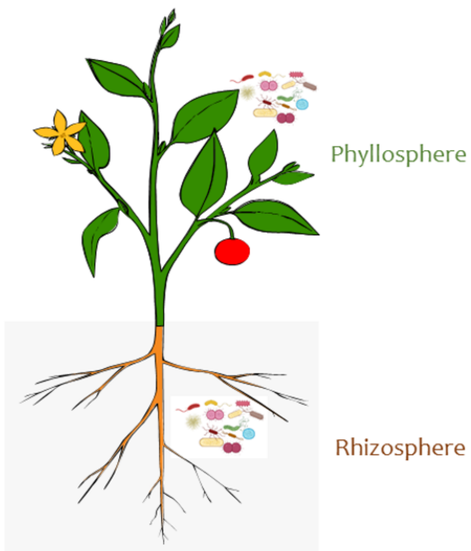
# Metagenomic assembly of complex ecosystems with highly accurate long-reads

Nicolas Maurice, Claire Lemaitre, Clémence Frioux, Riccardo Vicedomini

# Metagenomic assembly of complex ecosystems with highly accurate long-reads

Nicolas Maurice[1,2], Claire Lemaitre[1], Clémence Frioux[2], Riccardo Vicedomini[1]
[1] Inria/IRISA Genscale, Campus de Beaulieu, 35042 Rennes, France
[2] Inria/INRAE Pleiade, Campus de Talence, 33405 Bordeaux, France
nicolas.maurice@inria.fr

## Context




### Plant microbiota

Plants can recruit micro-organisms from the millions ($10^3$-$10^8$/g) found in soil, to regulate key functions such as :

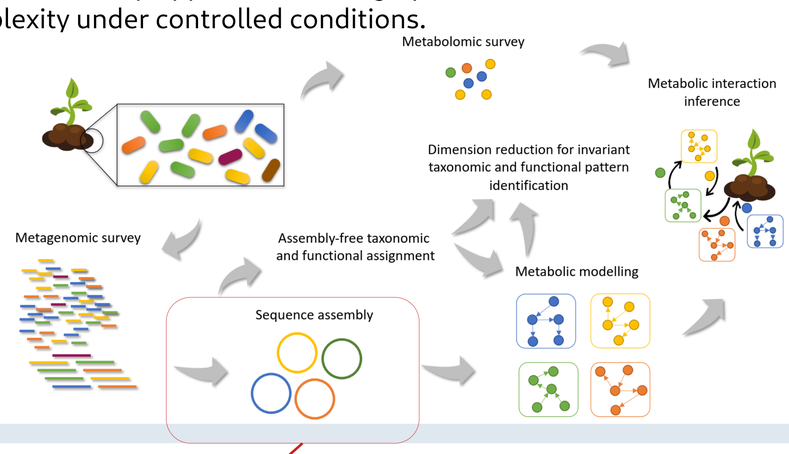
- Nutrition and growth,
- Abiotic stress,
- Biotic stress.

Still, both ecosystems are poorly caracterised, as culturomics, transcriptomics and short-read metagenomics cannot recover enough complete genomes needed to modelise and caracterise those communities.

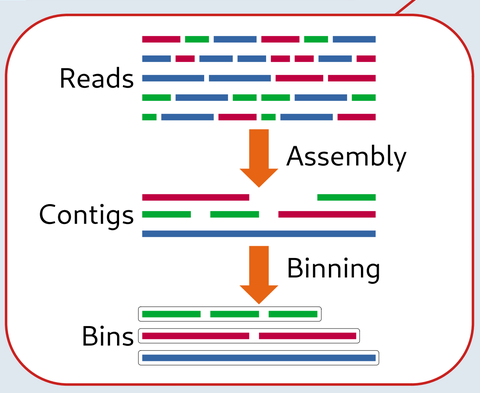
### MISTIC Project : project.inria.fr/mistic

Methodological development for microbial community dynamics modelling

- WP1 : Top down approach from large-scale data on natural communities,
- WP2 : Bottom up approach, building dynamic models of increased complexity under controlled conditions.




## Motivations



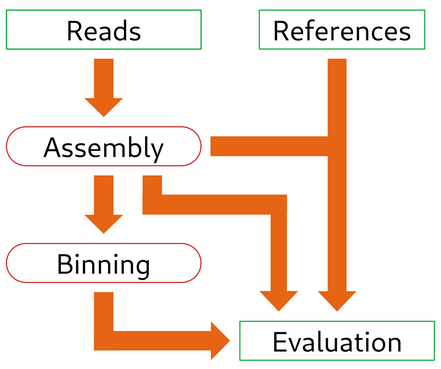

### Emerging sequencing technologies

| | Read type | Read length (bp) | Error rate |
|---|---|---|---|
| Illumina | Short reads | 151 bp | ~0.1 |
| ONT R9.4 | Long reads | ~20kb | ~10% |
| PacBio CLR | | ~10kb | ~10% |
| ONT R10.4 | Highly accurate long reads | ~20kb | ~1% |
| PacBio Hi-Fi | | ~10kb | ~1% |

Higher quality reads lead to higher quality assemblies.

