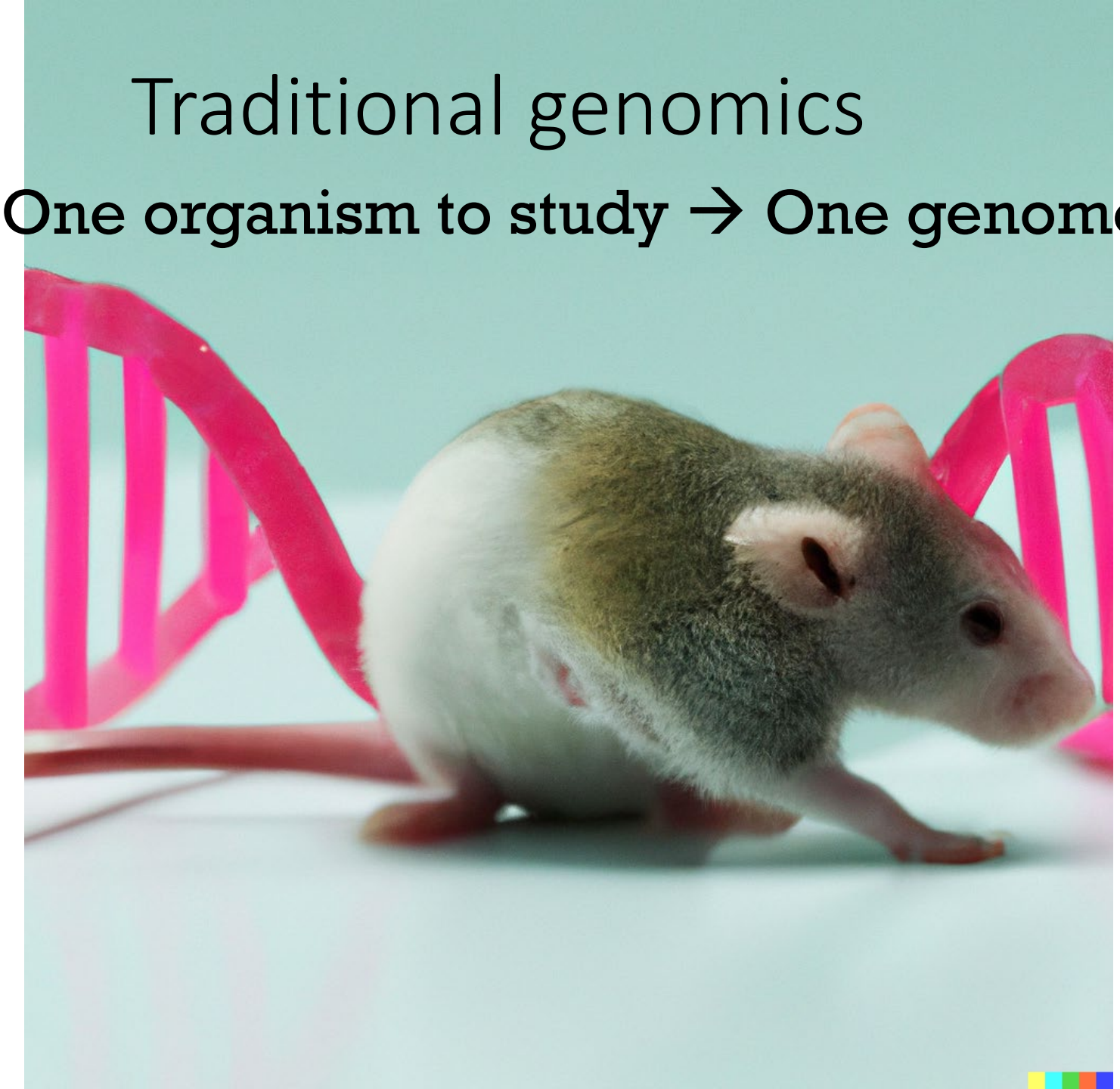
Introduction to Metagenomics

Renaud Van Damme, Department of Animal Biosciences, SLU

What’s Metagenomics ?

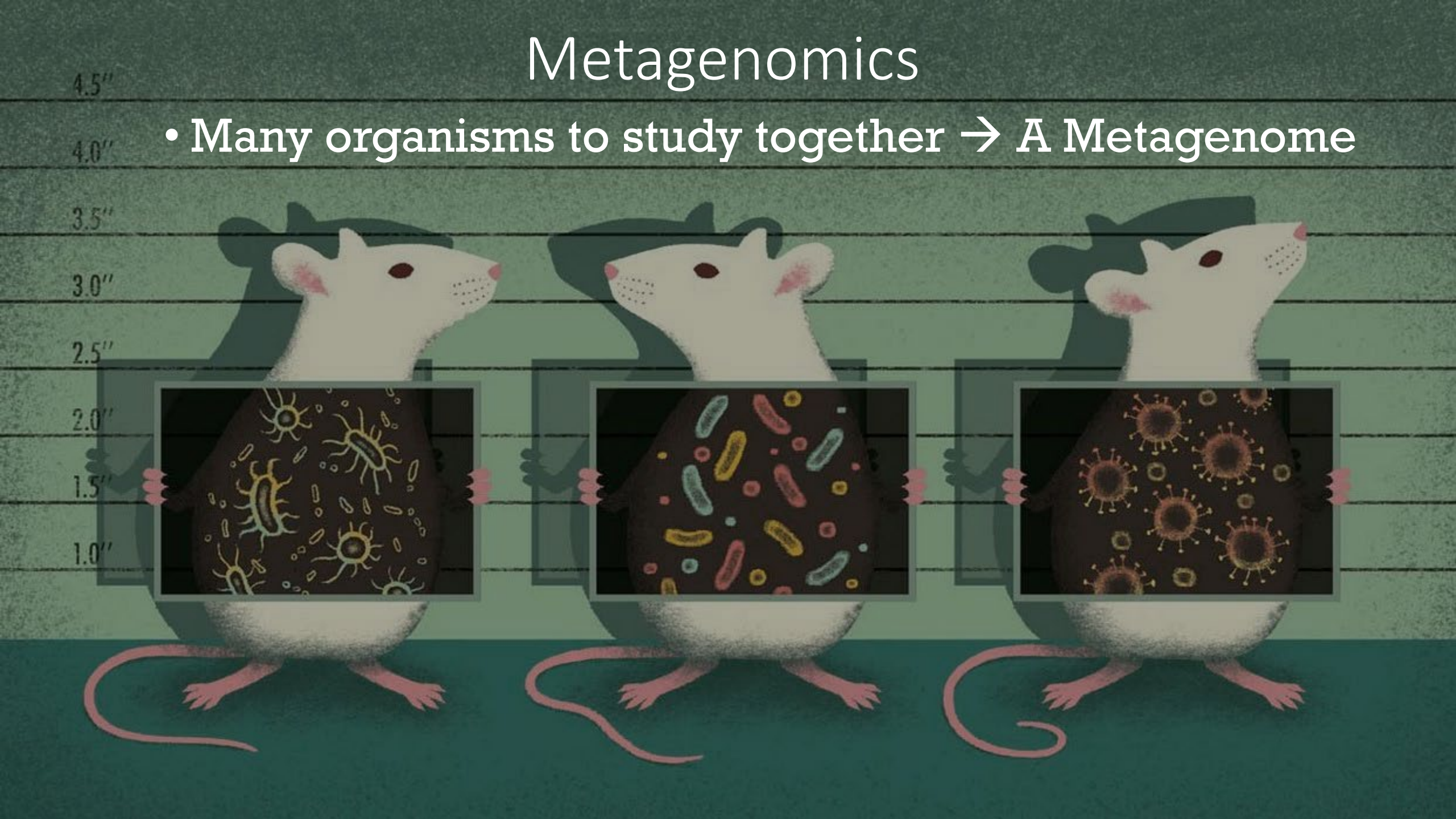
Traditional genomics

- **One organism to study → One genome**



Metagenomics

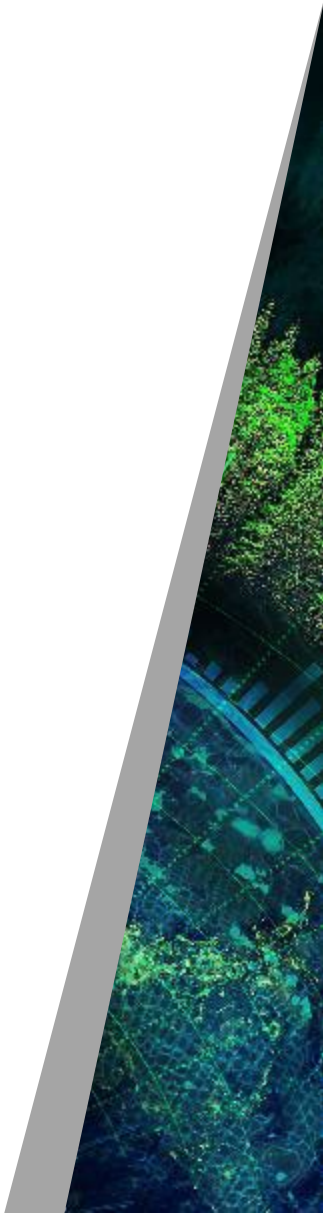
- **Many organisms to study together → A Metagenome**



