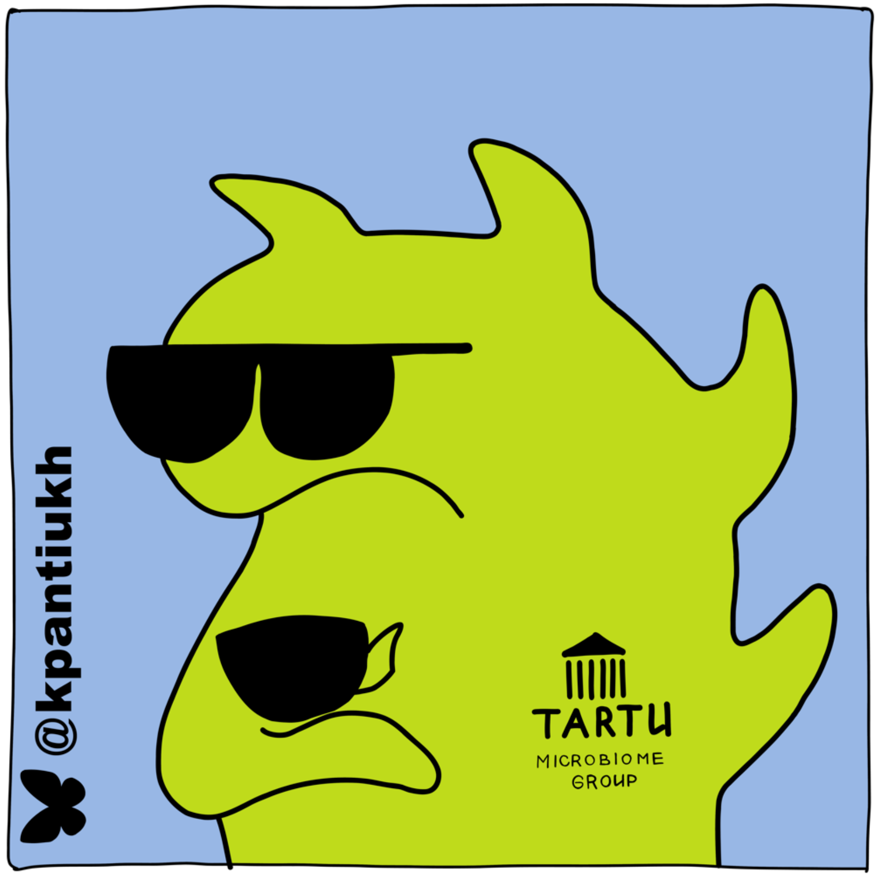


Metagenomics

Lecture 1

History of metagenomics, landmark studies, and development of the field. Course structure and AI policy

Kateryna Pantiukh
pantiukh@ut.ee

GitHub

Introduction

Kateryna Pantiukh

Writing my PhD thesis

University of Tartu, Estonia

Estonian Biobank

estonian genome center
university of tartu

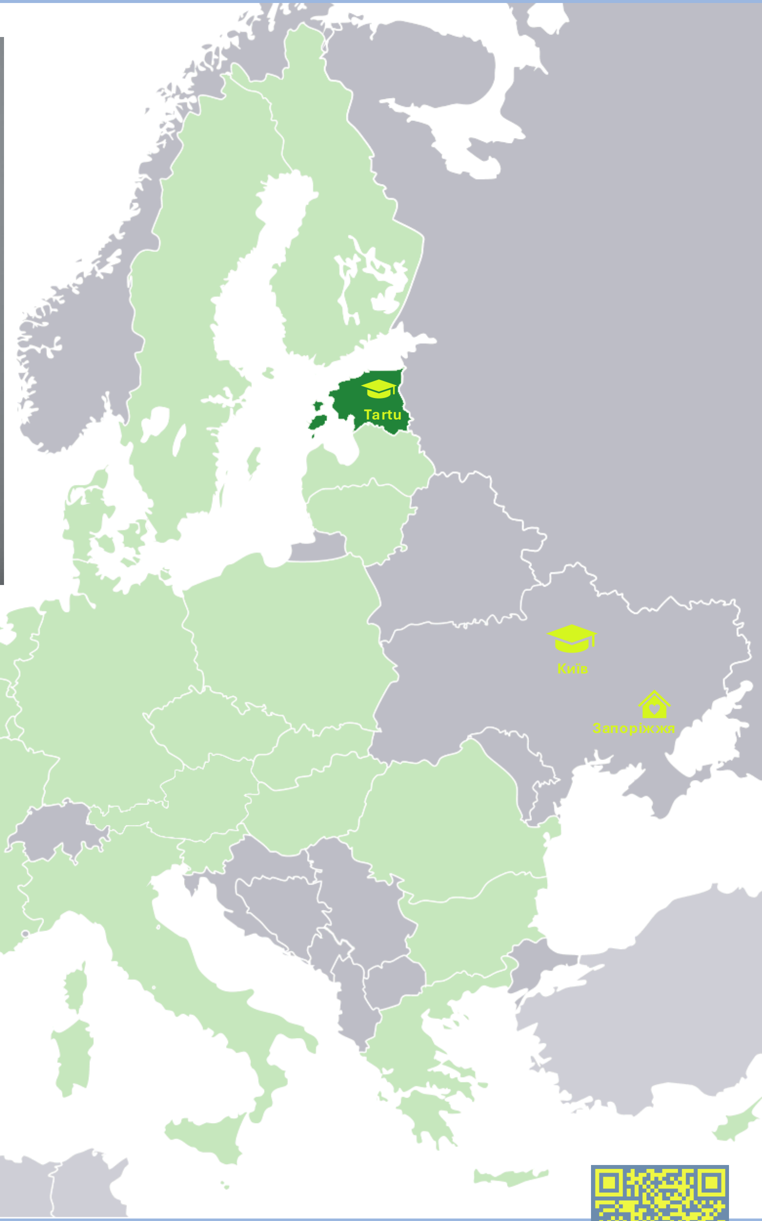
211 187

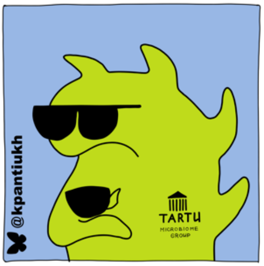
biobank participants

~20% of Estonian poplation

my research interests

Human gut microbiome
a community with big impact





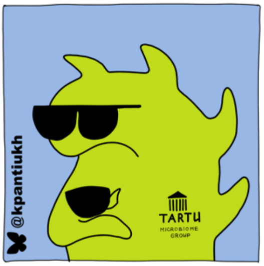
Introduction

estonian genome center
university of tartu

211 187

biobank participants

~20% of Estonian poplation

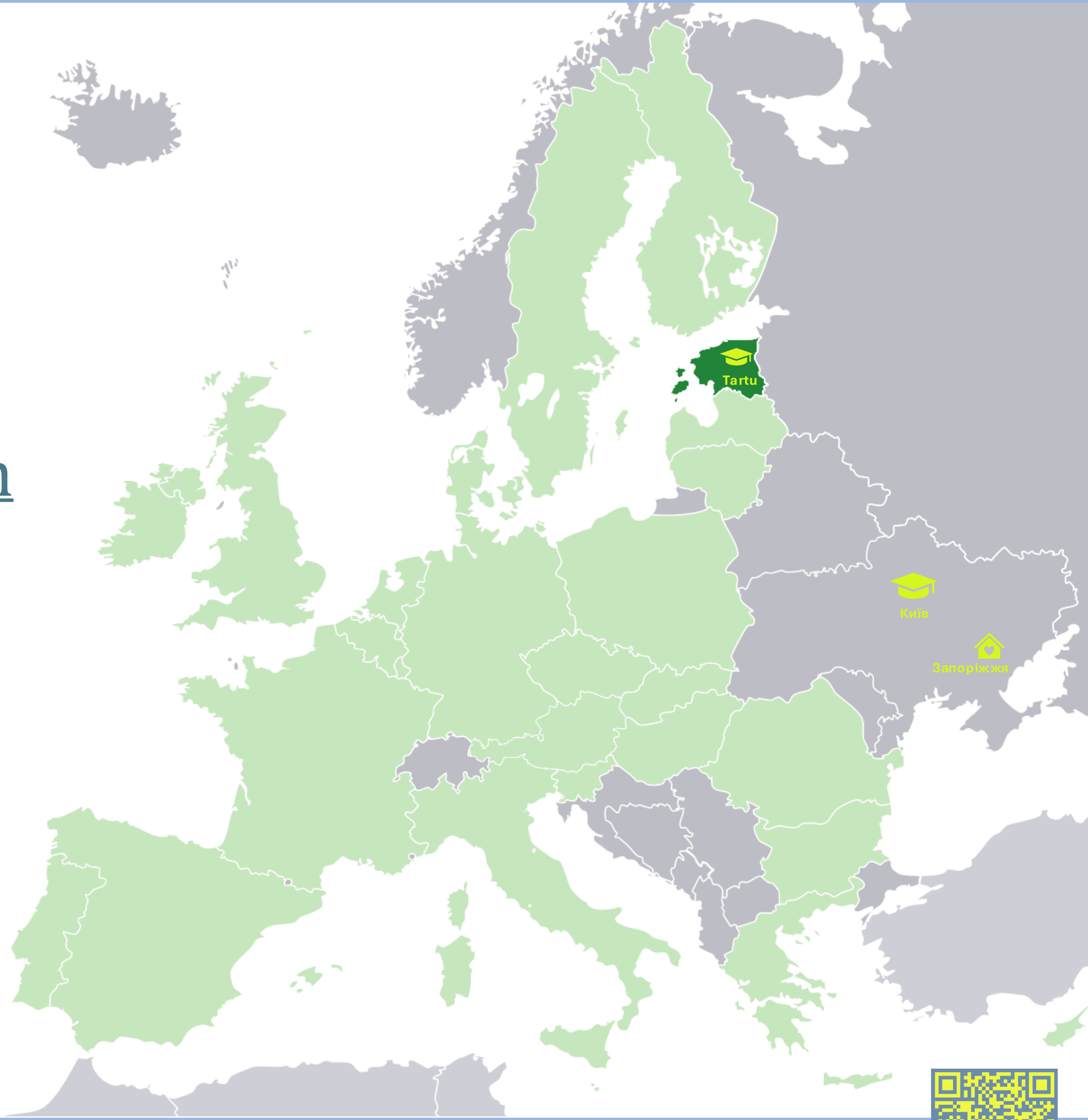


Introduction

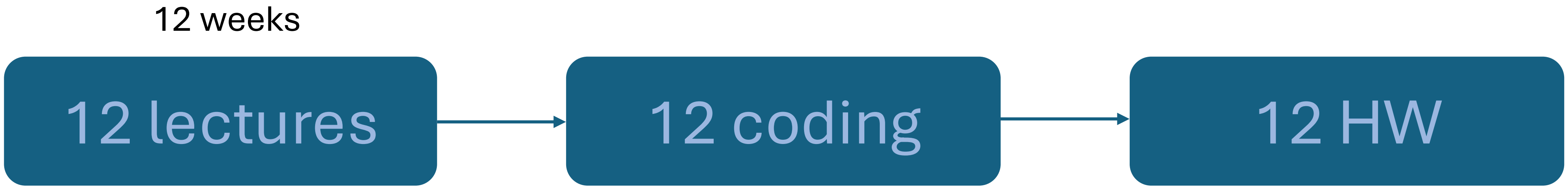
www.slido.com

2204536

https://app.sli.do/event/fsZSLCjqthSYtFLVhAN9aK



Course info



12 weeks

12 lectures

12 coding

12 HW

Background knowladge

01-09: practical exersize
10-11: final project preparation
12: selected project presentation

finalise the practical exercise task,
+ small supplementary tasks
+ optional development extensions

Grading Policy

Assignments: 60 points

12 weekly tasks * 5 points each

Final project: 40 points

20 points for the coding + 20 points for the written report

Extra (bonus) points: 10 points

Grading Policy

Assignments: 60 points

12 weekly tasks * 5 points each

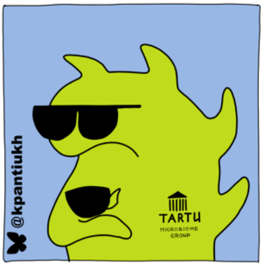
Final project: 40 points

20 points for the coding + 20 points for the written report

Extra (bonus) points: 10 points

Evaluation criteria:

