

What is metagenomics and why is it a big deal?



BIONOMICS-MMlab
MULTIPARTITE PARASITIC INTERACTIONS

Maliheh Mehrshad

Department of Aquatic Sciences and Assessment,



Who am I?





BIONOMICS-MMLab
MULTIPARTITE PARASITIC INTERACTIONS

www.bionomics-mmlab.com

Pushing the frontiers of multipartite parasitic interactions research

Current bionomics-mmlab members

Zahra Goodarzi – PhD student

Dr. Vinicius Silva Kavagutti – PostDoc

Dr. Vesna Grujcic – PostDoc

Lauren Davies – PhD student

Former lab members

Maryan Resaei-some – PostDoc

Fanny Persson – Field assistant

Armand Stoe – MSc student

Remco Hoogervorst – Internship student

Ezhilarasan Mani Ezhilan – MSc student



What is metagenomics and why is it a big deal?



BIONOMICS-MMlab
MULTIPARTITE PARASITIC INTERACTIONS

Maliheh Mehrshad

Department of Aquatic Sciences and Assessment,



Let's answer some questions on menti!

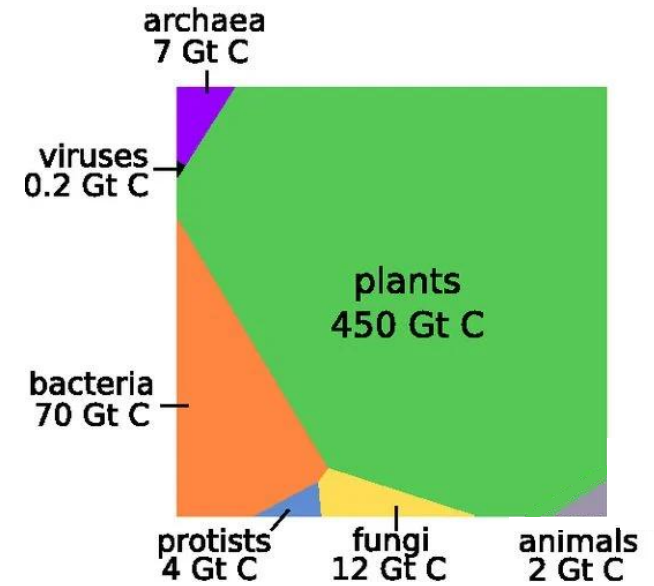
- Join at menti.com
- Use code 74440277

Microbes Matter!

Life on Earth has been microscopic for much of its ~4 billion year history.

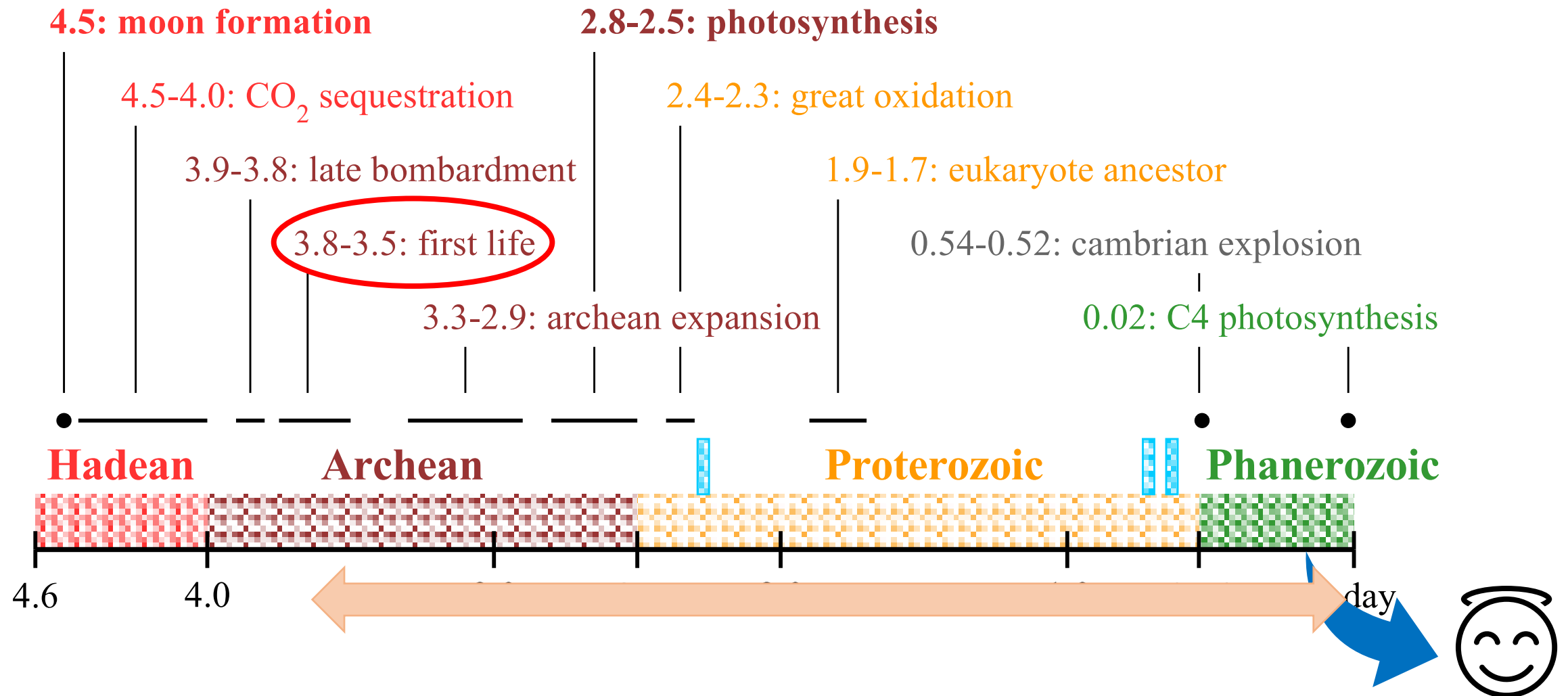
The metabolic activity of these organisms has left its mark.

- Great oxygenation event
- Photosynthesis
- Lignin and cellulose degradation
- Cycling of elements (C, N, S, Fe, ...)
- Greenhouse gas sink/emission
- Interact with plants and animals
- ...

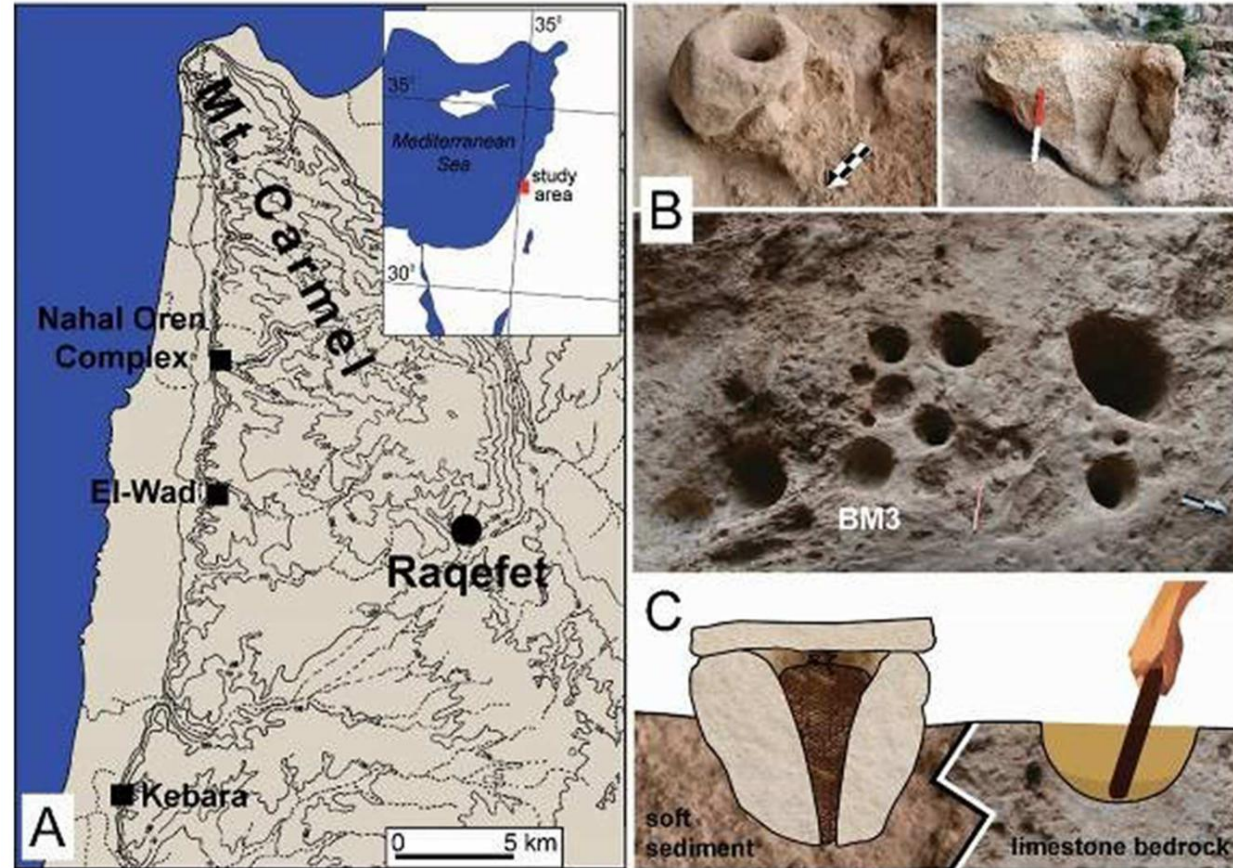


Bar-On *et.al.* PNAS, 2018

Life on the planetary timescale

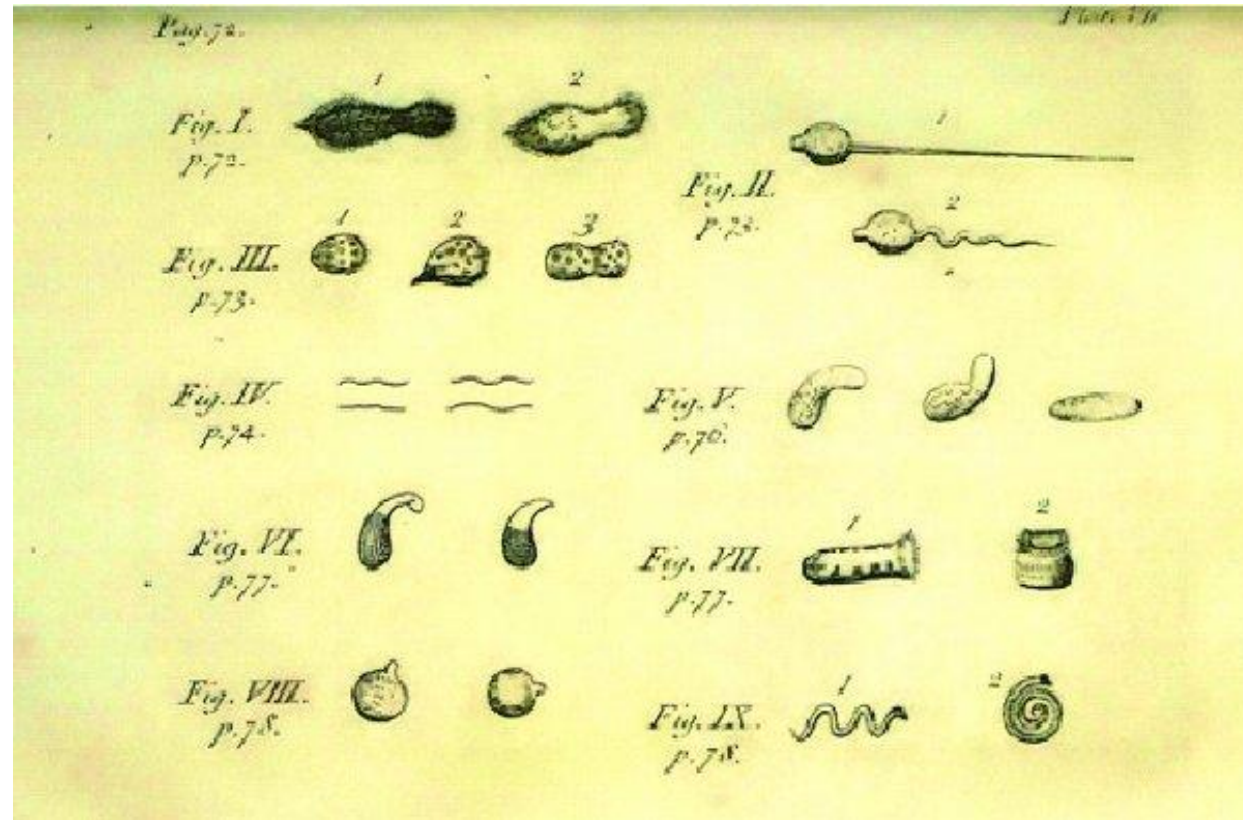


We knew microbial ecology before knowing microbes



Fermented beverage and food storage in 13,000 year-old stone mortars

The Unseen Majority



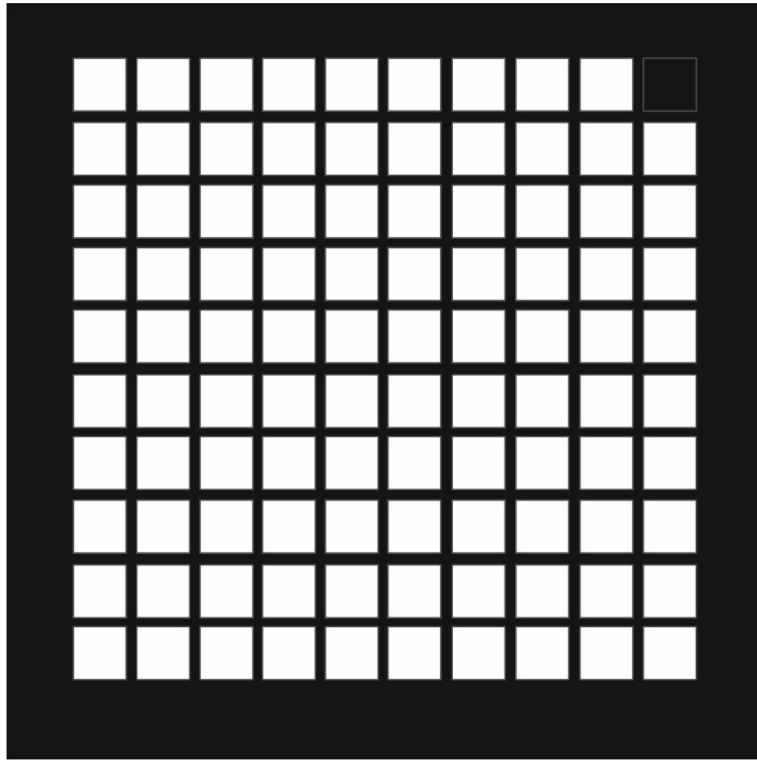
1676

The Unseen Majority: Great plate count anomaly

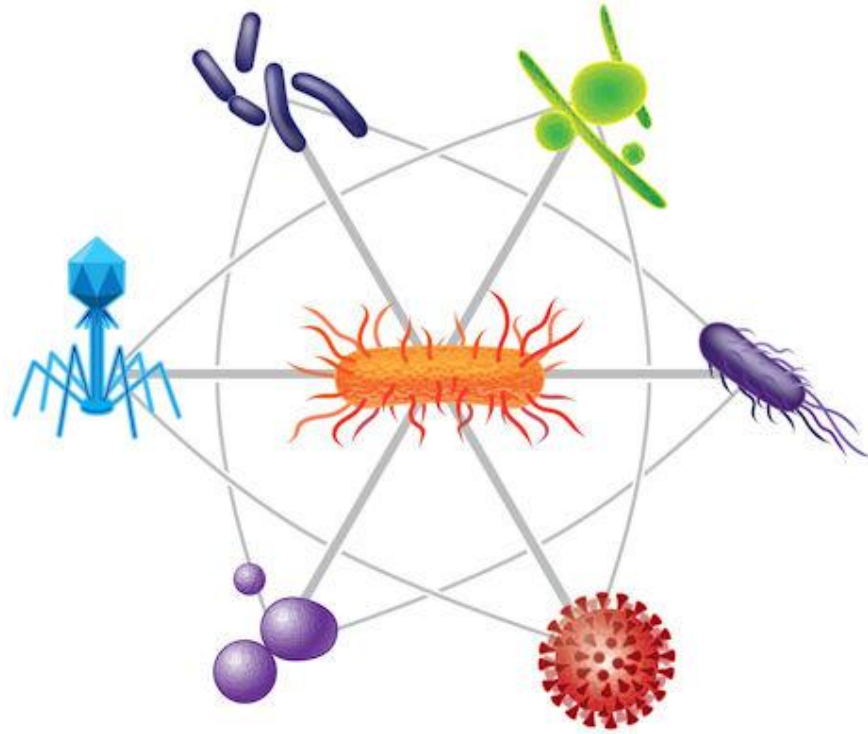


1985

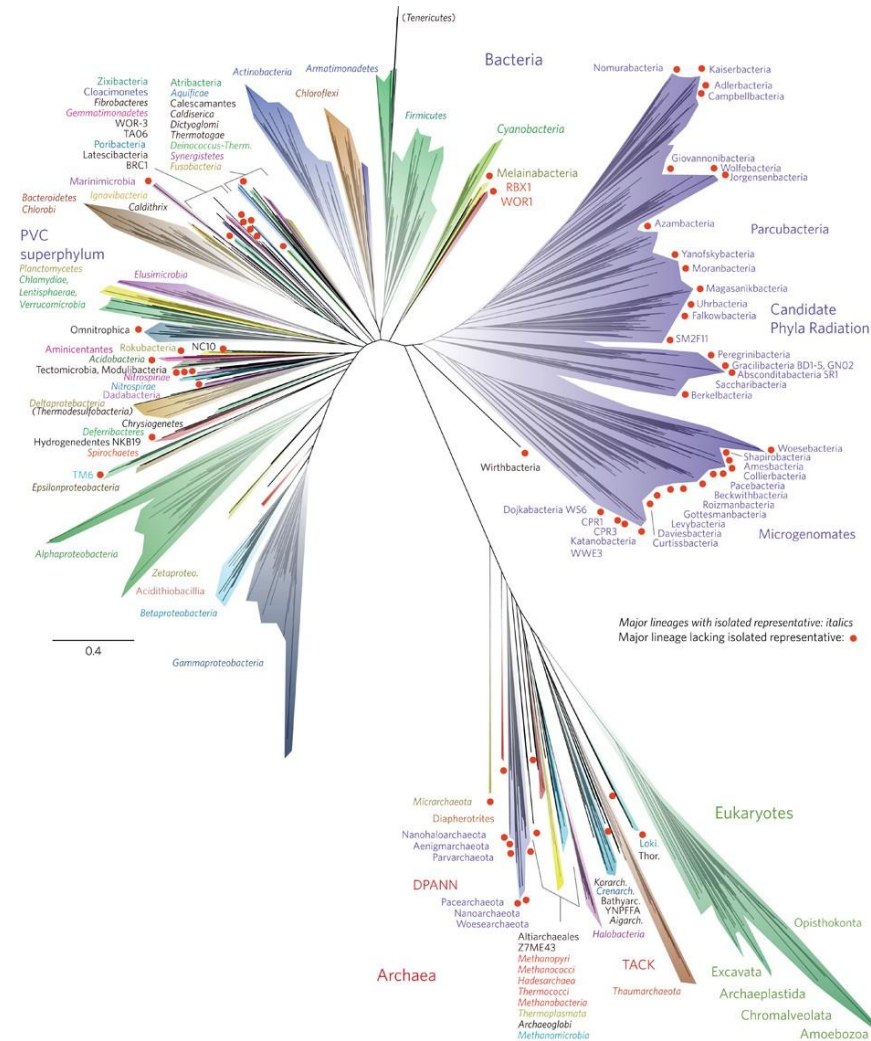
The Unseen Majority: Great plate count anomaly



1985



The Unseen Majority: metagenomics revolution



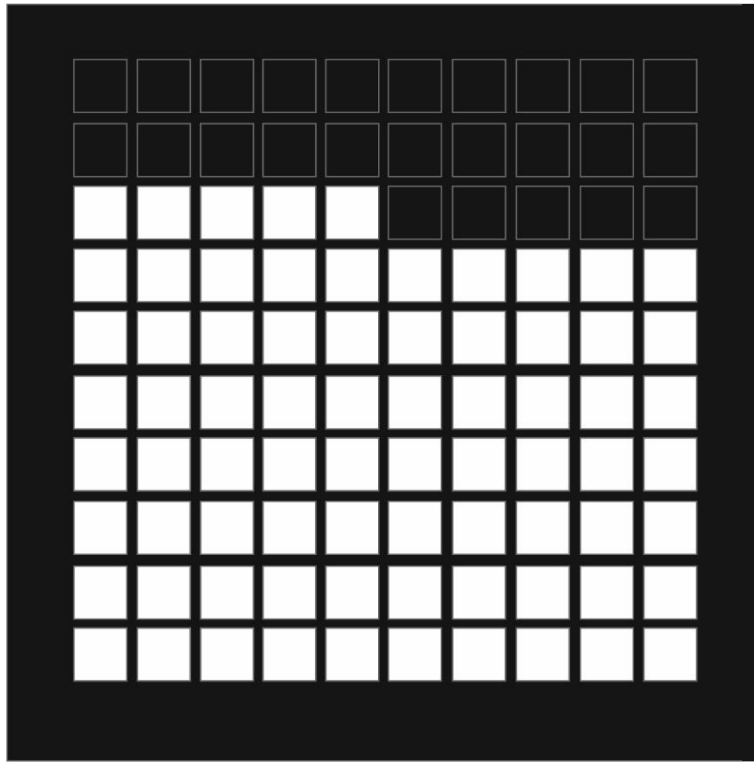
Hug et al., *Nat microbiol*, 2016

The Unseen Majority: metagenomics revolution

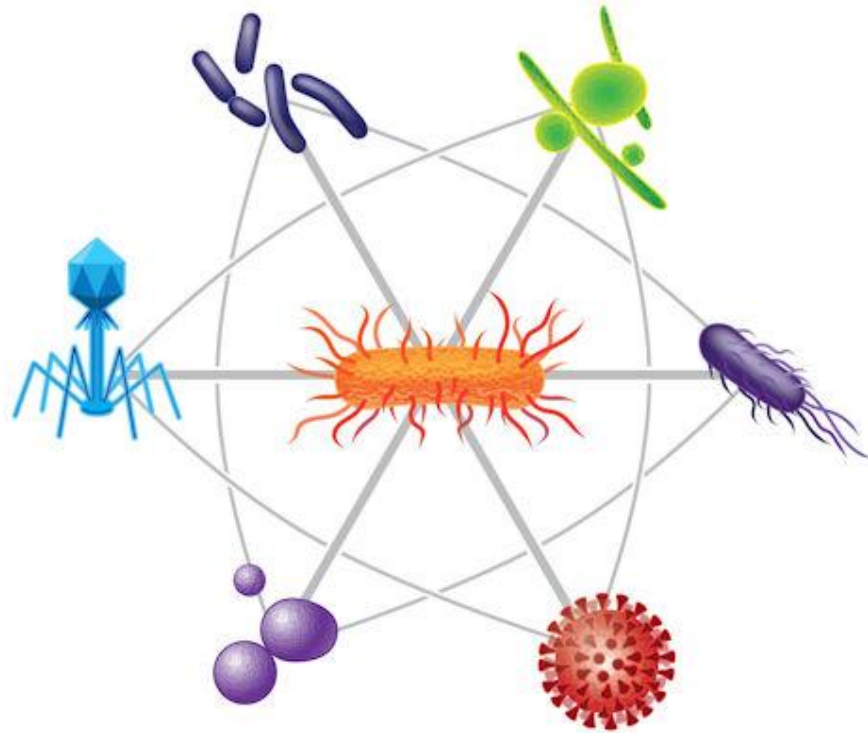


Now

The Unseen Majority: metagenomics revolution



Now



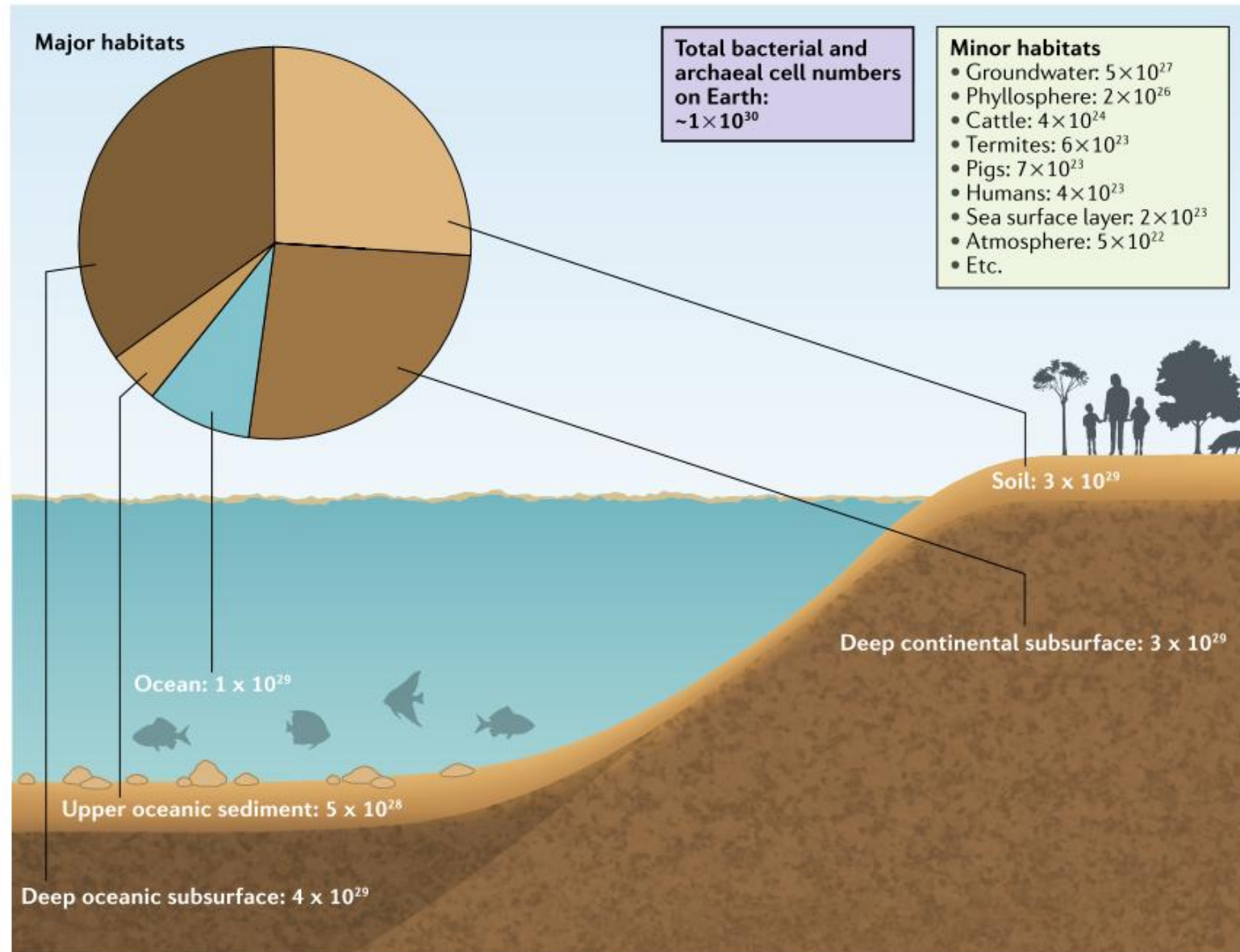
This enormous biomass is distributed in microscopic cells



