



Metagenomic assembly of complex ecosystems with highly accurate long-reads

Nicolas Maurice, Claire Lemaitre, Clémence Frioux, Riccardo Vicedomini

► To cite this version:

Nicolas Maurice, Claire Lemaitre, Clémence Frioux, Riccardo Vicedomini. Metagenomic assembly of complex ecosystems with highly accurate long-reads. Journées 2024 du PEPR Agroécologie et Numérique, Jan 2024, Rennes, France. pp.1-1, 2024. hal-04425626

HAL Id: hal-04425626

<https://inria.hal.science/hal-04425626>

Submitted on 30 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

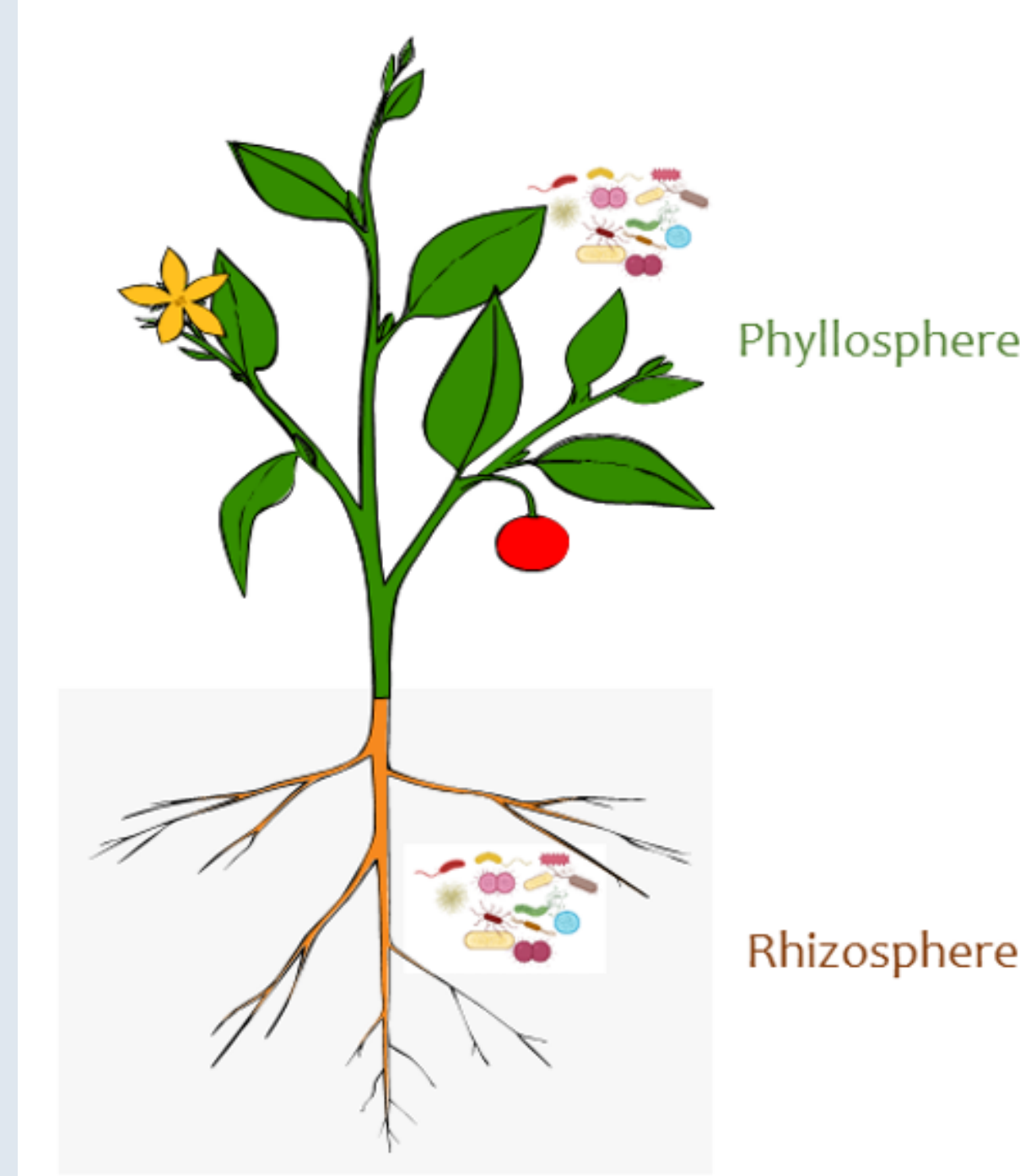
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Metagenomic assembly of complex ecosystems with highly accurate long-reads

