מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

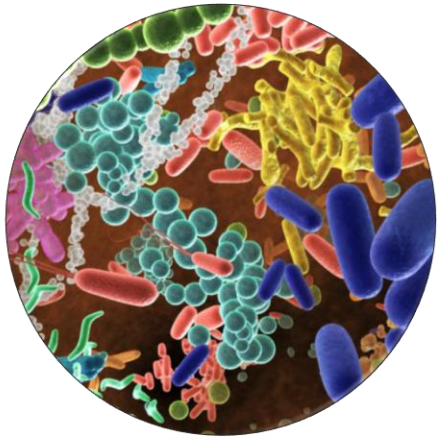
An introduction to Metagenomics

Bareket Dassa, Bioinformatics Unit

Introduction to Deep Sequencing Analysis course

2022

What is metagenomics?



Metagenomics explores complex **microbial communities,** which **cannot be cultured,** using high-throughput gene-level methods

(In Greek, *meta* means “transcendent”)

Metagenomics goes beyond the pure culture and single genome approaches

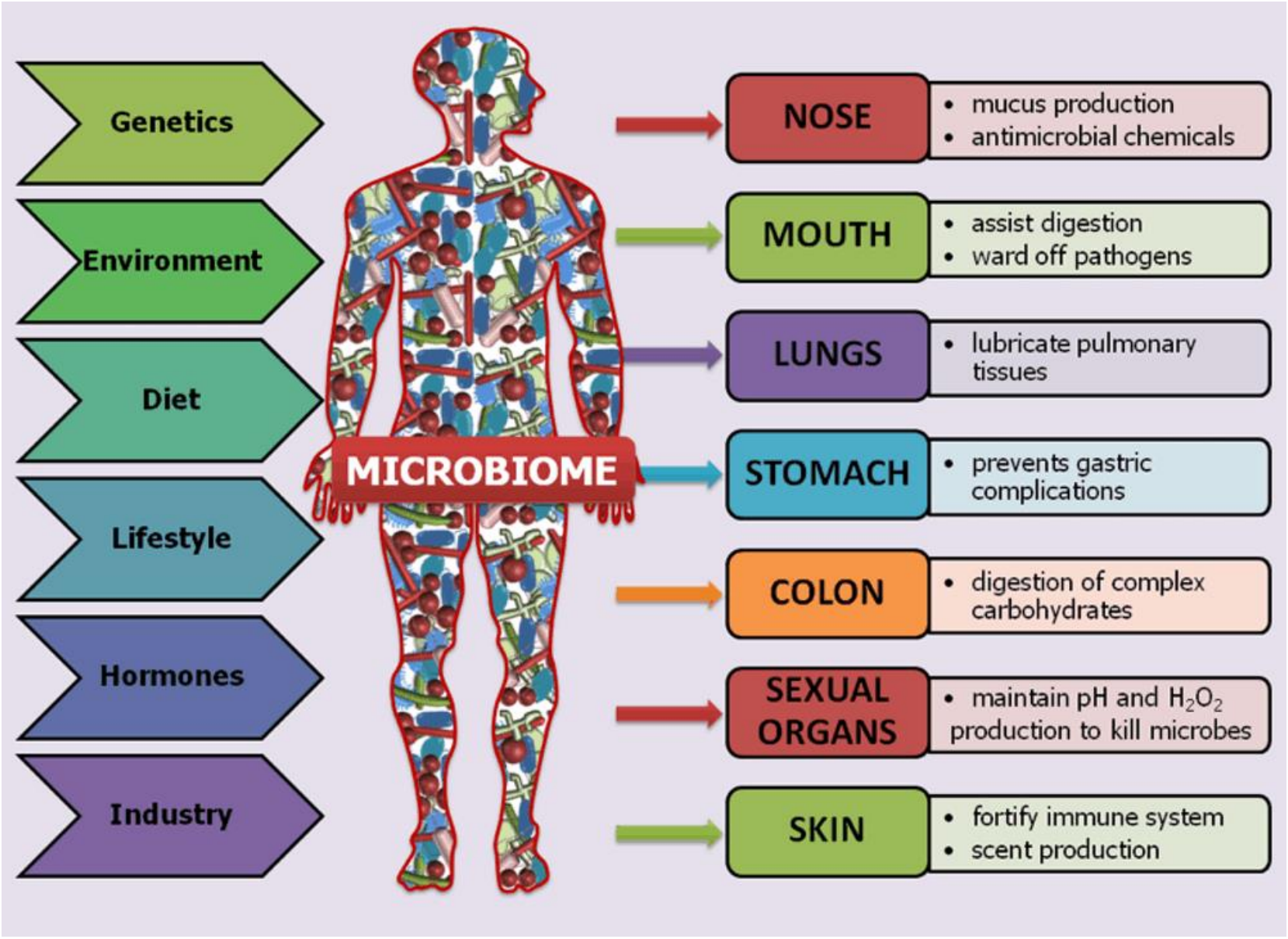
Microbial diversity

Microbes are everywhere, and are essential for life on Earth:

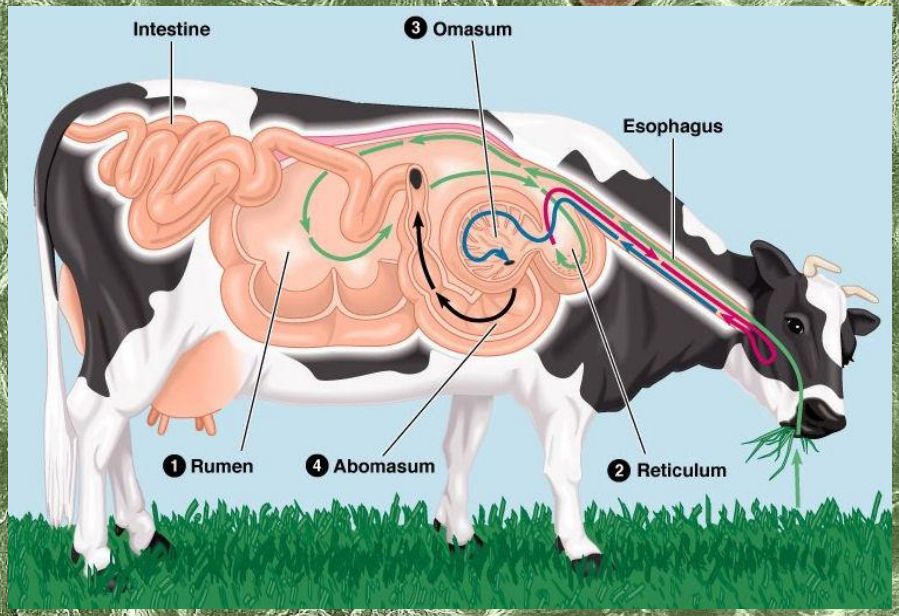
- biosphere cycles
- plant nutrients
- remediate toxins
- human diet…

Thermophilic microbial mat, Yellowstone National Park (Marissa Fessenden)

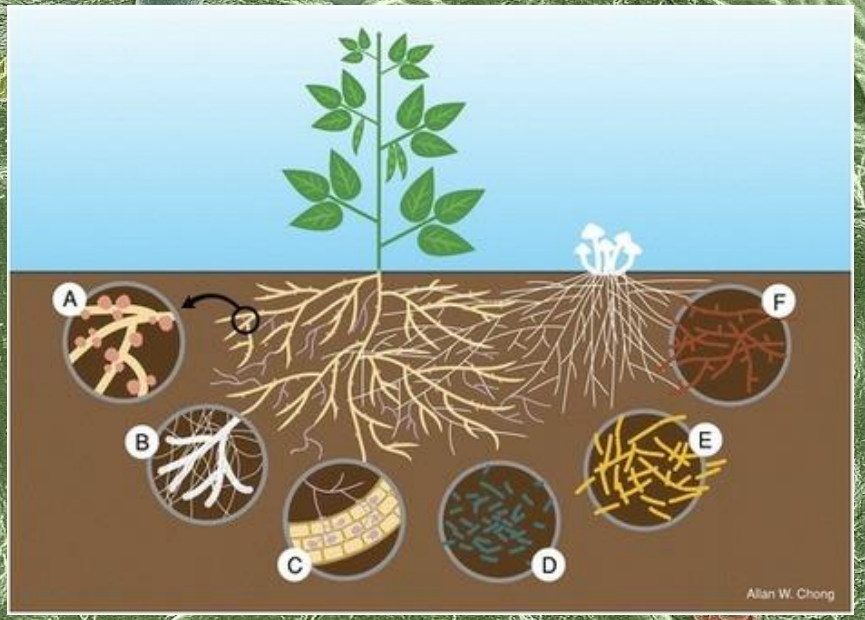
The human microbiome



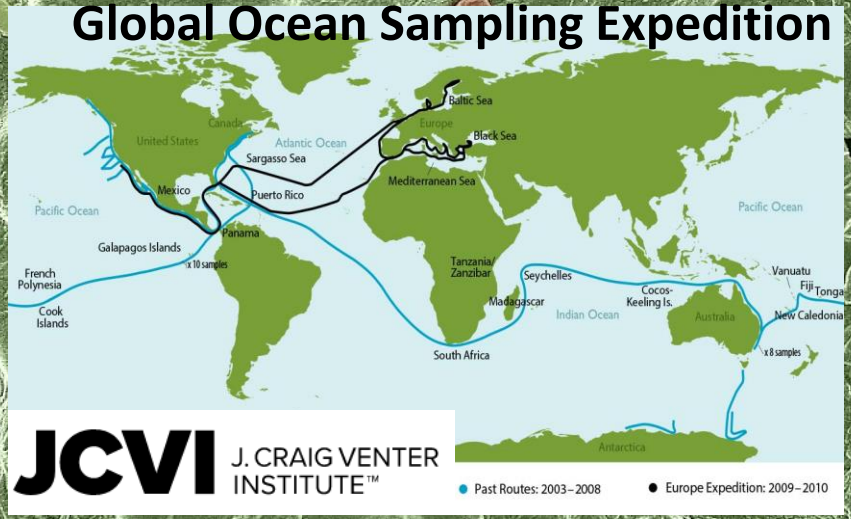
Animal-related rumen microbiome



Soil microbiome



Global Ocean Sampling Expedition

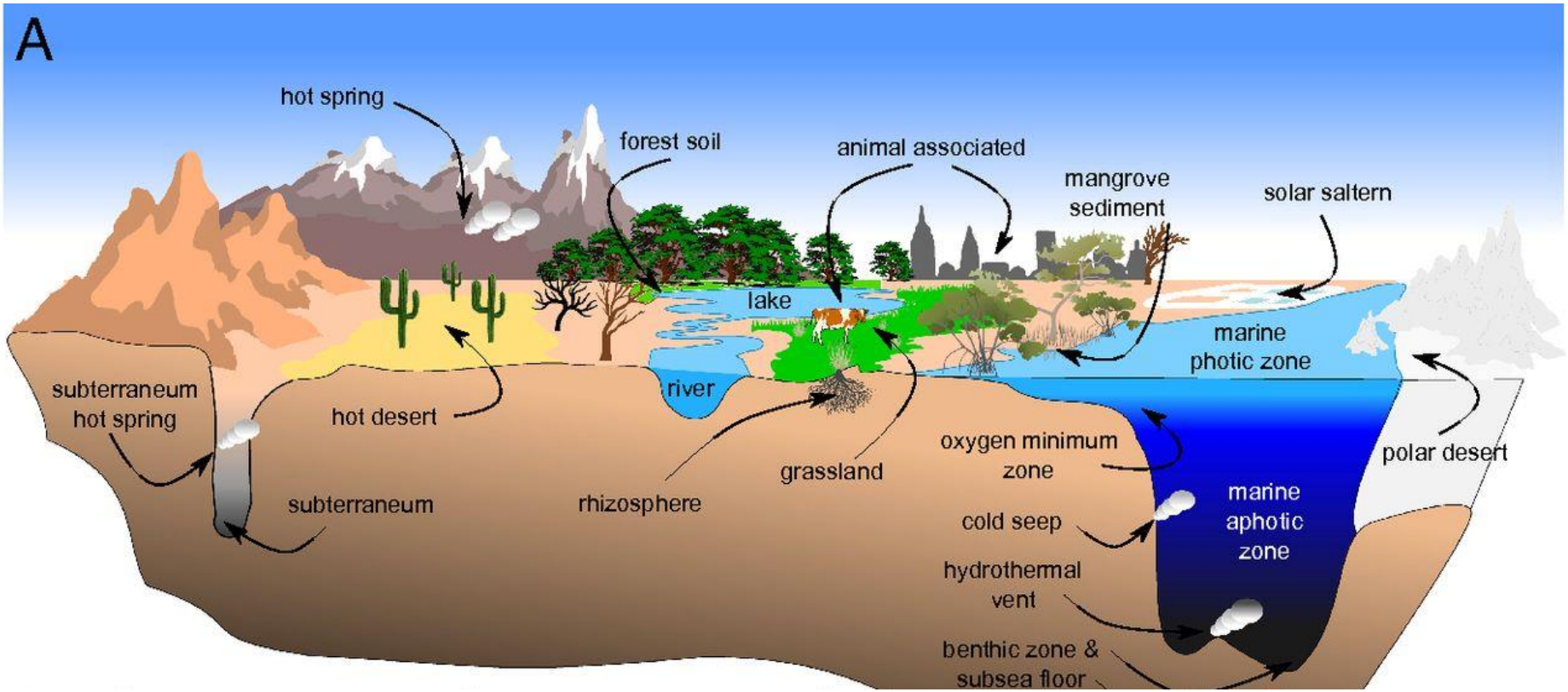


Microbes colonizing the surface of grass

https://www.creeveylab.org/2017/03/detecting-microbial-niches-in.html

Exploring microbiomes from diverse habitats

A microbiome is a community in an environment.
We explore the huge biodiversity and complex communities in oceans, soils, animals, the human body



https://doi.org/10.1073/pnas.1817554116

Why metagenomics?

