGs:** MAGs (metagenome-assembled genomes) are not the same thing as isolate genomes. They are composite representative genomes of closely related genomic lineages.

Some tools:
- anvi'o (interactive manual curation of bins; and much more)
- CONCOCT (kmer-based and coverage-based binning; also incorporated in anvi'o)
- COCACOLA (kmer-based, coverage-based, and incorporates paired-read linkage of contigs)
- MetaBAT2 (kmer-based and coverage-based binning tool)
- BinSanity (primarily coverage-based, optional second round kmer-based binning tool)
- checkm (genome-level taxonomy; and much more)
- DASTool (a tool for evaluating bins recovered by different methods)
- DESMAN (tool aimed at resolving strains)

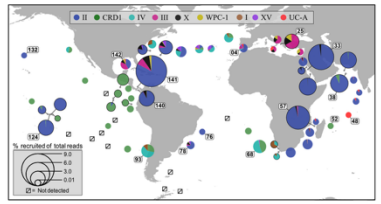
Some common genomics stuff

Phylogenomics
Comparative genomics
Pangenomics
Env. distributions

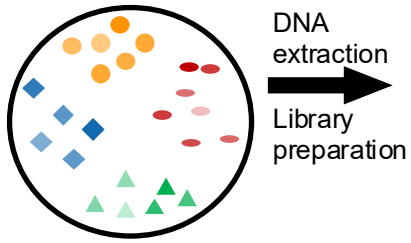
Some tools:
- anvi'o (integrated HMMs for common single-copy gene sets; integrated pangenomic workflow for identifying orthologs via OrthoMCL)
- PanOCT (identifies orthologs utilizing synteny information)
- StrainPhlAn/PanPhlAn (tools for strain-level analyses)
- MUSCLE (alignment software)
- FastTree (very fast, pseudo-maximum likelihood tree builder)
- RAxML (maximum likelihood tree builder)
- Mauve (whole-genome alignment)







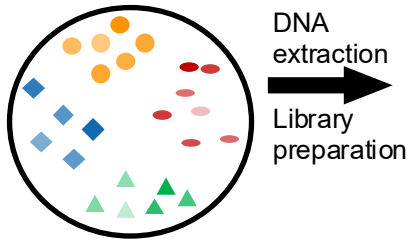
astrobiomike.github.io



Microbial community

DNA extraction

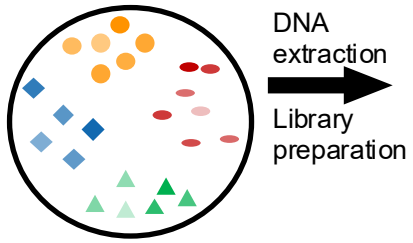
- Sufficiently high quality and <u>quantity</u> needed
  - >1ng (Illumina)
  - >50-400ng (Oxford Nanopore Technologies – ONT / PacBio)
    - Now: ca. 1ng



Microbial community

DNA extraction / library preparation

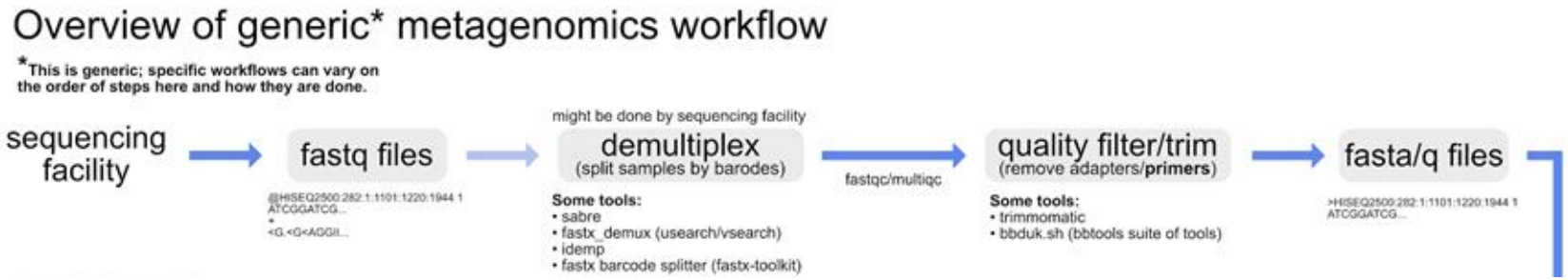
- Sufficiently high quality and quantity needed
- Extracted DNA is sheared into smaller fragments (inserts)
  - Illumina: ~300-600 bp; PacBio: ~20 kbp; ONT: Mbp fragments possible



DNA extraction

Library preparation

Microbial community



Storage of DNA sequence information

- National Center for Biotechnology Information (NCBI)
- European Nucleotide Archive (ENA)
- DNA Data Bank of Japan (DDBJ)

