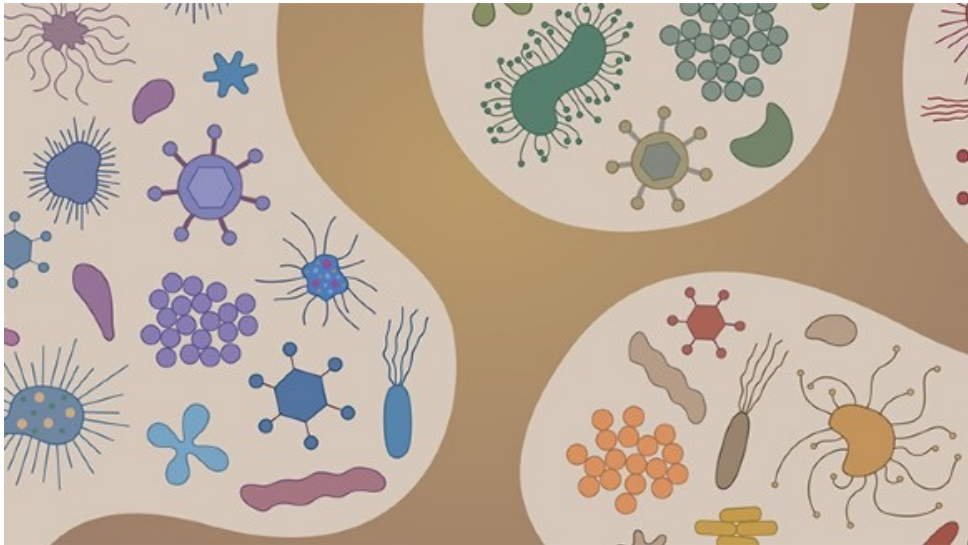


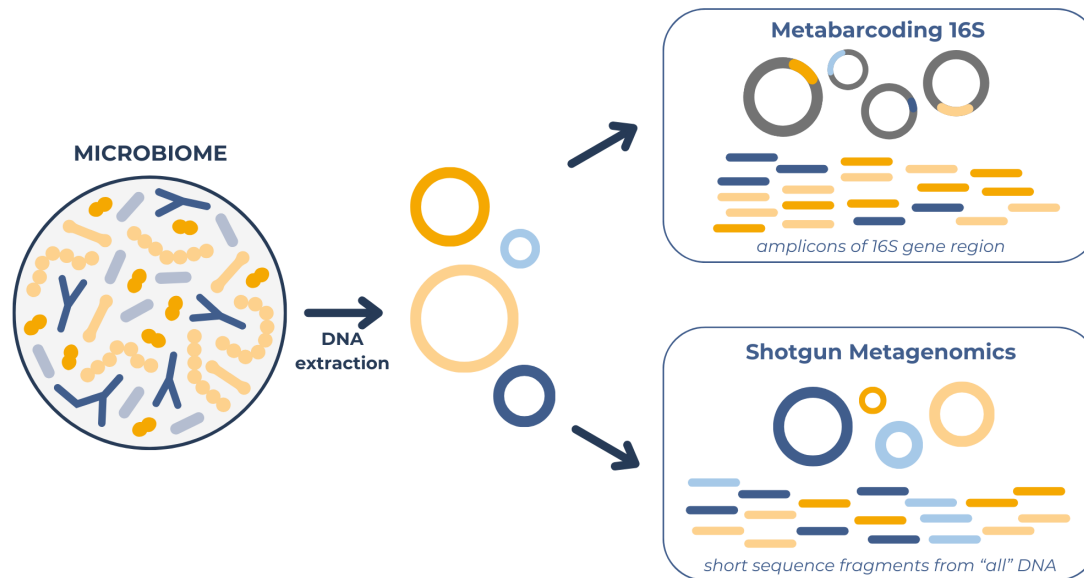
Analysis of Shotgun Metagenomics data

Juliette Hayer



Microbiome data

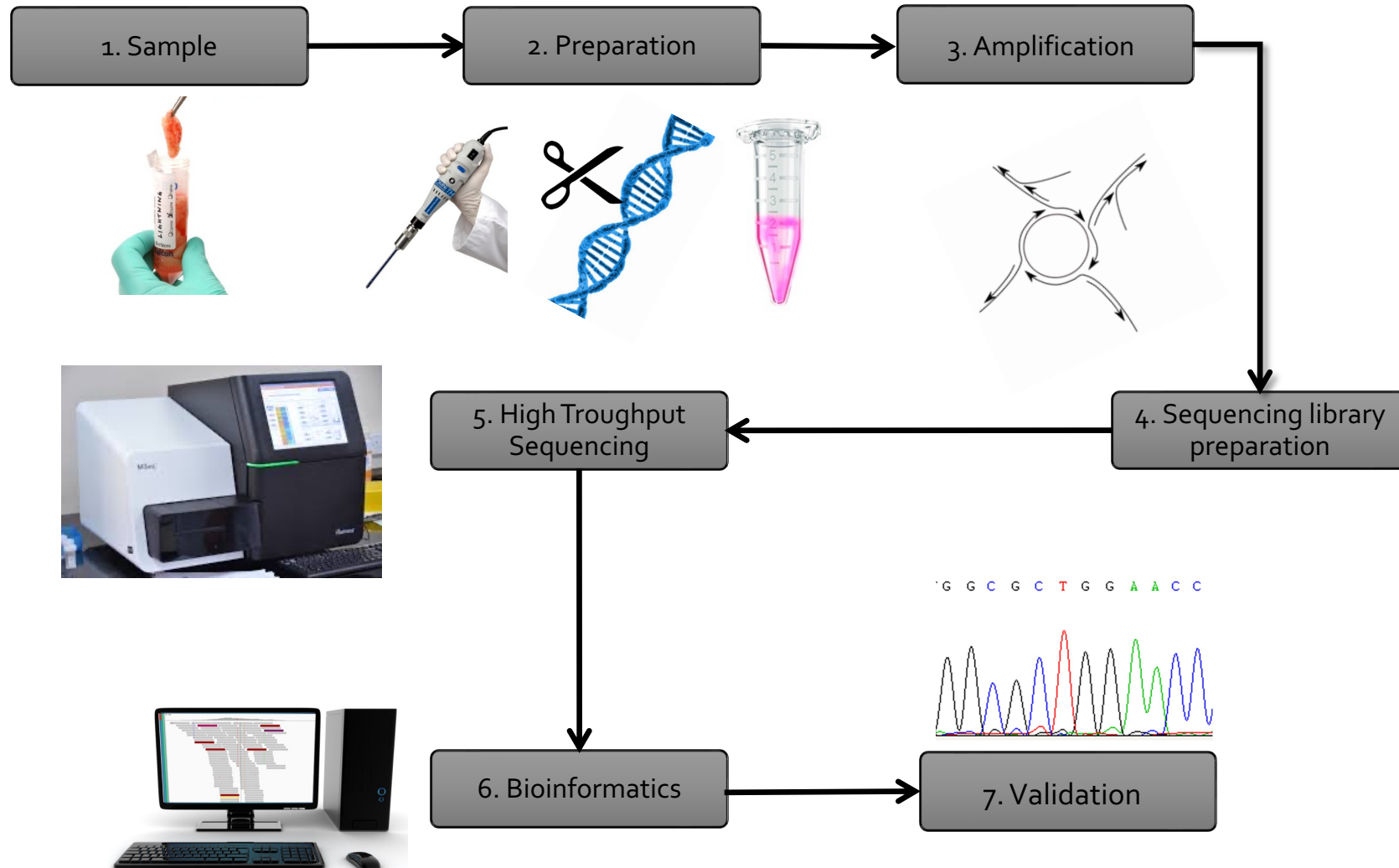
- Microbiome data can be gathered from different environments such as soil, water or the human gut.
- The biological interest lies in general in the question how the microbiome present at a specific site influences this environment.
- To study a microbiome, we need to use indirect methods like metagenomics or metatranscriptomics.



Metagenomics approach

- **Metagenomic samples** contain DNA from different organisms at a specific site, where the sample was collected.
- Metagenomic data can be used to find out which organisms coexist in that niche and which genes are present in the different organisms.
- The investigation of microorganisms present at a specific site and their relative abundance is also called “**microbial community profiling**”.
- The main objective is to identify the microorganisms that are present within the given sample. This can be achieved for all known microbes, where the DNA sequence specific for a certain species is known.
- For that we try to **identify the taxon** to which each individual read belongs.

Methodology: from samples to data



Taxonomic profiling

- The question:
 - Which organisms are present, and in what relative abundance?
- How to achieve: classification of individual reads to reference taxa to extract the relative abundances of the different taxa

The reported result may differ from the true composition because of:

- Unequal genome sizes.
- Unequal sequencing coverage.
- Conserved regions shared between organisms.
- Contamination.
- Database composition.
- Different classification thresholds.

```
Sample reads
├─ Organism A: 60%
├─ Organism B: 30%
└─ Organism C: 10%
```