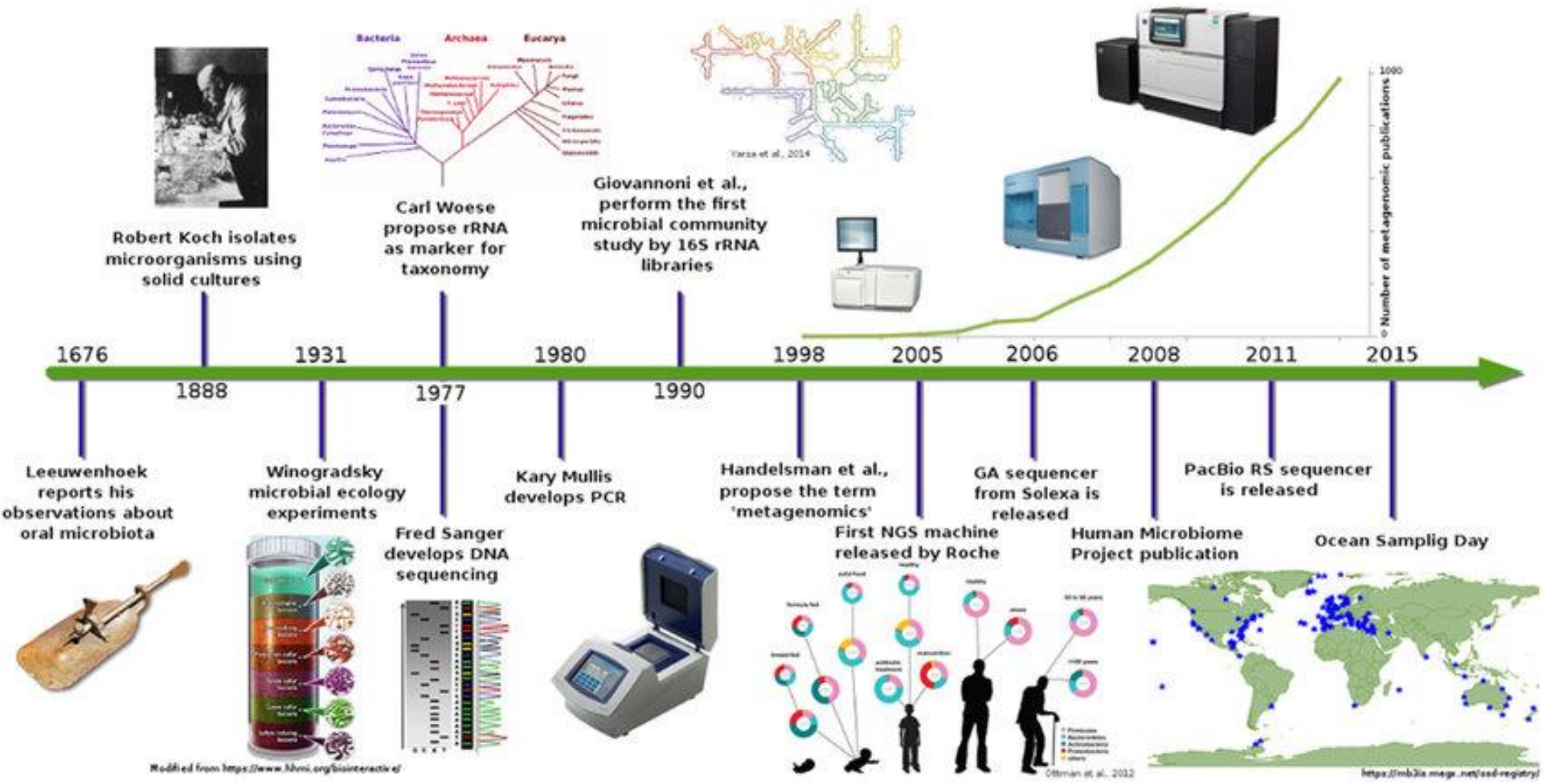
Why traditional genomics is not enough?

The (unlimited) **microbial diversity** is a largely underexplored because
Most microbial communities cannot be cultured

Using metagenomics tools we can explore:

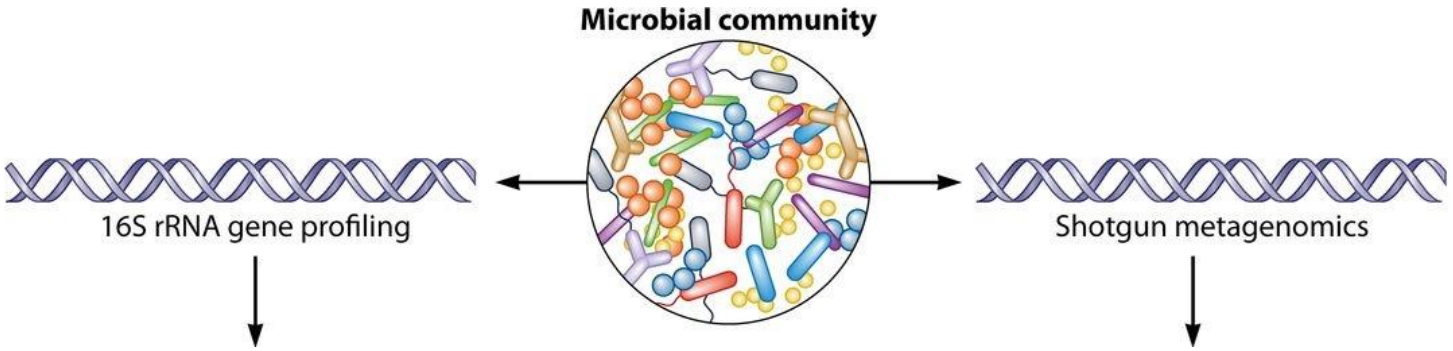
- What is the **composition** of microbial communities? Which species? Abundancies? **Function**? Dynamics?
- Can microbiome provide **novel biomarkers** for diseases, personalized profiles?

Metagenomics timeline and milestones



https://www.researchgate.net/figure/Metagenomics-timeline-and-milestones-Timeline-showing-advances-in-microbial-communities_fig1_289524171

Metagenomics strategies



(1) 16S rRNA / targeted / amplicon sequencing

Taxonomy profiling

Who is there?

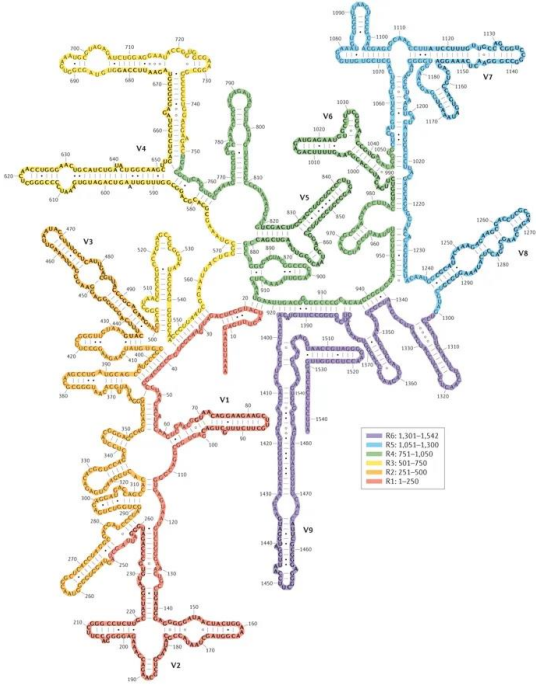
(2) Whole metagenome shotgun (WGS) sequencings

Taxonomy & **Functional** analysis

What are they doing?

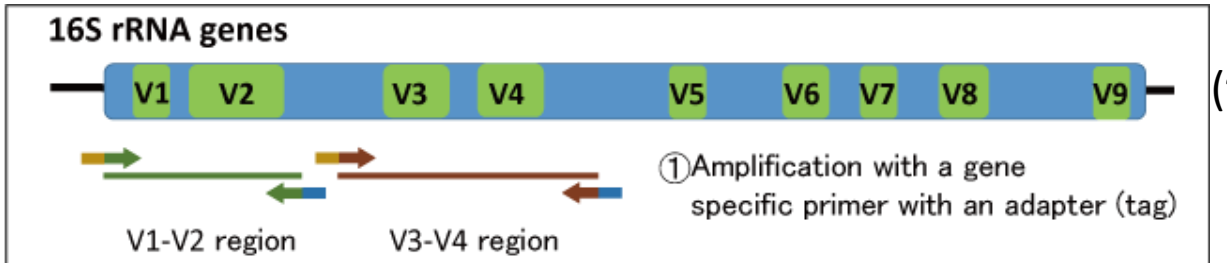
16S rRNA sequencing

Target the gene *encoding* the small-subunit ribosomal RNA (16S) locus
which is **a taxonomically informative marker** for Bacteria and Archaea (28S rRNA for Fungal)