Context

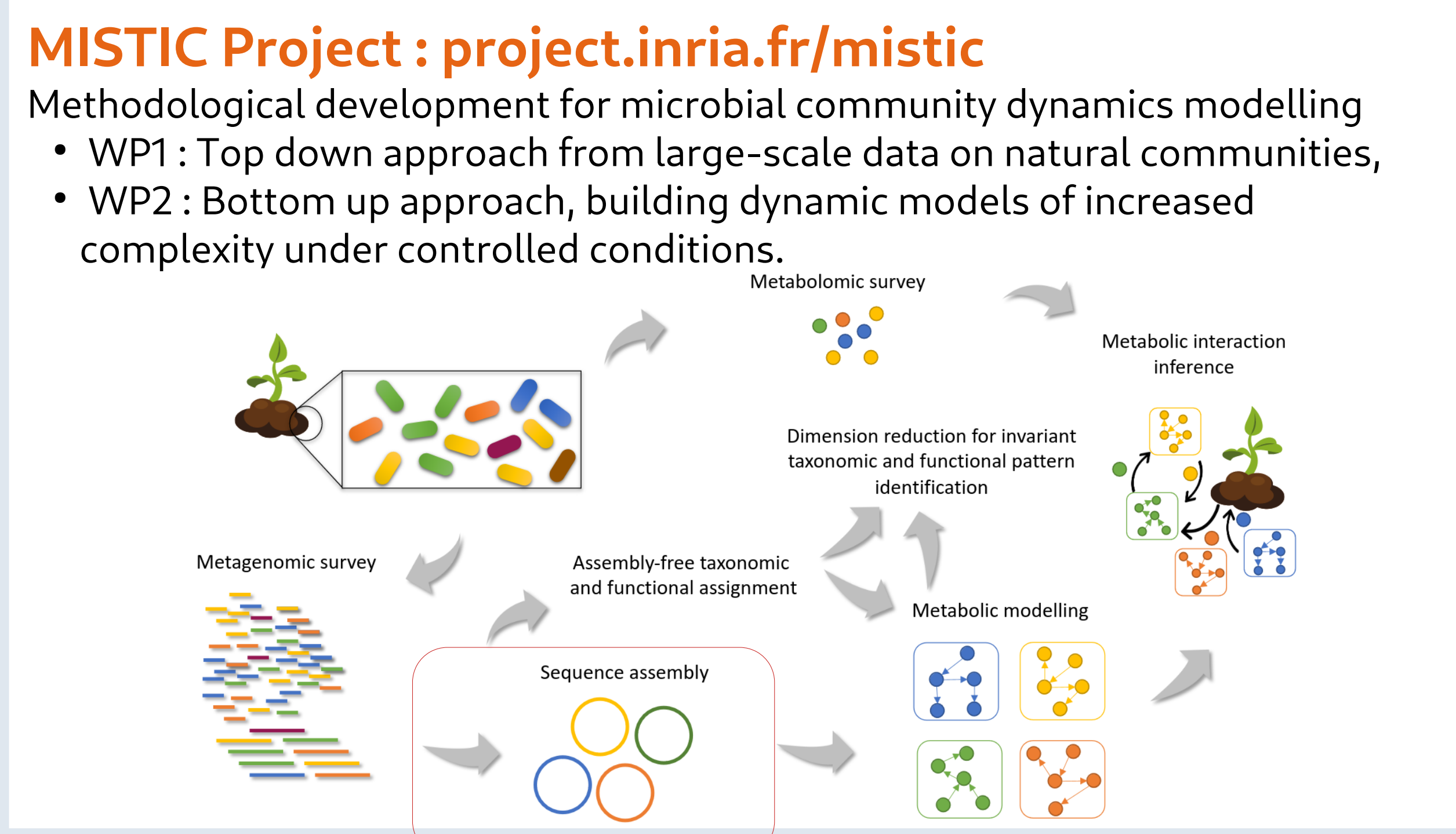


Plant microbiota

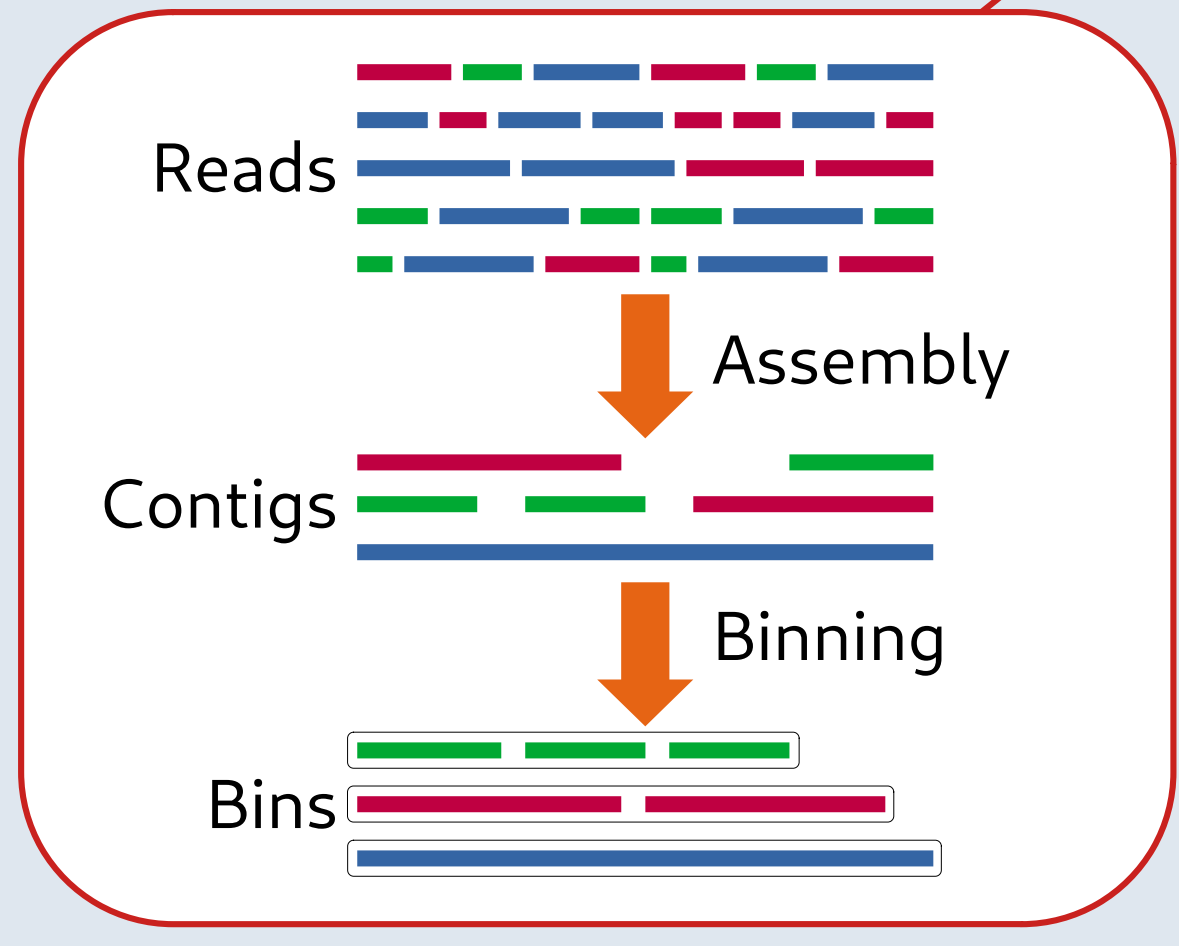
Plants can recruit micro-organisms from the millions (10^3 - 10^8 /g) found in soil, to regulate key functions such as :