- A: 90-100 points
- B: 75-89 points
- C: 65-74 points
- D: 50-64 points
- F: below 50 points



Grading Policy

Assignments: 60 points

12 weekly tasks * 5 points each

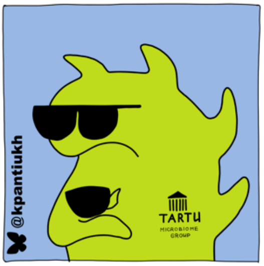
Final project: 40 points

20 points for the coding + 20 points for the written report

Extra (bonus) points: 10 points

Evaluation criteria:

- A: 90-100 points
- B: 75-89 points
- C: 65-74 points
- D: 50-64 points
- F: below 50 points



Grading Policy

Assignments: 60 points

12 weekly tasks * 5 points each

Final project: 40 points

20 points for the coding + 20 points for the written report

Extra (bonus) points: 10 points

Evaluation criteria:

- A: 90-100 points
- B: 75-89 points
- C: 65-74 points
- D: 50-64 points
- F: below 50 points

No retake option is offered, as the course relies primarily on continuous assessment and coding-based assignments

Grading Policy

Assignments: 60 points

12 weekly tasks * 5 points each

Final project: 40 points

20 points for the coding + 20 points for the written report

Extra (bonus) points: 10 points

Evaluation criteria:

- A: 90-100 points
- B: 75-89 points
- C: 65-74 points
- D: 50-64 points
- F: below 50 points

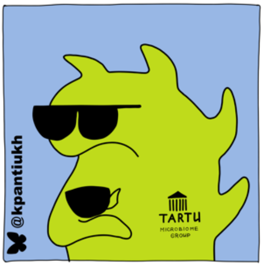
No retake option is offered, as the course relies primarily on continuous assessment and coding-based assignments

Students who have not earned any points by the end of the **fourth lecture** will no longer be evaluated

AI Policy

The use of AI tools (such as ChatGPT, GitHub Copilot, and similar systems) **is allowed** and actively encouraged as **learning companions** and **coding assistants**, provided they are used responsibly and with a clear awareness of their limitations.

Copilot is recommended for use within Visual Studio Code
(5 extra points may be earned for helping other students with installation)



AI Policy

The use of AI tools (such as ChatGPT, GitHub Copilot, and similar systems) **is allowed** and actively encouraged as **learning companions** and **coding assistants**, provided they are used responsibly and with a clear awareness of their limitations.

Copilot is recommended for use within Visual Studio Code
(5 extra points may be earned for helping other students with installation)

AI-generated answers can be **incomplete, misleading, or factually incorrect**, even when they appear confident and well-written.

"THINKING LOG"
process-oriented documentation

extrapoints for each document

Questions?

If you have questions:

Slack:

pantiukh@ut.ee
pantiukh@gmail.com

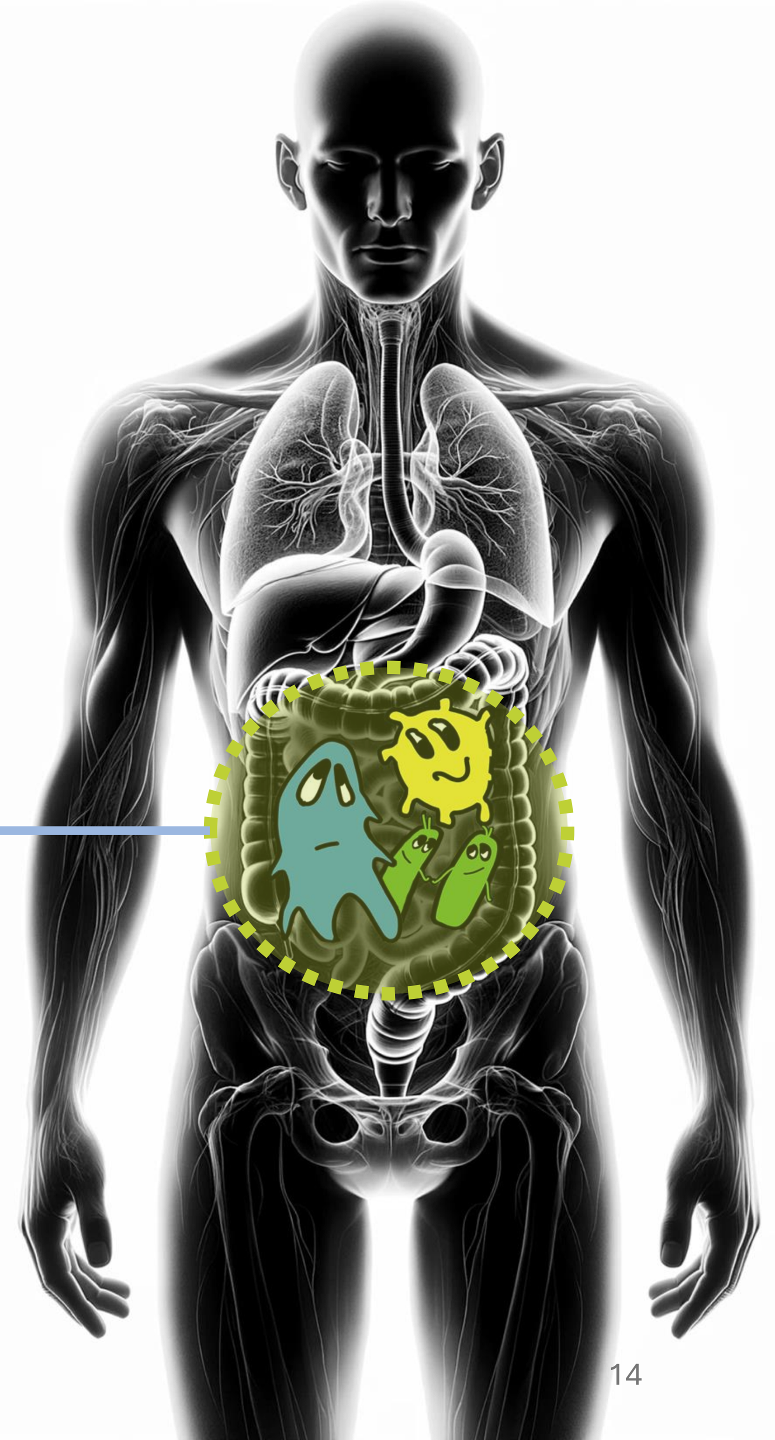
X/BlueSky: kpantiukh

Microbiome

Community of microorganisms

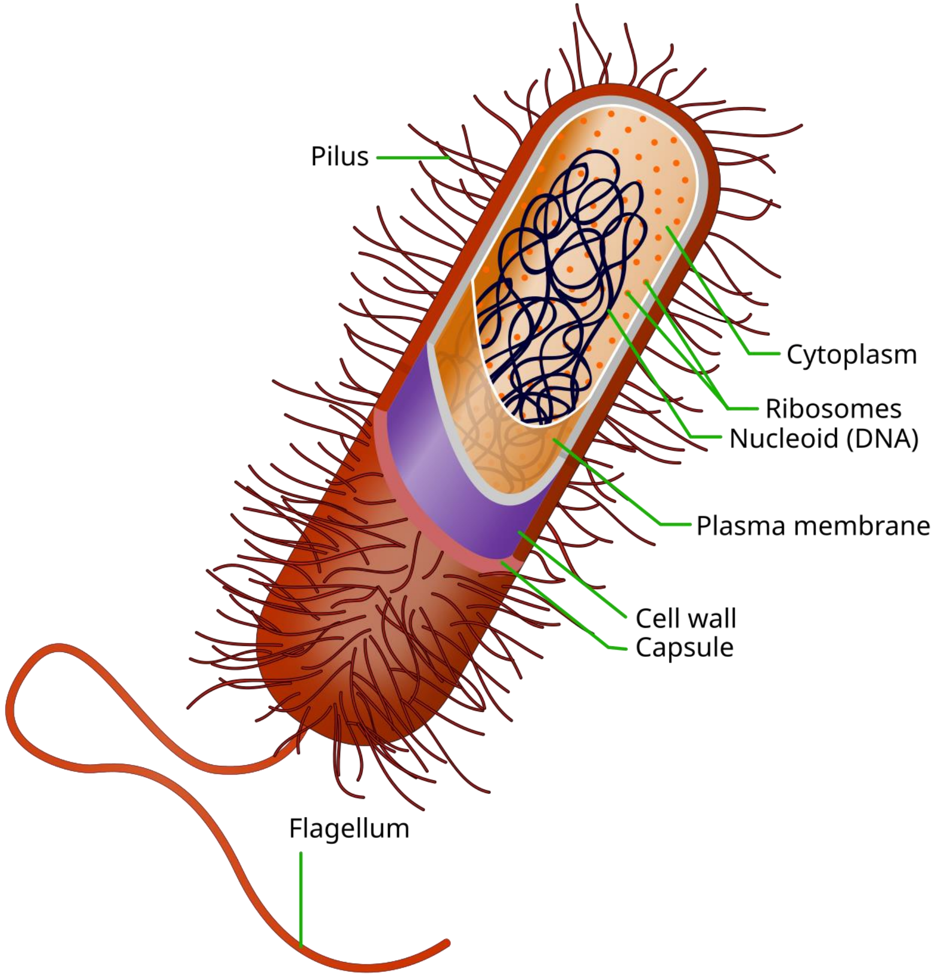
Bacteria
Archaea
Viruses
Fungi
Single cell
eukaryotes

MICROBIOME



Bacteria

prokaryote microorganisms



Pilus

Cytoplasm

Ribosomes

Nucleoid (DNA)

Plasma membrane

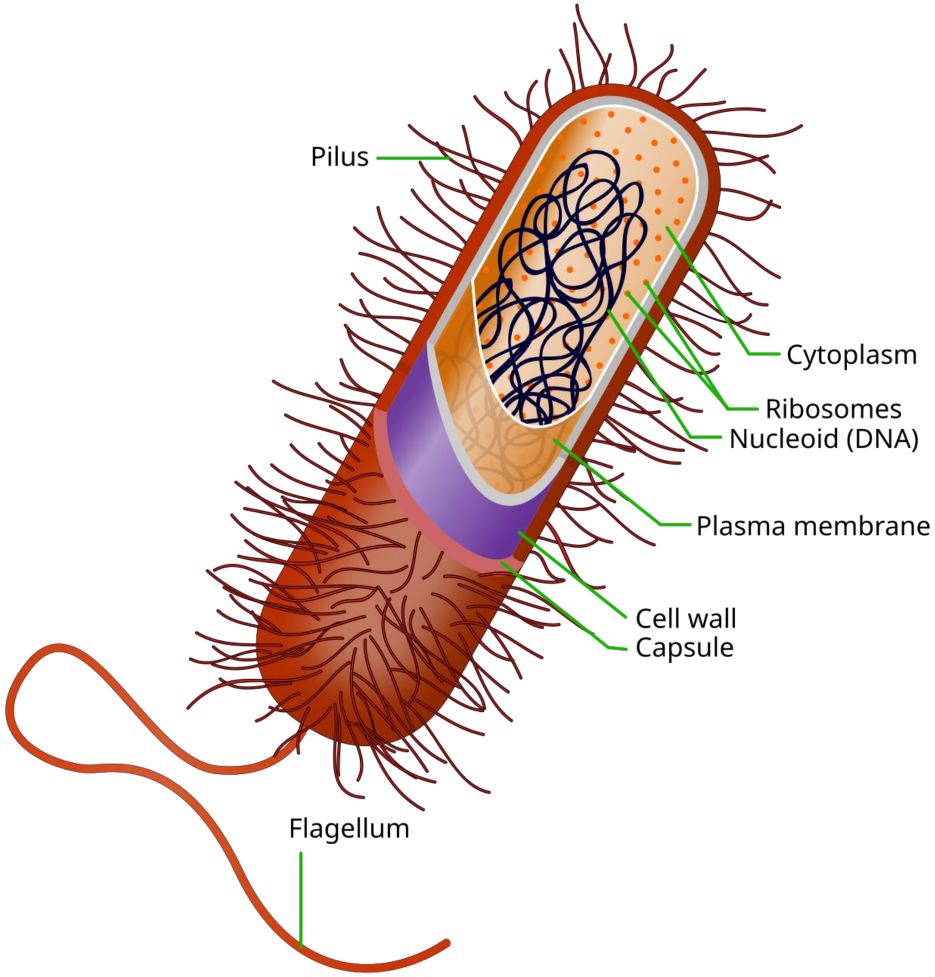
Cell wall

Capsule

Flagellum

Bacteria

prokaryote microorganisms



•**Single-celled life**
Bacteria are unicellular organisms.

