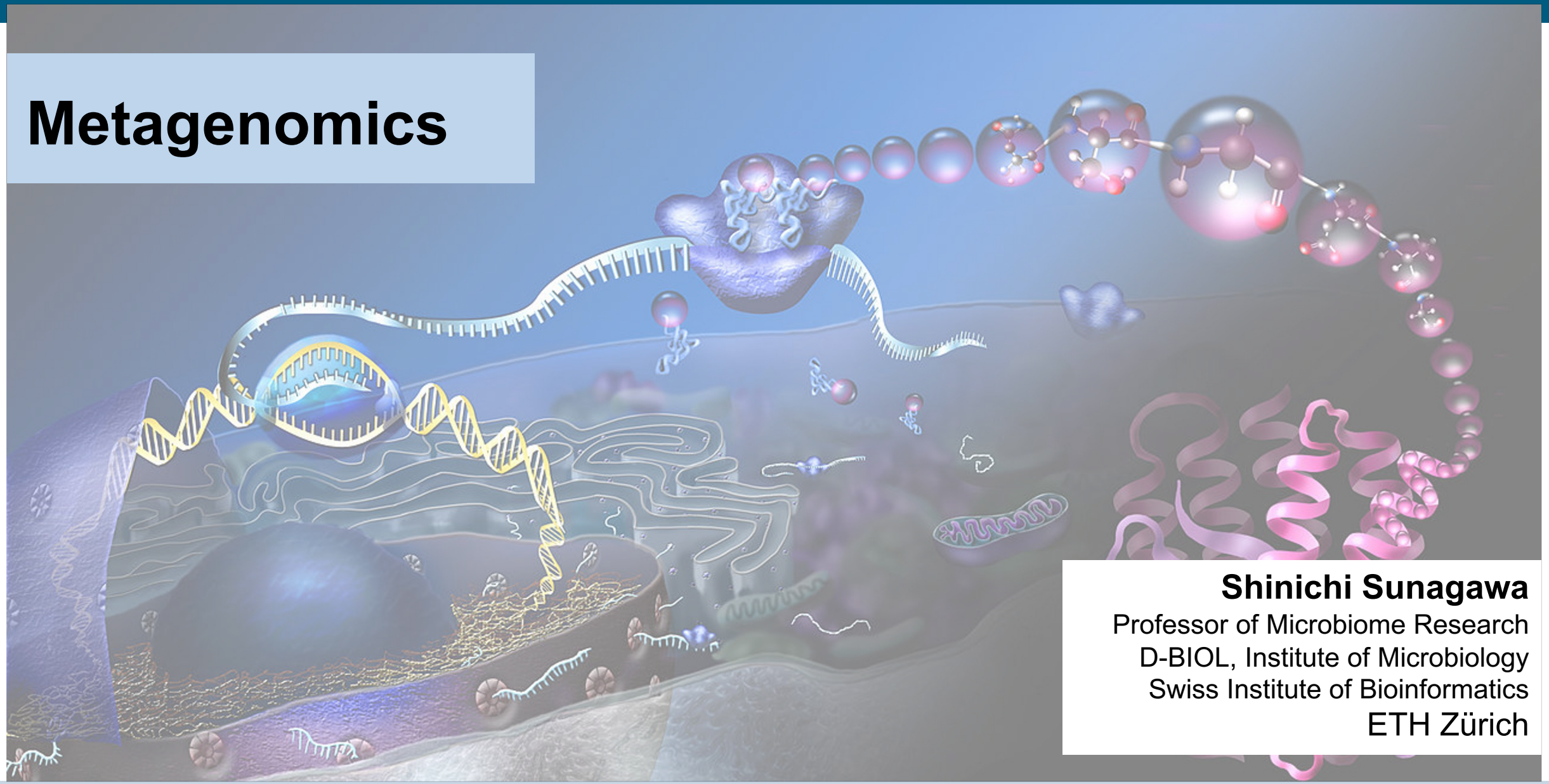
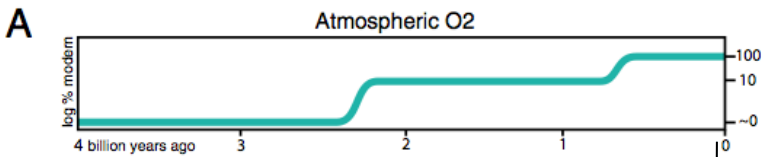
Metagenomics



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



Microorganisms

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

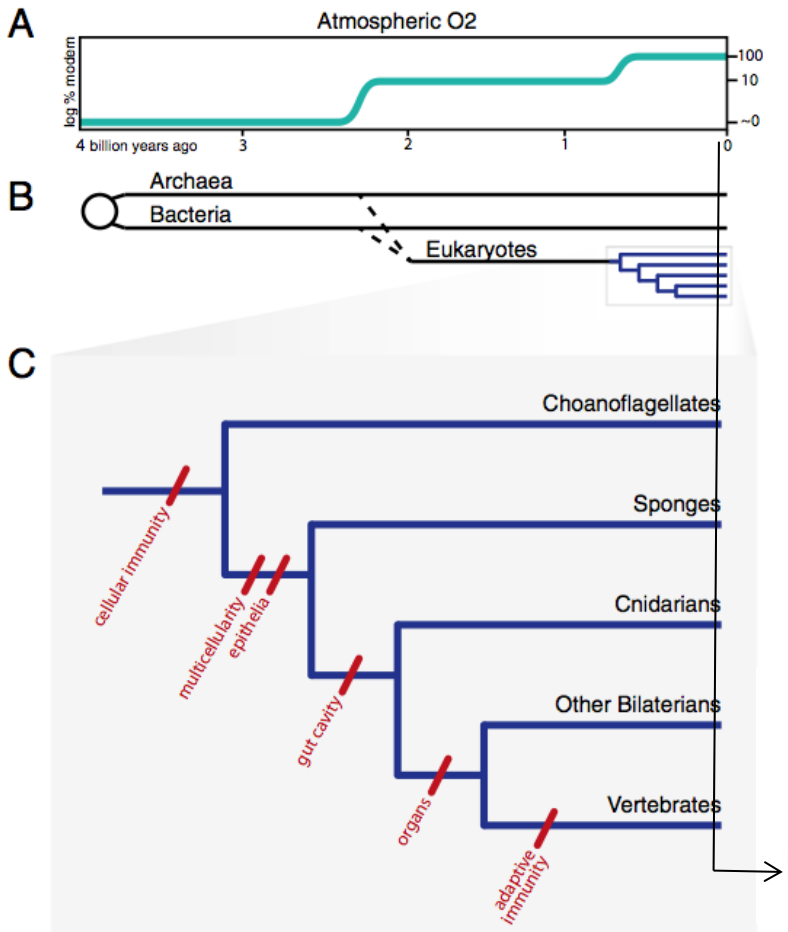
Significance (examples):

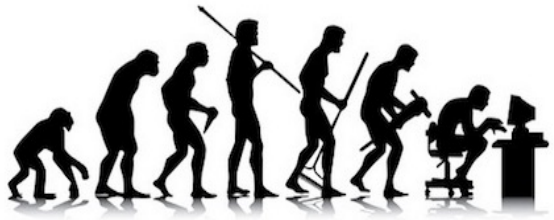
- biogeochemistry: e.g., photosynthesis by microbes, carbon fixation/export, nitrogen fixation
- health: help us digest food, provide essential vitamins, prime the immune system

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

Evolution and significance of microbiomes

From the origin of life to today





13 mio years

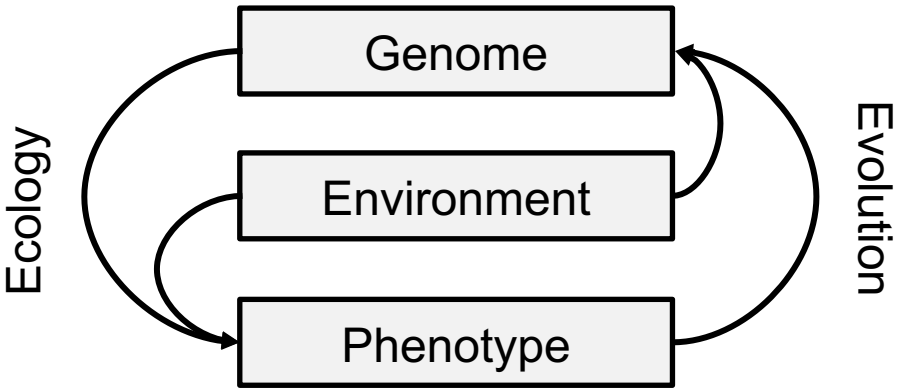
Microorganisms

- 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

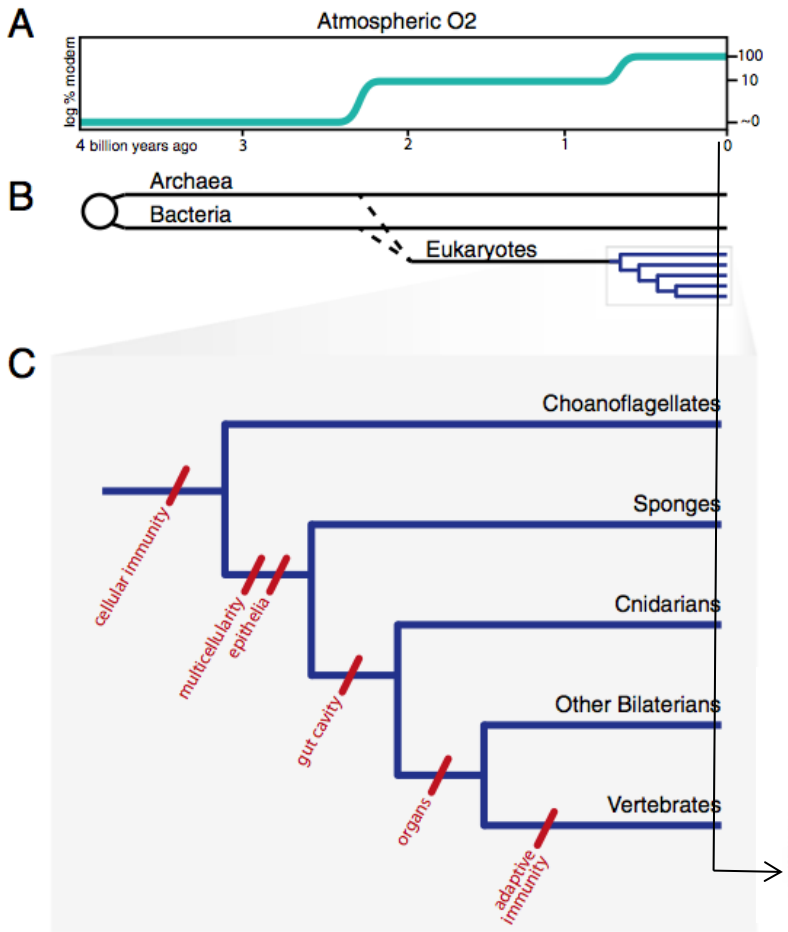
Single organism-centric view

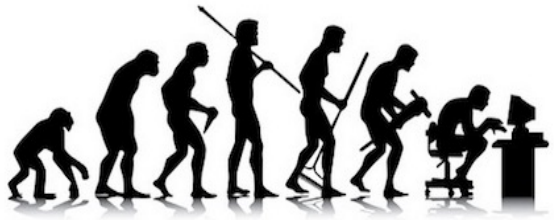


McFall-Ngai et al., PNAS, 2013; Venn et al., Science, 2014

Evolution and significance of microbiomes

From the origin of life to today





13 mio years

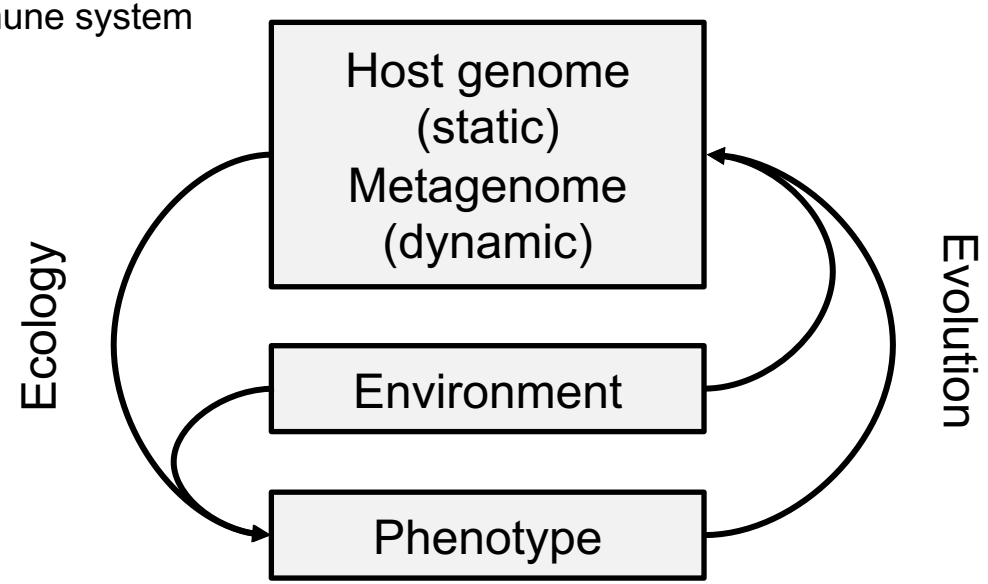
Microorganisms

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

Significance (examples):

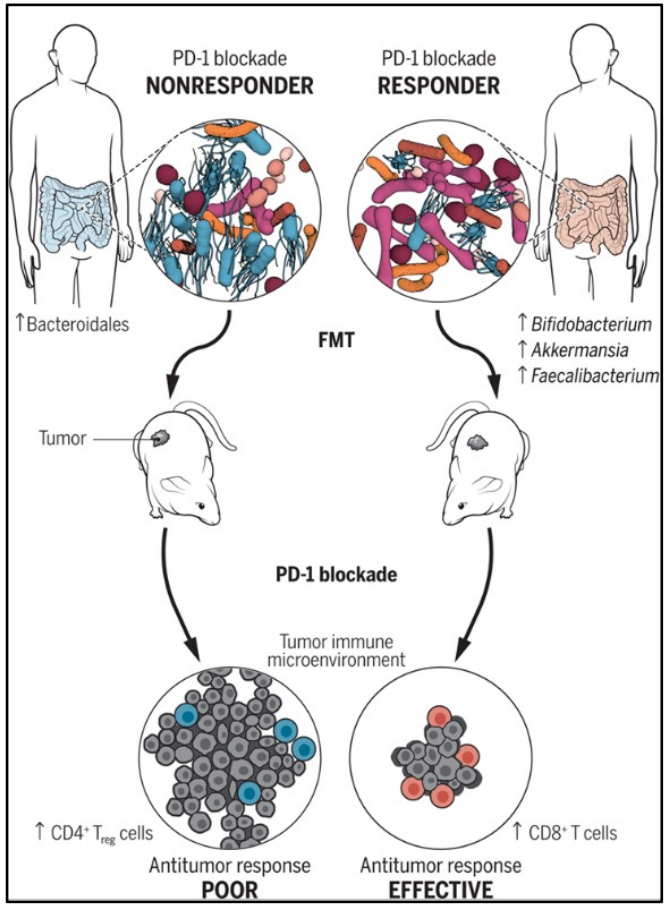
- biogeochemistry: e.g., photosynthesis by microbes, carbon fixation/export, nitrogen fixation
- health: help us digest food, provide essential vitamins, prime the immune system