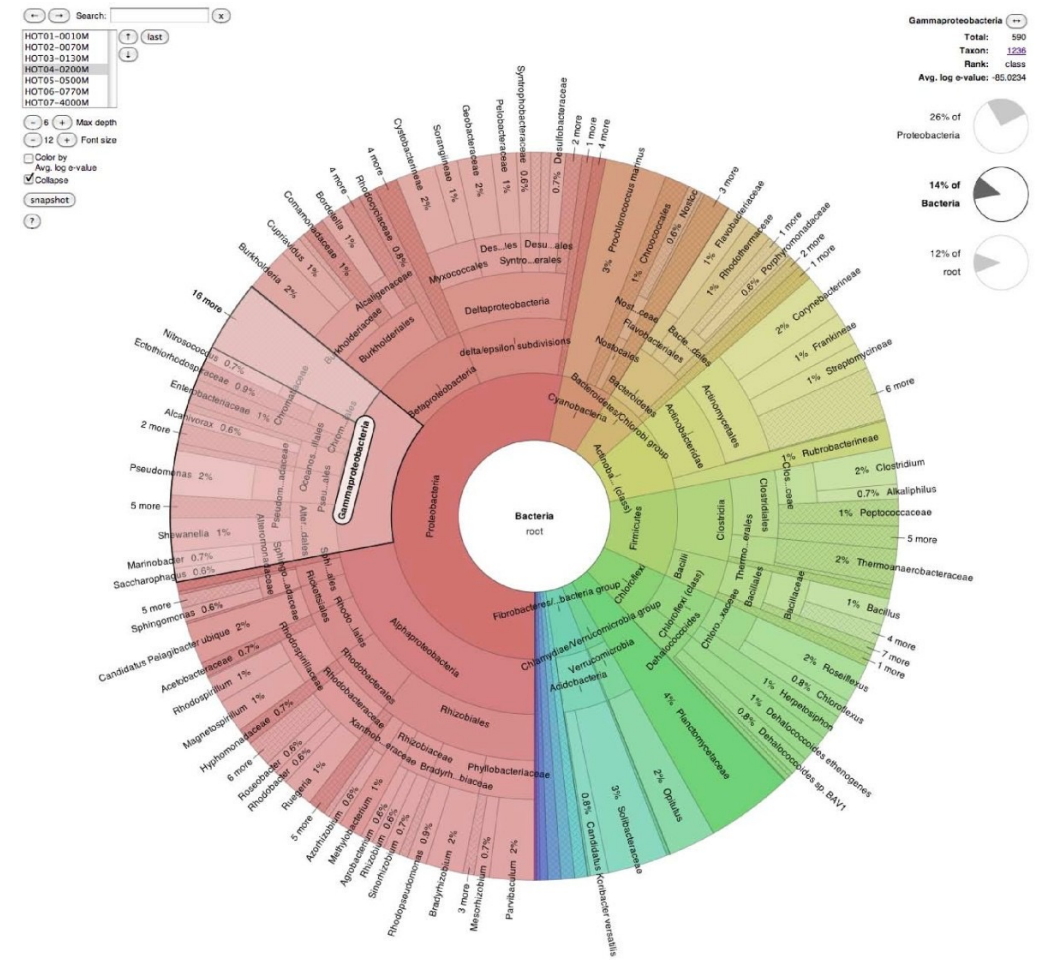
```
Reads → Database 1 → Species A: 60%, Species B: 30%
Reads → Database 2 → Species A: 45%, Species B: 25%, Unknown: 30%
```

Taxonomic classification and profiling methods

Approach	Principle	Example tools	Strengths	Limitations
k-mer classification	Matches short sequence fragments to reference genomes	Kraken2	Very fast and sensitive	Strongly database-dependent
Abundance estimation	Re-estimates species abundances from classified reads	Bracken	Improves abundance estimates after Kraken2	Inherits classification errors
Abundance estimation based on ANI and k-mer statistics	estimates genome-to-metagenome containment average nucleotide identity (ANI) through zero-inflated Poisson <i>k</i> -mer statistics, enabling ANI-based taxa detection	Sylph	Very fast and efficient Useful for low-abundance organisms Scales large database	Database dependant Limited strain-level resolution
Marker-gene profiling	Uses clade-specific conserved markers	MetaPhlAn	More conservative and interpretable	May miss poorly characterized organisms
Read mapping	Maps reads against reference genomes	Bowtie2, minimap2	Useful for targeted organisms and strain analysis	Requires suitable references