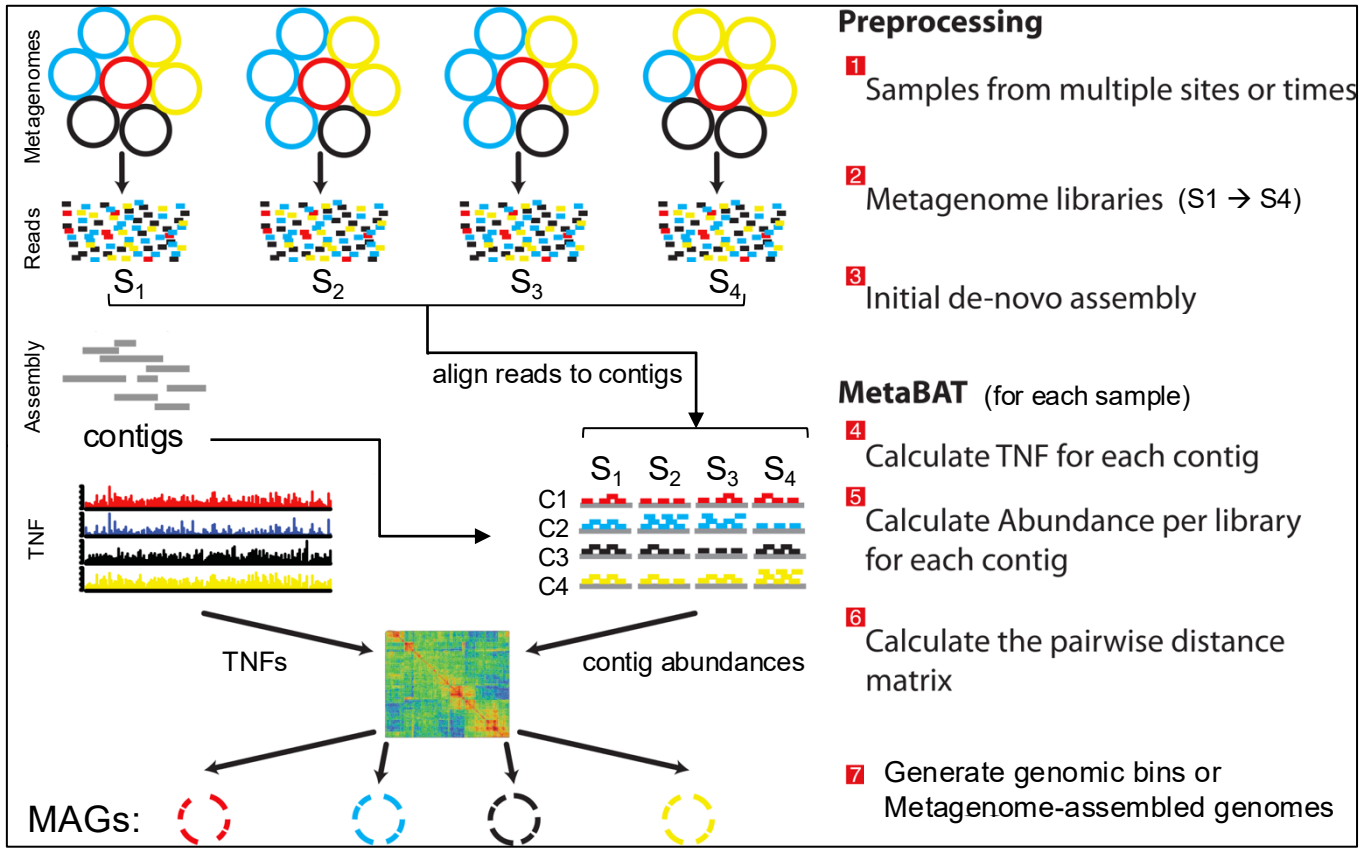
INSDC

International Nucleotide Sequence Database Collaboration

Databases

Data type	DDBJ	ENA European Nucleotide Archive	NCBI
Next Generation reads	Sequence Read Archive		Sequence Read Archive
Assembled Sequences	DDBJ	European Nucleotide Archive	GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

Binning contigs into metagenome-assembled genomes



From contigs to metagenome-assembled genomes (MAGs)

Distance matrices between contigs of the same sample based on (next slides):

a) Tetranucleotide frequencies (TNFs)
b) Abundances of contigs within and across samples

Identify clusters of highly correlated contigs:
→ metagenome-assembled genomes (MAGs)

After Kang et al., PeerJ, 2015

Genomic repeats – classroom activity I

Consider the sentence:

`"a cat is bigger than a fox is bigger than a pig is bigger than a yak is bigger than"`

Assume this was a circular molecule with the sequence:

`acatisbiggerthanafoxisbiggerthanapigisbiggerthanayakisbiggerthan`

and available sequencing technology allowed to sequence pieces (reads) with a maximum length (k) of 4 (e.g., `acat`, `cati`, etc.).