$\sim 1.2 \times 10^{30}$ bacterial/archaeal cells exist in the “big five” habitats of Earth

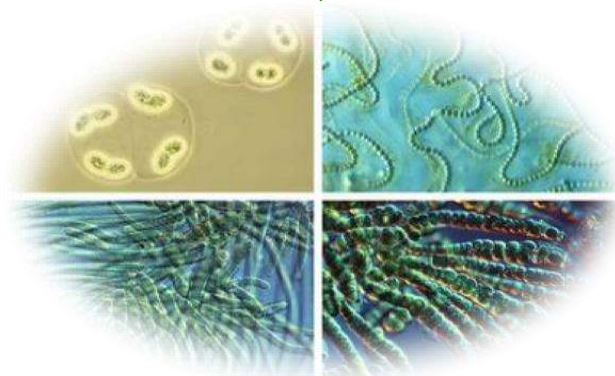
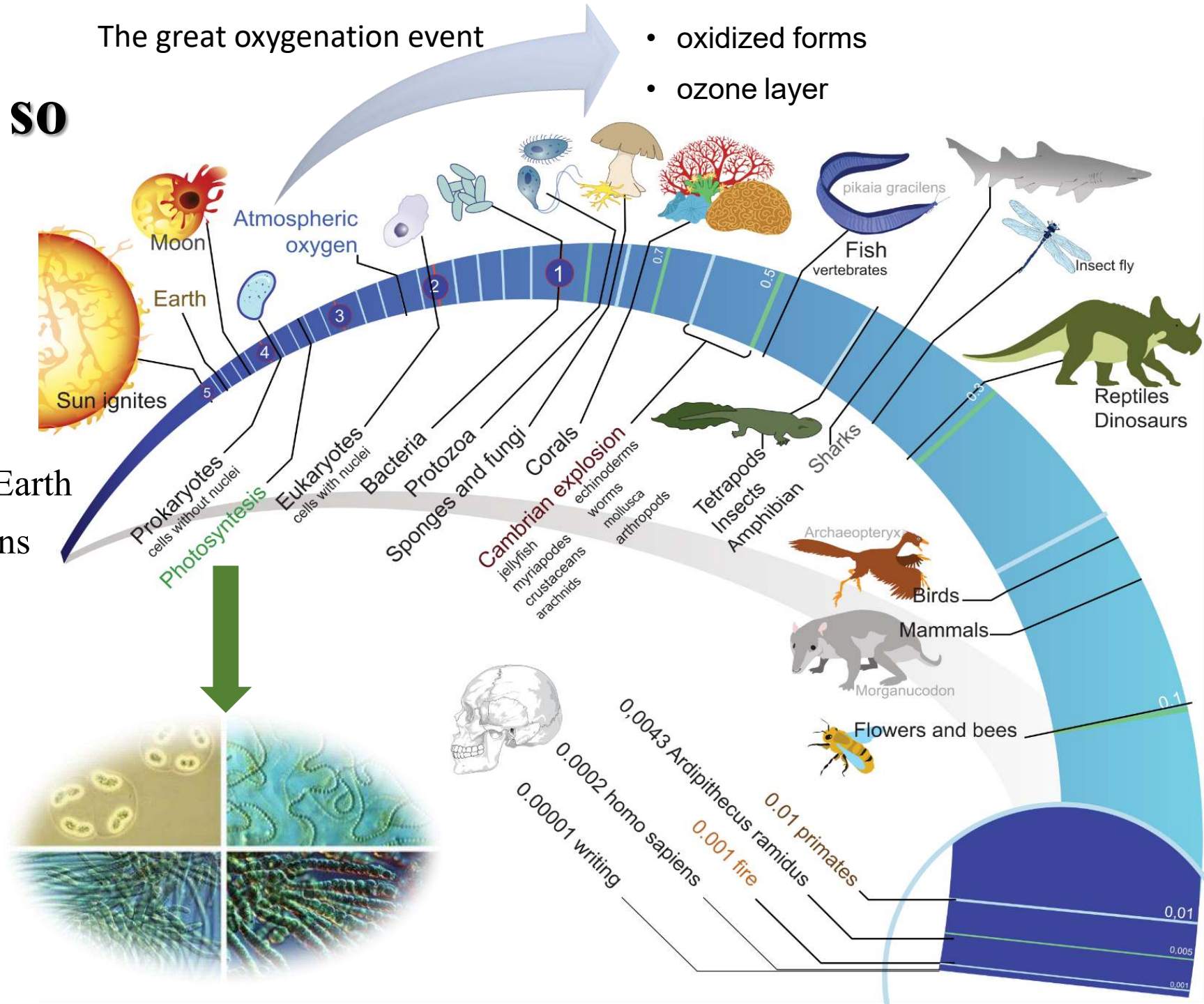
if all the 1×10^{31} viruses on earth were laid end to end, they would stretch for 100 million light years

Astronomical numbers!



Why are microbes so diverse?

- Evolved early
- Initially access to all habitats on Earth
- Survived a large span of conditions
- More habitats, more niches
- Short generation times
- Inter-species gene transfer



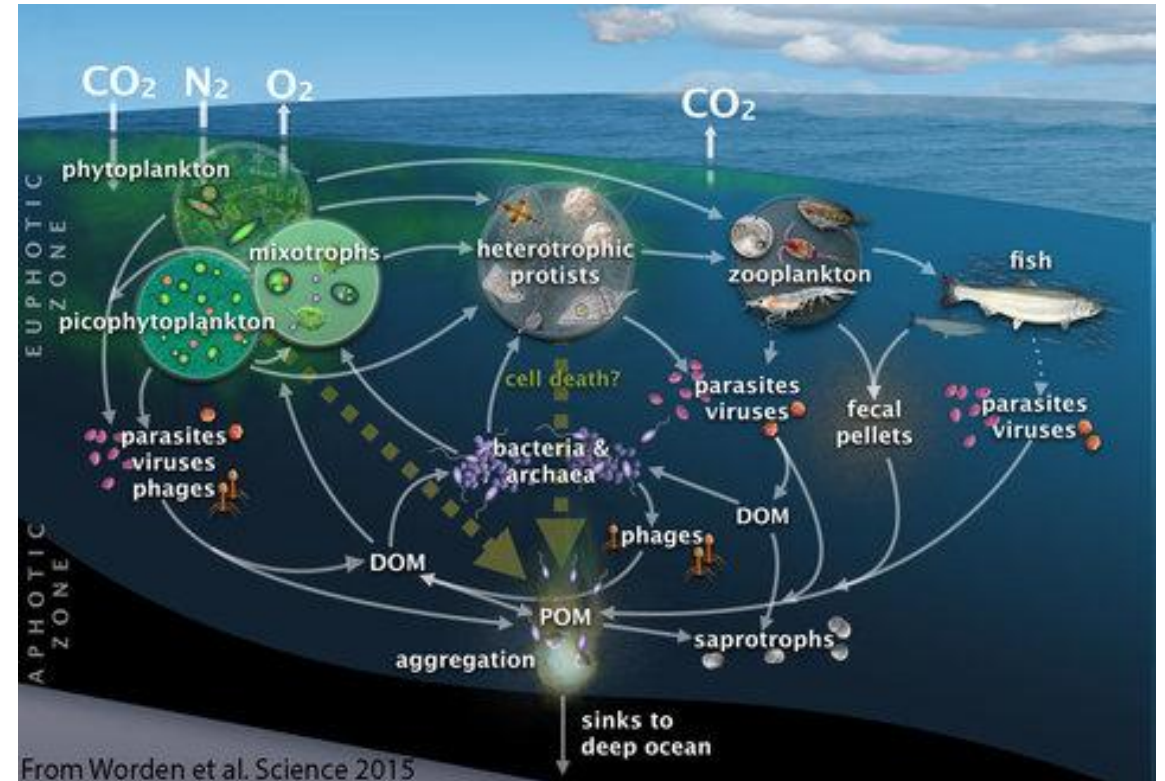
Microbes in carbon cycle

Microbes transfer an enormous flow of carbon through

Trophic interaction
Metabolism

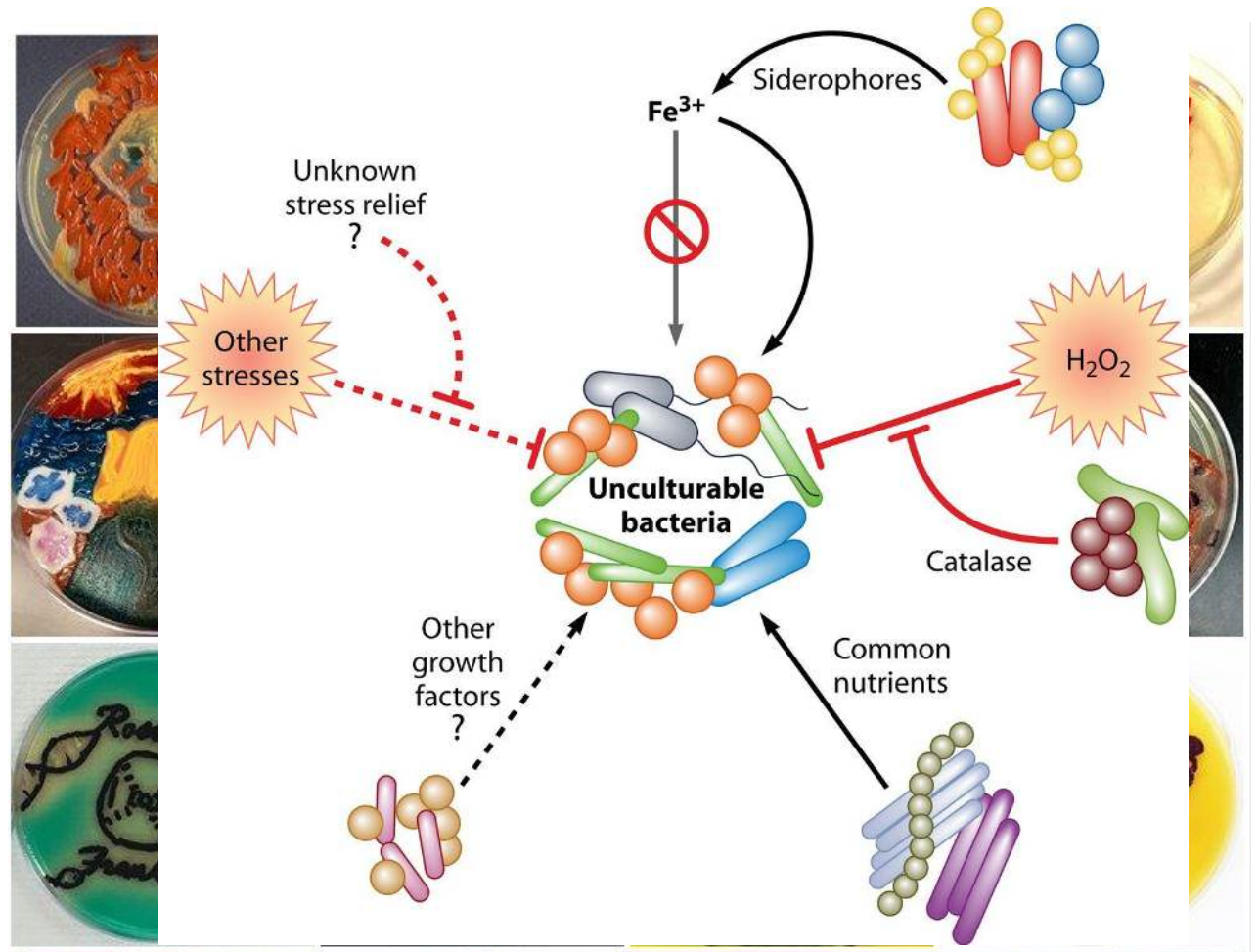
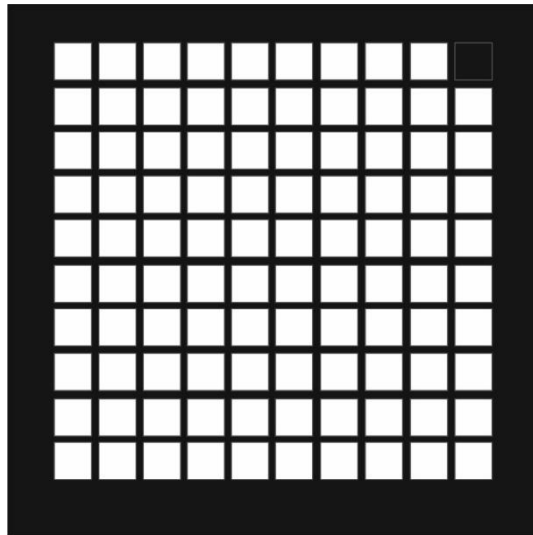
Greenhouse gas sink/emission control
Methane
Nitrous Oxide
CO₂

Climate impact

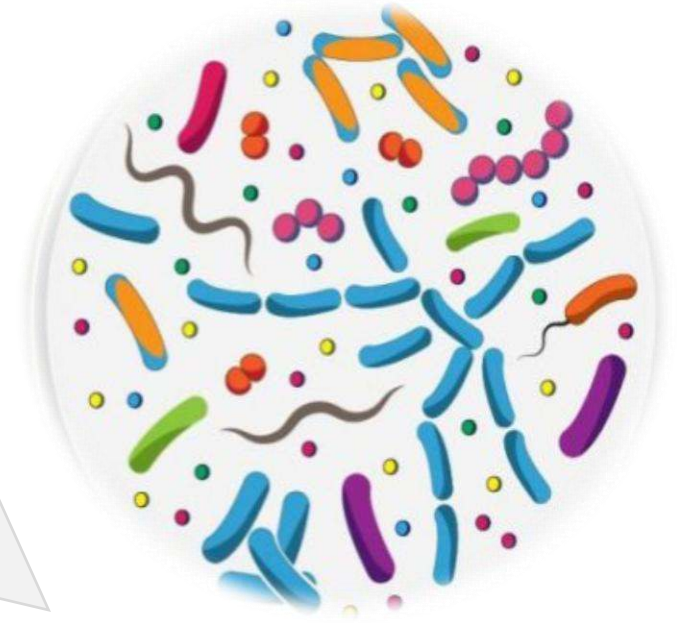
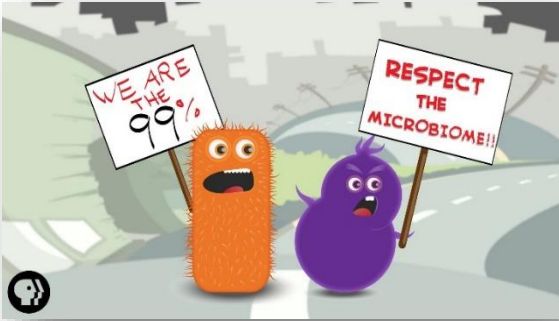


Challenge: Great plate count anomaly

Using conventional cultivation techniques only 0.1-1% of prokaryotes are Culturable in laboratory conditions.



Microbiome analysis

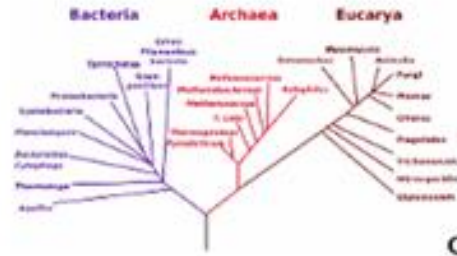


Desired efficiency anything more that **0.1-1%**





Robert Koch isolates microorganisms using solid cultures



Carl Woese propose rRNA as marker for taxonomy



Giovannoni et al., perform the first microbial community study by 16S rRNA libraries

1676

1888

1931

1977

1980

1990

1998

2005

2006

2008

2011

2015

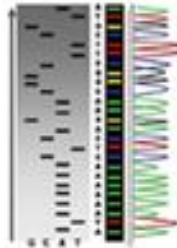
Leeuwenhoek reports his observations about oral microbiota



Winogradsky microbial ecology experiments



Fred Sanger develops DNA sequencing



Kary Mullis develops PCR



Handelsman et al., propose the term 'metagenomics'

First NGS machine released by Roche

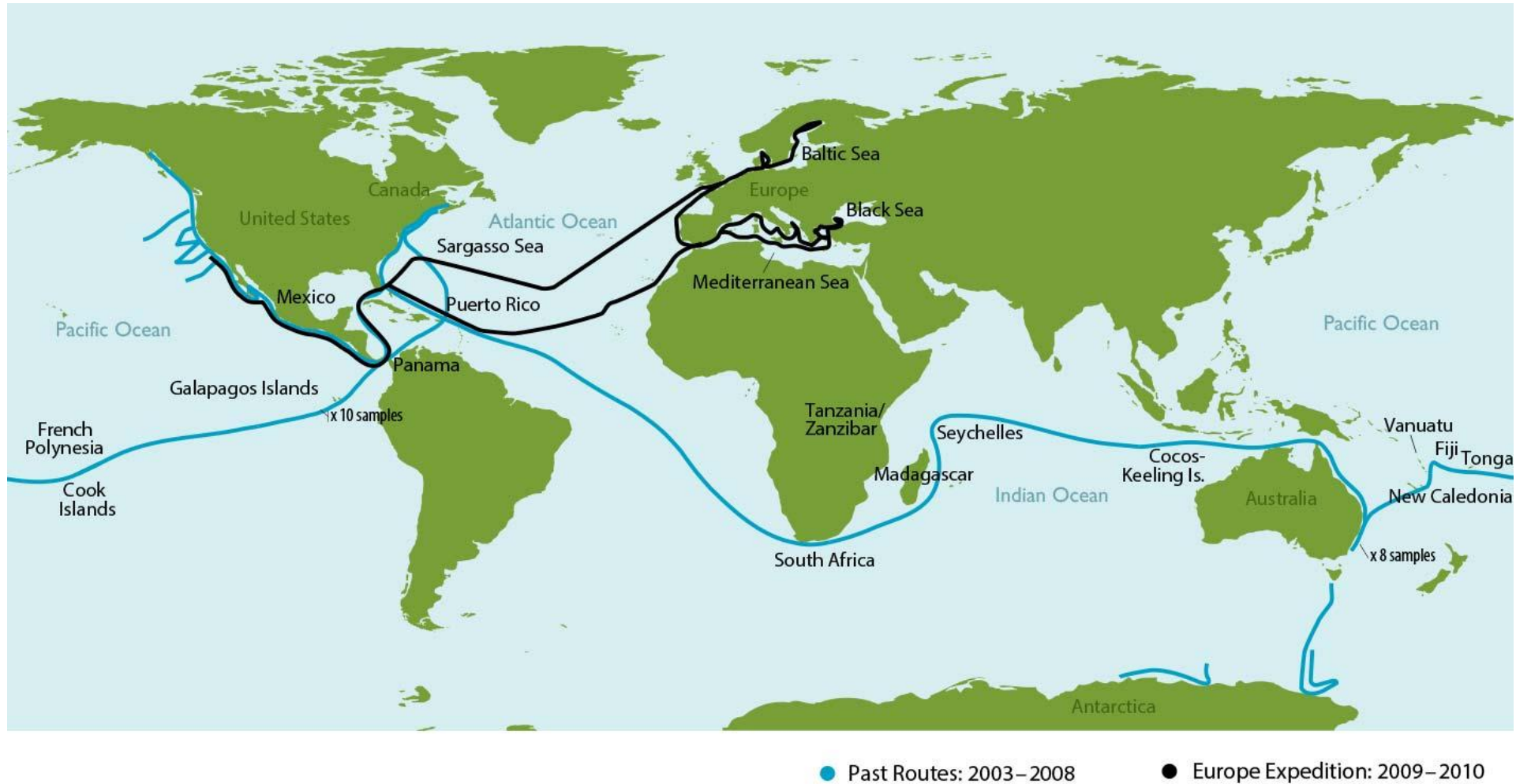
GA sequencer from Solexa is released

Human Microbiome Project publication

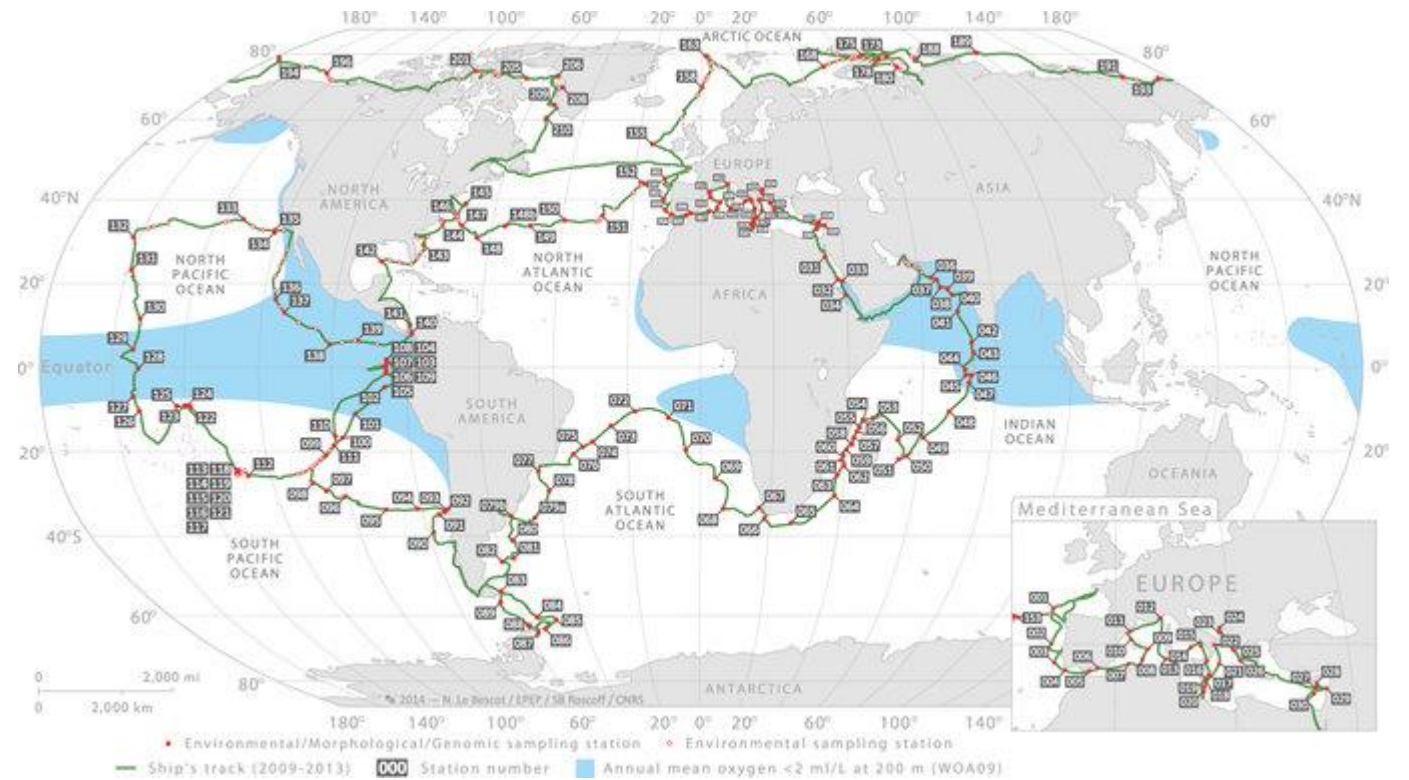
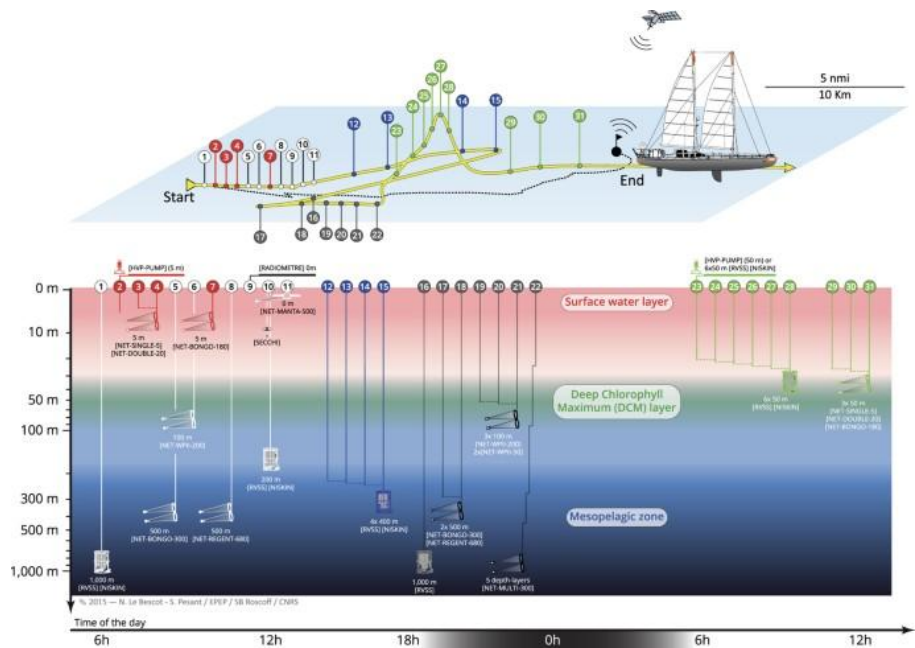
PacBio RS sequencer is released