•**No sex**
do not reproduce sexually -> species concept.

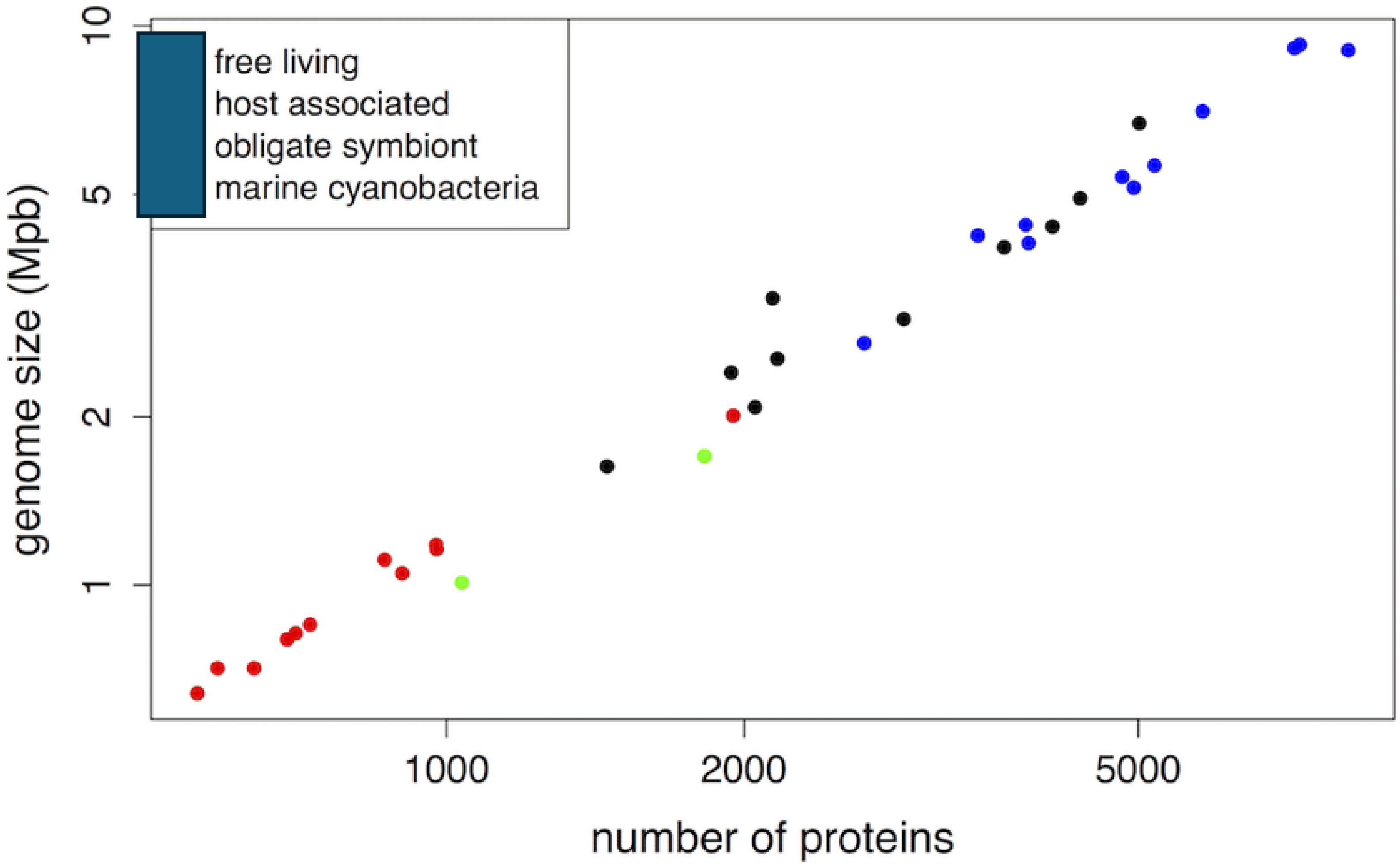
•**Cell size**
A typical bacterial cell is about **0.5–5 µm** in length. Small enough to be invisible eyes

•**Genome**
Single **circular DNA** molecule, usually **~1–6 Mbp** in size, encoding a few hundred to several thousand genes.

•**Extra genetic elements**
Many bacteria carry **plasmids** (small circular DNA, ~1–200 kbp) and interact with **bacteriophages** (viruses of bacteria), which move genes around and drive rapid evolution.