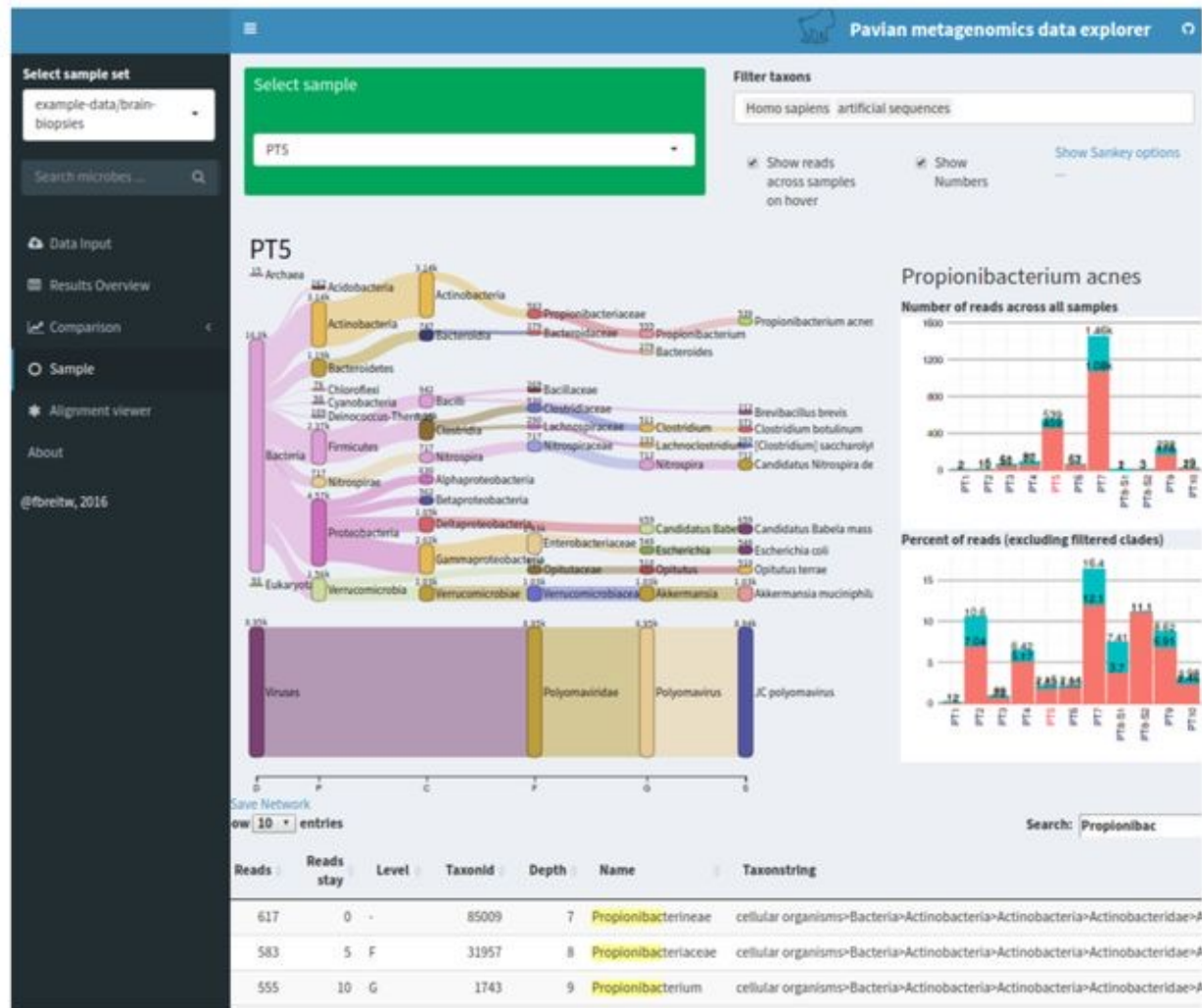
Taxonomic profiling: outputs

- For detection of highly distant species (e.g. viruses); we use methods at protein level:
 - Diamond (blastx)
 - Kaiju (kmer-based protein search)
- Reports of abundance from Kraken2 and Bracken
- Visualisation with Krona



Taxonomic profiling: visualisation with Pavian

A

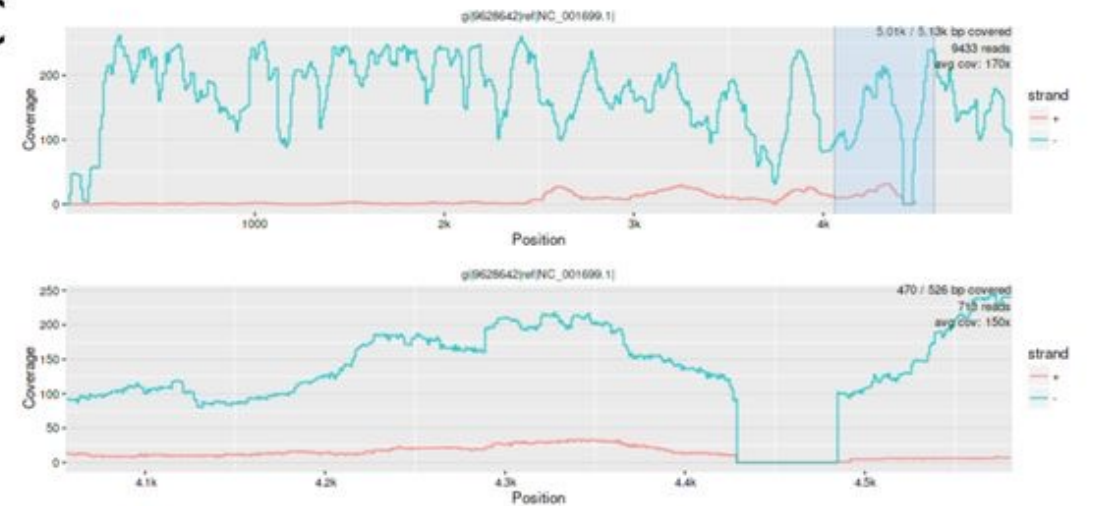


B

Show 10 rows Column visibility PDF Excel CSV Copy Search:

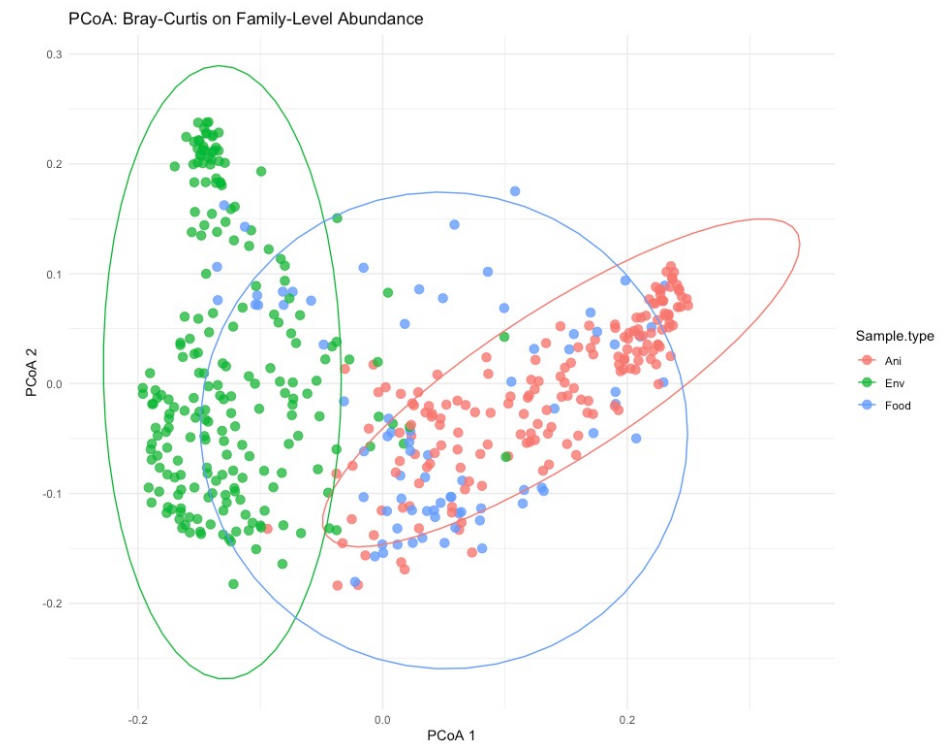
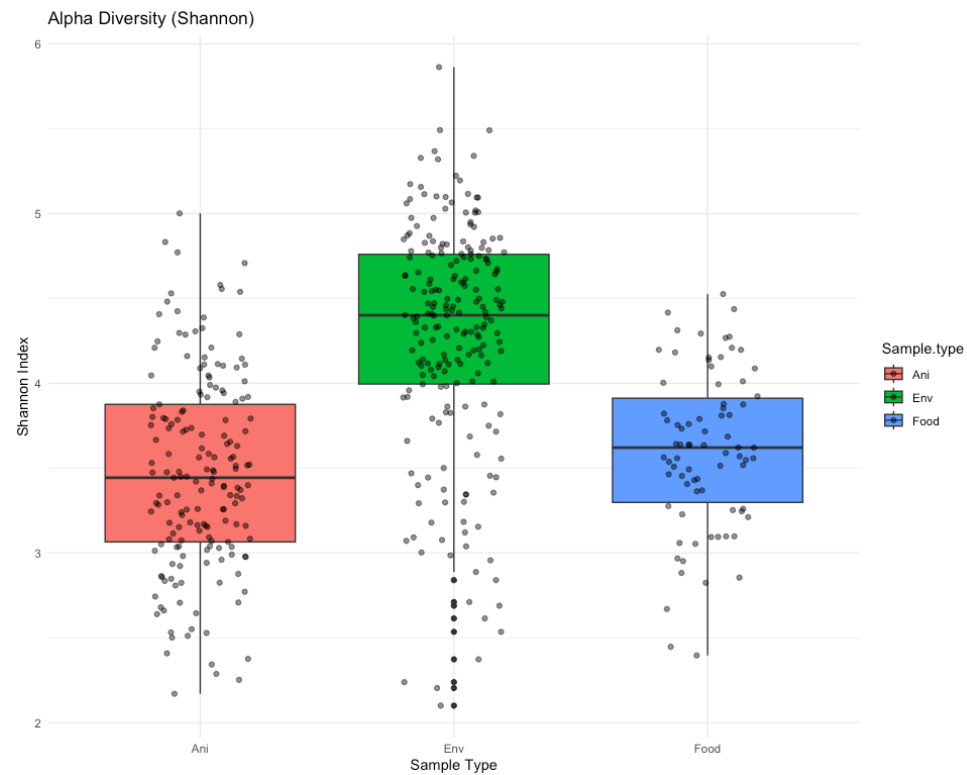
Name	Level	Taxonid	PT1	PT2	PT3	PT4	PT5	PT6	PT7	PT8-S1	PT8-S2	PT9	PT10
Saccharomyces cerevisiae	S	4932	1.075%		0.01725%	0.1773%	0.004227%			19.23%	4.167%	23.43%	49.57%
Propionibacterium acnes	S	1747	0.7168%	16.13%	0.8625%	12.41%	1.94%	3.215%	12.75%	3.846%	12.5%	8.26%	5.507%
Mycobacterium tuberculosis complex	-	77643							0.03549%		58.17%	0.0464%	
JC polyomavirus	S	10632					37.81%	0.846%					
Bacteria	D	2	1.434%	0%	1.208%	20.04%	6.797%	2.594%	1.29%	0%	0%	0.6032%	1.159%
Deitlia acidovorans SPH-1	-	398578	1.075%	14.52%	1.501%				1.514%	11.54%		0.1856%	
Propionibacterium acnes ATCC 11828	-	1091045		4.839%	0.1725%	3.014%	0.3382%	0.3384%	4.082%	3.846%		1.763%	1.159%

