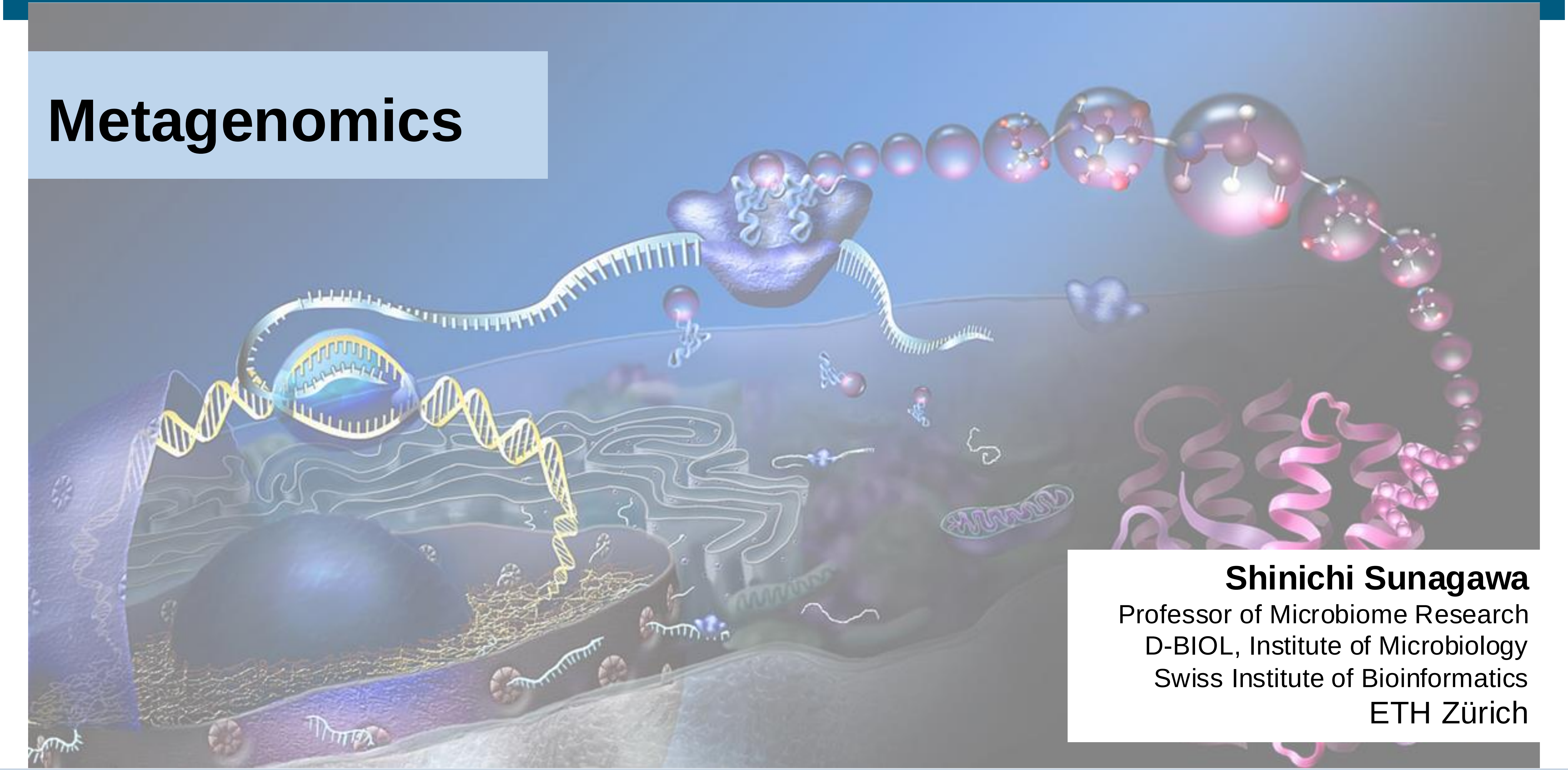
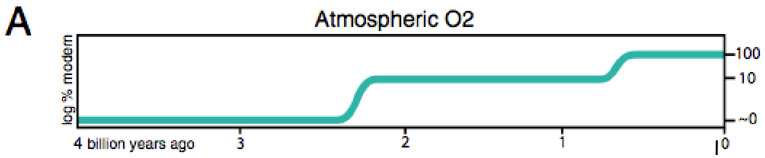


Metagenomics

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

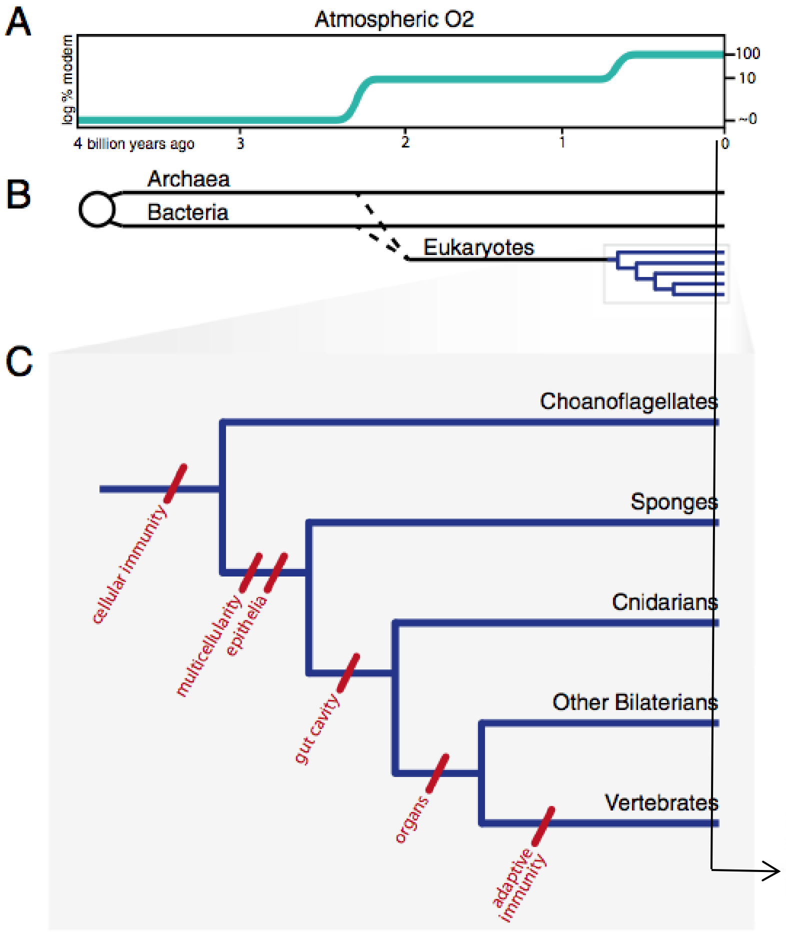
Significance (examples):

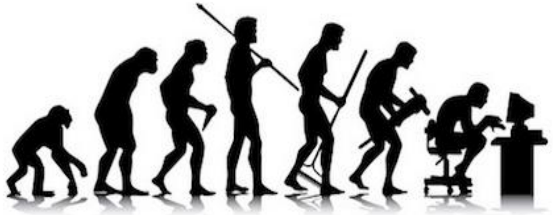
- biogeochemistry: e.g., photosynthesis by microbes, carbon fixation/export, nitrogen fixation
- health: help us digest food, provide essential vitamins, prime the immune system

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

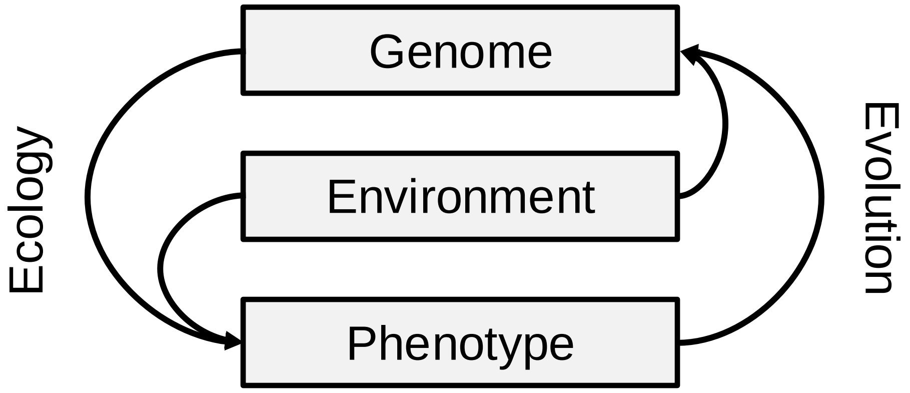
Microorganisms

- 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

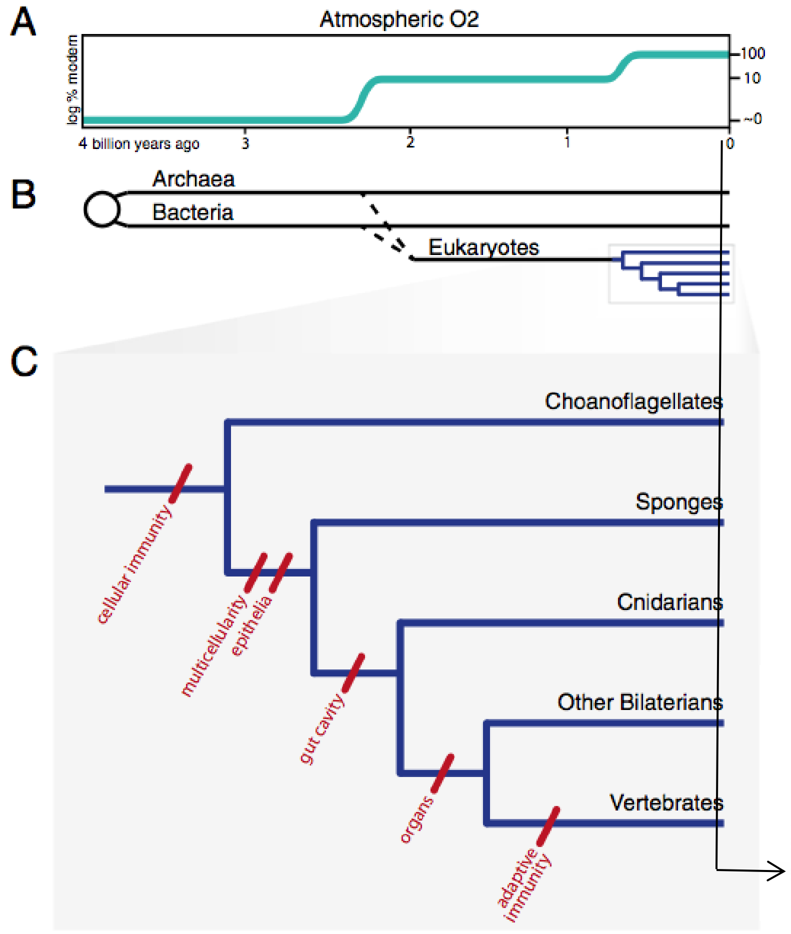
Single organism-centric view

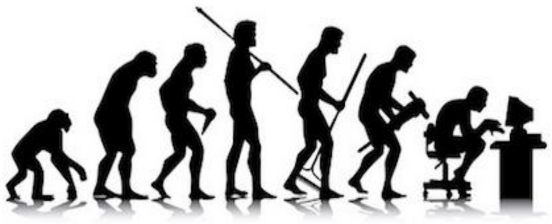


McFall-Ngai et al., PNAS, 2013; Venn et al., Science, 2014

Evolution and significance of microbiomes

From the origin of life to today





13 mio years

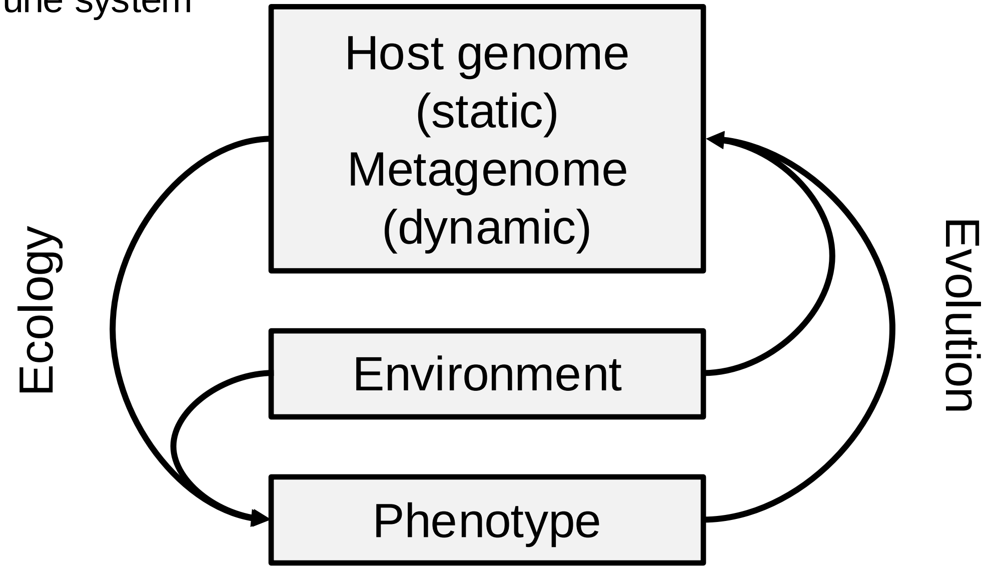
<u>Microorganisms</u>

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

<u>Significance (examples):</u>

- 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



Host genome (static)
Metagenome (dynamic)

Environment

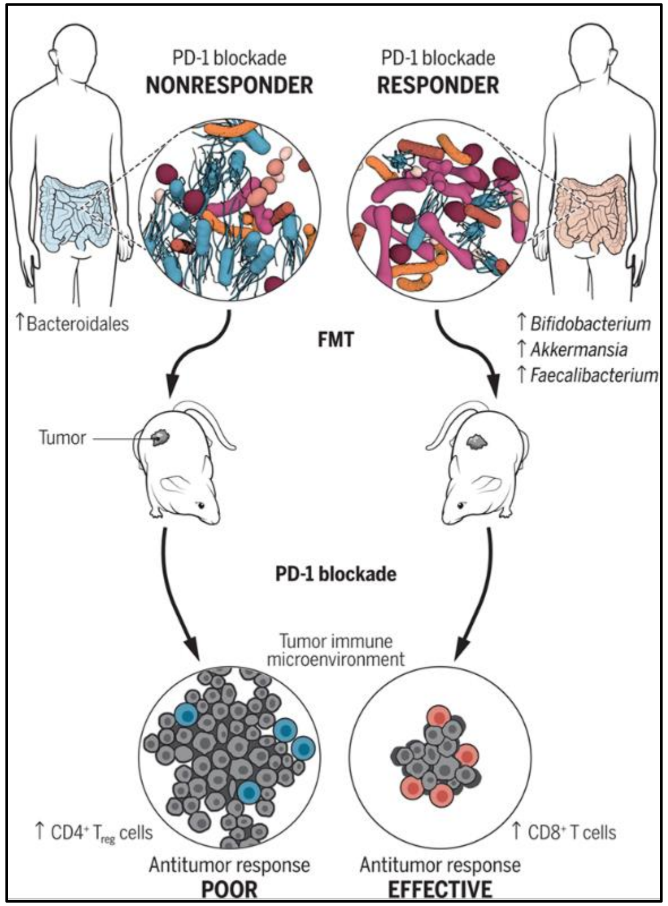
Phenotype

Ecology

Evolution

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

Describing microbial communities – Example 1



GRAPHIC: V. ALTOUNIAN/*SCIENCE*

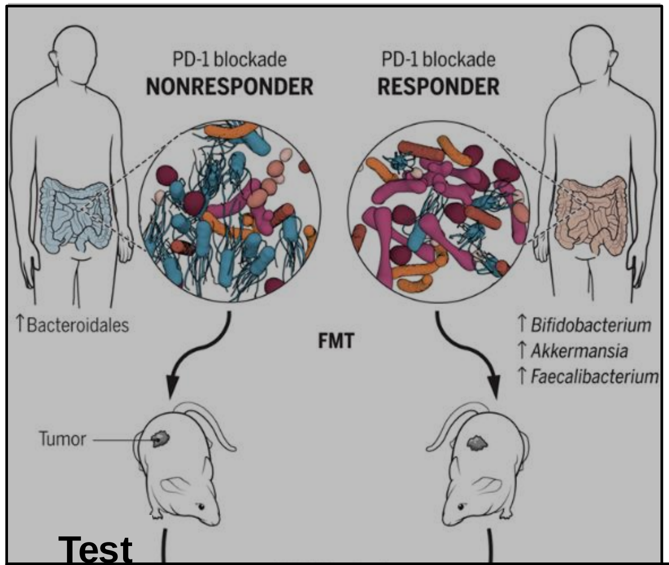
Gut microbial community compositions

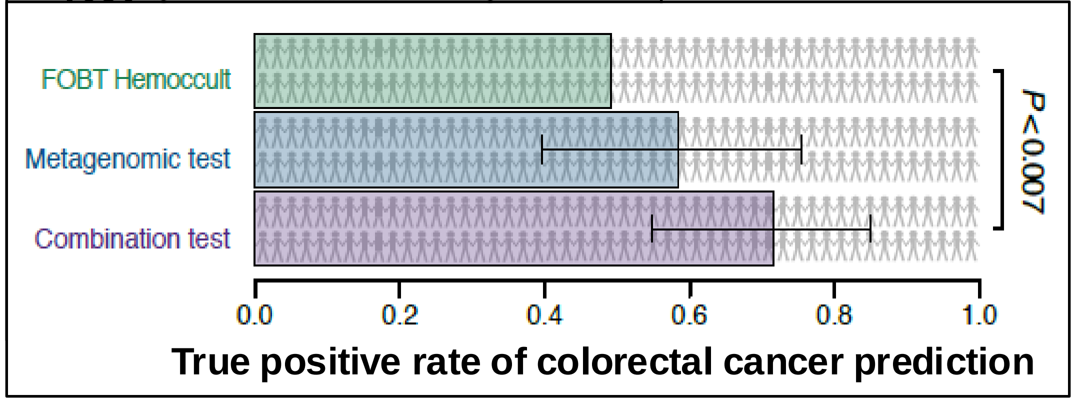
- can alter efficacy of treatments

→ 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

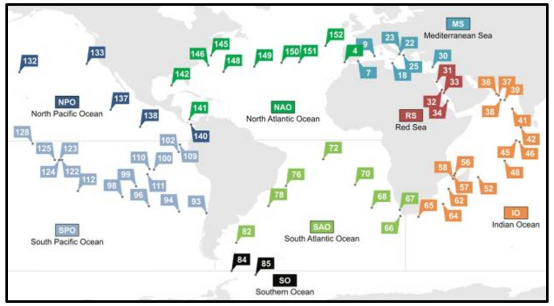
Routy et al., Gopalakrishnan et al., and Matson et al., Science 2018