Ocean Sampling Day

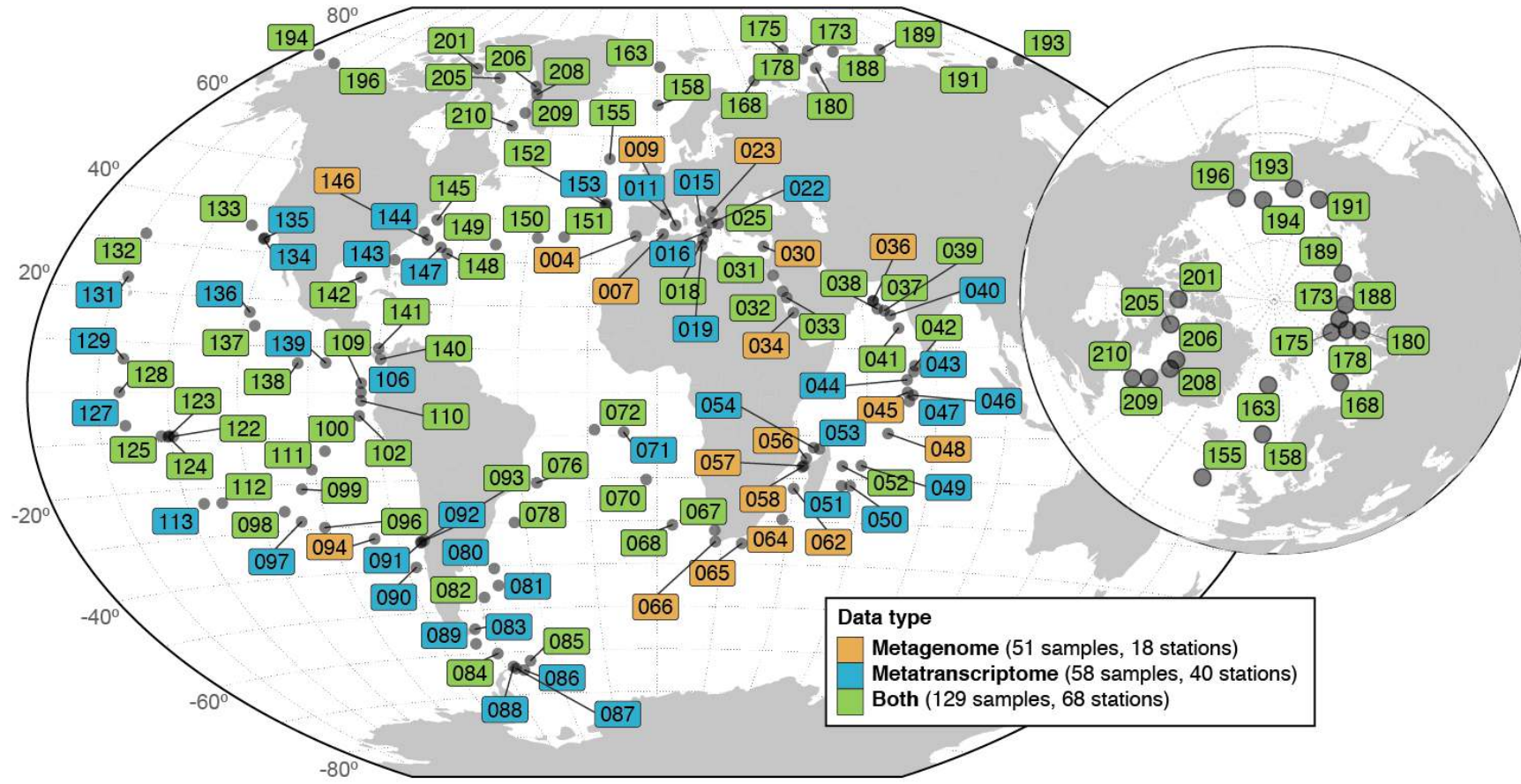
Global Ocean Sampling



Tara Ocean Sampling



Tara Ocean Sampling

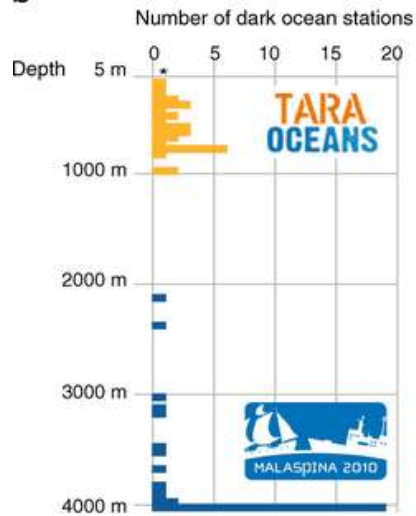


Malaspina Sampling

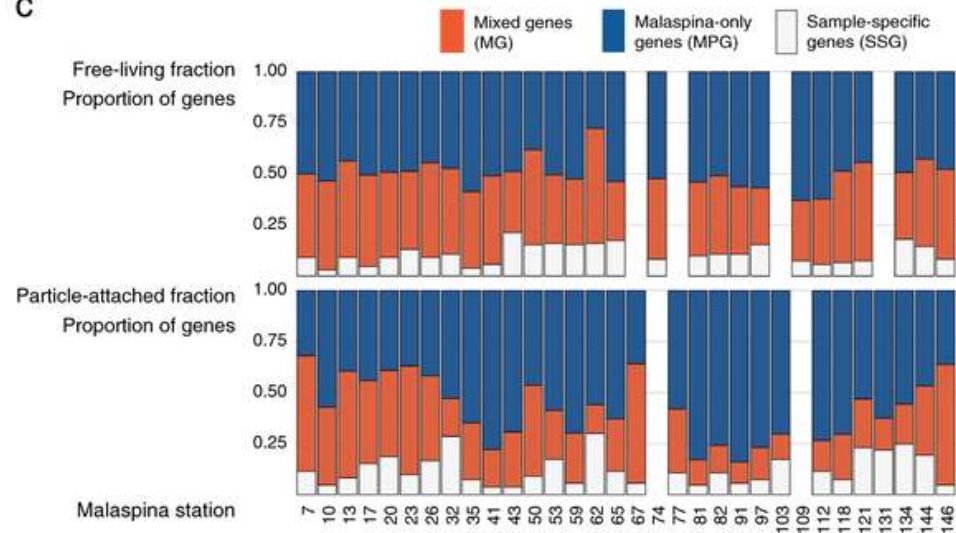
a



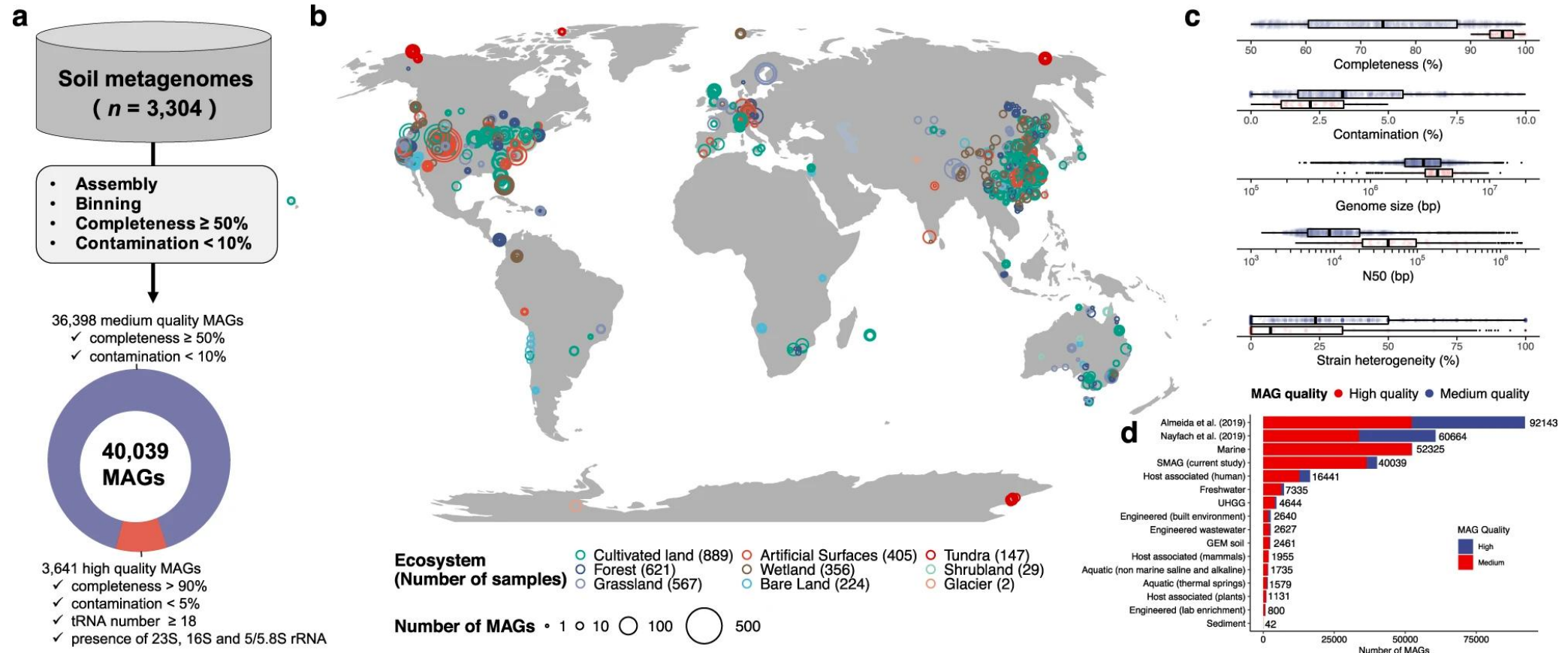
b



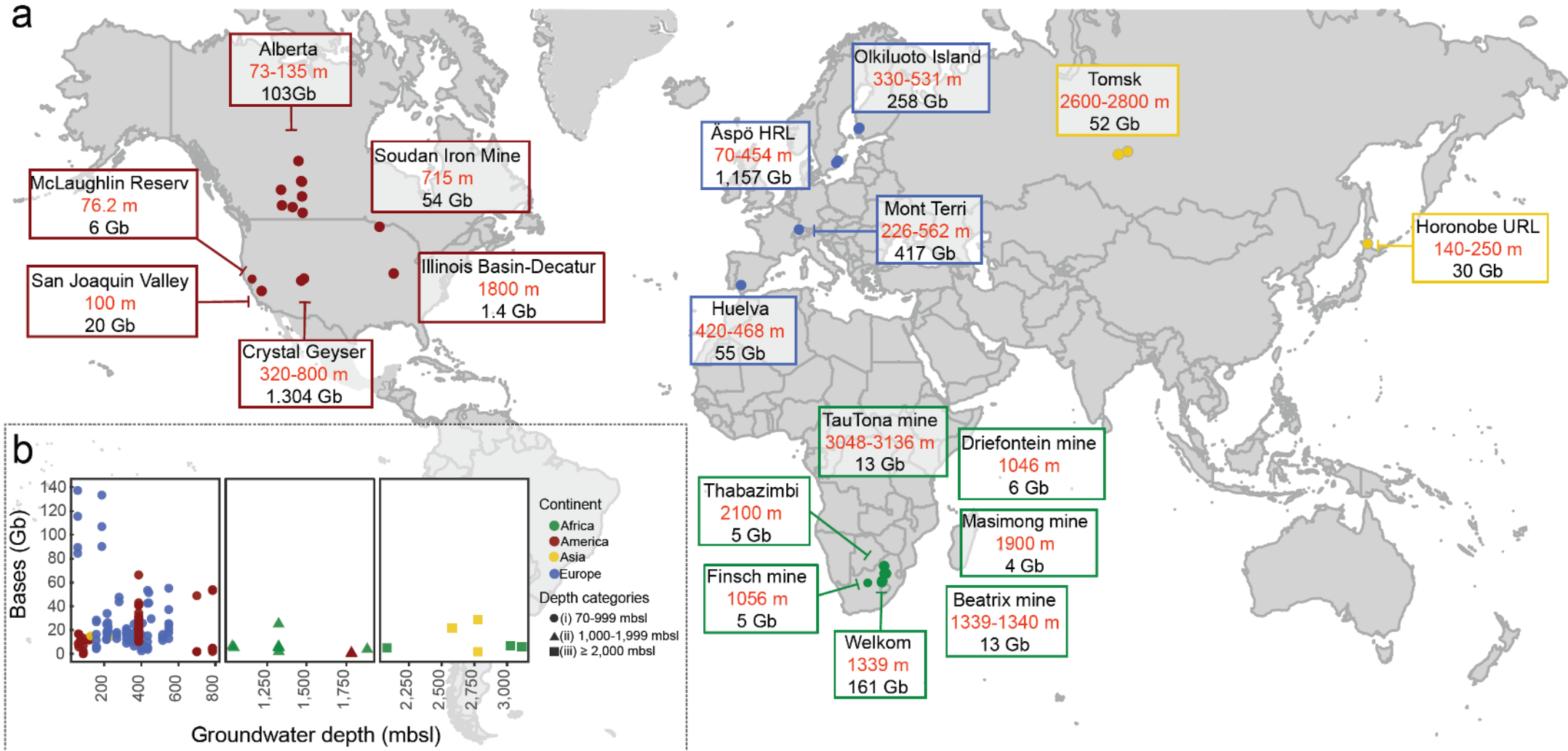
c



Soil metagenomes



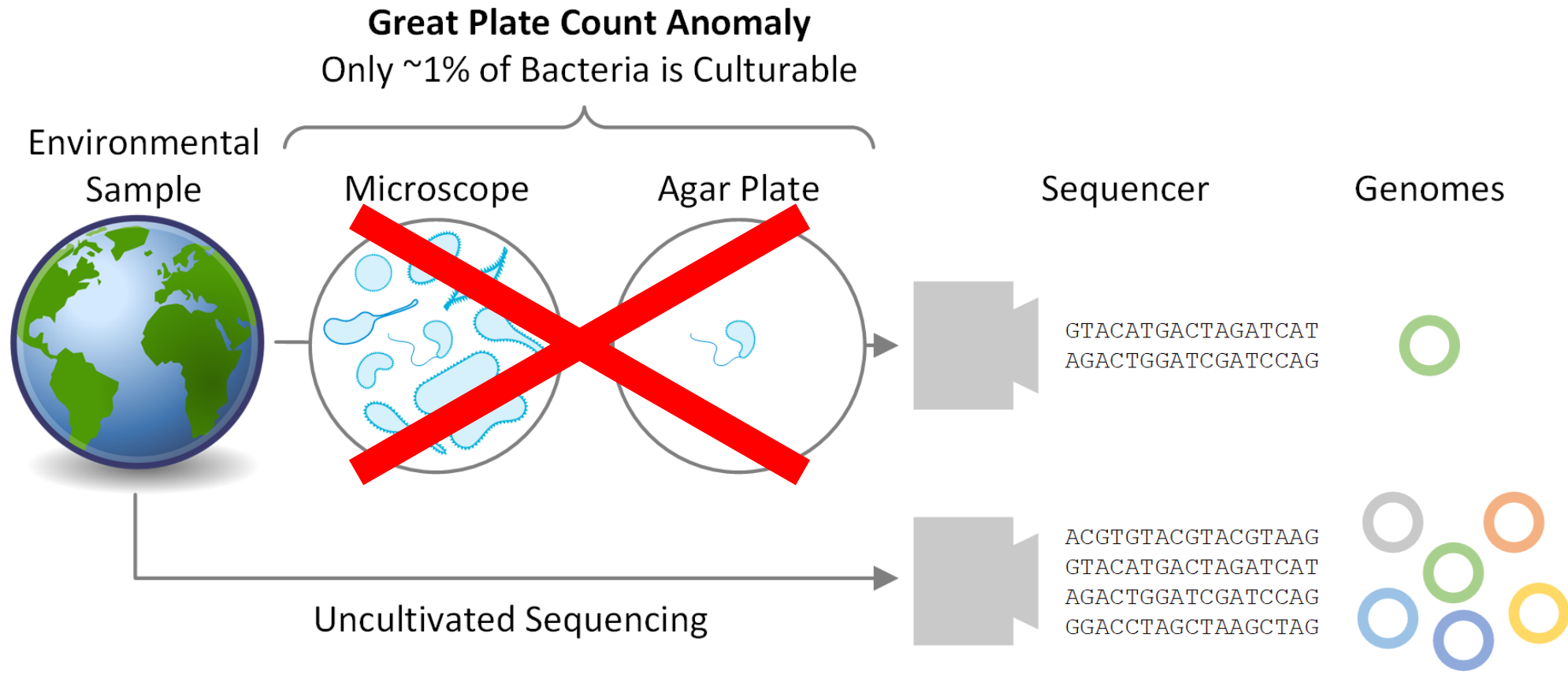
Deep groundwater metagenomes



How is Metagenomics done and what does it tell us?

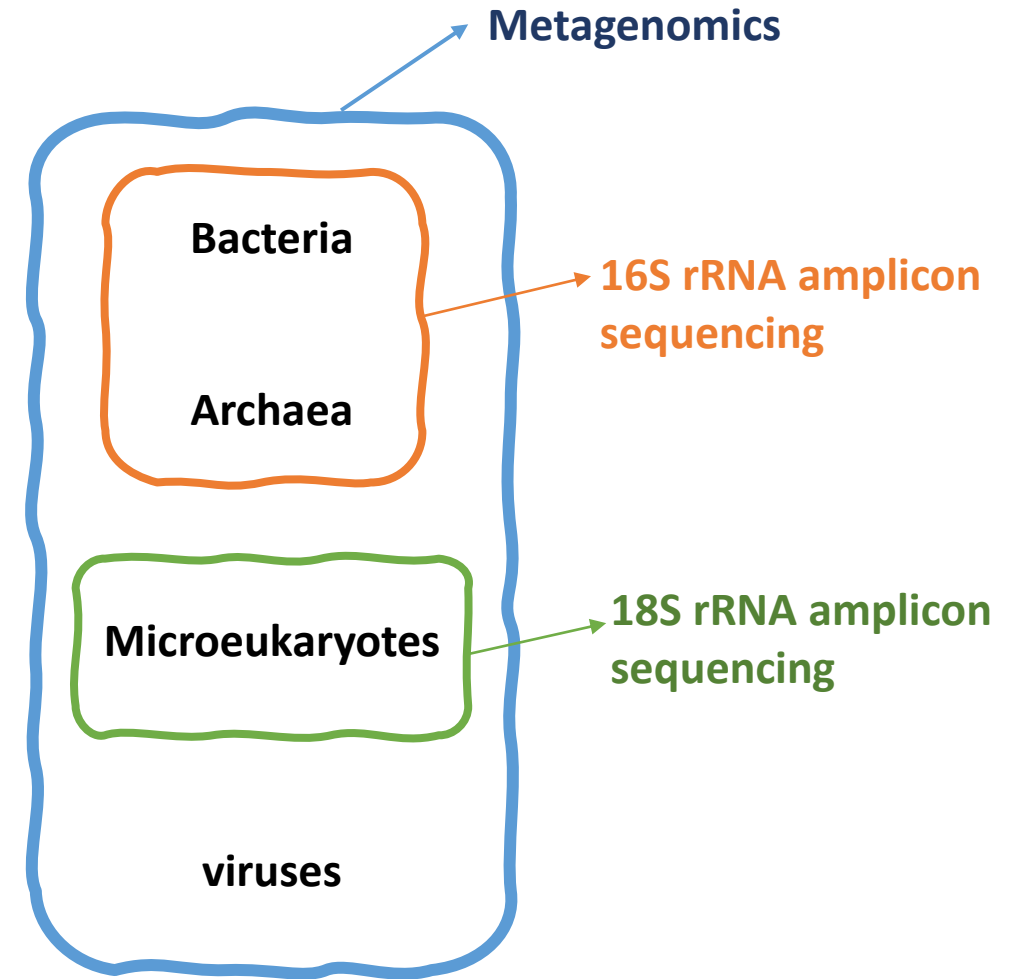
How is Metagenomics done and what does it tell us?

Capturing the unseen Majority or their footprint...

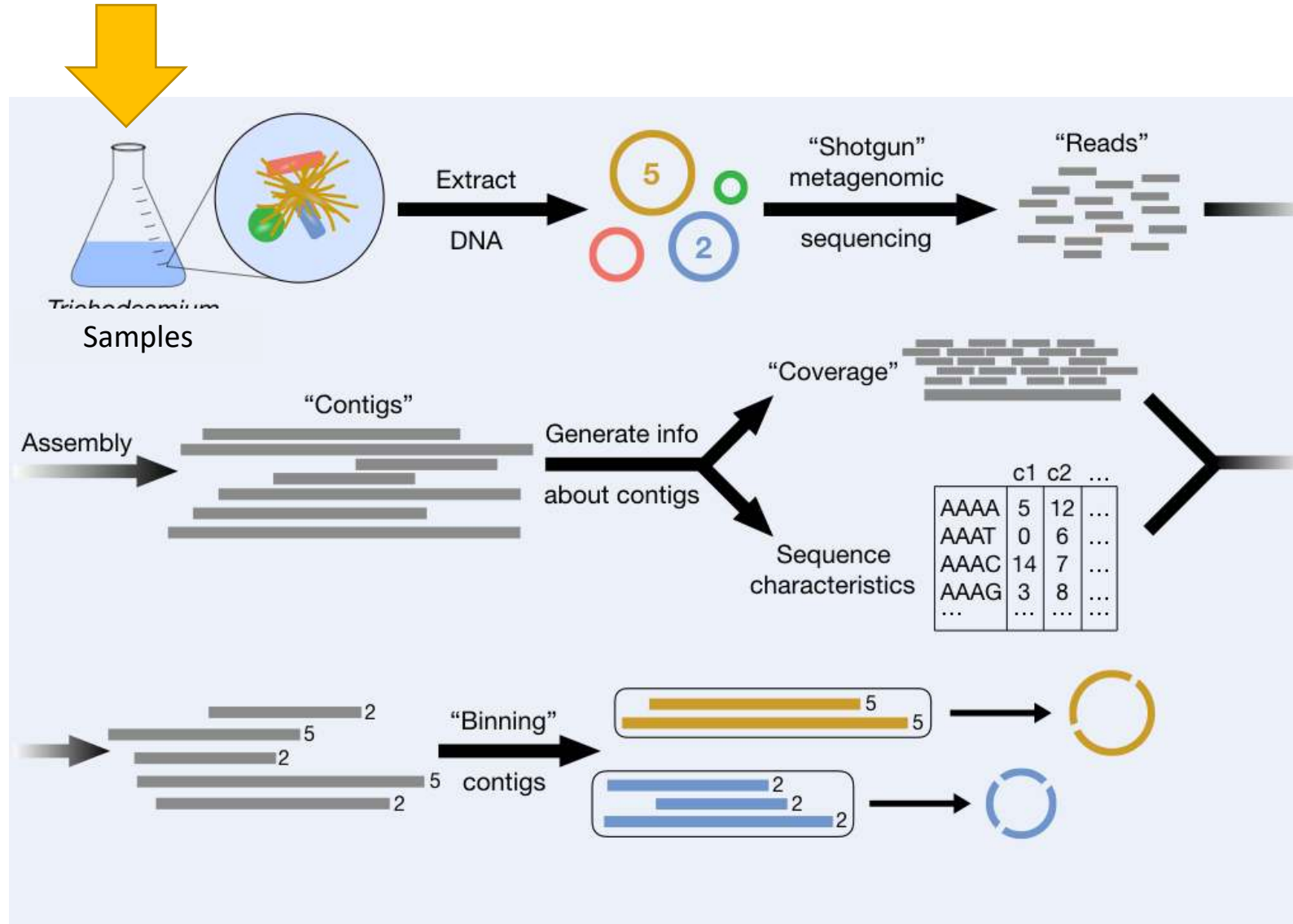


Amplicon vs. Metagenomics

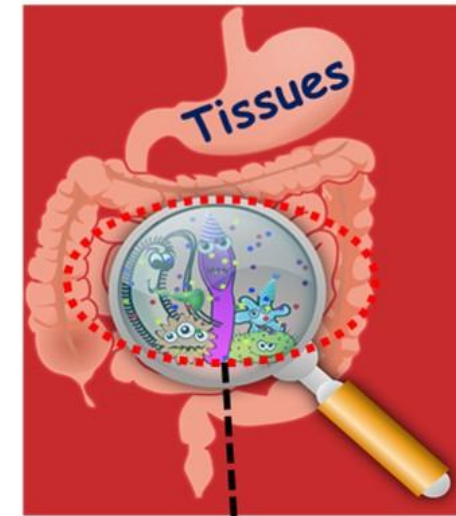
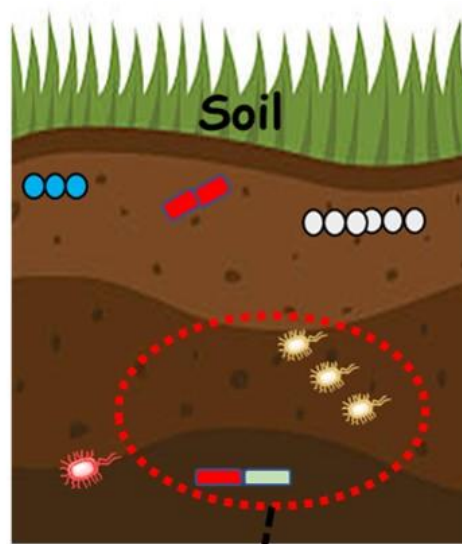
- Less complex
 - Better coverage
 - More samples
- Extensive database
- Same fragment
 - Comparable phylogenetic info
- Qualitative
- PCR and primer bias
- Limited phylogenetic info
- Limited functional information



Metagenomics Workflow



What to think of when designing a experiment...