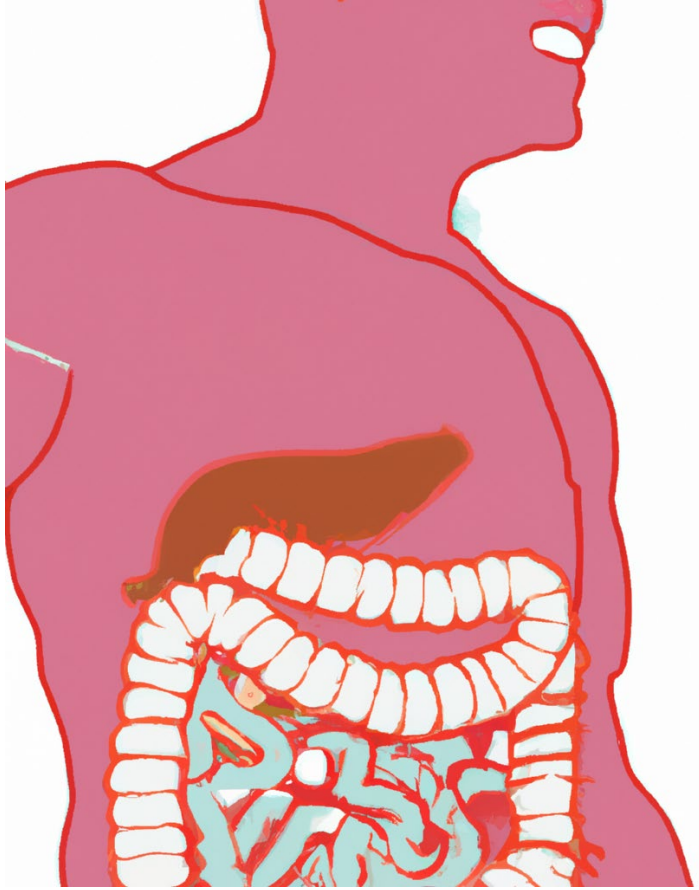
Why make our life complicated ?

Why make our life complicated ?

- It is hard/impossible to separate the organisms:
- Isolating (e.g. making a pure culture) microbes is hard
- We want to study interaction
- We want to study organisms in situ
- It's mainly about the microbes



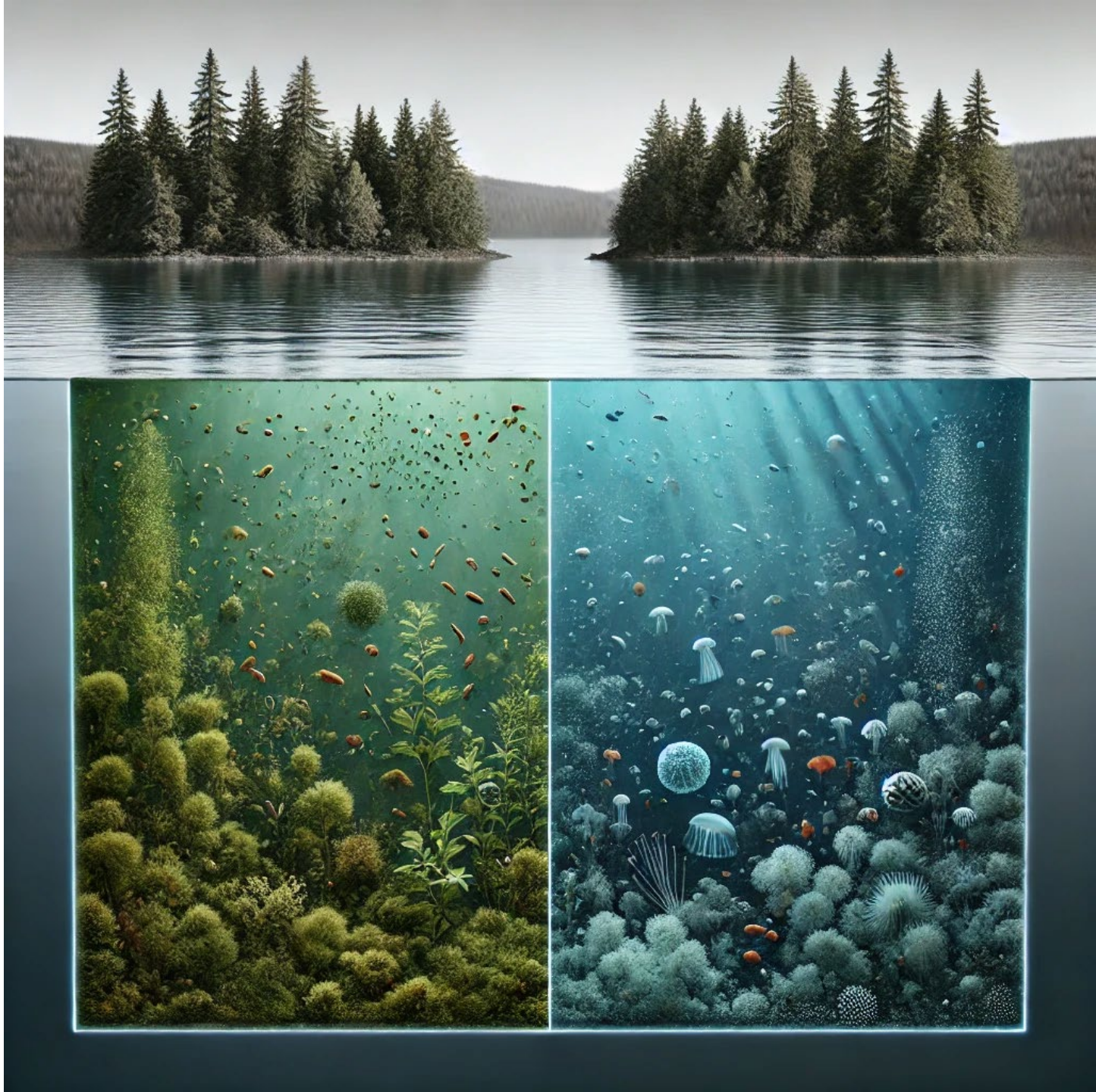
Human Gut example



- 10*10^13 cells (1.3 x human cells)
- About 0.5% of dry weight
- ~400 species in every gut
- But only ~20% cultivated
- 90% of disease in humans can be related to microbiomes
- Most nutrients that end up in your blood stream have been somewhat metabolized by some microbe

A lake

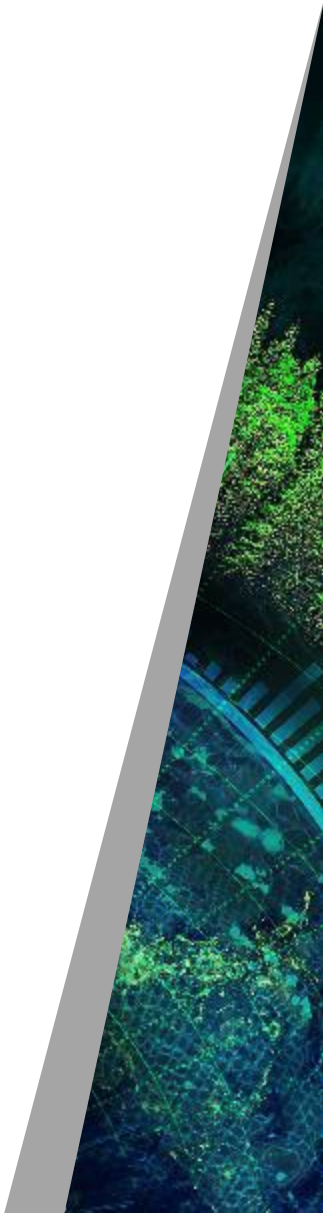
- 1.35*10^17 cells (a milion per mL)
  - About 135kg of dry weight
  - ~700 species in the water
  - 3.000.000 encoded genes
  - Only a couple of handfull are cultivated.



Why does it matter?

They are everywhere
They do a lot of stuff:

- 90% of disease in humans can be related to microbiomes
- Most nutrients that end up in your blood stream have been
somewhat metabolized by some microbe
- Microbes do 2/3rds of carbon fixation in aquatic environments
- But also the majority of carbon emissions!



What information can I find?