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

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



Motivations



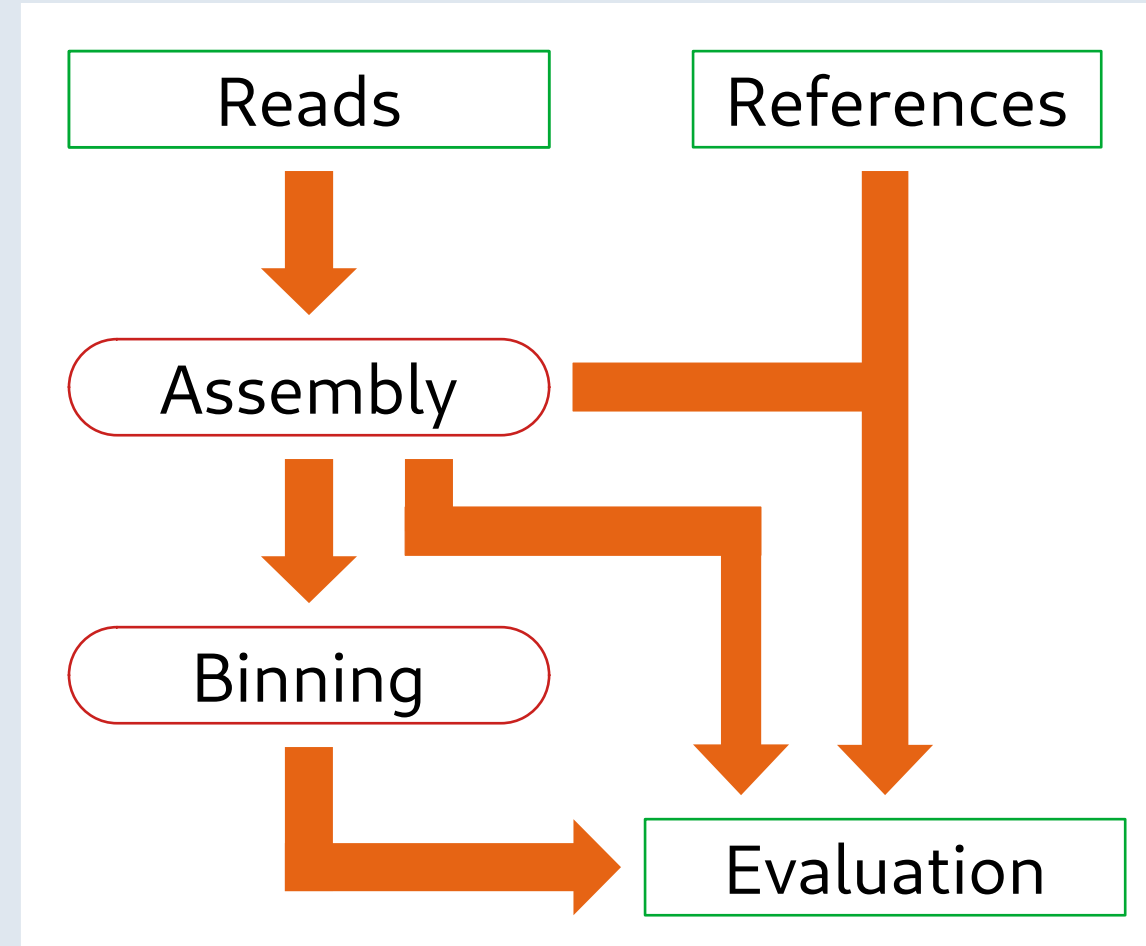
Emerging sequencing technologies

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

- Metaflye¹,
- Hifiasm-meta²,
- MetaMDBG³.