Optimized sample collection, preparation, and DNA extraction

Scientific question

Sample type

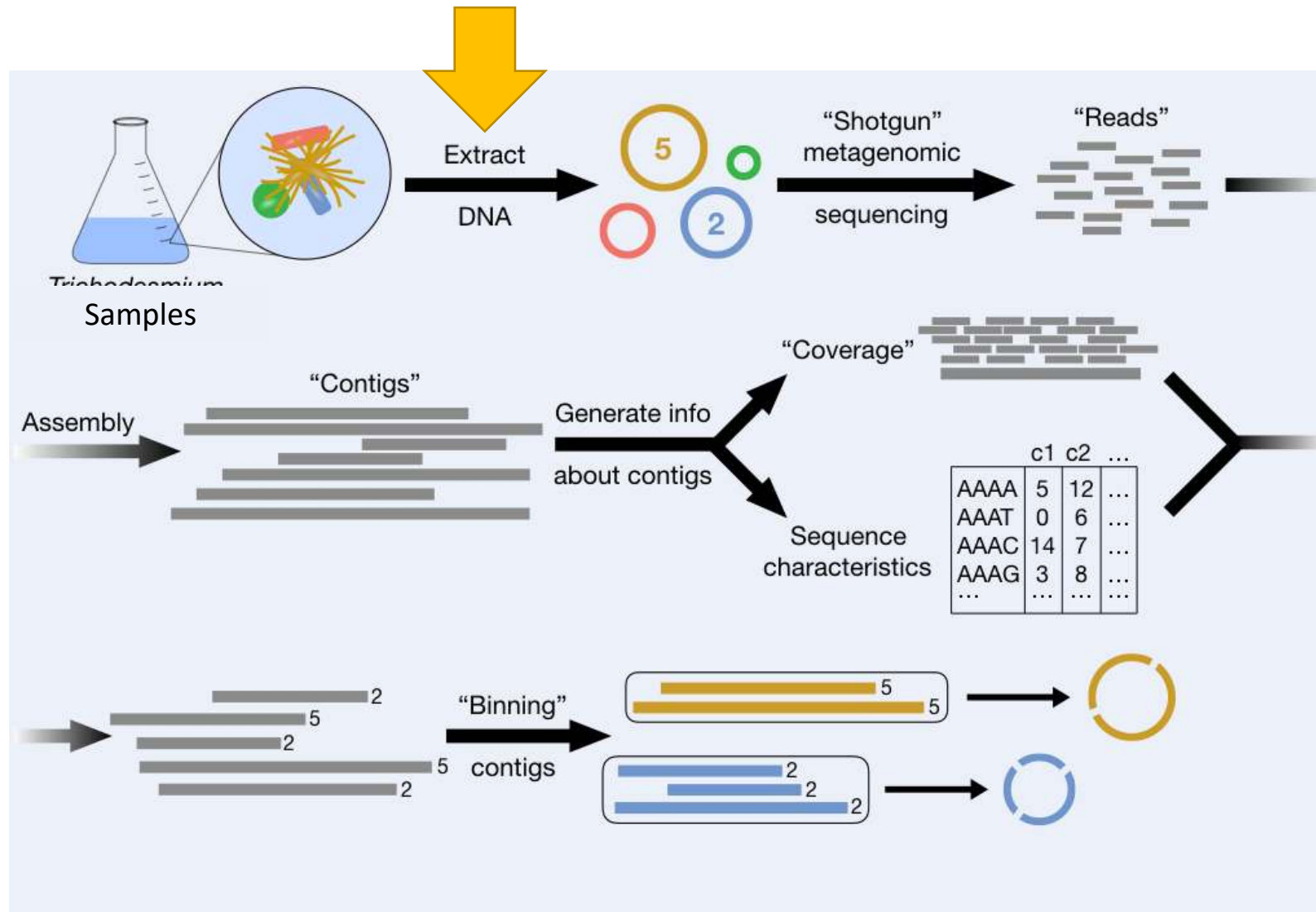
Sample should be representative

Remember all metagenomics values are RELATIVE

Process samples fast
and preserve DNA properly

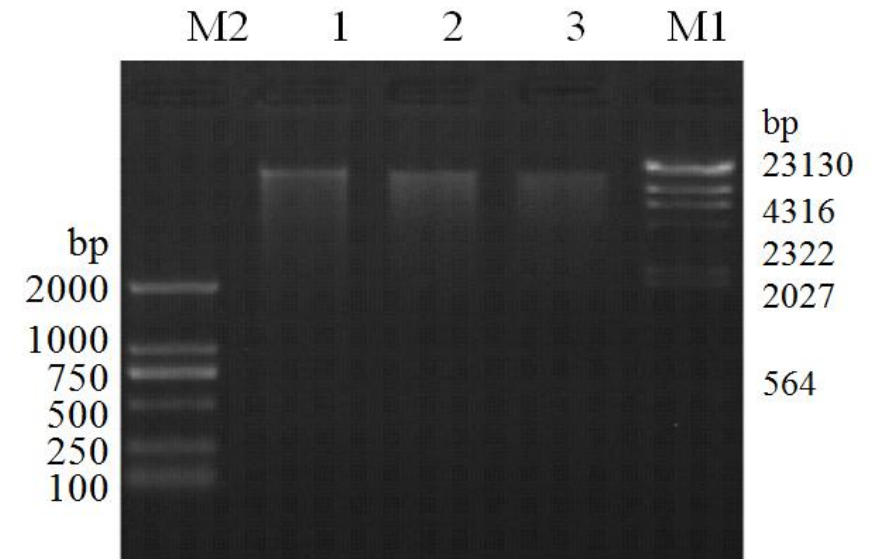


Metagenomics Workflow

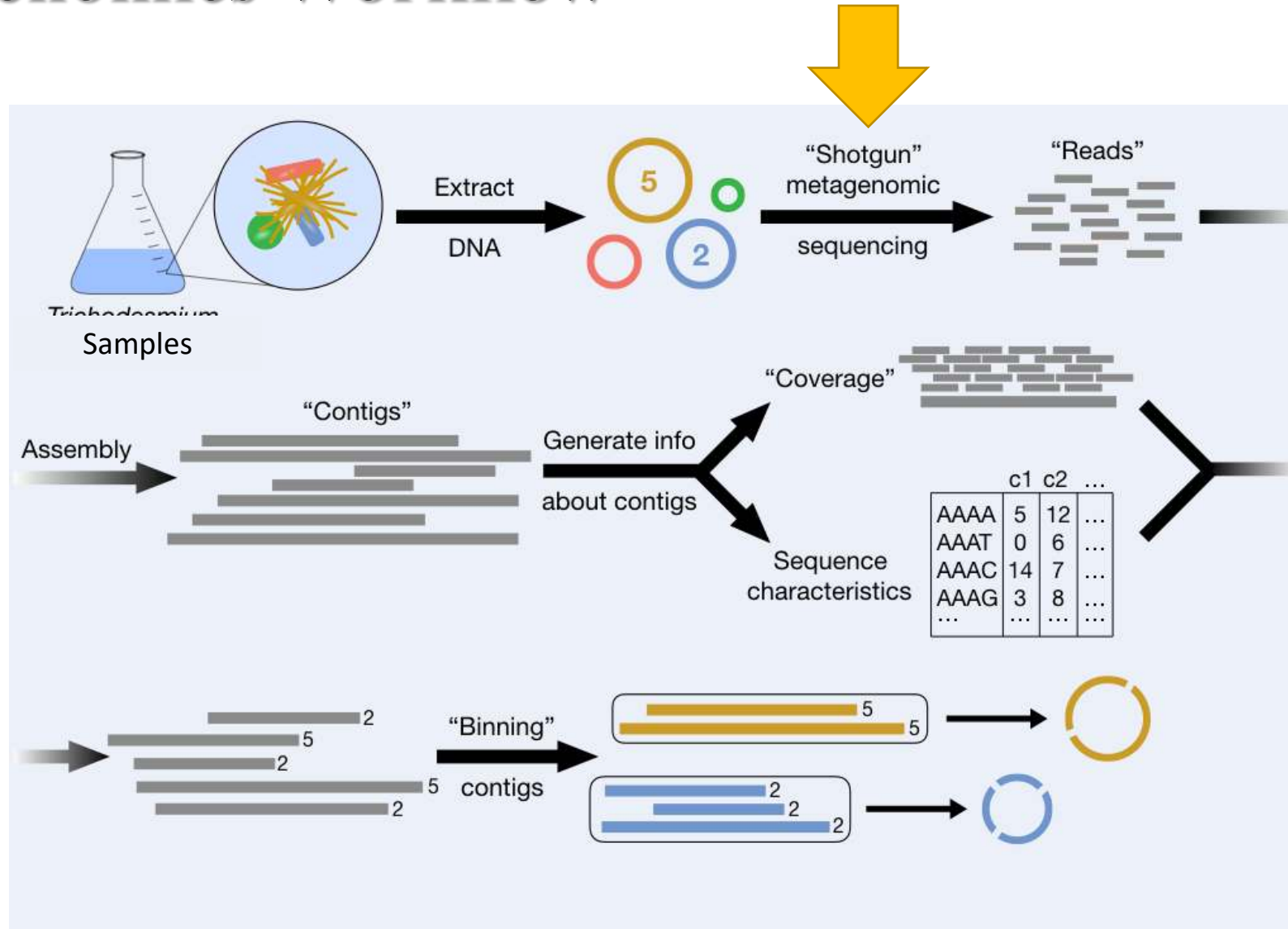


DNA recovery method impacts the output

- Choice of DNA extraction method
 - Consistent method for all experiments we want to compare
- DNA extraction quality
 - Gel electrophoresis
 - The integrity and size of genomic DNA
 - Spectrophotometry
 - Pure DNA has an A260/A280 ratio of 1.7–1.9
 - DNA concentration has been determined using nanodrop
 - Fluorometry concentration measurements



Metagenomics Workflow



Sequencing

Short read

illumina[®]

Long read



Consider the sequencing quality, read length, and price

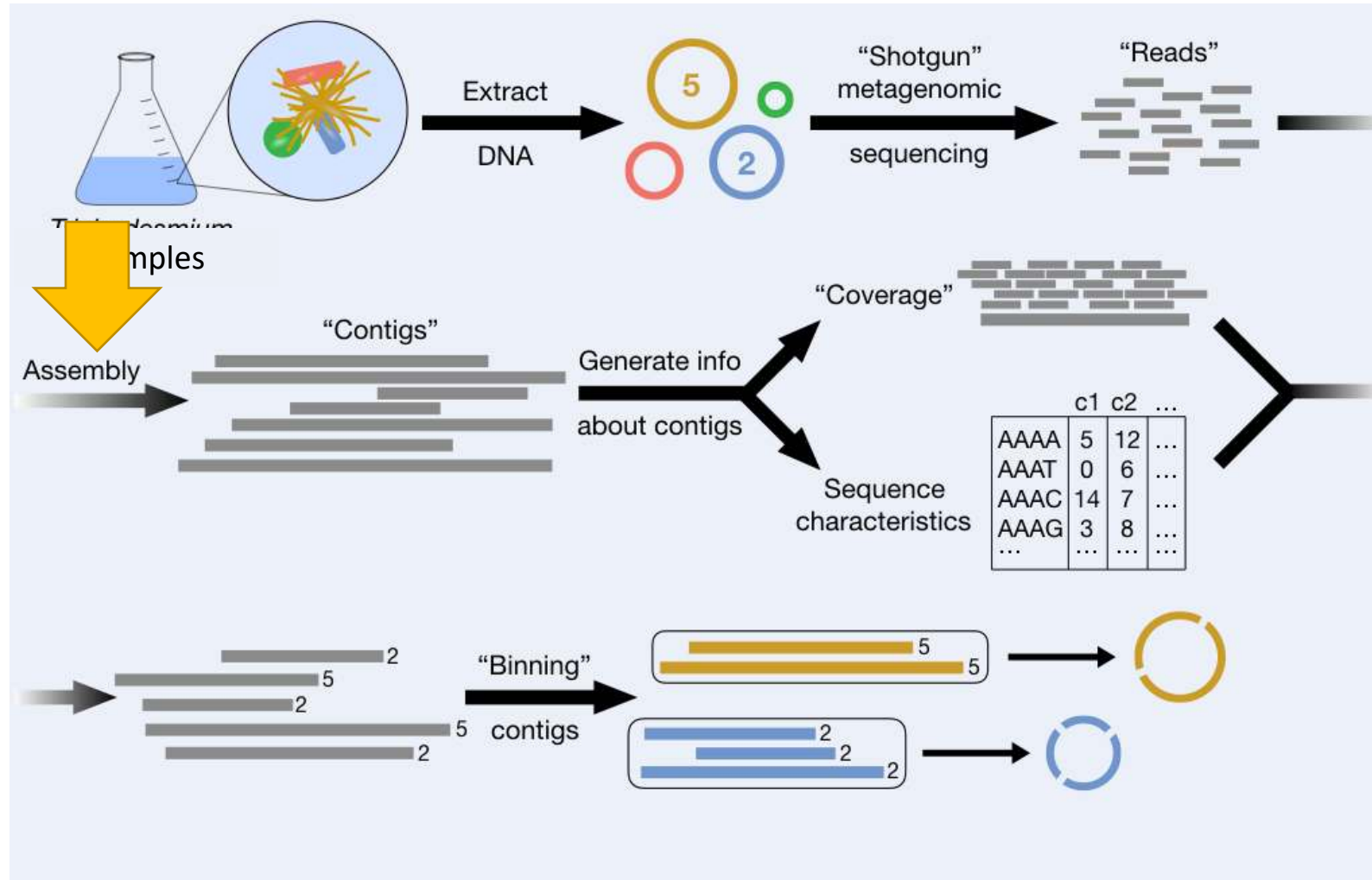


Deeper sequencing = higher resolution also = computationally intensive

Let's have a break and come back in 20 min

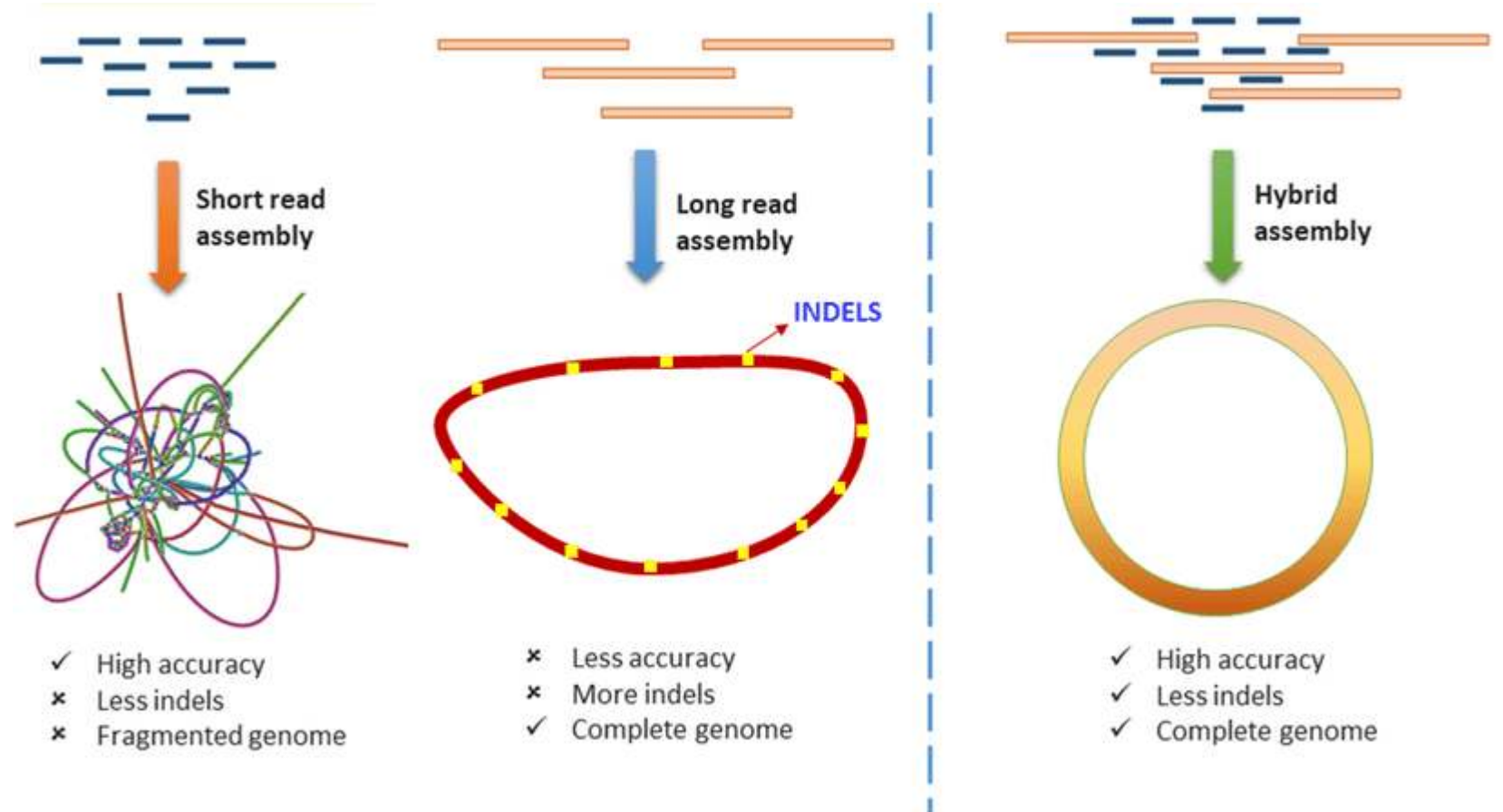


Metagenomics Workflow



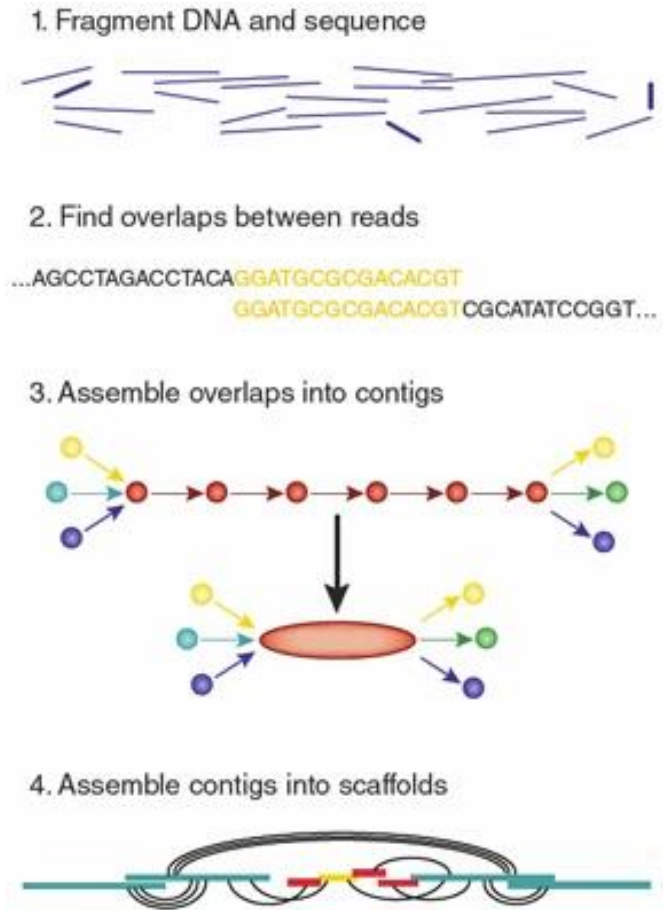
Sequence analysis

Assembly

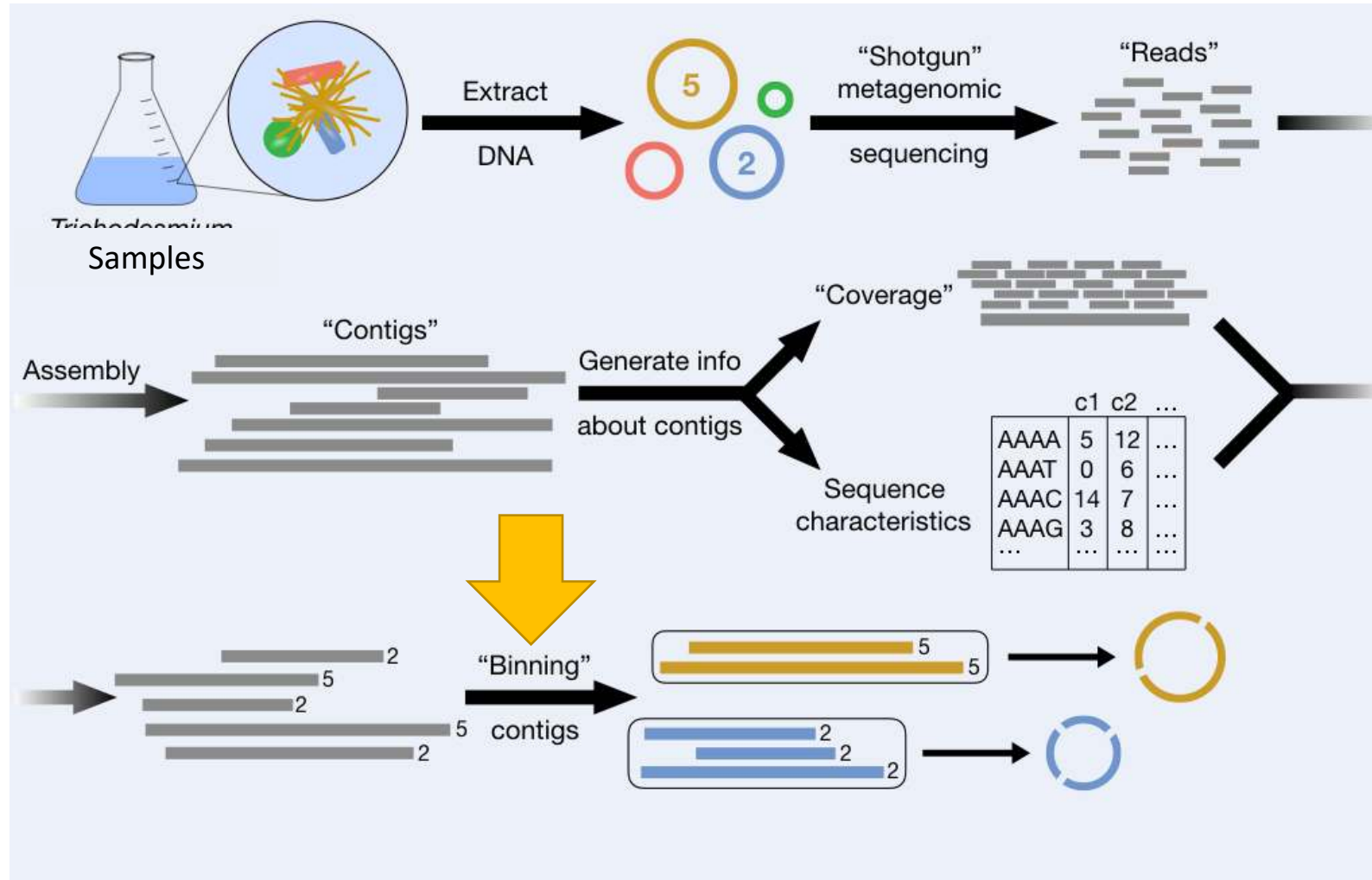


Assembly tools

- **MEGAHIT:** makes use of succinct *de Bruijn* graphs (SdBG; Bowe *et al.*, 2012), which are compressed representation of *de Bruijn* graphs.
- **metaSPAdes:** first constructs the *de Bruijn* graph of all reads using SPAdes, transforms it into the assembly graph using various graph simplification procedures, and reconstructs paths in the assembly graph that correspond to long genomic fragments within a metagenome.