C

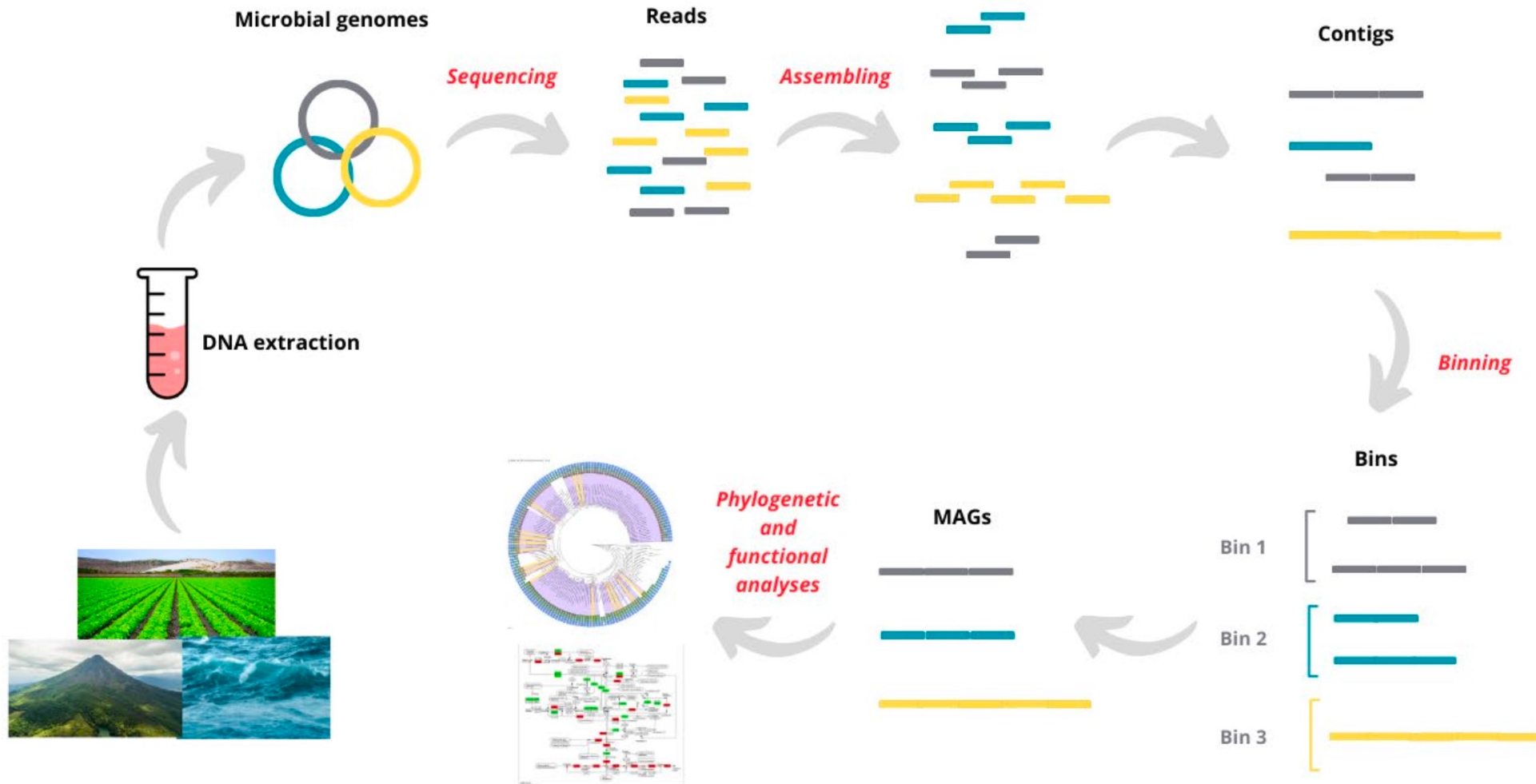


Diversity metrics

- From the Bracken abundance:
 - compute alpha and beta diversity; for further statistical analysis



Metagenomics: MAGs reconstruction process



Why to assemble?

Assembly combines overlapping reads into longer sequences called contigs.

Longer contigs provide more context for:

- Gene prediction.
- Operon reconstruction.
- AMR gene context.
- Mobile genetic element analysis.
- Genome binning.
- Recovery of MAGs.



Metagenomics assembly

Metagenomic assembly is further complicated by

- the large volume of data produced
 - the quality of the sequence
 - the unequal representation of members of the microbial community
 - the presence of closely related microorganisms with similar genomes
 - the presence of several strains of the same microorganism
 - an insufficient amount of data for minor community members
-
- The different assemblers (e.g. **metaSPAdes**, **MEGAHIT**) have different computational characteristics and their performance varies according to the microbiome
 - => shown in by both rounds of Critical Assessment of Metagenome Interpretation initiative (CAMI)

About the co-assembly

- To reduce the differences in coverage between samples, we can use a **co-assembly** approach, where reads from all samples are aligned together

Pros of co-assembly	Cons of co-assembly
More data	Higher computational overhead
Better/longer assemblies	Risk of shattering your assembly
Access to lower abundant organisms	Risk of increased contamination

Co-assembly is reasonable if:

- Same samples
- Same sampling event
- Longitudinal sampling of the same site
- Related samples

Example where co-assembly would NOT be recommended:

Samples from different patients => **individual assembly** is preferred.