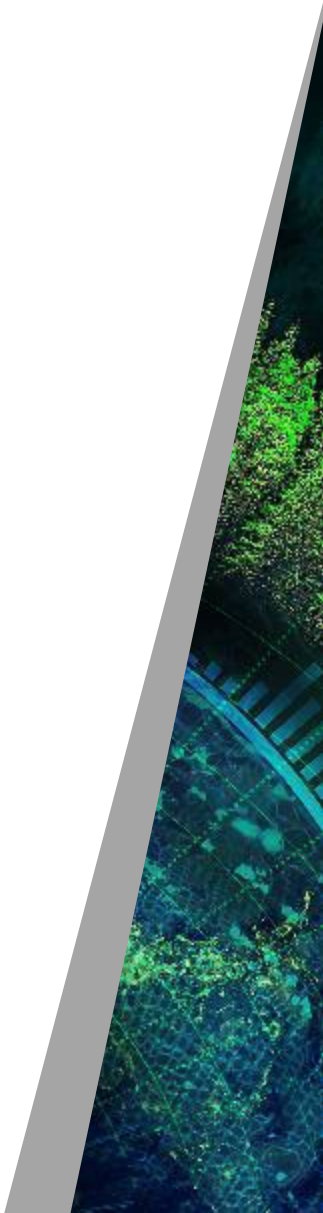
Different Sequencing Strategies give different information

- **Who is in the environment? (taxonomy)**
- **How many of each? (abundance)**
- **What can they do? (genetic potential)**
- **What are they doing? (expression analysis)**
- **What do they eat? (metabolism)**
- **Where do they come from? (evolution and ecology)**

The strategies

The strategies

- **Targeted, a.k.a. amplicon/metabarcoding/edna**
- **Shotgun sequencing:**
  - Short reads (HiSeq, NovaSeq, PacBio Onso)
  - Long reads (PacBio, Nanopore)
- **Metatranscriptomic**
- **Functional metagenomics**



Two Main Strategies

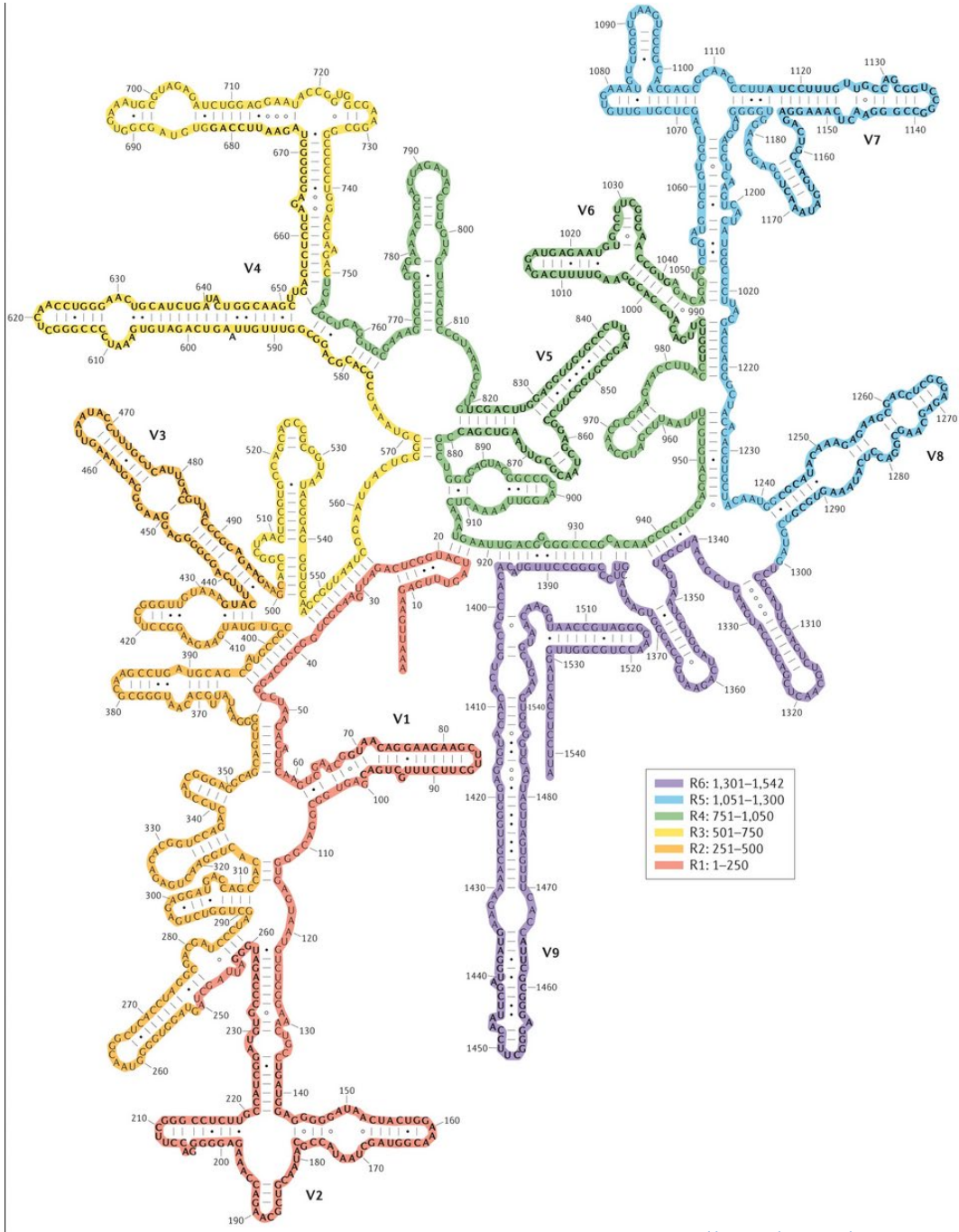
Shotgun	Targeted
Explorative approach	Pick 1 gene to sequence
Sequence everything then rebuild the puzzles	Snapshot of a function or taxonomy
Many different methodologies	Much cheaper and more depth
2 Sub approaches (assembly based and assembly-free)	Data base relient

Metabarcoding a.k.a Targeted sequencing

Targeted: 16s rRNA

The 16S rRNA gene contains conserved and variable regions, allowing for the identification of bacteria and archaea at different taxonomic levels.

- Small genomic region → Cheaper sequencing while maintaining good
- Extensive reference databases → good for finding what we know exist, bad for novel discovery
- Can be easily amplified with universal primers, making it accessible for most sequencing platforms.



https://doi.org/10.1038/nrmicro3330
Nature Reviews | Microbiology

Targeted: 16s rRNA

Short read sequencing

- 150-300bp (usually v3-v4 regions)
- High base accuracy
- High number of reads leading to High depth and taxonomic resolution (up to species level)
- Challenging to differentiate closely related species (95% identity same genus, 97% same species and 99% strain)

Long read sequencing

- Entire gene (~1.5 kb)
- Lower accuracy or equal to Illumina (PacBio)
- Less reads → less depth for complex communities
- Full length gene allow for better classification of related species and strain-level differences (can even reach divergent copies within genome*)
- Tools still maturing

DB: https://www.arb-silva.de/

* https://doi.org/10.1038/s41467-019-13036-1

Other genes used for metabarcoding

1. ITS (Internal Transcribed Spacer)

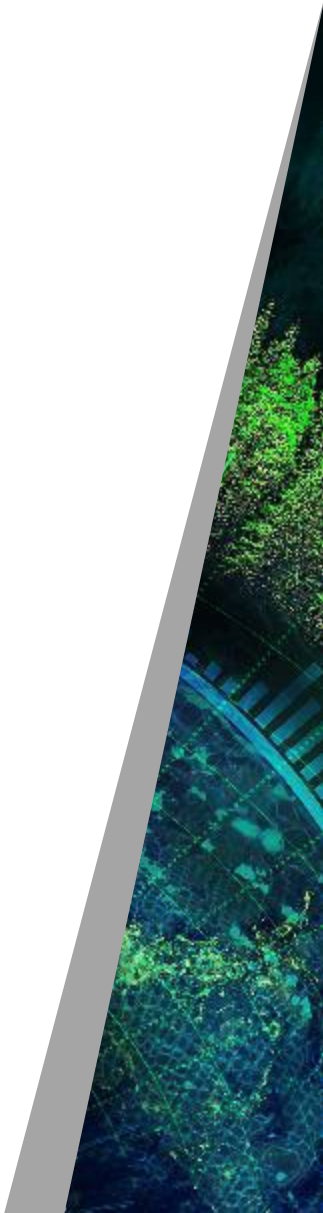
- **Fungi**
- **Applications**: Widely used in fungal diversity studies, plant-fungal interactions, and environmental microbiology.
- **DB:** https://unite.ut.ee/repository.php