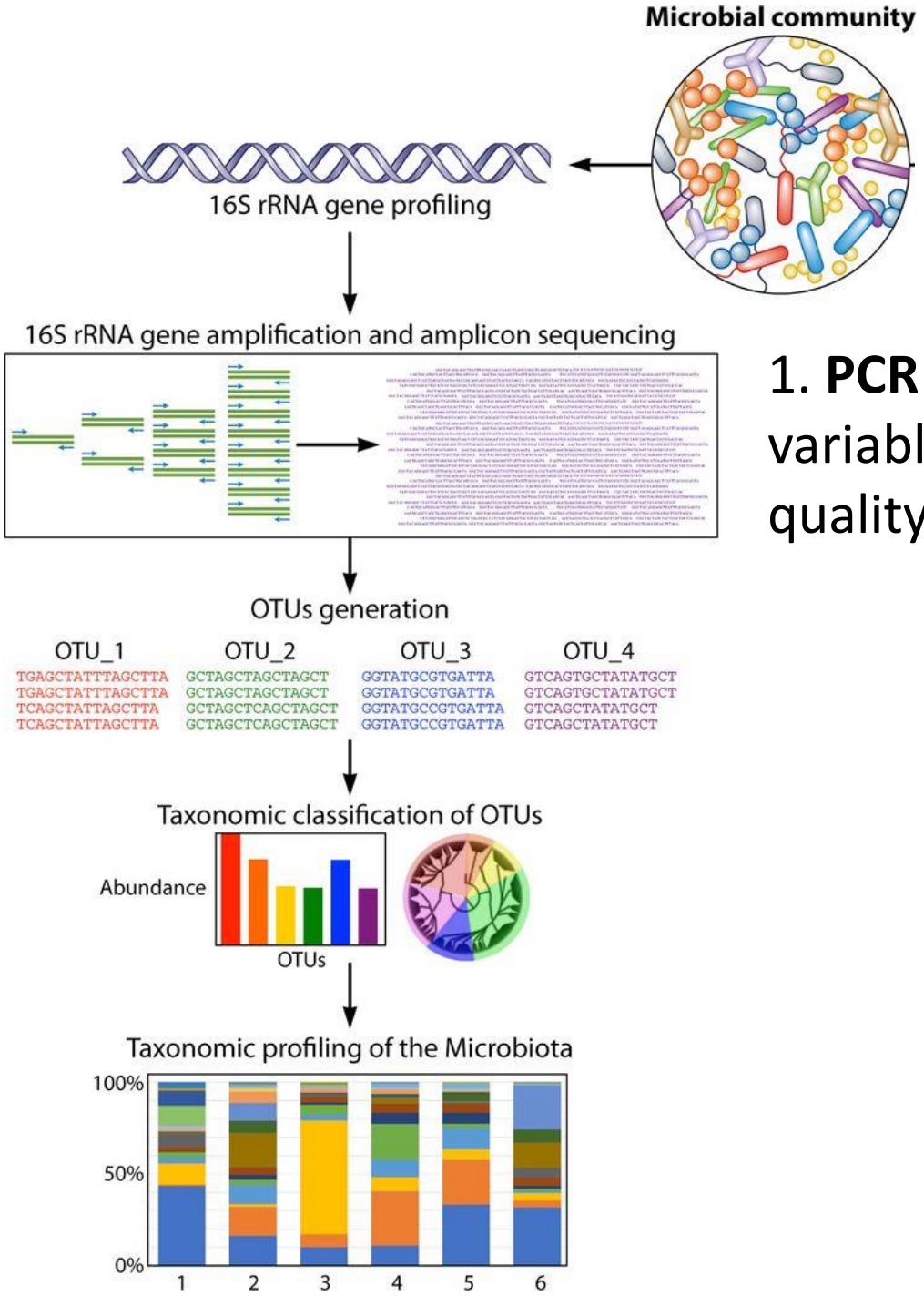
Nature Reviews | Microbiology

Secondary structure of the 16S rRNA of *E. coli*



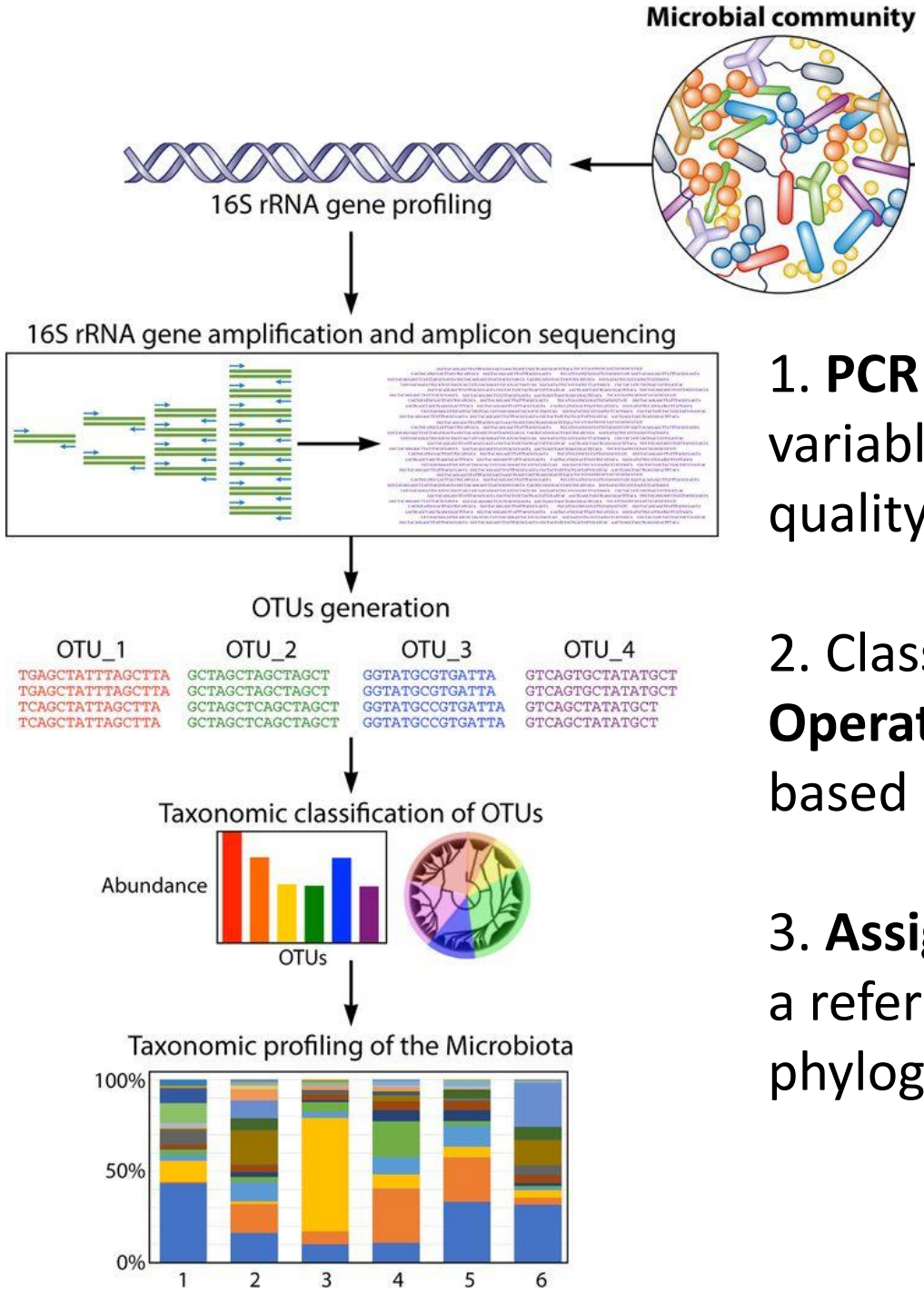
(~1.6 kb)

- An essential gene, found in all species
- Present in multiple copies within a genome
- Has conserved and variable (unique) regions



16S rRNA profiling

1. **PCR Amplification**, sequencing of variable regions in the 16S rRNA gene, quality filtering



16S rRNA profiling

1. **PCR Amplification**, sequencing of variable regions in the 16S rRNA gene, quality filtering

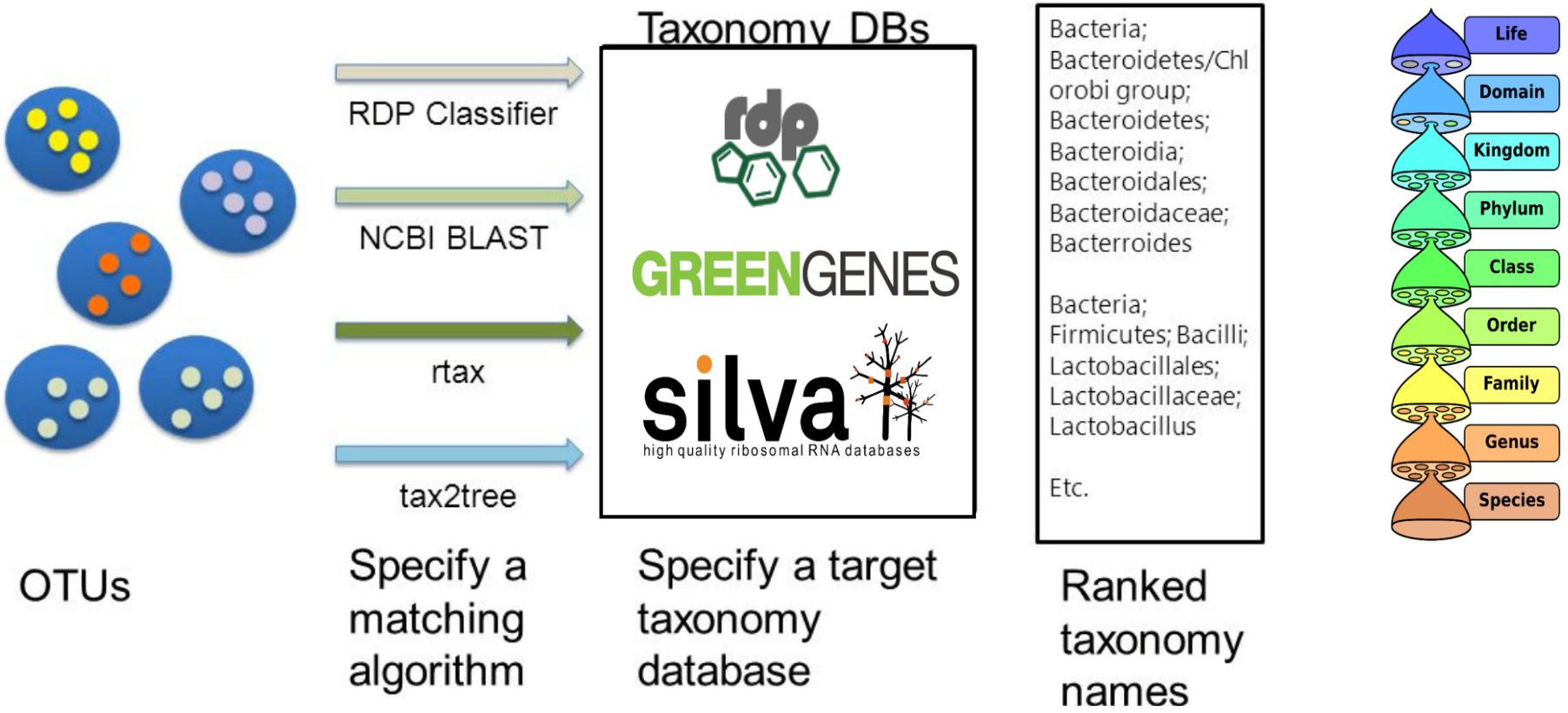
2. Classify reads into **Operational Taxonomic Units (OTUs)** based on sequence identity

3. **Assign taxonomy** - compare OTUs to a reference database, built a phylogenetic tree

How 16S rRNA taxonomic classification works?

Clustering (binning) similar sequences into
'Operational Taxonomic Units' (OTUs)

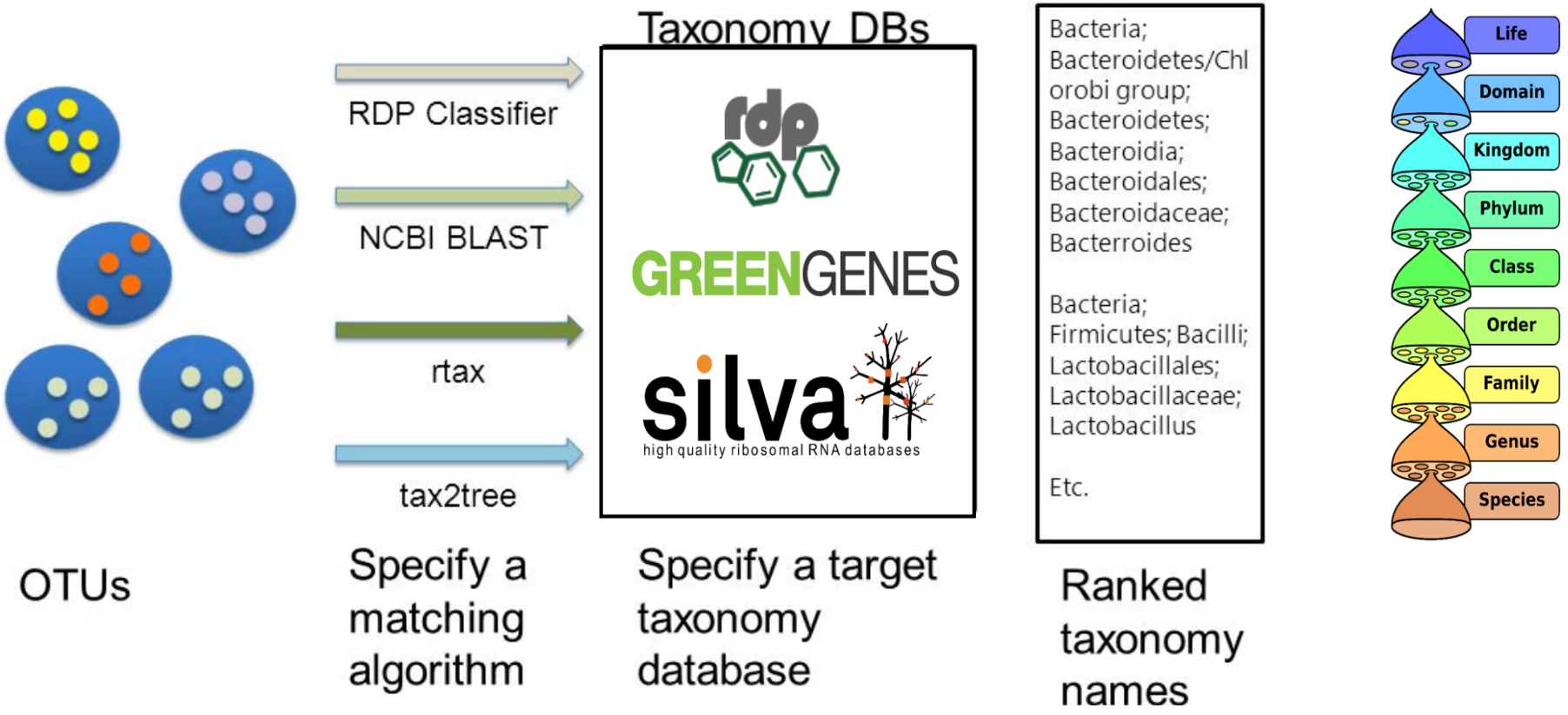
Comparing OTUs to a **reference database** of curated sequences with known taxonomic composition:



How 16S rRNA taxonomic classification works?

Determine the closest taxonomic affiliation with some degree of confidence, using:

- **Sequence** alignment similarity
- **Classifiers** (word composition, naïve Bayesian)

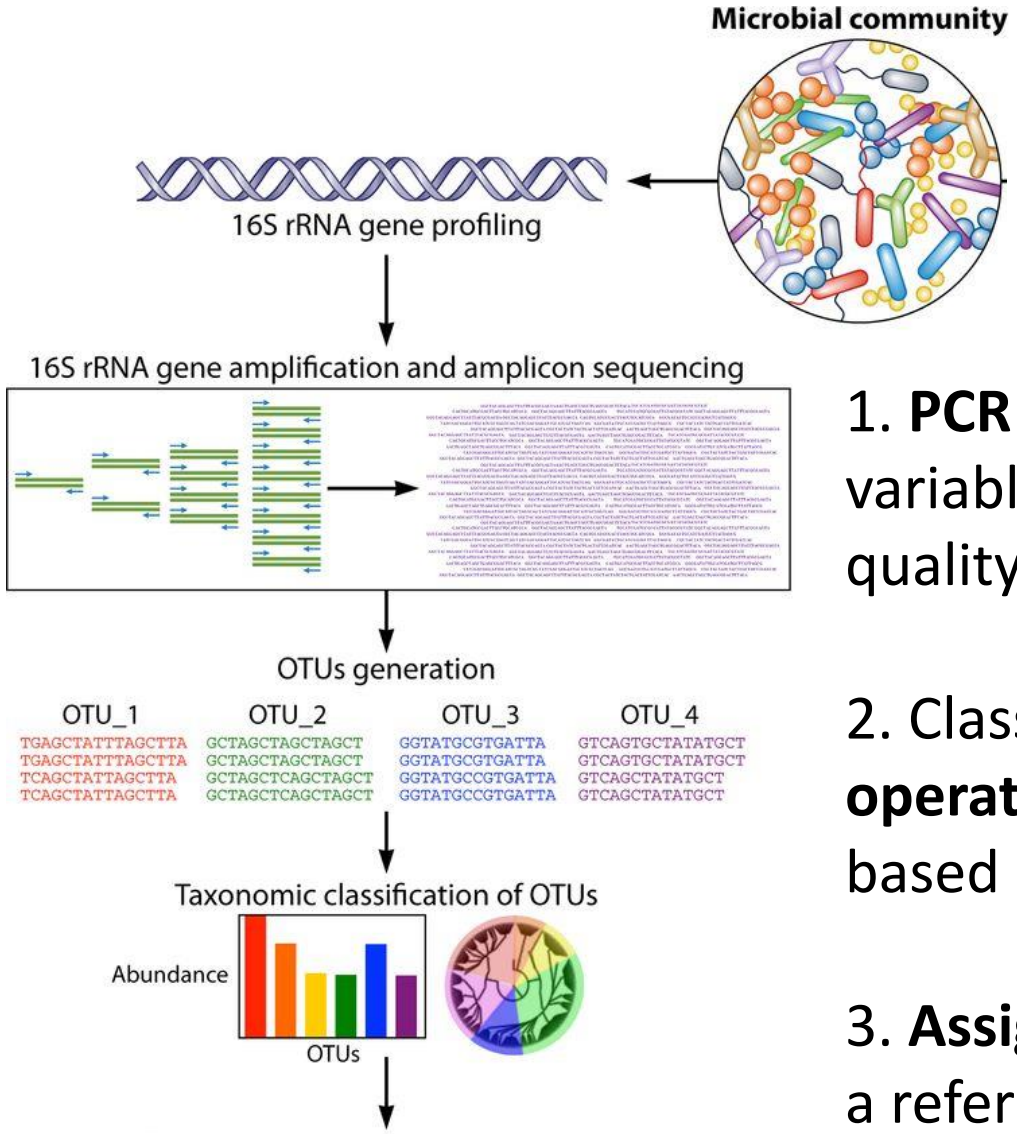


Limitations of 16S rRNA taxonomic classification

- The common **reference databases** lack sequences for most uncultivated taxa, biased toward medically relevant species