You can reconstruct the original sequence by overlapping 3-mers (e.g., `acat` and `cati` overlap by the 3-mer `cat`, so they can be merged into `acati`.

The number of k-mers is proportional to the copy number at which they occurr in the genome.

Questions:

- To what extent would you be able to reconstruct the original sequence?
- What is the unresolvable repeat sequence?
- What is the minimum read length required to resolve the sequence completely?

Genomic repeats – classroom activity I

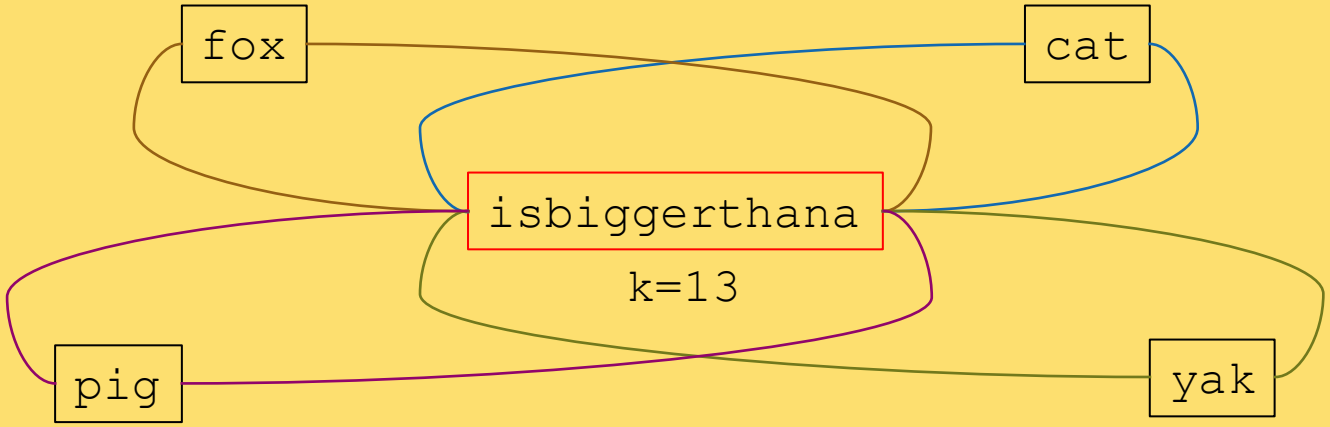
Consider the sentence:

"a cat is bigger than a fox is bigger than a pig is bigger than a yak is bigger than"

Assume this was a circular molecule with the sequence:

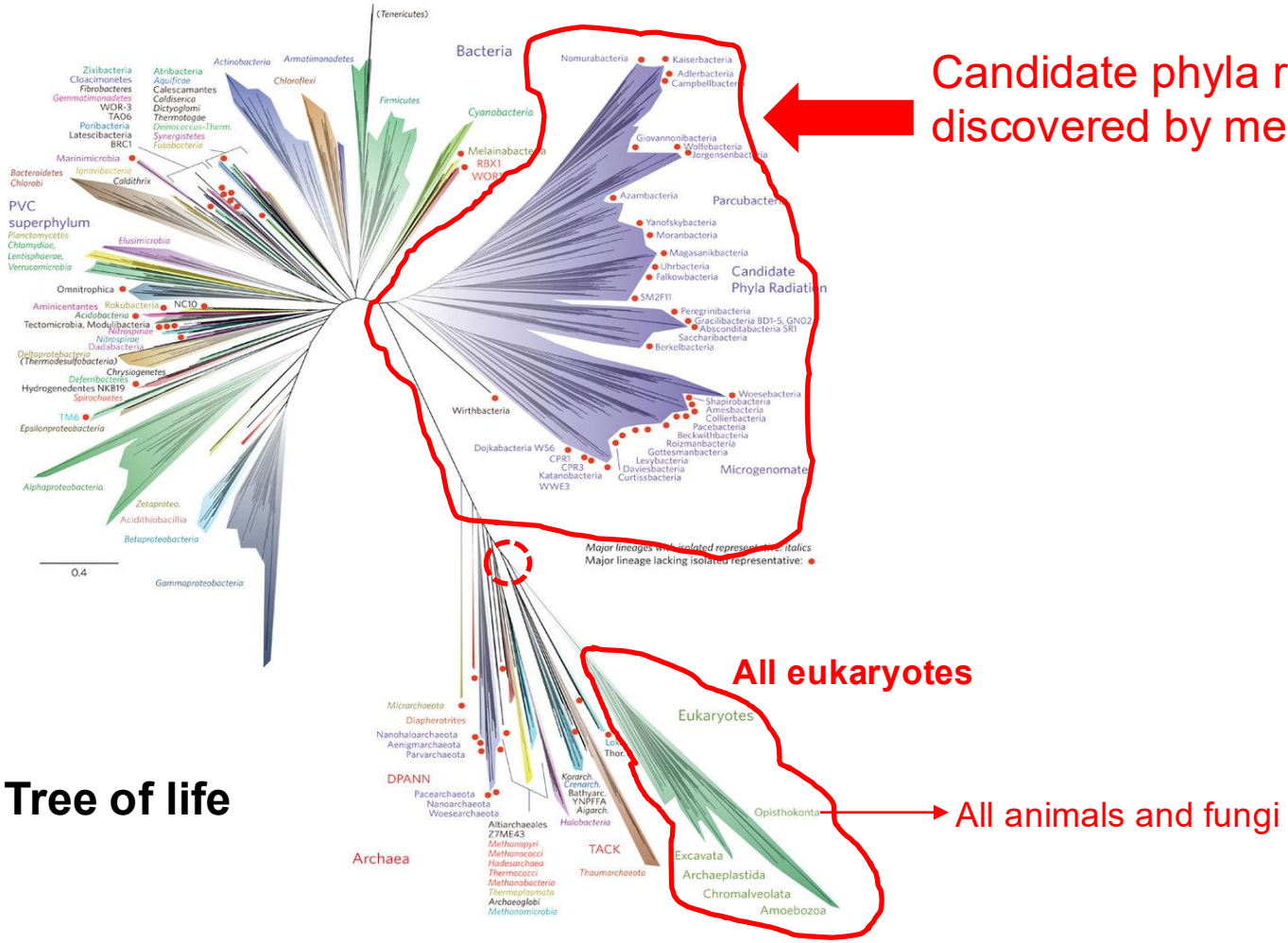
acatisbiggerthanafoxisbiggerthanapigisbiggerthanayakisbiggerthan