2. 18S rRNA

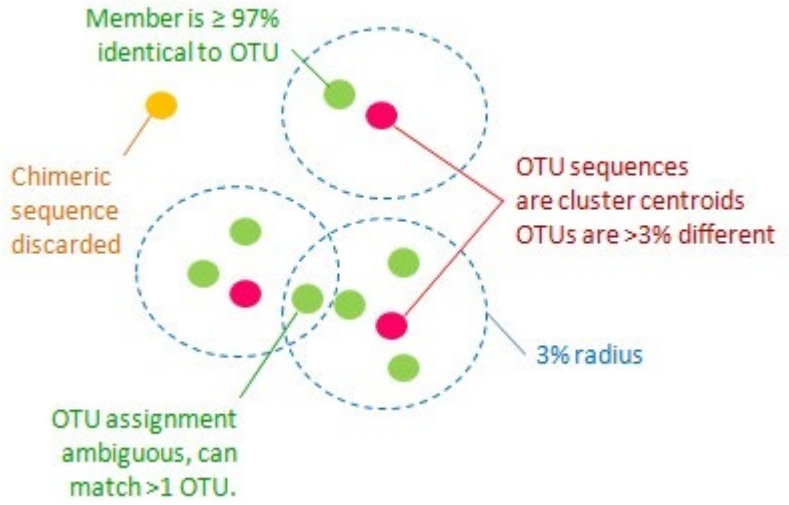
- **Eukaryotes**
- **Applications**: Commonly used in studies of eukaryotic microbial communities, environmental monitoring, and marine biodiversity.
- **DB:** https://www.arb-silva.de/

3. COI (Cytochrome c Oxidase I)

- **Animals**
- **Applications**: Used in biodiversity assessments, conservation studies, and species identification.
- **DB:** https://www.ibol.org/phase1/bold/



Clustering sequence: OTU



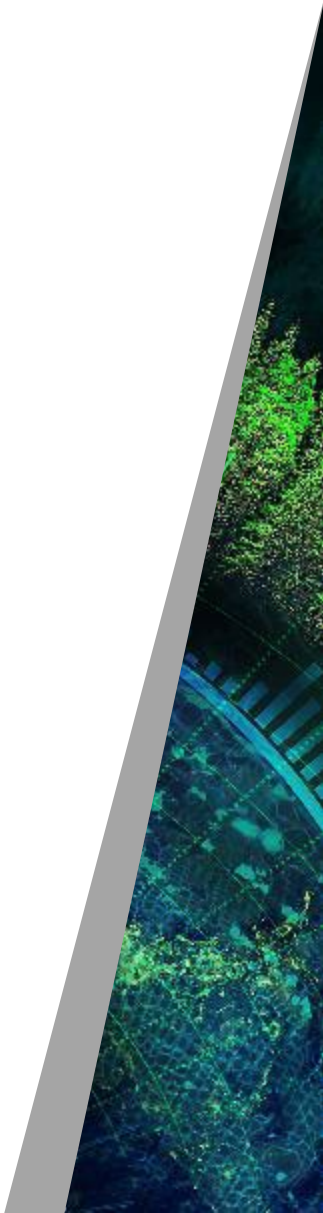
https://drive5.com/usearch/manual/uparseotu_algo.html

Take into account:
- intraspecific variability
- sequencing errors
- 97% similarity: species threshold
- two types: de novo and reference-based

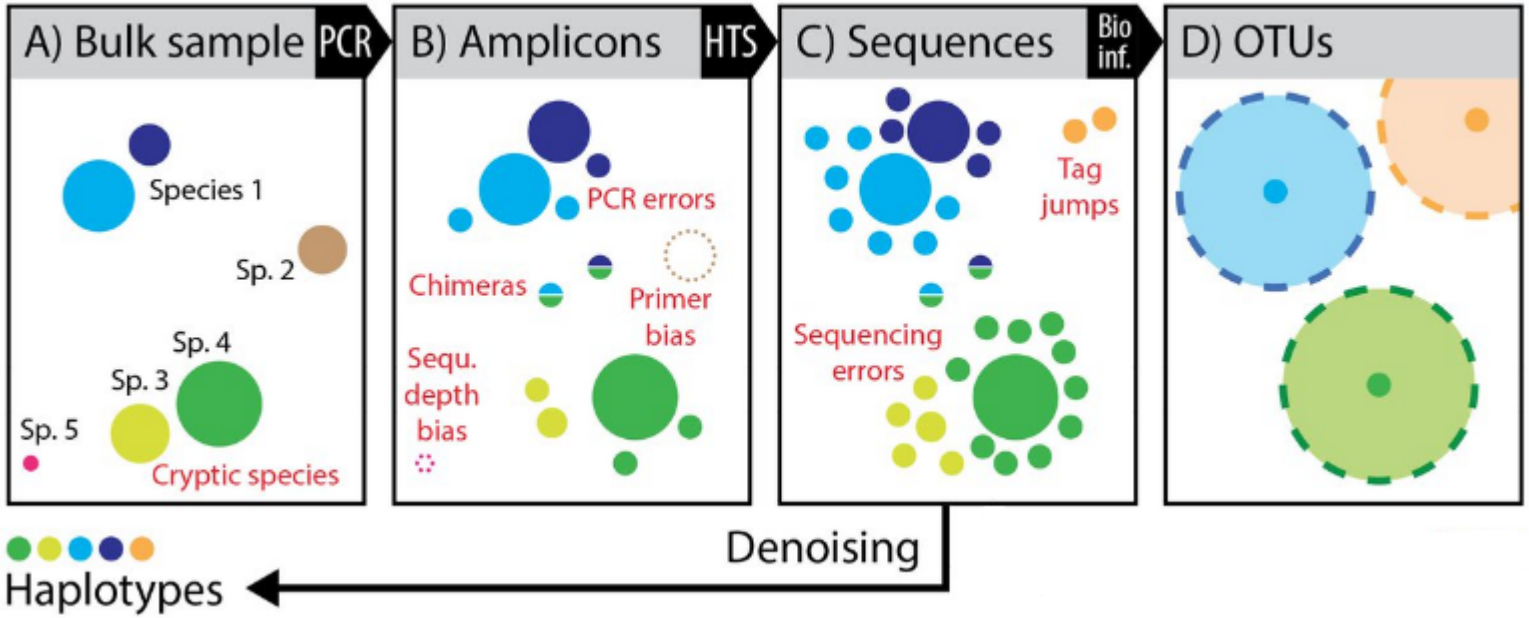
Clustering sequence: ASV

- Amplicon sequence variants (today's tutorial)
  The "true" biological sequences are inferred from reads
  "Biological sequences are discriminated
  from errors on the basis of, in part, the
  expectation that biological sequences are
  more likely to be repeatedly observed than
  are error-containing sequences".

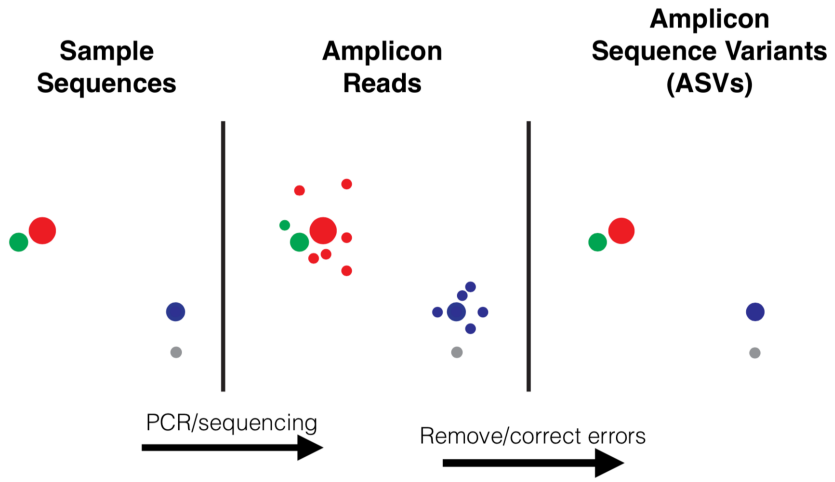
- Further reading
  https://www.nature.com/articles/ismej2017119