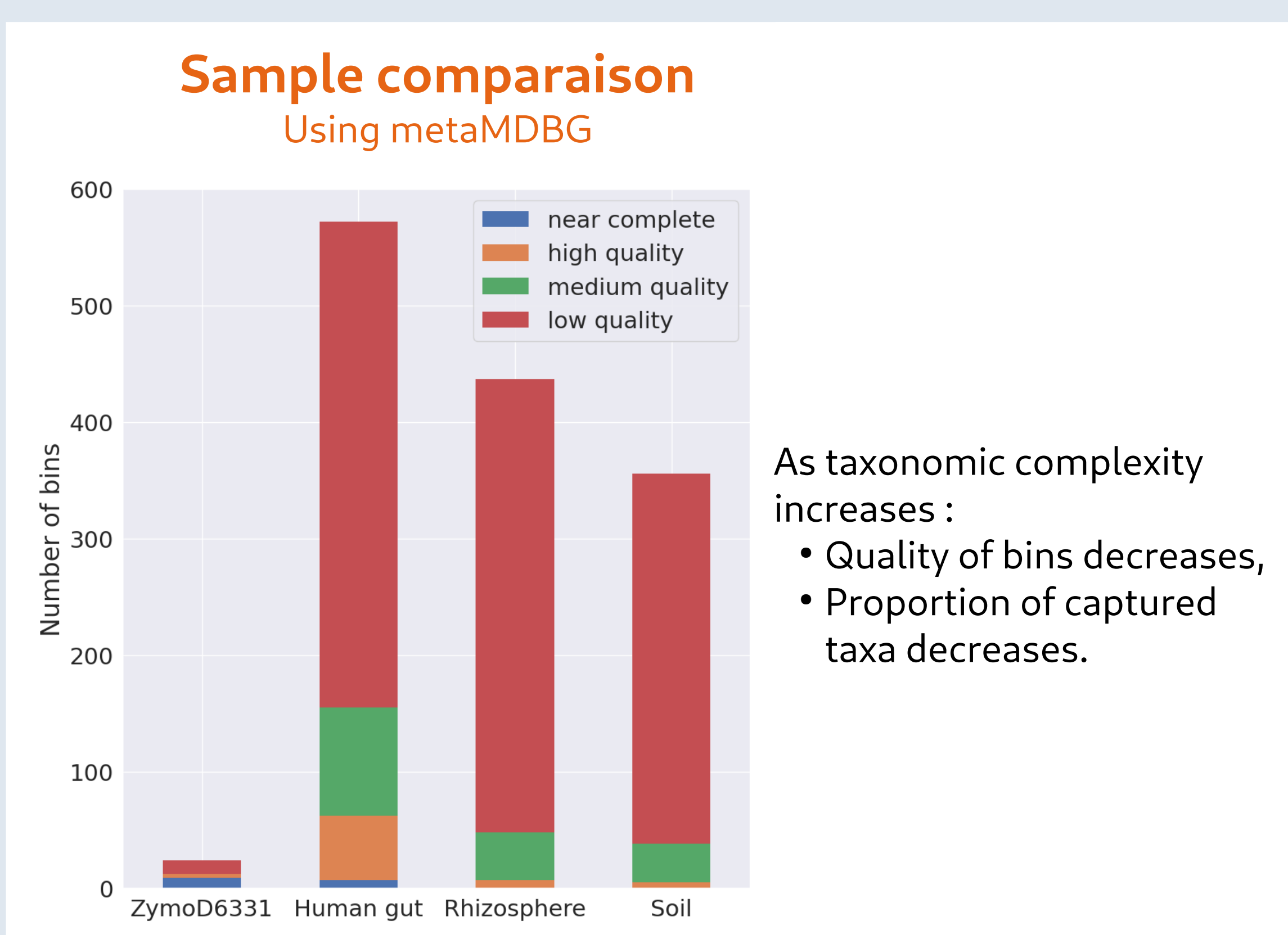
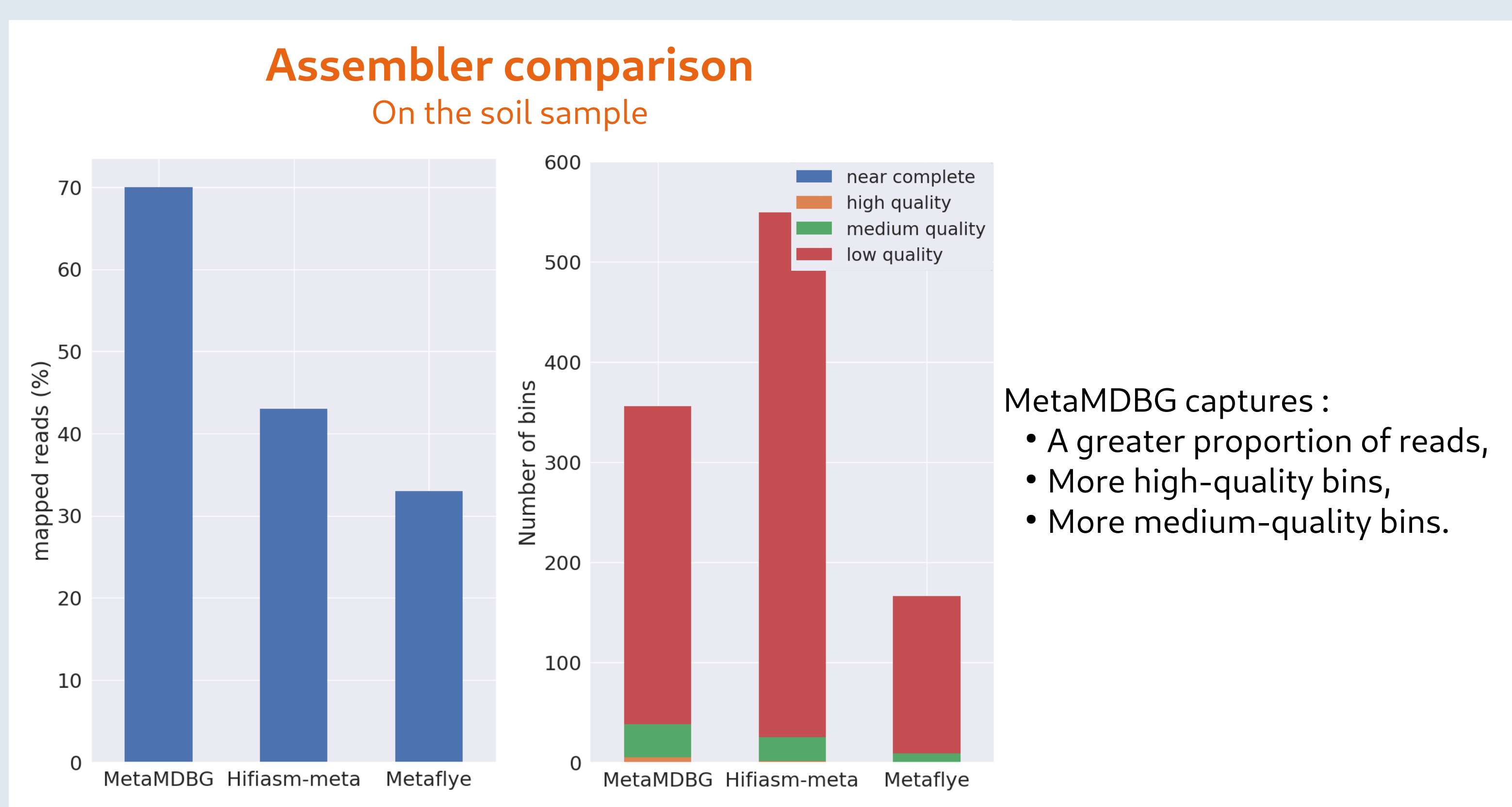
Datasets

	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 available : 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



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