Holobiont view



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

Describing microbial communities – Example 1



GRAPHIC: V. ALTOUNIAN/*SCIENCE*

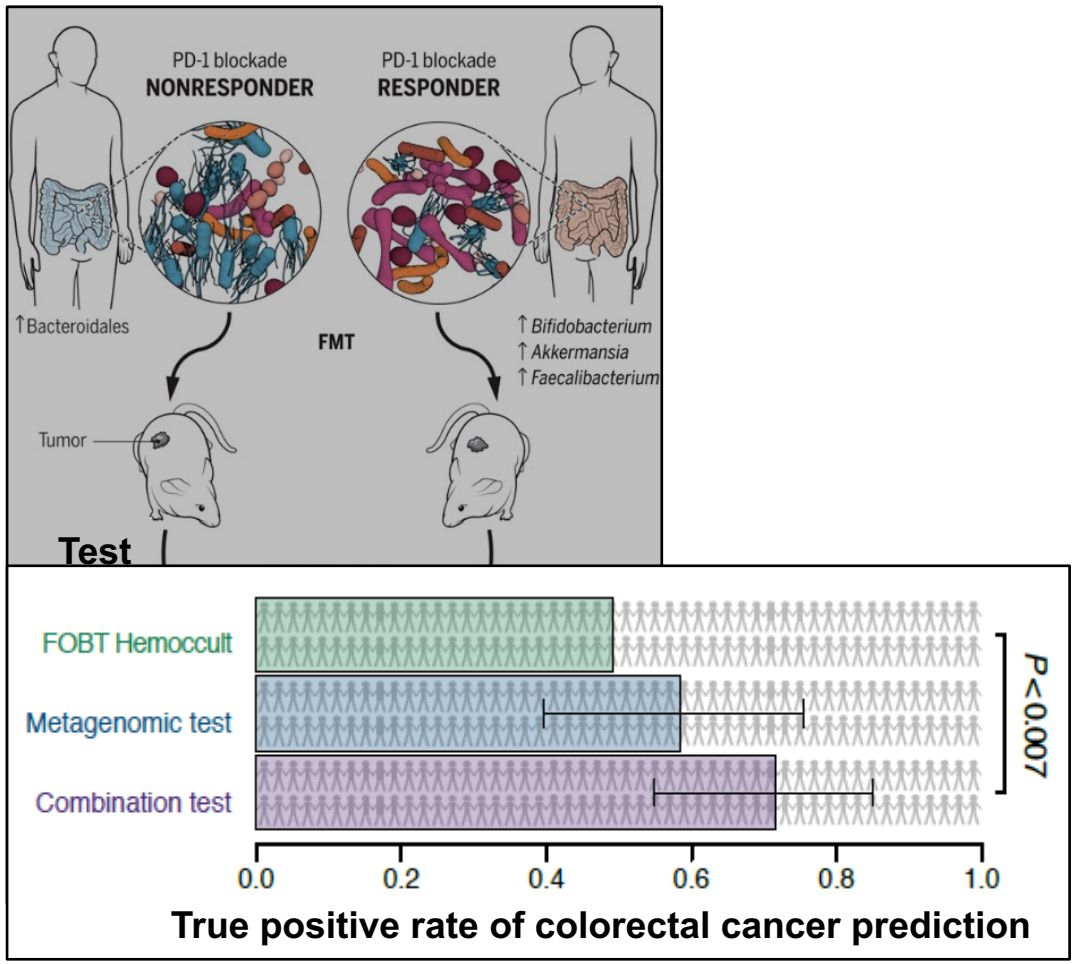
Gut microbial community compositions

- 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. Science 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

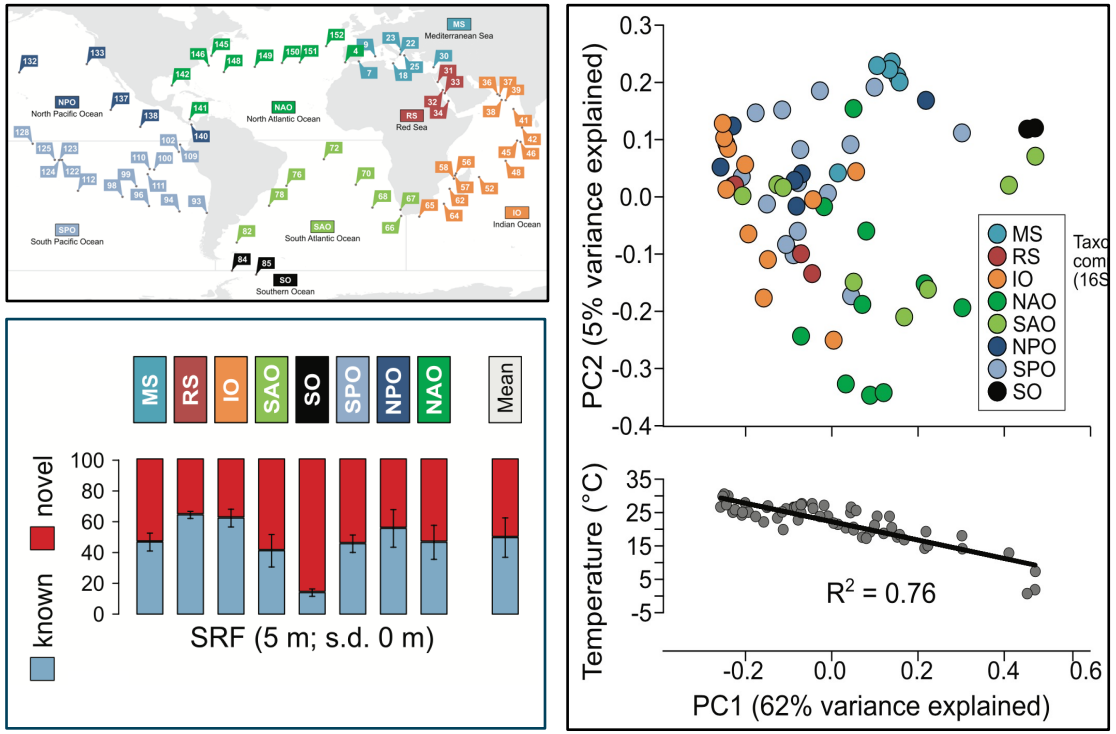
Routy et al., Gopalakrishnan et al., and Matson et al., Science 2018

- can be indicative for diseases

→ Statistical models of fecal microbiota composition can predict colorectal cancer

Zeller et al., MSB, 2014; Wirbel et al., Nat Med, 2019

Describing microbial communities – Example 2



Ocean microbial community compositions

- reveal previously unknown organisms and genes (left bottom)

→ implying novel taxa, enzymes and functions

Paoli et al., Nature, 2022

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

→ but strongly driven by temperature (bottom right)

Sunagawa et al., Science, 2015

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

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

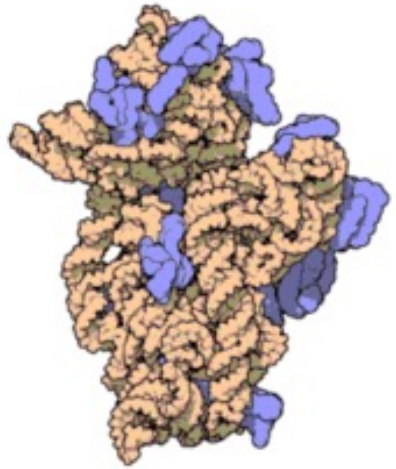
- Phenotypic characteristics
  - morphology, physiology/metabolism, ecology, exchange of genetic material
- Molecular characteristics
  - DNA-DNA hybridization
  - **DNA sequences of individual genes (e.g., 16S rRNA gene) or complete genomes**

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

The 16S rRNA gene

- 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

30S small subunit of ribosomes in prokaryotes

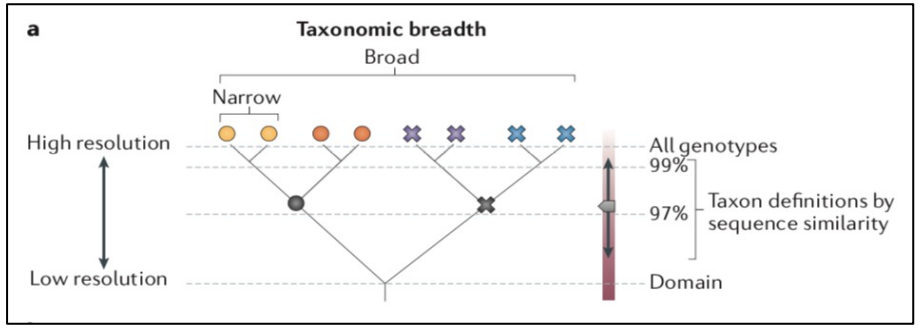


Hanson et al., Nat Rev Microbiol, 2012

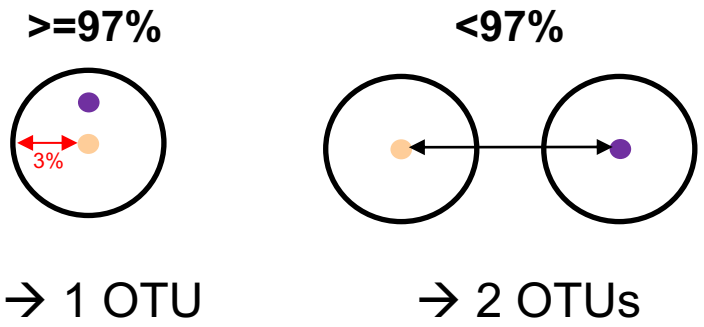
16S rRNA-based Operational Taxonomic Units (OTUs)

- 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

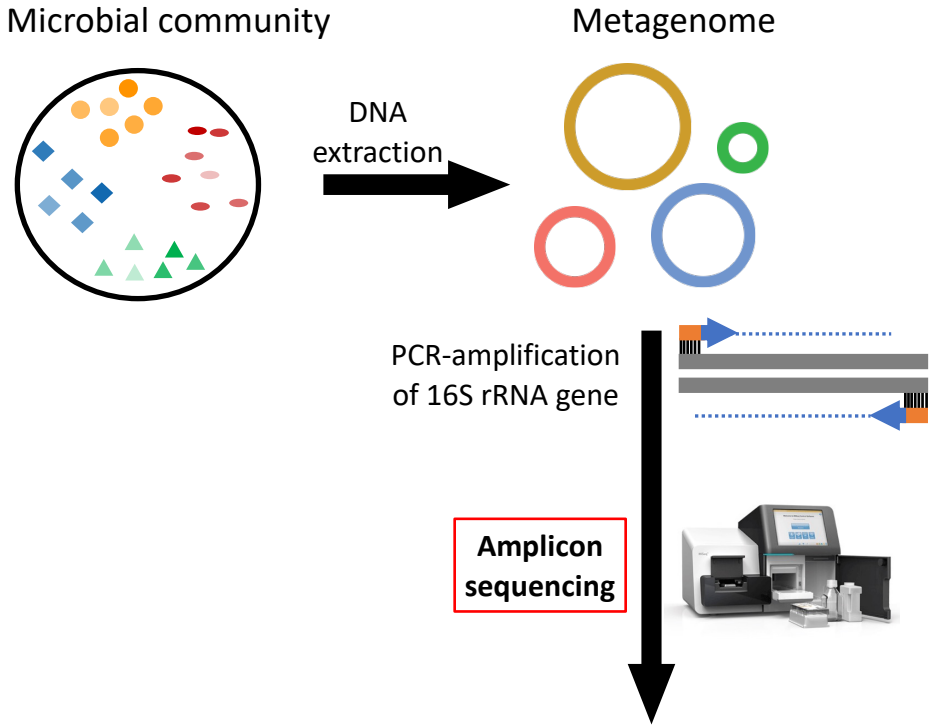
- Owing to lack of prokaryotic species definition, 97% sequence similarity is often used to define 'species'-like:
  "**Operational Taxonomic Units" (OTUs)**



Identity of 16S rRNA gene sequences



Amplification of 16S rRNA gene fragments by PCR



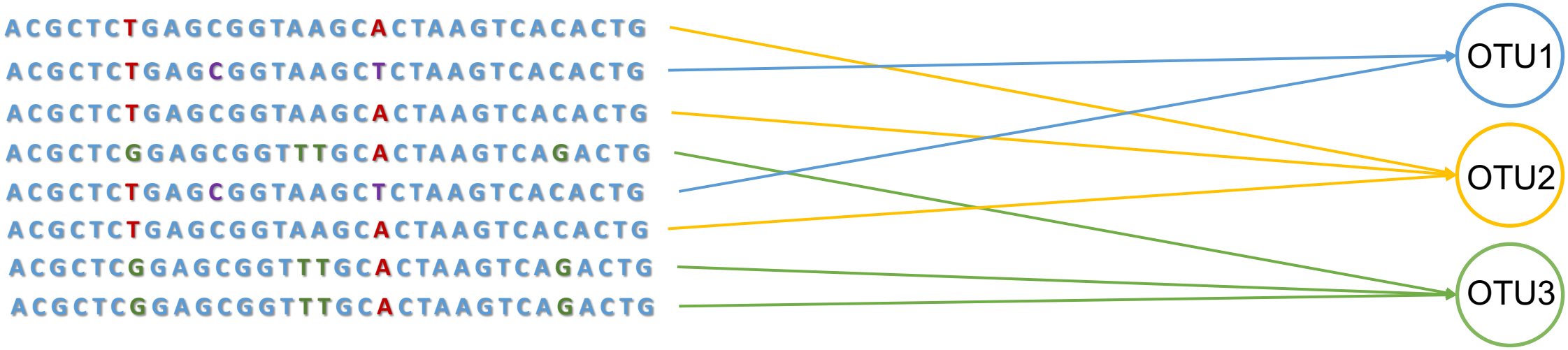
16S rRNA amplicons

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')