**How to exploit those highly accurate long-reads to recover high-quality genomes from complex environements such as soil and rhizosphere ?**

## Mapler : gitlab.inria.fr/mistic/mapler




### Pipeline

Built with Snakemake and Slurm integration.

Benchmarks three highly accurate long reads assemblers :

- Metaflye[1],
- Hifiasm-meta[2],
- MetaMDBG[3].

### Datasets

| | Total size (Gb) | Median size (Kb) | Taxonomic complexity |
|---|---|---|---|
| Zymo D6331 | 18.0 | 8.1 | Mock community |
| Human gut | 15.2 | 8.4 | + |
| Rhizosphere | 19.5 | 5.8 | ++ |
| Soil | 13.4 | 6.0 | +++ |

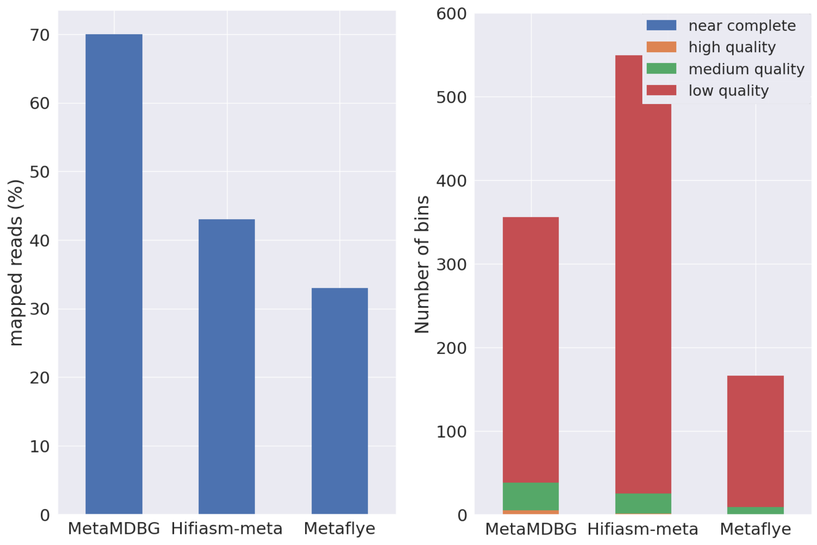
### Evaluation metrics

- Aligning contigs on reference genomes, if aviable : NGA50 and genome fraction,
- Aligning reads on contigs : counting reads that map to the assembly,
- Evaluating the bins completeness and contamination with marker genes, then categorizing bins by level of quality.

## Mapler : Results

### Assembler comparison

On the soil sample

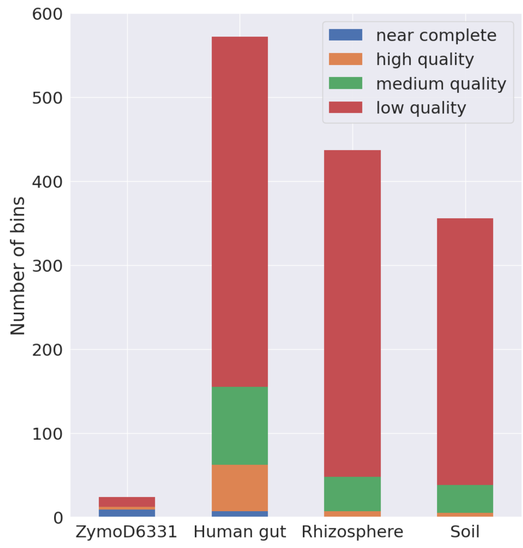



MetaMDBG captures :

- A greater proportion of reads,
- More high-quality bins,
- More medium-quality bins.

### Sample comparaison

Using metaMDBG

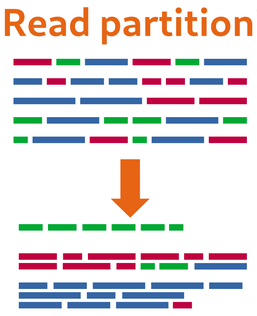



As taxonomic complexity increases :

- Quality of bins decreases,
- Proportion of captured taxa decreases.

## Methodological avenues of improvement

### Read partitioning

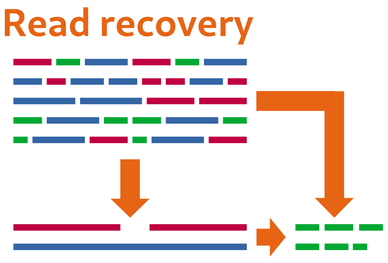


By partitioning the reads into smaller sets via :

- Taxonomic assignation,
- Composition-based clustering,
- Abundance-based clustering.

We could reduce the complexity of each subset, simplifying the assemblies.

### Read recovery



By mapping the reads on contigs, we could extract unassembled reads, and assemble those separately with reduced complexity.

### Hybrid assembly or binning

Highly accurate long reads could be complemented by a higher coverage of short reads, which could be used to better assess the abundance of each long-read, leading to a better assembly or binning.

[1] M.Kolmogorov et Al. metaFlye: scalable long-read metagenome assembly using repeat graphs, 2020
[2] X. Feng et Al. Metagenome assembly of high-fidelity long reads with hifiasm-meta, 2022
[3] G. Benoit et Al. High-quality metagenome assembly from long accurate reads with metaMDBG, 2024