k=15



→ "isbiggerthana" is the repeat sequence and 15-mers are needed to determine the correct order of fragments

Insights by reconstructing microbial community genomes



Candidate phyla radiation discovered by metagenomics

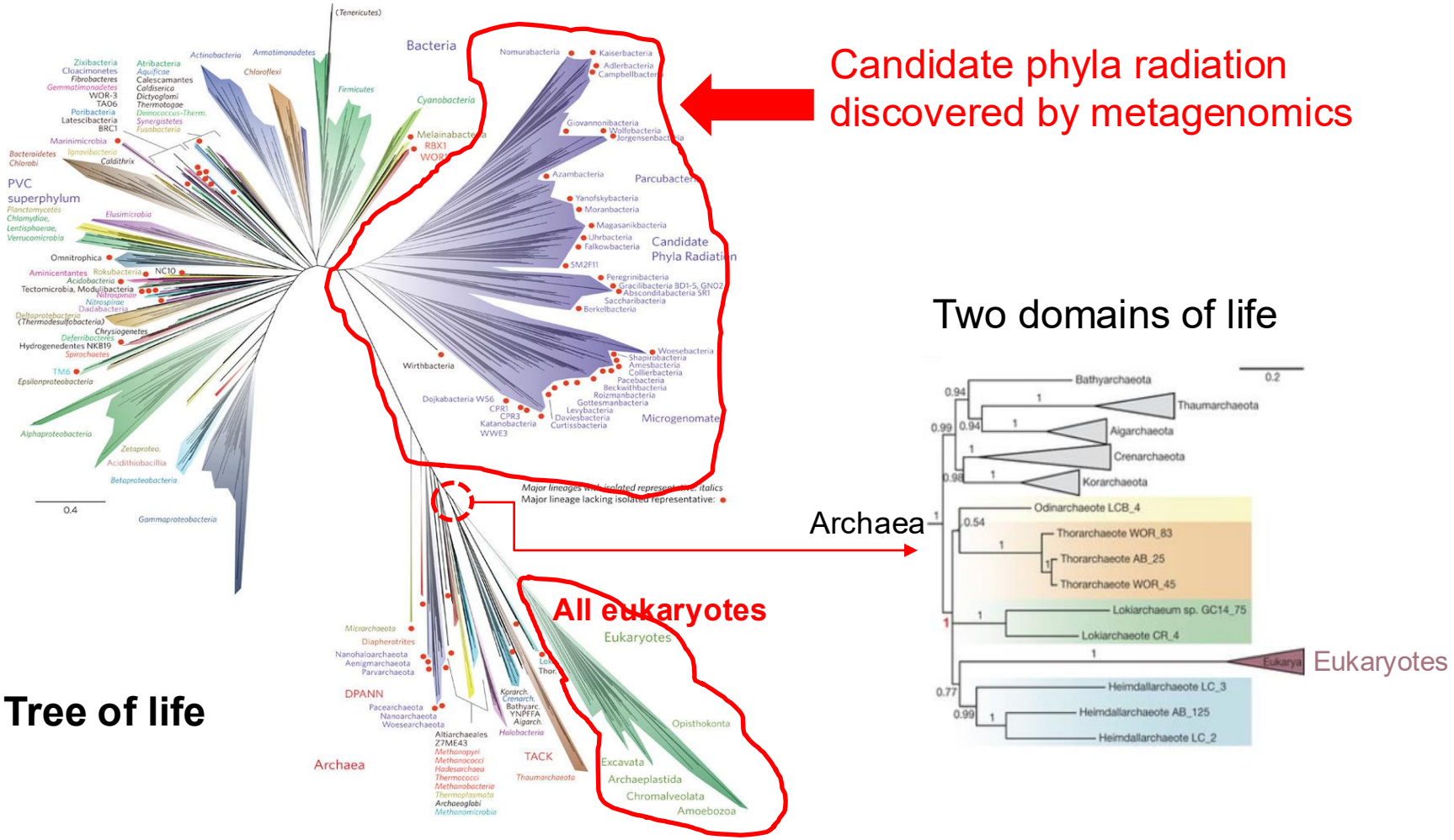
All eukaryotes

All animals and fungi

Tree of life

Hug et al., Nat. Microbiol., 2016; Zaremba-Niedzwiedzka et al., Nature, 2017; Martijn et al., Nature, 2018

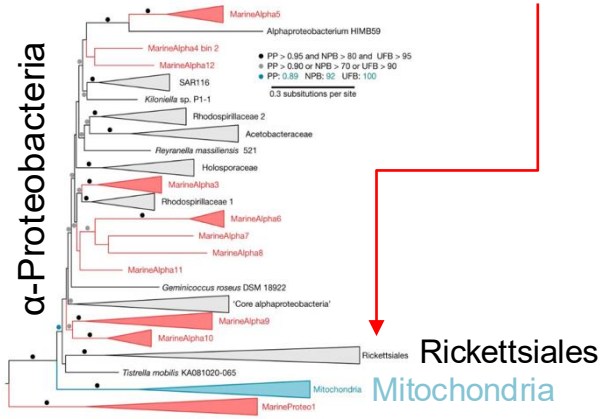
Insights by reconstructing microbial community genomes