OTU vs ASV



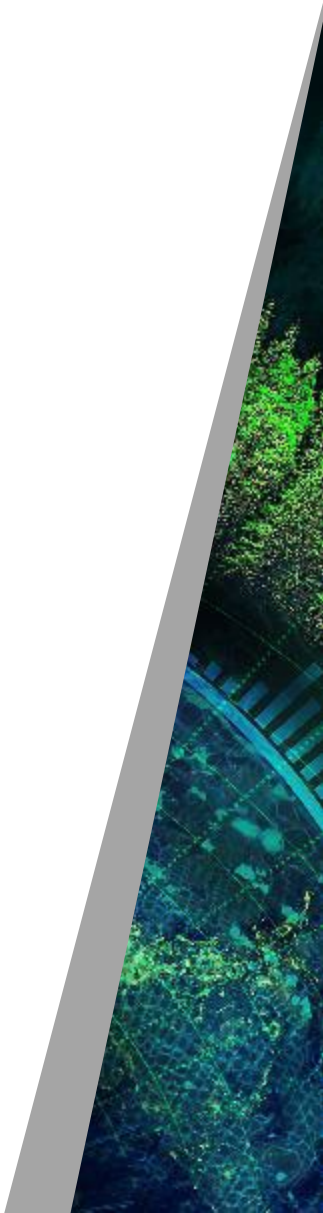
https://trainings.migale.inrae.fr/posts/2022-06-07-module20/content/slides.html#55



https://en.wikipedia.org/wiki/Amplicon_sequence_variant

For metabarcoding (16s)

- https://benjjneb.github.io/dada2/ (ASV)
- https://mothur.org/ (OTU)
- https://qiime2.org/ (OTU)



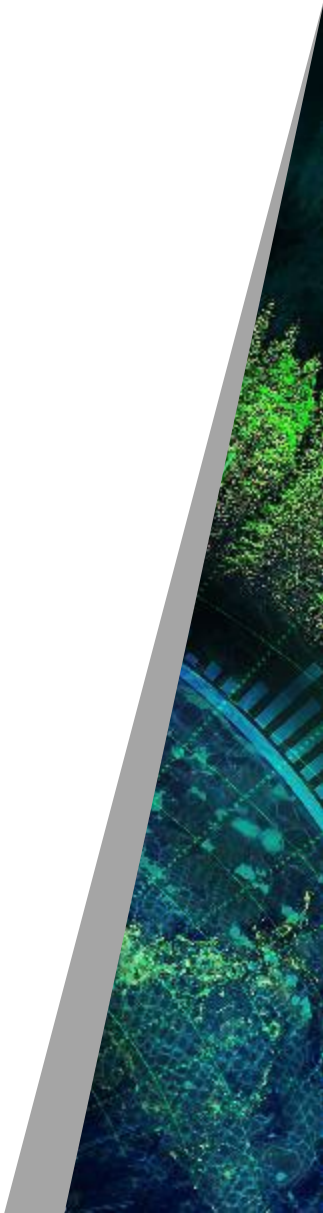
Shotgun a.k.a Whole Genome Sequencing

Whole Genome sequencing Analysis Methods

Assembly-free	Assembly based
map/classify reads	Assembling reads into contigs
Direct annotation	Binning the contigs by similarity
Heavily database reliant	Annotation of the bins
Faster, easier	Slow, tricky
More noisy	More specific

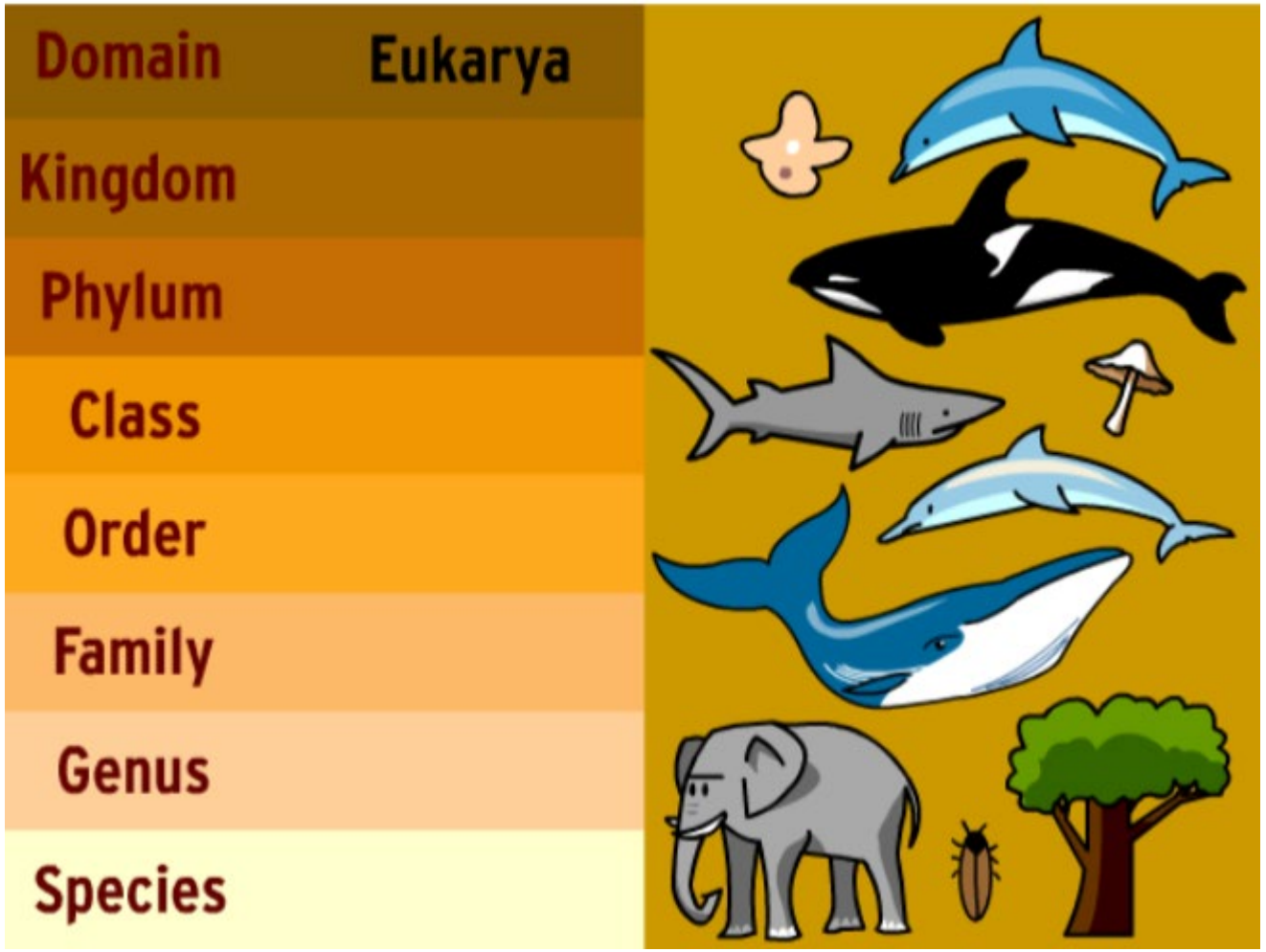
Assembly free

- The inventory approach:
  - Annotating the reads directly
  - Heavily relying on databases
  - Good for well known microbiomes (western European poop, agricultural soils)
  - Dangerous for others as methods tend to overclassify



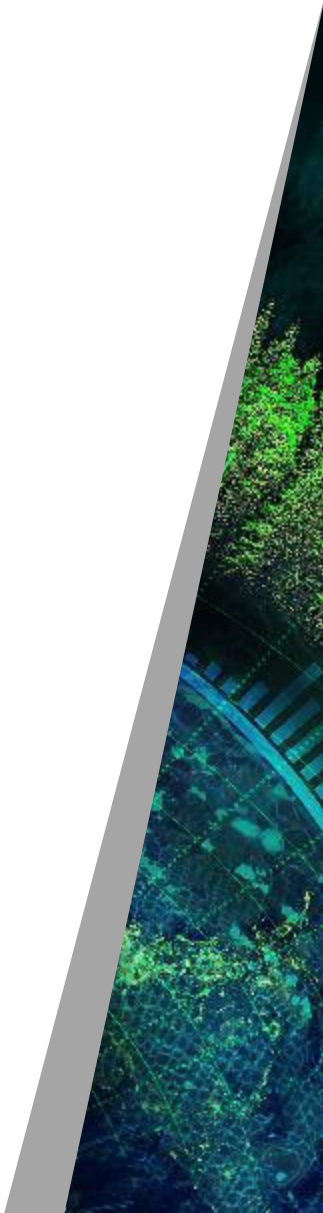
Assembly free

- Taxonomically :
  - Kraken
  - Kaiju
  - Functionally :
  - Blast …
  - Megan
  - Humann