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

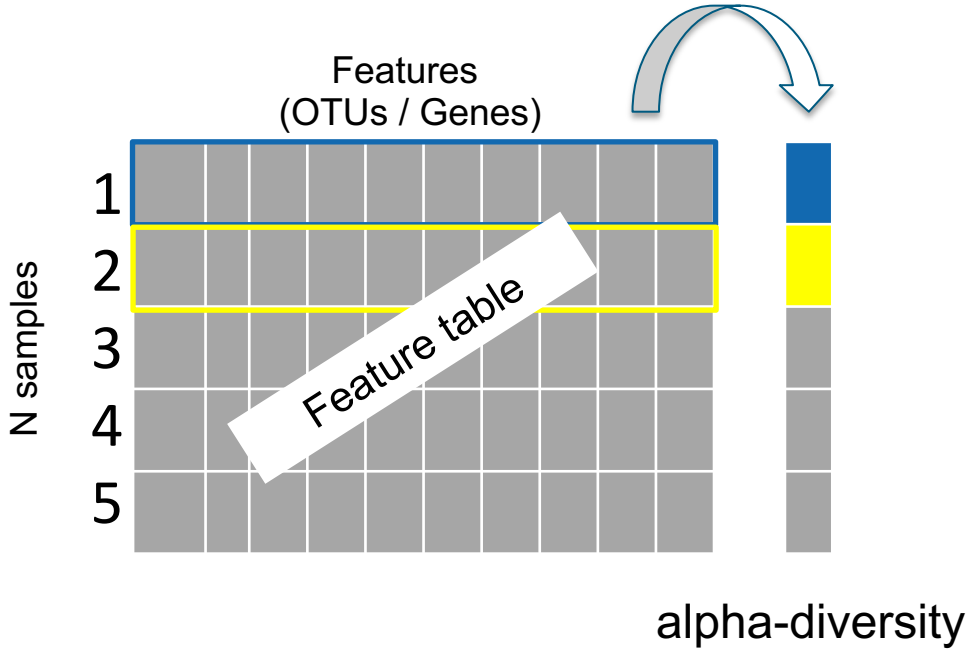
where $$H'_{max} = -\sum_{i=1}^{R} \frac{1}{R} ln \frac{1}{R} = \ln R$$

that is, every species is equally likely

Concept of alpha diversity - Summary

Microbial community structure

- microbial taxonomy and operational taxonomic units: **taxonomic profiling**
- diversity within a microbial community: **alpha diversity**



Summary - Part I

- Metagenomics facilitates the study of microorganisms, many of which have not been cultivated yet

- Taxonomic marker genes sequences are used to:
  - define operational taxonomic units
  - study the phylogenetic relatedness of bacteria and archaea
  - quantify the composition of microbial communities

- Alpha diversity (within sample diversity) is a function of richness and evenness

Overview of the Metagenomics part

Microbial community structure

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

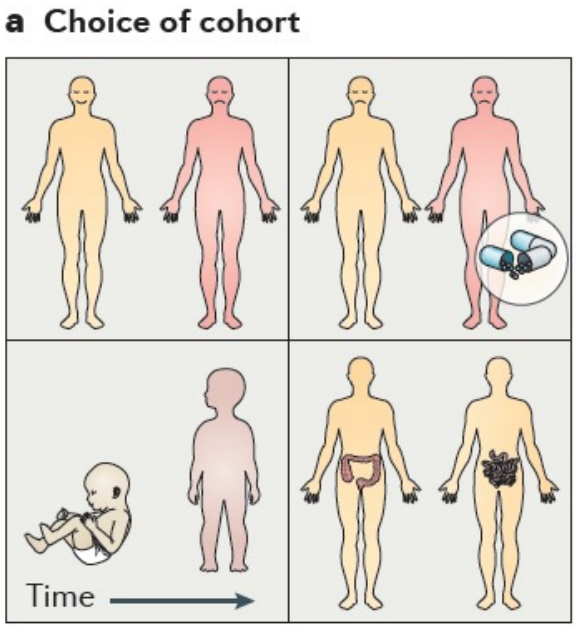
Differences between microbial communities

- 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:

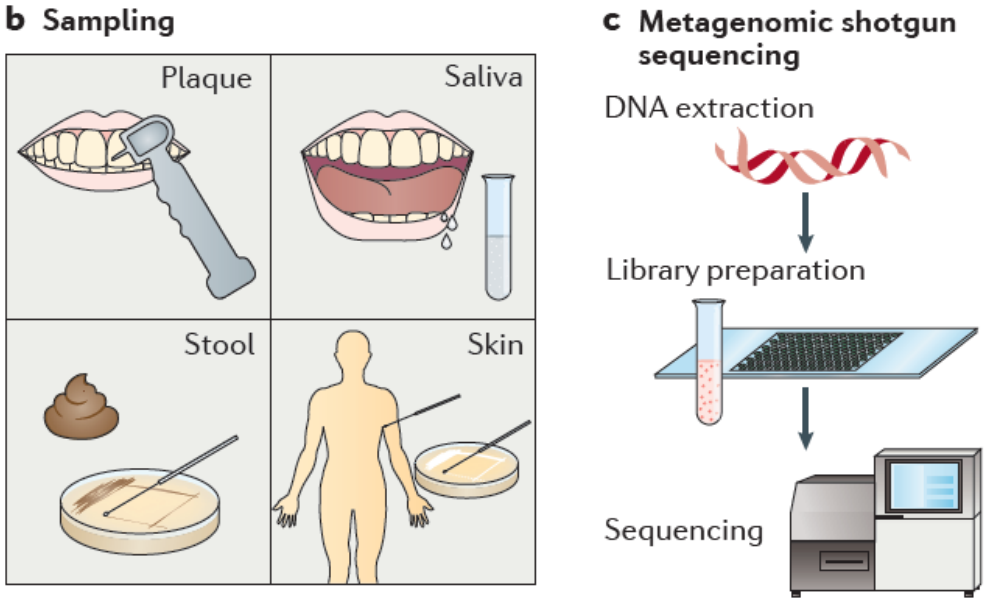
- groups of individuals and/or health states
- differential response to drugs (or nutrition)
- organismal development (or disease progression)
- differences between body sites

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

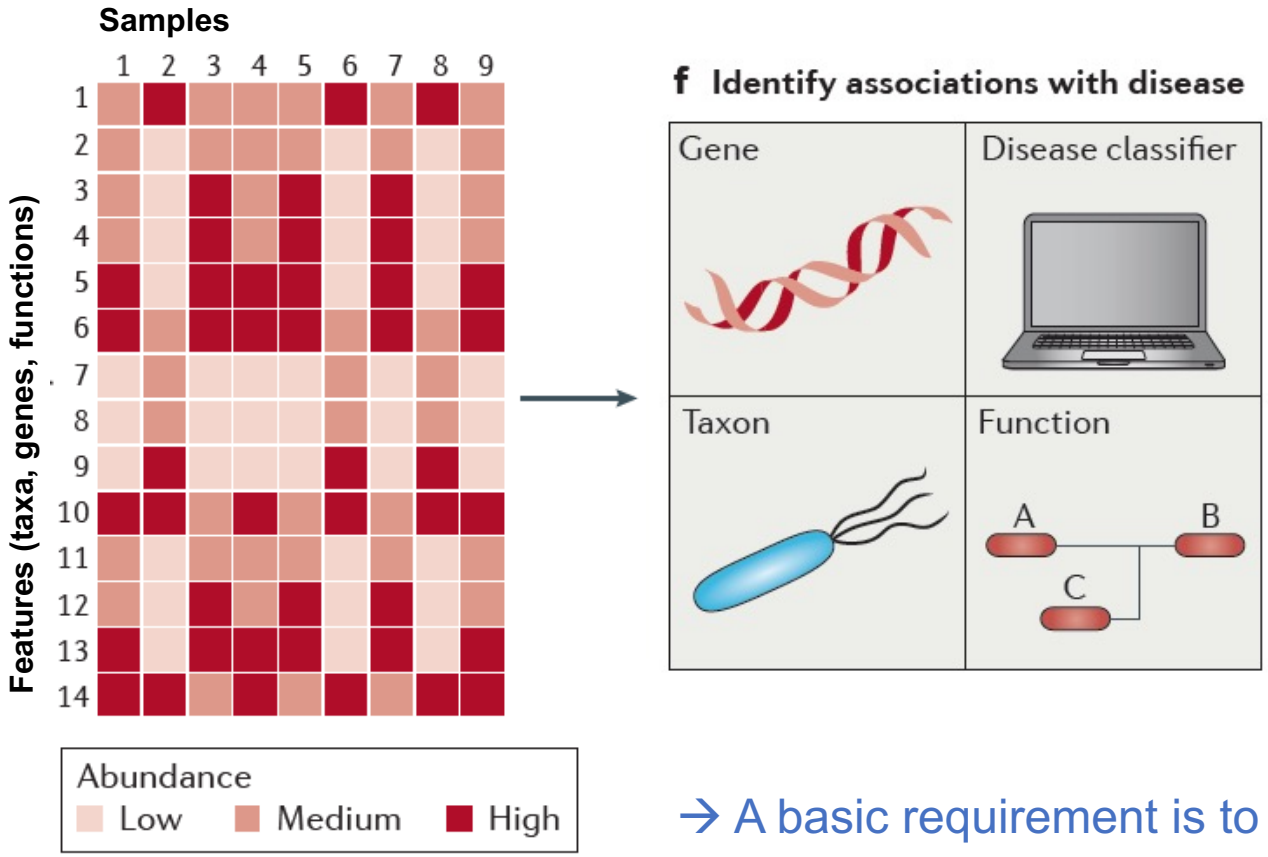
Microbiome-wide association studies are analogous to GWAS



- Microbial community DNA is extracted from samples and randomly sheared into fragments
- DNA fragments are “repaired” and used to prepare sequencing libraries
- Libraries are subjected to high throughput sequencing

Wang and Jia, NRM, 2016

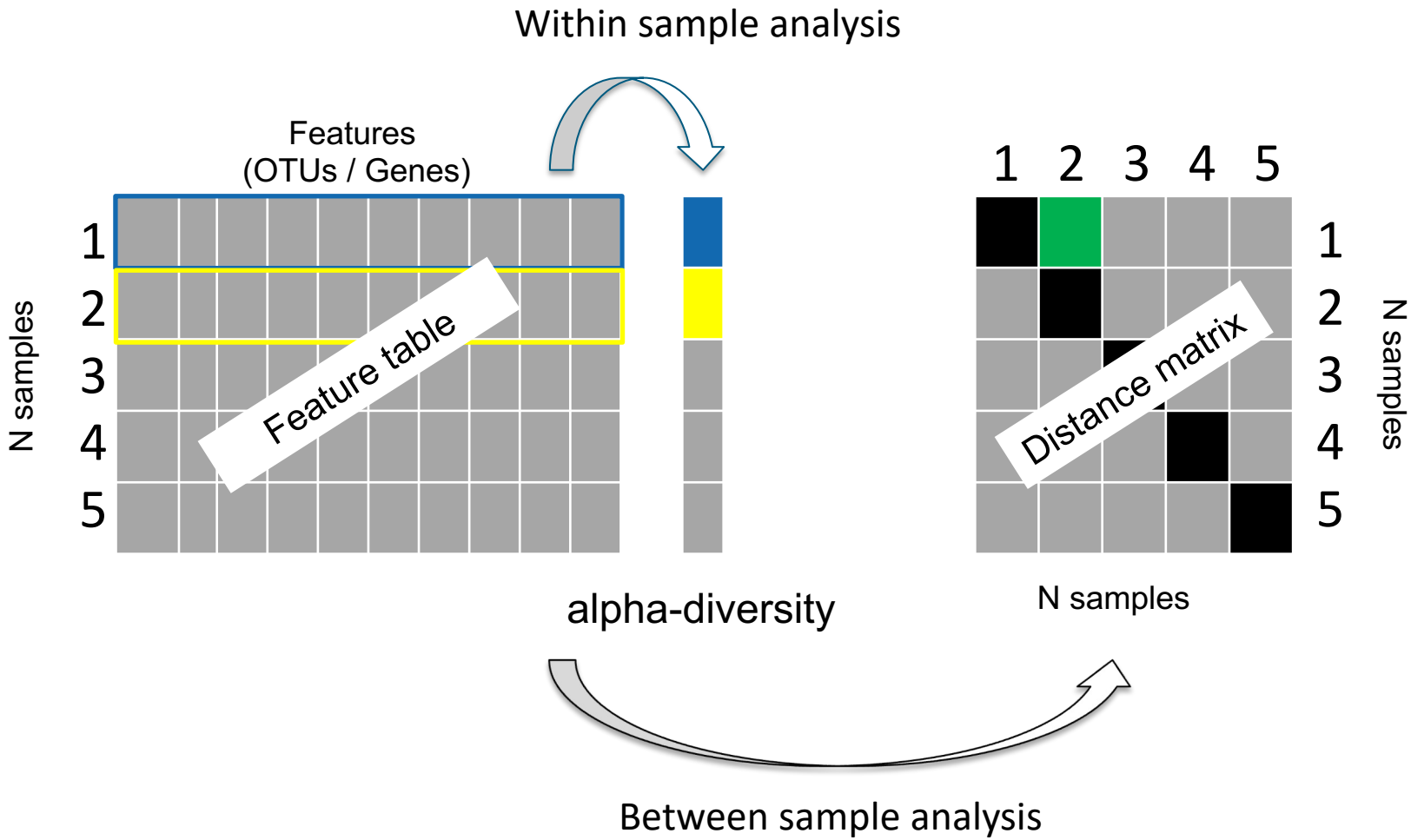
Microbiome-wide association studies are analogous to GWAS



- 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

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

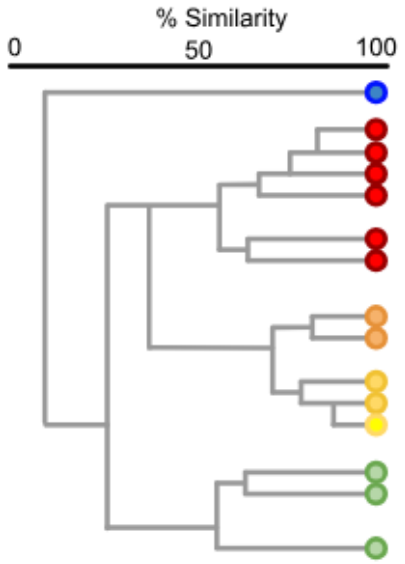
a_i = abundance of taxon *i* in sample *a*, and
c_i = abundance of taxon *i* in sample *c*

For additional practice, download the spread sheet named "Exercise – beta diversity" from Moodle, and calculate all pairwise Jaccard and Bray-Curtis distances for the example data.