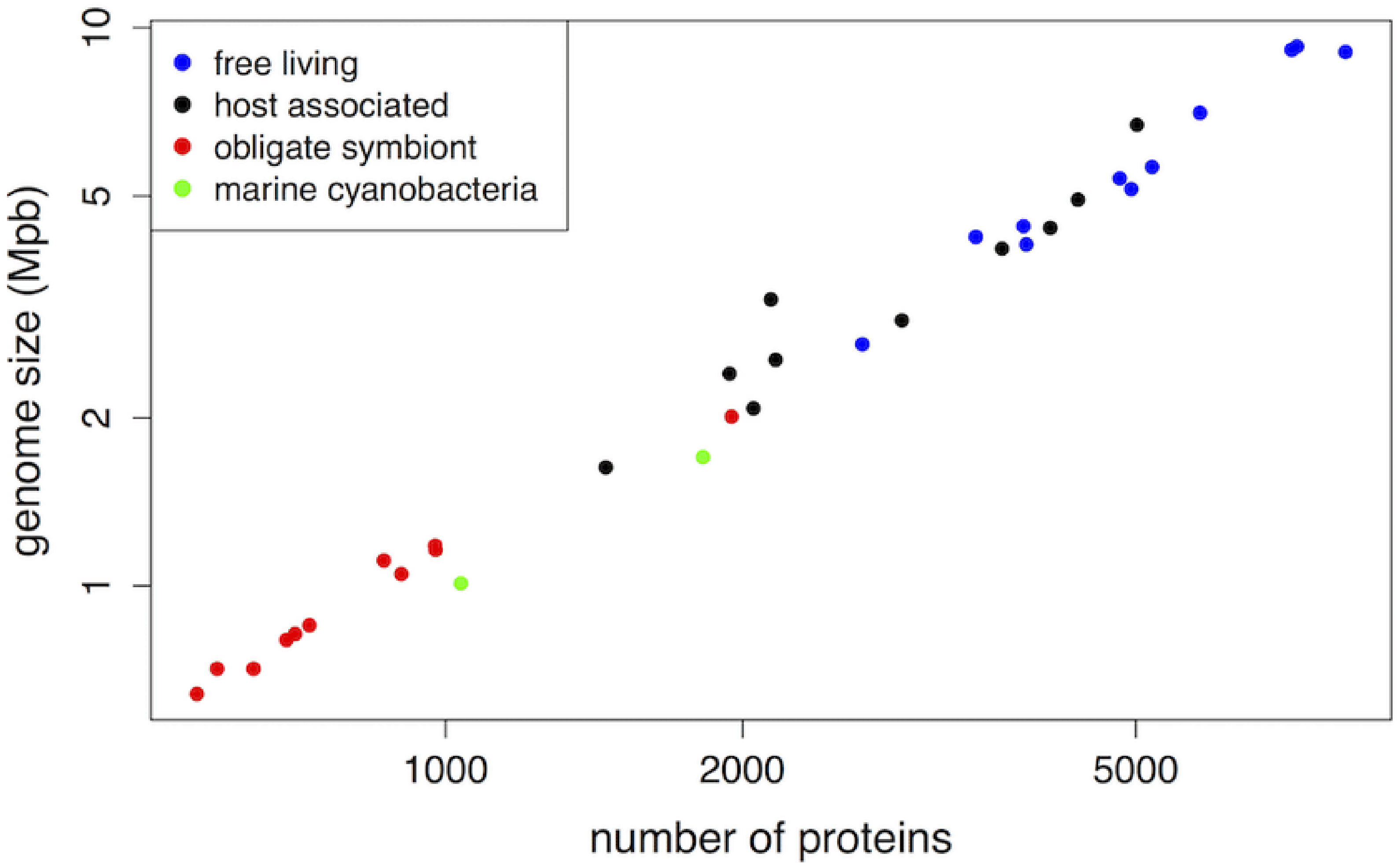
Bacteria

Genome size and environment



Bacteria

Genome size and environment

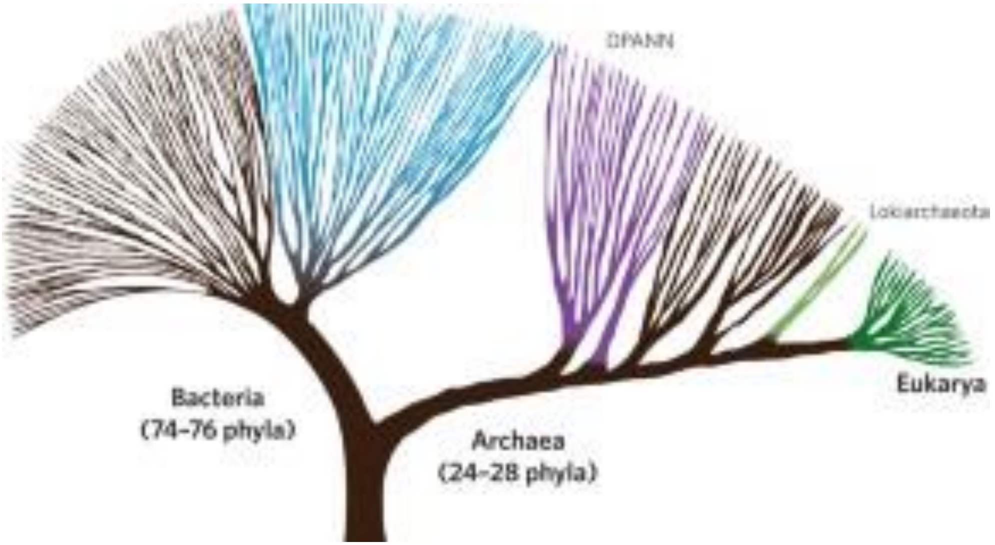


Archaea

prokaryote microorganisms

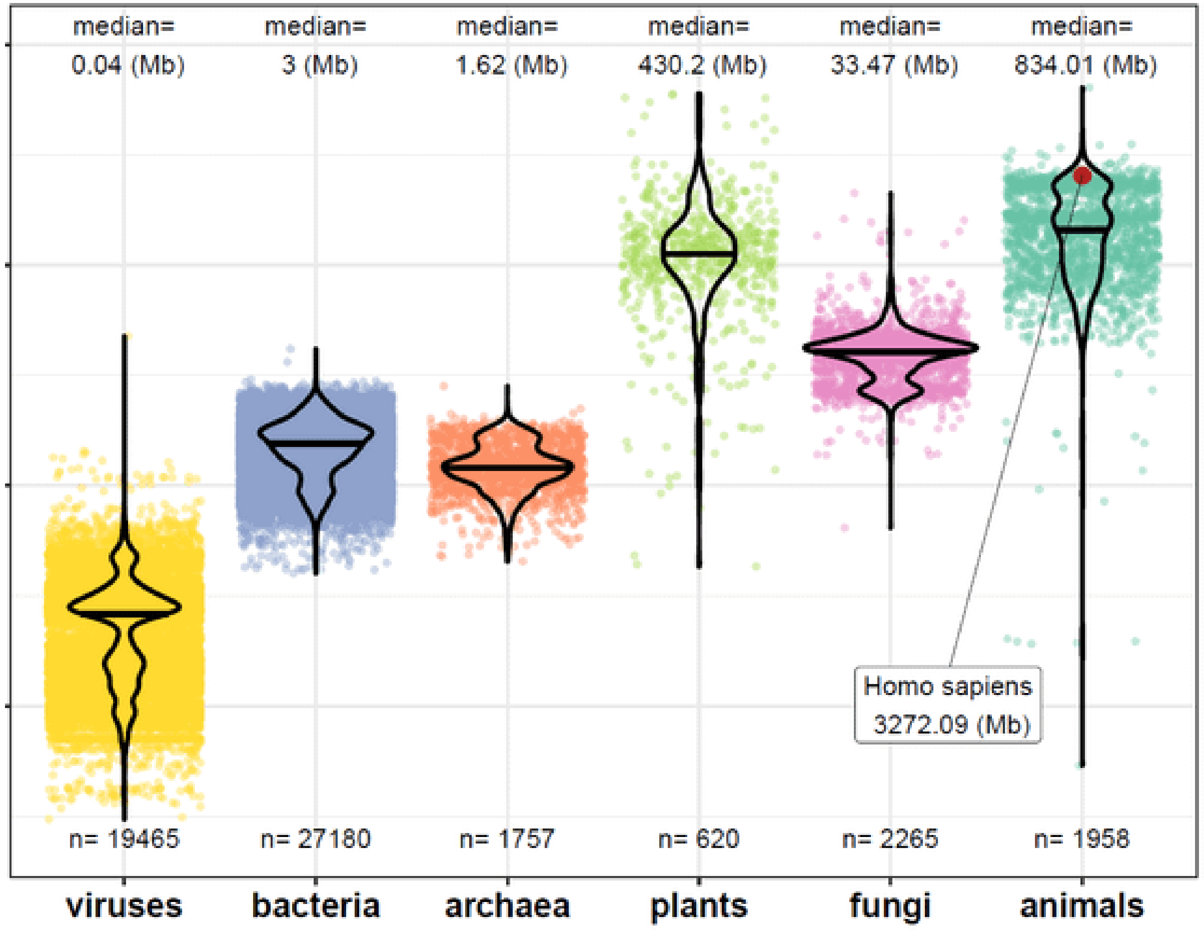
In 1977, Woese and Fox proposed the Archaea as a new domain of life and that the tree of life is divided into three branches — the Eukarya, Bacteria and Archaea

Human and animal gut - methanogens



Microbiome

Community of microorganisms



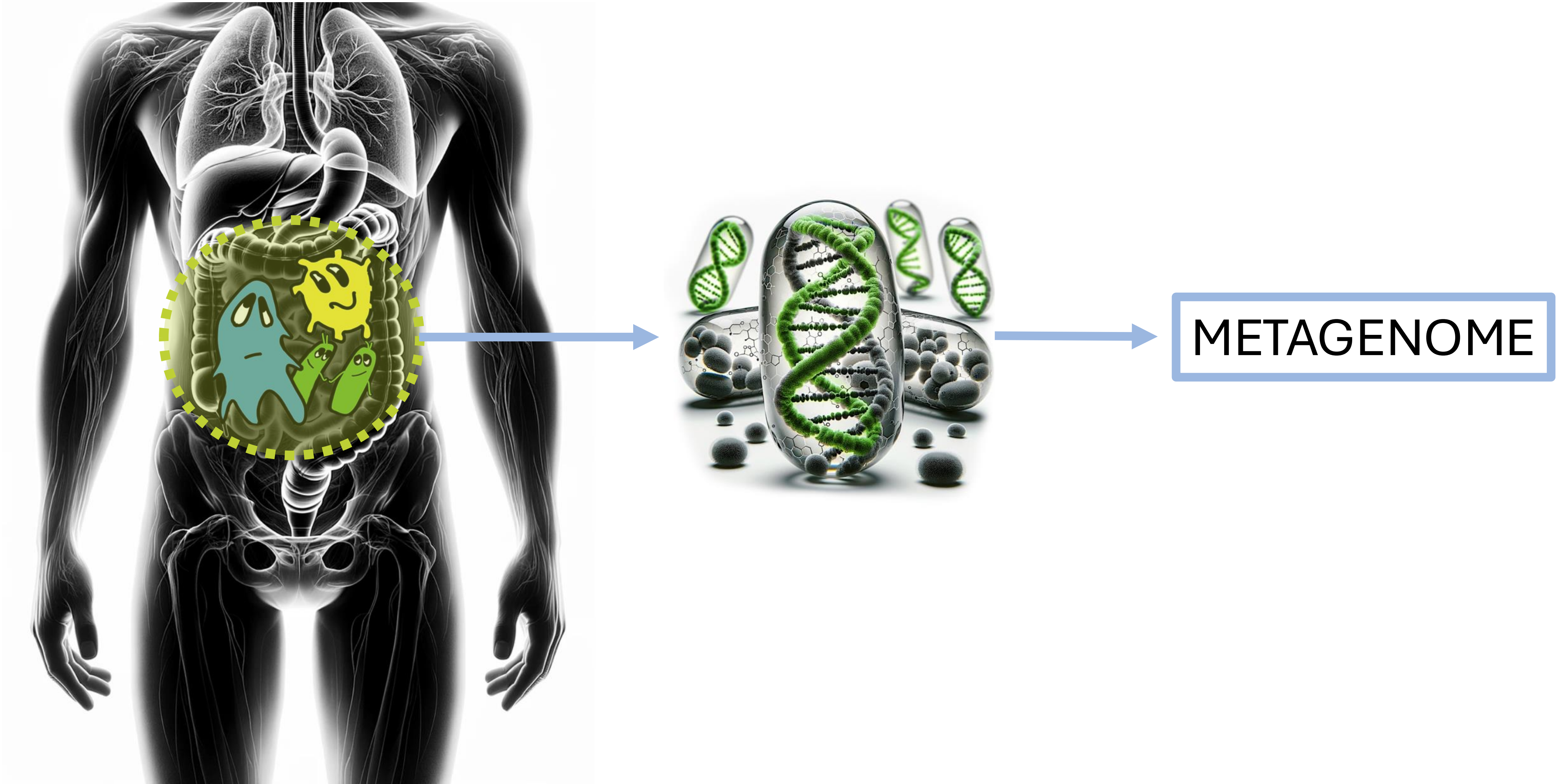
MICROBIOME

Bacteria - 99%

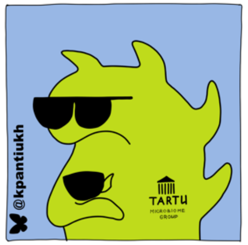
Archaea
Viruses
Fungi
Single cell eukaryotes
- 1%

METAGENOME

a mixture of DNA from all microorganisms in the community



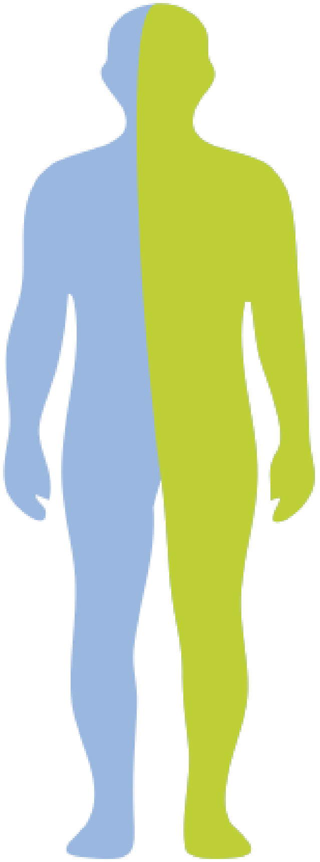
METAGENOME



Human body

30 trillion
Human cells

43%

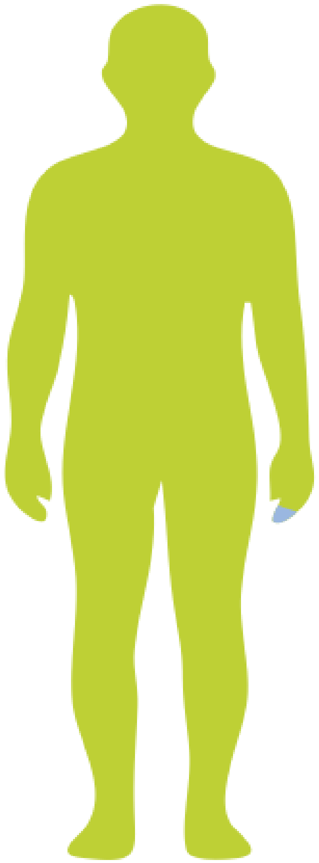


39 trillion
Bacterial cells
(mostly gut bacteria)

Human body

20 000
Human genes

1%

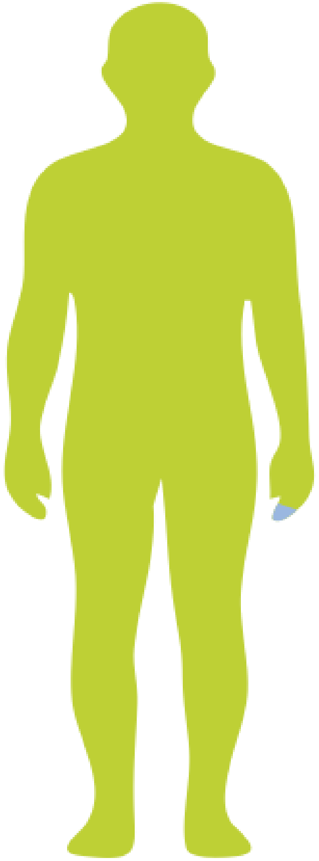


2-20 mln.
Bacterial genes

Human body

20 000
Human genes

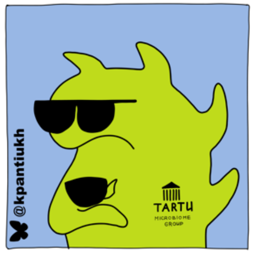
1%



2-20 mln.
Bacterial genes

Human body

We don't pay enough attention to 99% of the genes in our body



Human body

We don't pay enough attention to 99% of the genes in our body

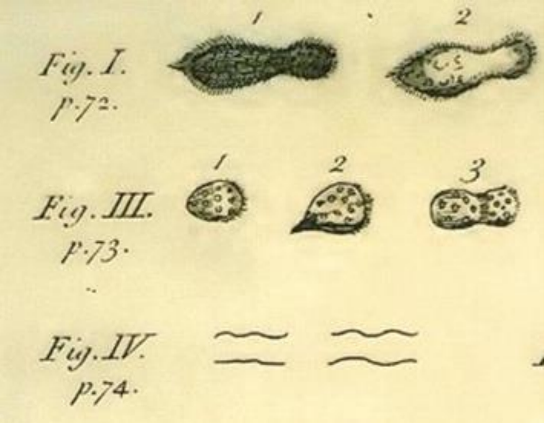
It is the 99% that, unlike our own genome, **CAN BE CHANGED**