Tools

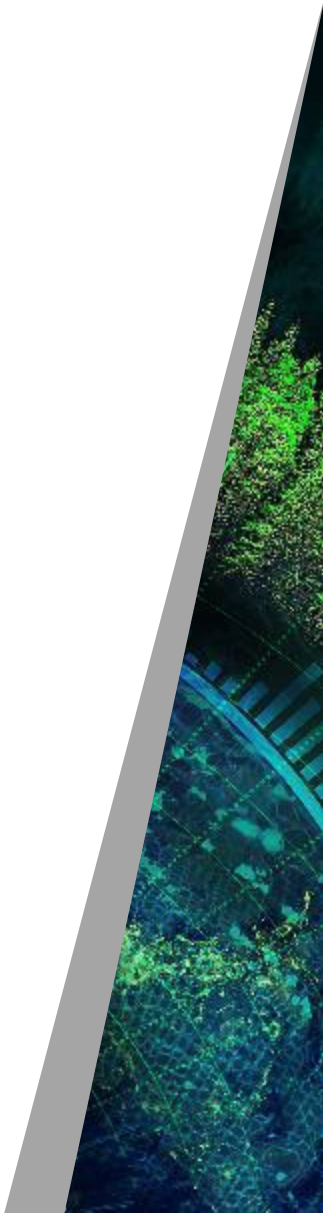
- Kraken2/Bracken
  - https://github.com/DerrickWood/kraken2
  - https://github.com/jenniferlu717/Bracken
  - DB: https://benlangmead.github.io/aws-indexes/k2
- Kaiju
  - https://github.com/bioinformatics-centre/kaiju
  - DB: https://bioinformatics-centre.github.io/kaiju/downloads.html
- Krona
  - https://github.com/marbl/Krona/wiki



Assembly based

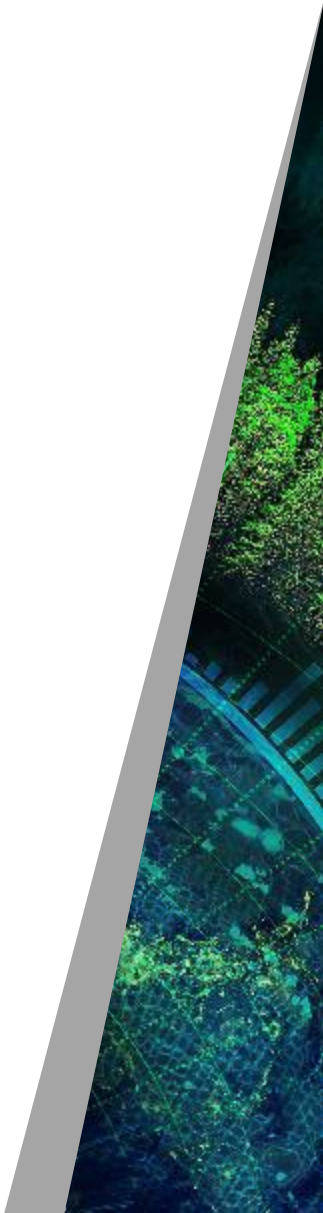
- Turning the reads into genes and more! (aka The Contig):
  - Full length coding sequences
  - Pieces of genomes
  - Or anything else made of DNA

- Increases specificity of previous approaches, but might lose some sensitivity (e.g. not everything assembles

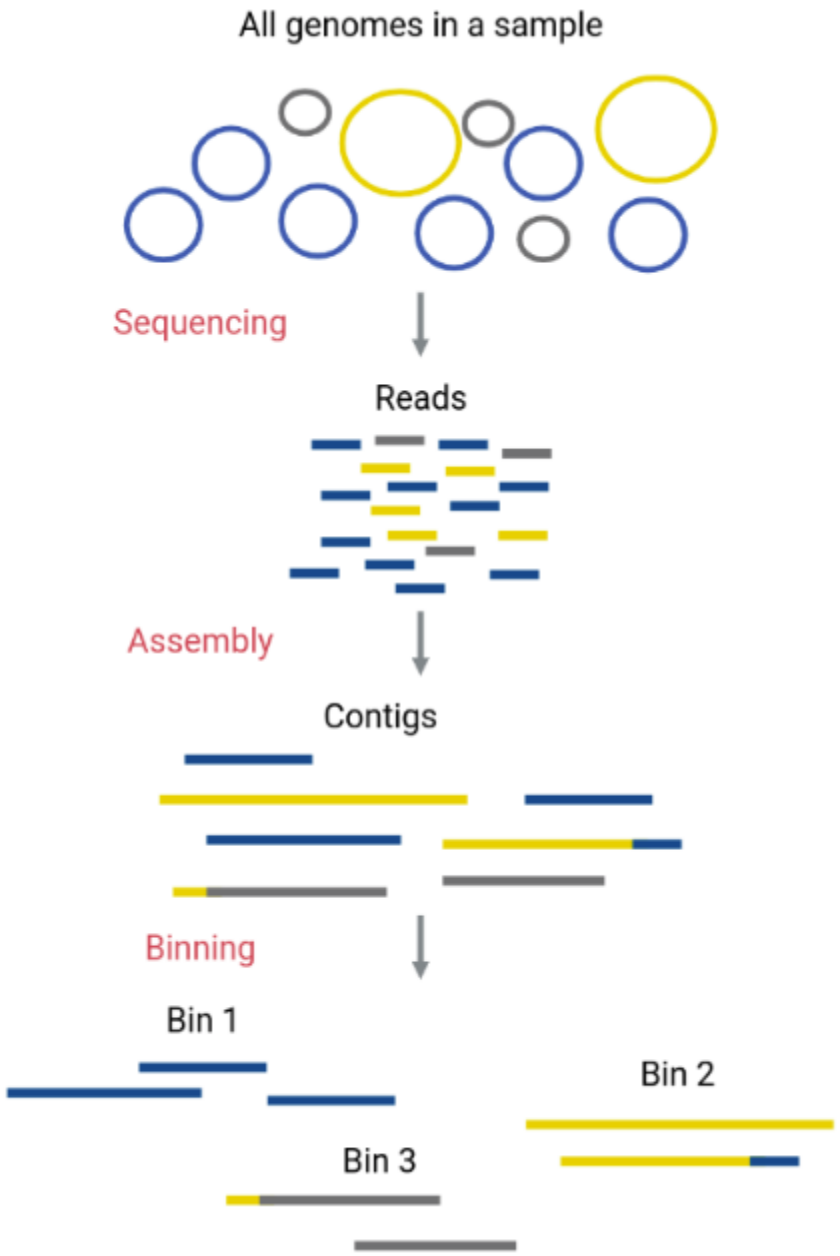


Assembly based

- Genome-resolved metagenomics:
  - Cluster contigs into bins
  - Identify possible genomes
  - Quantify completeness/redundancy
  - Assign taxonomy



Binning


All genomes in a sample
Sequencing
Reads
Assembly
Contigs
Binning
Bin 1
Bin 2
Bin 3

Gene Atlas

Gene Atlas Approach

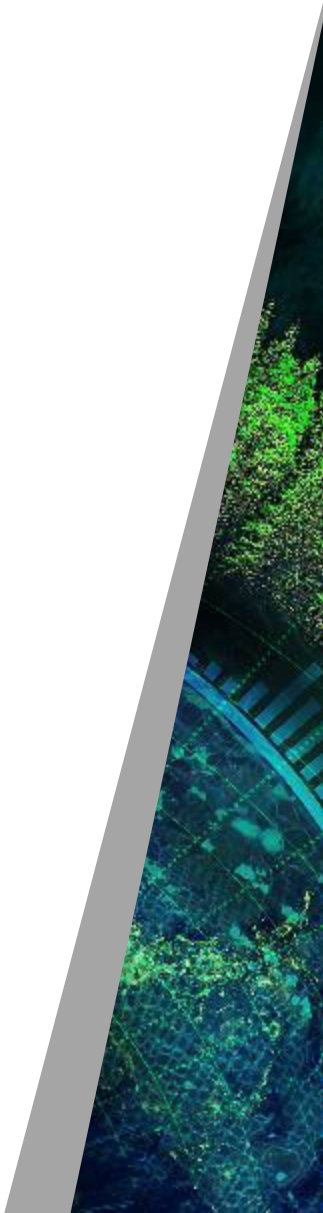
If you only care about genes:
- Predict genes and annotate genes:
- Functionally
- Taxonomically
- Gene clustering:
- By similarity
- Into COGs
- Into families
- Quantify and analyse diversity

Big Data: Problems

Storage

- We want to sequence the soil of grapevines with nanopore MinIon (rough output between 100Gb and 300Gb of raw files) once basecalled it becomes 5-15Gb of fastq reads
- To have a good experiment we need at least 5 samples per type of soil and grapevine.
- Let say you want to try 3 different grapevine breed through 2 farms that leave you with the need to store at minimum:
- 3,15 Tb of sequencing data for 1 minimal experiment