Only ~11,000 bacterial and archaeal species have been classified
It has been estimated that it would take >1,000 years to classify all of the remaining species

- Inaccurate assignments of the **classification** algorithm (may be improved by **long-read** amplicon sequencing)



	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	5309	97	4920	147
OTU 2	138	23	100	73
OTU 3	65	2455	27	2505
OTU 4	775	90	737	140
OTU 5	1	90	6	140

16S rRNA profiling

1. **PCR Amplification**, sequencing of variable regions in the 16S rRNA gene, quality filtering

2. Classify reads into **operational taxonomic units (OTUs)** based on sequence identity

3. **Assign taxonomy** - compare OTUs to a reference database, built a phylogenetic tree

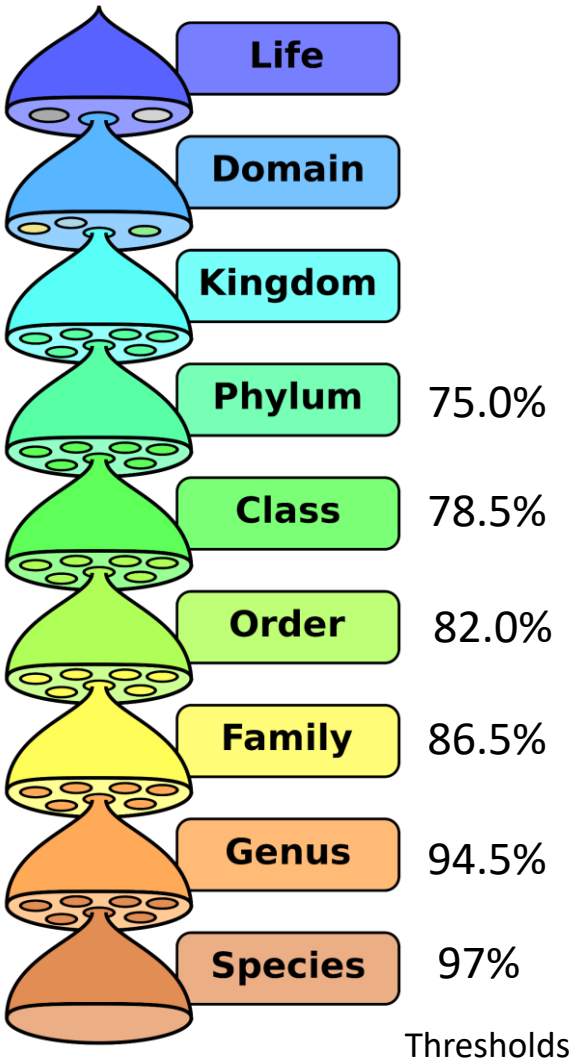
4. Compare diversity between samples

16S rRNA taxonomic classification

OTU table:

		Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__S24-7; g__; s__	5309	97	4920	147
OTU 2	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__; g__; s__	138	23	100	73
OTU 3	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__Allobaculum; s__	65	2455	27	2505
OTU 4	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae	775	90	737	140
OTU 5	Unassigned	1	90	6	140

Taxonomic hierarchy
(criteria are standardized but biologically subjective)

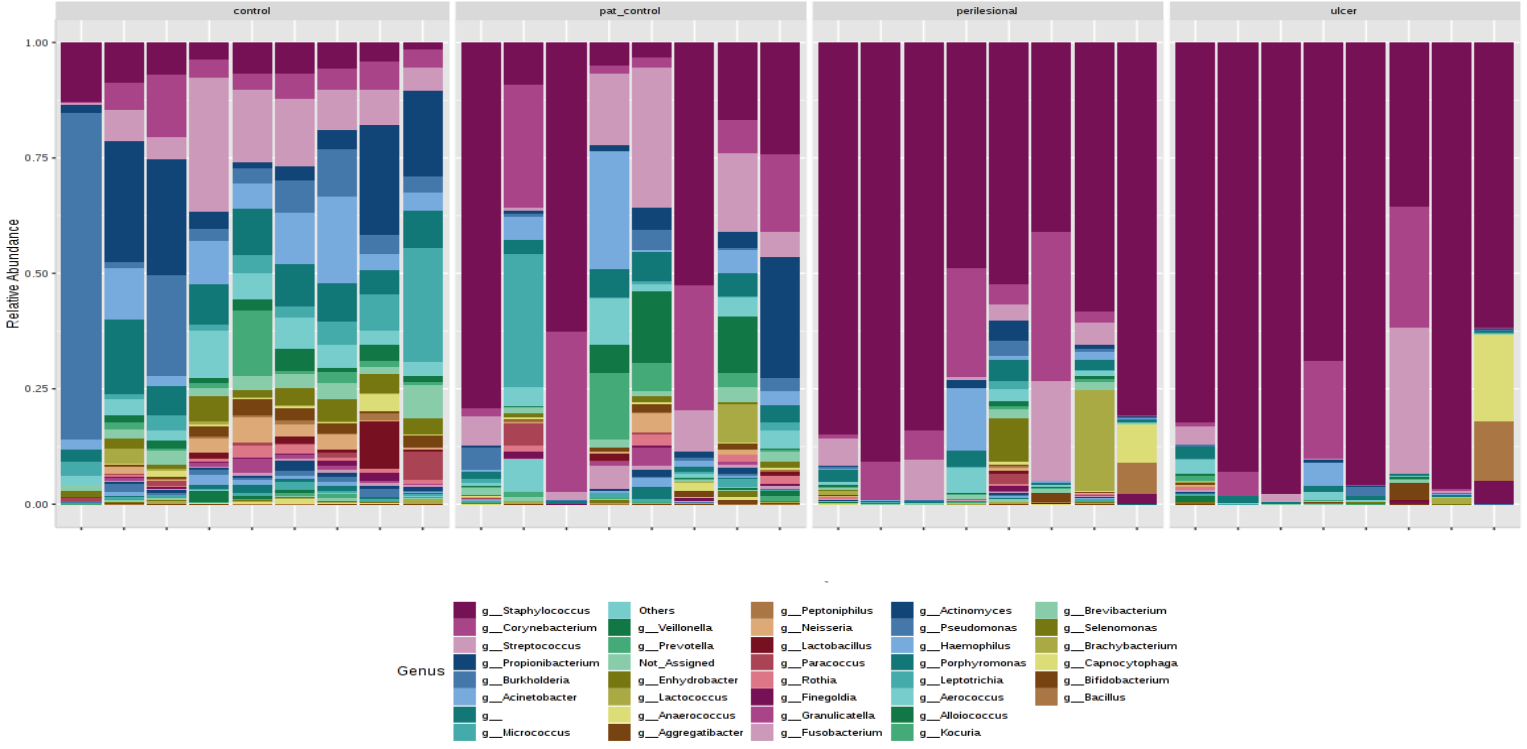


Visualizing taxonomic hierarchical structure

OTU table:

	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	5309	97	4920	147
OTU 2	138	23	100	73
OTU 3	65	2455	27	2505
OTU 4	775	90	737	140
OTU 5	1	90	6	140

Compare *relative* abundances (proportions) of OTUs at genus-level**:**

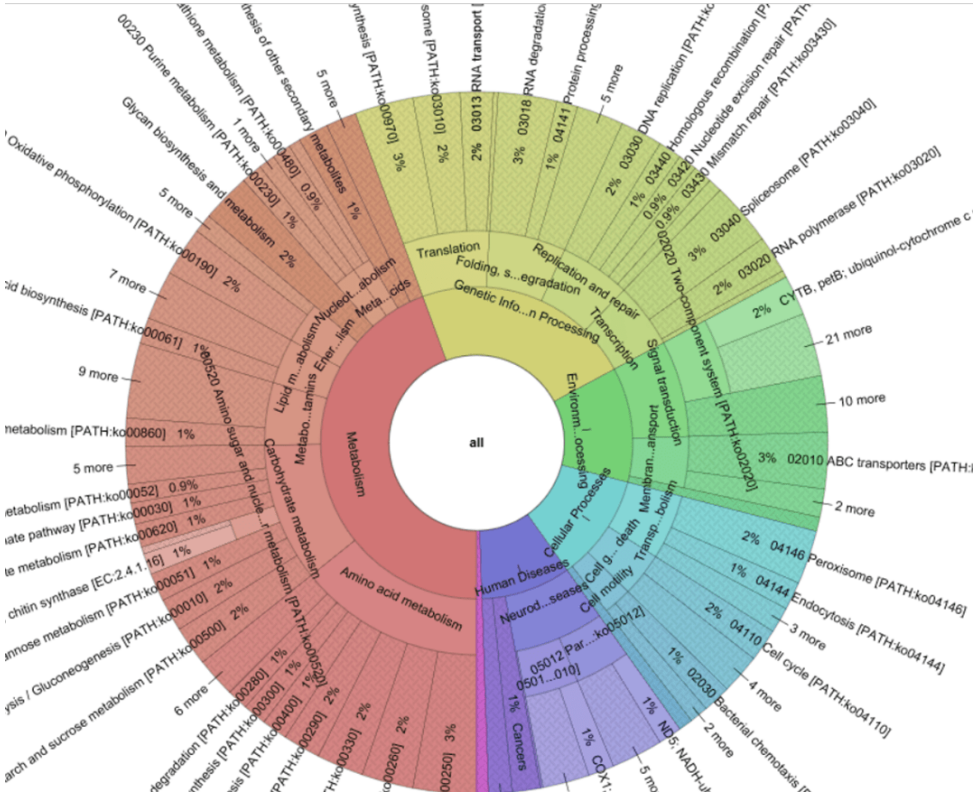


Visualizing taxonomic hierarchical structure

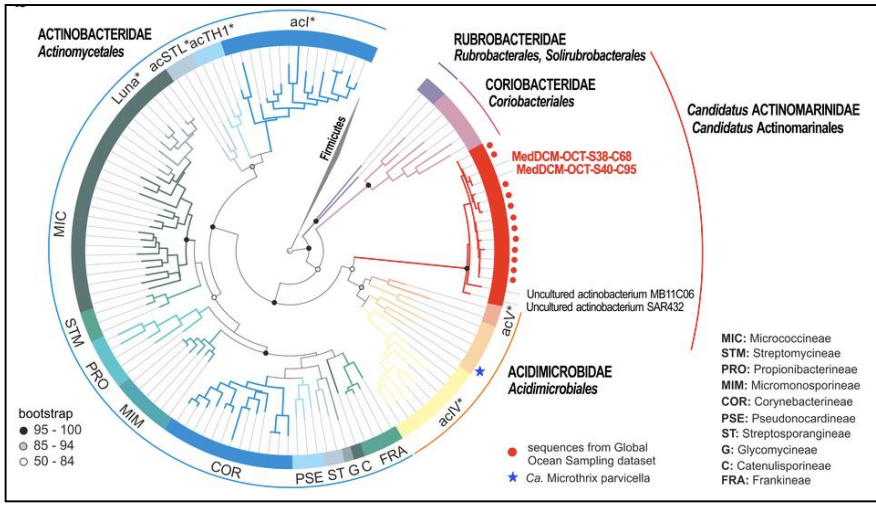
OTU abundance table:

	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	5309	97	4920	147
OTU 2	138	23	100	73
OTU 3	65	2455	27	2505
OTU 4	775	90	737	140
OTU 5	1	90	6	140