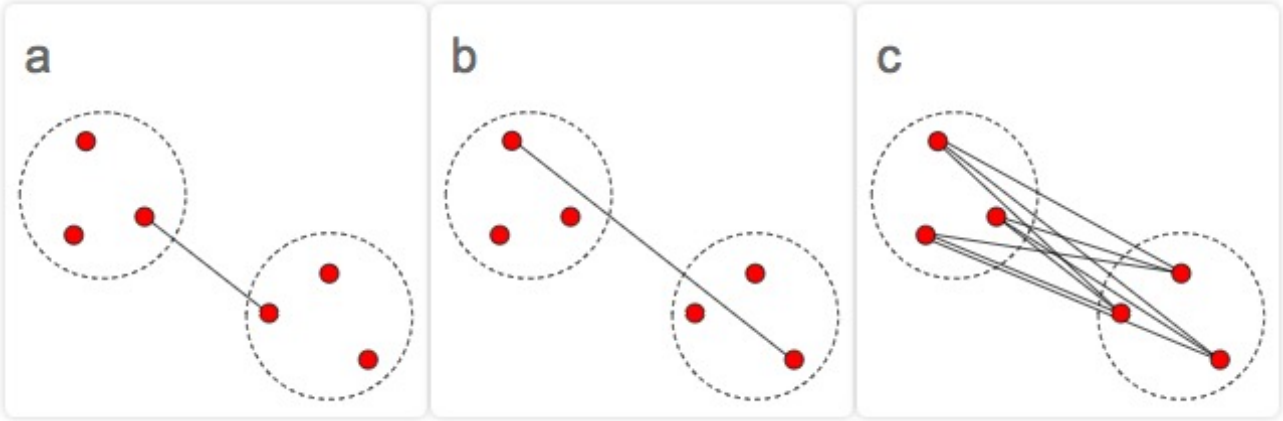
Visualize dissimilarities between microbial communities

- 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

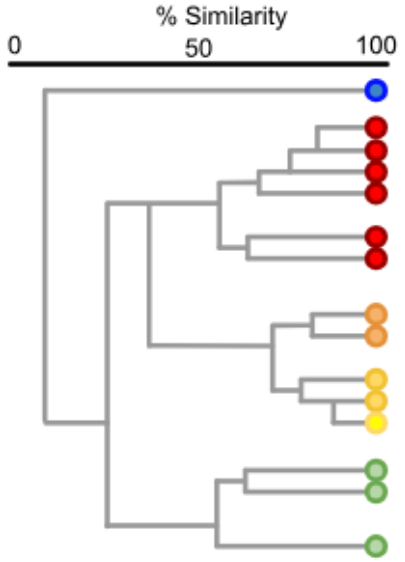


single complete average

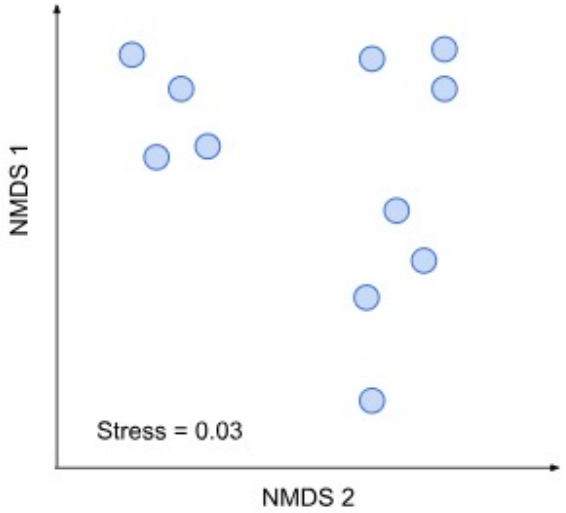
Visualize dissimilarities between microbial communities

- 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

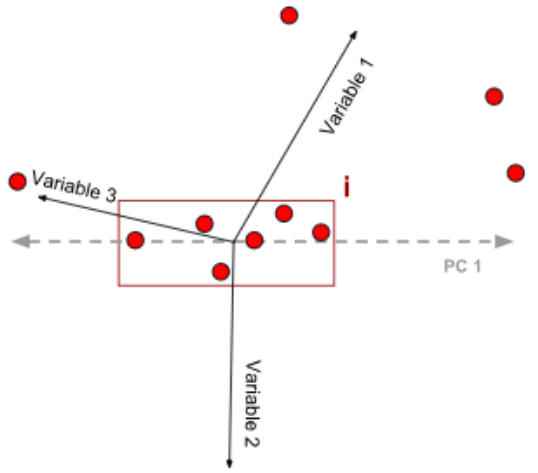
Hierarchical clustering



Non-metric dimensional scaling (NMDS)



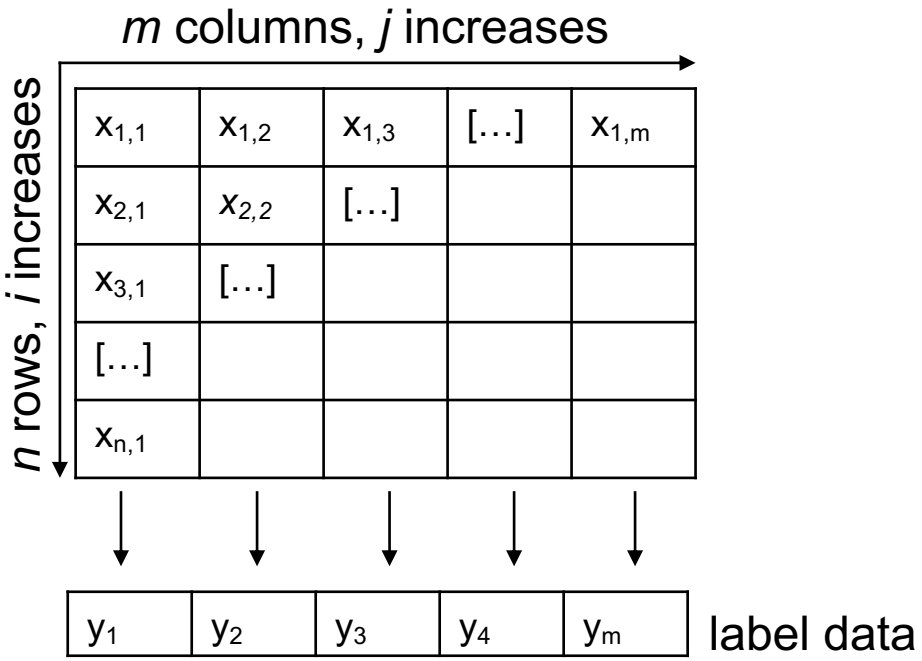
Principal component or coordinate analysis (PCA or PCoA)



Generalization and notation

Matrix m x n

where element x*i,j* is in row *i* and column *j*, and max(*i*) = *n* and max(*j*) = *m*

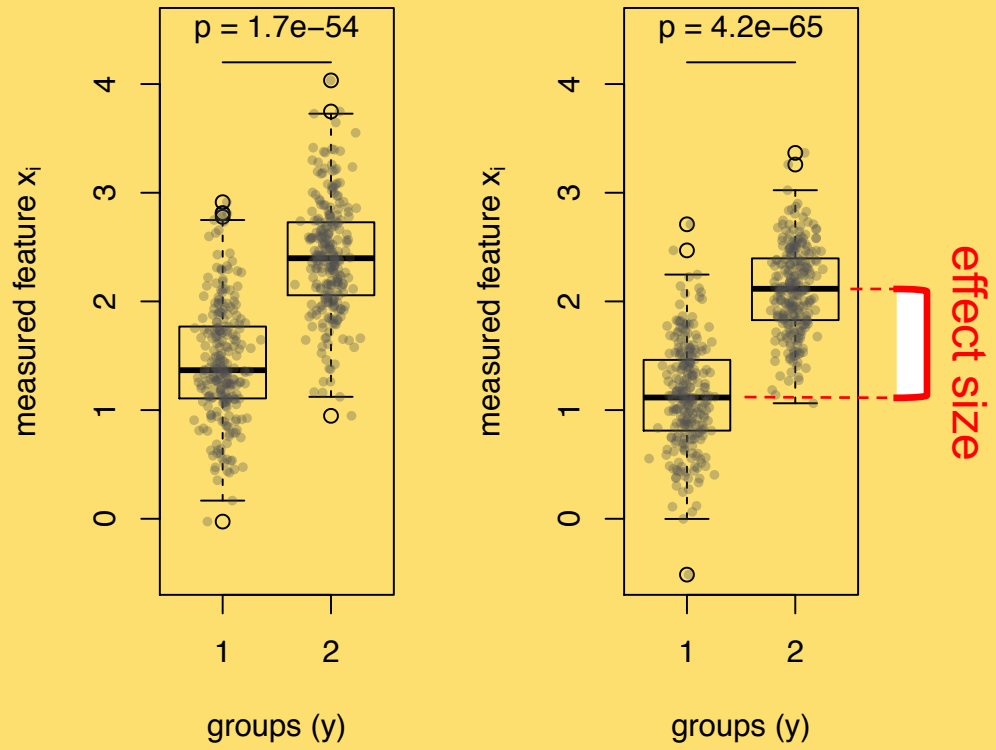


- Feature data **x** (or observations, predictors):
  - *i*: rows → feature, *j*: columns → samples
  - $\mathbf{x}_i$ denotes the vector for the *i*-th feature
  - $\mathbf{x}_{ij}$ denotes *i*-th feature in *j*-th sample
- Label data **y** (or dependent variable, response)
  - vector of length *m*
- Example: labels for y are 1=healthy, 2=diseased

Label	binary	binary
y_1=healthy	1	h
y_2=healthy	1	h
y_3=diseased	2	d
y_4=healthy	1	h
[…]	[…]	[…]

Determine differentially abundant features

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



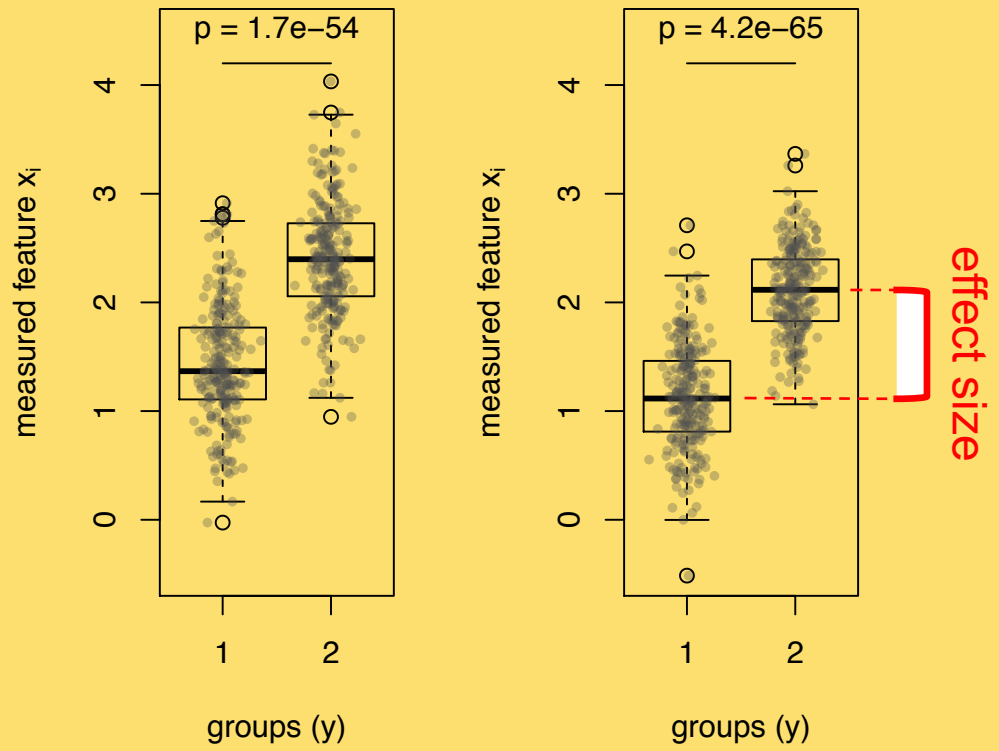
Question 1: in a clinical trial, you observe differences in the taxonomic composition of stool samples from healthy vs. diseased individuals. Assuming it to be a true effect, what do you expect from sampling additional individuals?