Microbiome

history overview

Microbiology

1670s–1680s, **Antonie van Leeuwenhoek** and MO description via microscope



200 years

Pasteur, Koch and Petri

1860s–1890s

cultivation with liquid media and agar

100 years

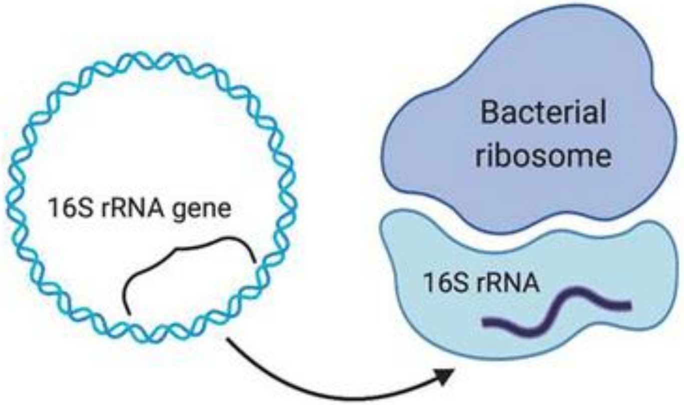
PCR & sequencing

1950s - 2000s

Community profiling

Based on the 16S rRNA gene

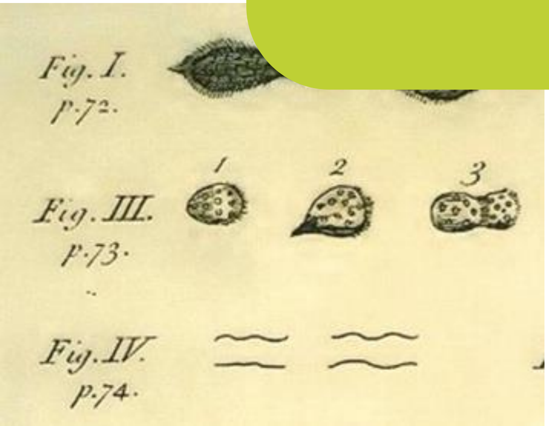
**** for known species only***



Microbiome

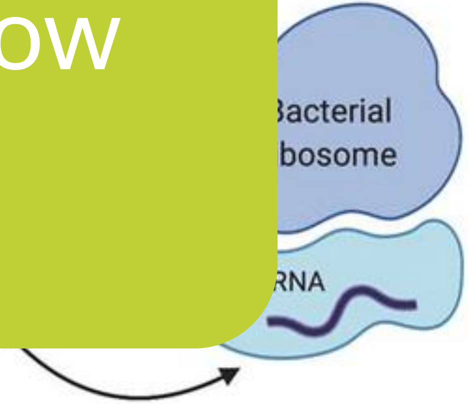
history overview

Microbi
1670s–16
Antonie
Leeuwer
and MO
via micro

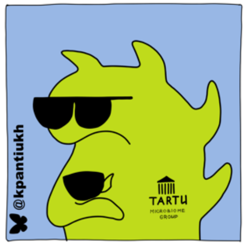


Pasteur, Koch

PCR & sequencing
1950s - 2000s
Community profiling
gene
y

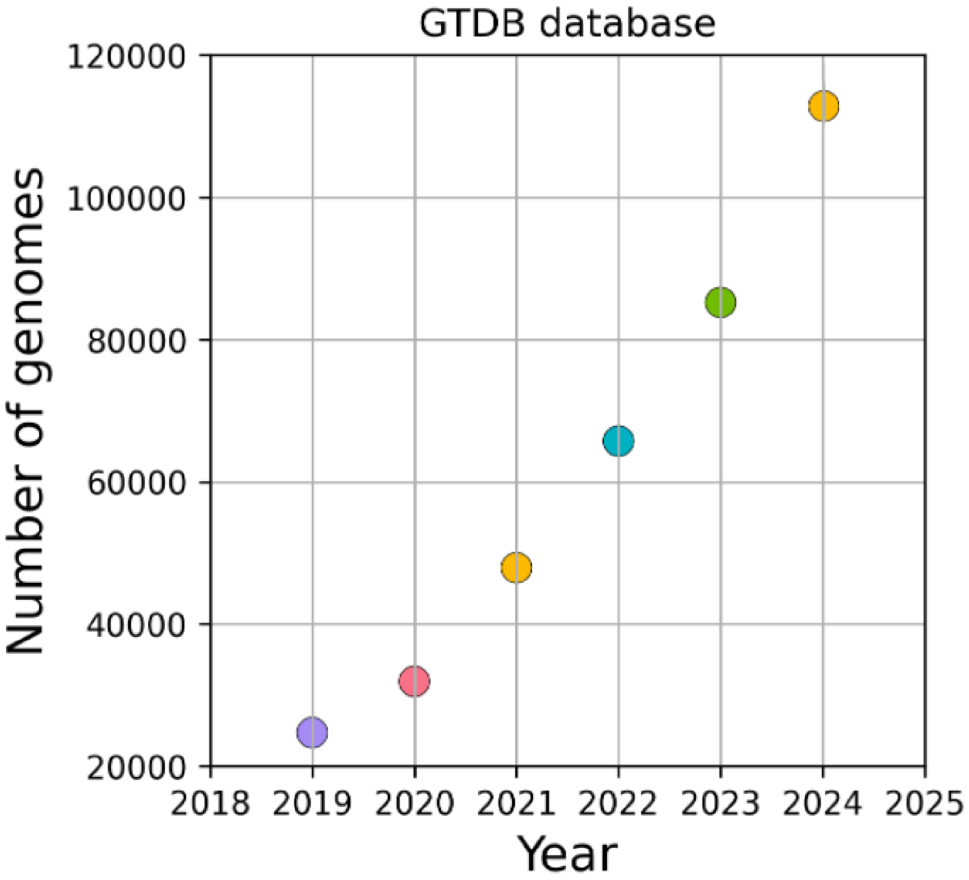


How many species did we know before the year 2000?



Microbiome

history overview



before 2000

fewer than

5,000

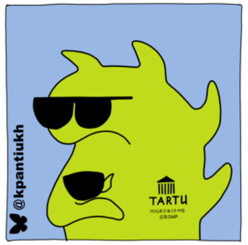
bacterial species are known

2025

143,614

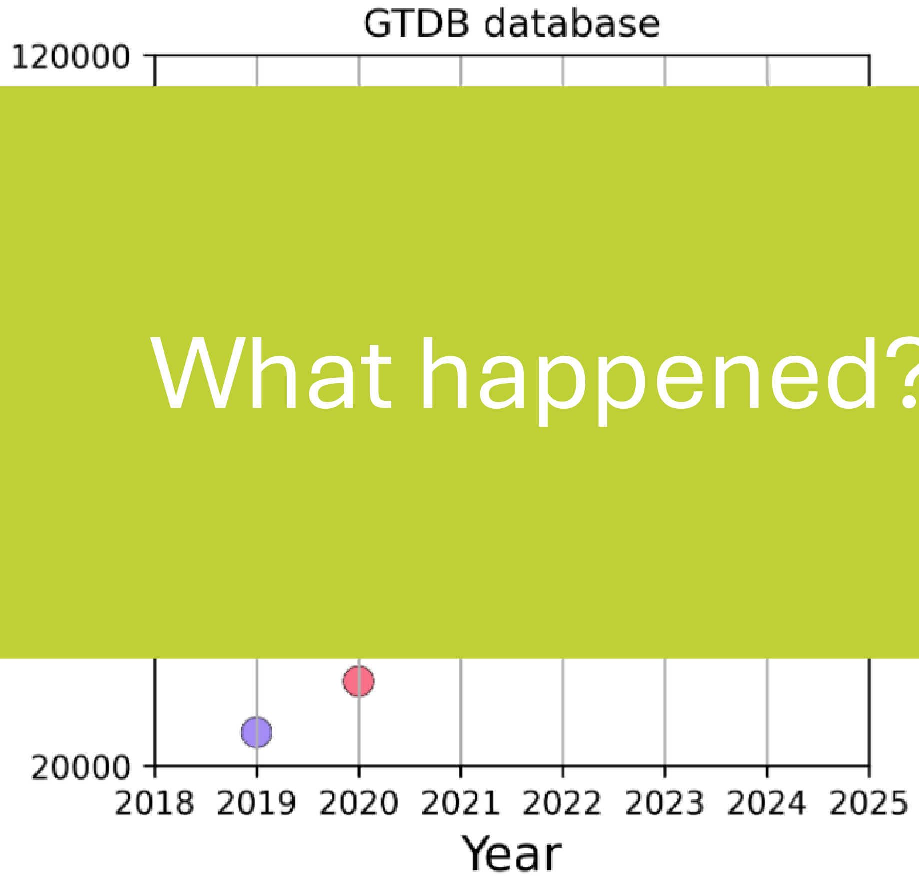
bacterial species are known

https://gtdb.ecogenomic.org



Microbiome

history overview



2025

What happened?