From contigs to MAGs: principles of binning

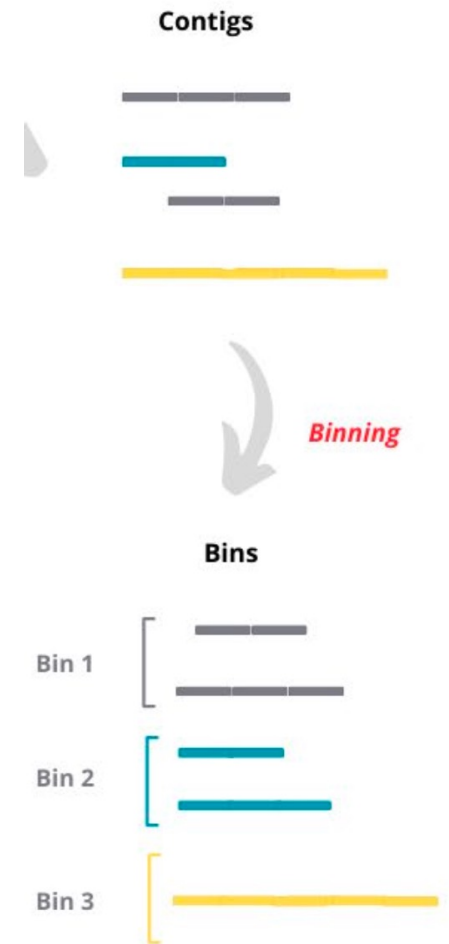
A Metagenome-Assembled Genome is a draft microbial genome reconstructed from metagenomic sequencing data.

It is not necessarily a complete genome and should not automatically be interpreted as an isolate genome.

Binning is used for grouping contigs that are likely to originate from the same organism; using various features/signals:

- Sequence composition, particularly tetranucleotide frequency.
- Relative abundance and coverage.
- Presence of genome-specific marker genes.

Binning tools such as MetaBAT2, CONCOCT and MaxBin2 use sequence composition and coverage information to group contigs into bins.

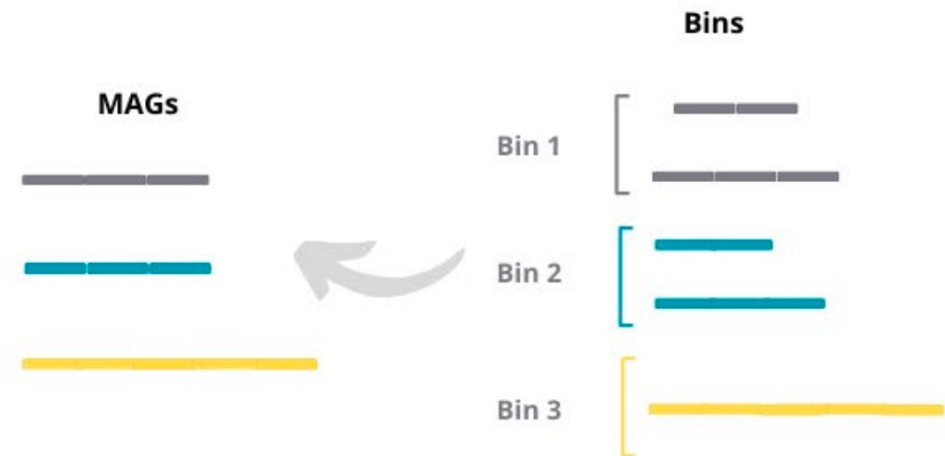


From contigs to MAGs: bin refinement

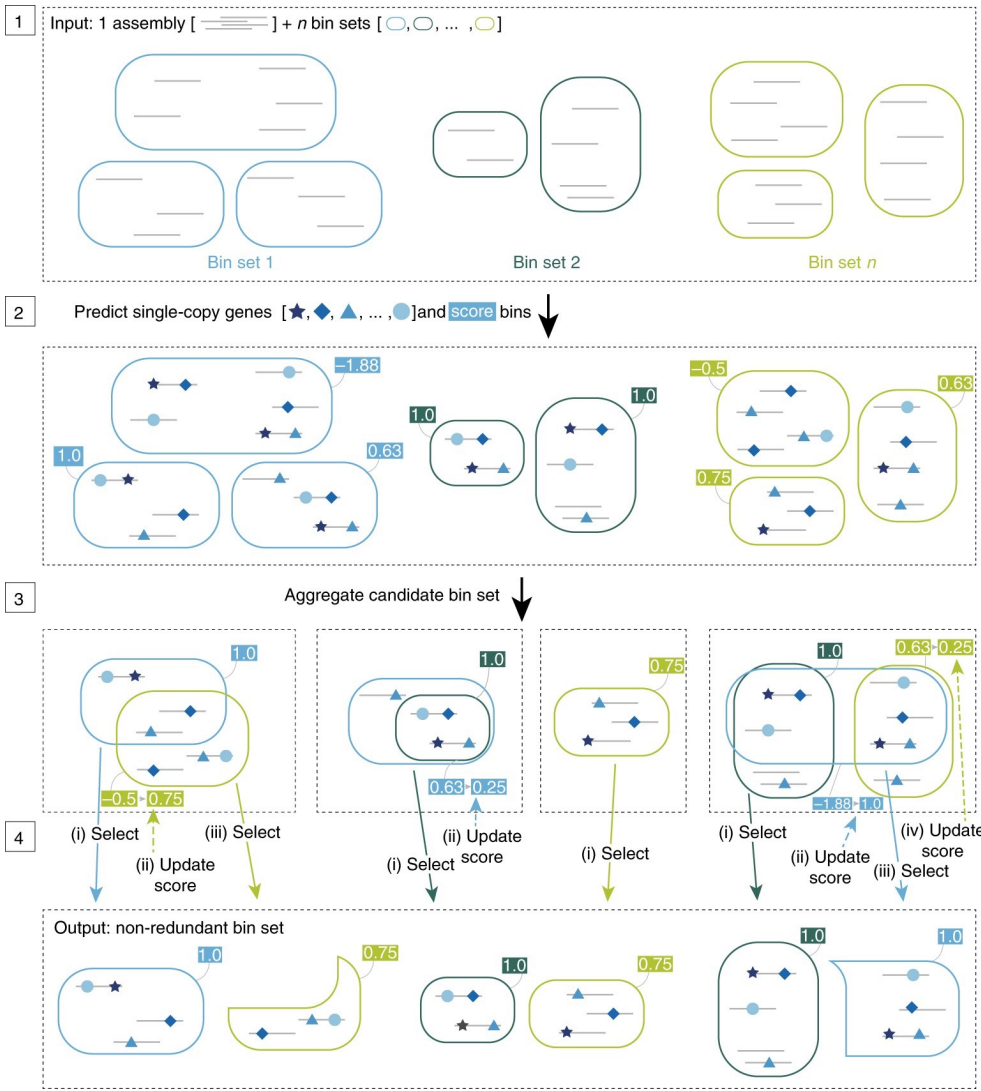
- Different bidders can assign contigs differently; no single algorithm performs optimally for every genome or community.
- Bin refinement integrates several binning results to select or construct better-quality, less redundant genome bins.