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

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
- Predictive modeling approaches can be used to classify unknown 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)

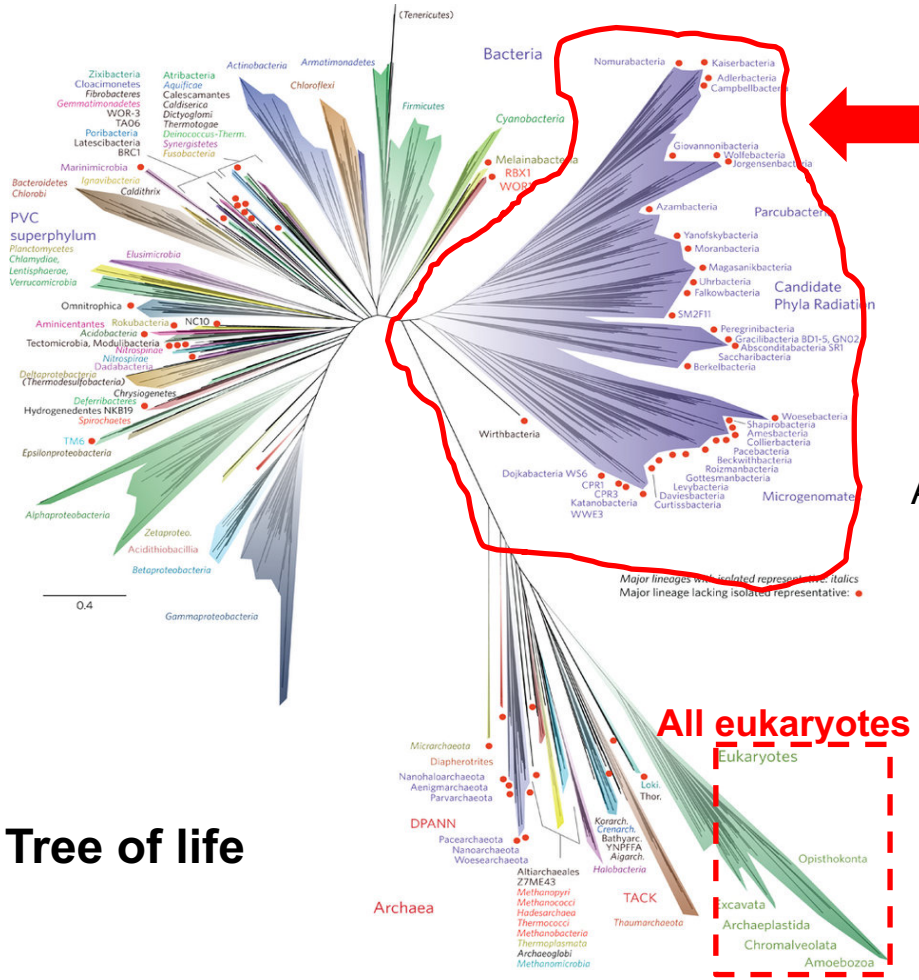
Working with microbial community genes and genomes

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

Reconstruction and annotation of microbial community genomes

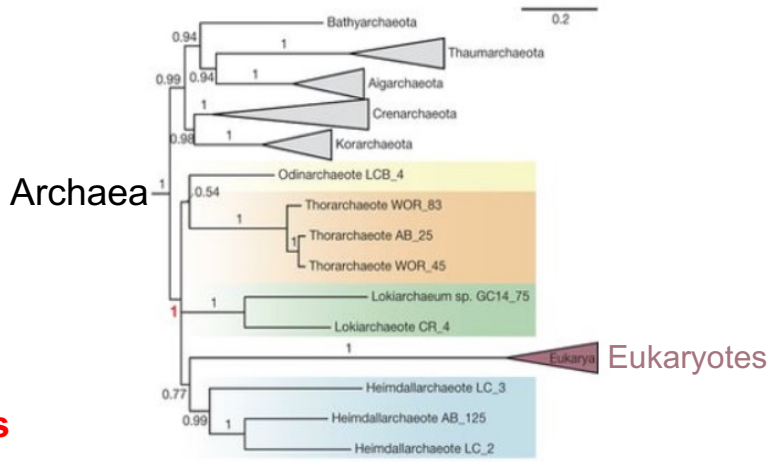
- Most organisms in microbial communities have not been isolated and cultured
  - However, we can sequence microbial community DNA, reconstruct genomes and predict protein sequences / structures
- Genomes reconstructed from natural environment capture microbial diversity on Earth
  - New data challenge long-standing concepts
- Predicted genes inform about functional capabilities and other traits of organisms
  - “Who is there?” → “What can they do?”
- Genomic information enable discovery of new enzymes and microbial 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

Insights by reconstructing microbial community genomes

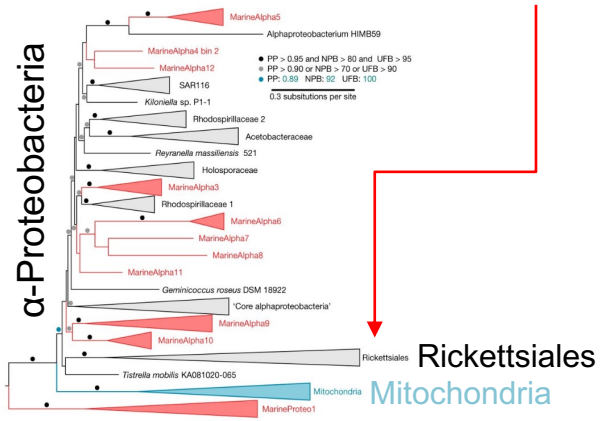


Candidate phyla radiation discovered by metagenomics

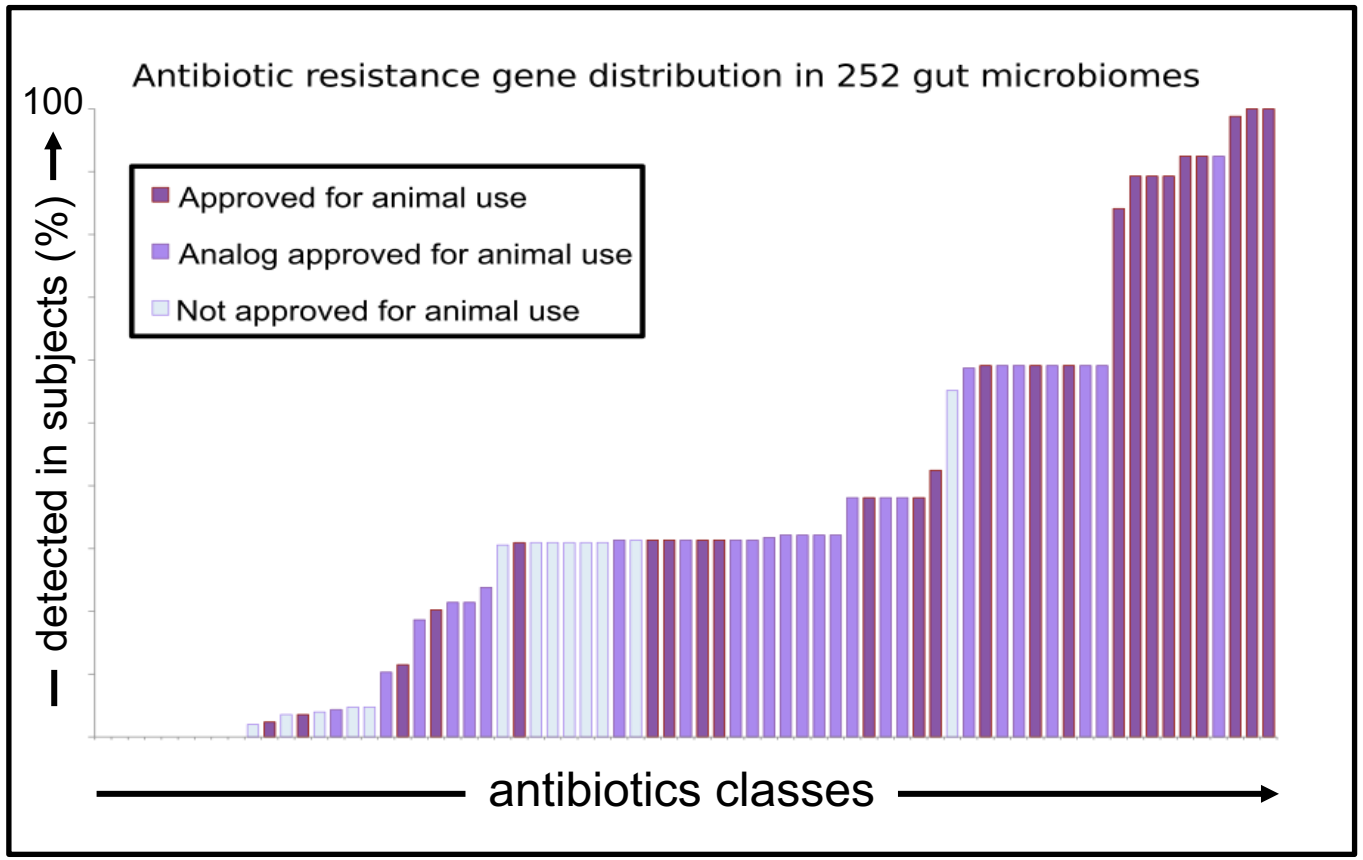
Two domains of life?

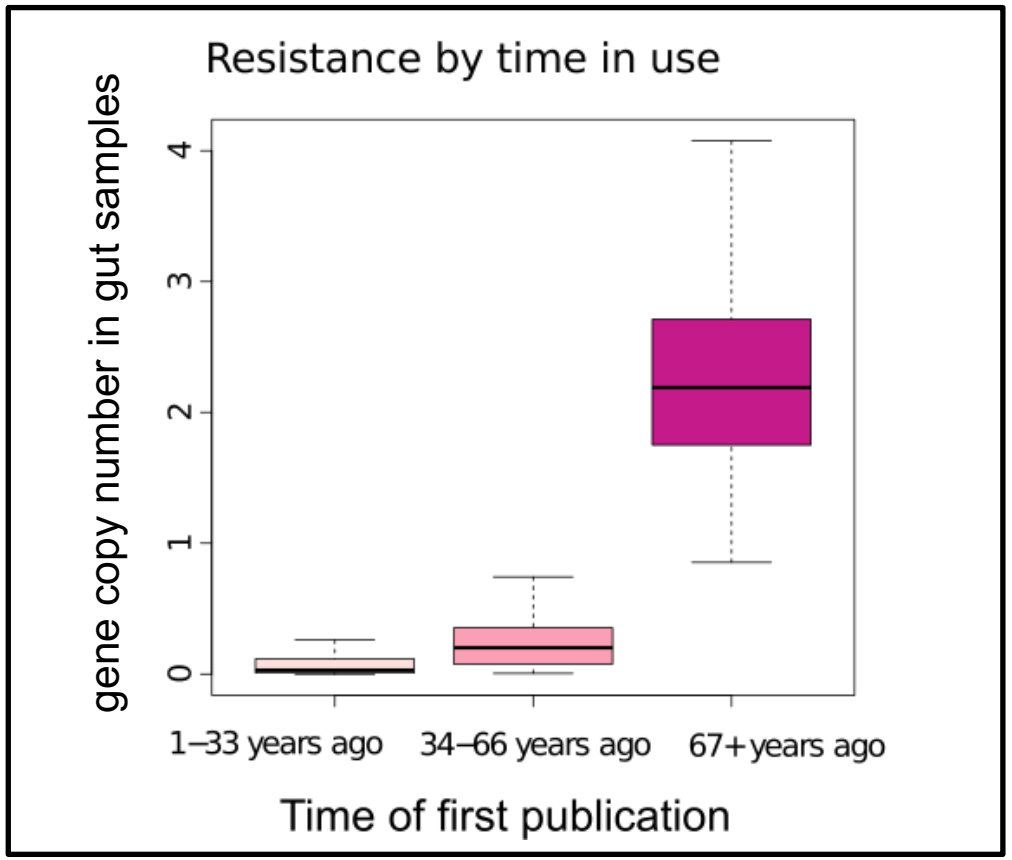


Mitochondrial origin not within Rickettsiales?

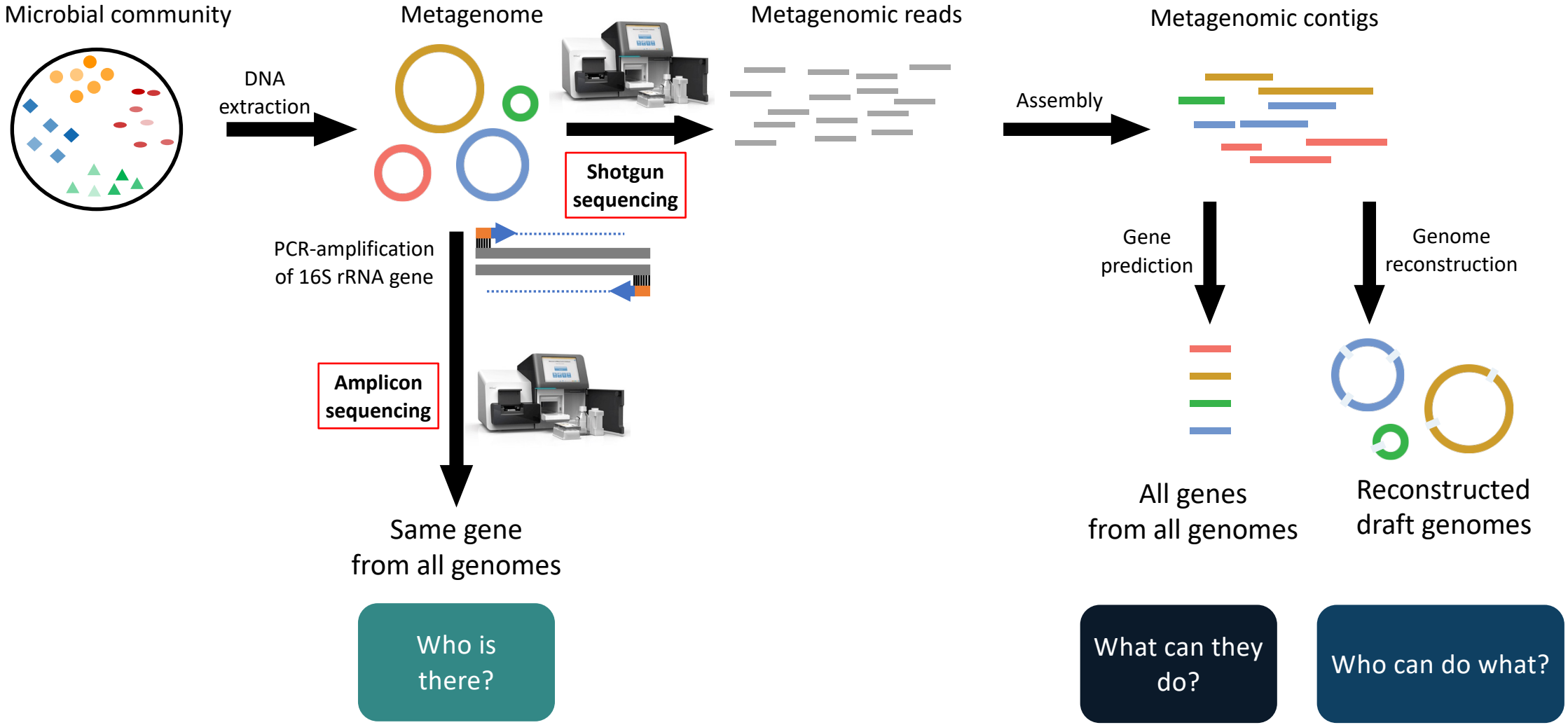


Insights by quantifying microbial gene abundances





Sequencing microbial community DNA



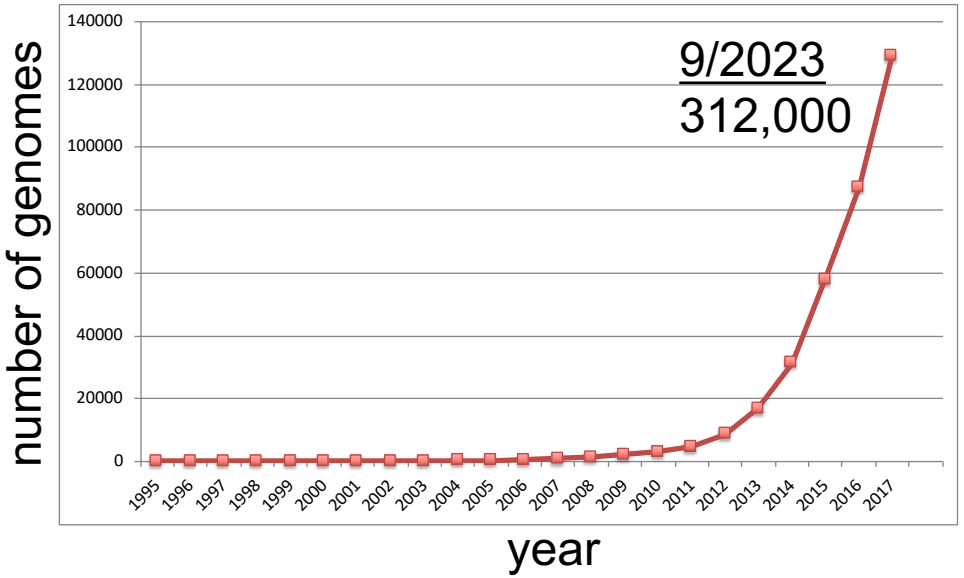
Sequencing of microbial isolate genomes

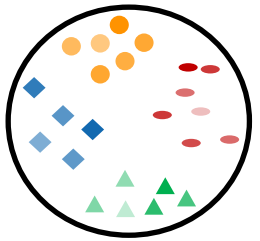
- First bacterial genome (1995): *Haemophilus influenzae* *Fleischmann et al. 1995*
- Followed by many isolated pathogens of diseases (e.g., plague, anthrax, tuberculosis, Lyme disease)
- Many isolates of important non-pathogenic species: e.g., *Prochlorococcus*, *Lactobacillus*, *Bradyrhizobium*

- 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

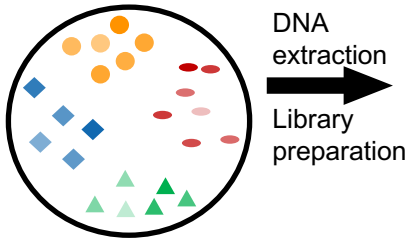




Microbial community

→However, most bacteria and archaea have not been isolated and sequenced.