before 2000

fewer than

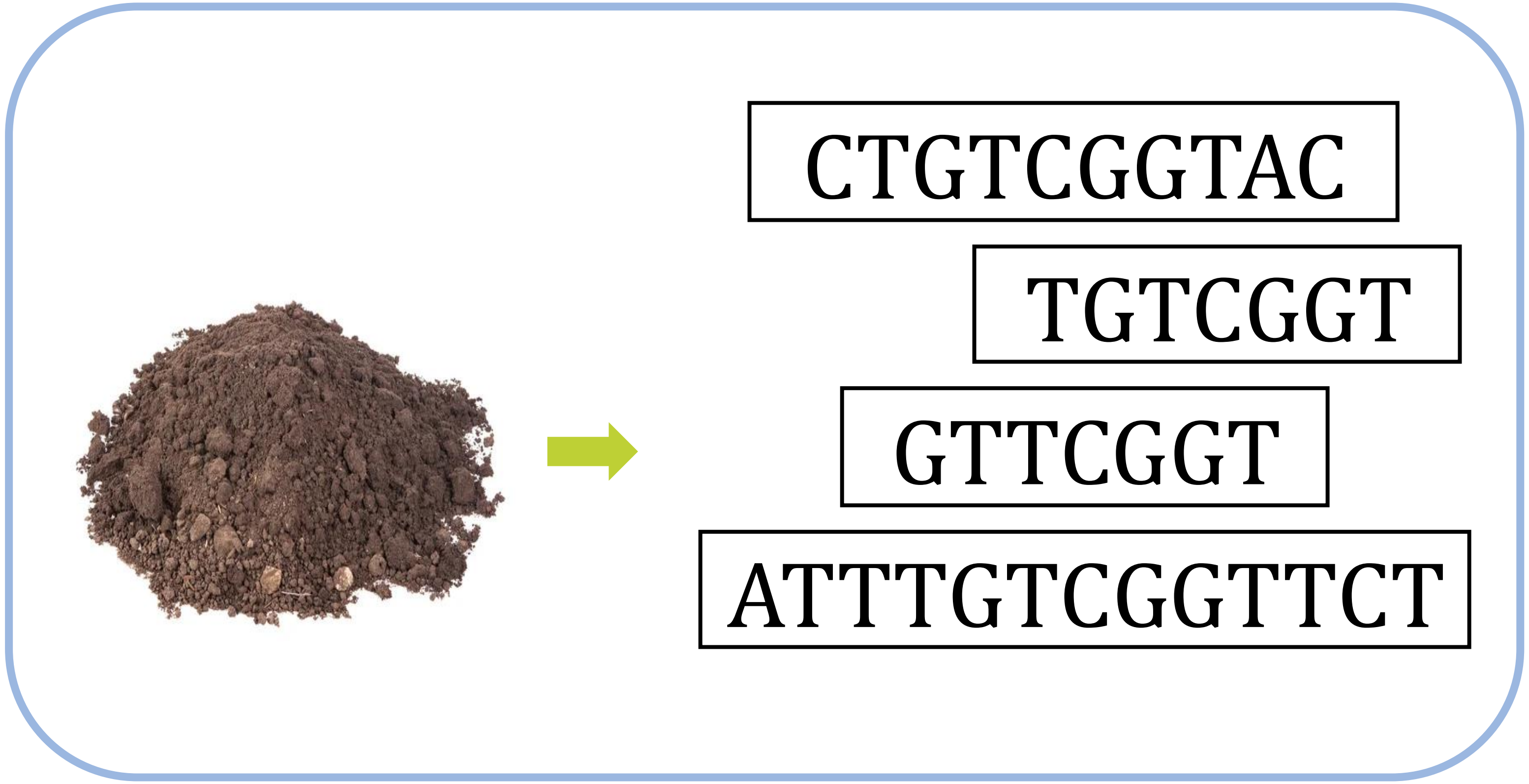
5,000

bacterial species are known

https://gtdb.ecogenomic.org

METAGENOME

a mixture of DNA from all microorganisms in the community



CTGTCGGTAC

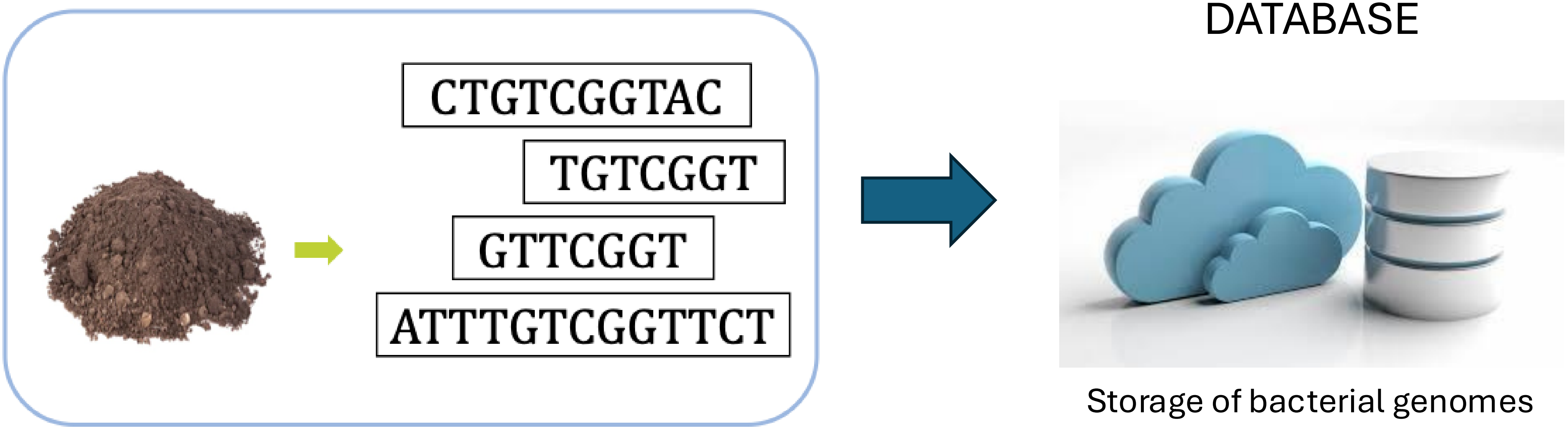
TGTCGGT

GTTCGGT

ATTTGTCGGTTCT

METAGENOME

a mixture of DNA from all microorganisms in the community



CTGTCGGTAC

TGTCGGT

GTTCGGT

ATTTGTCGGTTCT

DATABASE

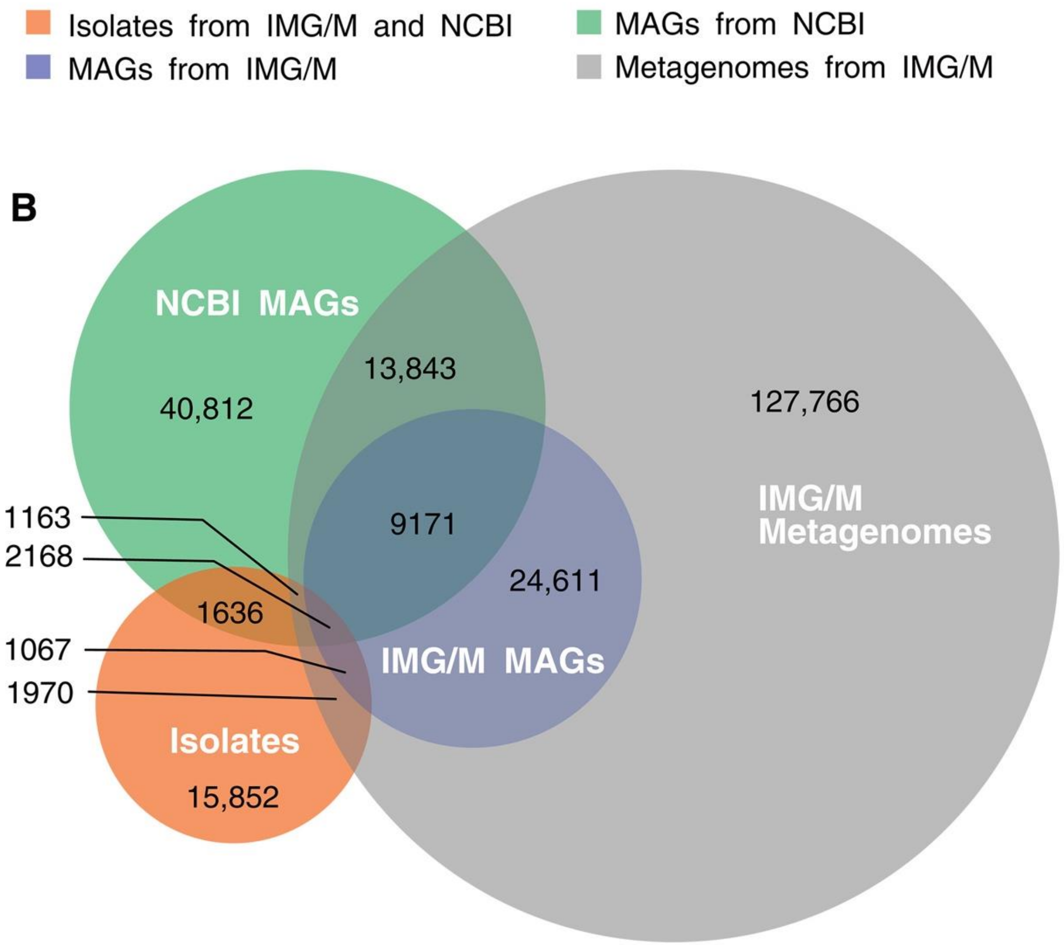
Storage of bacterial genomes

Bacterial genome sources

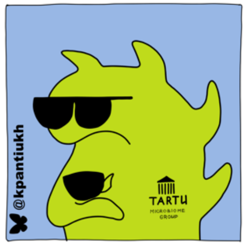
Bacterial isolate genomes represent **9.73%** of the total estimated diversity

Efforts to recover MAGs expanded the known diversity to **49%**

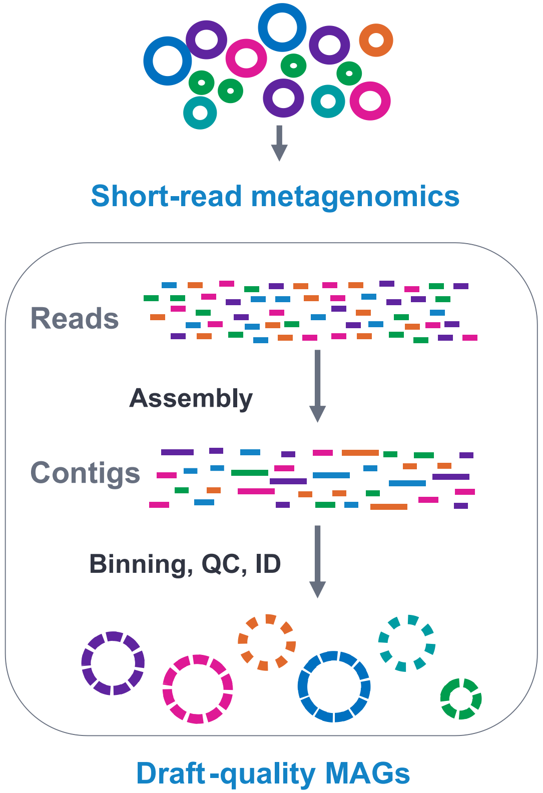
42% of bacterial diversity lacks genomic representation in public databases



https://www.science.org/doi/10.1126/sciadv.adq2166



Bacterial genome sources



SPIRE

EMBL-EBI | MGnify

MGnify

1. New species

2. Huge numbers of **genomes per species**

METAGENOME

a mixture of DNA from all microorganisms in the community

DATABASE

Storage of bacterial genomes

→

Community profiling

Based on the 16S rRNA gene sequencing

Based on short metagenomics

METAGENOME

a mixture of DNA from all microorganisms in the community

DATABASE

Storage of bacterial genomes

→

Community profiling

Based on the 16S rRNA gene sequencing

Based on short metagenomics

↓

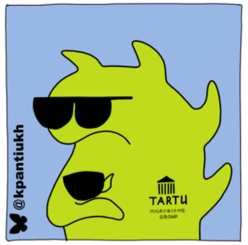
ABUNDANCE TABLE

Abundance table

community profiling

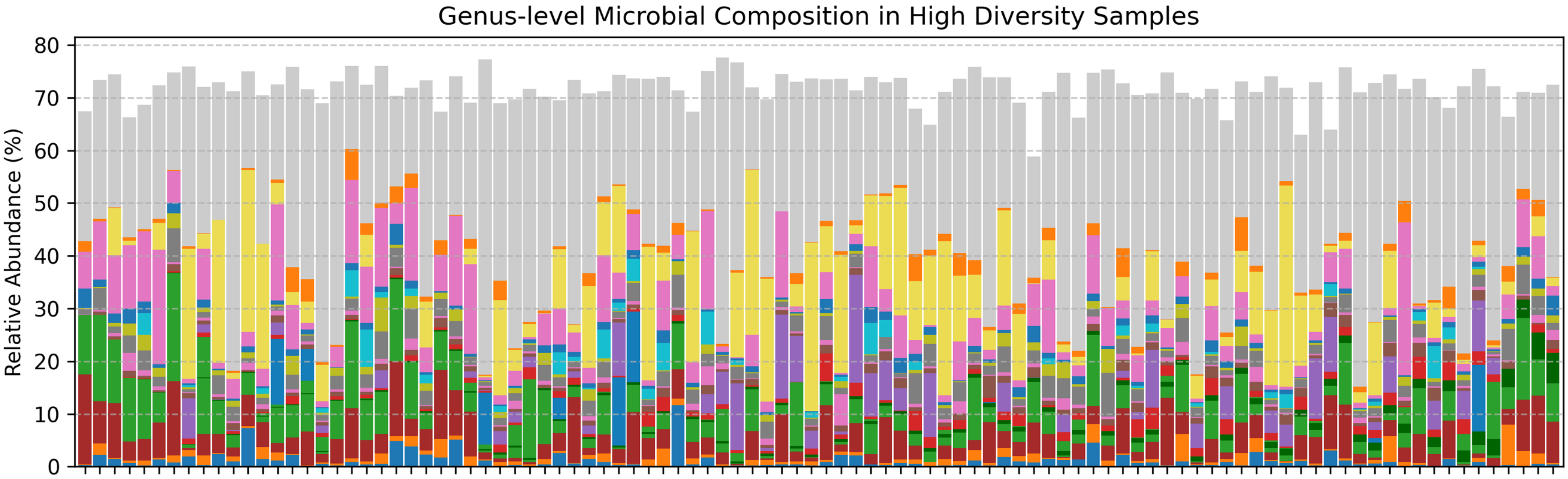
```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACG
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATT
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAG
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATC
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTT
```