Metagenomics Workflow

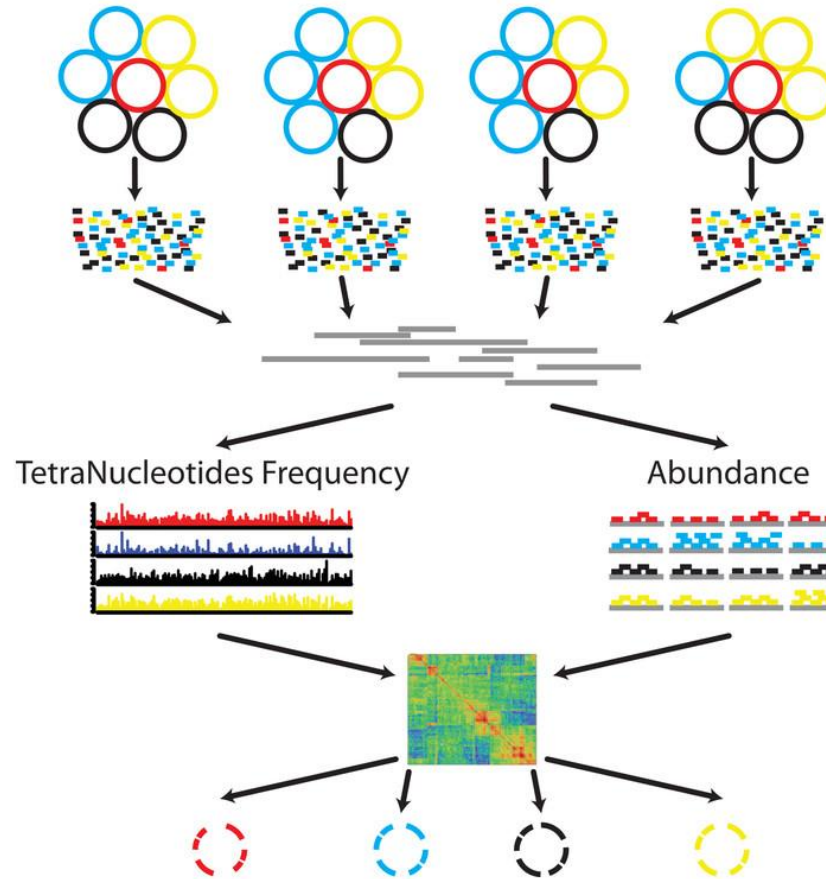


Sequence analysis

Binning

MAG (Metagenome assembled genome)

MAG quality check



Preprocessing

- 1 Samples from multiple sites or times
- 2 Metagenome libraries
- 3 Initial de-novo assembly using the combined library

MetaBAT

- 4 Calculate TNF for each contig
- 5 Calculate Abundance per library for each contig
- 6 Calculate the pairwise distance matrix using pre-trained probabilistic models
- 7 Forming genome bins iteratively

Binning tools

- **MetaBAT2:** uses the same raw TNF and abundance (ABD) scores
- **CONCOCT:** does unsupervised binning of metagenomic contigs by using nucleotide composition - kmer frequencies - and coverage data for multiple samples.
- **MaxBin:** algorithm utilizes two different genomic features: tetranucleotide frequencies and scaffold coverage levels to populate the genomic bins using single-copy maker genes and an expectation-maximization algorithm.

Genome/MAG quality check

- CheckM provides robust estimates of genome completeness and contamination by using collocated sets of genes that are ubiquitous and single-copy within a phylogenetic lineage.
- Assessment of genome quality using plots depicting key genomic characteristics (e.g., GC, coding density) which highlight sequences outside the expected distributions of a typical genome.
- CheckM also identifies genome bins that are likely candidates for merging based on marker set compatibility, similarity in genomic characteristics, and proximity within a reference genome tree.
- <https://ecogenomics.github.io/CheckM/>

Genome completeness standards

Table 1 Genome reporting standards for SAGs and MAGs

Criterion	Description
Finished (SAG/MAG)	
Assembly quality ^a	Single contiguous sequence without gaps or ambiguities with a consensus error rate equivalent to Q50 or better
High-quality draft (SAG/MAG)	
Assembly quality ^a	Multiple fragments where gaps span repetitive regions. Presence of the 23S, 16S, and 5S rRNA genes and at least 18 tRNAs.
Completion ^b	>90%
Contamination ^c	<5%
Medium-quality draft (SAG/MAG)	
Assembly quality ^a	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion ^b	≥50%
Contamination ^c	<10%
Low-quality draft (SAG/MAG)	
Assembly quality ^a	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion ^b	<50%
Contamination ^c	<10%

This is a compressed set of genome reporting standards for SAGs and MAGs. For a complete list of mandatory and optional standards, see **Supplementary Table 1**.

^aAssembly statistics include but are not limited to: N50, L50, largest contig, number of contigs, assembly size, percentage of reads that map back to the assembly, and number of predicted genes per genome. ^bCompletion: ratio of observed single-copy marker genes to total single-copy marker genes in chosen marker gene set. ^cContamination: ratio of observed single-copy marker genes in ≥2 copies to total single-copy marker genes in chosen marker gene set.

Genome taxonomy

- Taxonomy and nomenclature
- <https://gtdb.ecogenomic.org/>
- <https://ncbiinsights.ncbi.nlm.nih.gov/2021/12/10/ncbi-taxonomy-prokaryote-phyla-added/>

Sequence analysis

Annotation

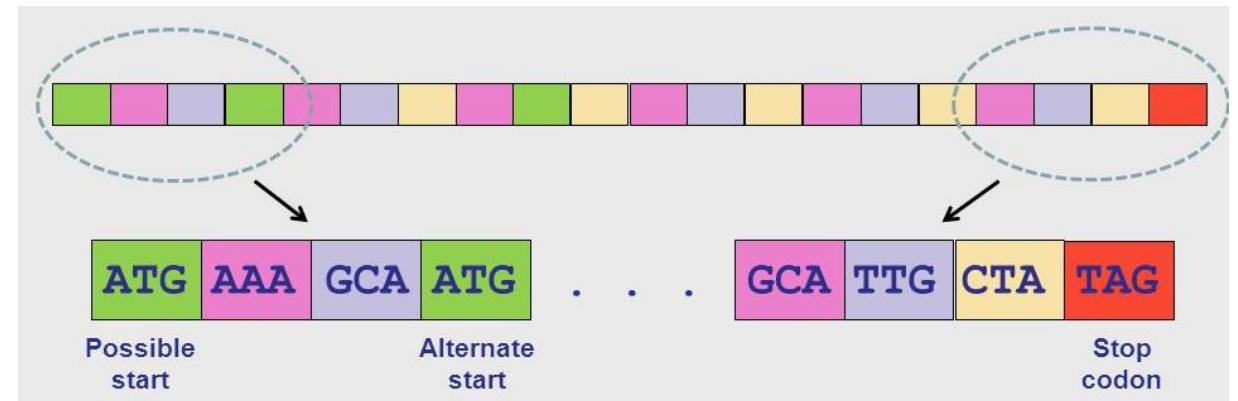
Gene prediction => ORF finding

Prodigal

Function assignment

BLAST

HMM models



Lists of IDs lists of names

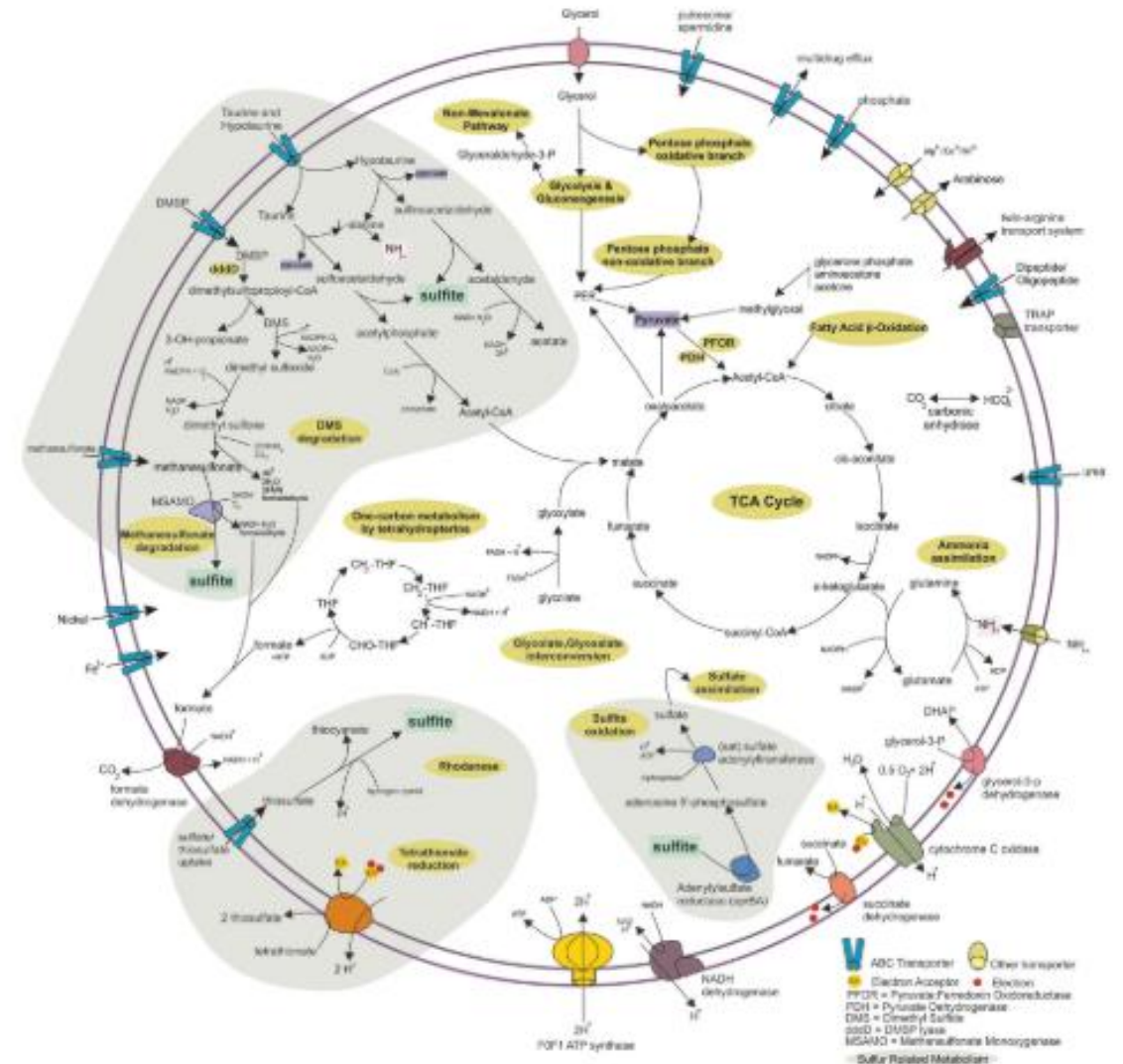
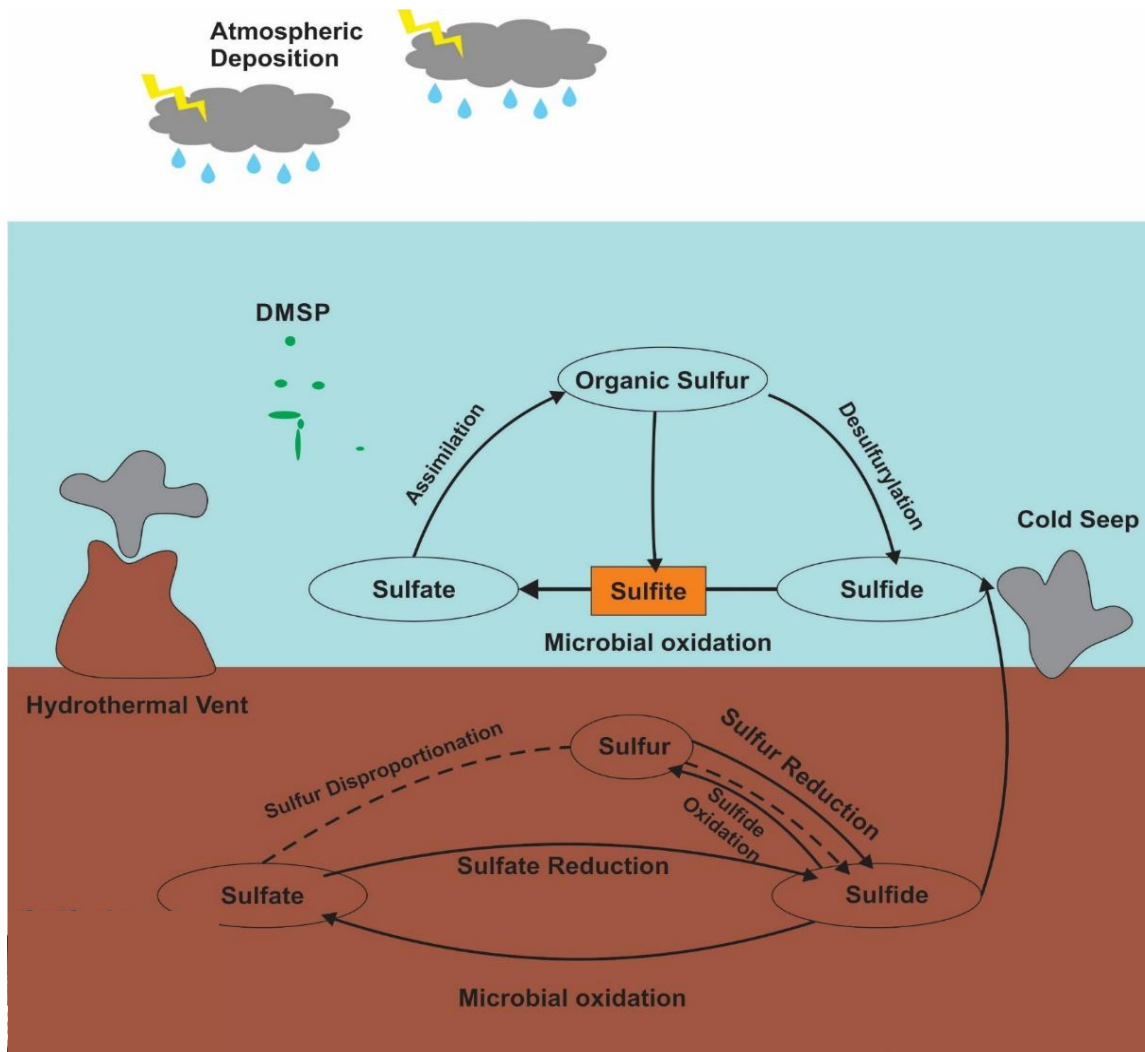
- BioCyc
- KEGG
- Ensembl Bacteria
- Kbase
- IMG
- PATRIC

```
# emapper version: emapper-1.0.3-40-g41a8498 emapper DB: 2.0
# command: ./emapper.py -m diamond -i Khw_5.faa --cpu 10 --output Khw_5.faa.emapper
# time: Mon Oct 21 04:20:39 2019
#query_name      seed_eggNOG_ortholog      seed_ortholog_evalue      seed_ortholog_score      best_tax_level Preferred_name G0s
      EC      KEGG_ko KEGG_Pathway      KEGG_Module      KEGG_Reaction      KEGG_rclass      BRITE      KEGG_TC CAZy      BiGG_Reacti
on
DFHNPLIN_00005  455436.DS989811_gene1935      8.9e-40 170.6      Alteromonadaceae      coaX      2.7.1.33,6.3.4.15 k
o:K01947,ko:K03525      ko00770,ko00780,ko01100,map00770,map00780,map01100      M00120      R01074,R02971,R03018,R04391,R05145R
C00002,RC00017,RC00043,RC00070,RC00096,RC02896      ko00000,ko00001,ko00002,ko01000      Bacteria      lmu
YA@1224,1S99V@1236,46848@72275,COG1521@1,COG1521@2      NA|NA|NA      H      Catalyzes the phosphorylation of pantothena
te (Pan), the first step in CoA biosynthesis
DFHNPLIN_00006  455436.DS989811_gene1934      8.3e-94 350.5      Alteromonadaceae      birA      G0:0000166,G0:0000976,G0:00
00984,G0:0001017,G0:0001067,G0:0003674,G0:0003676,G0:0003677,G0:0003690,G0:0003824,G0:0004077,G0:0005488,G0:0005515,G0:0005
524,G0:0005575,G0:0006082,G0:0006464,G0:0006732,G0:0006766,G0:0006767,G0:0006768,G0:0006790,G0:0006807,G0:0008144,G0:000815
0,G0:0008152,G0:0009058,G0:0009102,G0:0009108,G0:0009110,G0:0009305,G0:0009374,G0:0009987,G0:0016053,G0:0016874,G0:0016879,
G0:0017053,G0:0017076,G0:0017144,G0:0018130,G0:0018271,G0:0019538,G0:0019752,G0:0019842,G0:0030554,G0:0031406,G0:0032553,G0
:0032555,G0:0032559,G0:0032787,G0:0032991,G0:0033218,G0:0033293,G0:0034641,G0:0035639,G0:0036094,G0:0036211,G0:0042364,G0:0
042802,G0:0042803,G0:0043167,G0:0043168,G0:0043170,G0:0043177,G0:0043412,G0:0043436,G0:0043565,G0:0043603,G0:0043604,G0:004
4212,G0:0044237,G0:0044238,G0:0044249,G0:0044260,G0:0044267,G0:0044271,G0:0044272,G0:0044281,G0:0044283,G0:0046394,G0:00464
83,G0:0046983,G0:0048037,G0:0050662,G0:0051186,G0:0051188,G0:0071704,G0:0072330,G0:0097159,G0:0097367,G0:0140096,G0:1901265
,G0:1901360,G0:1901362,G0:1901363,G0:1901564,G0:1901566,G0:1901576,G0:1901681,G0:1990837      2.7.1.33,6.3.4.15      ko:
K01947,ko:K03524,ko:K04096      ko00770,ko00780,ko01100,map00770,map00780,map01100      M00120      R01074,R02971,R03018,R04391
,R05145 RC00002,RC00017,RC00043,RC00070,RC00096,RC02896      ko00000,ko00001,ko00002,ko01000,ko03000      Bac
teria      lmuMCC@1224,1RNGC@1236,465FU@72275,COG0340@1,COG0340@2,COG1654@1,COG1654@2      NA|NA|NA      K      Acts both a
s a biotin-- acetyl-CoA-carboxylase ligase and a biotin-operon repressor. In the presence of ATP, BirA activates biotin to
form the BirA-biotinyl-S'-adenylate (BirA-bio- 5'-AMP or holoBirA) complex. HoloBirA can either transfer the biotinyl moiety
to the biotin carboxyl carrier protein (BCCP) subunit of acetyl-CoA carboxylase, or bind to the biotin operator site and
inhibit transcription of the operon
DFHNPLIN_00007  455436.DS989811_gene1933      1.6e-94 352.8      Alteromonadaceae      murB      G0:0000166,G0:0000270,G0:00
03674,G0:0003824,G0:0005488,G0:0005575,G0:0005622,G0:0005623,G0:0005737,G0:0005829,G0:0006022,G0:0006023,G0:0006024,G0:0006
807,G0:0008150,G0:0008152,G0:0008360,G0:0008762,G0:0009058,G0:0009059,G0:0009252,G0:0009273,G0:0009987,G0:0016043,G0:001649
1,G0:0016614,G0:0016616,G0:0022603,G0:0022604,G0:0030203,G0:0034645,G0:0036094,G0:0042546,G0:0043167,G0:0043168,G0:0043170,
G0:0044036,G0:0044038,G0:0044085,G0:0044237,G0:0044260,G0:0044267,G0:0044271,G0:0044272,G0:0044444,G0:0044464,G0:0045229,G0:0048037,G0
:0050660,G0:0050662,G0:0050789,G0:0050793,G0:0050794,G0:0051128,G0:0055114,G0:0065007,G0:0065008,G0:0070589,G0:0071554,G0:0
071555,G0:0071704,G0:0071840,G0:0097159,G0:1901135,G0:1901137,G0:1901265,G0:1901363,G0:1901564,G0:1901566,G0:1901576      1.3
.1.98      ko:K00075      ko00520,ko00550,ko01100,map00520,map00550,map01100      R03191,R03192      RC02639      ko00000,ko0
0001,ko01000,ko01011      iECED1_1282.ECED1_4683,iECUMN_1333.ECUMN_4498_AT6,iLF82_1304.LF82_1416,iNRG857_1313
--More-- (0%)
```

KEGG

- The Kyoto Encyclopedia of Genes and Genomes is a resource for understanding high-level functions of a biological system from molecular-level information.
-
- Tools for analysis of large-scale molecular datasets generated by high-throughput experimental technologies.
- Home page: <https://www.kegg.jp/>

From Protein to metabolism ...



The chase takes time ...

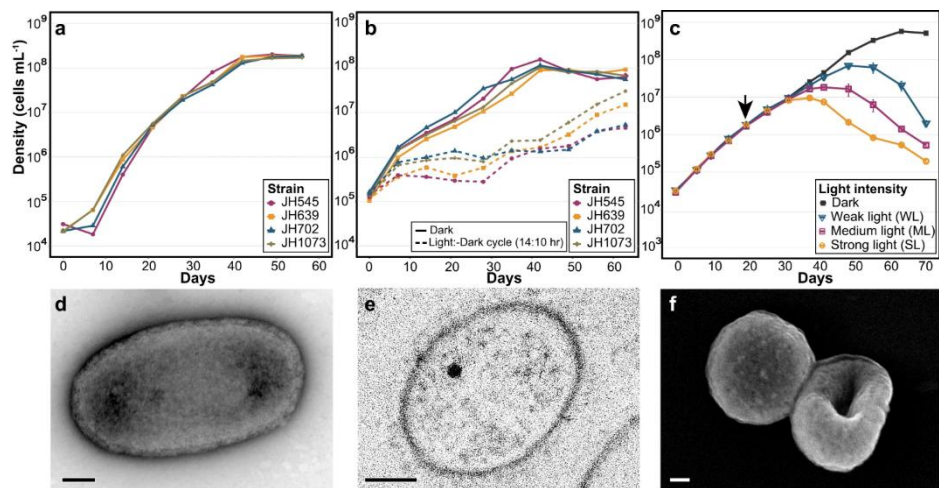
The case of SAR202

Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 7979–7984, July 1996
Microbiology

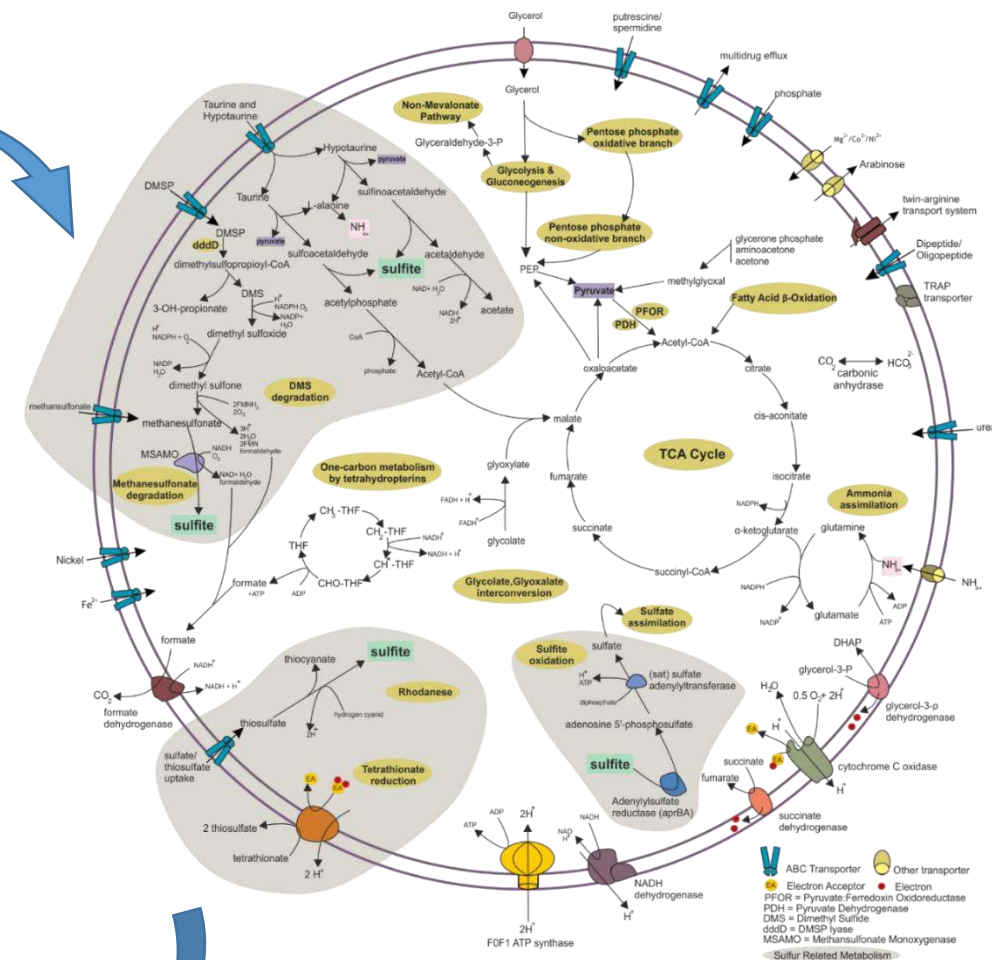
16S rRNA genes reveal stratified open ocean bacterioplankton populations related to the Green Non-Sulfur bacteria
(molecular ecology/phylogeny/thermophily)