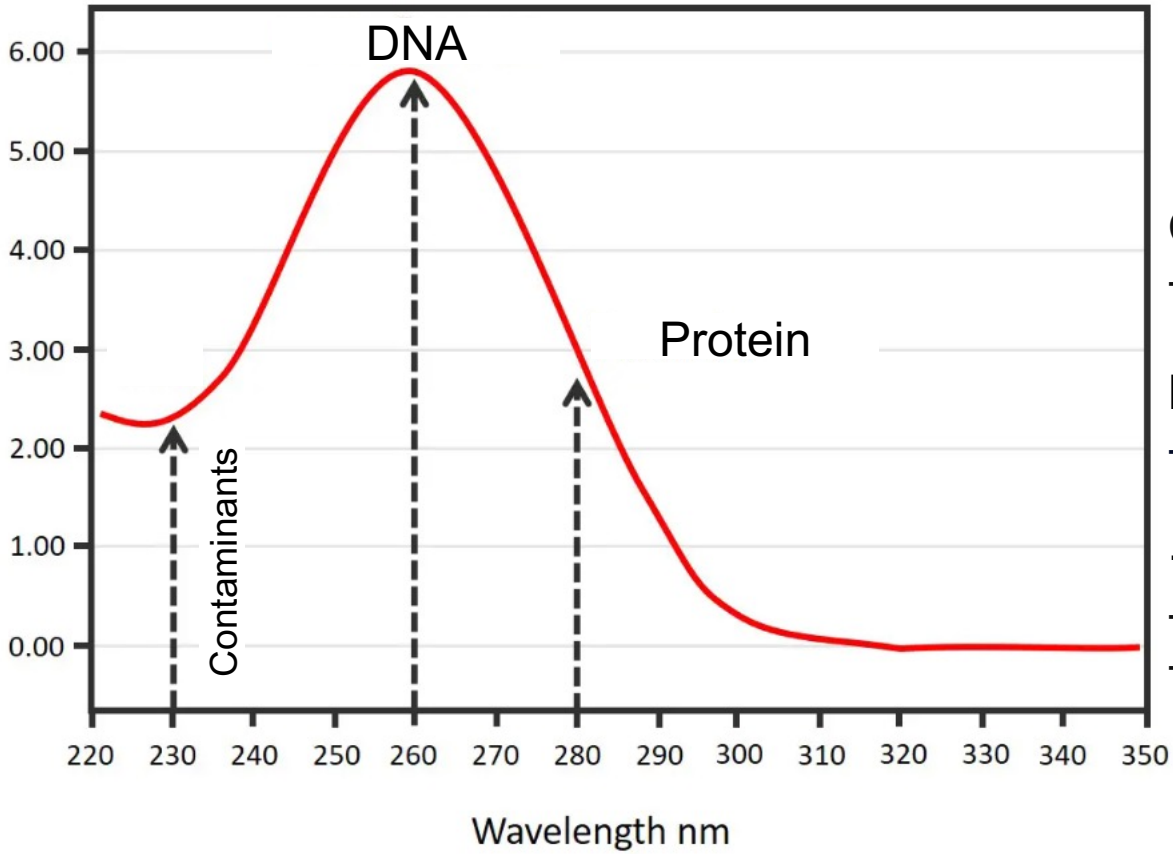
Metagenomics provides access, in principle, to all genomic information within a microbial community. This allows us to ask: “what can they do?”, in addition to: “who is there?”.



Microbial community

DNA extraction

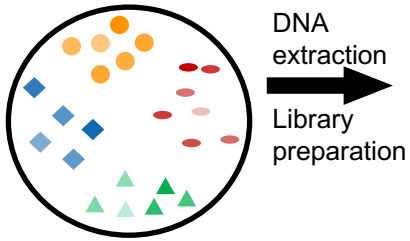
- Sufficiently high quality and quantity needed



Contaminants:
- e.g., phenol, carbohydrates, EDTA

Protein:
- tyrosine and tryptophan

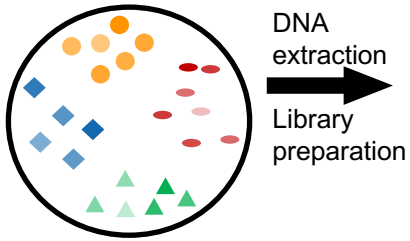
→ DNA quality:
- 260/280 ratio
- 260/230 ratio



Microbial community

DNA extraction

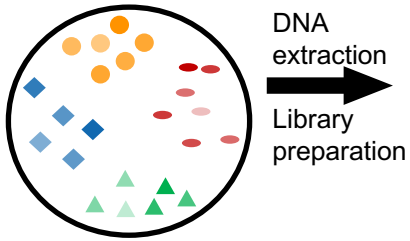
- Sufficiently high quality and quantity needed
  - >1ng (Illumina); >50-400ng (Oxford Nanopore Technologies – ONT / PacBio)
  - lower amounts possible after amplification of DNA



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: no limit



Microbial community

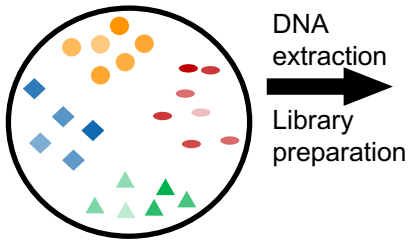
DNA extraction / library preparation

- Sufficiently high quality and quantity needed
- Extracted DNA is sheared into smaller fragments (inserts)
  - Illumina: ~300-600 bp; PacBio: ~20 kbp; ONT: no limit
- Adapters (of known sequences) are added to inserts
  - To allow for multiplexing samples; as templates for sequencing primers

Example (Illumina library):

- Index sequences (i5, i7) allow for multiplexing
- Illumina adapters (P5, P7) as template for forward and reverse (i.e., paired-end) sequencing primers

```
5'- AATGATACGGCGACCACCGAGATCTACACNNNNNNNNACACTCTTTCCCTACACGACGCTCTTCCGATCT-insert-AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG -3'
3'- TTACTATGCCGCTGGTGGCTCTAGATGTGNNNNNNNNTGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA-insert-TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGNNNNNNNNTAGAGCATACGGCAGAAGACGAAC -5'
       Illumina P5              i5          Truseq Read 1                        Truseq Read 2              i7        Illumina P7
```



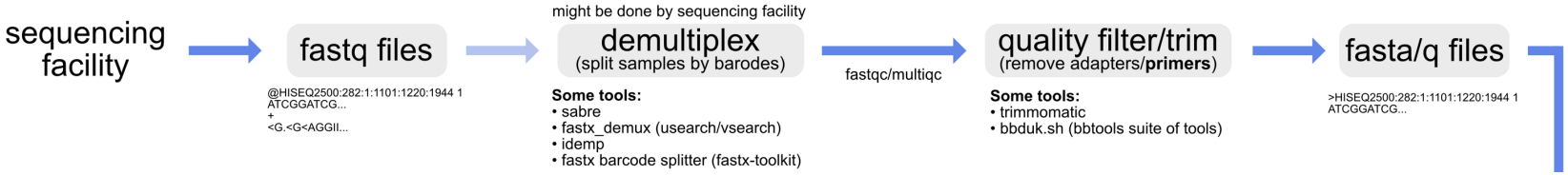
DNA extraction

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.



From raw sequencing reads to high-quality reads

- Removal of multiplexing and sequencing adapters
- Residual control DNA sequences (e.g., “PhiX spike-ins”)
- Removal of sequences from non-target organisms (contamination)
- Removal of low-quality bases (“trimming”) from sequencing reads

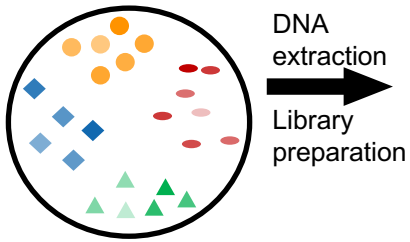
Base calling quality (*phred*) scores

$Q = -10 \log_{10} P$

Probability error: $P = 10^{-Q/10}$

Probability truth: $1 - P$

Quality score	% Correct Base
40	99.99
30	99.9
20	99
10	90



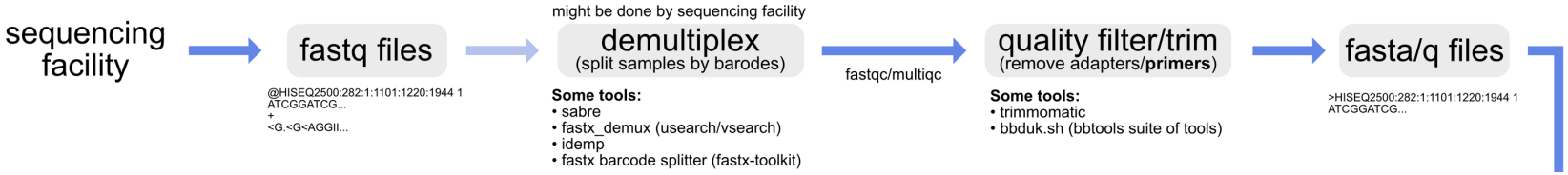
DNA extraction

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.



From raw sequencing reads to high-quality reads

- Standard format for sequencing reads (FASTA/Q)
- https://en.wikipedia.org/wiki/FASTQ_format

Example:

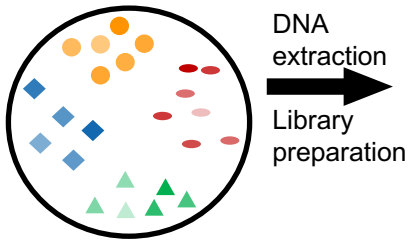
two “forward” reads:

```
@071112_SLXA-EAS1_s_7:5:1:817:345
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+071112_SLXA-EAS1_s_7:5:1:817:345
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
@071112_SLXA-EAS1_s_7:5:1:801:338
GTTCAGGGATACGACGTTTGTATTTTAAGAATCTGA
+071112_SLXA-EAS1_s_7:5:1:801:338
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII6IBI
```

two “reverse” reads:

```
@071112_SLXA-EAS1_s_7:5:1:817:345
AAGTTACCCTTAACAACTTAAGGGTTTTCAAATAGA
+071112_SLXA-EAS1_s_7:5:1:817:345
IIIIIIIIIIIIIIIIIIIDIIIIIII>IIIIII/
@071112_SLXA-EAS1_s_7:5:1:801:338
AGCAGAAGTCGATGATAATACGCGTCGTTTTATCAT
+071112_SLXA-EAS1_s_7:5:1:801:338
IIIIIIIIIIIIIIIIIIIIGII>IIIII-I)8I
```

Each read = 4 rows:
- sequence header
- **nucleotide sequence**
- header repeated
- encoded quality score



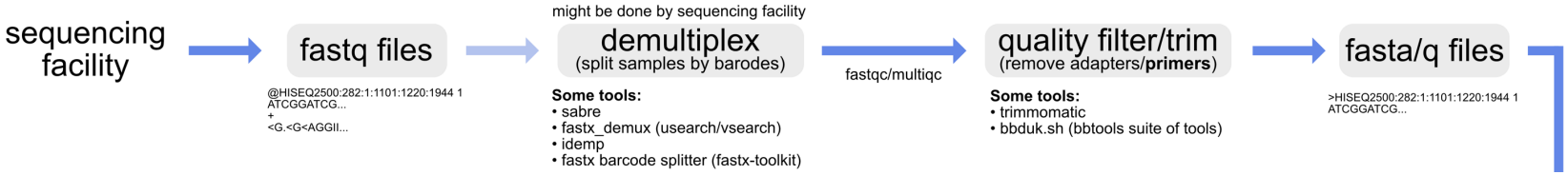
DNA extraction

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.