	Tom	Mary
Species 1	0.1	3.7
Species 2	0.1	0.2
Species 3	2.3	0.0



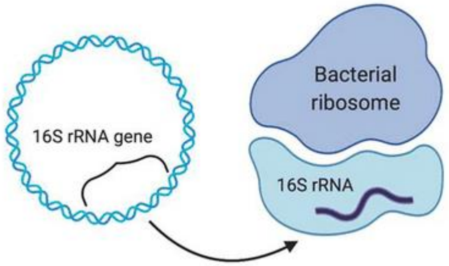
Abundance table

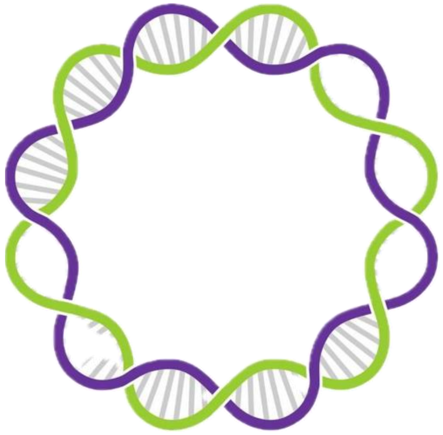
community profiling



Abundance table

and very different ways to get it





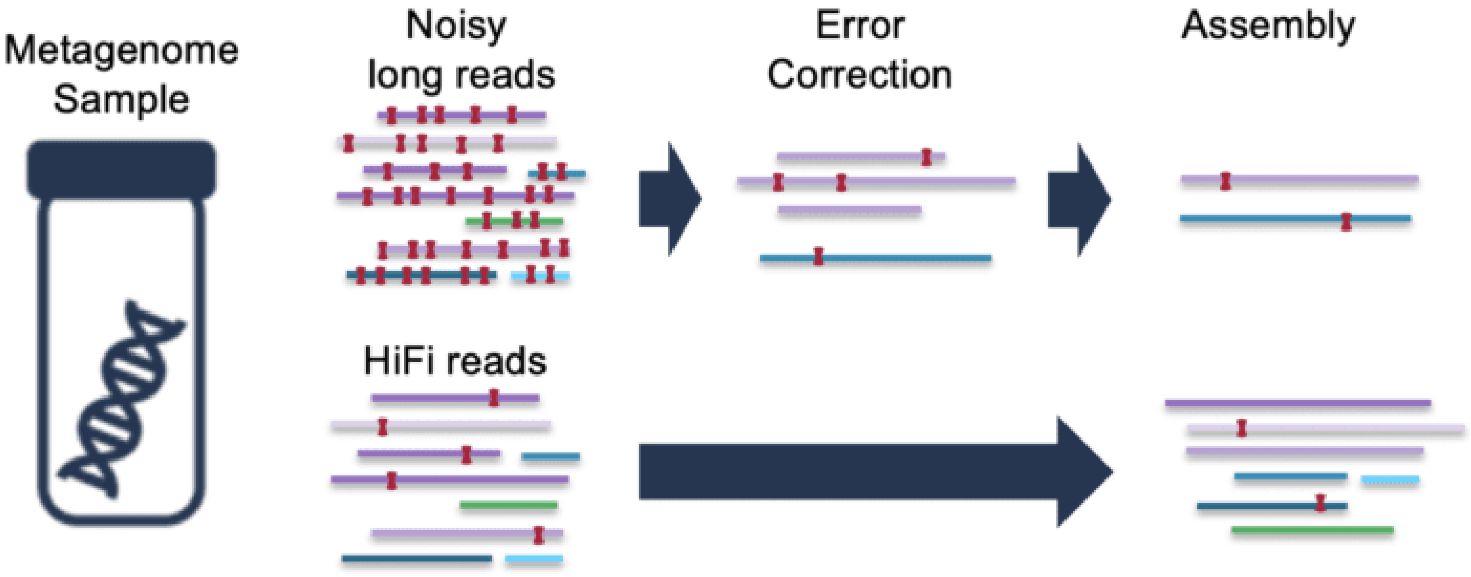
16S rRNA gene sequencing

- cheap, simple

Based on metagenomics
- short reads/long reads/hi-C reads

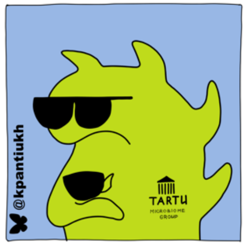
Long reads

assembly and profiling



- Species averaging
- Missing low abundance species
- Low contig purity

- Resolved strains
- Improved assembly completeness
- High contig purity



Metagenomic data

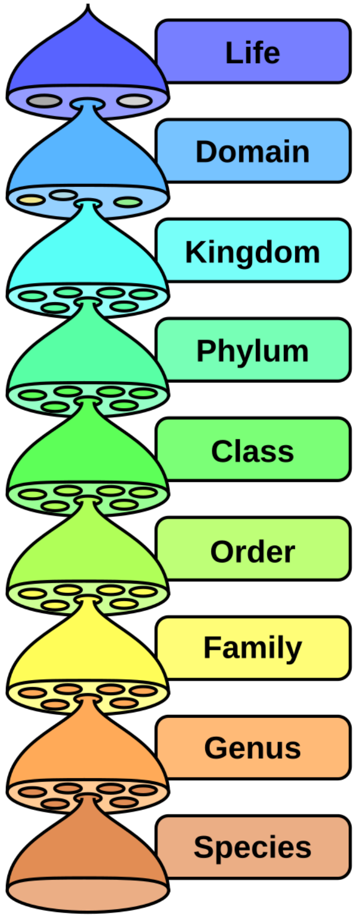
what data do you need?

Abundance Table + ???

Taxonomic data

description of species from abundance data

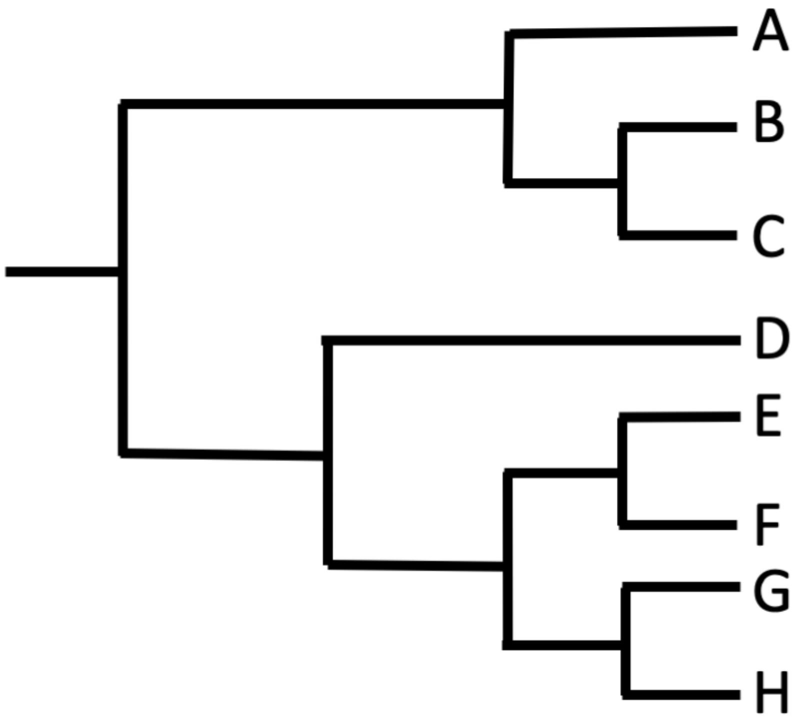
What level to select?