(100+5)*5*3*2=3150

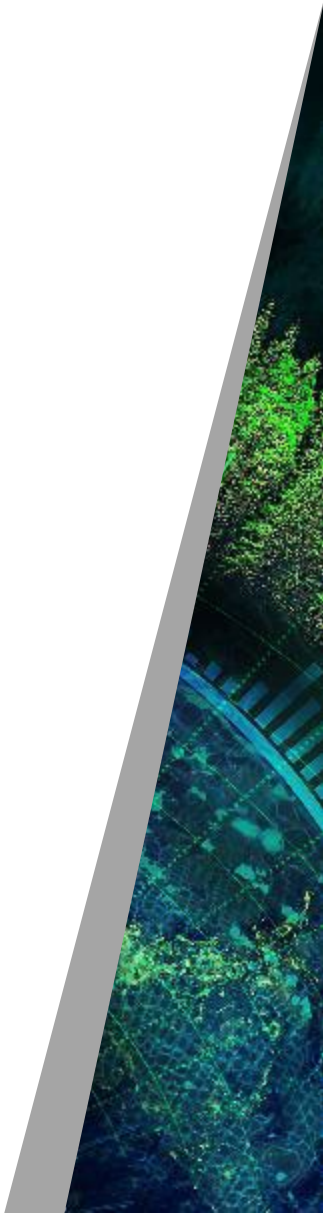


Time

To reproduce a typical computational biology paper with minimal expertise it takes 280 Hours.
https://doi.org/10.1371/journal.pone.0080278

That include:

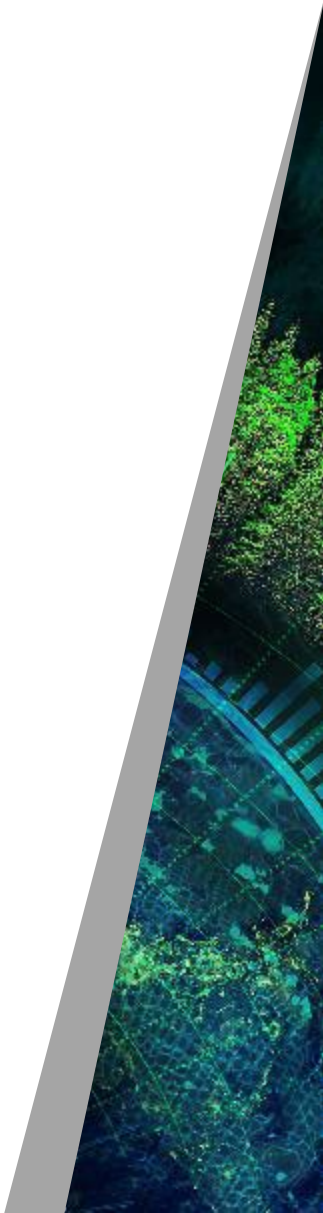
- Understanding the paper and material used
- Installation, set up
- Finding parameters
- Workflow validation
- Computing



Reproducibility

1. Access to the tools mentioned in the paper
2. Difficulty of installation
3. Version issues (machine and software)

https://doi.org/10.1371/journal.pbio.3000333



The issues of Big Data:

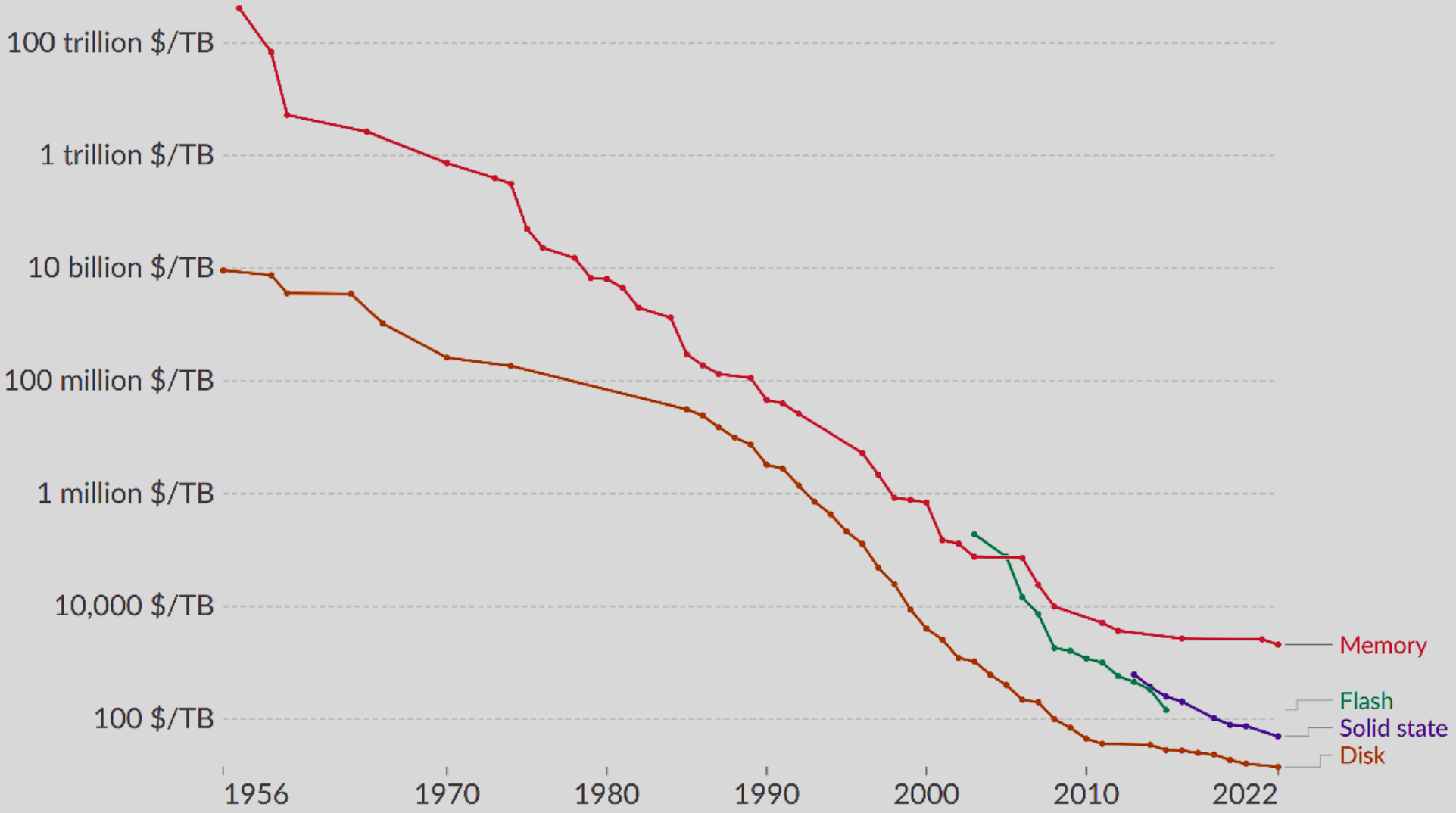
1. Storage
2. Reproducibility
3. Computing Power
4. Time
5. Money
6. Consistency
7. Usability

Big Data: Solutions

Historical cost of computer memory and storage

This data is expressed in US dollars per terabyte (TB). It is not adjusted for inflation.

Our World in Data



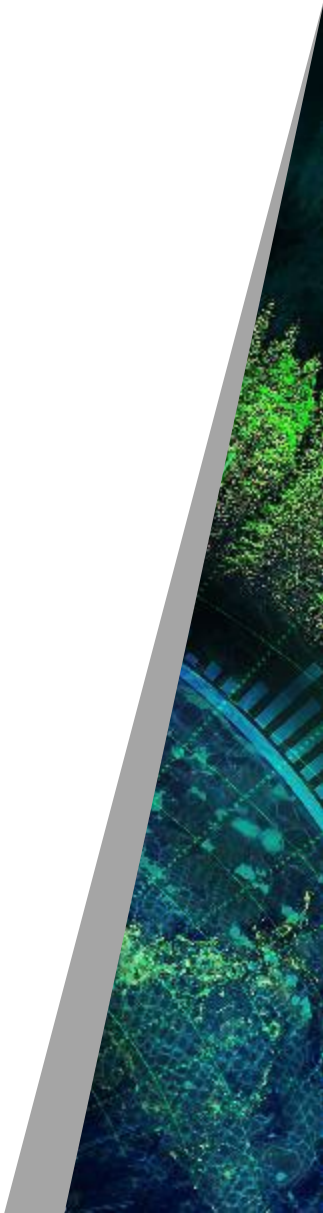
Source: John C. McCallum (2022)

OurWorldInData.org/technological-change • CC BY

Note: For each year, the time series shows the cheapest historical price recorded until that year.