Candidate phyla radiation discovered by metagenomics

Two domains of life

Mitochondrial origin not within Rickettsiales

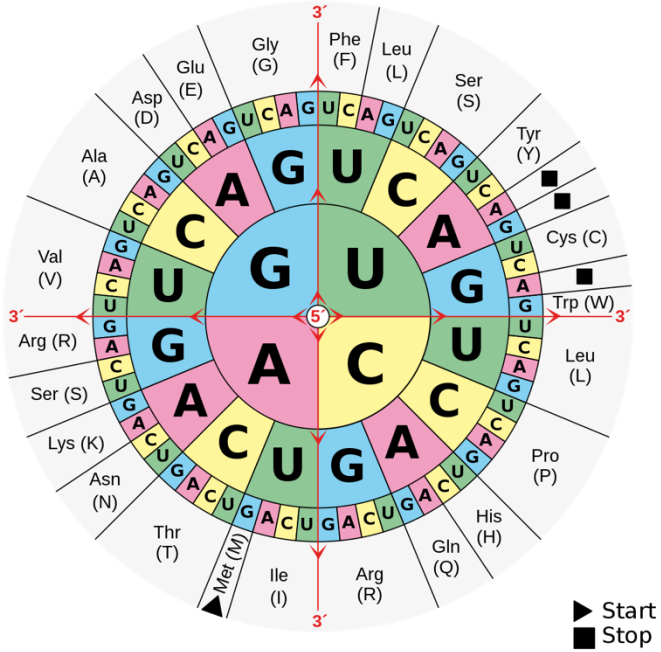


Genome annotation – protein coding genes

Gene prediction

- Identify protein-coding (and non-coding) sequences in a (meta)genome

- *Ab initio* - using only the genomic DNA sequence
  - most simple approach: find (large) open reading frames (ORFs)
  - search for signals (specific sequences) of protein coding regions



Genome annotation – protein coding genes

Example – finding Open Reading Frames (ORFs)