Objectives:

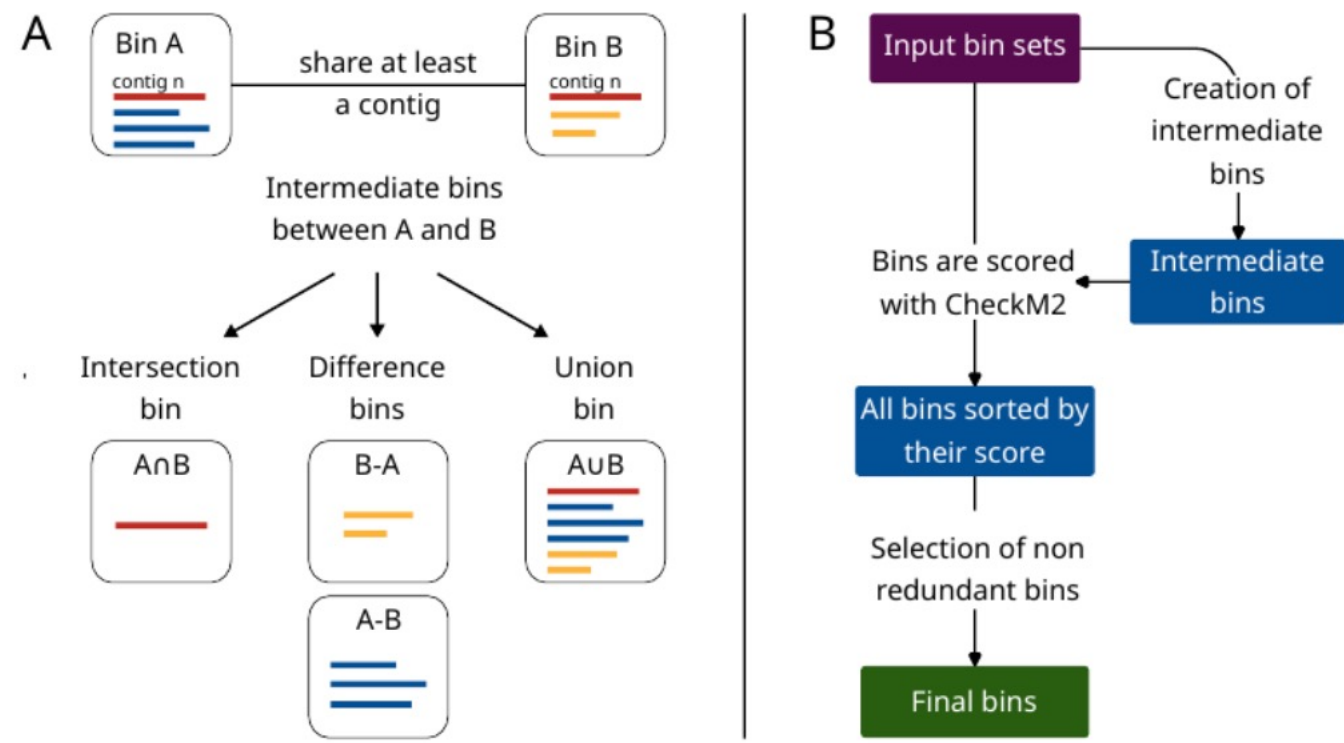
- Reduce contamination.
- Improve completeness.
- Resolve conflicting contig assignments.
- Remove redundant bins.
- Recover complementary contigs.



Bin refinement tools



DAS Tool; Sieber et al., Nature Microbiol. 2018



Binette; Mainguy et al., JOSS. 2025

Quality control of MAGs

- Refinement does not guarantee a biologically correct genome.
- A higher completeness score may come at the cost of increased contamination.
- Hybrid bins can be chimeric if contigs from different organisms are combined.

Refined MAGs should be checked using:

- Completeness and contamination.
- Coverage consistency.
- Taxonomic consistency.
- Marker-gene duplication.
- GC-content and coverage patterns.

Bin refinement combines alternative binning results to obtain cleaner, more complete and less redundant MAGs => but every refined MAG still requires independent quality control.

Bioinformatics workflow