Storage of DNA sequence information

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

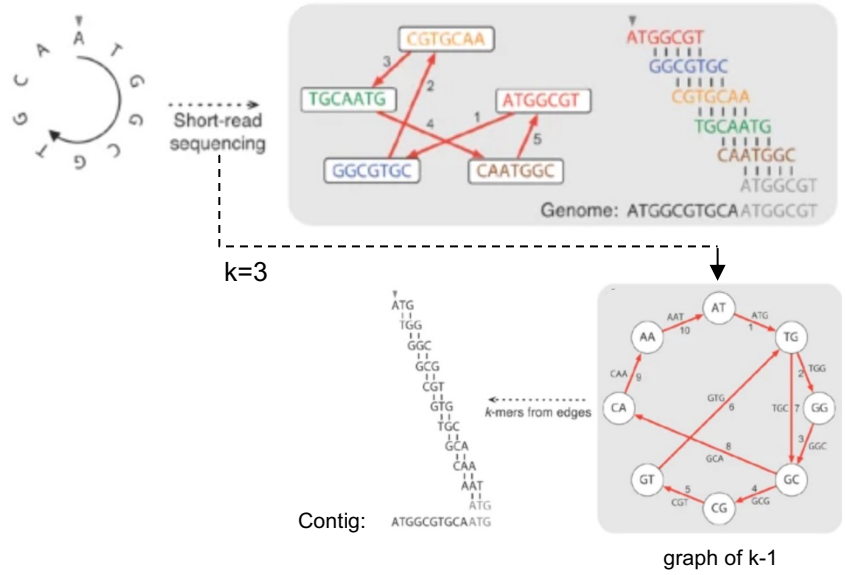
Databases

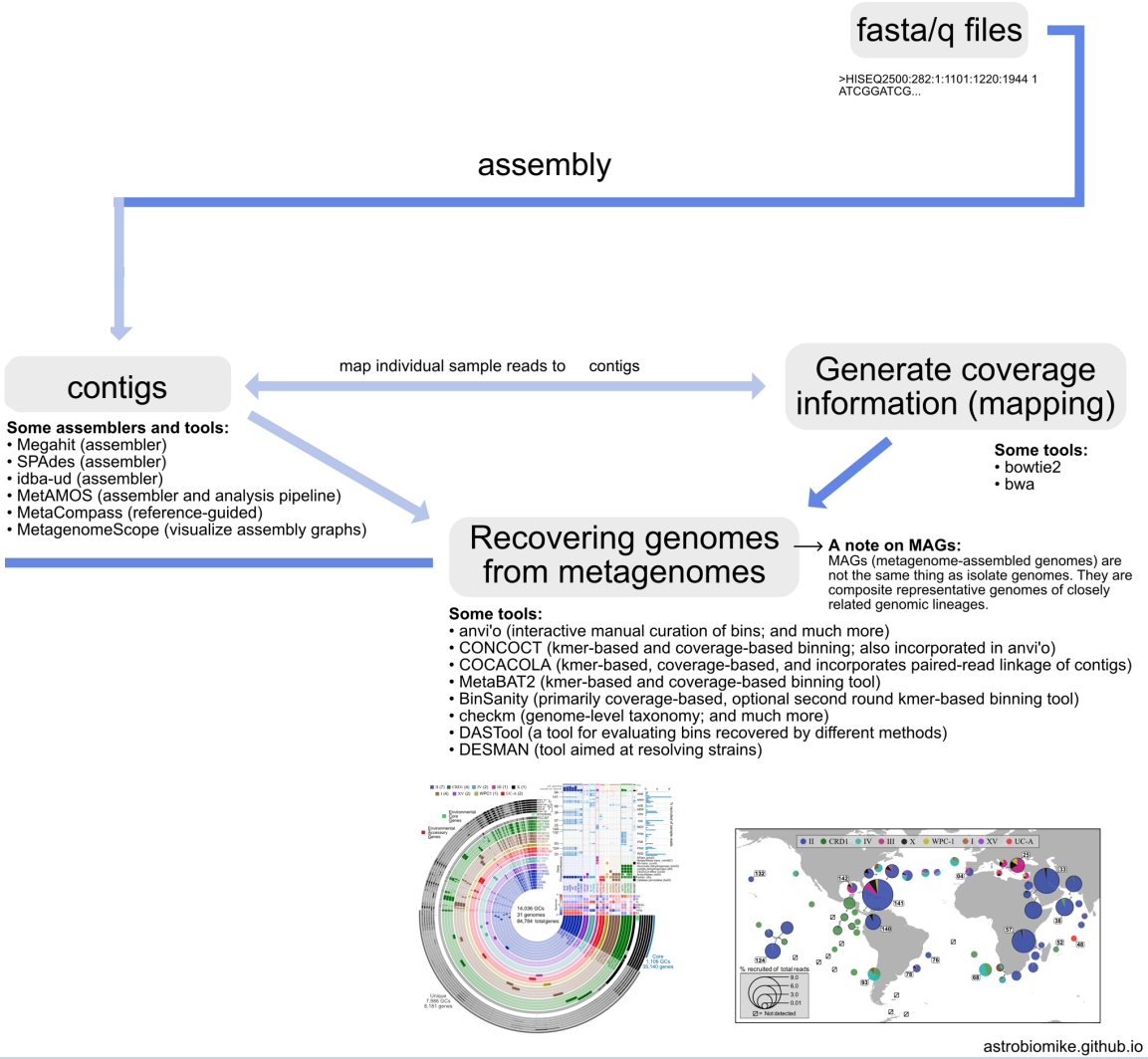
Data type	DDBJ	EMBL-EBI	NCBI
Next Generation reads	Sequence Read Archive		Sequence Read Archive
Assembled Sequences	DDBJ	European Nucleotide Archive	GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

Overview of generic* metagenomics workflow

Assembly of reads into contigs

- Traditionally, all-by-all alignemts and shortest “path” through reads = contig
- Today, reads are reduced to k-mers to find shortest paths through all k-mers

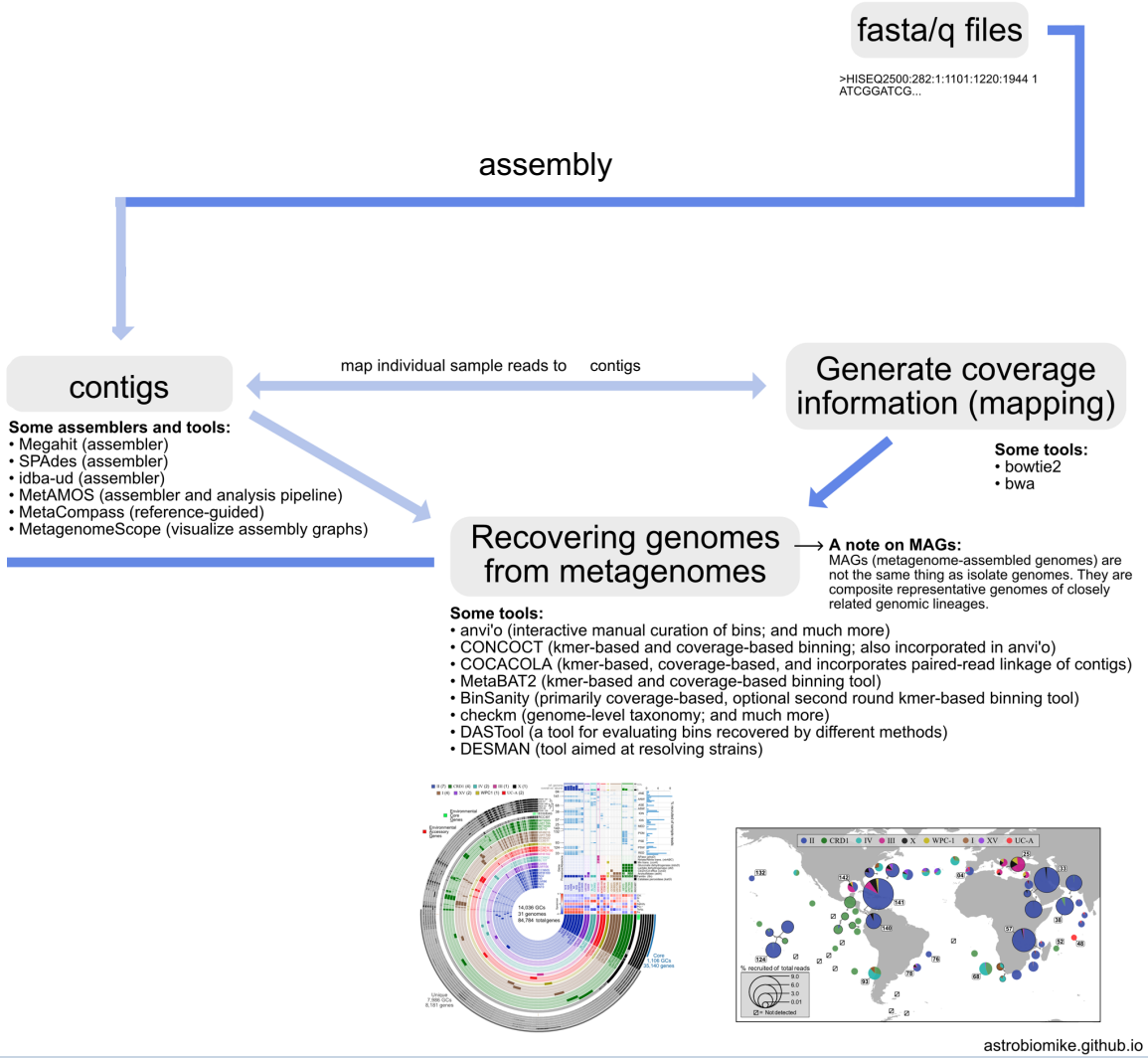




Overview of generic* metagenomics workflow

Assembly of reads into contigs

- Traditionally, all-by-all alignemts and shortest “path” through reads = contig
- Today, reads are reduced to k-mers to find shortest paths through all k-mers
- Metagenomic *de-novo* assemblies produce many fragments of genomes (i.e., contigs from different genomes)

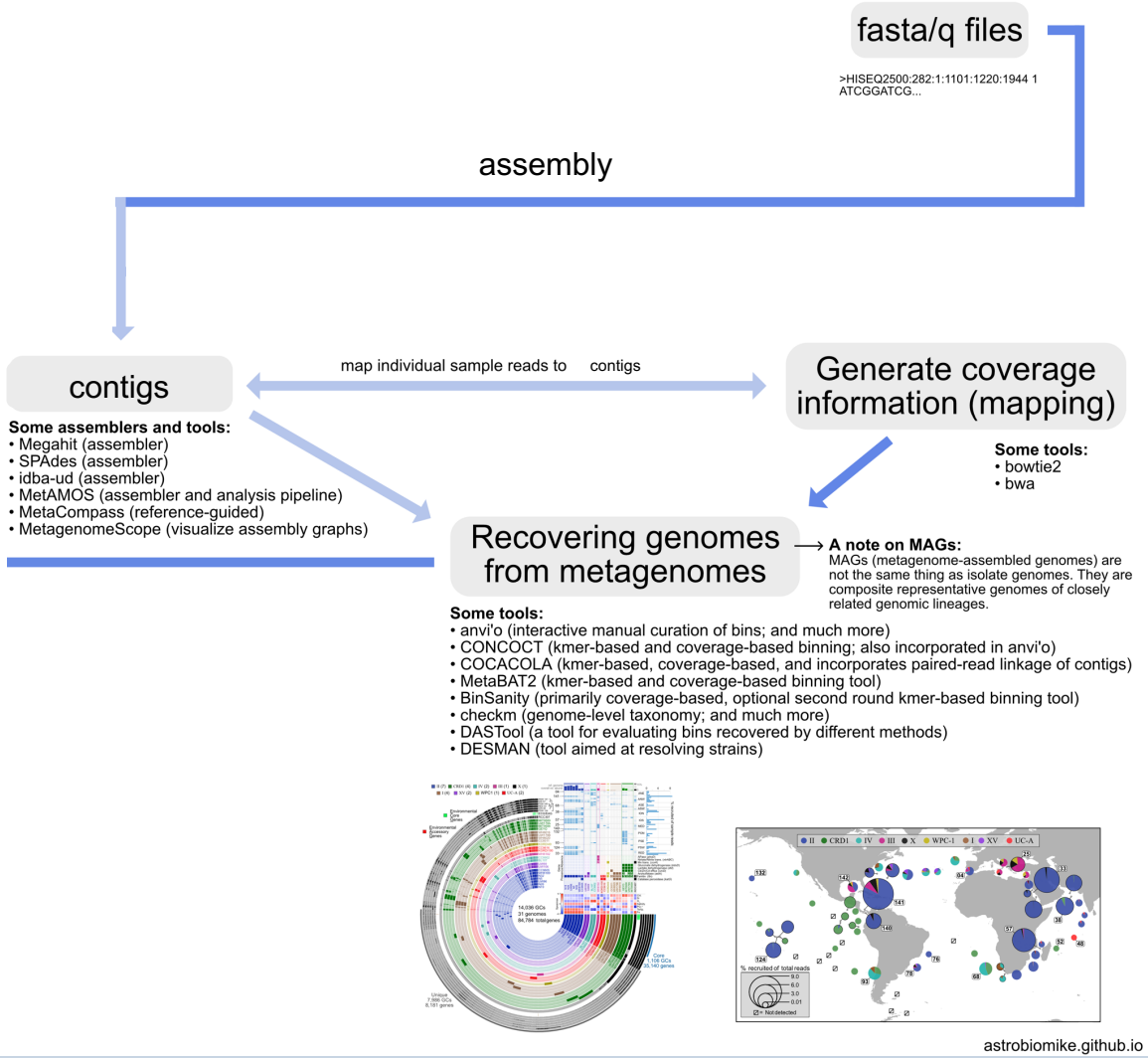


Overview of generic* metagenomics workflow

Assembly of reads into contigs

- Traditionally, all-by-all alignemts and shortest “path” through reads = contig
- Today, reads are reduced to k-mers to find shortest paths through all k-mers
- Metagenomic *de-novo* assemblies produce many fragments of genomes (i.e., contigs from different genomes)

- Contigs are “binned” into metagenome-assembled genomes (MAGs)



astrobiomike.github.io

Quality of metagenome-assembled genomes

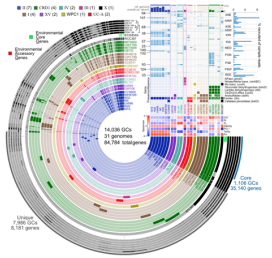
Quality of MAGs

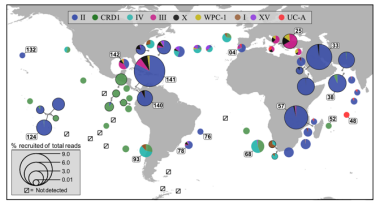
- How do we assess if:
  - a) contigs were binned correctly?
    → contamination
  - b) all contigs of a genome were identified?
    → completeness

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)





astrobiomike.github.io

Quality of metagenome-assembled genomes

Quality of MAGs

- How do we assess if:
  - a) contigs were binned correctly?
    → contamination
  - b) all contigs of a genome were identified?
    → completeness
- Genome Taxonomy Database (GTDB)
  - Contains genomes from isolates and environmental genomes, i.e., MAGs (and SAGs)
  - Uses single copy marker genes (scMGs) present in all bacteria (120) and all archaea (53) to provide a 'ground truth' for expected content in reconstructed genomes

GTDB R214 spans 402,709 genomes organized into 85,205 species clusters.