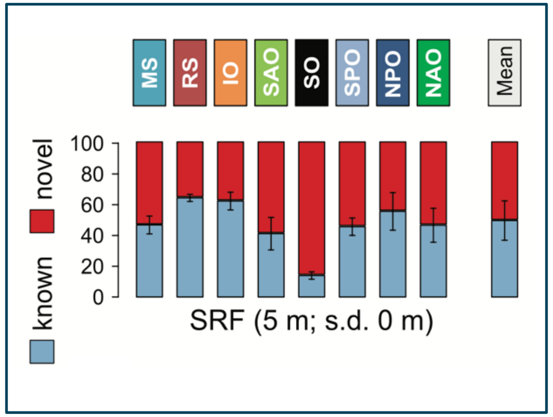
- can be indicative for diseases

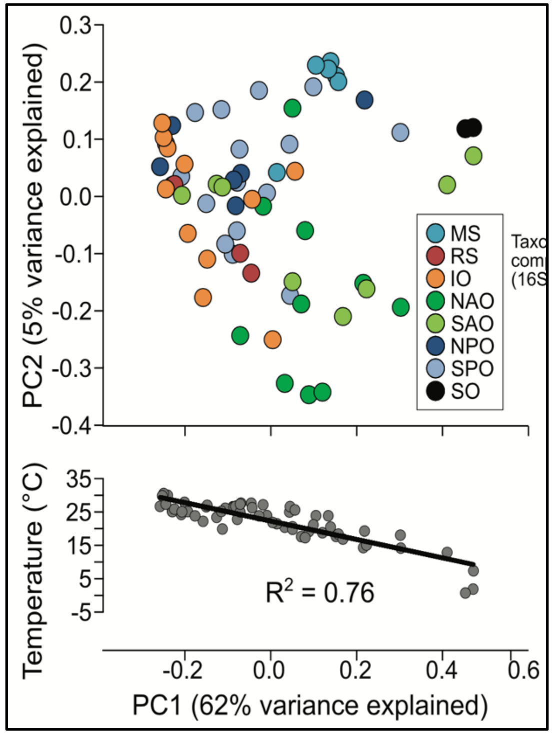
→ Statistical models of fecal microbiota composition can predict colorectal cancer

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

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

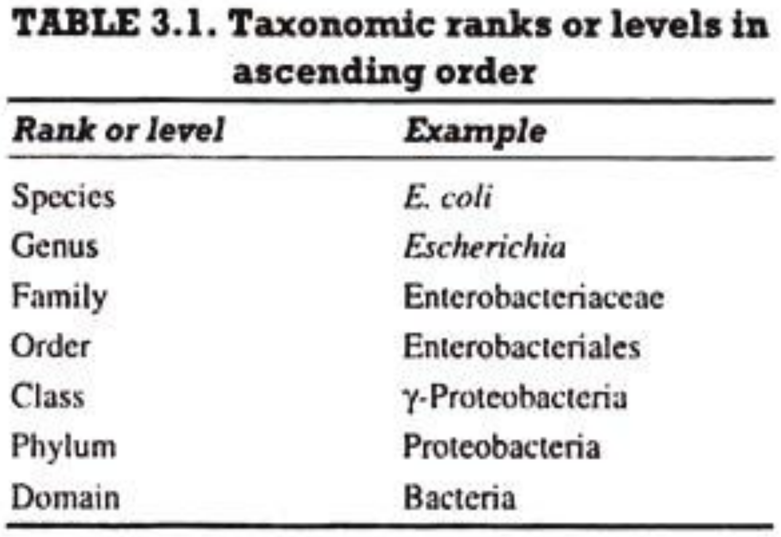


TABLE 3.1. Taxonomic ranks or levels in ascending order

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

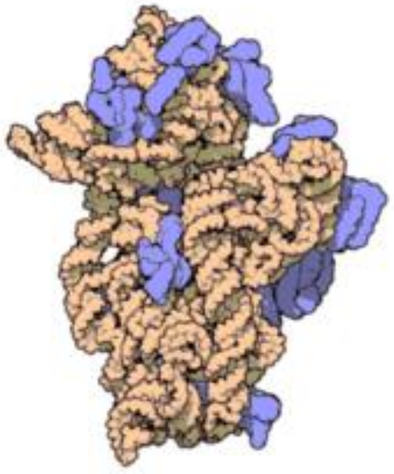
- Phenotypic characteristics
  - morphology, physiology/metabolism, ecology, exchange of genetic material
- Molecular characteristics
  - DNA-DNA hybridization
  - **DNA sequences of individual genes (e.g., 16S rRNA gene) or complete genomes**

→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

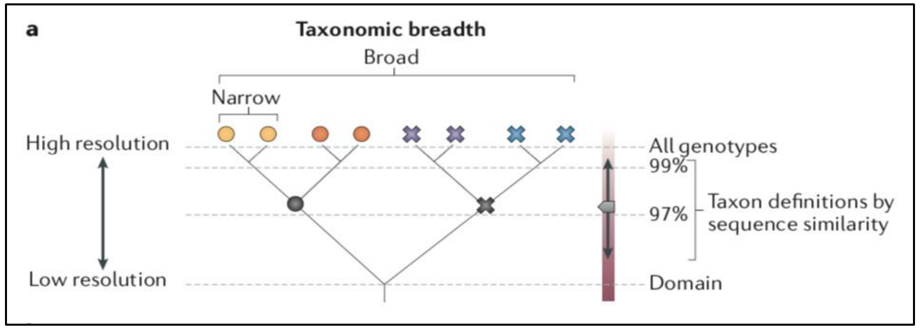


Hanson et al., Nat Rev Microbiol, 2012

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

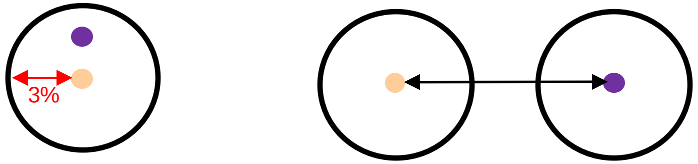
- Owing to lack of prokaryotic species definition, 97% sequence similarity is often used to define ‘species’-like: “**Operational Taxonomic Units” (OTUs)**



Identity of 16S rRNA gene sequences

>=97%

<97%

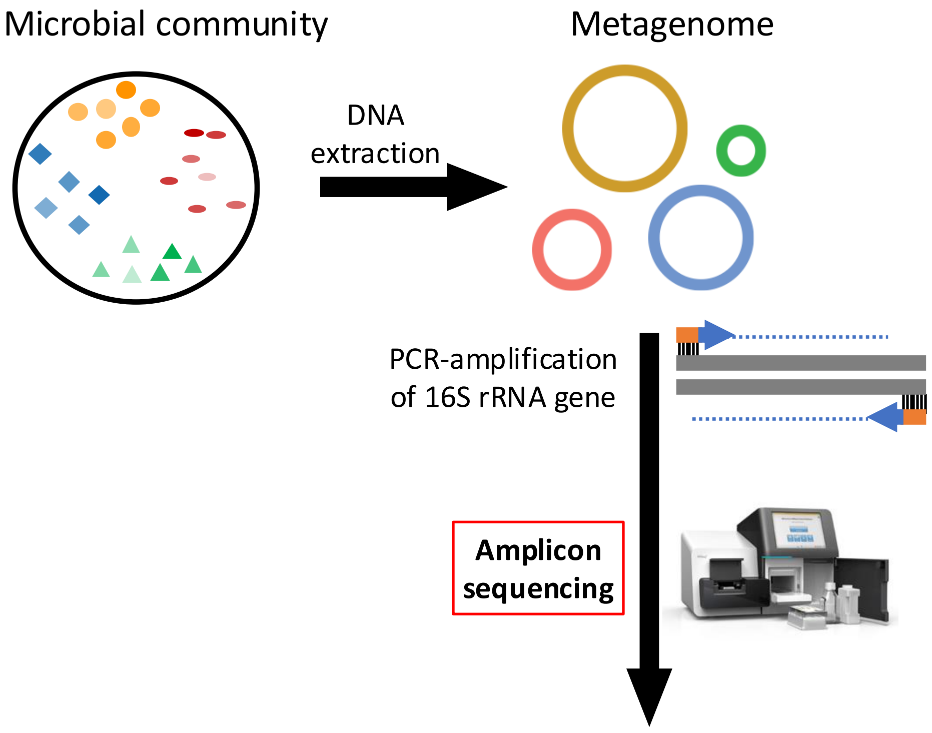


→ 1 OTU

→ 2 OTUs

Hanson et al., Nat Rev Microbiol, 2012

Amplification of 16S rRNA gene fragments by PCR



Who is there?

16S rRNA amplicons

ACGCTCTGAGCGGTAAGCACTAAGTCACACTG

ACGCTCTGAGCGGTAAGCTCTAAGTCACACTG

ACGCTCTGAGCGGTAAGCACTAAGTCACACTG

ACGCTCGGAGCGGTTTGCACTAAGTCAGACTG

ACGCTCTGAGCGGTAAGCTCTAAGTCACACTG

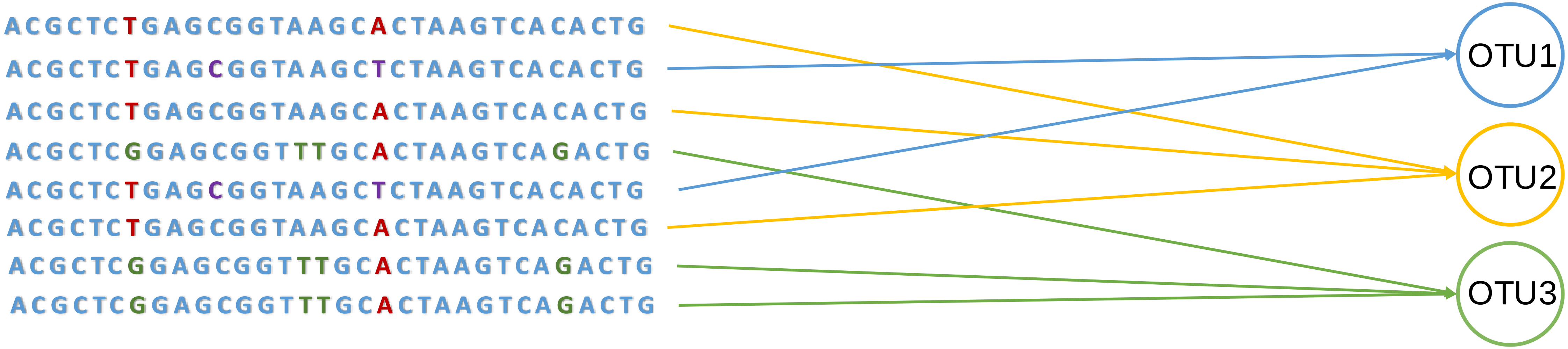
ACGCTCTGAGCGGTAAGCACTAAGTCACACTG

ACGCTCGGAGCGGTTTGCACTAAGTCAGACTG

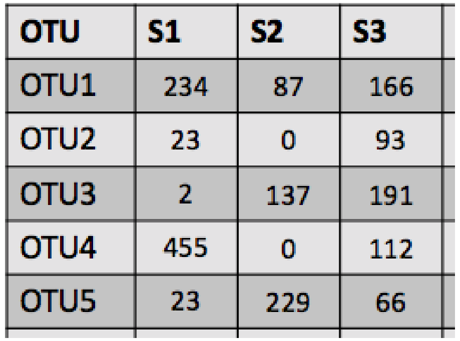
ACGCTCGGAGCGGTTTGCACTAAGTCAGACTG

Quantification of OTU abundances

All amplicons are aligned to best matching OTU and counted


ACGCTCTGAGCGGTAAGCACTAAGTCACACTG
ACGCTCTGAGCGGTAAGCTCTAAGTCACACTG
ACGCTCTGAGCGGTAAGCACTAAGTCACACTG
ACGCTCGGAGCGGTTTGCACTAAGTCAGACTG
ACGCTCTGAGCGGTAAGCTCTAAGTCACACTG
ACGCTCTGAGCGGTAAGCACTAAGTCACACTG
ACGCTCGGAGCGGTTTGCACTAAGTCAGACTG
ACGCTCGGAGCGGTTTGCACTAAGTCAGACTG
OTU1
OTU2
OTU3

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

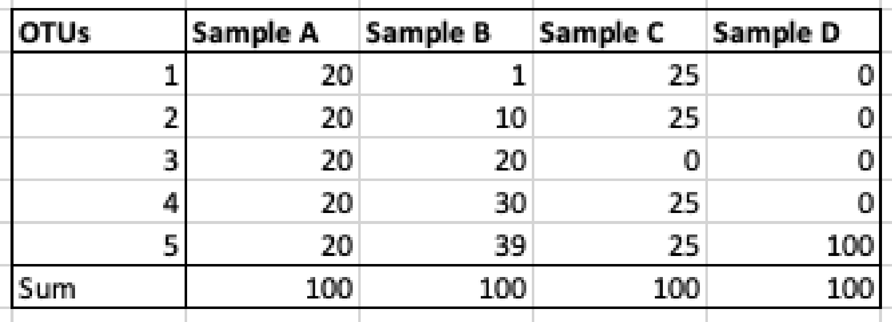


OTU	S1	S2	S3
OTU1	234	87	166
OTU2	23	0	93
OTU3	2	137	191
OTU4	455	0	112
OTU5	23	229	66

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

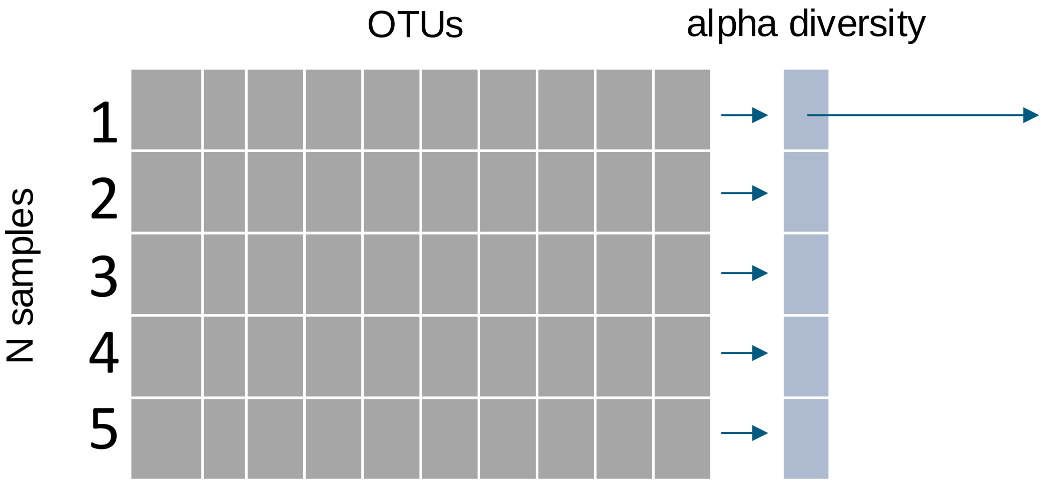
$$H'_{max} = -\sum_{i=1}^{R} \frac{1}{R} ln \frac{1}{R} = \ln R$$

that is, every species is equally likely

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

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

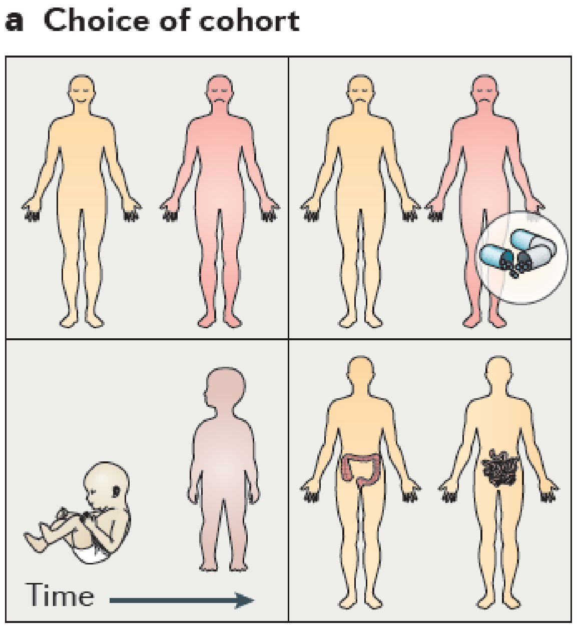
Differences between microbial communities

- 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:

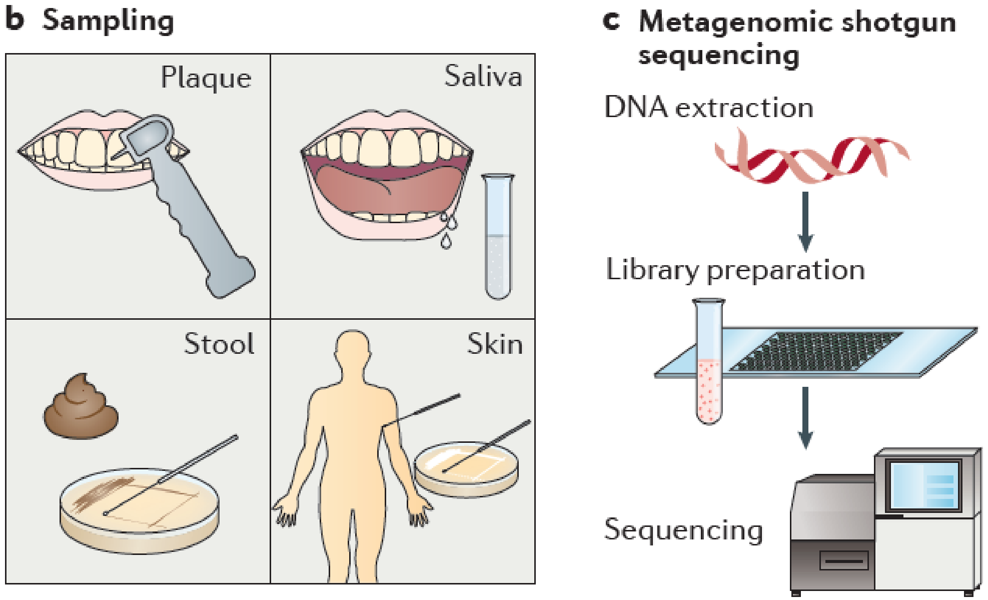
- groups of individuals and/or health states
- differential response to drugs (or nutrition)
- organismal development (or disease progression)
- differences between body sites

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

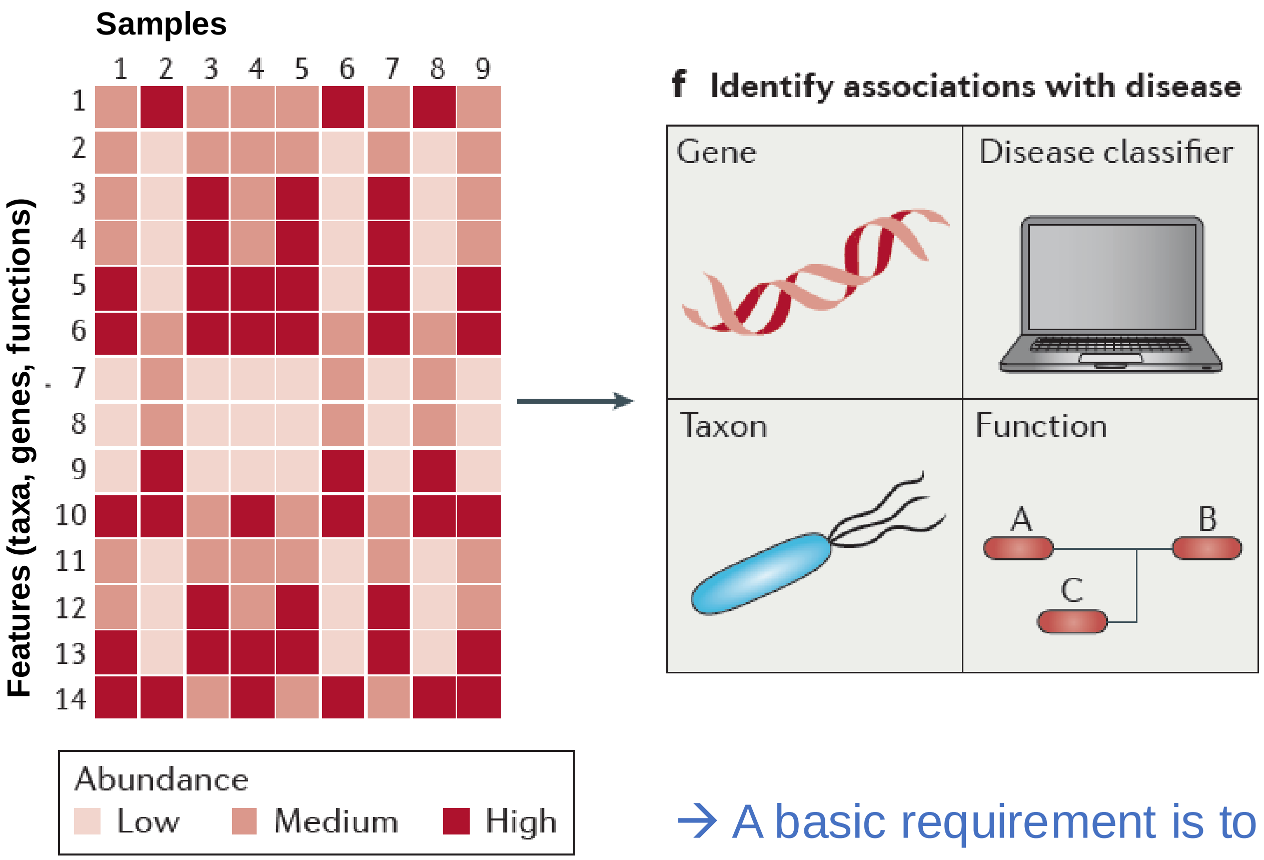
Microbiome-wide association studies are analogous to GWAS



- 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”)

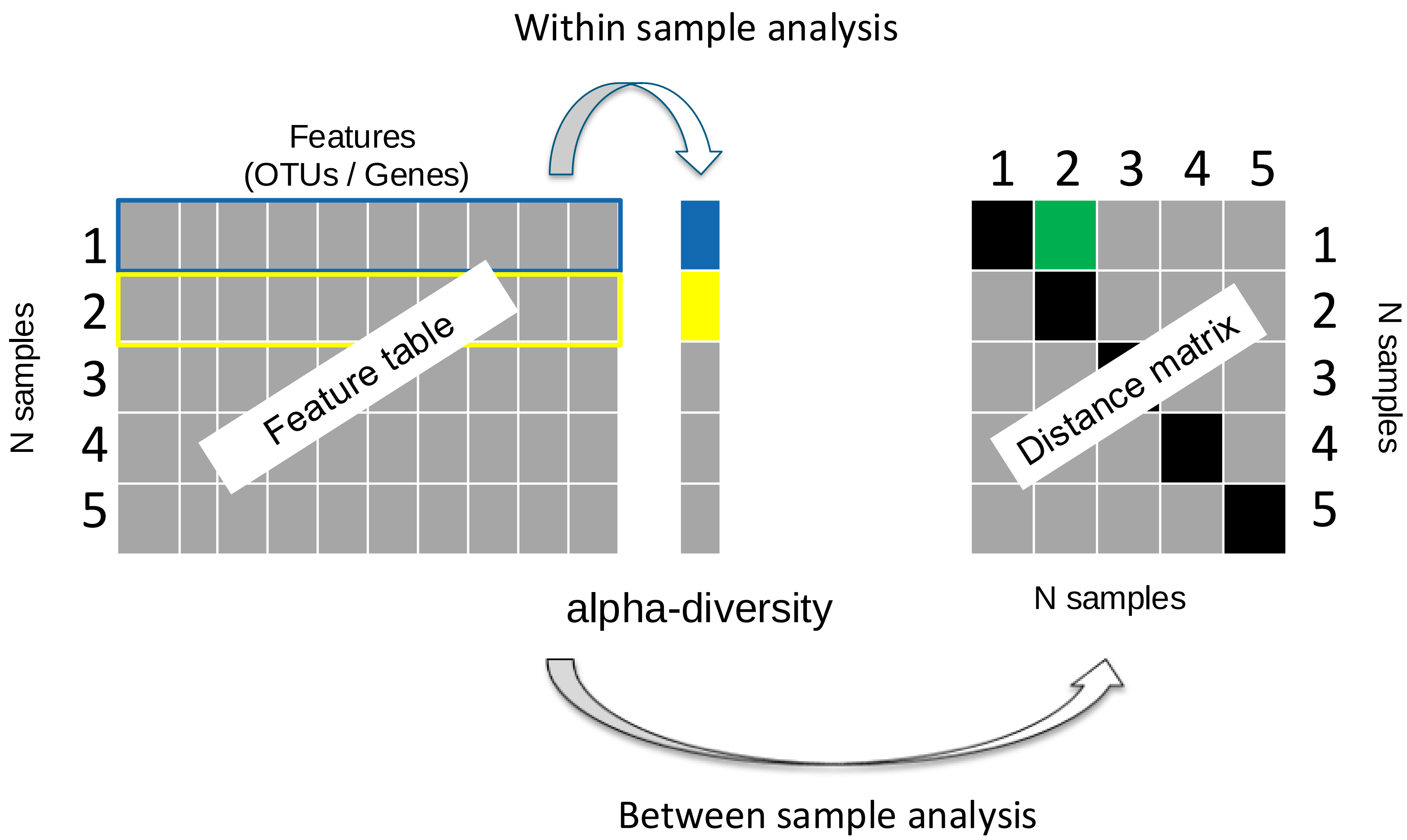
- Abundance tables are analyzed to determine differentially abundant features, e.g., between groups of samples, to identify biomarkers

- Machine learning is used to classify samples and/or to identify relationships between the microbiome and clinical/environmental phenotypes

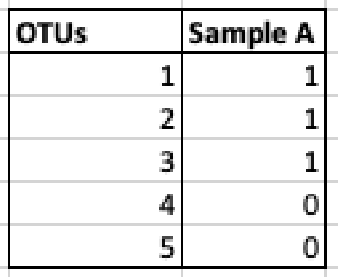
→ A basic requirement is to quantify the differences between samples

Wang and Jia, NRM, 2016

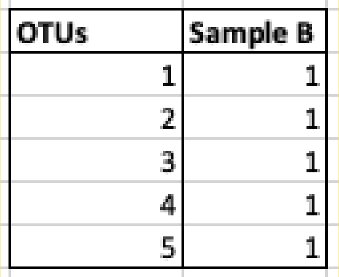
How many of which OTUs/genes are found in a sample?
How similar are the OTUs/gene compositions between samples?



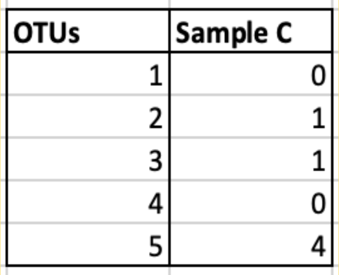
In-class task 2: beta diversity



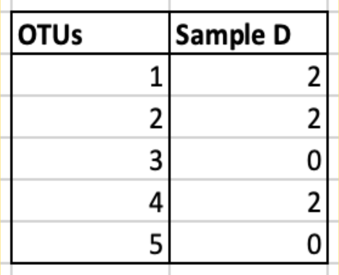
OTUs	Sample A
1	1
2	1
3	1
4	0
5	0



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

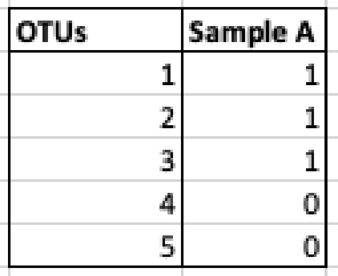


OTUs	Sample D
1	2
2	2
3	0
4	2
5	0

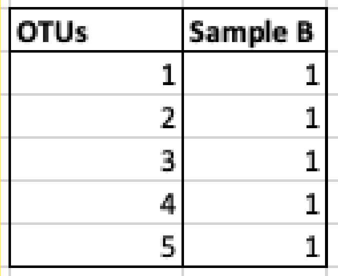
→ In pairs, please discuss how pairwise similarities of samples A, B, C, and D could be quantified?

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

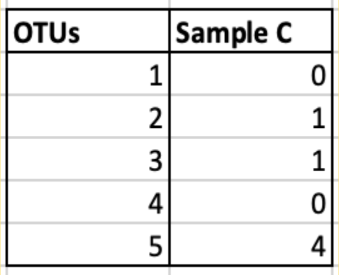
In-class task 2: beta diversity



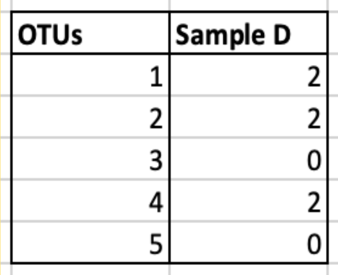
OTUs	Sample A
1	1
2	1
3	1
4	0
5	0



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

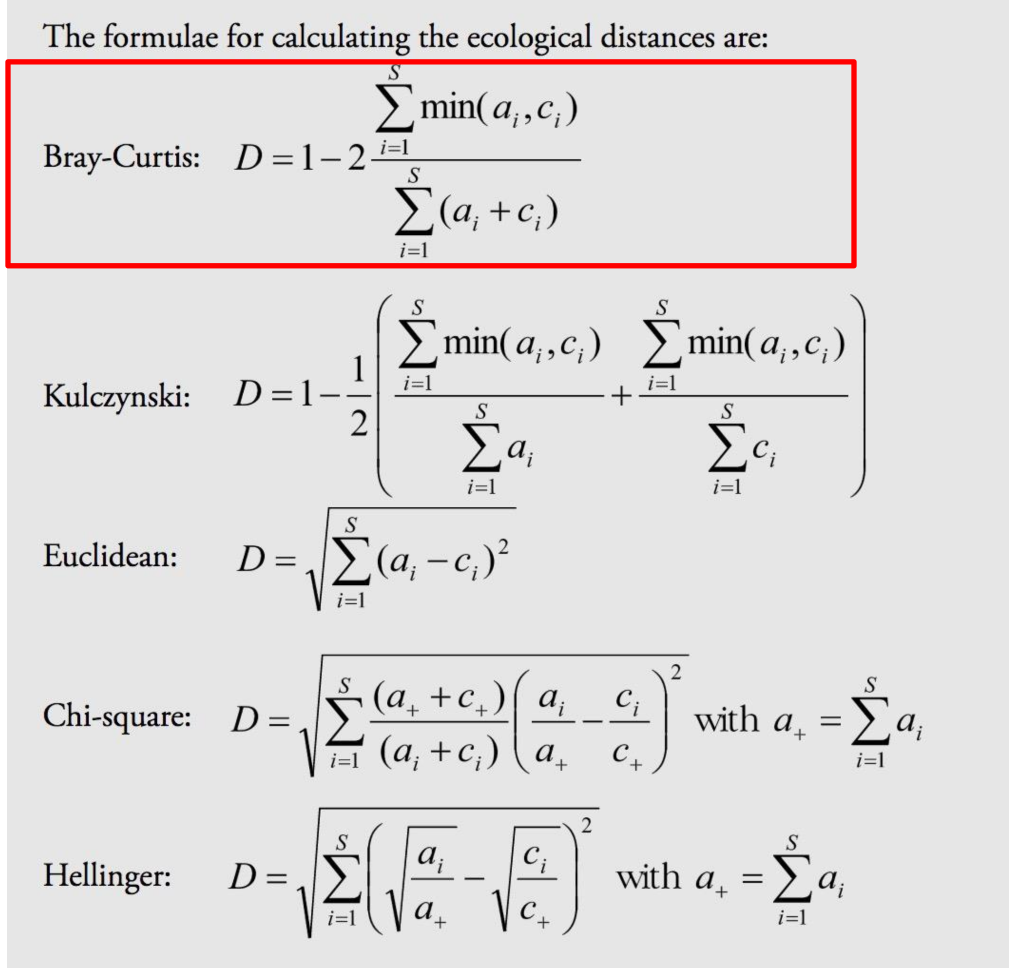
b) Differences in the abundance of OTUs shared between samples are not accounted for

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

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

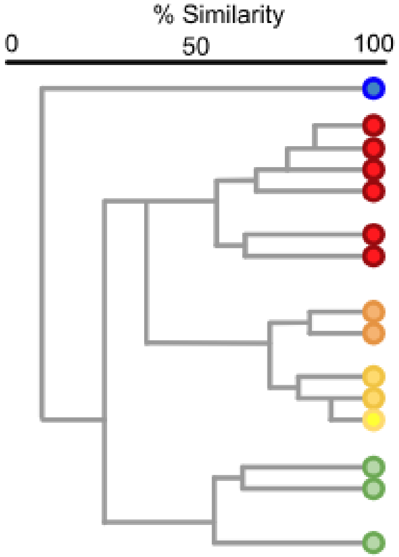
a_i = abundance of taxon *i* in sample *a*, and
c_i = abundance of taxon *i* in sample *c*

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.