- Sequence has 6 possible translations from nucleotide to amino acid sequence

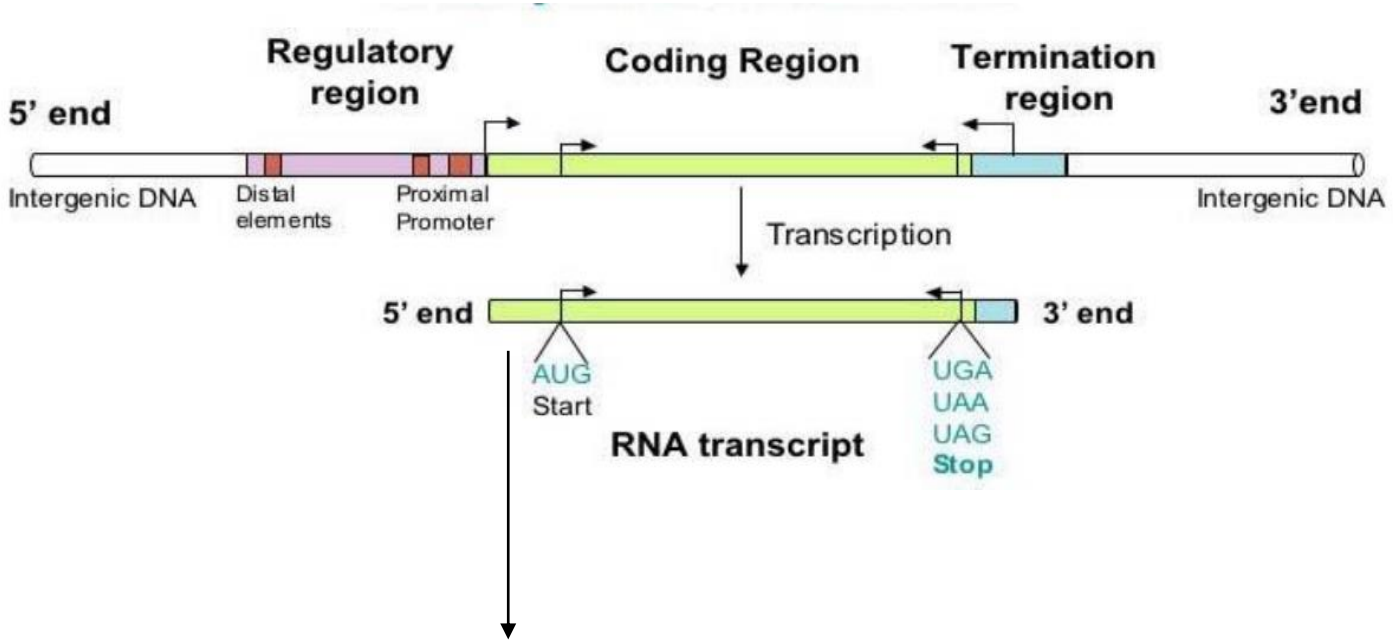
…AGC TTT TCA TTC TGA CTG CAA CGG GCA ATA TGT CTC TGT GTG GAT TAA AAA AAG AGT GTC TGA TAG CAG C…
…A GCT TTT CAT TCT GAC TGC AAC GGG CAA TAT GTC TCT GTG TGG ATT AAA AAA AGA GTG TCT GAT AGC AGC…
…AG CTT TTC ATT CTG ACT GCA ACG GGC AAT ATG TCT CTG TGT GGA TTA AAA AAA GAG TGT CTG ATA GCA GC…

…G CTG CTA TCA GAC ACT CTT TTT TTA ATC CAC ACA GAG ACA TAT TGC CCG TTG CAG TCA GAA TGA AAA GCT…
…GCT GCT ATC AGA CAC TCT TTT TTT AAT CCA CAC AGA GAC ATA TTG CCC GTT GCA GTC AGA ATG AAA AGC T…
…GC TGC TAT CAG ACA CTC TTT TTT TAA TCC ACA CAG AGA CAT ATT GCC CGT TGC AGT CAG AAT GAA AAG CT…

- An ORF is a sufficiently large region between a start and a stop codon

Genome annotation – protein coding genes

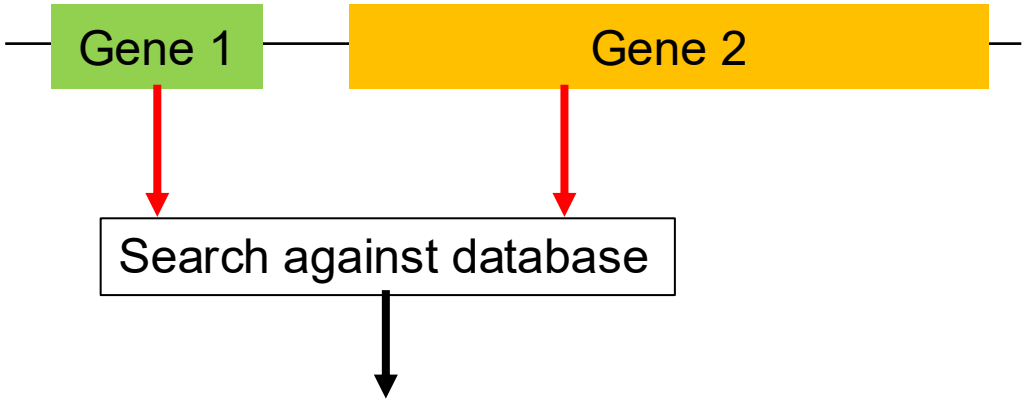
Prokaryotic gene structure



The 5'-UTR (untranslated region):

- from transcription start site to -1 bp of start codon
- contains **ribosome binding site** (RBS)