STEPHEN J. GIOVANNONI*, MICHAEL S. RAPPE, KEVIN L. VERGIN, AND Nanci L. ADAIR

1996

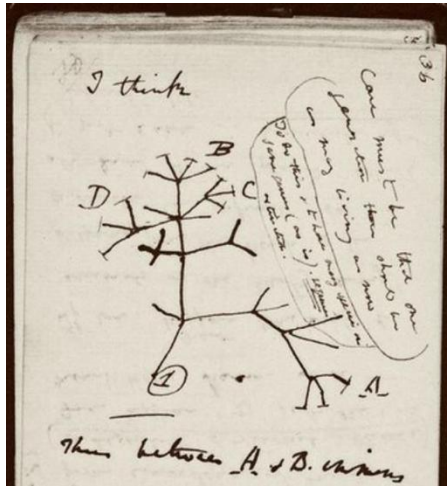


Lim et. al. *Nat. Commun.*,
2023

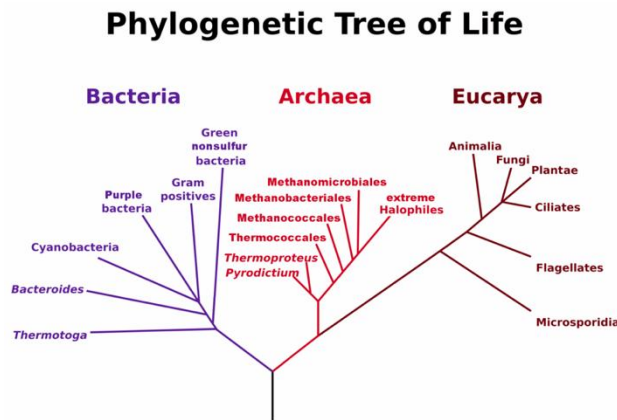


Mehrshad et. al. *ISME J.*,
2018

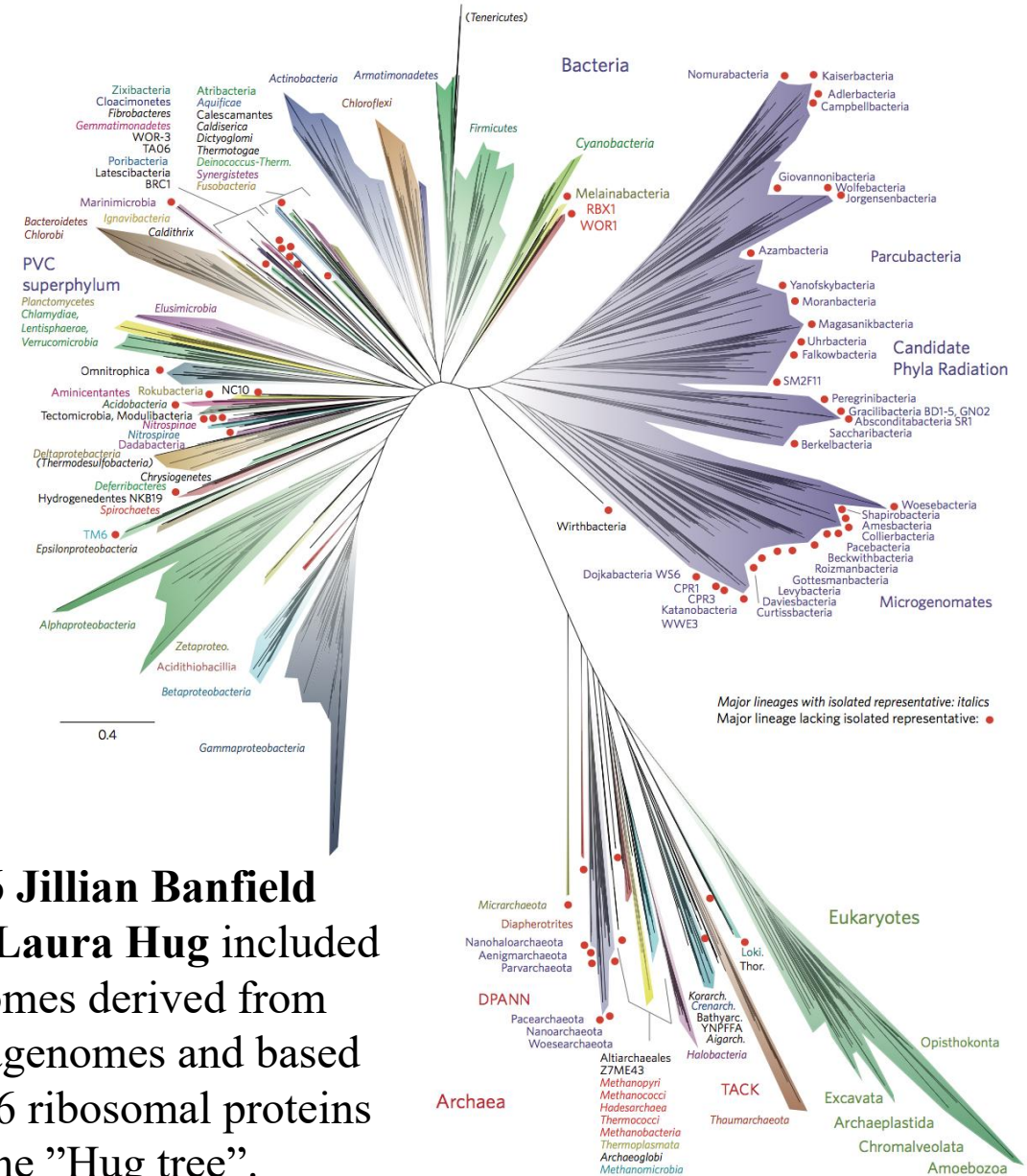
Tree of Life has evolved



1837 Charles Darwin's ideas on evolution, species descend from common ancestors and evolve over time.



1990 Carl Woese tree with LUCA and three domains. Based on rRNA gene. Later elaborated by Norman Pace.

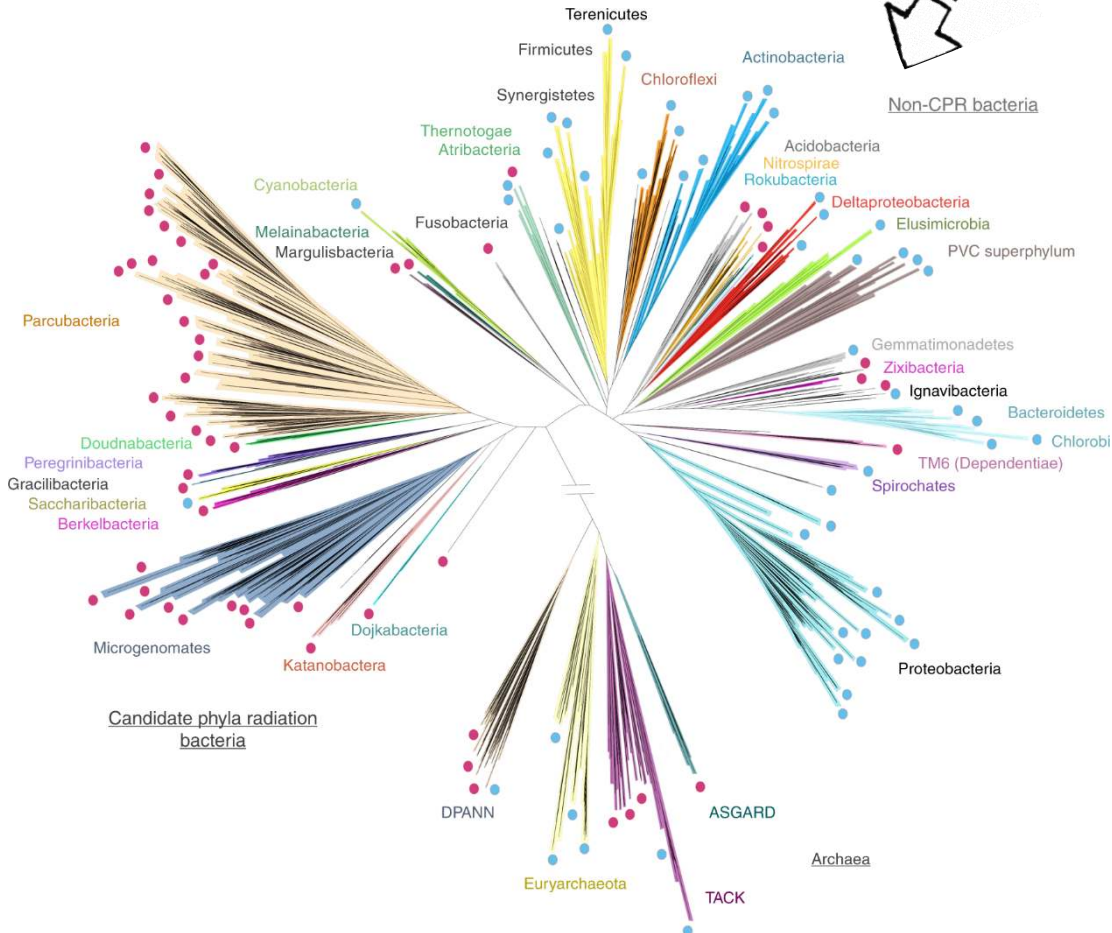


2016 Jillian Banfield and Laura Hug included genomes derived from metagenomes and based on 16 ribosomal proteins for the "Hug tree".

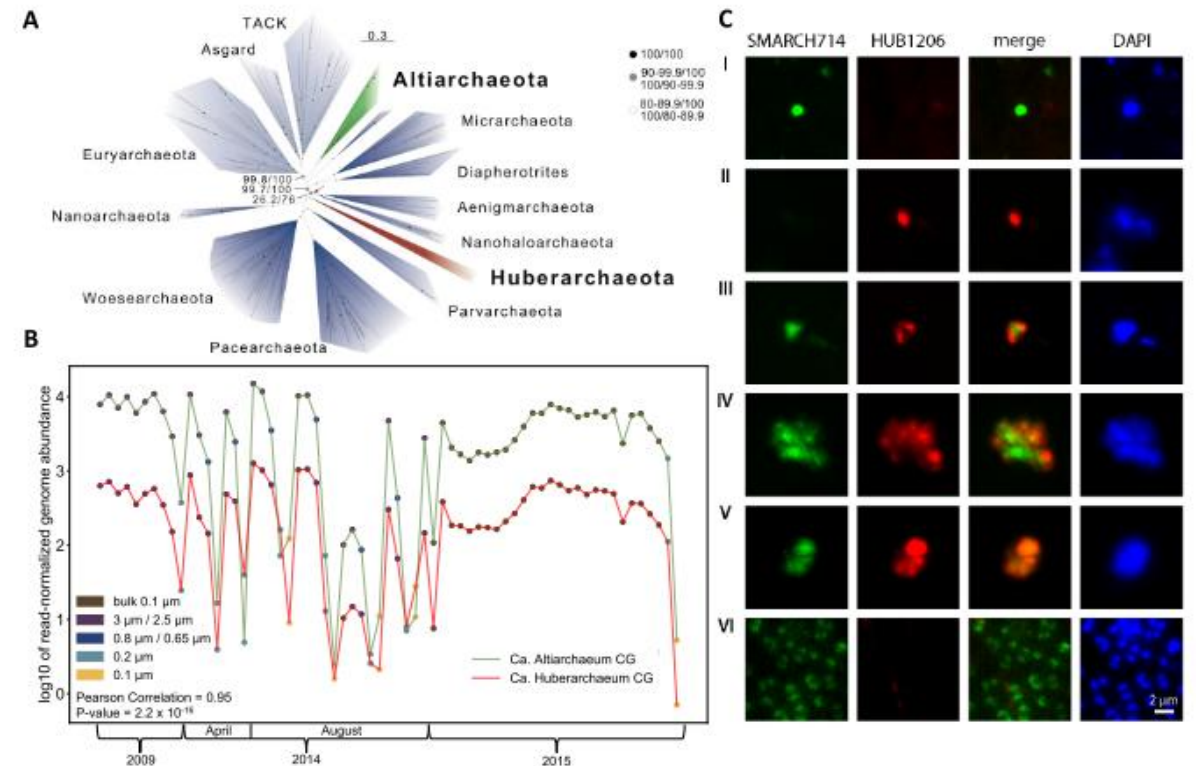
(Hug et al., 2016. Nature Microbiology)

smallest genomes with epi-symbiotic lifestyle

CPR & DPANN



Burstein et. al. *Nat Commun*, 2019

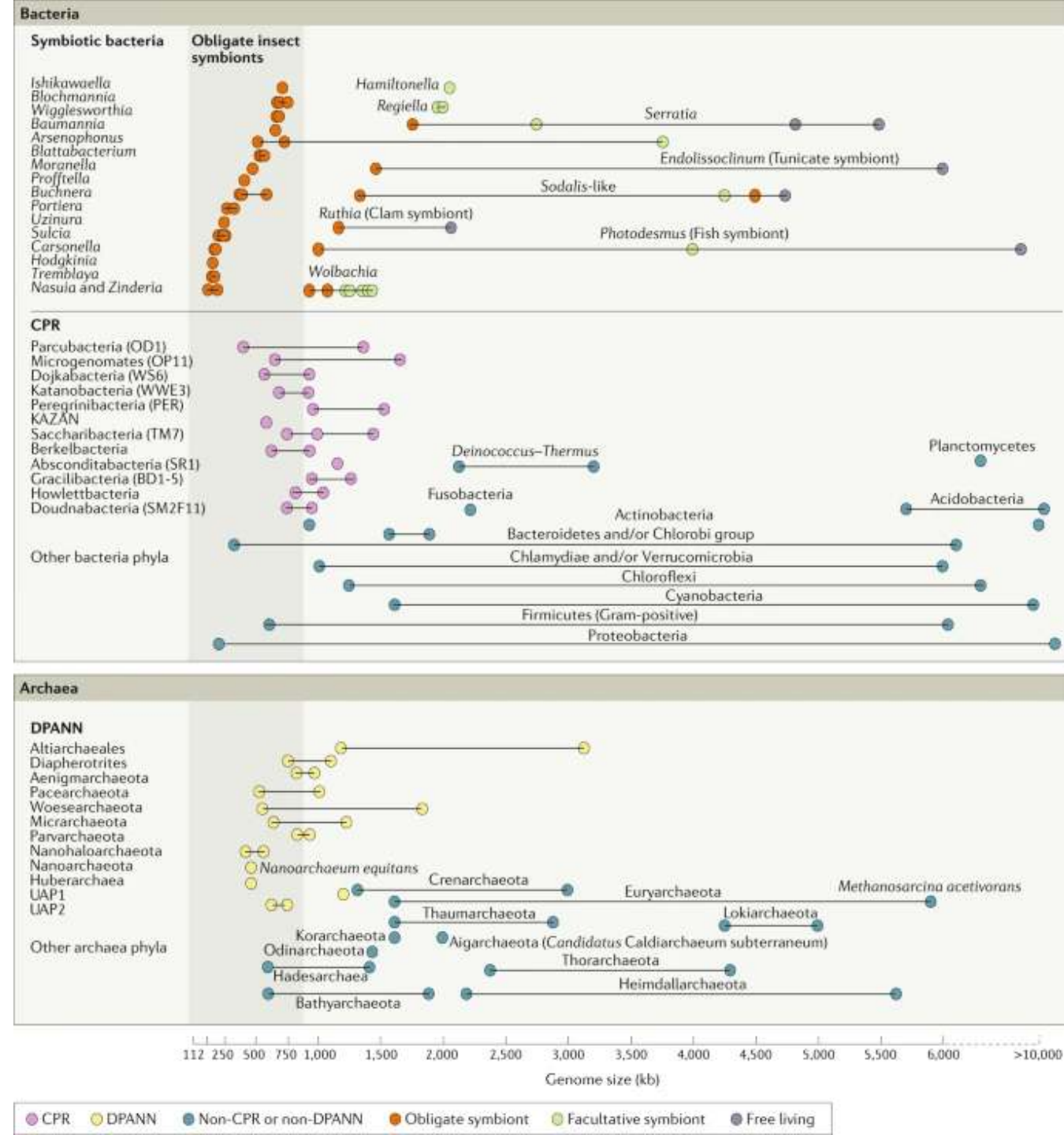


Schwank et. al. *ISMEJ*, 2019

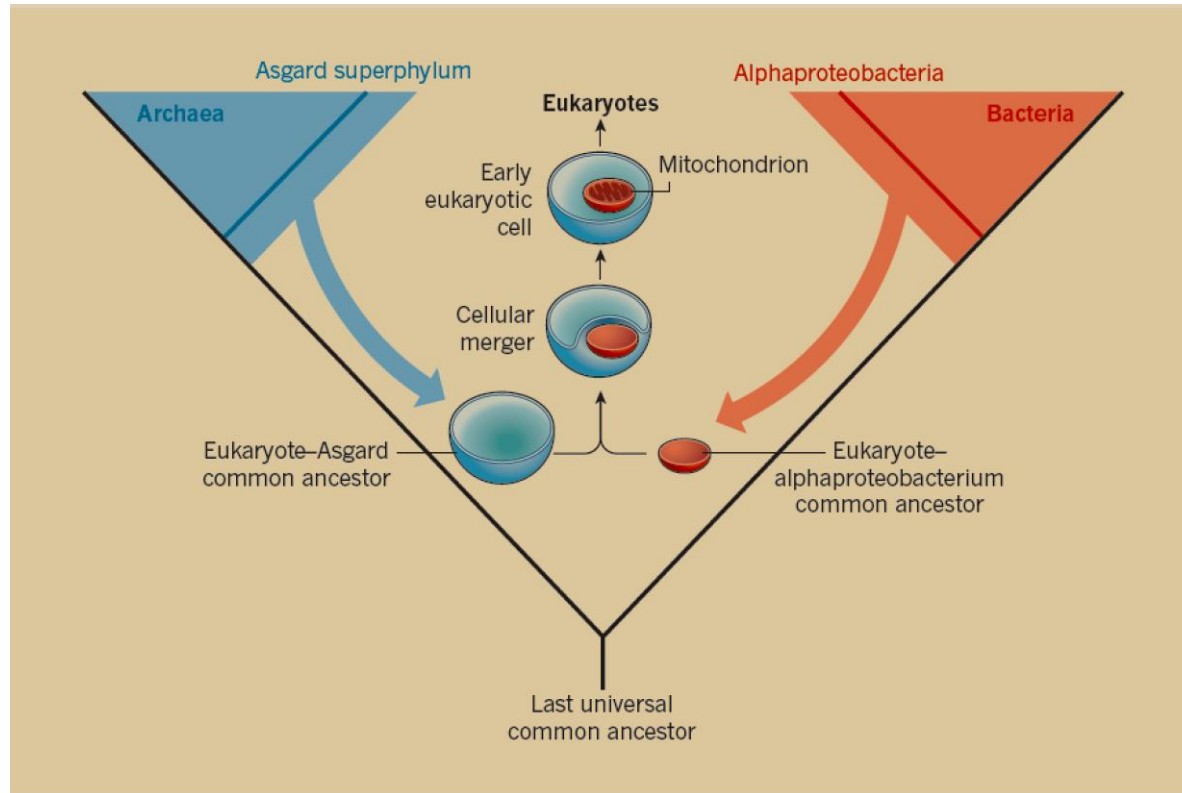
Remarkable aspect of the tree of life

- Candidate phyla radiation (CPR)
- DPANN (an acronym of the names of the first included phyla, ‘Candidatus Diapherotrites’, ‘Candidatus Parvarchaeota’, ‘Candidatus Aenigmarchaeota’, Nanoarchaeota and ‘Candidatus Nanohaloarchaeota’)

- Small genomes
- Small cell sizes
- Notable gaps in core metabolic potential
- Mostly symbiotic lifestyle
- Their ecological role is not yet well understood

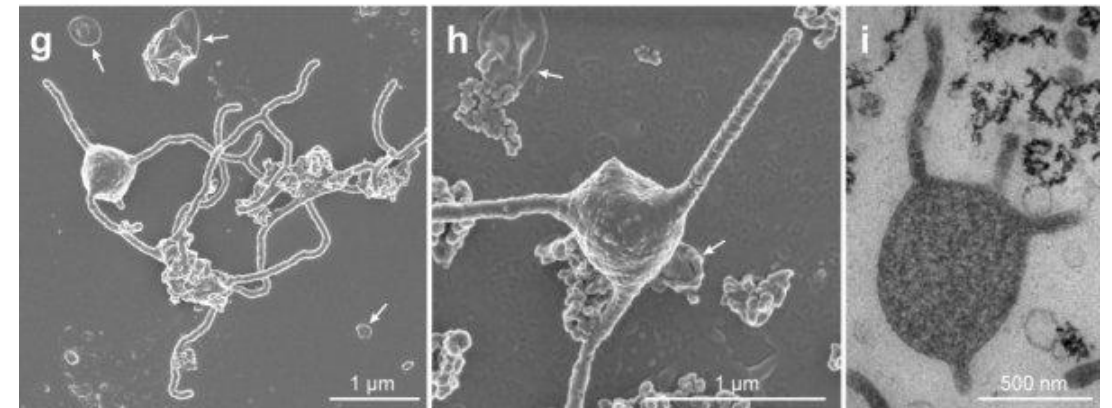
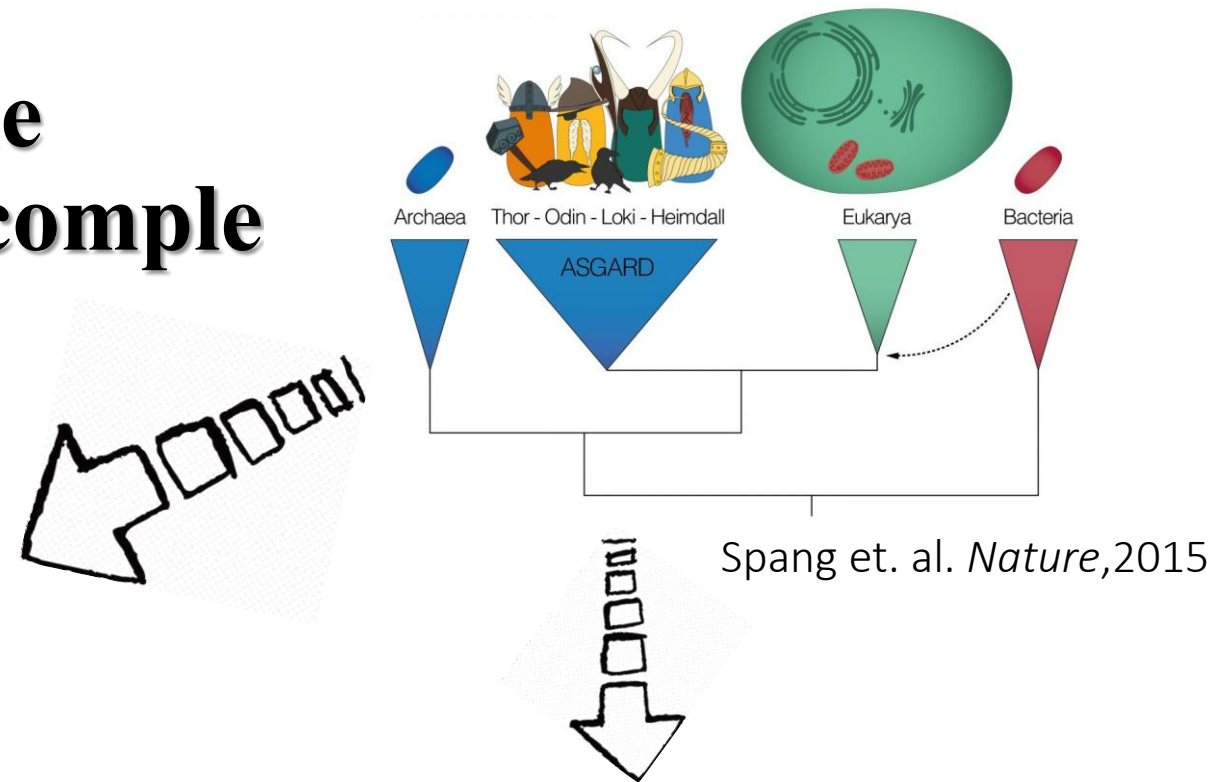