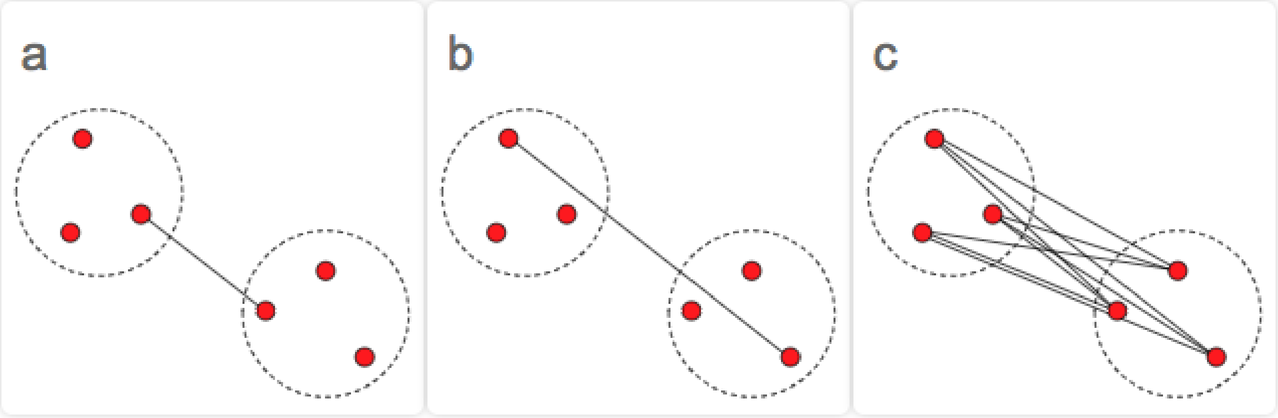
Visualize dissimilarities between microbial communities

- 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

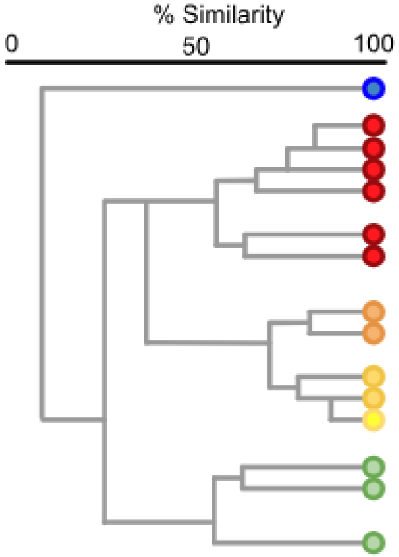


single complete average

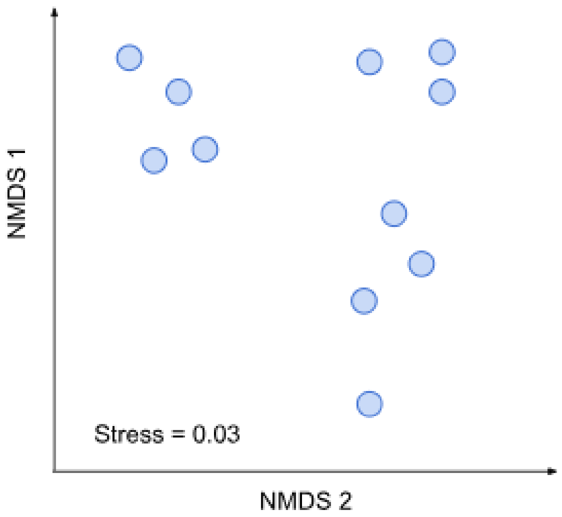
Visualize dissimilarities between microbial communities

- 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

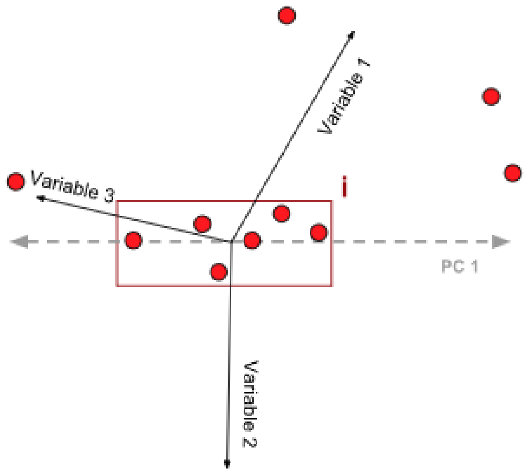
Hierarchical clustering



Non-metric dimensional scaling (NMDS)



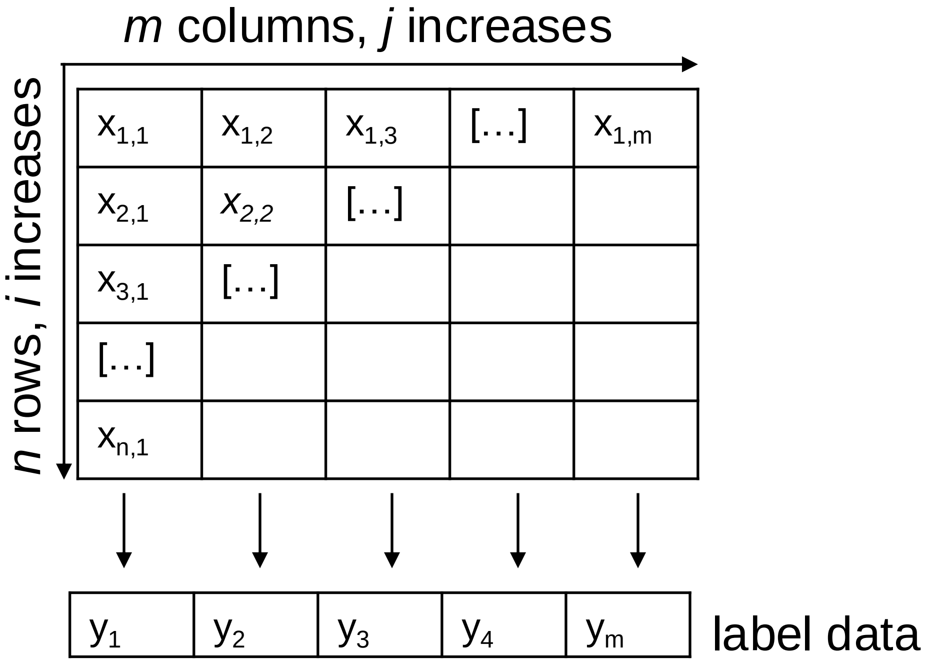
Principal component or coordinate analysis (PCA or PCoA)



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*



m columns, *j* increases

n rows, *i* increases

$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	[…]	$x_{1,m}$
$x_{2,1}$	$x_{2,2}$	[…]		
$x_{3,1}$	[…]			
[…]				
$x_{n,1}$				

y_1	y_2	y_3	y_4	y_m

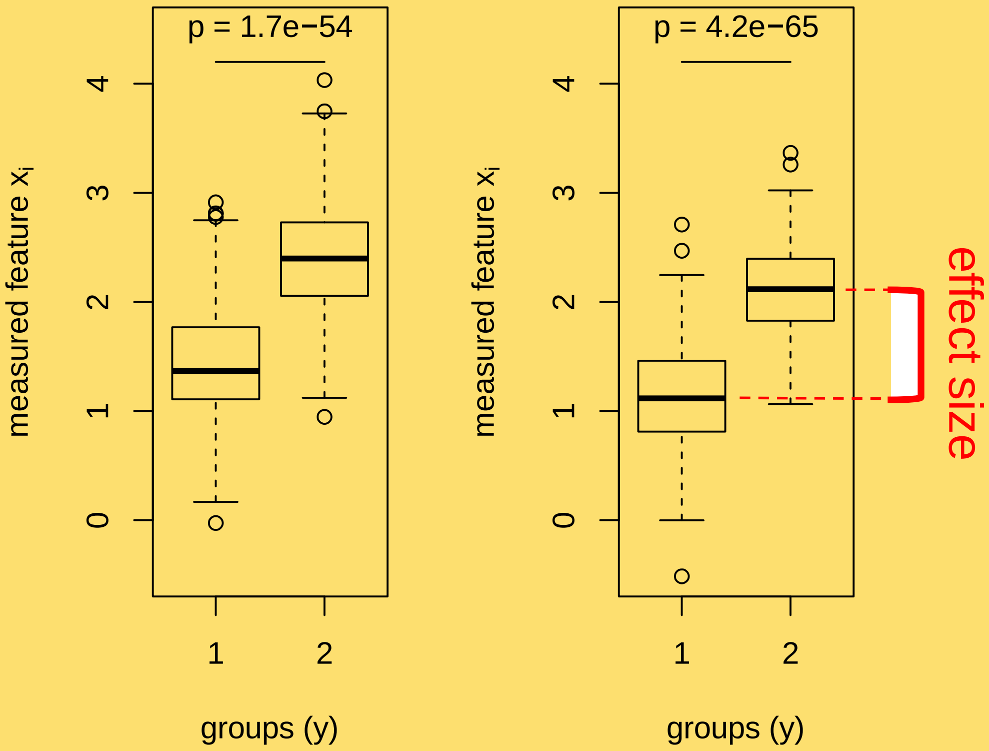
label data

- 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

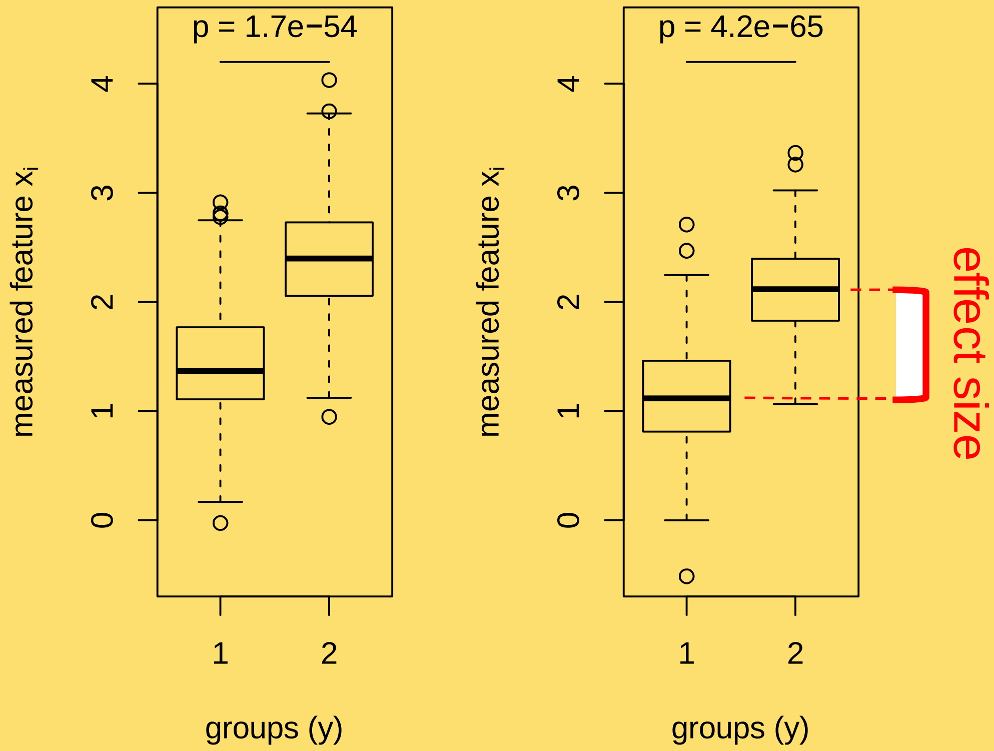
[✓ b) The p-value associated with these changes to decrease]

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)

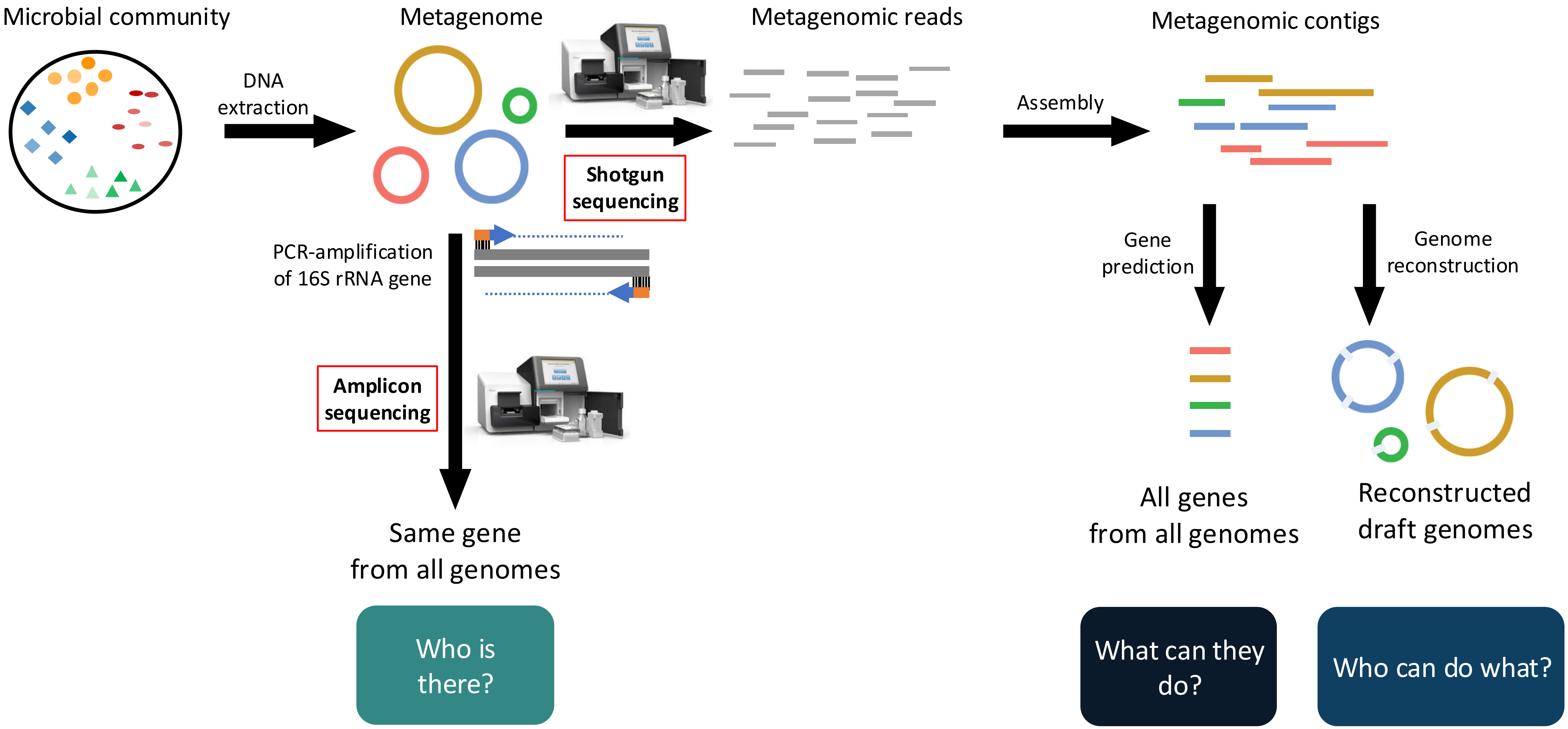
Working with microbial community genes and genomes

- 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

Sequencing microbial community DNA



Microbial community

DNA extraction

Metagenome

Shotgun sequencing

Metagenomic reads

Assembly

Metagenomic contigs

PCR-amplification of 16S rRNA gene

Amplicon sequencing

Same gene from all genomes

Who is there?

Gene prediction

Genome reconstruction

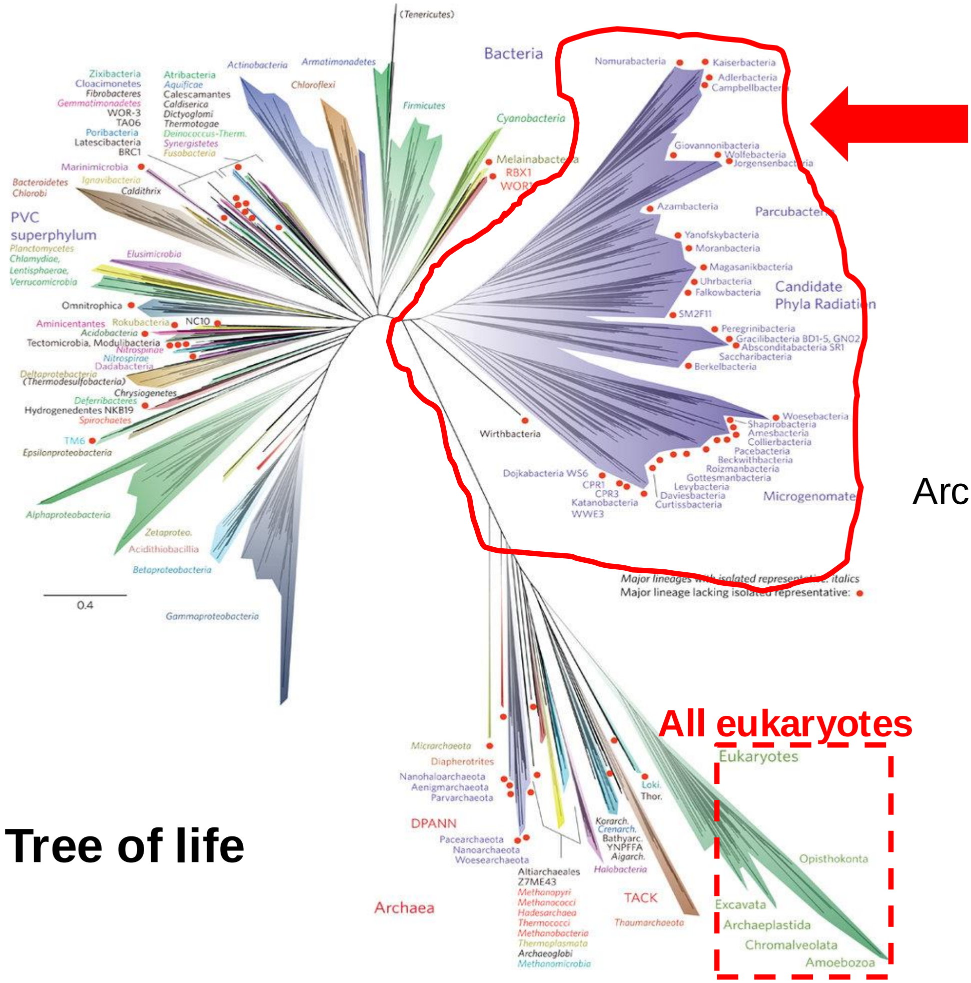
All genes from all genomes

Reconstructed draft genomes

What can they do?

Who can do what?

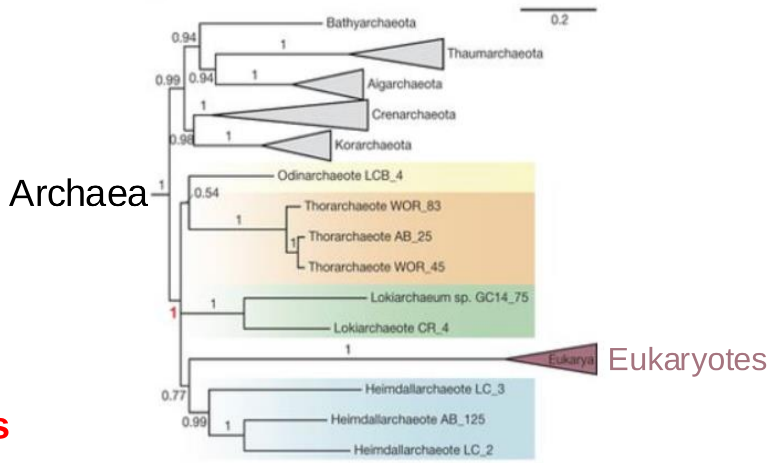
Insights by reconstructing microbial community genomes



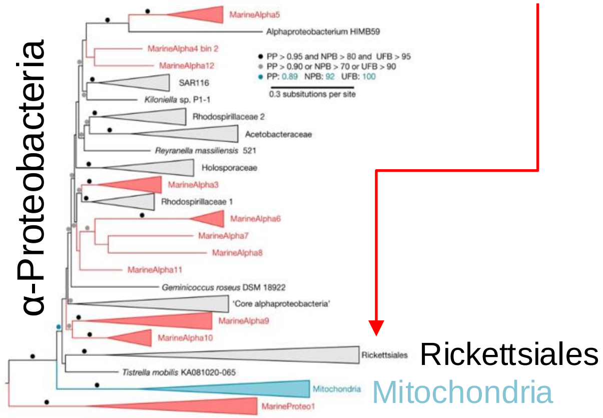
Candidate phyla radiation discovered by metagenomics

Tree of life

Two domains of life?