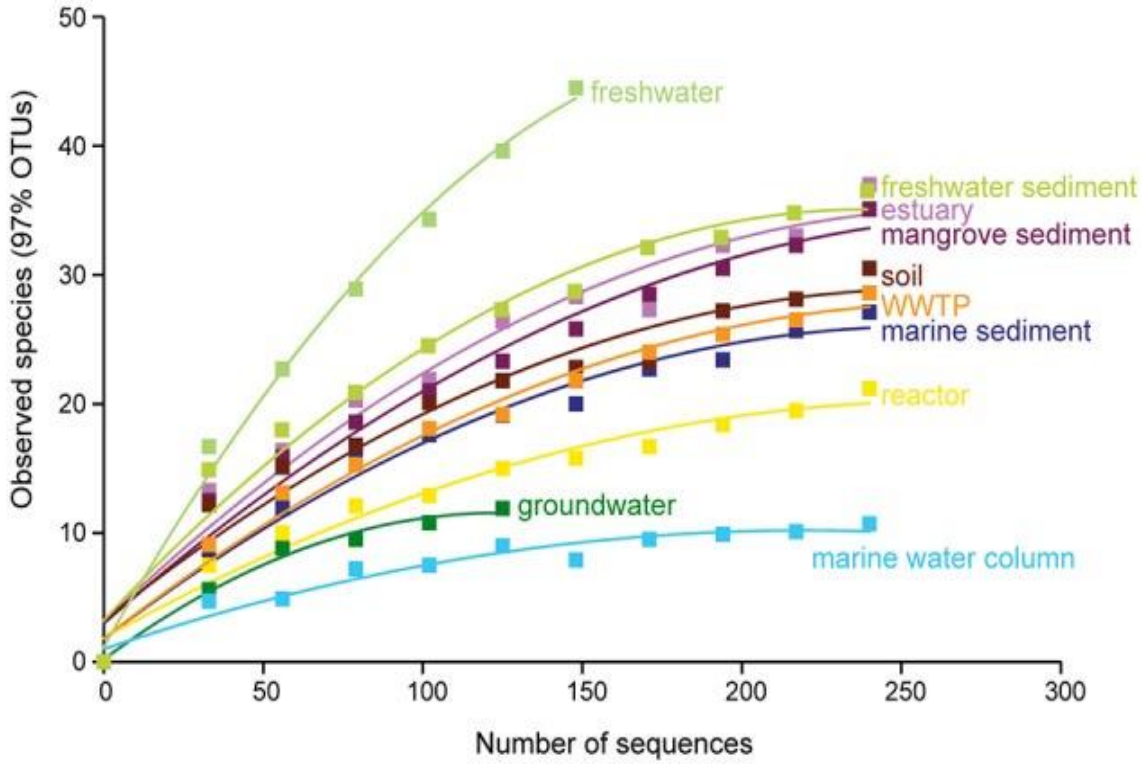
Interactive charts using Krona



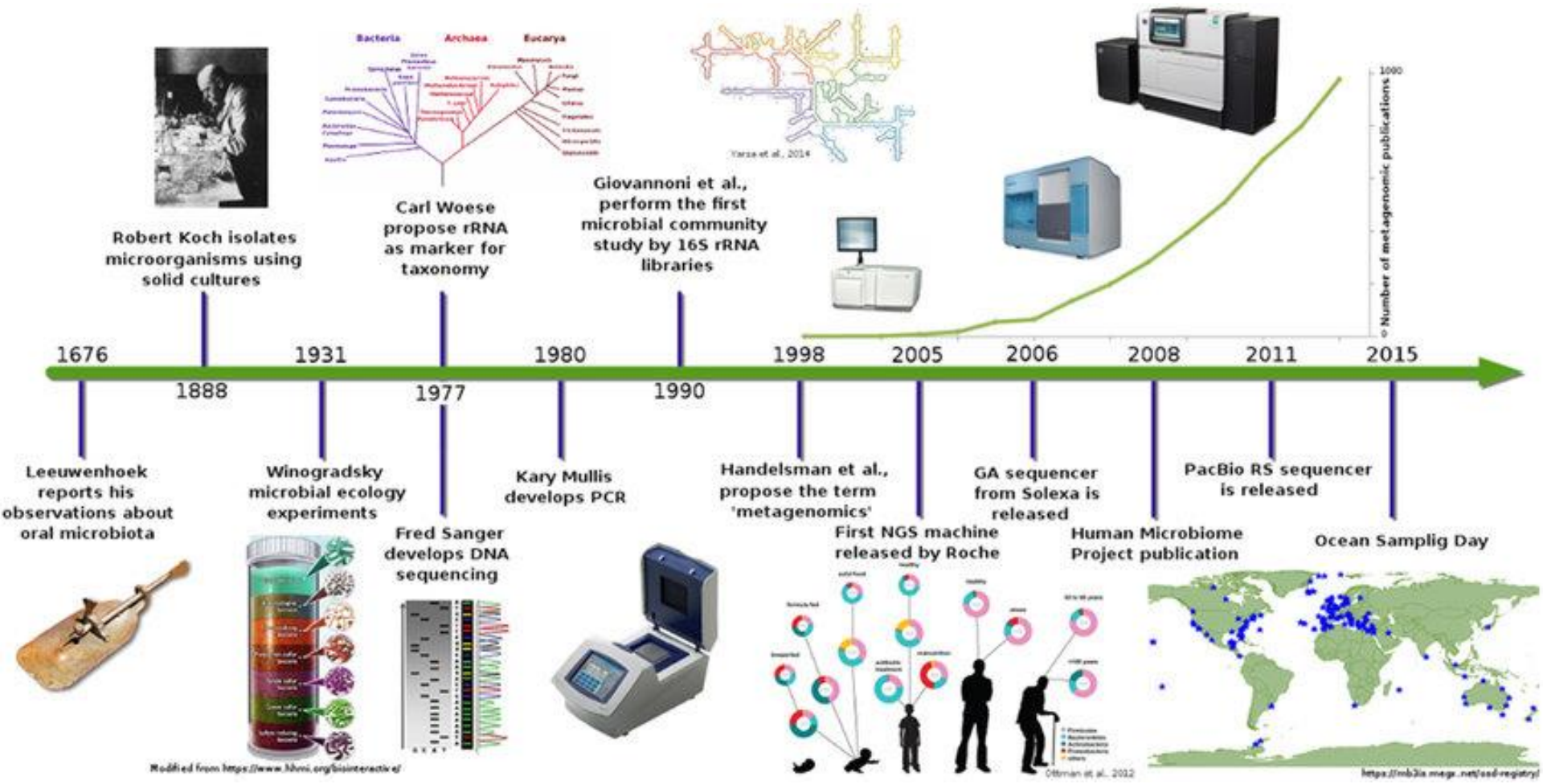
Draw a phylogenetic tree



Estimate species diversity using “Rarefaction Curves”

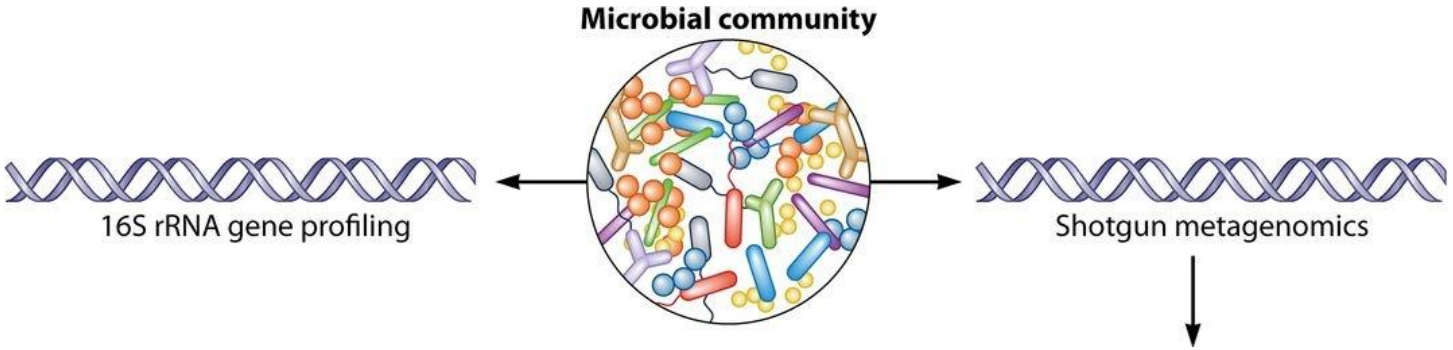


Metagenomics timeline and milestones



https://www.researchgate.net/figure/Metagenomics-timeline-and-milestones-Timeline-showing-advances-in-microbial-communities_fig1_289524171

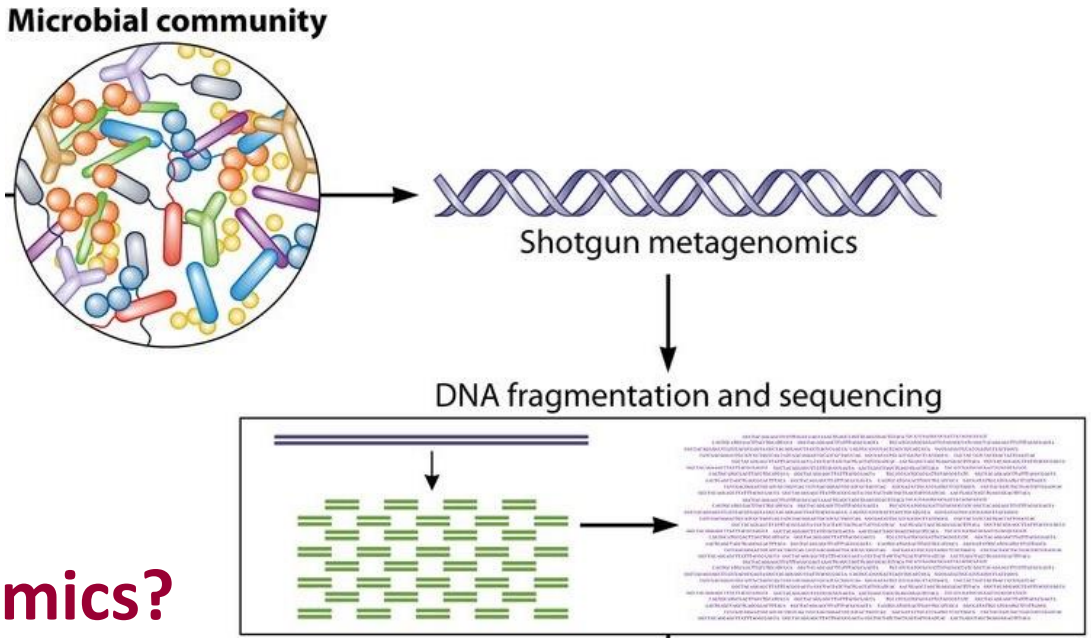
Metagenomics strategies



Shotgun metagenomics

Untargeted (‘shotgun’) sequencing of all (‘meta-’) microbial genomes ‘genomics’ present in a microbial community

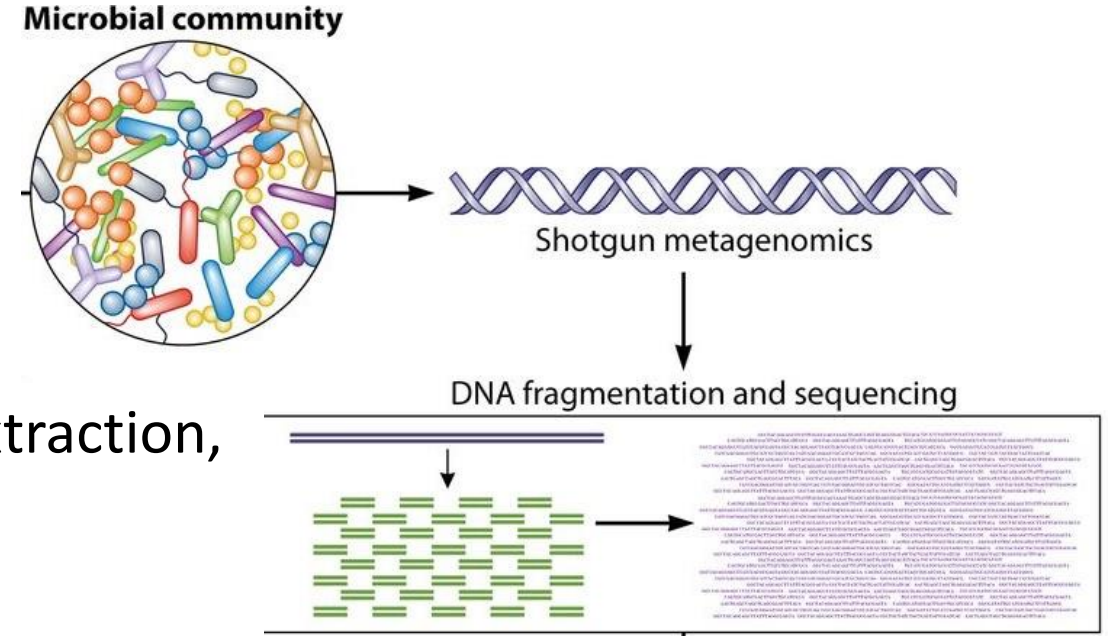
Profiling the **taxonomic** composition and **functional** potential



Why shotgun metagenomics?