Functional annotation of genes



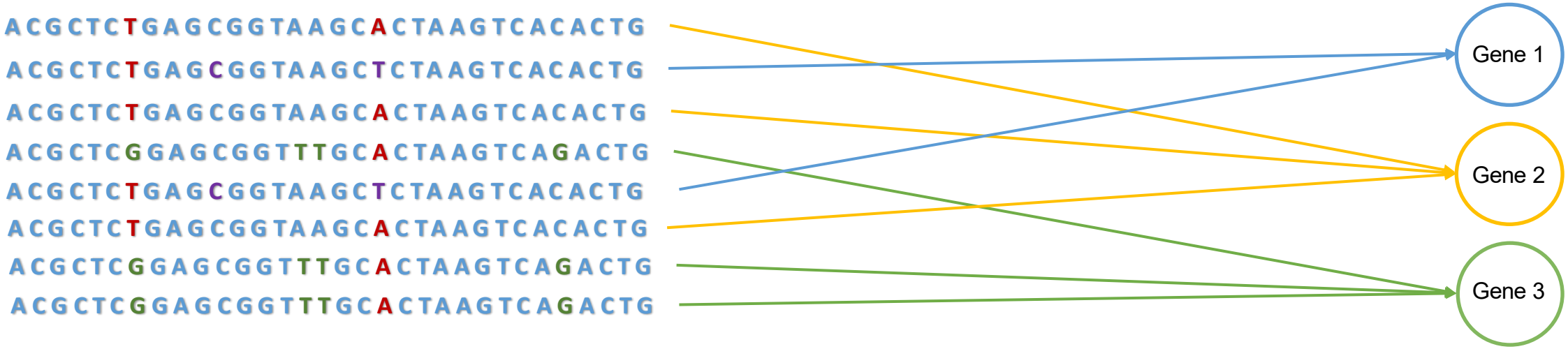
	DB	Identifier
Database 1:	KEGG	K00847
Database 2:	COG	COG1940
Database 3:	Pfam	PF00480

Goal: assign to each gene its function

- Gene sequences can be searched against different databases that store information on known gene functions
- Widely used databases:
  - Kyoto Encyclopedia of Genes and Genomes (KEGG)
  - Cluster of Orthologous Groups (COG)
  - Protein Family domains (Pfam)
  - Comprehensive Antibiotic Resistance Database (CARD)
  - Many more…

Quantification of gene abundances

All **metagenomic reads** are aligned to best matching gene



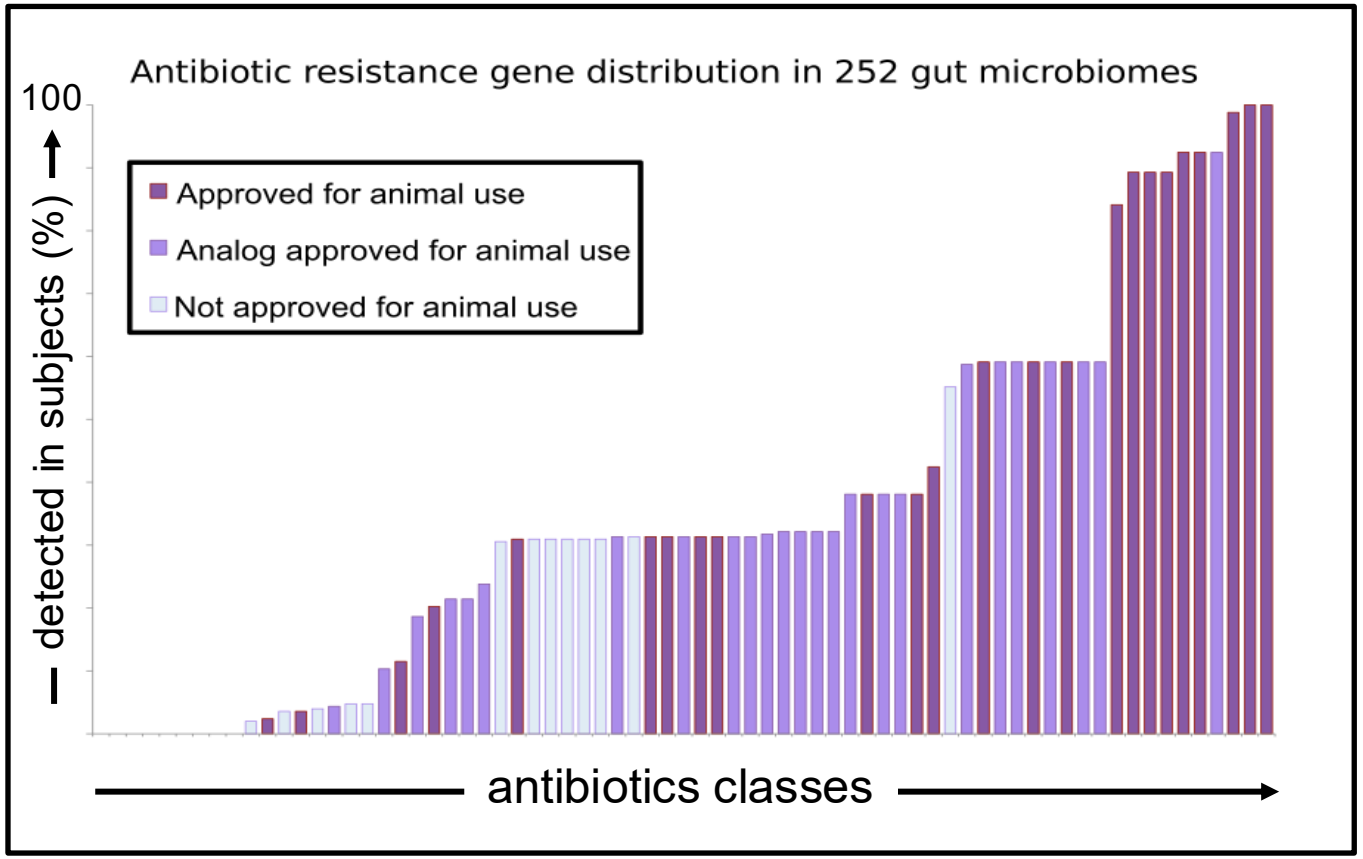
The result is a gene count table, summarizing read counts for each gene for each sample