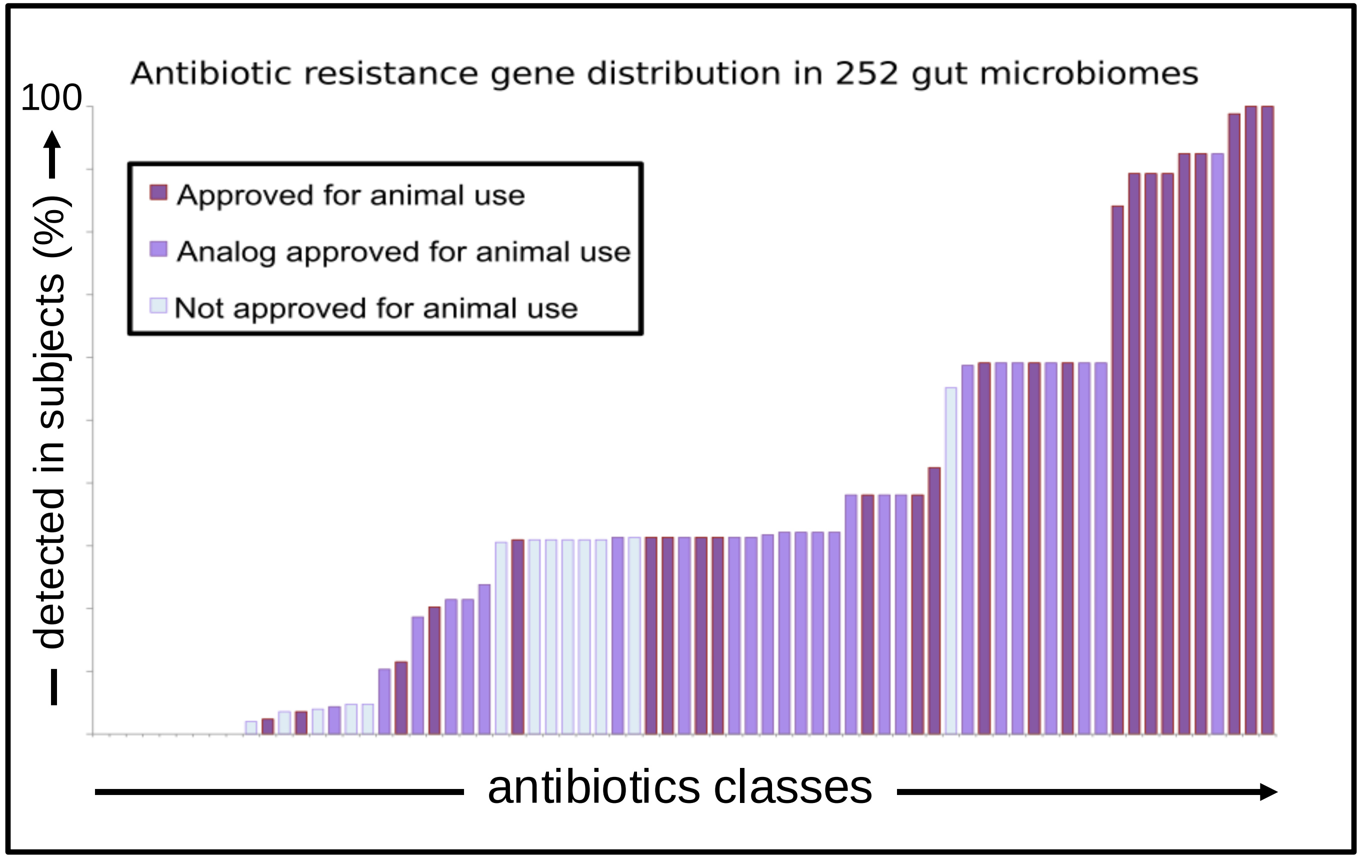


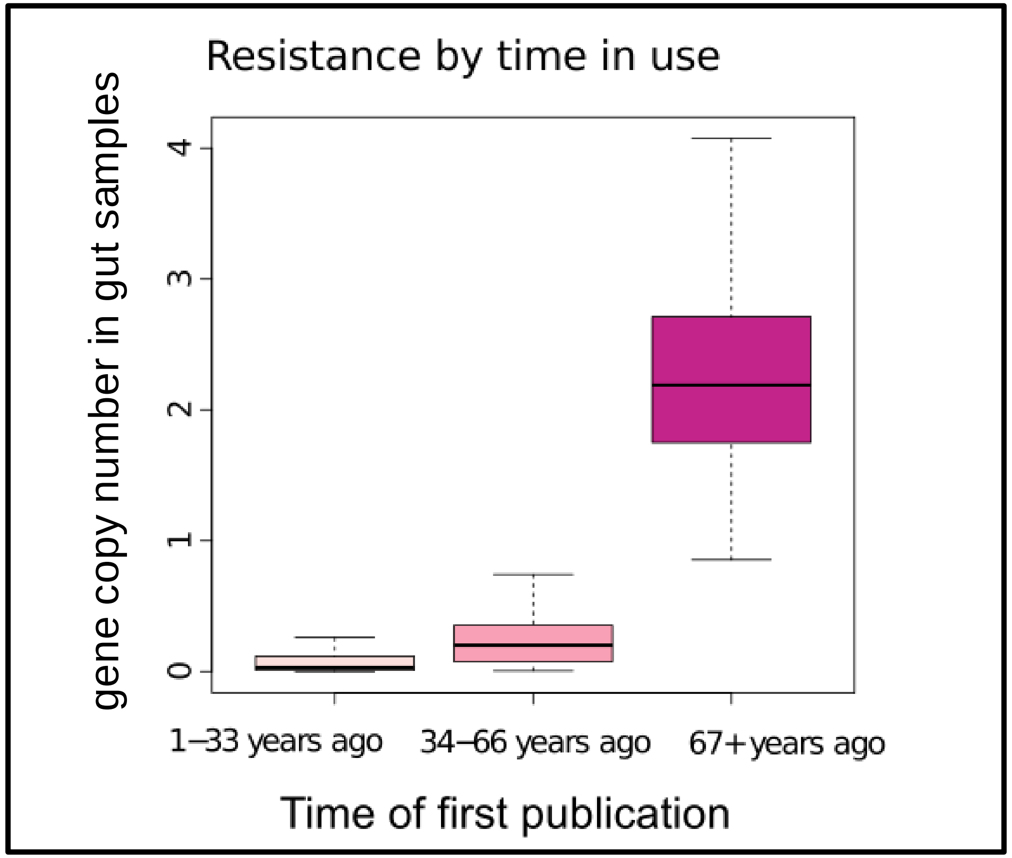
Mitochondrial origin not within Rickettsiales?



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

Insights by quantifying microbial gene abundances





Forslund, et al., Genome Research 2013, Bioassays 2014

Sequencing of microbial isolate genomes

- First bacterial genome (1995): *Haemophilus influenzae*

Fleischmann et al. 1995

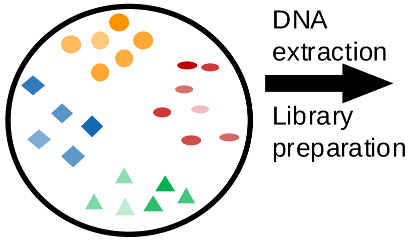
- Followed by many isolated pathogens of diseases (e.g., plague, anthrax, tuberculosis, Lyme disease)
- Many isolates of important non-pathogenic species: e.g., *Prochlorococcus*, *Lactobacillus*, *Bradyrhizobium*

- 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)

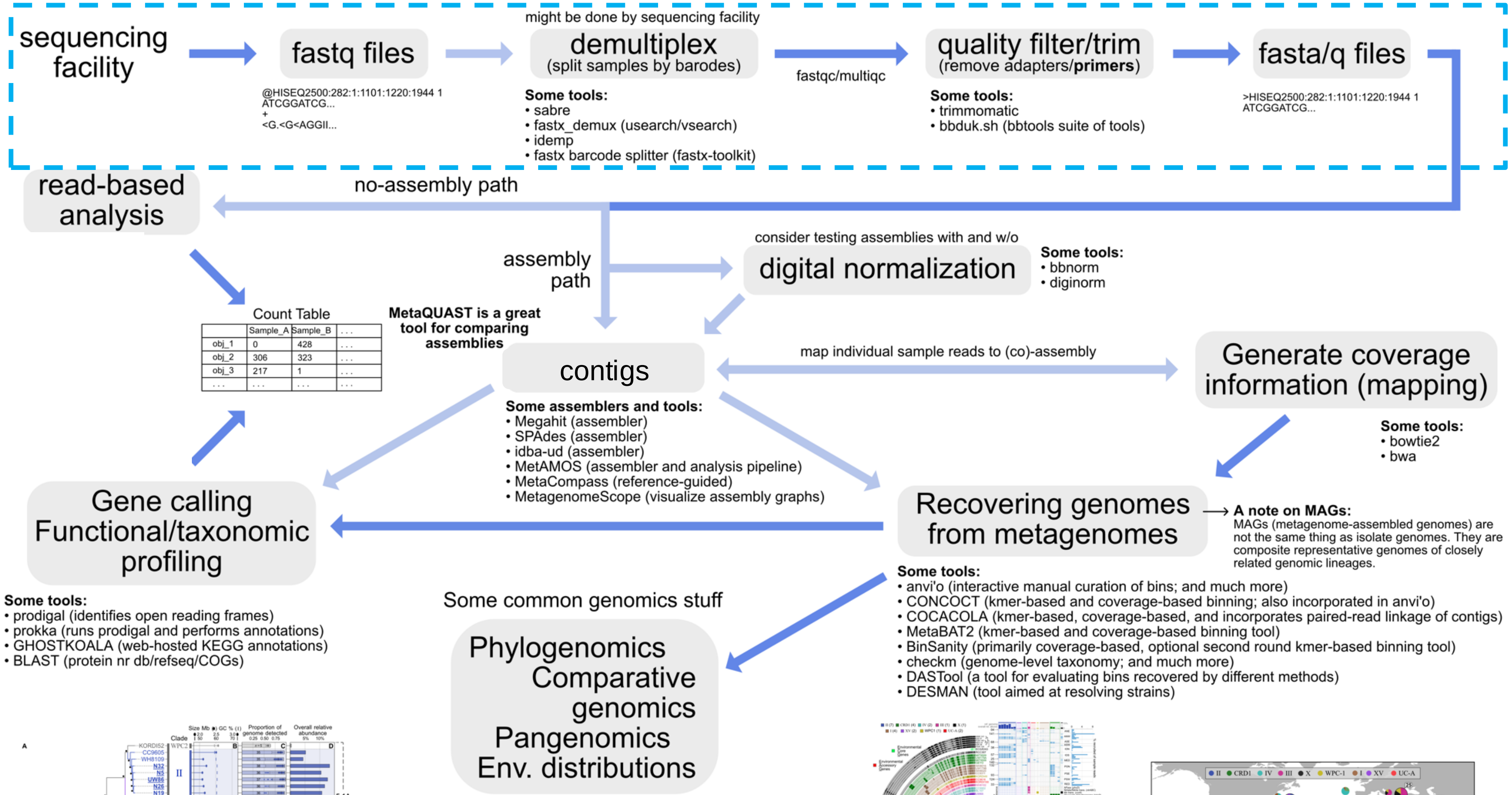
Overview of generic* metagenomics workflow

*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.



Microbial community


sequencing facility
fastq files
might be done by sequencing facility
demultiplex
(split samples by barodes)
fastqc/multiqc
quality filter/trim
(remove adapters/primers)
fasta/q files
read-based analysis
no-assembly path
assembly path
consider testing assemblies with and w/o
digital normalization
contigs
map individual sample reads to (co)-assembly
Generate coverage information (mapping)
Gene calling
Functional/taxonomic profiling
Recovering genomes from metagenomes
Some common genomics stuff
Phylogenomics
Comparative genomics
Pangenomics
Env. distributions

```
@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)