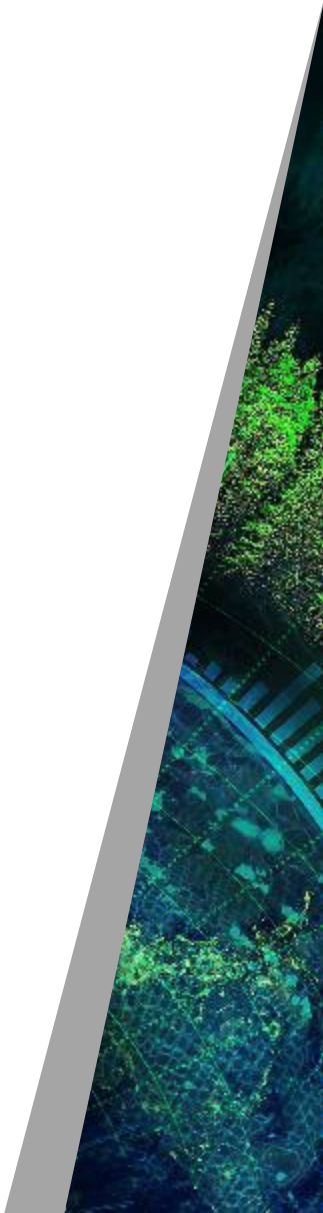
Workflow Manager

- One command runs the entire workflow
- Parallelisation is automatic and scheduled based on the available resources
- Allow reuse of code in other scripts
- Continuous checkpoints for resuming and expanding the pipelines



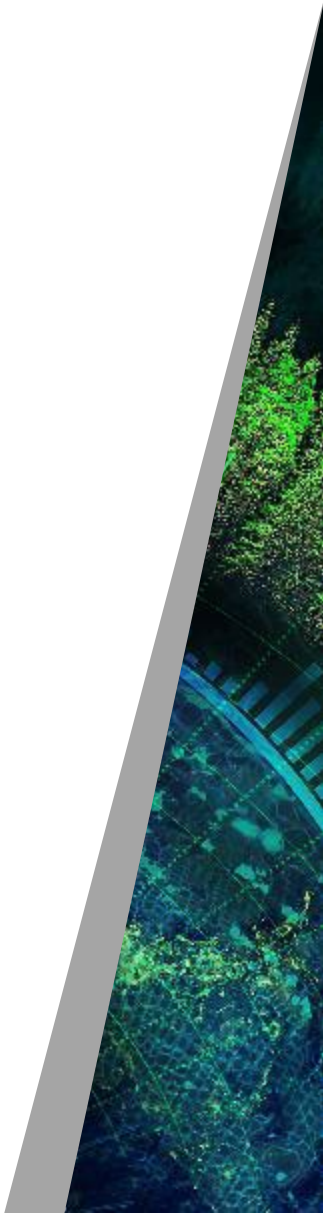
Usability - Workflow Manager

- One command runs the entire workflow
- Parallelisation is automatic and scheduled based on the available resources
- You can write your internal code in any language (bash, R, Python)
- The intermediary files are handled
- Allow reuse of code in other scripts
- Continuous checkpoints for resuming and expanding the pipelines



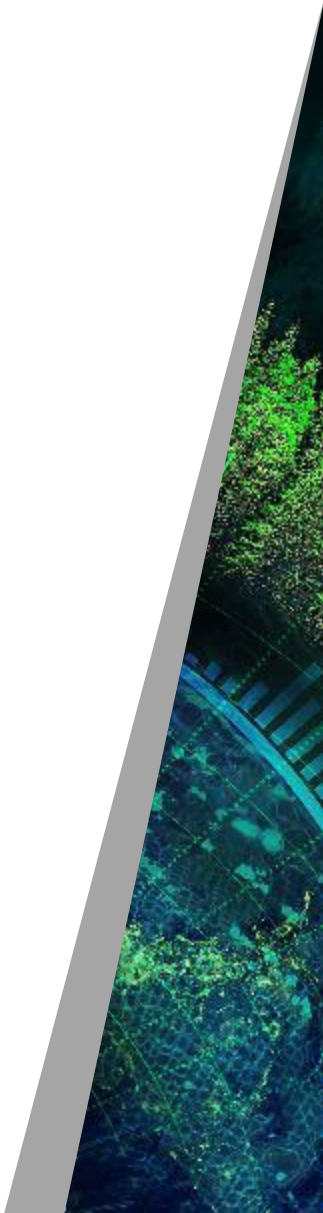
Reproducibility - Workflows Manager

- Pipeline written in stone and version controlled (Git)
- Support for containers and in some cases conda environment
- The application handling and the configuration/deployment are separated
- The only parameters to change are the resources and environment



Consistency

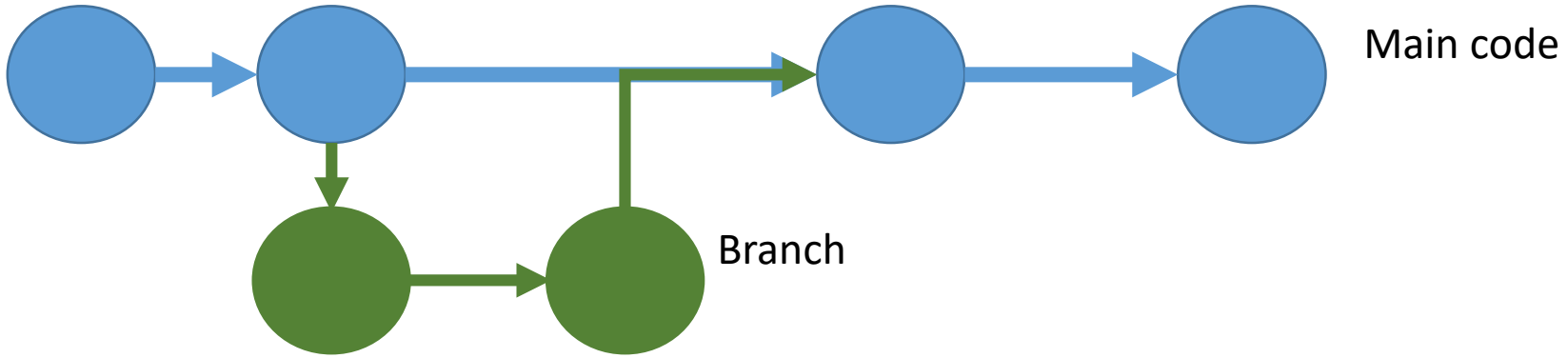
- **Version control (git, bitbucket, github, gitlab)**
- **Container (Apptainer, Docker, Podman)**
- **Open source**



Tools for a simpler life

Git

- version control system that handles all from small to very large projects
- Simple to use
- Track changes and save specific versions, those version can be changed but are then new versions



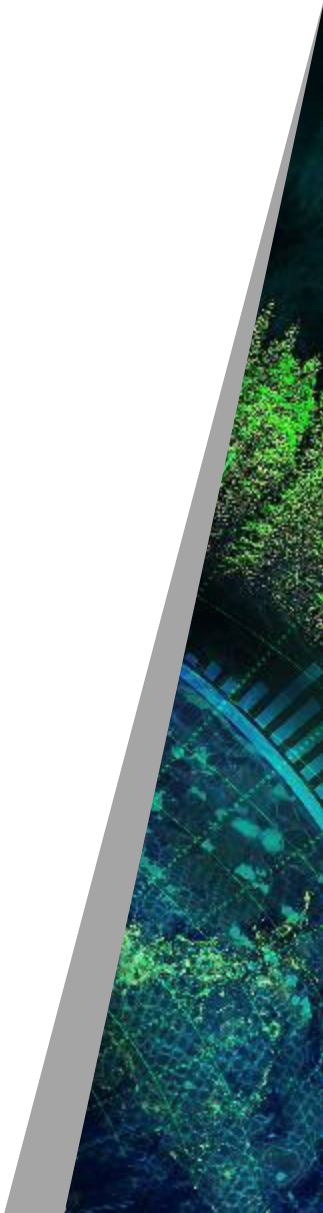
Container vs Virtual Machine

- **Lighter: MB vs GB**
- **Faster startup as it virtualises a process and not an OS**
- **Immutable: does not change over time**
- **Composable: The output of one can be fed directly in the next**
- **Transparent**

Genomic Workflows

Genomic workflows

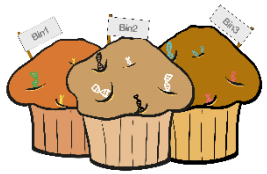
- Data analysis apps to retrieve information from datasets
- Huge parallelisation, creating numerous job over a cluster
- Mix of tools and scripts
- Complex configurations and dependencies interactions between said tools



Workflow Manager

SnakeMake	Nextflow
CL oriented tool	CL oriented tool
Rules defined using file name pattern	Can manage any data structure
Built-in support Apptainer	Support all major container runtimes
Custom script for cluster	Built-in support for cluster and cloud services
No support for source code management system	Built-in support for Git/GitHub, manage pipeline revisions
Python based	Groovy/JVM based

For metagenomics (shotgun)



- https://github.com/RVanDamme/MUFFIN
- https://nf-co.re/mag/2.4.0
- https://github.com/metagenome-atlas/atlas
- …

SCIENCE AND EDUCATION FOR SUSTAINABLE LIFE