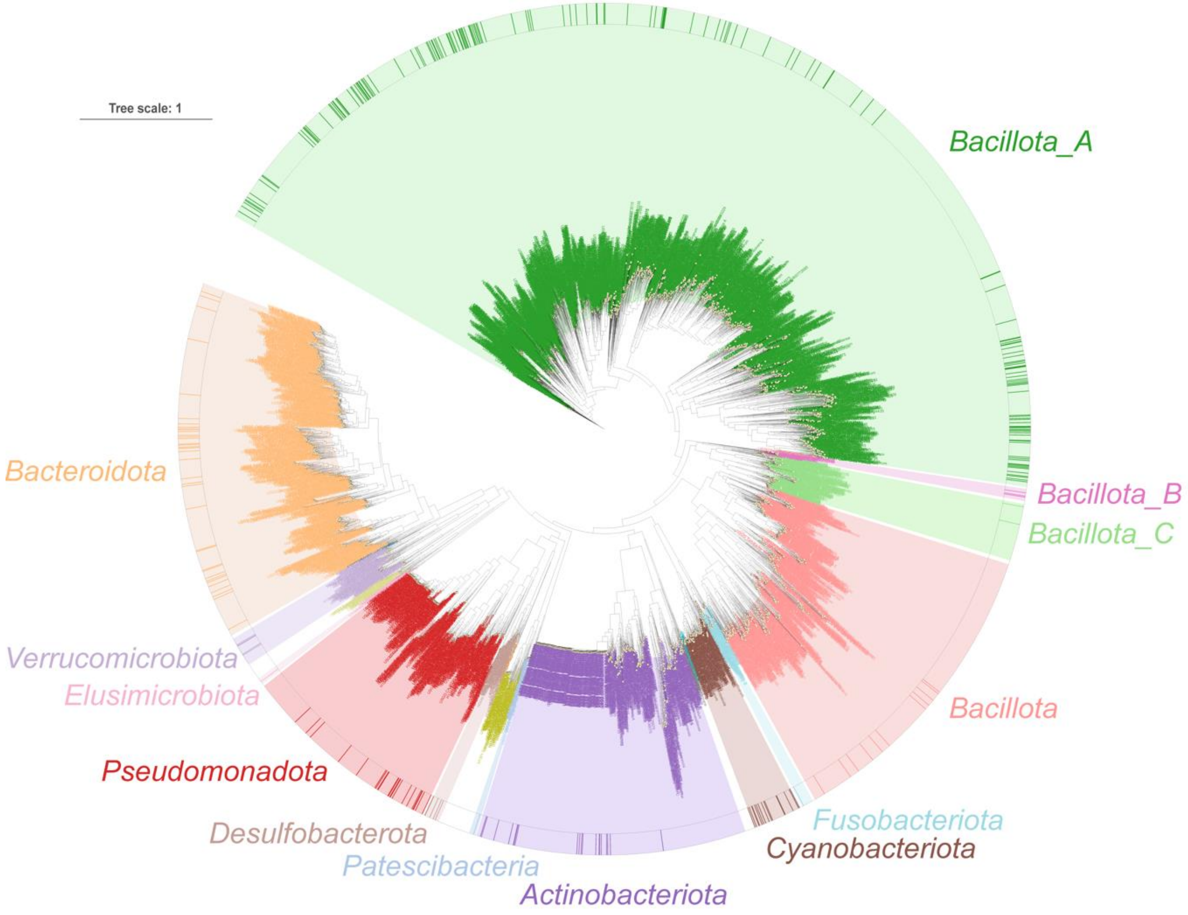


	Tom	Mary
Species 1	0.1	3.7
Species 2	0.1	0.2
Species 3	2.3	0.0

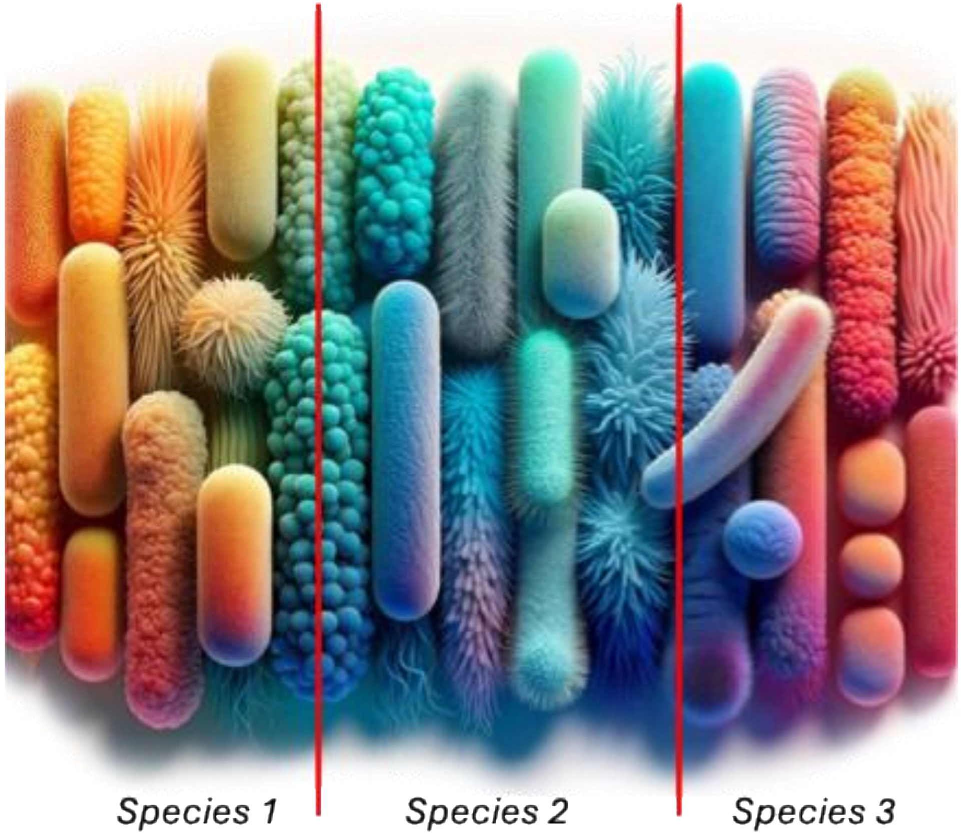
Taxonomic data

Trees

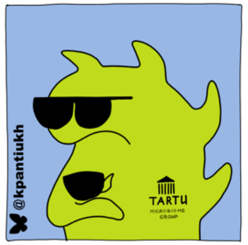




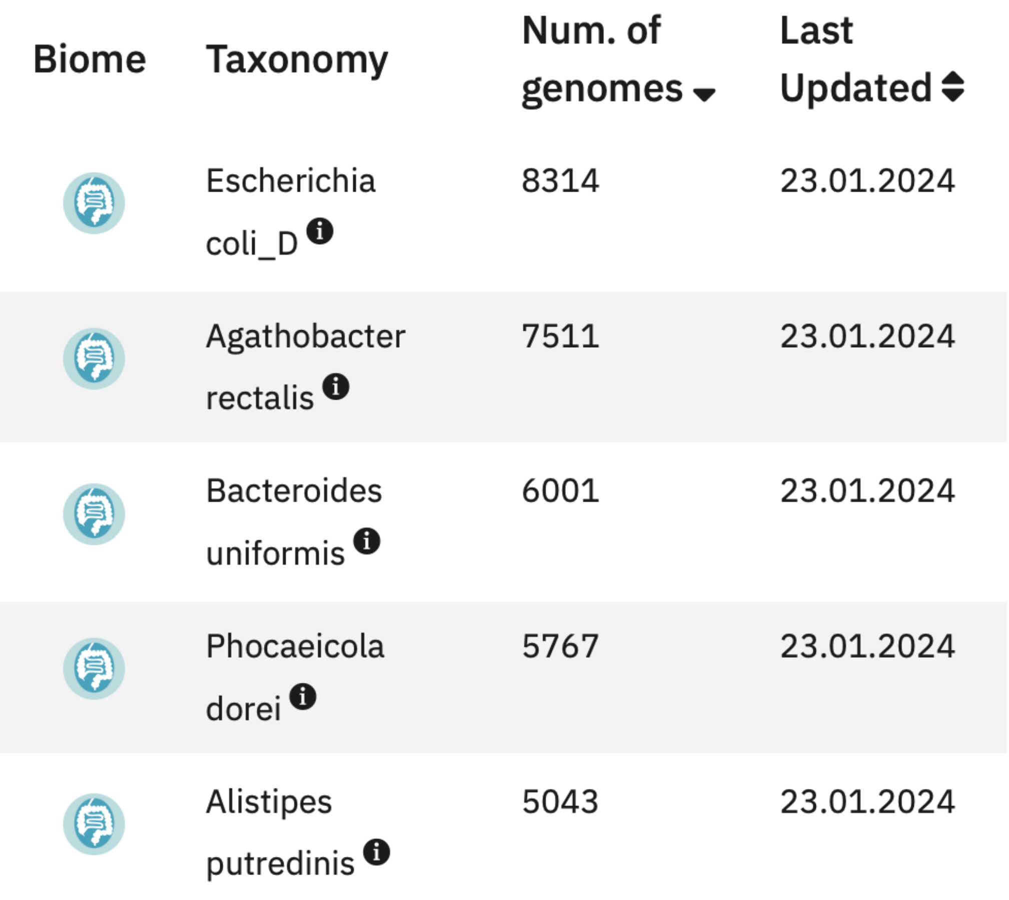
The Problem. Where is a species border?



Bacteria with
genome similarity > 95%
are different species



The Problem. Where is a species border?

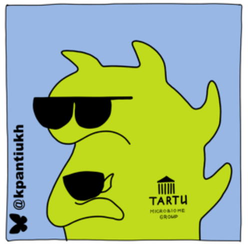


Biome	Taxonomy	Num. of genomes ▾	Last Updated ⬍
	Escherichia coli_D ⓘ	8314	23.01.2024
	Agathobacter rectalis ⓘ	7511	23.01.2024
	Bacteroides uniformis ⓘ	6001	23.01.2024
	Phocaeicola dorei ⓘ	5767	23.01.2024
	Alistipes putredinis ⓘ	5043	23.01.2024

Bacteria with
genome similarity > 95%
are different species

Sub-species level

Strain?



Metagenomic data

what data do you need?

Abundance Table + Taxonomic data + ???

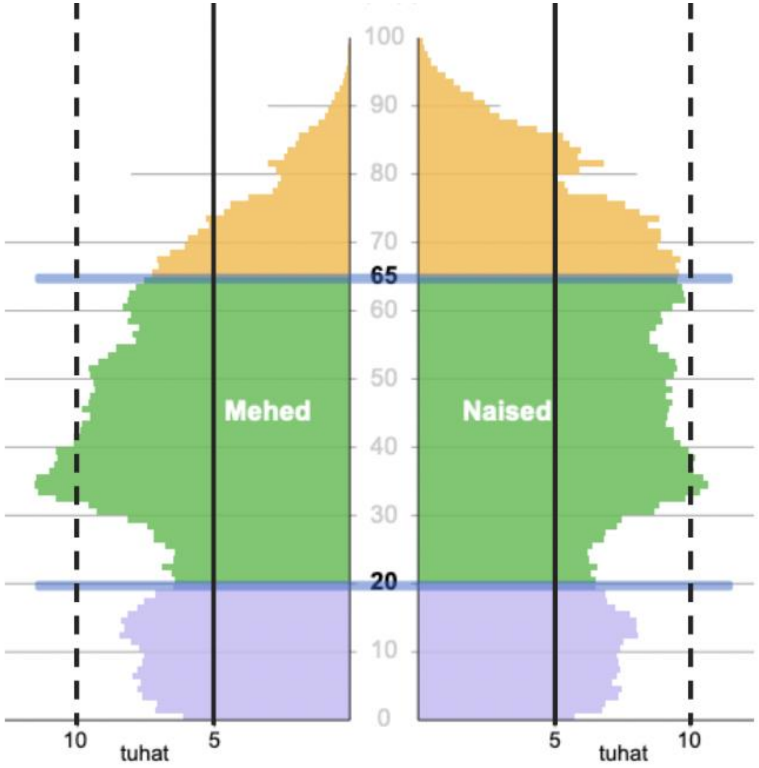
Meta data

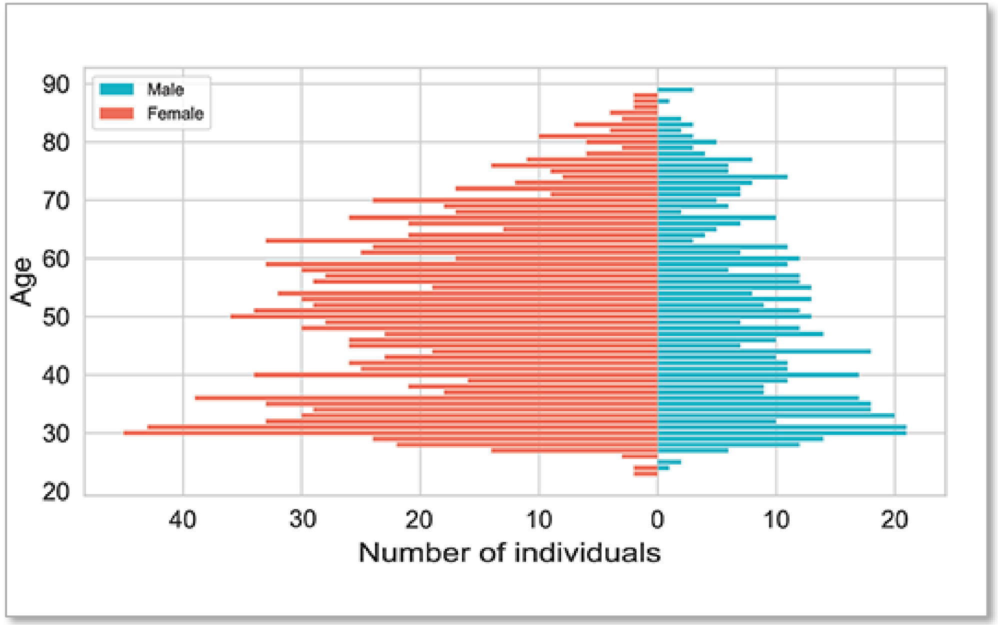
description of samples from abundance data

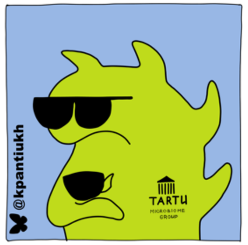
	Tom	Mary
Gender	1	0
Age	24	36
BMI	19	23

Meta data

Population piramids

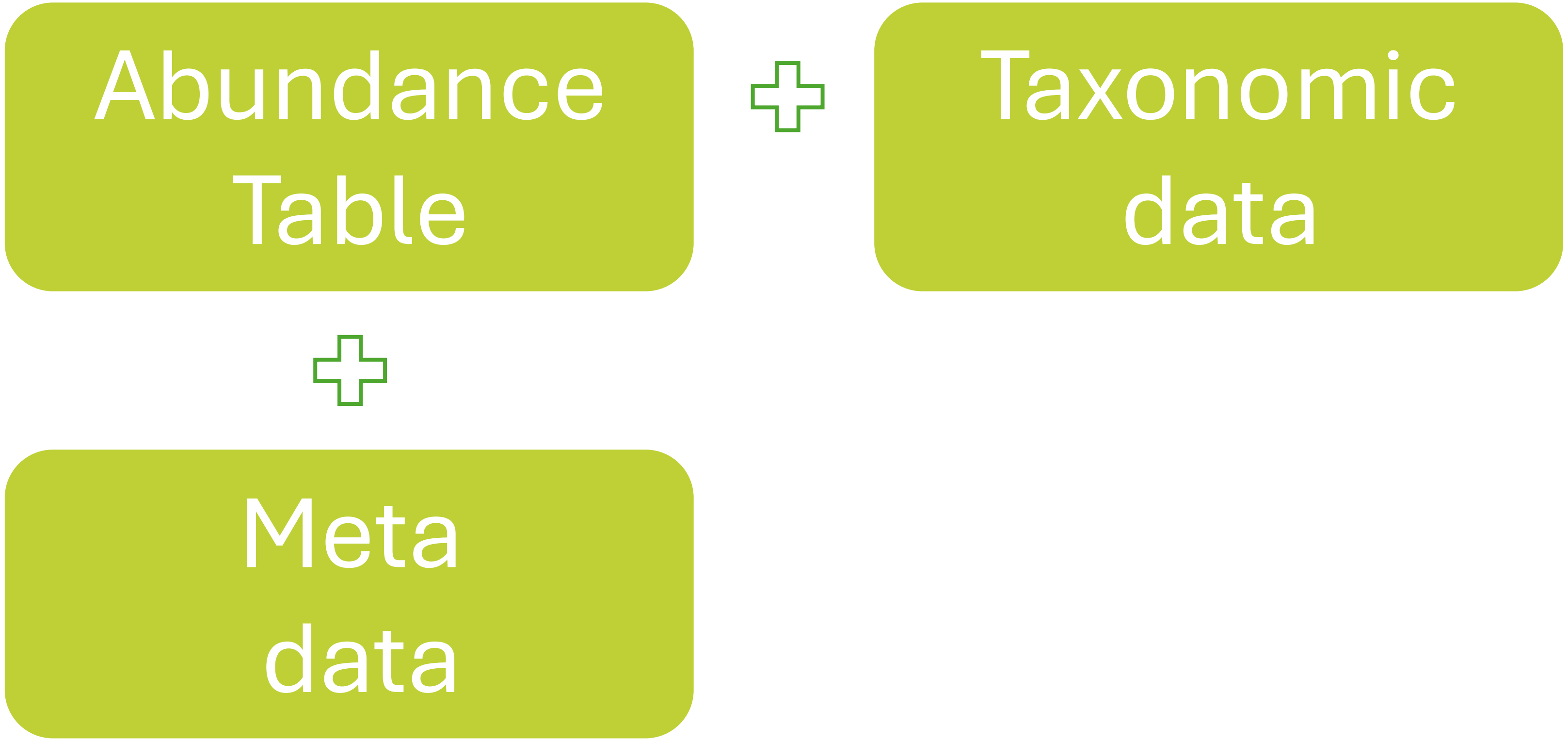






Metagenomic data

what data do you need?



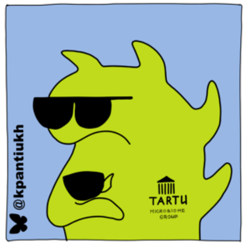
Abundance Table

+

Taxonomic data

+

Meta data



Metagenomic data

what data do you need?

DISCUSSION:
What question you can answer based on data provided?

Abundance Table

Taxonomic data

Meta data