```
>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
```

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

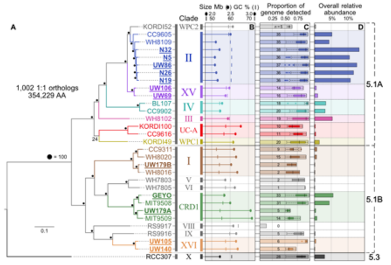
Some assemblers and tools:
- Megahit (assembler)
- SPAdes (assembler)
- idba-ud (assembler)
- MetAMOS (assembler and analysis pipeline)
- MetaCompass (reference-guided)
- MetagenomeScope (visualize assembly graphs)

Some tools:
- bowtie2
- bwa

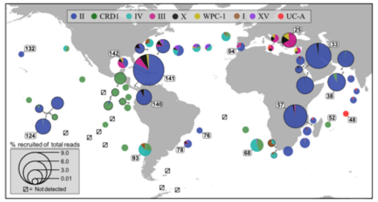
→ **A note on MAGs:**
MAGs (metagenome-assembled genomes) are not the same thing as isolate genomes. They are composite representative genomes of closely related genomic lineages.

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

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 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)

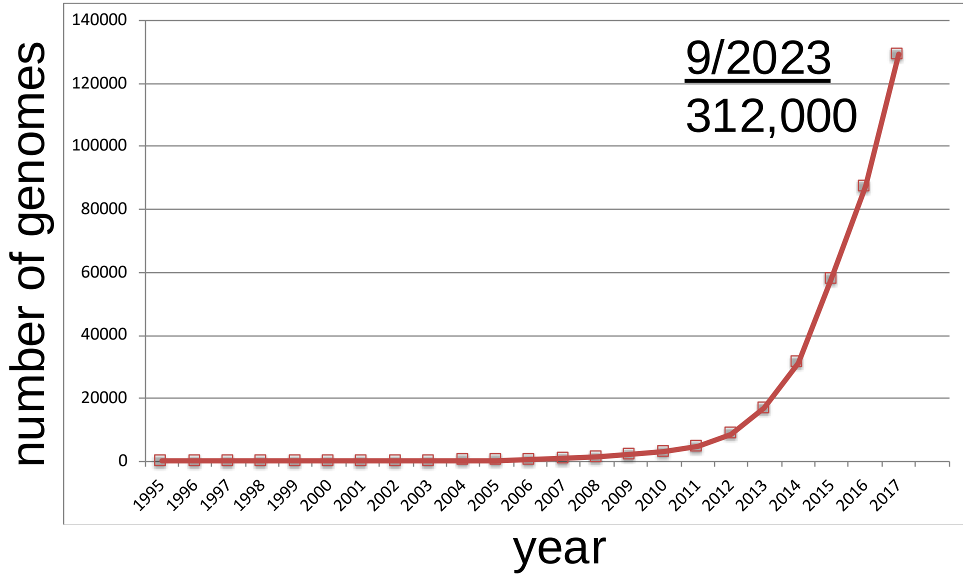


astrobiomike.github.io

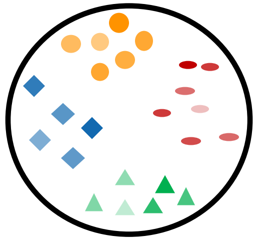
Modified from astrobiomike.github.io

Background: added value of metagenomics

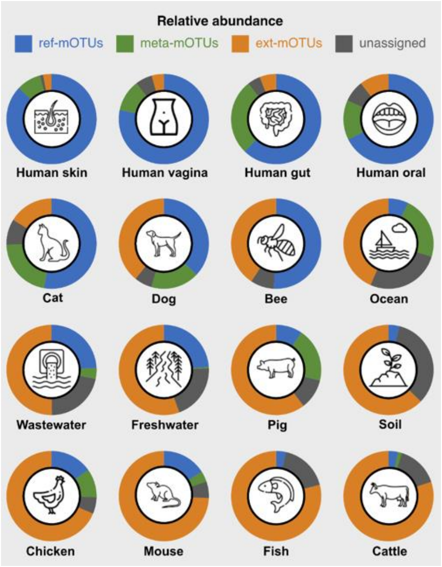
Microbial isolate genome sequences



9/2023
312,000

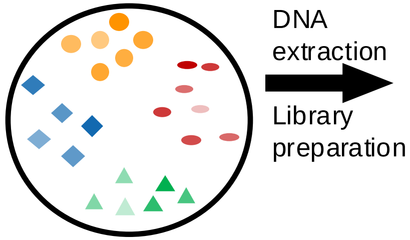


Microbial community



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

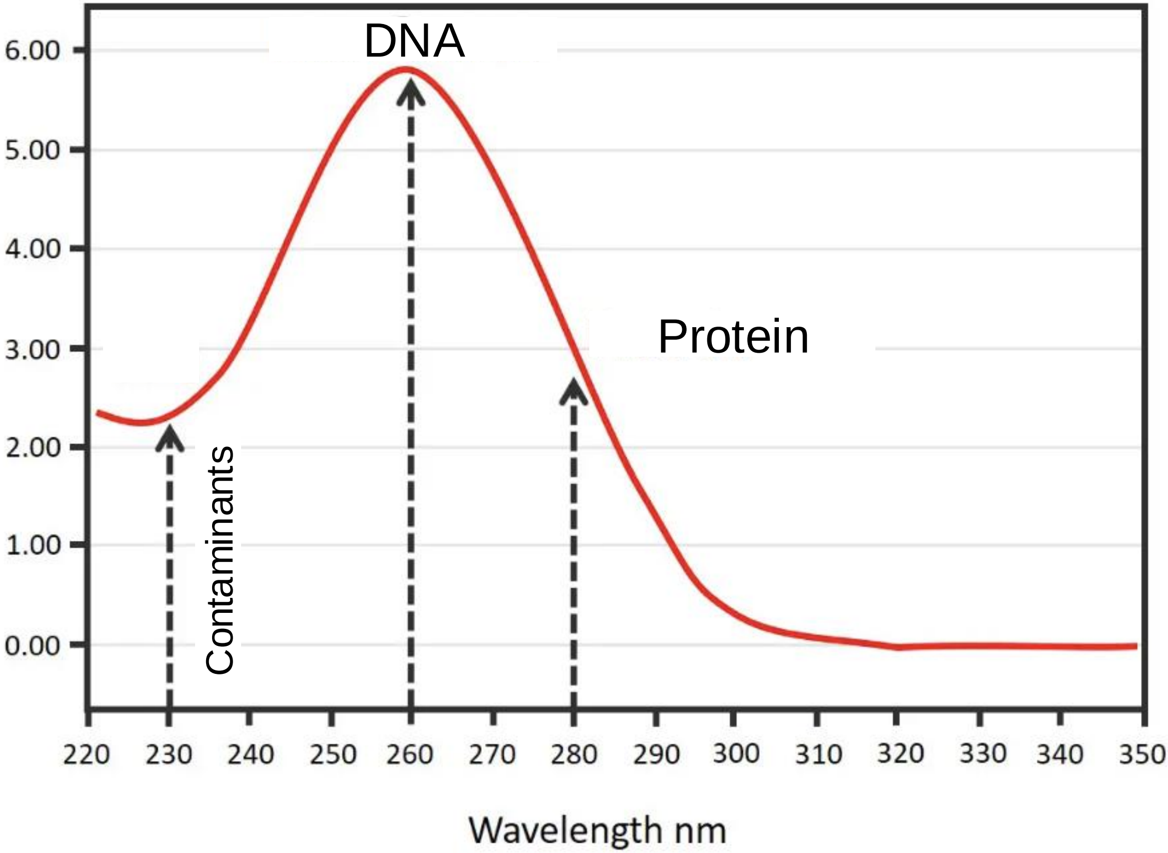
Metagenomics provides access, in principle, to all genomic information within a microbial community. This allows us to ask: “what can they do?”, in addition to: “who is there?”.



Microbial community

DNA extraction

- Sufficiently high quality and quantity needed



Contaminants:

- e.g., phenol, carbohydrates, EDTA

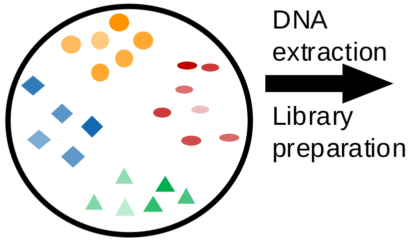
Protein:

- tyrosine and tryptophan

→ DNA quality:

- 260/280 ratio
- 260/230 ratio

Modified after astrobiomike.github.io

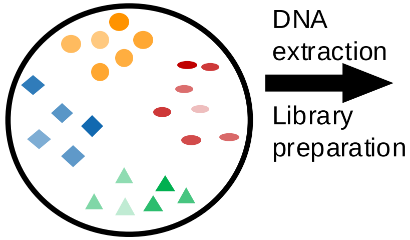


Microbial community

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

Modified after astrobiomike.github.io; https://teichlab.github.io/scg_lib_structs/methods_html/Illumina.html

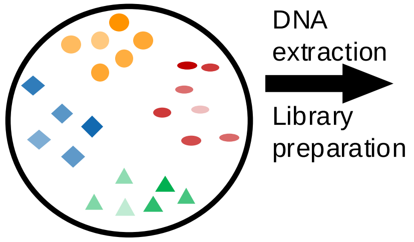


Microbial community

DNA extraction / library preparation

- Sufficiently high quality and <u>quantity</u> needed
- Extracted DNA is sheared into smaller fragments (inserts)
  - Illumina: ~300-600 bp; PacBio: ~20 kbp; ONT: no limit

Modified after astrobiomike.github.io; https://teichlab.github.io/scg_lib_structs/methods_html/Illumina.html



Microbial community

DNA extraction / library preparation

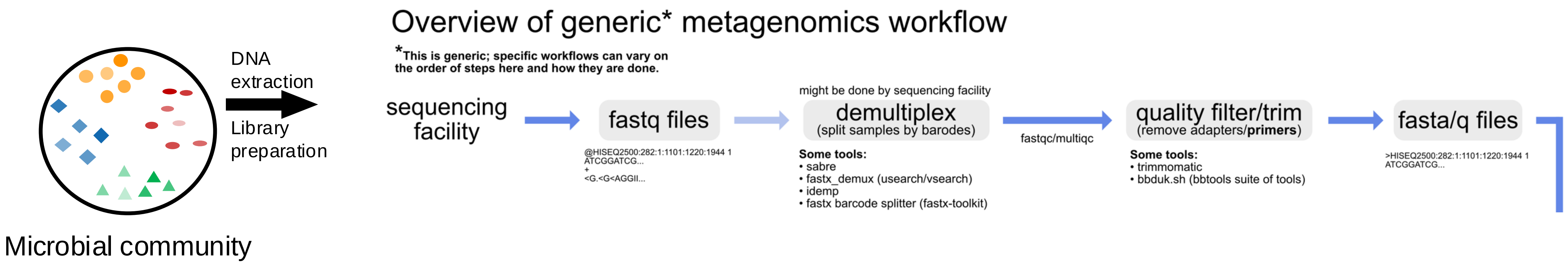
- Sufficiently high quality and <u>quantity</u> needed
- Extracted DNA is sheared into smaller fragments (inserts)
  - Illumina: ~300-600 bp; PacBio: ~20 kbp; ONT: no limit
- Adapters (of known sequences) are added to inserts
  - To allow for multiplexing samples; as templates for sequencing primers

Example (Illumina library):

- Index sequences (i5, i7) allow for multiplexing
- Illumina adapters (P5, P7) as template for <u>forward</u> and <u>reverse</u> (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
```

Modified after astrobiomike.github.io; https://teichlab.github.io/scg_lib_structs/methods_html/Illumina.html



Overview of generic* metagenomics workflow

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

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

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

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

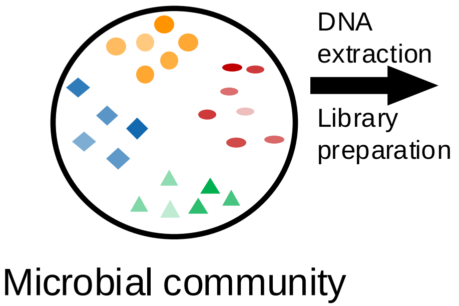
$Q = -10 \log_{10} P$

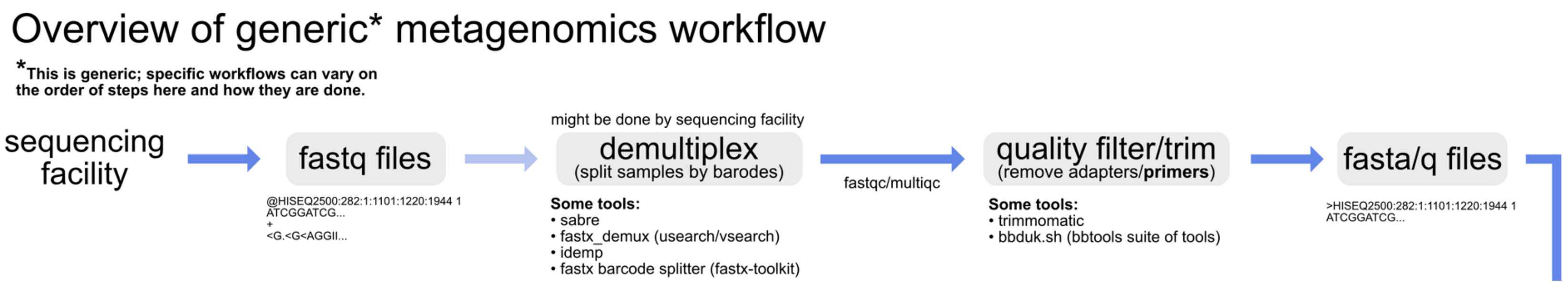
Probability error: $P = 10^{-Q/10}$

Probability truth: $1 - P$

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

Modified after astrobiomike.github.io



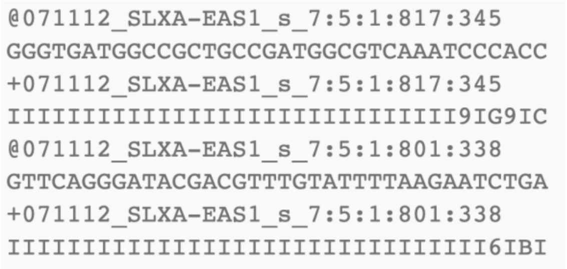

Overview of generic* metagenomics workflow
*This is generic; specific workflows can vary on the order of steps here and how they are done.
sequencing facility
fastq files
@HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G.<G<AGGII...
might be done by sequencing facility
demultiplex
(split samples by barodes)
Some tools:
• sabre
• fastx_demux (usearch/vsearch)
• idemp
• fastx barcode splitter (fastx-toolkit)
fastqc/multiqc
quality filter/trim
(remove adapters/primers)
Some tools:
• trimmomatic
• bbduk.sh (bbtools suite of tools)
fasta/q files
>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...

From raw sequencing reads to high-quality reads

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

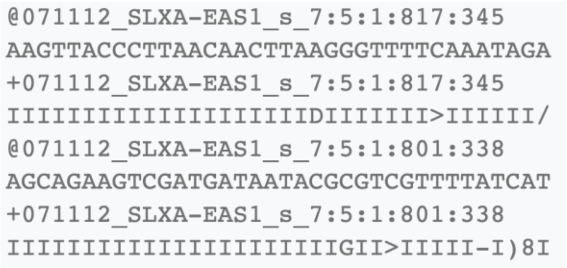
Example:

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
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII6IBI
```

two “reverse” reads:



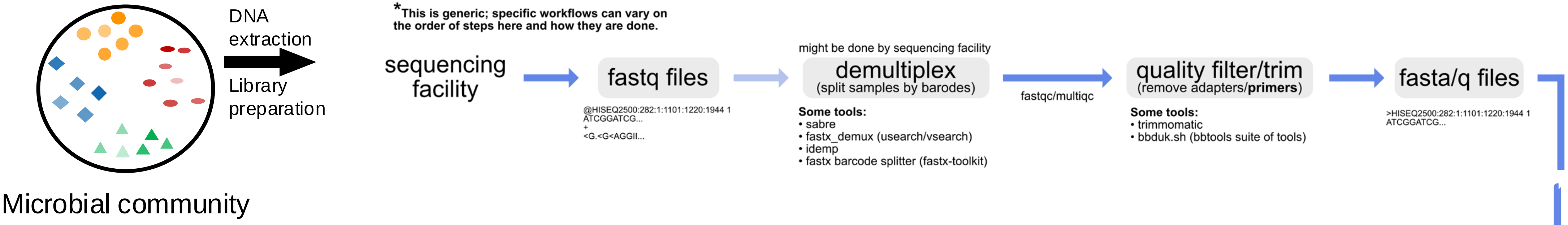
```
@071112_SLXA-EAS1_s_7:5:1:817:345
AAGTTACCCTTAACAACTTAAGGGTTTTCAAATAGA
+071112_SLXA-EAS1_s_7:5:1:817:345
IIIIIIIIIIIIIIIIIIIIDIIIIIII>IIIIII/
@071112_SLXA-EAS1_s_7:5:1:801:338
AGCAGAAGTCGATGATAATACGCGTCGTTTTATCAT
+071112_SLXA-EAS1_s_7:5:1:801:338
IIIIIIIIIIIIIIIIIIIIIIGII>IIIII-I)8I
```

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

Modified after astrobiomike.github.io

Overview of generic* metagenomics workflow

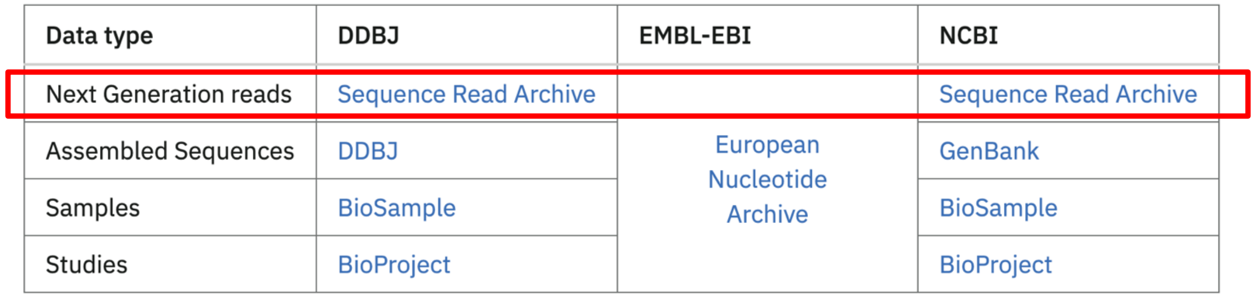
*This is generic; specific workflows can vary on the order of steps here and how they are done.


DNA extraction
Library preparation
Microbial community
sequencing facility
fastq files
@HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G.<G<AGGII...
might be done by sequencing facility
demultiplex
(split samples by barodes)
Some tools:
• sabre
• fastx_demux (usearch/vsearch)
• idemp
• fastx barcode splitter (fastx-toolkit)
fastqc/multiqc
quality filter/trim
(remove adapters/primers)
Some tools:
• trimmomatic
• bbduk.sh (bbtools suite of tools)
fasta/q files
>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...

Storage of DNA sequence information

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

Databases



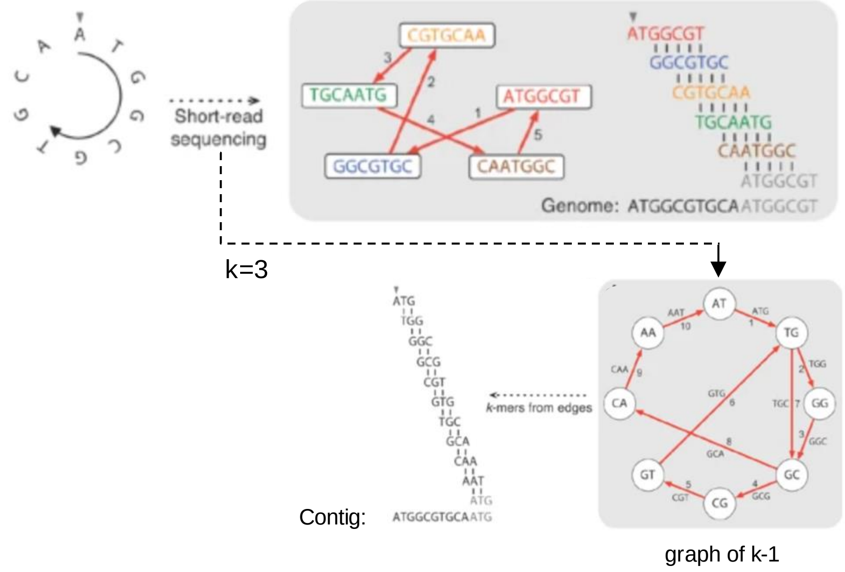
Data type	DDBJ	EMBL-EBI	NCBI
Next Generation reads	Sequence Read Archive		Sequence Read Archive
Assembled Sequences	DDBJ	European Nucleotide Archive	GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

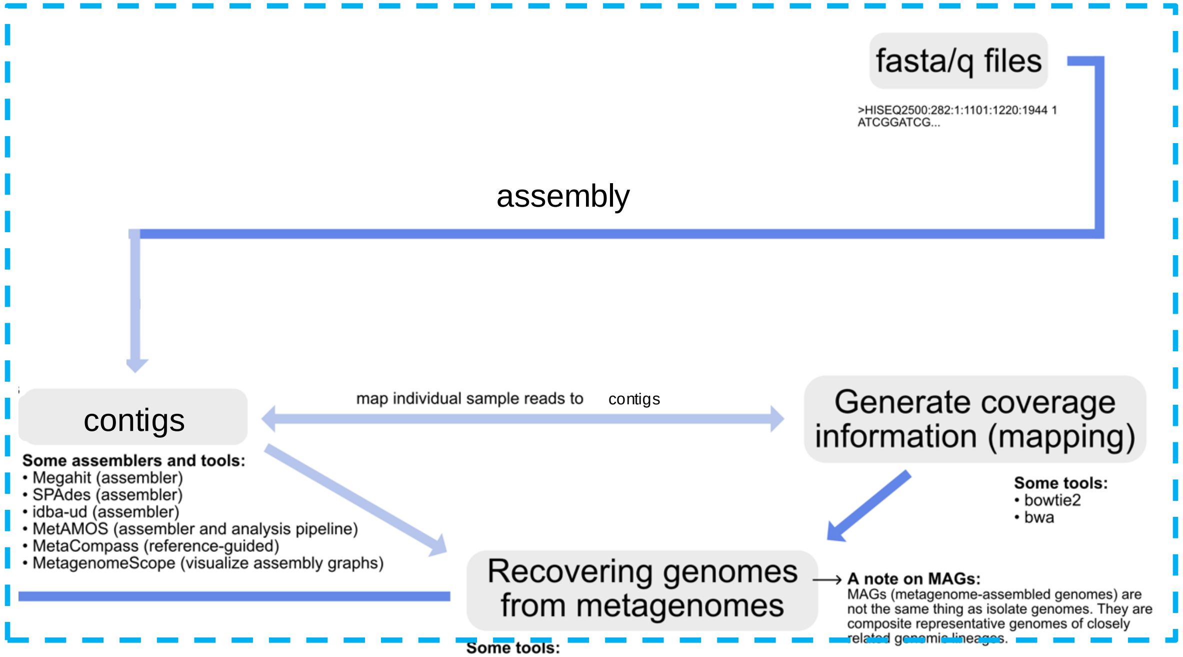
Modified after astrobiomike.github.io

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




fasta/q files

>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...

assembly

contigs

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

map individual sample reads to contigs

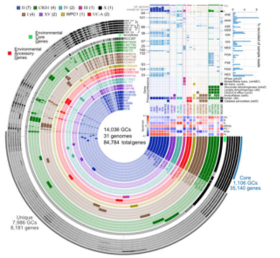
Generate coverage information (mapping)

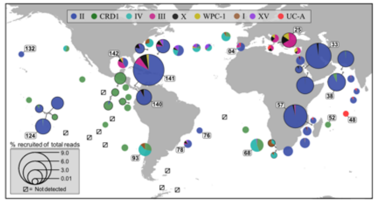
Some tools:
- bowtie2
- bwa

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)





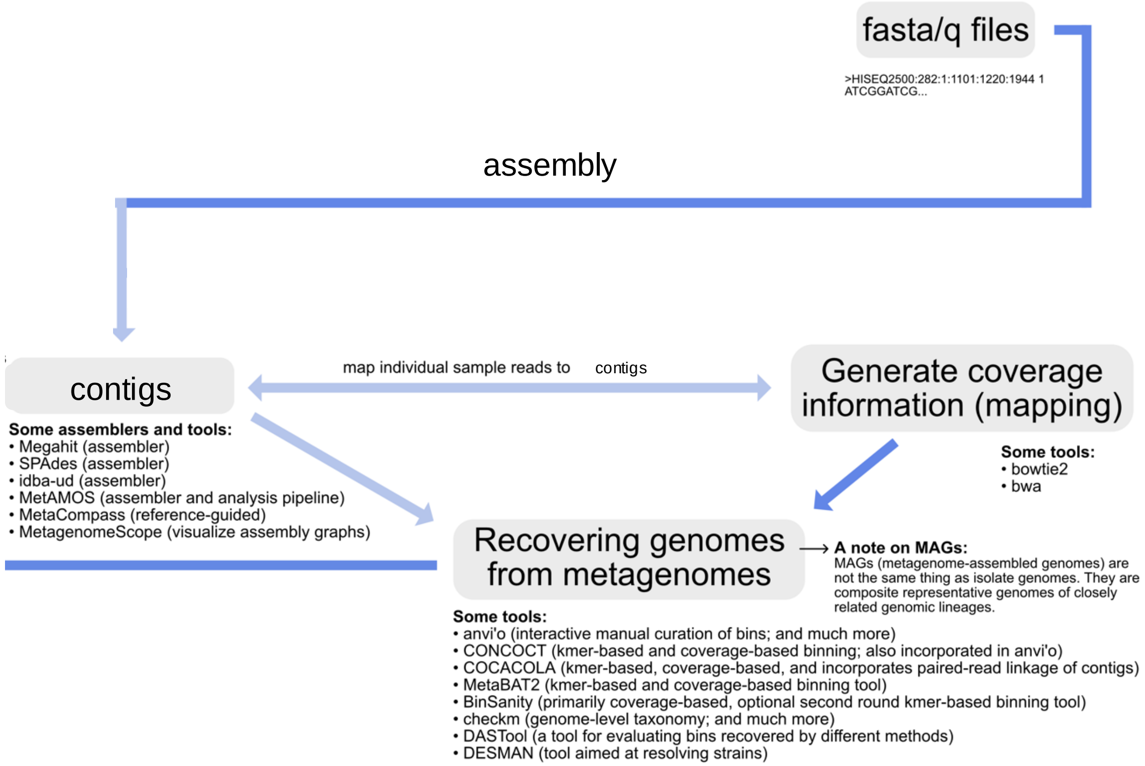
astrobiomike.github.io

Modified after astrobiomike.github.io and Compeau, Pevzner and Tesler, Nat Biotechnol, 2011

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)

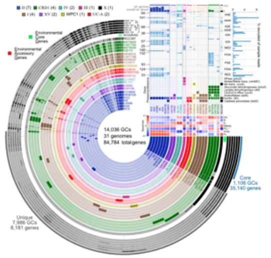

fasta/q files
>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
assembly
contigs
map individual sample reads to contigs
Generate coverage information (mapping)
Recovering genomes from metagenomes

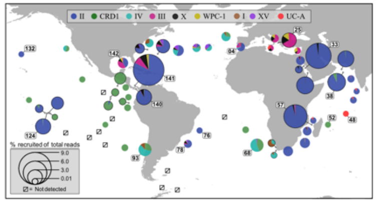
Some assemblers and tools:
- Megahit (assembler)
- SPAdes (assembler)
- idba-ud (assembler)
- MetAMOS (assembler and analysis pipeline)
- MetaCompass (reference-guided)
- MetagenomeScope (visualize assembly graphs)

Some tools:
- bowtie2
- bwa

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

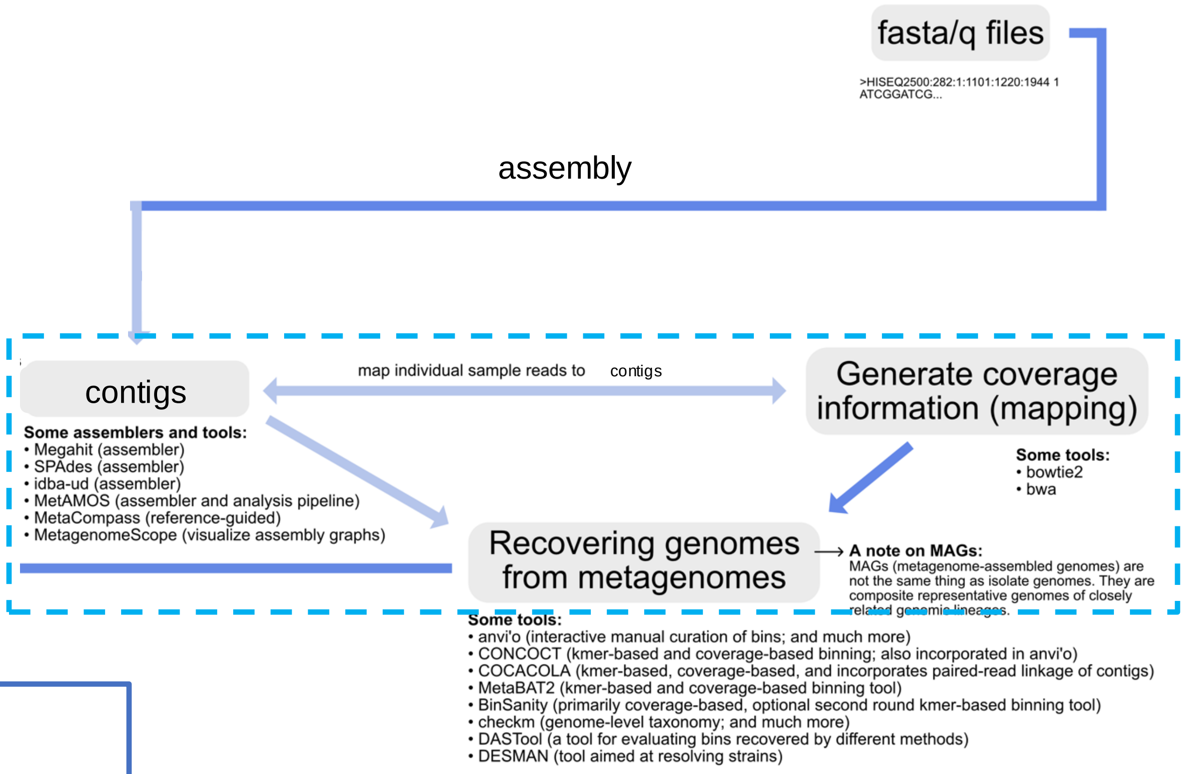
Modified after astrobiomike.github.io

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)

→ How do we group (bin) these contigs to recover the original genomes they originated from?

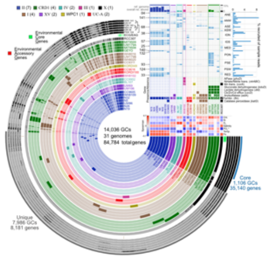

fasta/q files
>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
assembly
contigs
map individual sample reads to contigs
Generate coverage information (mapping)
Recovering genomes from metagenomes

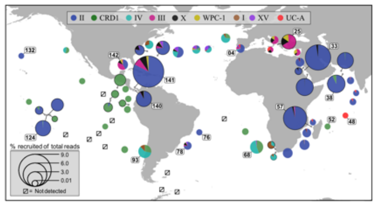
Some assemblers and tools:
- Megahit (assembler)
- SPAdes (assembler)
- idba-ud (assembler)
- MetAMOS (assembler and analysis pipeline)
- MetaCompass (reference-guided)
- MetagenomeScope (visualize assembly graphs)

Some tools:
- bowtie2
- bwa

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)





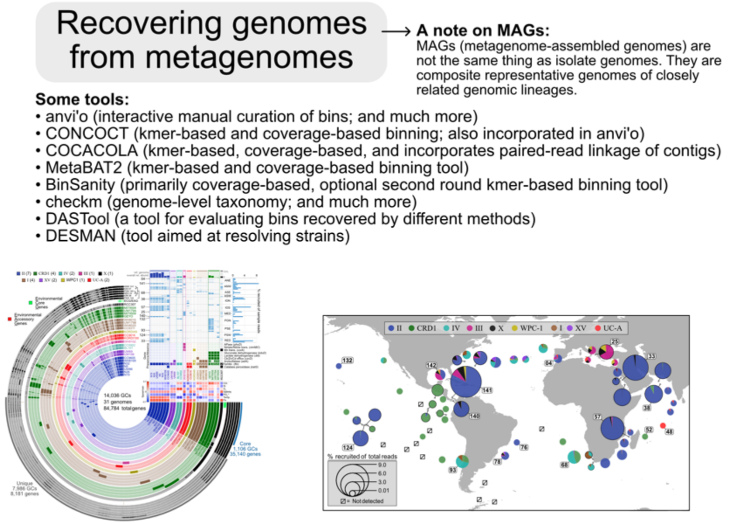
astrobiomike.github.io

Modified after 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

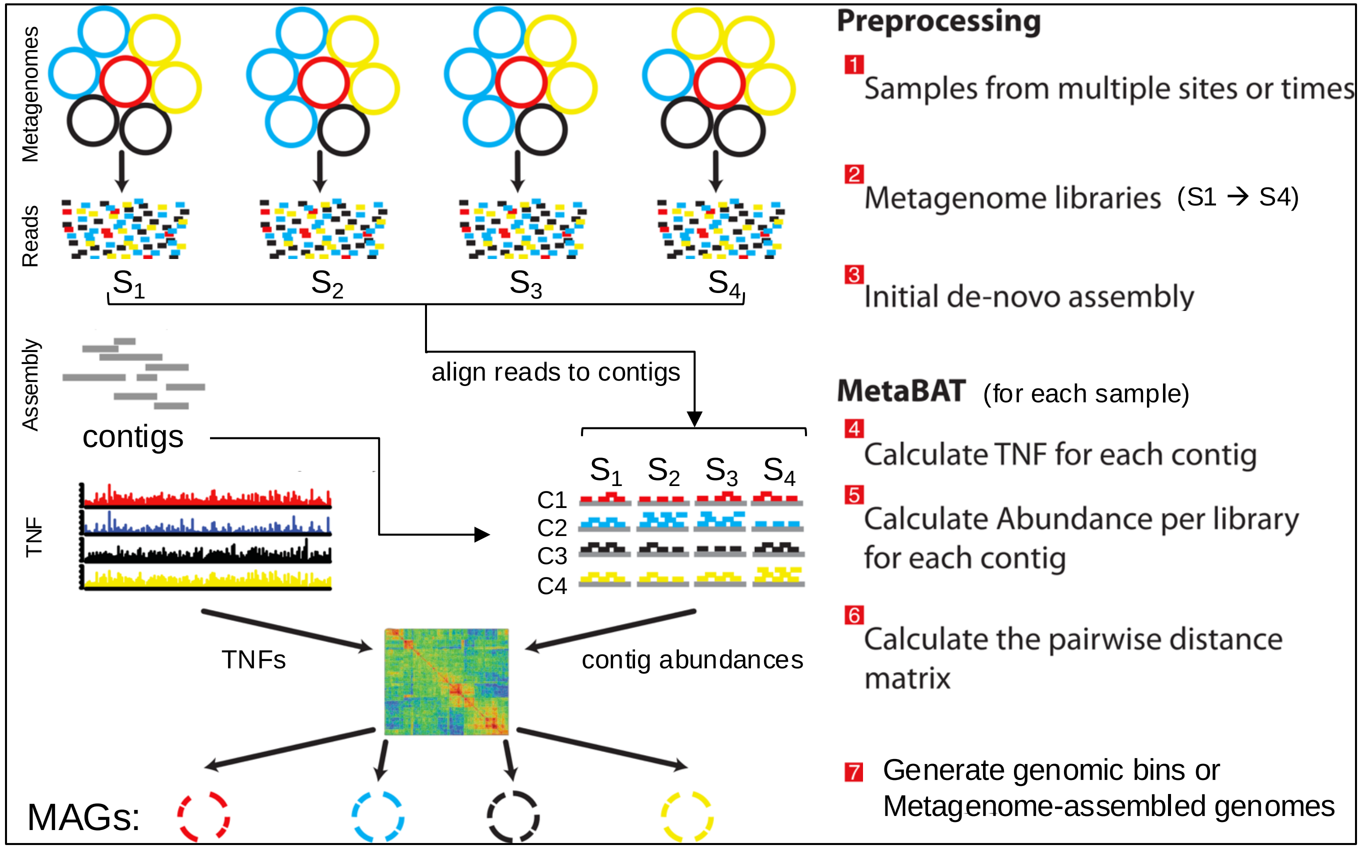
→ **A note on MAGs:** MAGs (metagenome-assembled genomes) are not the same thing as isolate genomes. They are composite representative genomes of closely related genomic lineages.