- Taxonomic profiling at higher resolution
- Discover new enzymes / pathways
- Antibiotic genes
- Monitor outbreak of human pathogens

Shotgun metagenomics analysis workflow



1. Samples collection, DNA extraction, processing and sequencing

collect sufficient microbial biomass

minimize contamination

Shotgun metagenomics analysis workflow

1. Samples collection, DNA extraction, processing and sequencing

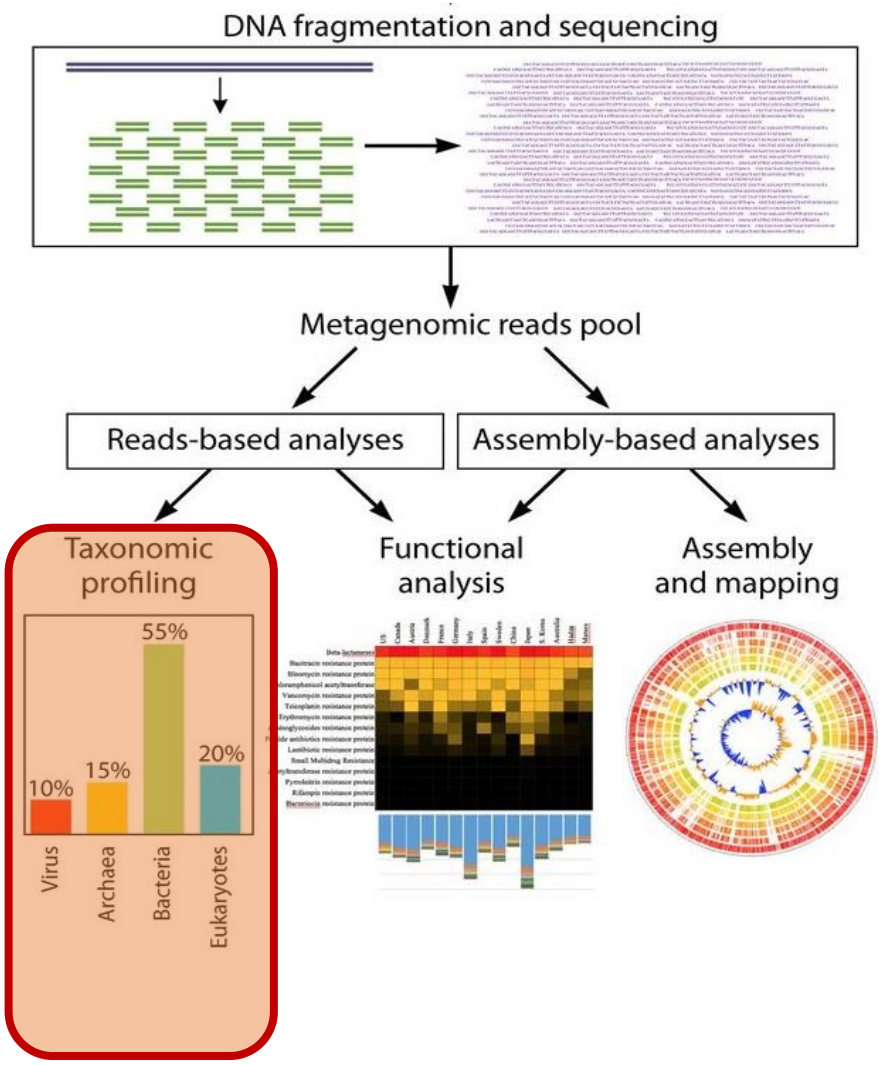
collect sufficient microbial biomass

minimize contamination

2. Library preparation and sequencing

3. Sequence analysis to profile **taxonomic, functional and genomic** features

4. Statistical analysis and validation



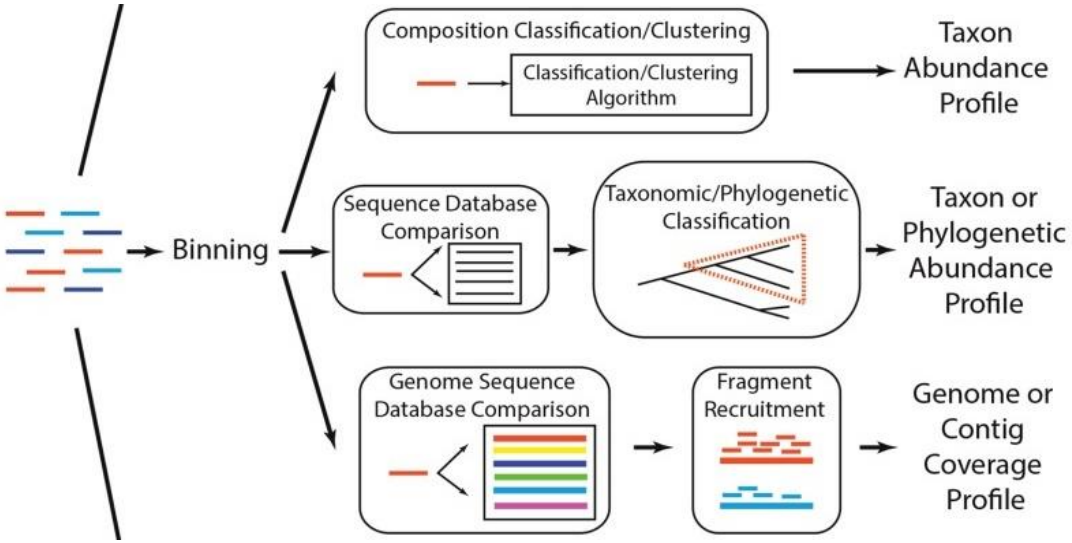
A. Taxonomic binning using shotgun metagenomics

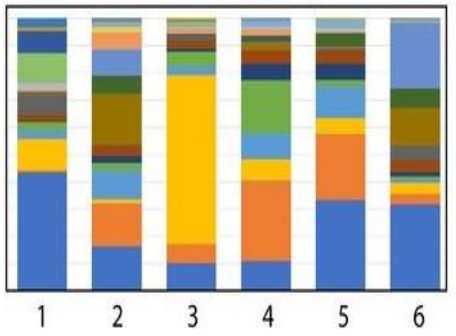
Mapping of reads to known genomes, and counting abundances is *computationally demanding!*

Various algorithms exist, for example:

- Compare to **clade-specific marker genes** as a taxonomic reference
  (i.e MetaPhlAn - based on BLAST, ~17,000 reference genomes)

- Mapping with **k-mer** matching to speed up the computation
  (i.e Kraken)



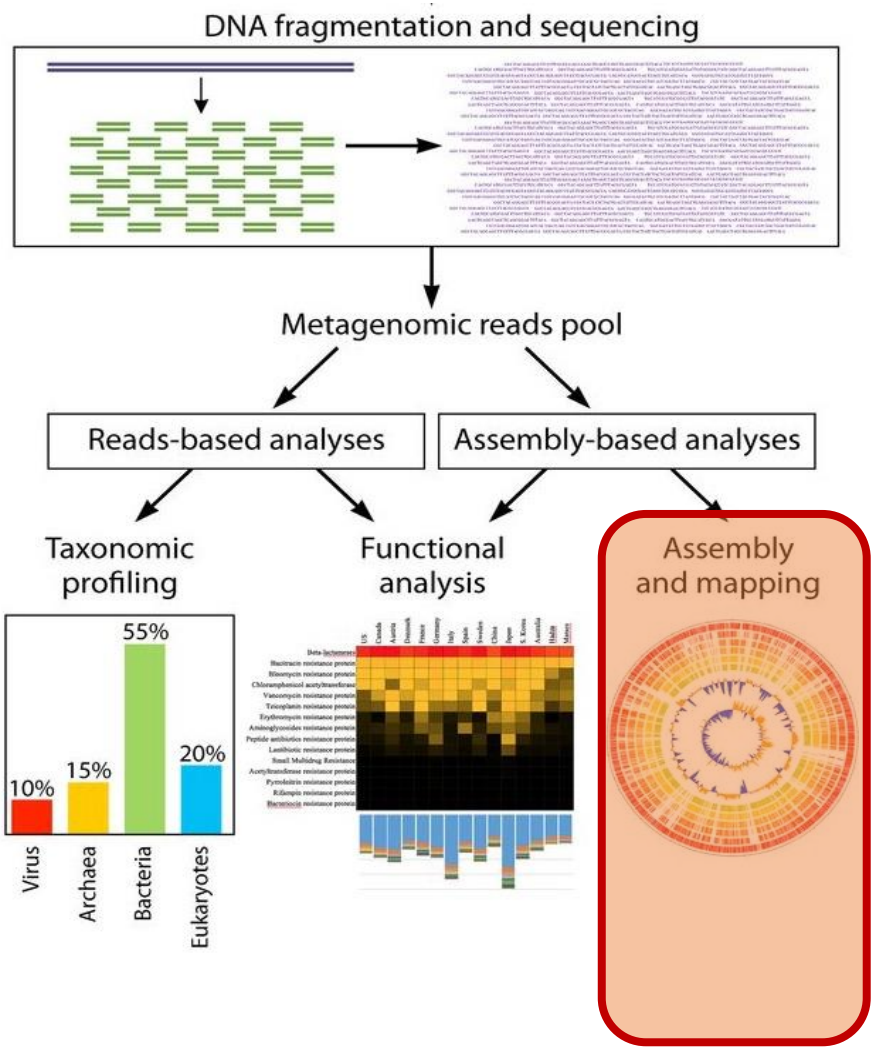


A. Taxonomic binning using shotgun metagenomics

Limitation:

- Accuracy of methods depends on the reference genome
- Non-equal sampling of genomes

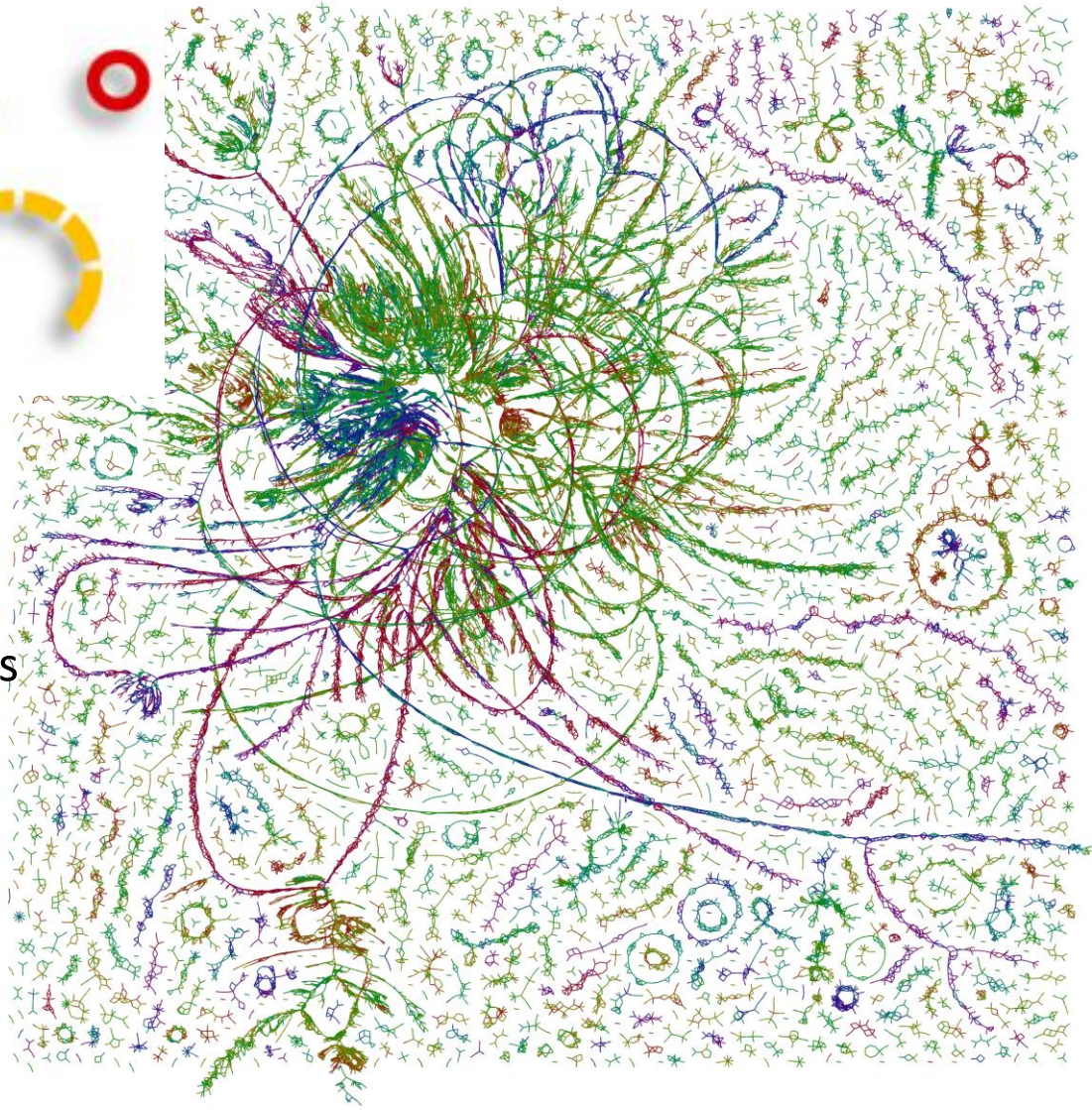
Shotgun metagenomics analysis workflow



B. Metagenome assembly

Reconstruction of contigs from mixed reads from different species

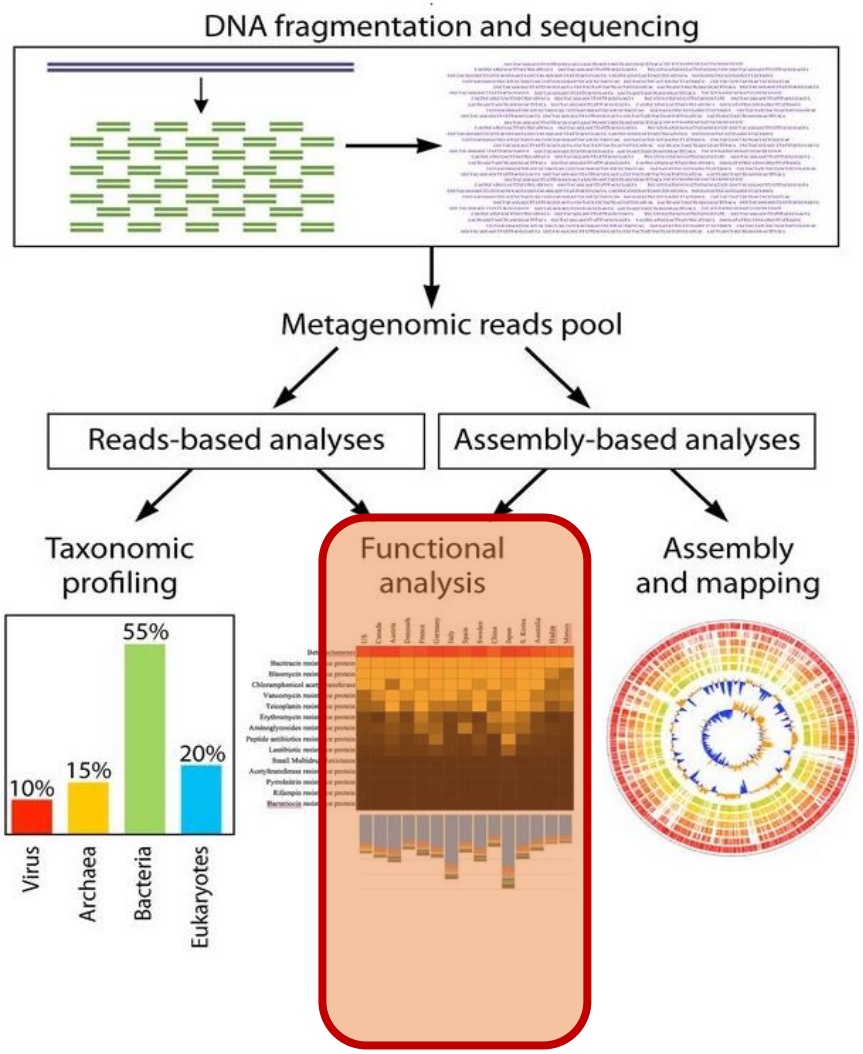
Using a reference genome, or
De novo assembly