	Bacteria	Archaea	Total
Phylum	161	20	181
Class	488	60	548
Order	1,624	148	1,772
Family	4,262	508	4,772
Genus	19,153	1,586	20,739
Species	80,789	4,416	85,205

Quality of metagenome-assembled genomes

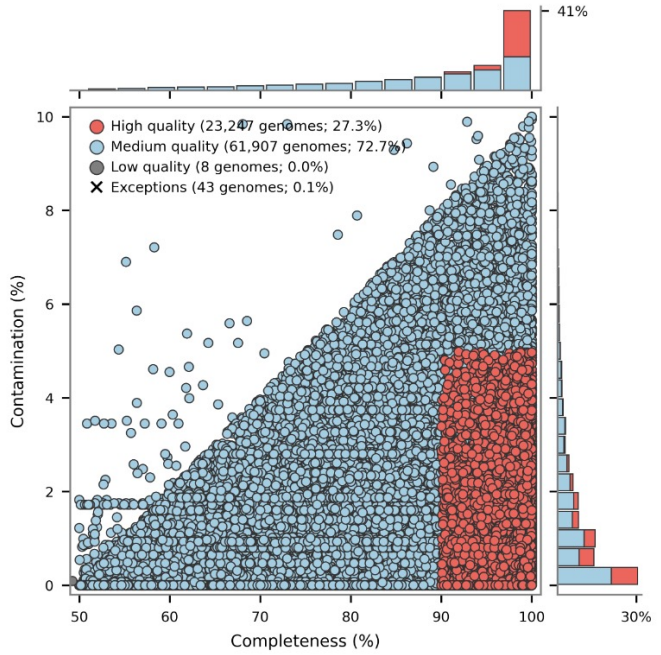
Quality of MAGs

- How do we assess if:
  - a) contigs were binned correctly?
    → contamination
  - b) all contigs of a genome were identified?
    → completeness
- Genome Taxonomy Database (GTDB)
  - Contains genomes from isolates and environmental genomes, i.e., MAGs (and SAGs)
  - Uses single copy marker genes (scMGs) present in all bacteria (120) and all archaea (53) to provide a 'ground truth' for expected content in reconstructed genomes

→ Completeness (fraction of scMGs detected)
→ Contamination (fraction of scMGs from non-target taxa)

Commonly used criteria for genome quality

Completeness/contamination:
High: ≥ 90% / ≤ 5%
Good: ≥ 70% / ≤ 10%
Medium: ≥ 50% / ≤ 10%



Parks et al, Genome Res, 2015

Taxonomic annotation of metagenome-assembled genomes

Quality of MAGs

- How do we assess if:
  - a) contigs were binned correctly?
    → contamination
  - b) all contigs of a genome were identified?
    → completeness
- Genome Taxonomy Database (GTDB)
  - Contains genomes from isolates and environmental genomes, i.e., MAGs (and SAGs)
  - Uses single copy marker genes (scMGs) present in all bacteria (120) and all archaea (53) to provide a 'ground truth' for expected content in reconstructed genomes

→ Completeness (fraction of scMGs detected)
→ Contamination (fraction of scMGs from non-target taxa)

Taxonomic annotation of MAGs

- How do we assess if:
  - which taxon the genome belongs to?
  - the reconstructed genome represents a 'novel' species?

Genome Taxonomy Database (GTDB) – toolkit (Tk)

- Compares genome to genomes in the GTDB
  - scMGs from genome are extracted
  - Genomes are classified based on placement of genome in GTDB reference tree
  - If identity to any reference genome is <95%

→ Novel species

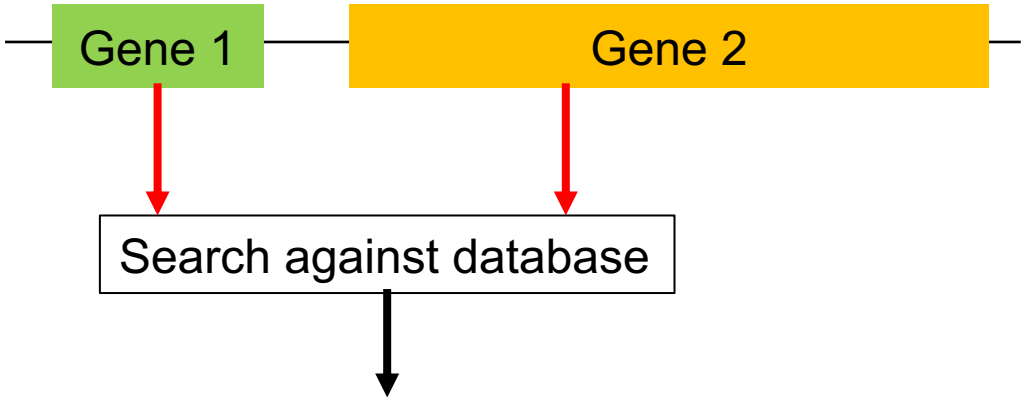
GTDB TK

Notation:

GTDB TAXONOMY
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Synechococcales;f__Cyanobiaceae;g__Synechococcus_C;s__
d__Bacteria;p__Eremiobacterota;c__UBP9;o__UBA4705;f__;g__;s__

Parks et al, Genome Res, 2015

Functional annotation of genes



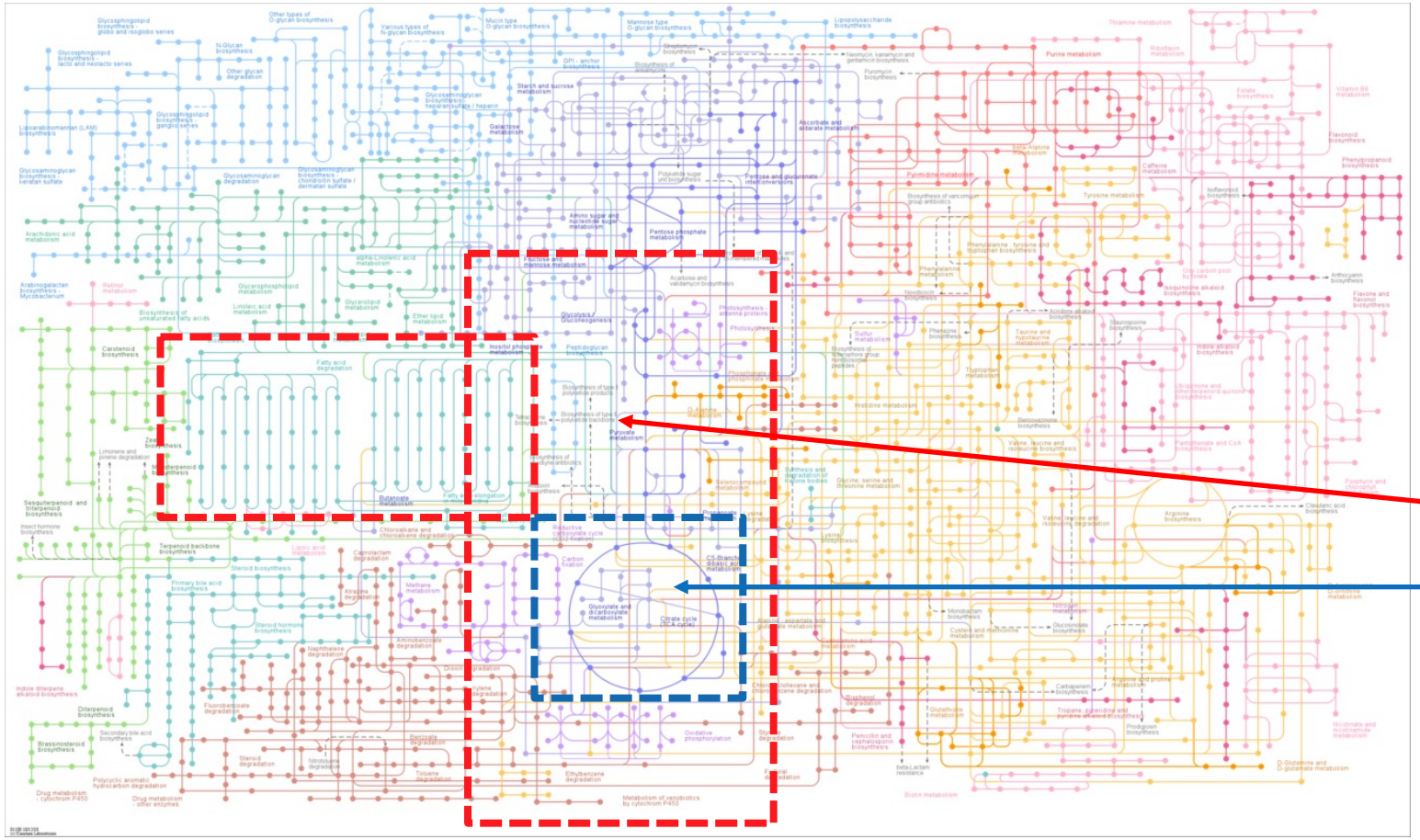
	DB	Identifier
Database 1:	KEGG	K00847
Database 2:	COG	COG1940
Database 3:	Pfam	PF00480

Goal: assign a function to a gene

- Gene sequences can be searched by many different tools using different algorithms against many different databases that store sequences of functionally annotated genes
- Commonly 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…

Example: KEGG database

Kyoto **E**ncyclopedia of **G**enes and **G**enomes



Map of known metabolic reactions

- Nodes = compounds
- Connections = reactions catalyzed by known enzymes
- Enzymes grouped into KOs = KEGG orthologous groups
- Map divided into:
  - pathways and
  - modules

Functional annotation of genes: KEGG database

First level (Pathways)

CARBON METABOLISM

CARBON FIXATION

METHANE METABOLISM

[…]

Second level (Modules)

M00009 – Citrate cycle (Krebs cycle)

[…]

Third level (KOs: KEGG orthologous groups)

K01647

K05942

K01681

K01682

K00031

K00030

K00164

K00658

K00382

K00174

K00175

K00177

[…]

Forth level (genes)

GeneID i in species y

GeneID i+1 in species y+1

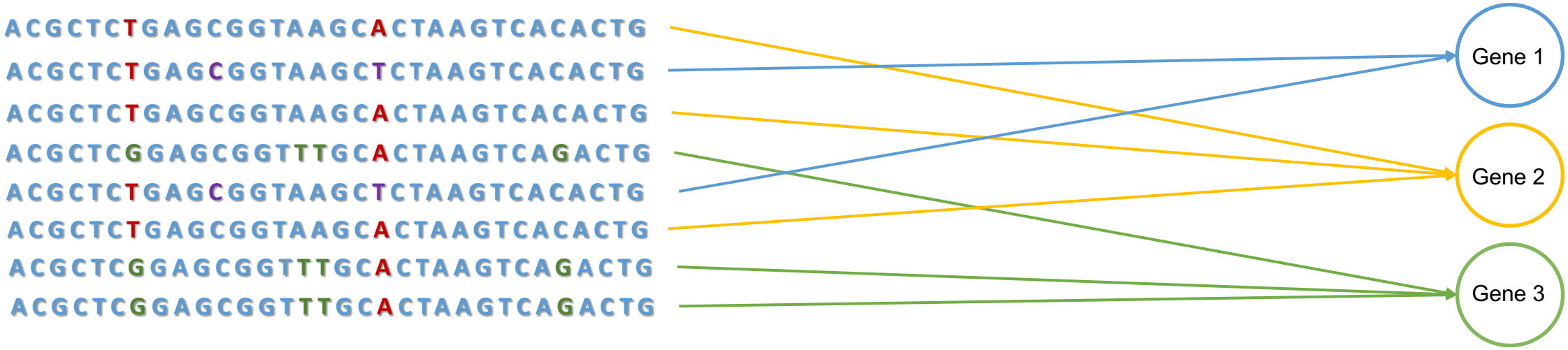
[…]

→ Genes are members of orthologous groups
→ Orthologous groups are members of modules
→ Modules are members of pathways

After: Tatusov, Koonin, Lipman, Science, 1997

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

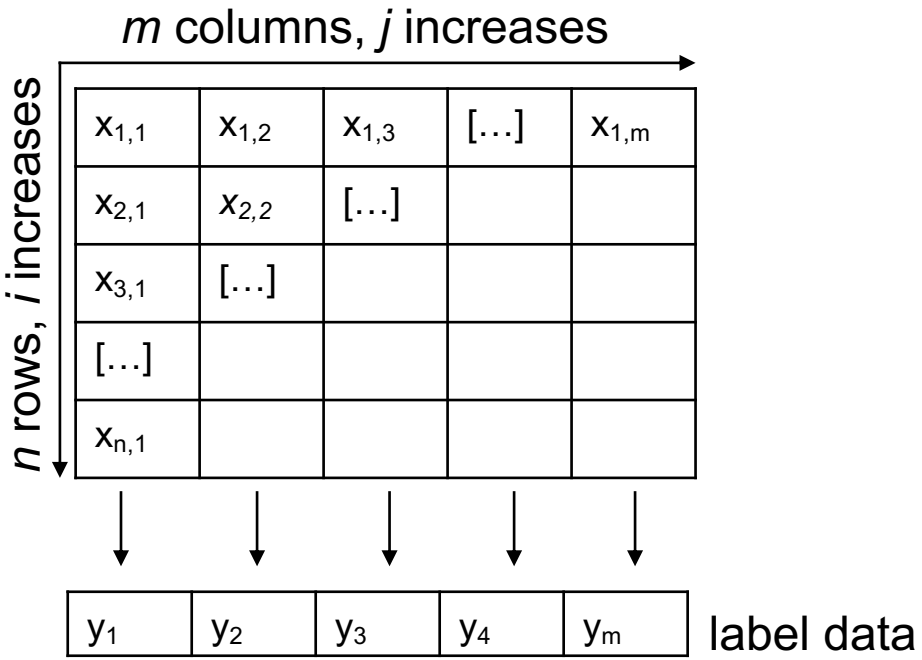
KO abundance tables can be summarized into Module abundance tables

Module abundance tables can be summarized into Pathway abundance tables

Generalization and notation

Matrix m x n

where element x*i,j* is in row *i* and column *j*, and max(*i*) = *n* and max(*j*) = *m*



- Feature data **x** (or observations, predictors):
  - *i*: rows → feature, *j*: columns → samples
  - $\mathbf{x}_i$ denotes the vector for the *i*-th feature
  - $\mathbf{x}_{ij}$ denotes *i*-th feature in *j*-th sample
- Features can be OTUs, genes, KOs, […]
- Label data **y** (or dependent variable, response)
  - vector of length *m*
- Example: labels for y are 1=healthy, 2=diseased

Label	binary	binary
y_1=healthy	1	h
y_2=healthy	1	h
y_3=diseased	2	d
y_4=healthy	1	h
[…]	[…]	[…]

→ **Enables differential abundance testing**

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