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

Modified after astrobiomike.github.io

Binning contigs into metagenome-assembled genomes



Preprocessing

1 Samples from multiple sites or times

2 Metagenome libraries (S1 → S4)

3 Initial de-novo assembly

MetaBAT (for each sample)

4 Calculate TNF for each contig

5 Calculate Abundance per library for each contig

6 Calculate the pairwise distance matrix

7 Generate genomic bins or Metagenome-assembled genomes

From contigs to metagenome-assembled genomes (MAGs)

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

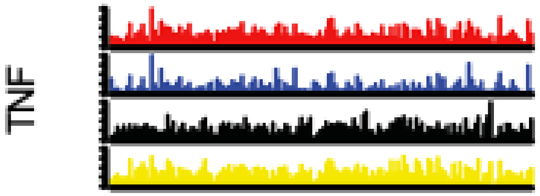
a) Tetranucleotide frequencies (TNFs)
b) Abundances of contigs within and across samples

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

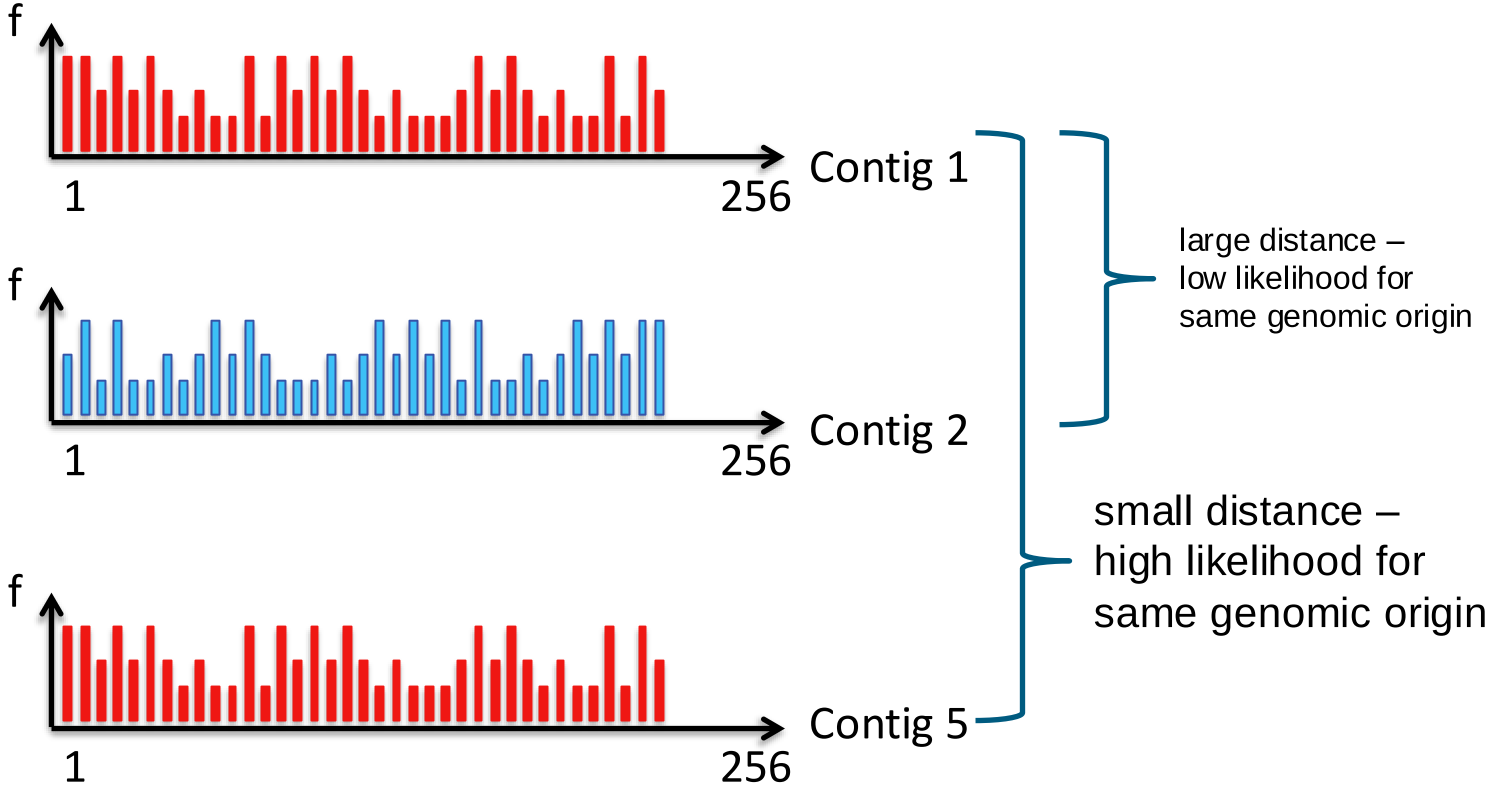
After Kang et al., PeerJ, 2015

TNF is constant within a genome and different between genomes

Tetranucleotide (k=4) frequencies



$[ATGC]^4$ = 256 possible combinations

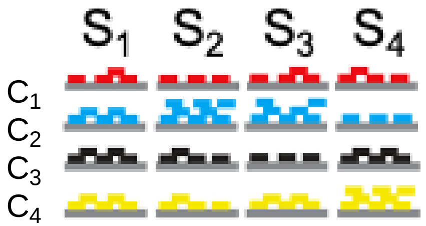


large distance – low likelihood for same genomic origin

small distance – high likelihood for same genomic origin

Teeling et al., Environ Microbiol, 2004

Contig abundances within and across samples



	S_1	S_2	S_3	S_4	Sn
C_1	5	5	5	5	
C_2	10	15	15	10	
C_3	10	5	5	10	
C_4	10	10	10	15	
C_5	5	5	5	5	
C_6	10	15	15	10	
C_7	10	10	10	15	
C_8	10	5	5	10	
C_n					

→ Contigs with similar abundance within samples **may** be of same genome origin

→ Contigs with high abundance correlations across samples **are likely** of same genome origin

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, codon usage, GC content) of protein coding regions

Genome annotation – protein coding genes

Example - Open Reading Frame (ORF) finding

- 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…

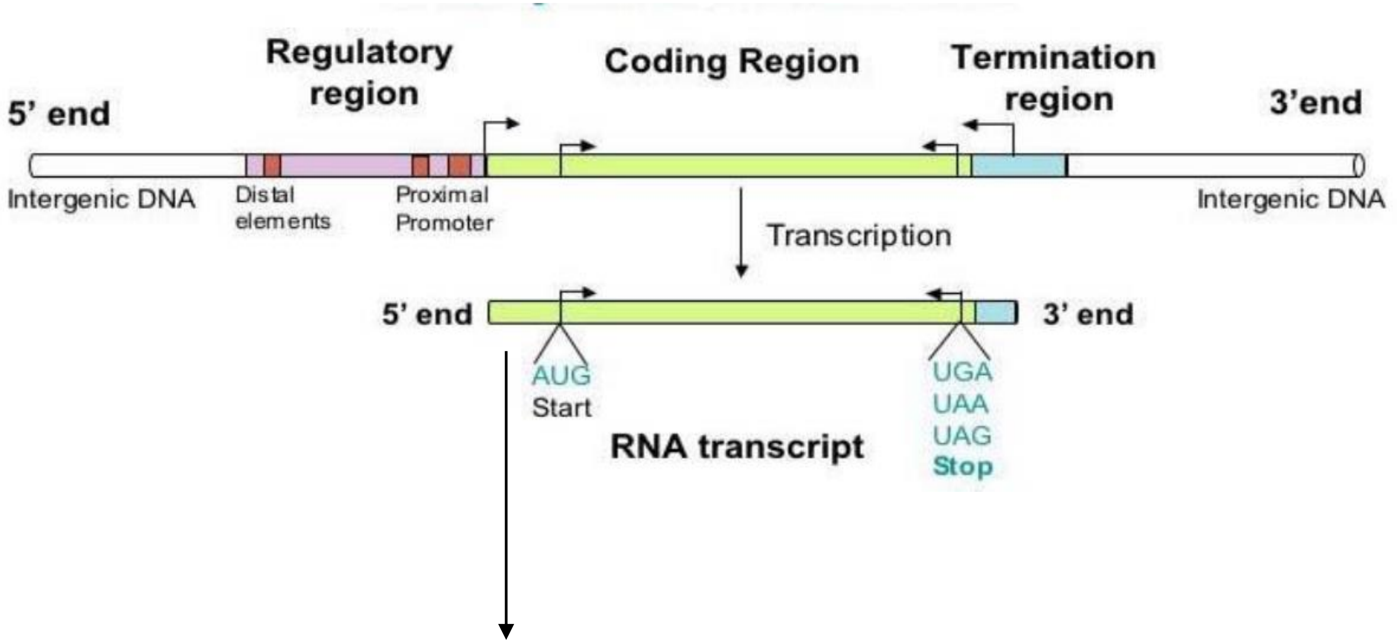
…GCT GCT ATC AGA CAC TCT TTT TTT AAT CCA CAC AGA GAC ATA TTG CCC GTT GCA GTC AGA ATG AAA AGC T…

…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

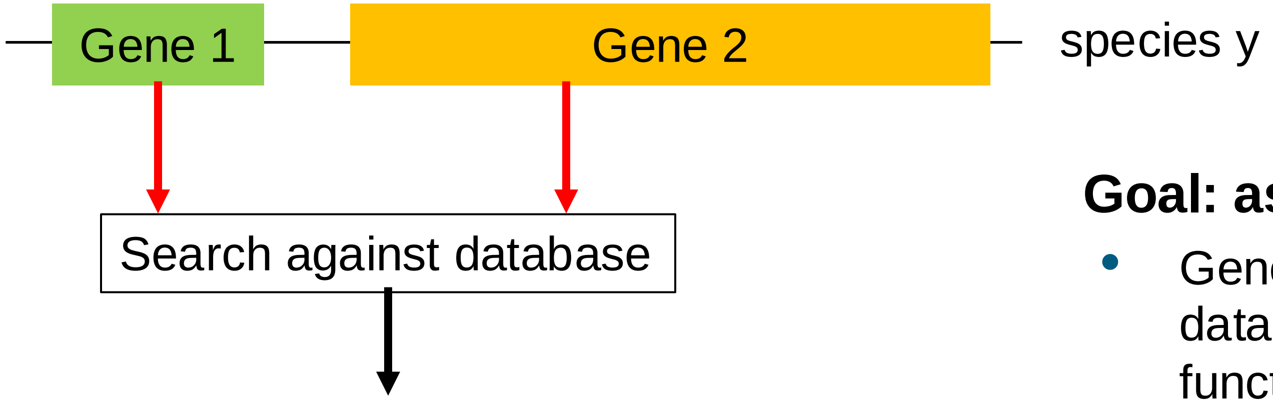
Prokaryotic gene structure



The 5'-UTR (untranslated region):

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

Functional annotation of genes

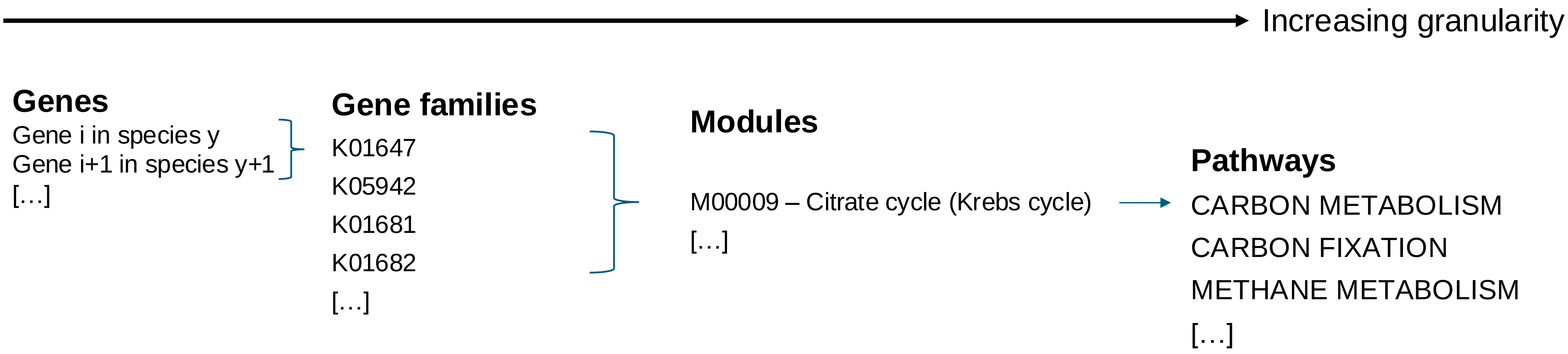


	DB	Identifier
Database 1:	KEGG	K00847
Database 2:	COG	COG1940
Database 3:	Pfam	PF00480

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…

Functional annotation of genes: KEGG database



Increasing granularity

Genes

Gene i in species y
Gene i+1 in species y+1
[…]

Gene families

K01647
K05942
K01681
K01682
[…]

Modules

M00009 – Citrate cycle (Krebs cycle)
[…]

Pathways

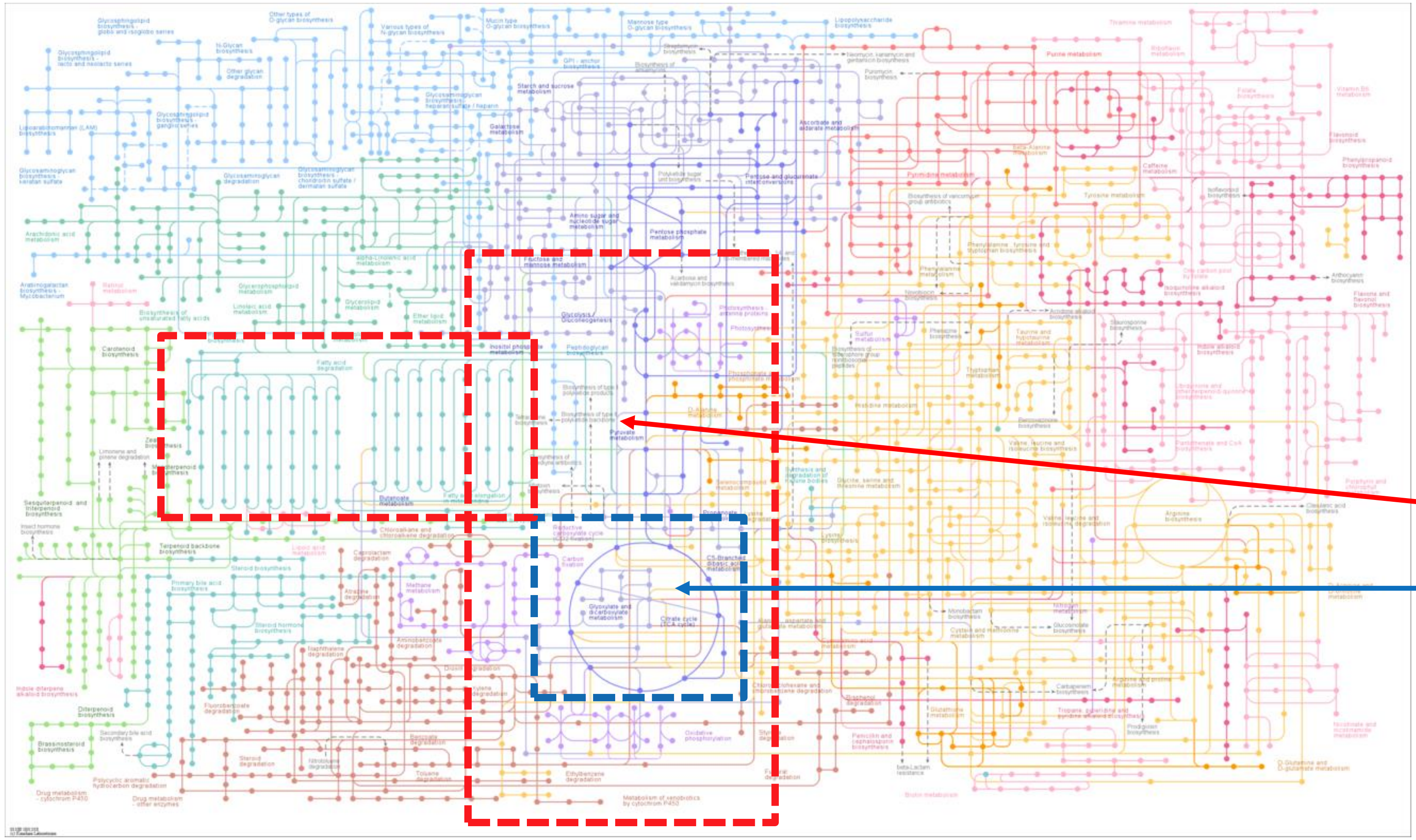
CARBON METABOLISM
CARBON FIXATION
METHANE METABOLISM
[…]

→ Genes are members of gene families
→ Gene families are members of modules
→ Modules are members of pathways

After: Tatusov, Koonin, Lipman, Science, 1997

Example: KEGG database

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



Map of known metabolic reactions

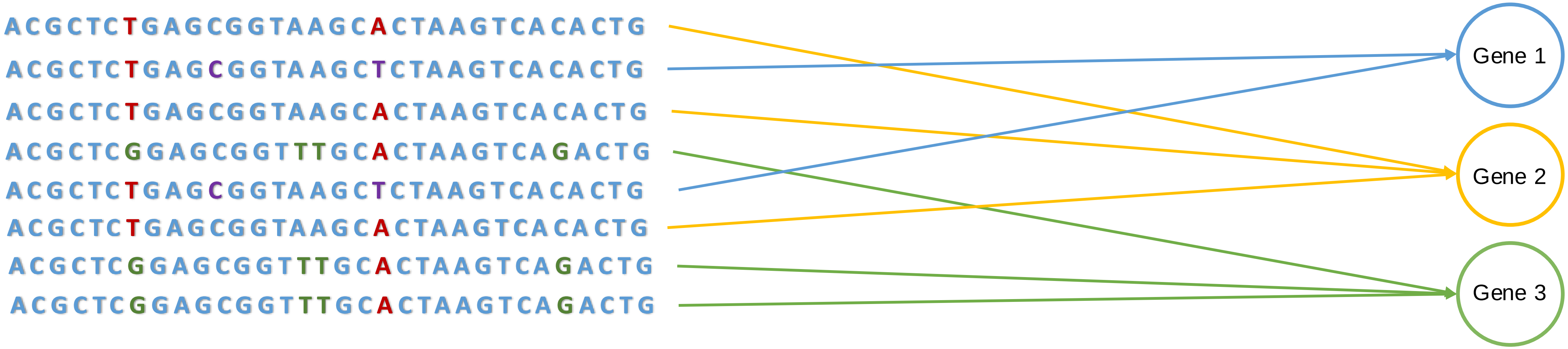
- Nodes = compounds
- Connections = reactions catalyzed by known enzymes
- Enzymes grouped into KOs = KEGG orthologous groups
- Map divided into:

  pathways and

  modules

Quantification of gene abundances

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


ACGCTCTGAGCGGTAAGCACTAAGTCACACTG
ACGCTCTGAGCGGTAAGCTCTAAGTCACACTG
ACGCTCTGAGCGGTAAGCACTAAGTCACACTG
ACGCTCGGAGCGGTTTGCACTAAGTCAGACTG
ACGCTCTGAGCGGTAAGCTCTAAGTCACACTG
ACGCTCTGAGCGGTAAGCACTAAGTCACACTG
ACGCTCGGAGCGGTTTGCACTAAGTCAGACTG
ACGCTCGGAGCGGTTTGCACTAAGTCAGACTG
Gene 1
Gene 2
Gene 3

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

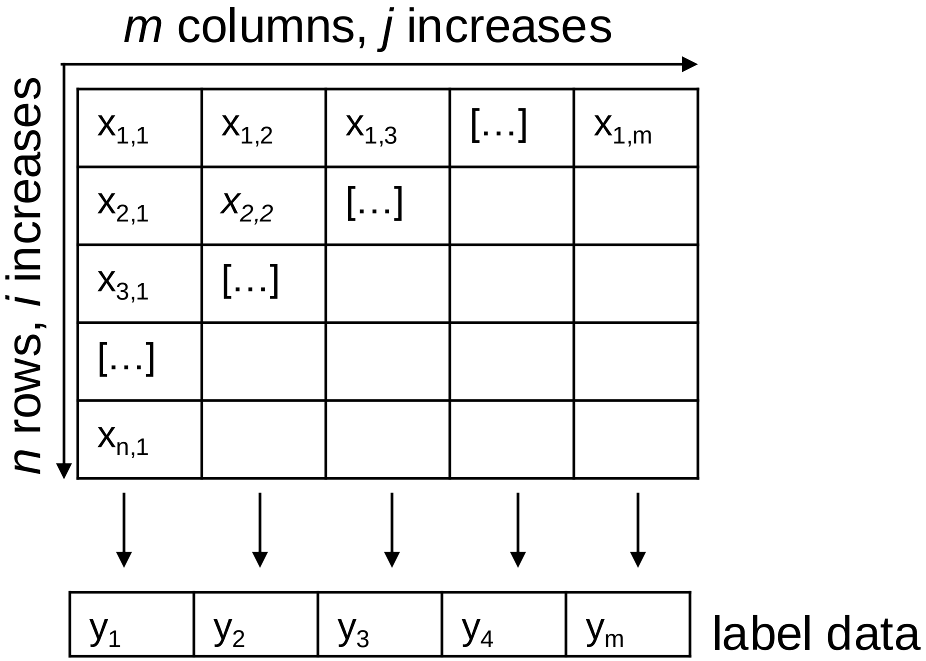
KO abundance tables can be summarized into Module abundance tables

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*



m columns, *j* increases

n rows, *i* increases

$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	[…]	$x_{1,m}$
$x_{2,1}$	$x_{2,2}$	[…]		
$x_{3,1}$	[…]			
[…]				
$x_{n,1}$				

y_1	y_2	y_3	y_4	y_m

label data

→ **Enables differential abundance testing**

- 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
[…]	[…]	[…]

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