- Complicated by sequencing errors and repetitive sequence
- Challenge in assembling closely related and low-abundance organisms
- Require sufficient genome coverage (>20×)

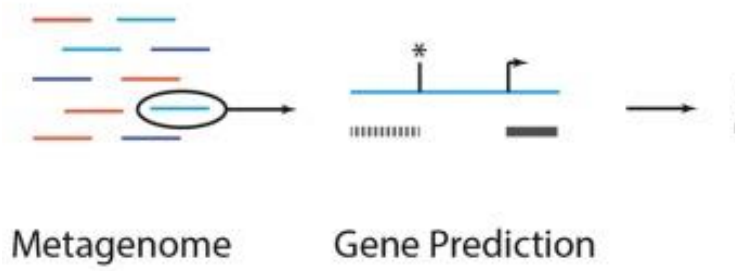
Shotgun metagenomics analysis workflow

Predict the **functional potential** of microbial communities using closest relatives references



C. Functional assignment using shotgun metagenomics

Step 1: **Predict coding sequences**



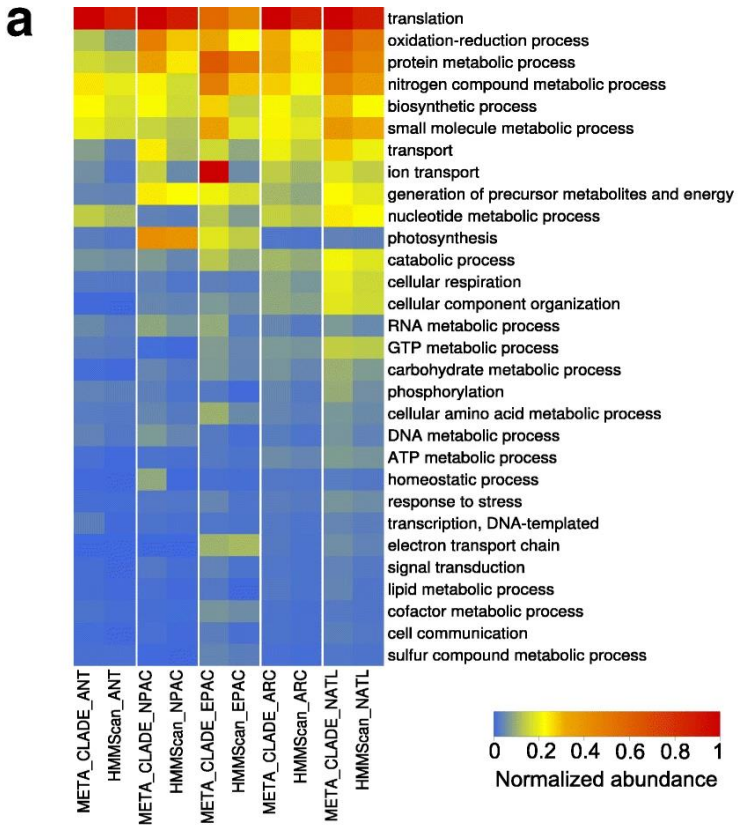
by either:

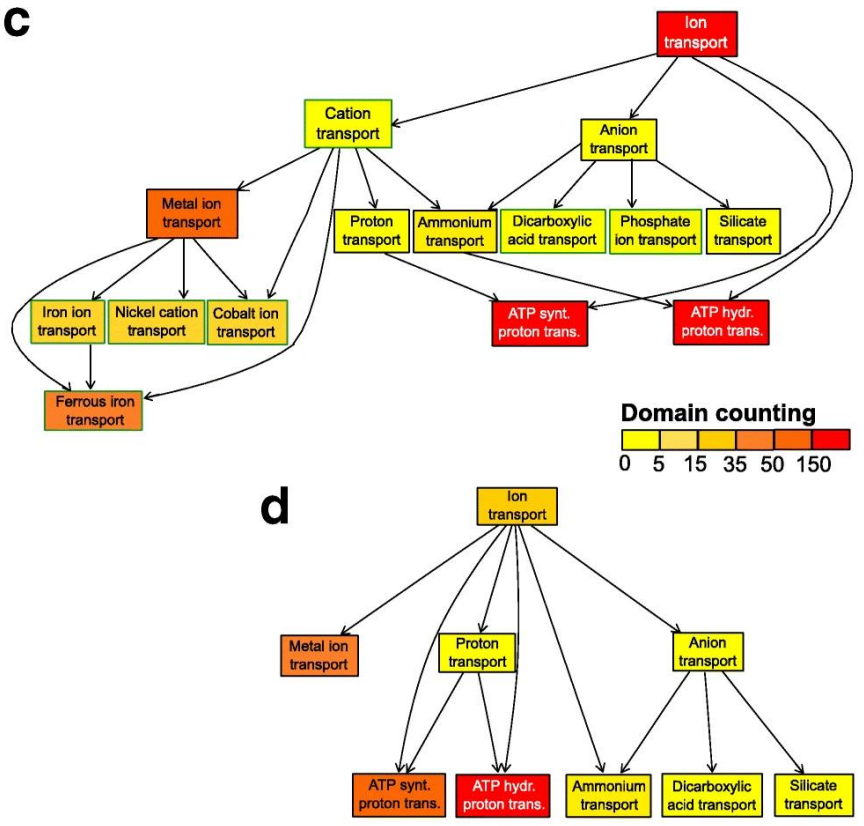
1. Mapping reads to a sequence database
2. Translating each read into all six possible protein coding frames

C. Functional assignment using shotgun metagenomics

Step 2: **Compare** the resulting peptides to a protein database
Using **sequence** or **motif**-based (Homology modeling) databases

Commonly used databases: KEGG, MetaCyc, EggNOG, Pfam, SEED, Phylofacts, UniProt, HUMAnN pipeline





Challenges in inferring genes / metabolic pathways

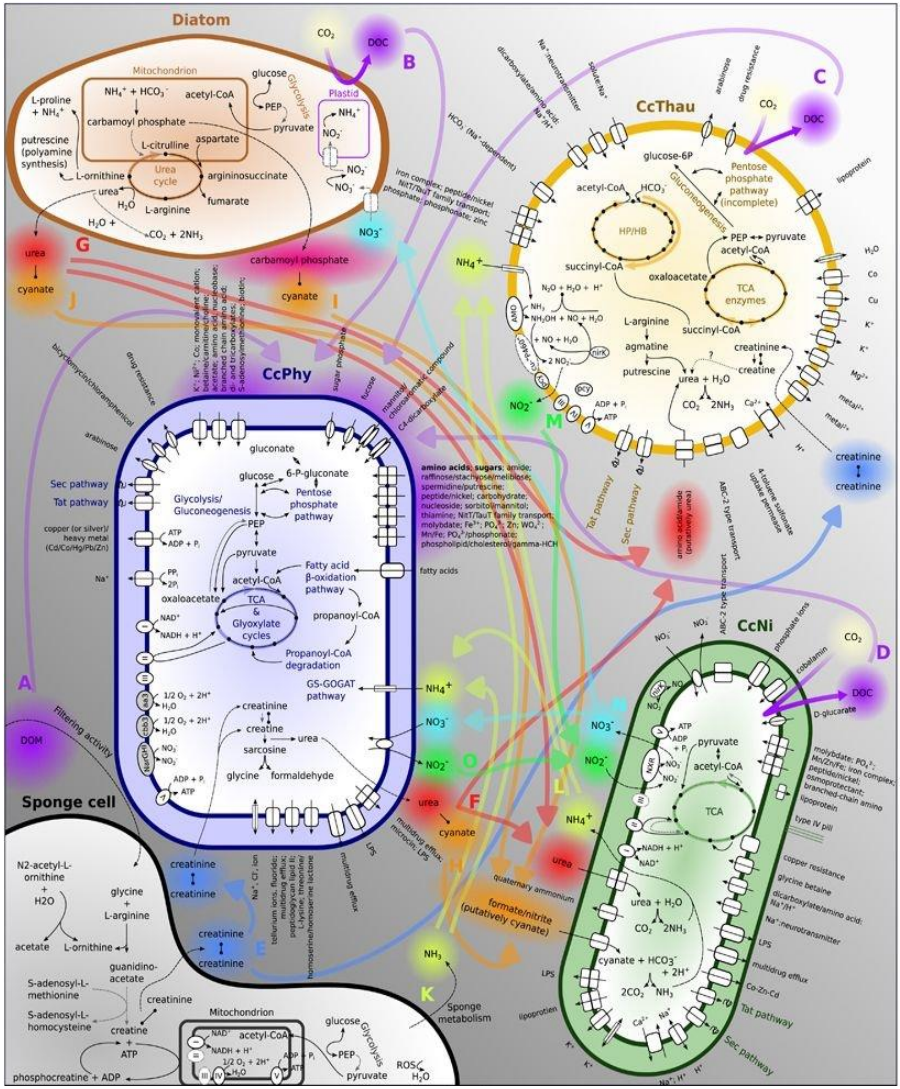
- **Biased** protein databases:

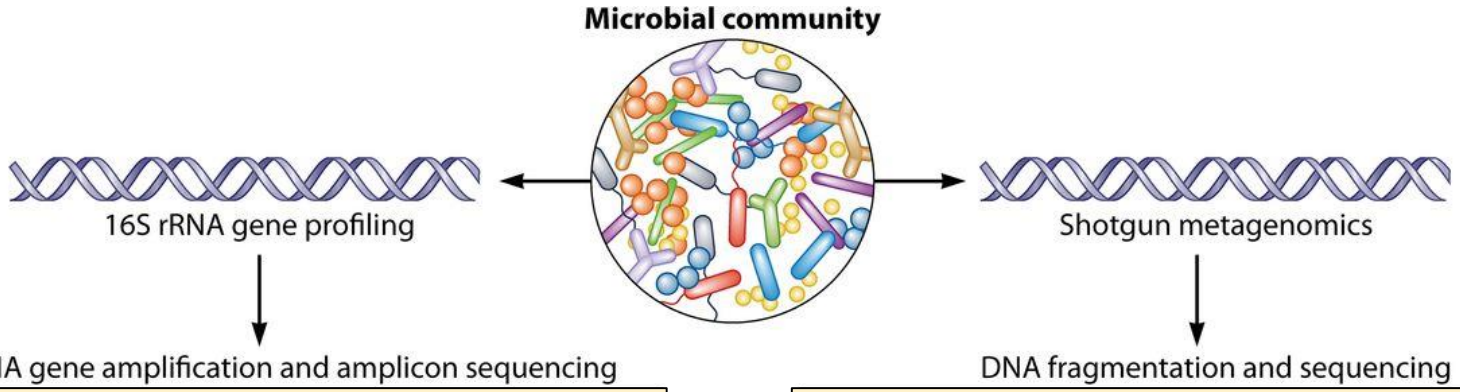
Highly conserved pathways and housekeeping functions are more represented in the databases

- The presence of a gene does not mean that it is expressed

(complemented with **meta-transcriptomics** and meta-proteomics)

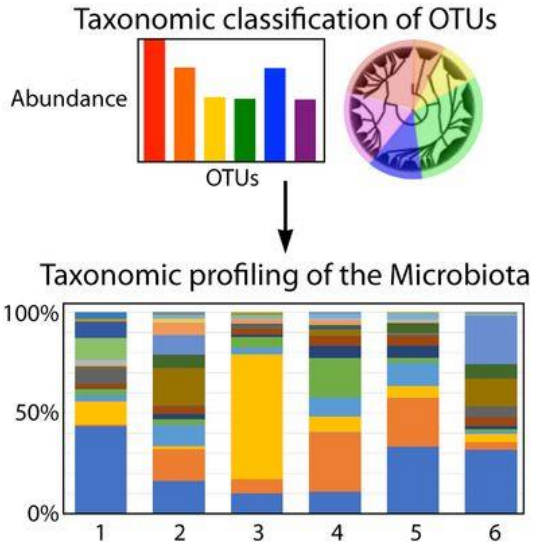
- May be improved with **long-read** sequencing platforms

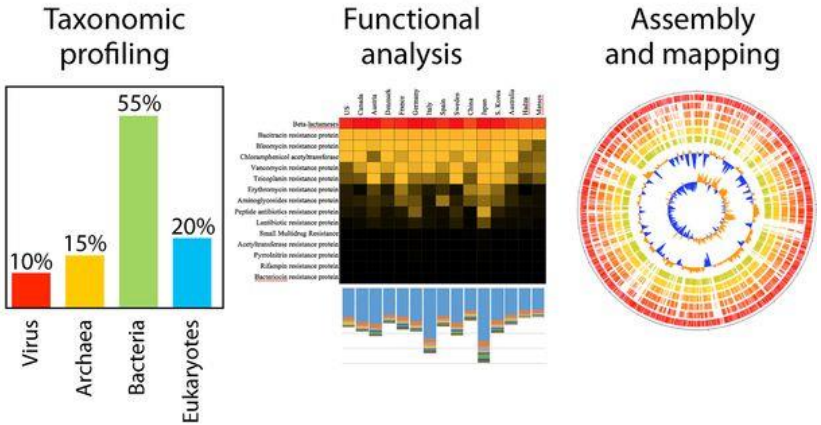




Less sequencing –> less expensive
Limited taxonomic resolution
Compliant with low biomass
Does not reveal functional pathways