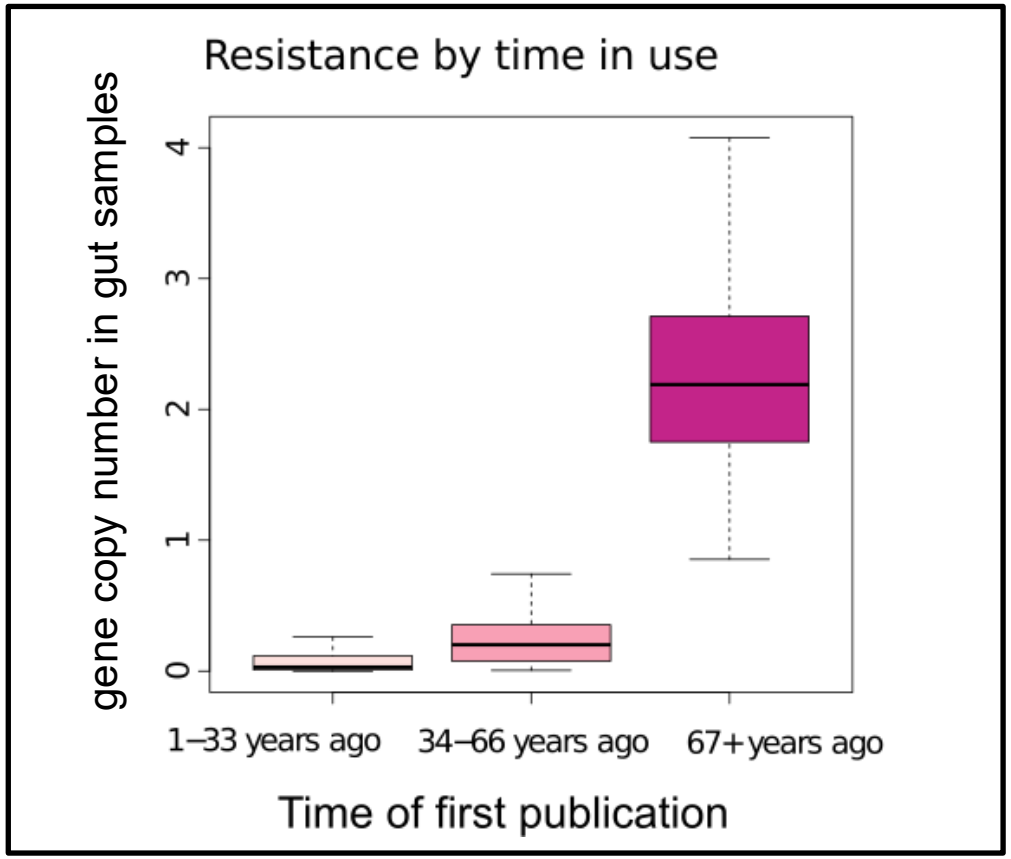
Gene count tables can be summarized into KO abundance tables

KO abundance tables can be summarized into Module abundance tables

Module abundance tables can be summarized into Pathway abundance tables

Insights by quantifying microbial gene abundances





Summary – Part III

- Metagenomic sequencing and genome reconstruction provides access to studying microbes in their natural environment where they live in complex communities
- Taxonomic annotation of a reconstructed genome provides information about its 'novelty'
- Prediction of genes and their annotation using different databases provides information about the functional capabilities of microorganisms
- Genes can be grouped into higher functional levels and profiled to study gene functional differences between microbial communities