Asgard archaea illuminate the origin of eukaryotic cellular comple



López-García et. al. *NatMicrobiol*,2019

<https://www.nature.com/articles/s41564-020-0710-4>

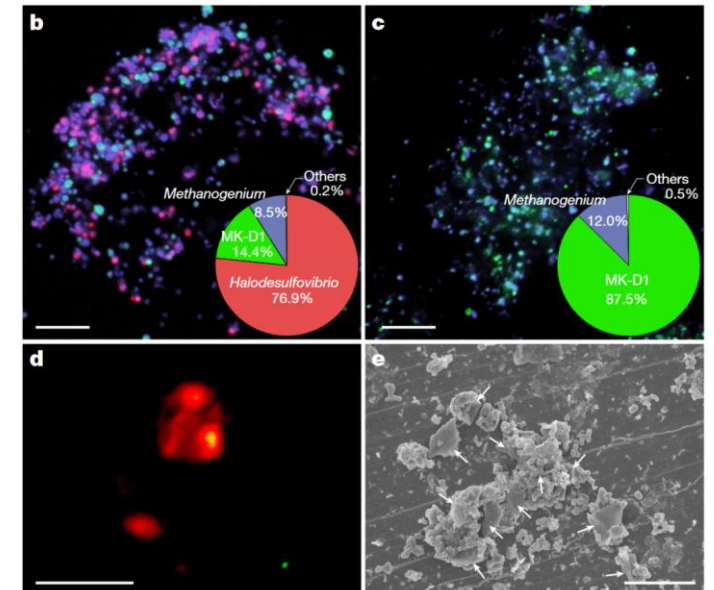


Imachi & Nobu et. al. *nature*,2020

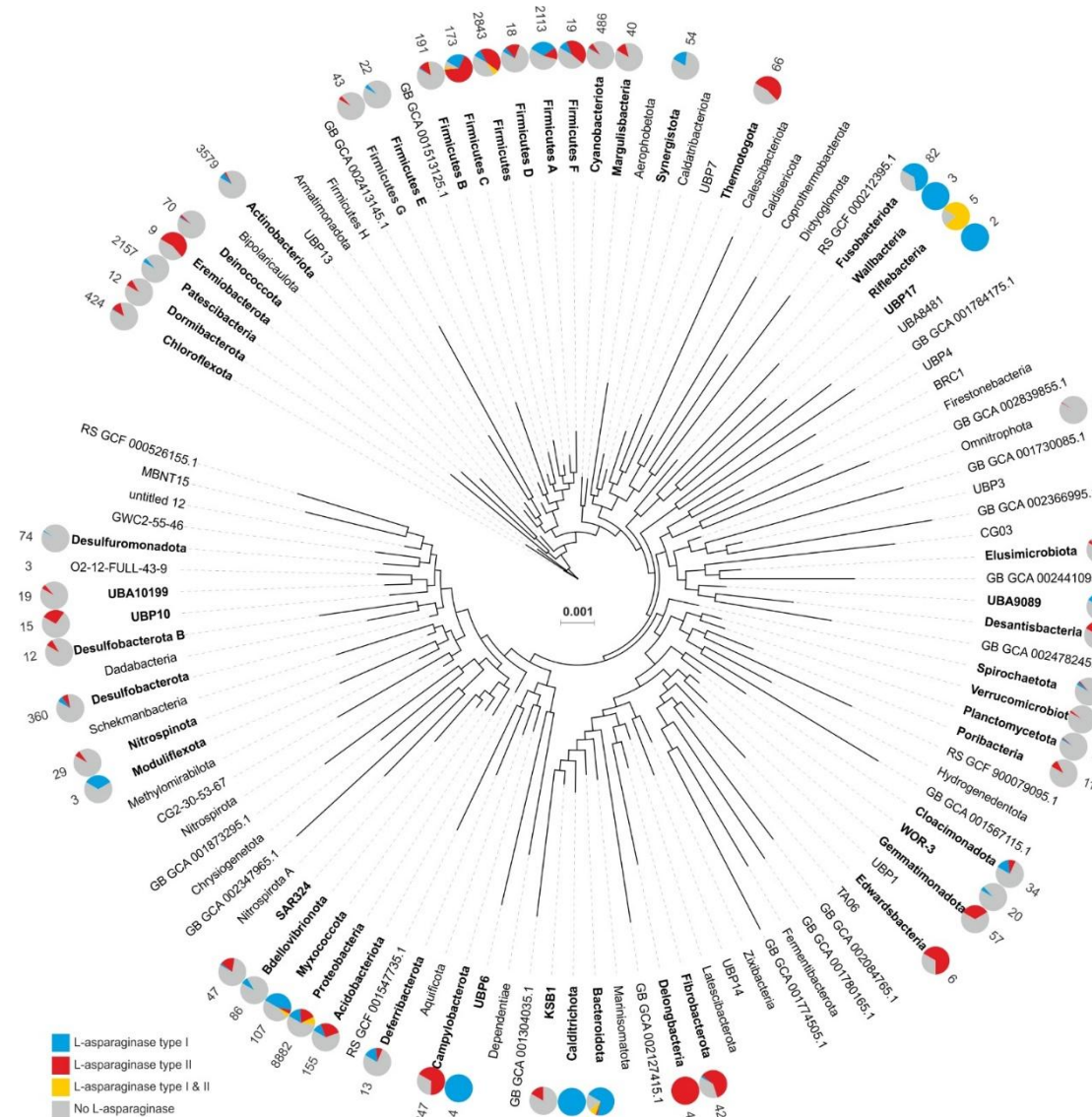
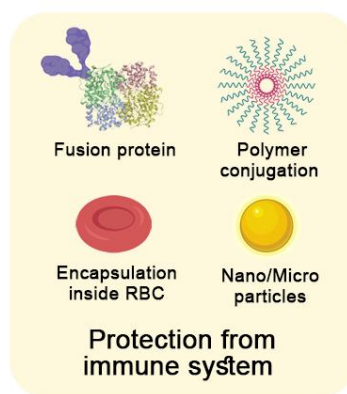
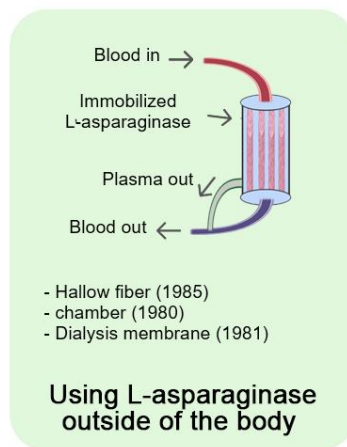
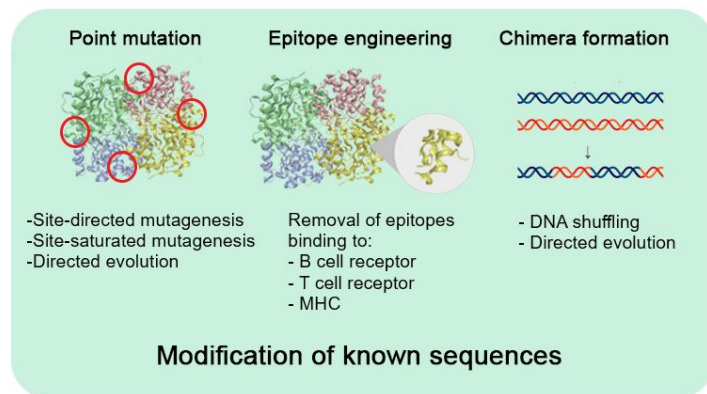
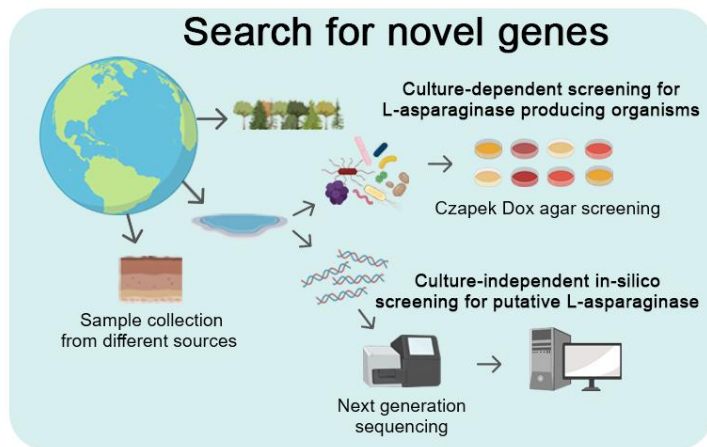
‘*CandidatusPrometheoarchaeum syntrophicum*’

- Pure co-culture of the target archaeon MK-D1 and Methanogenium after a 12-year study
- From bioreactor-based pre-enrichment of deep-sea sediments to a final 7 years of in vitro enrichment.
- Extremely slow growth rate and low cell yield.
- The culture consistently had a **30–60-day lag phase** and required more than 3 months to reach full growth: around 10^5 16S rRNA gene copies ml⁻¹
- The doubling time was estimated to be approximately **14–25 days**.

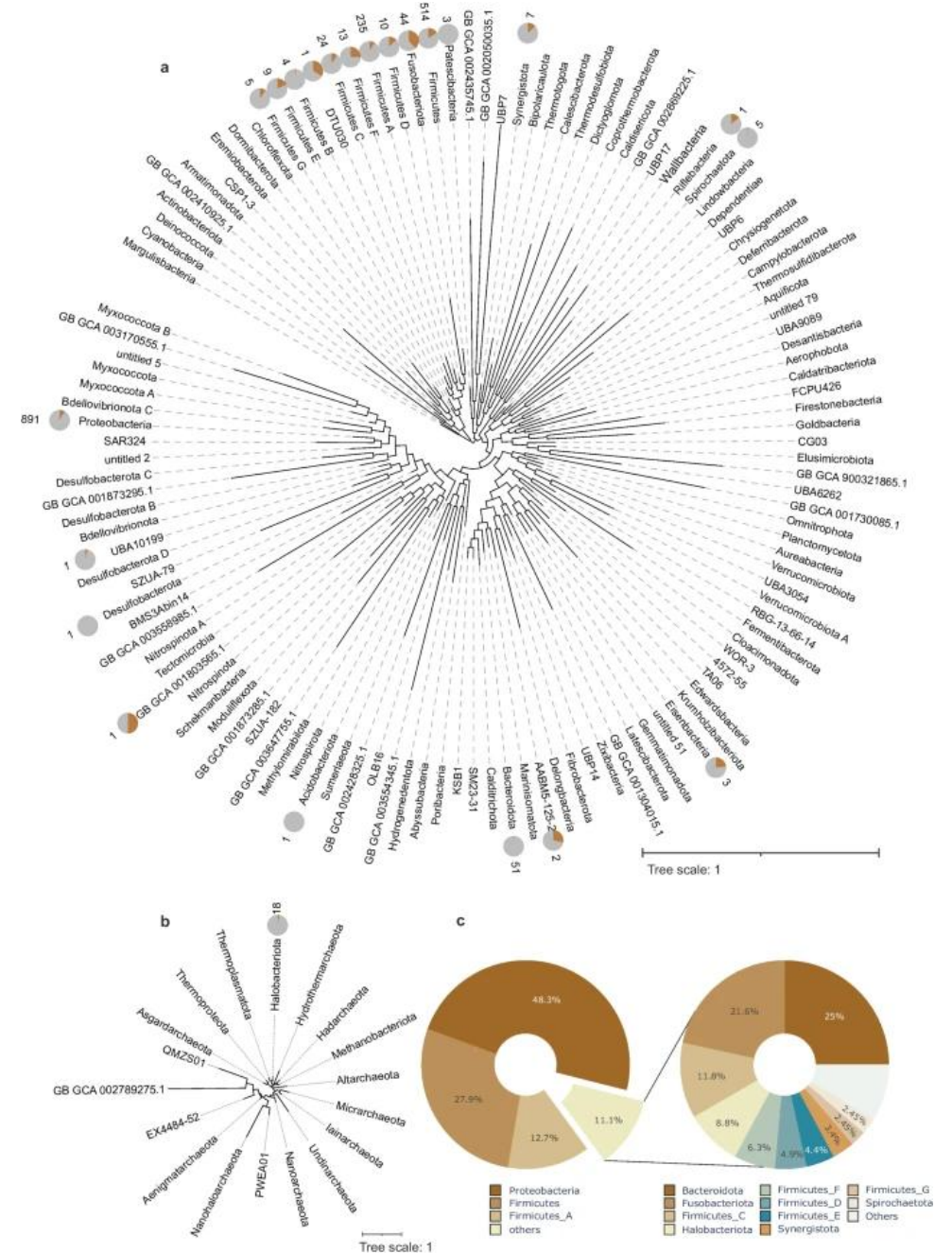
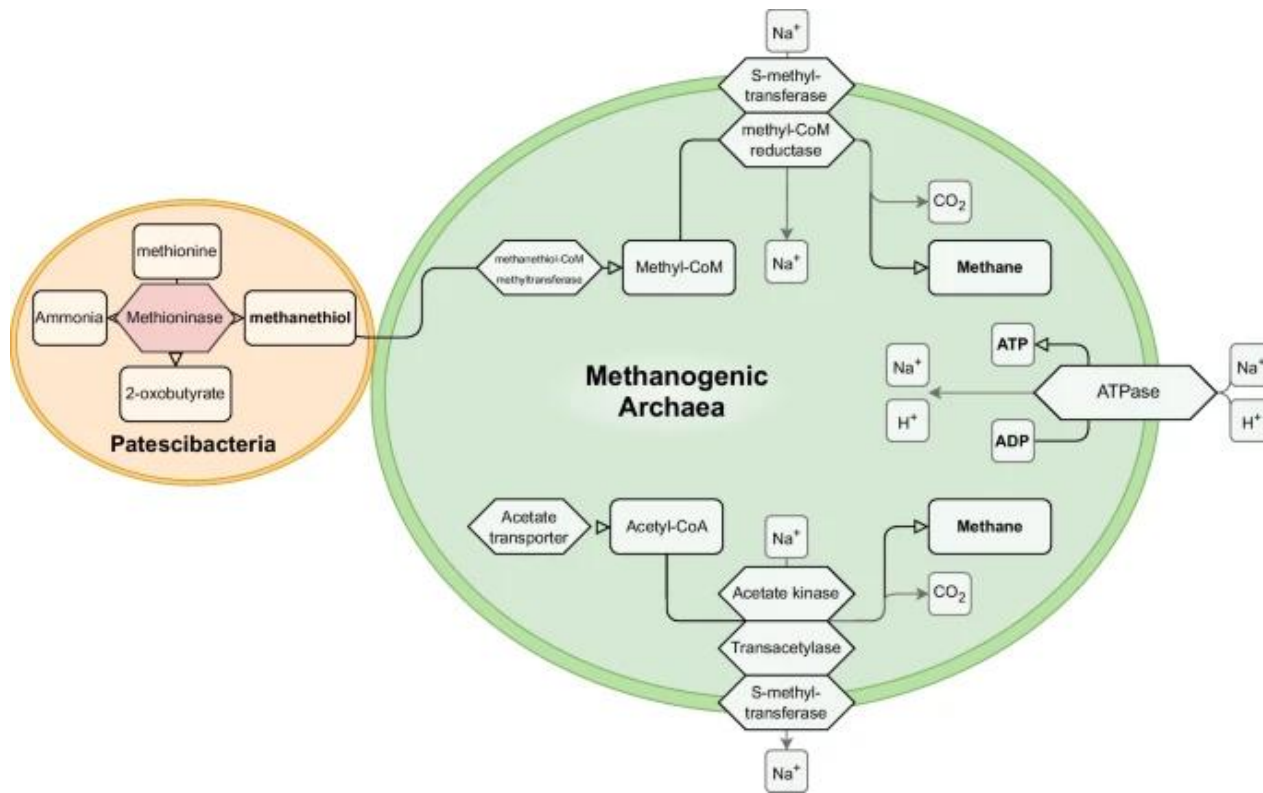
- Imachi, H., *et al. Nature*, 2020



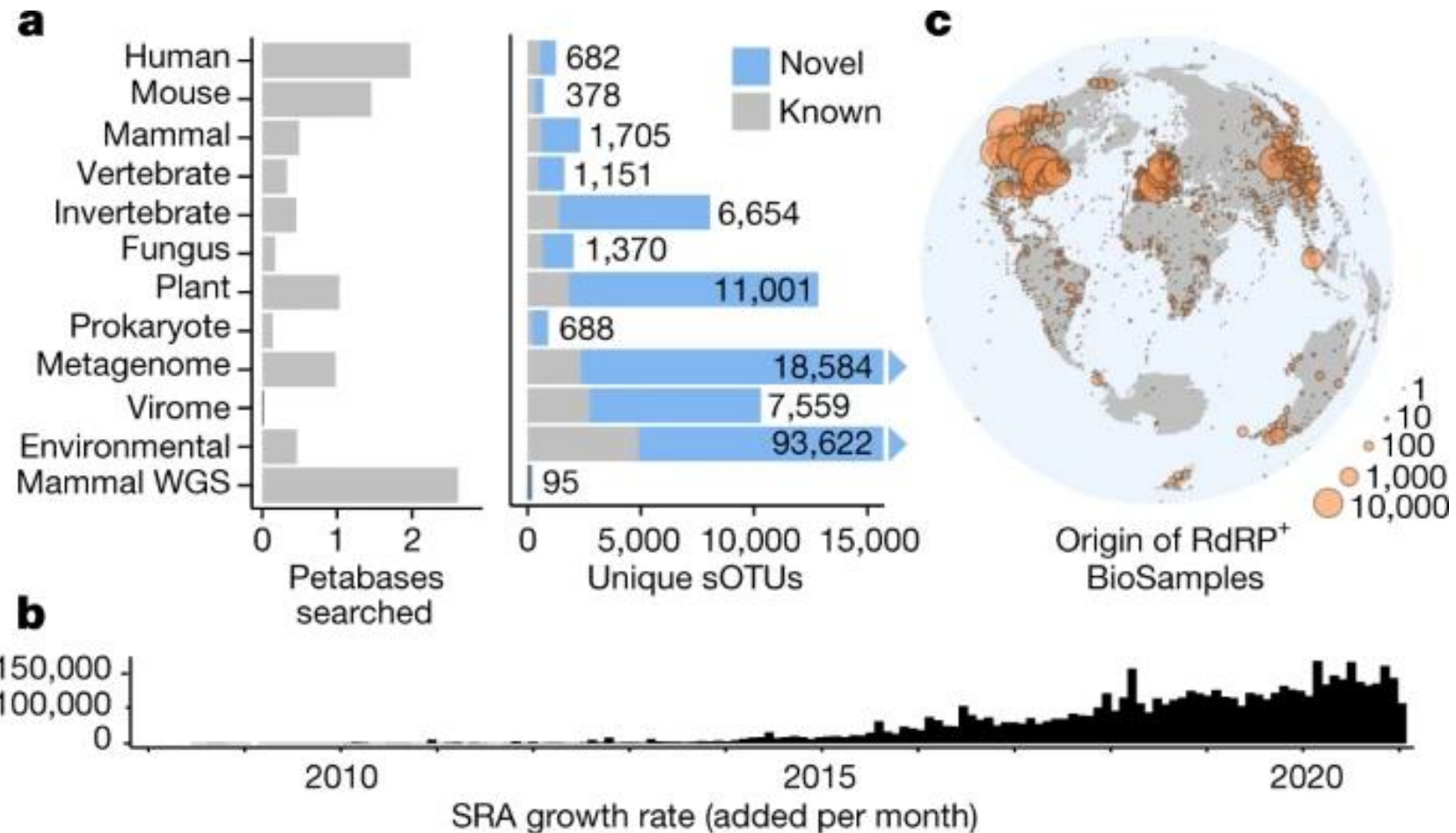
L-asparaginase

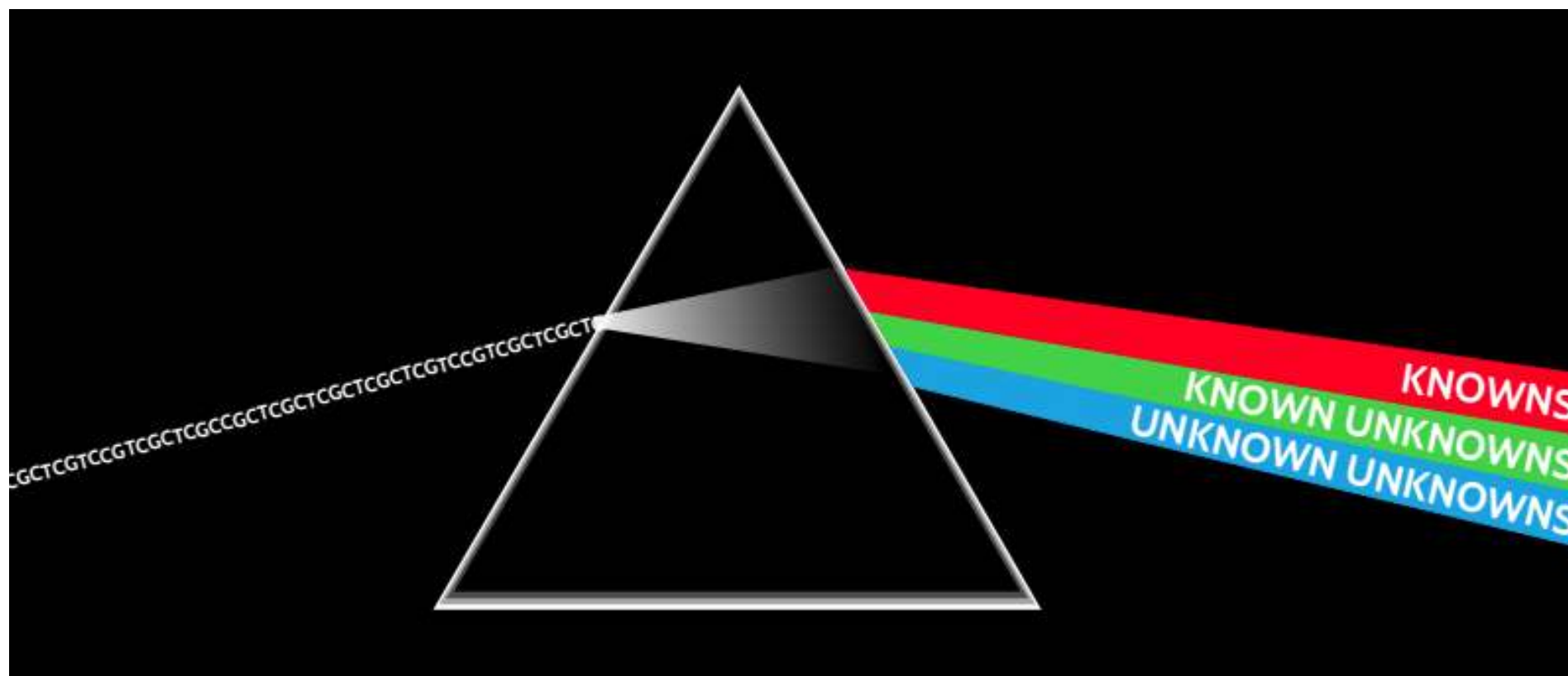


Methioninase involved in Metabolic Syntrophy

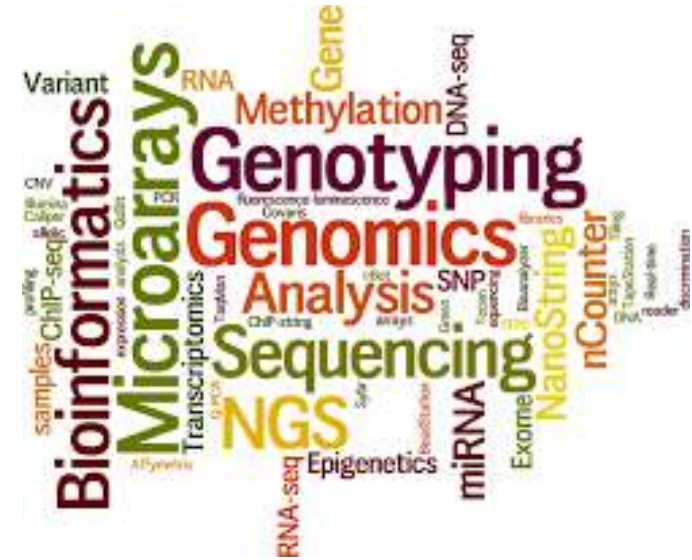


Petabase-scale sequence alignment catalyses viral discovery





Metagenomics



Metatranscriptome

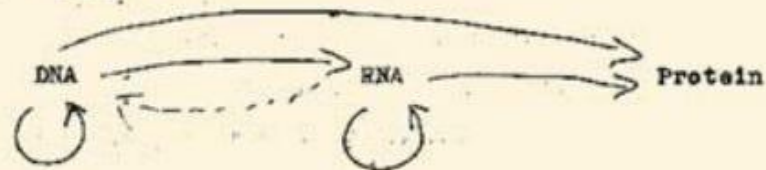
The transfer of information

Ideas on Protein Synthesis (Oct. 1956)

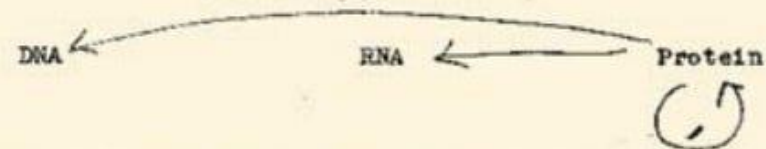
The Doctrine of the Triad.

The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it.

That is, we may be able to have



but never



where the arrows show the transfer of information.

Metatranscriptome

What to consider?

- Sampling
- Sample processing
- RNA extraction
- Replicates
- Library preparation (rRNA depletion)

Metatranscriptome

What to consider?