More sequencing –> more expensive
Can identify novel genes/genomes
Can reveal functional potential
High species resolution
Requires high biomass
Memory intensive





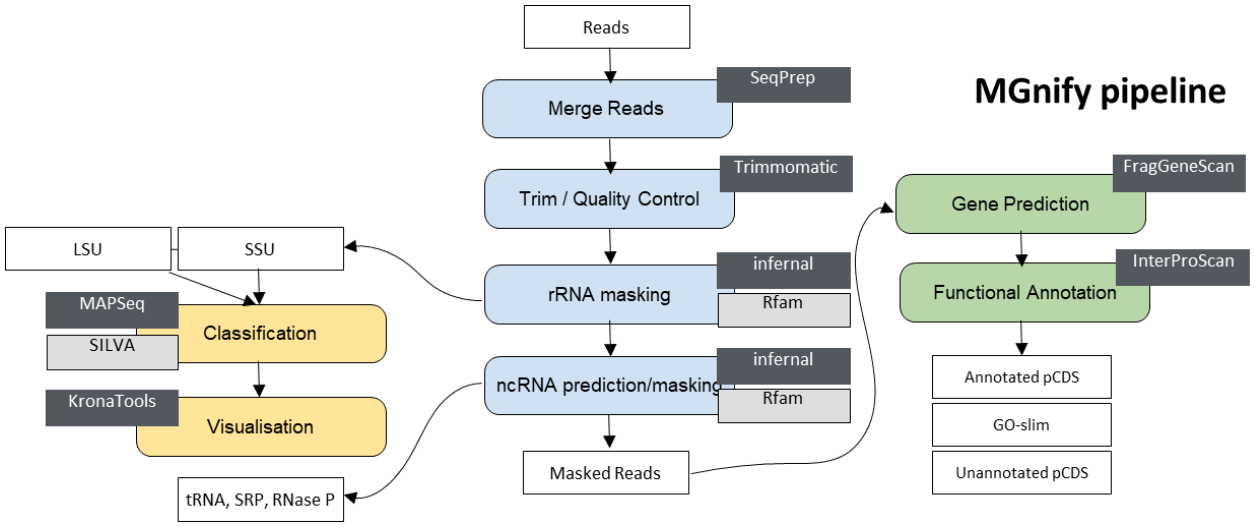
Metagenomic resources

- Metagenomic analysis require **specialized tools** and algorithms and substantial computing **resources**
- Public web pipelines provide generic analysis via **standardized** workflows

img

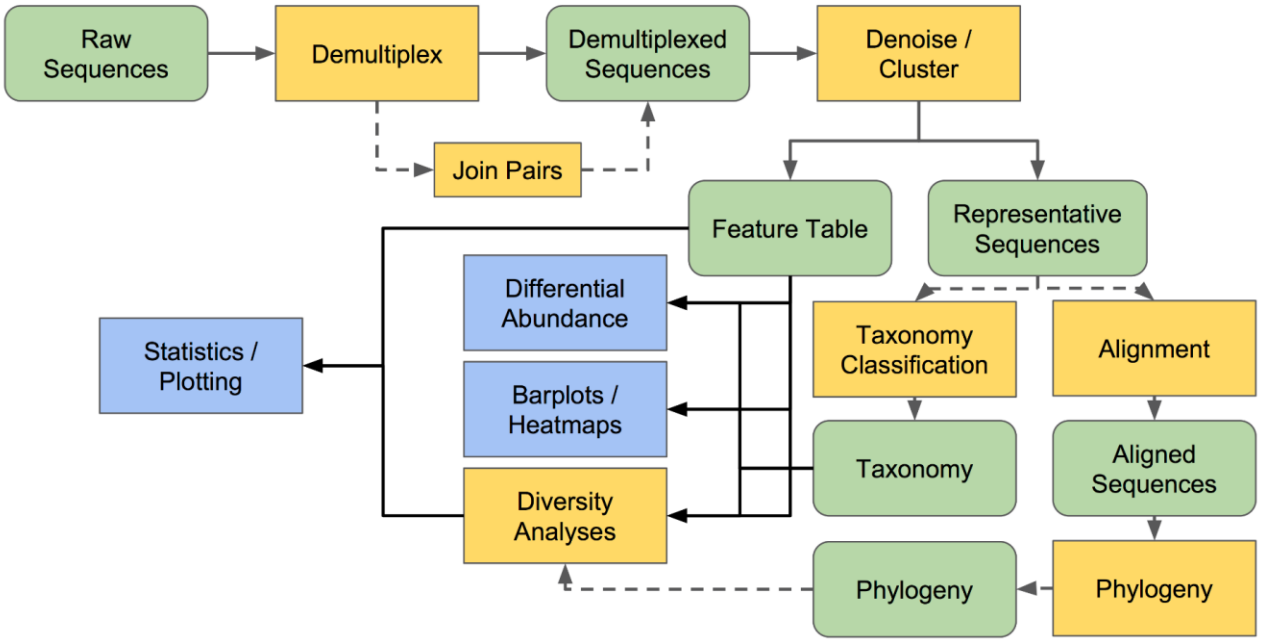
MG-RAST
metagenomics analysis server

MGnify

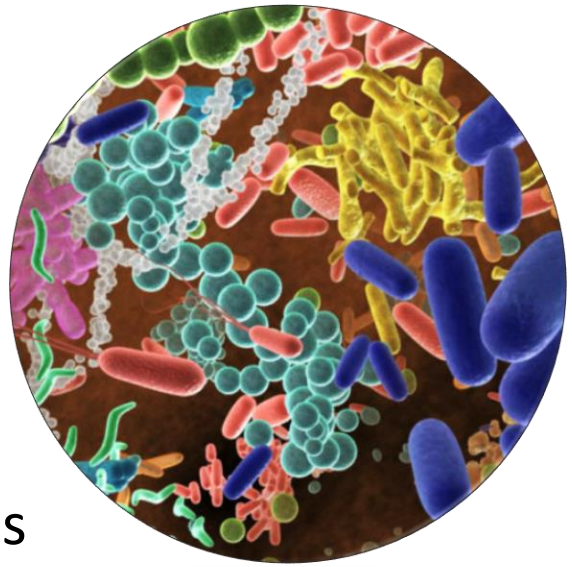


Available suites of computational pipelines

16S analysis workflow with **QIIME** 2



Summary



Metagenomics explores **complex** microbial communities **which cannot be cultured**

16S rRNA profiling detects the **taxonomic** composition and relative abundances of species

Shotgun sequencing explores in higher resolution **taxonomic** and **functional** diversity of microbial communities

Various bioinformatic tools deal with potential **experimental biases** and the try to reduce the **complexity of computational** analyses

For additional reading:

Shotgun metagenomics, from sampling to analysis
https://www.nature.com/articles/nbt.3935.pdf

An introduction to the analysis of shotgun metagenomic data
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4059276

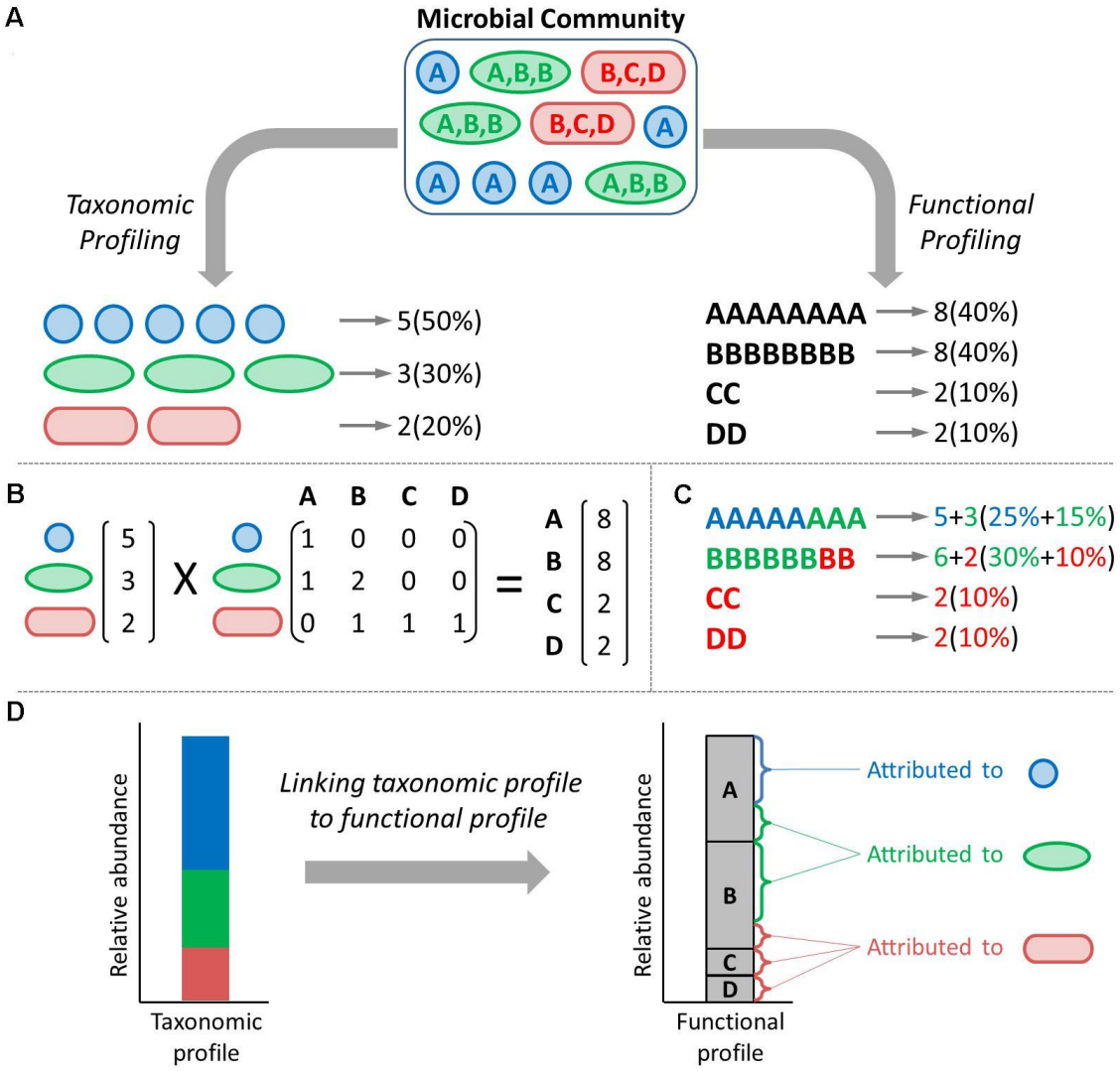
Jan 17th: NO CLASS

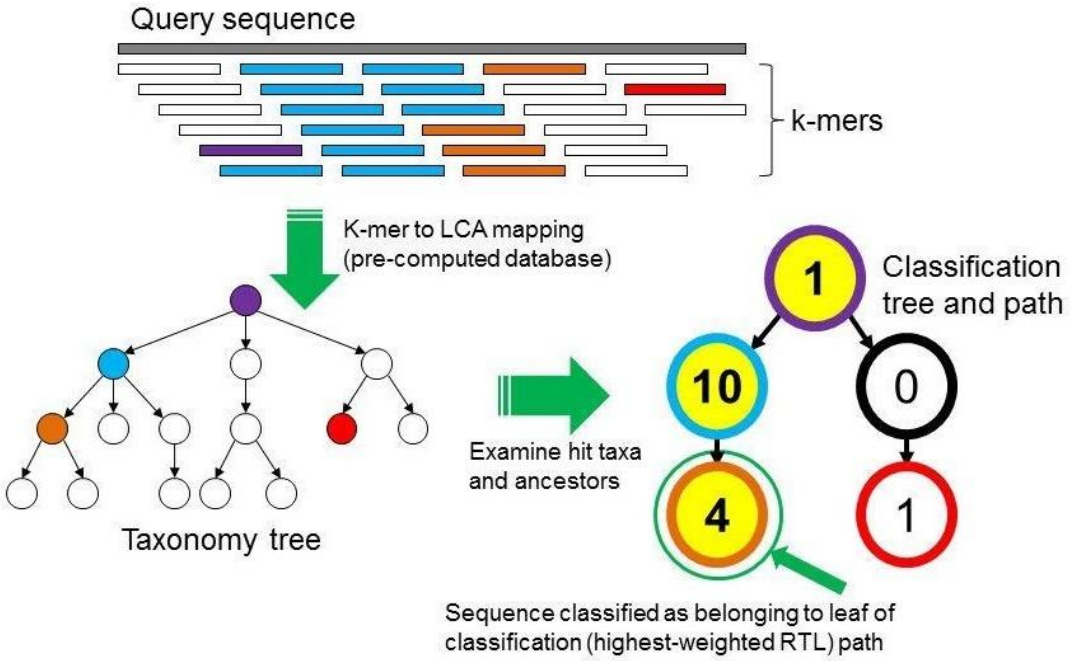
Jan 24th (10:15-12:00):
Example questions for the exam + summary (zoom)
Please send us questions!

Feb 24th (10:15-12:00):
(optional) **Open class for Q&A** (zoom)
Please send us questions ahead

Exams (Feinberg room B):
Schedule A: **27/02/2022**
Schedule B: **16/03/2022**

BURRITO: An Interactive Multi-Omic Tool for Visualizing Taxa–Function Relationships in Microbiome Data




Query sequence
k-mers
K-mer to LCA mapping
(pre-computed database)
Taxonomy tree
Examine hit taxa
and ancestors
1
Classification
tree and path
10
0
4
1
Sequence classified as belonging to leaf of
classification (highest-weighted RTL) path