- Sampling
- Sample processing
- RNA extraction
- Replicates
- Library preparation (rRNA depletion)

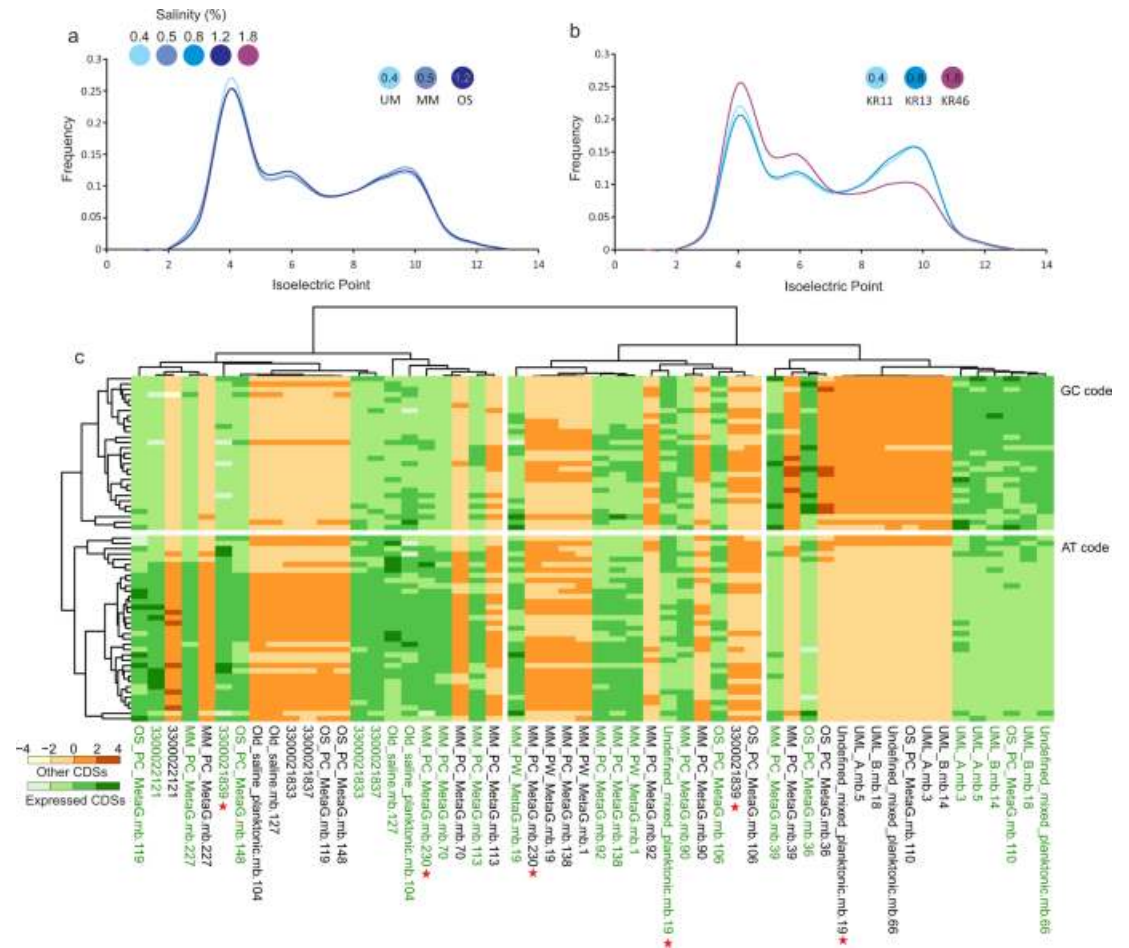
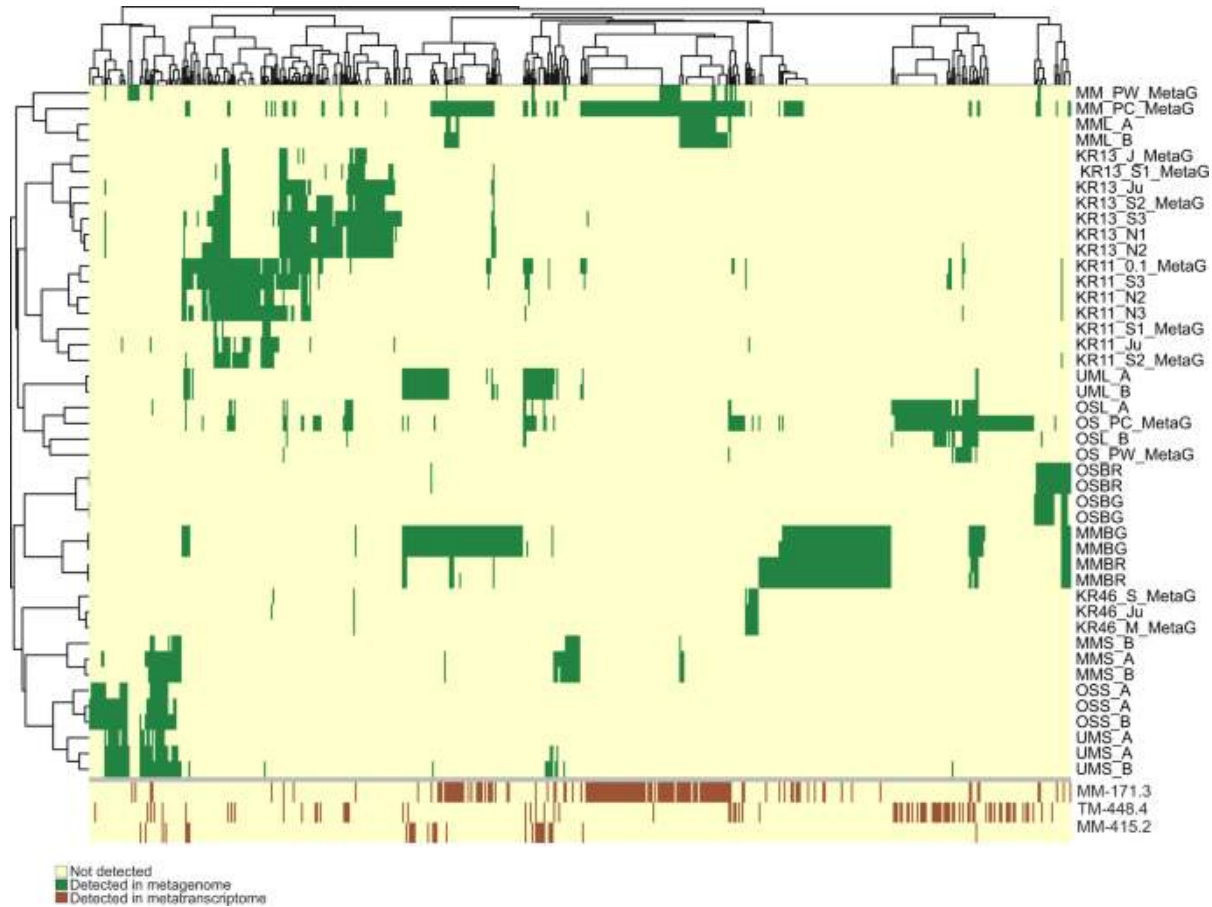
How to analyze?

- Remove rRNA or not?
- Assemble or not?
- Gene-resolved metatranscriptomics
- How to normalize?

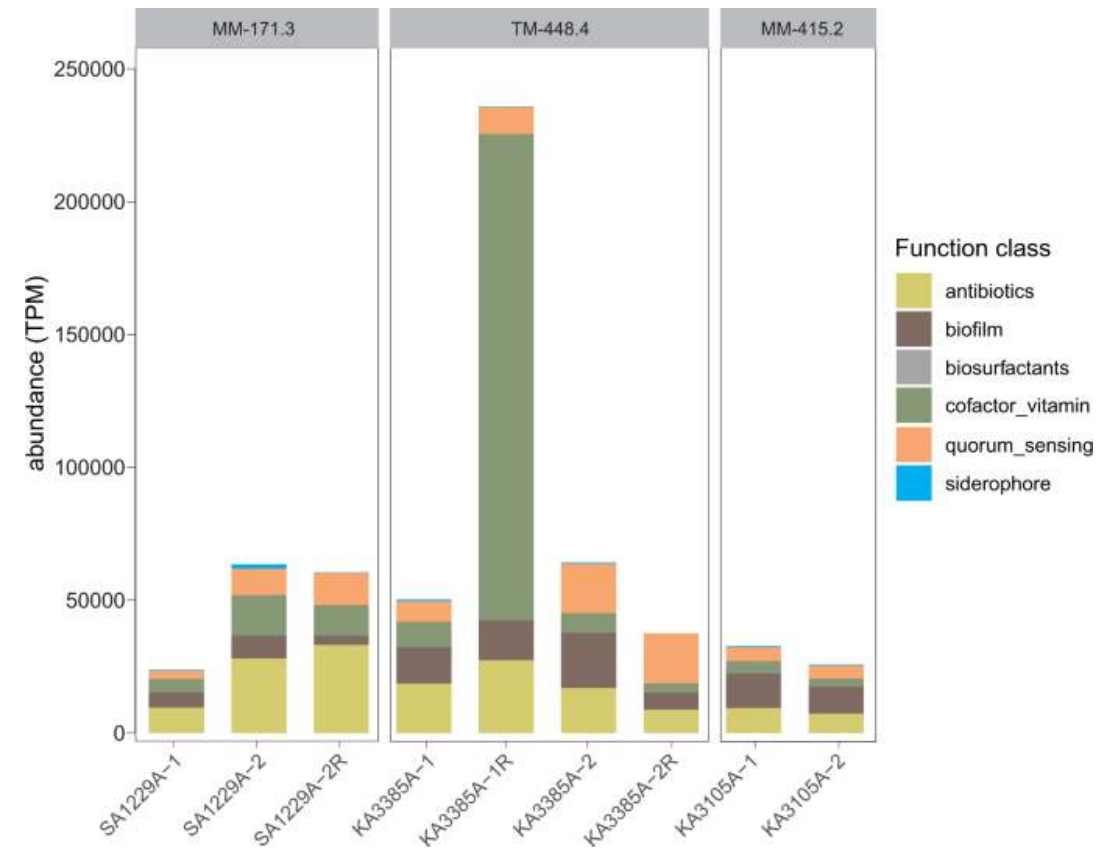
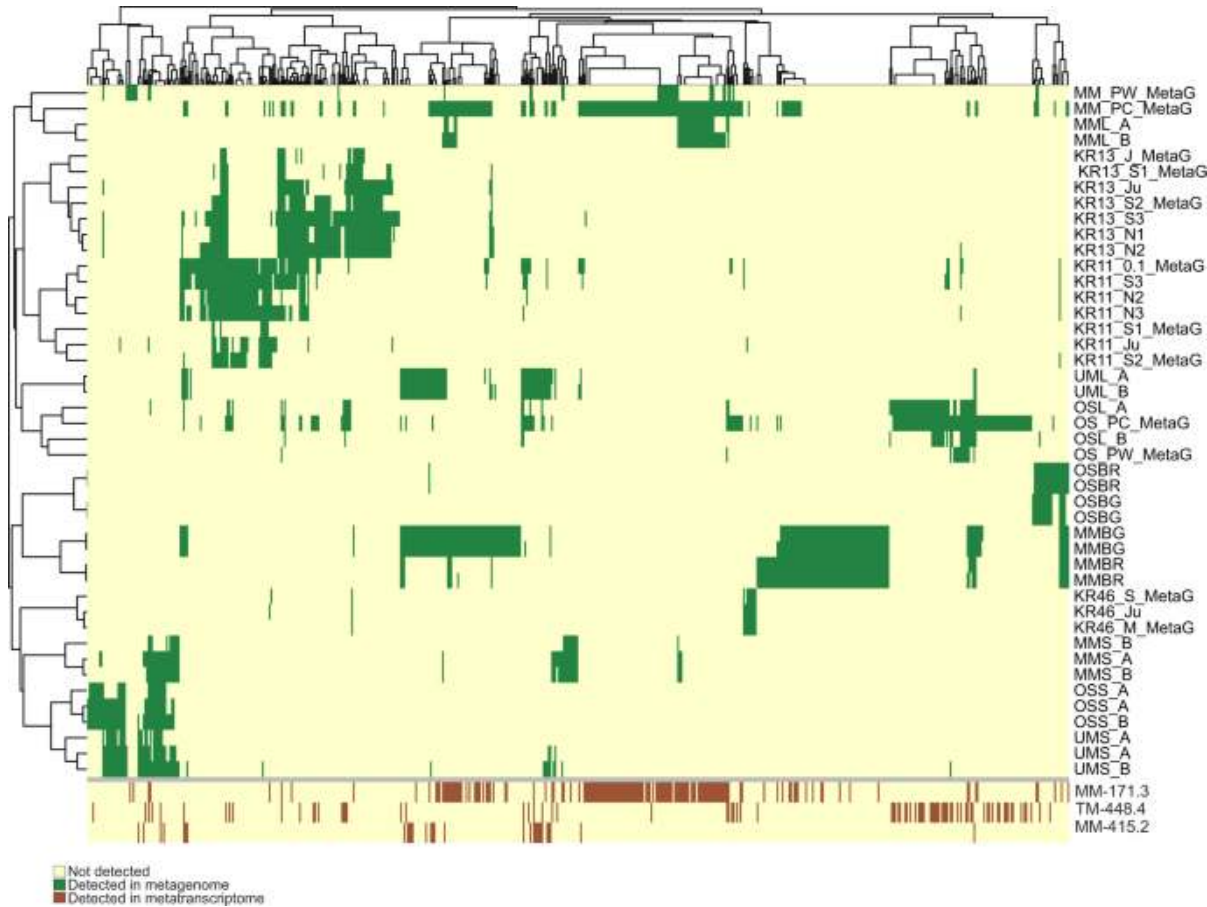
Some normalization methods

method	description
relative_abundance	(default) Percentage relative abundance of each genome, and the unmapped read percentage
mean	Average number of aligned reads overlapping each position on the genome
trimmed_mean	Average number of aligned reads overlapping each position after removing the most deeply and shallow-ly covered positions. See <code>--trim-min</code> / <code>--trim-max</code> to adjust.
coverage_histogram	Histogram of coverage depths
covered_bases	Number of bases covered by 1 or more reads
variance	Variance of coverage depths
length	Length of each genome in base pairs
count	Number of reads aligned to each genome. Note that supplementary alignments are not counted.
reads_per_base	Number of reads aligned divided by the length of the genome
anir	Average BLAST-like identity of mapped reads
rpkm	Reads mapped per kilobase of genome, per million mapped reads
tpm	Transcripts Per Million as described in Li et al 2010 https://doi.org/10.1093/bioinformatics/btp692

The transfer of information



The transfer of information

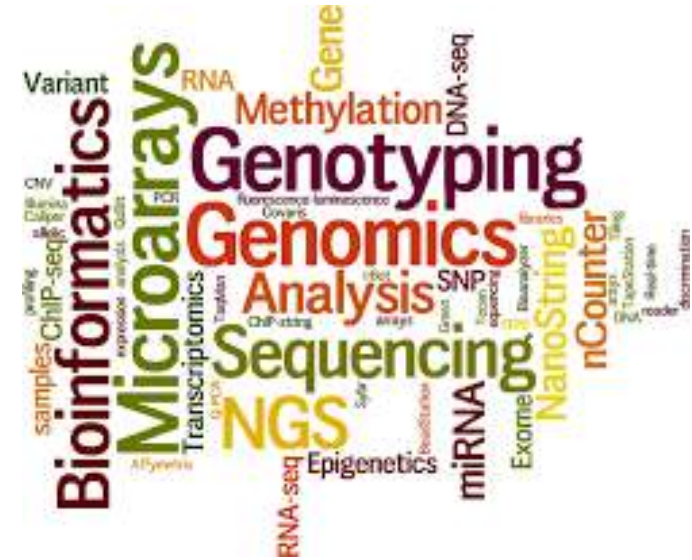


“multi-omics” approach to answer eco-evolutionary questions

Single-cell
amplified genomes

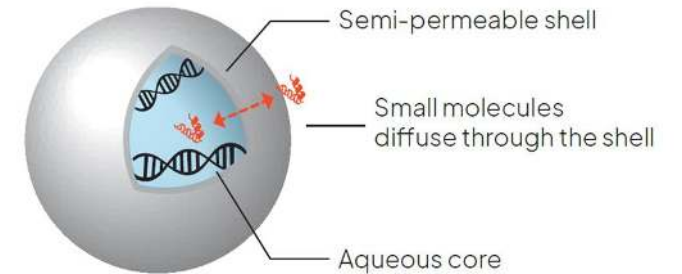
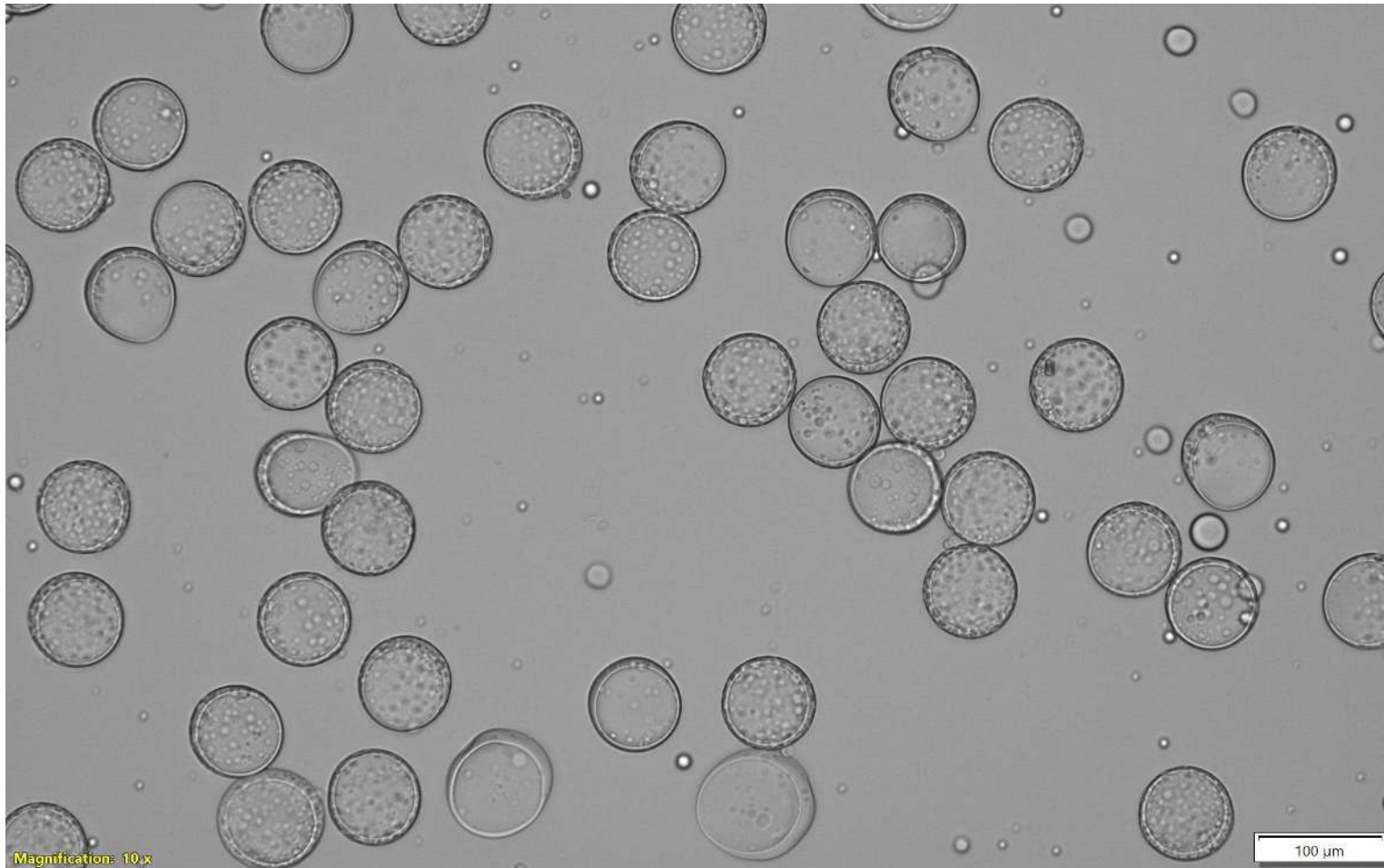


Metagenomics

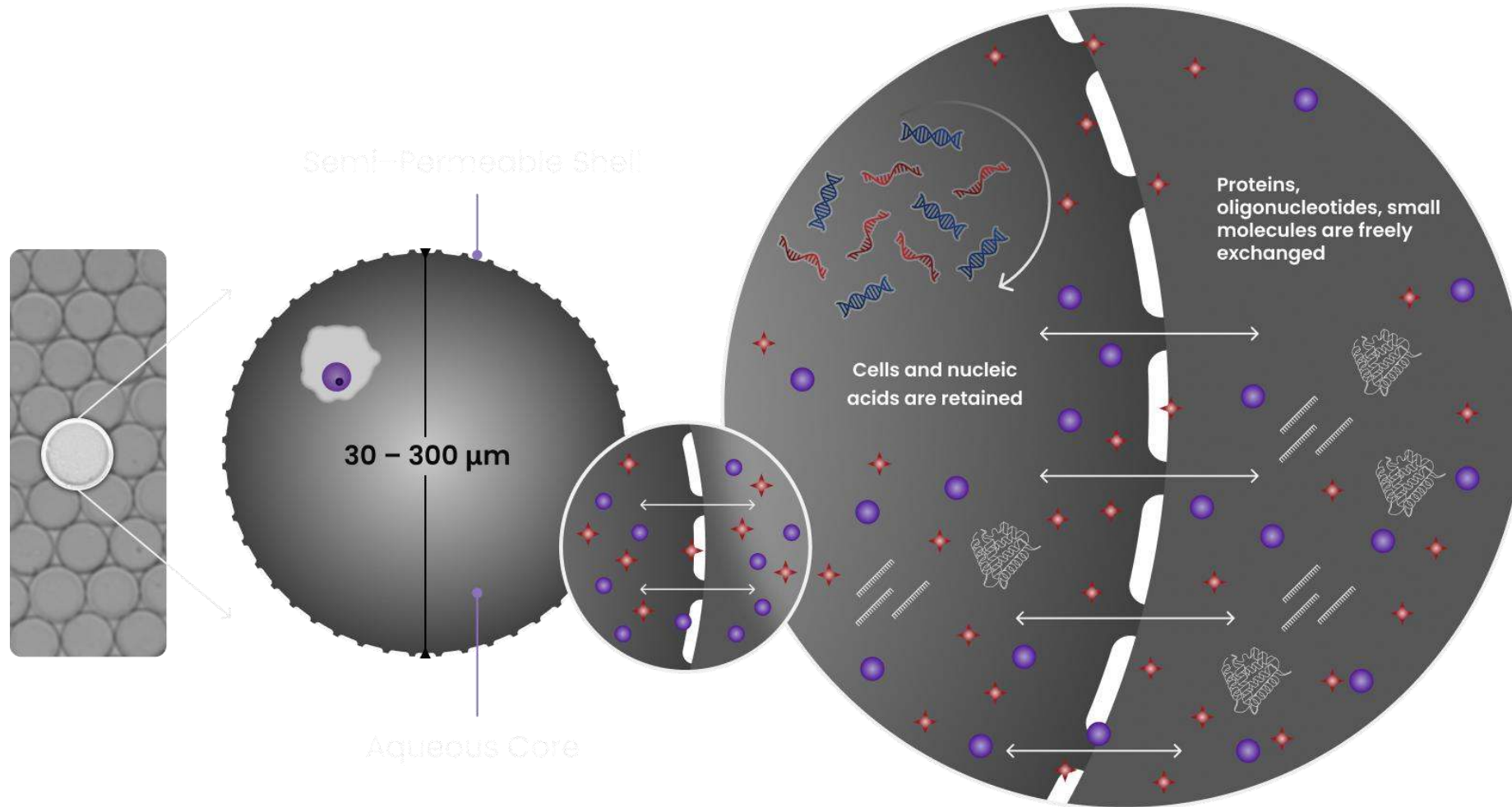


Metatranscriptome

Compartmentalized metagenomics

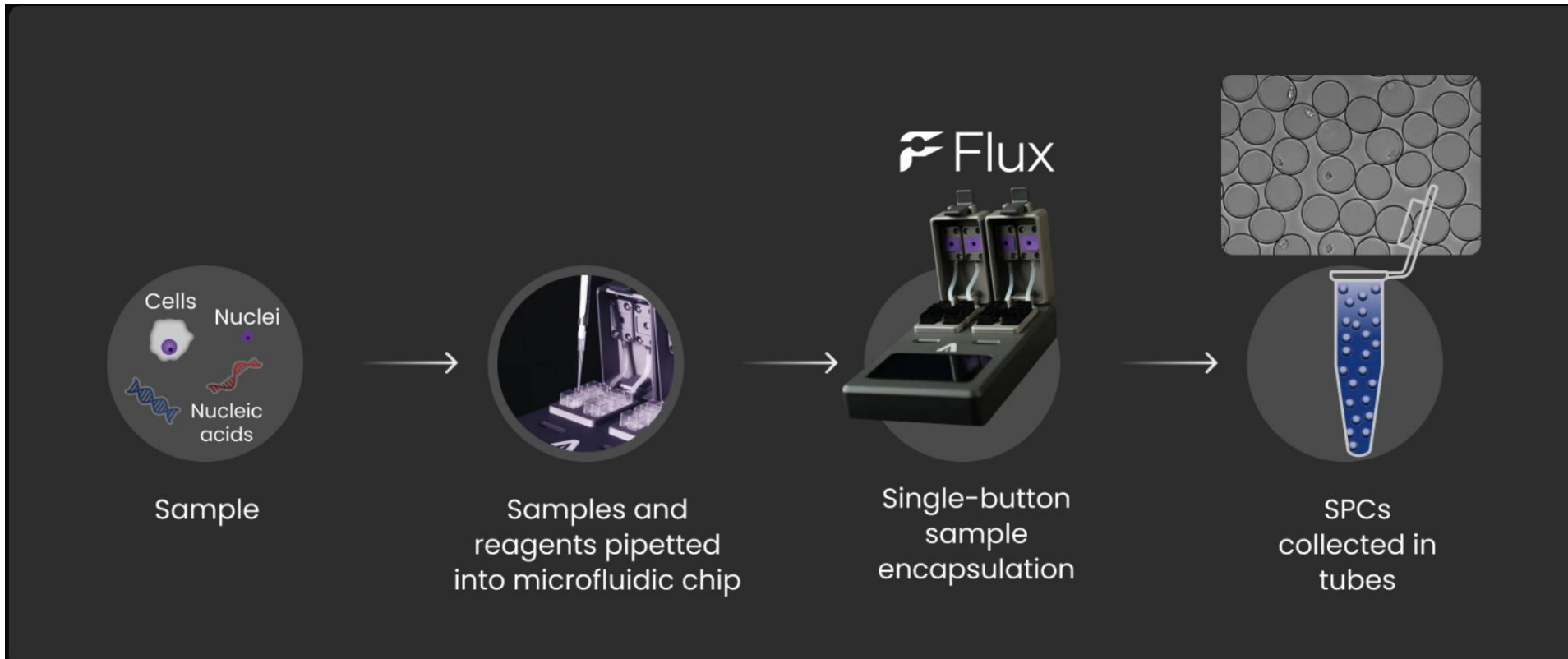


Semi permeable capsules



REAGENT EXCHANGE

Semi permeable capsules



Removed unpublished results

Metagenomics is it a big deal!?



BIONOMICS-MMlab
MULTIPARTITE PARASITIC INTERACTIONS

Maliheh Mehrshad

Department of Aquatic Sciences and Assessment,

