Enhancing Microbial Genome Reconstruction in Complex Environments by combining Short-and Long-read Sequencing

Carole Belliardo, Arthur Péré, Alain Franc, Jean-Marc Frigerio, Nicolas Maurice, Clémence Frioux, Claire Lemaitre, Samuel Mondy, Marc Bailly-Bechet, Pierre Abad, et al.

Enhancing Microbial Genome Reconstruction in Complex Environments by combining Short- and Long-read Sequencing

Carole Belliardo[1], Arthur Péré[2], Alain Franc[3], Jean-Marc Frigerio[3], Nicolas Maurice[4], Clémence Frioux[3], Claire Lemaitre[4], Samuel Mondy[5], Marc Bailly-Bechet[2], Pierre Abad[2], David J Sherman[3], Etienne GJ Danchin[2]

[1] Université Côte d'Azur, MSI, France,[2] INRAE, Université Côte d'Azur, Institut Sophia Agrobiotech, France; [3] Inria/INRAE Pleiade, Bordeaux, France;

[4] Inria/IRISA Genscale, Rennes, France; [5] INRAE, Institut AGRO Dijon, Université de Bourgogne, DIJON

carole.belliardo@univ-cotedazur.fr; etienne.danchin@intea.fr

CONTEXT

Soil is one of the most diverse microbial ecosystems [1,2], yet much of this **"dark matter"** remains unexplored. Advances in sequencing technologies have improved insights, but challenges persist in fully characterising soil microbial diversity [3].

We used **PacBio HiFi long-reads (LR)** and **Illumina short-reads (SR)** whole-genome sequencing (WGS) to reconstruct **metagenome-assembled genomes (MAGs)** from a soil sample (Fig.1). Metabarcoding analyses completed this study (Fig. 2) by assessing microbial diversity [4] and determining which taxa WGS effectively captured.

Material and Methods

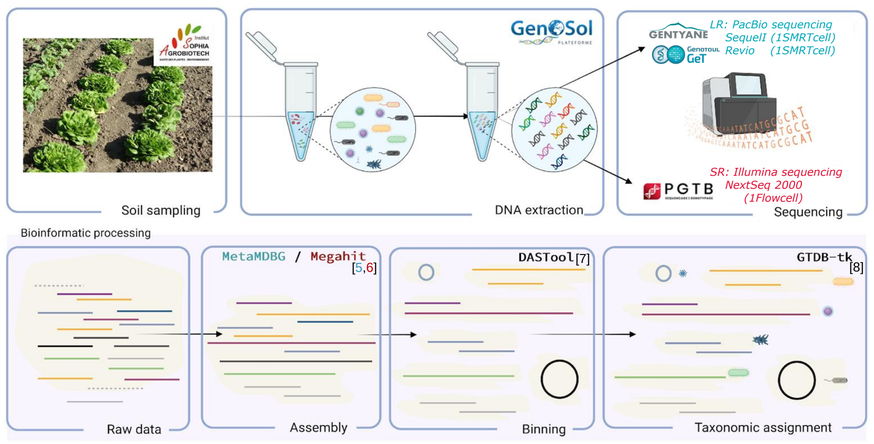

Figure 1 | Data Production and Processing Workflow Overview

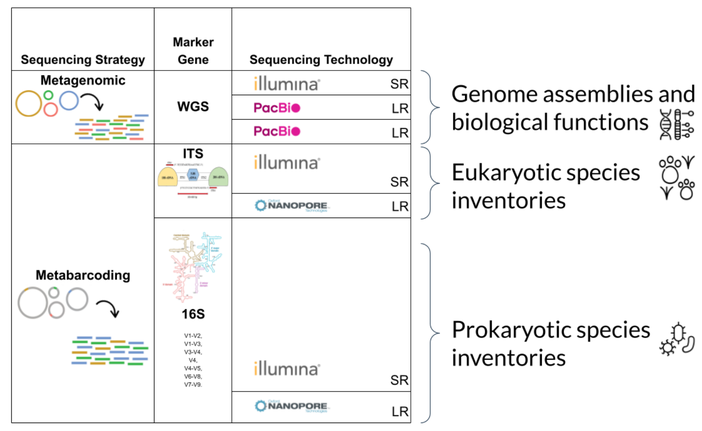

Figure 2 | Summary of generated datasets

HiFi LR Sequencing Delivers the Most Contiguous and Comprehensive Metagenomes

Long-read sequencing and assembly produce significantly longer contigs than short-read methods (Fig. 3). This improvement allows for a more in-depth investigation of genomes and microbial functions, revealing new insights into the complexity of microbial ecosystems.

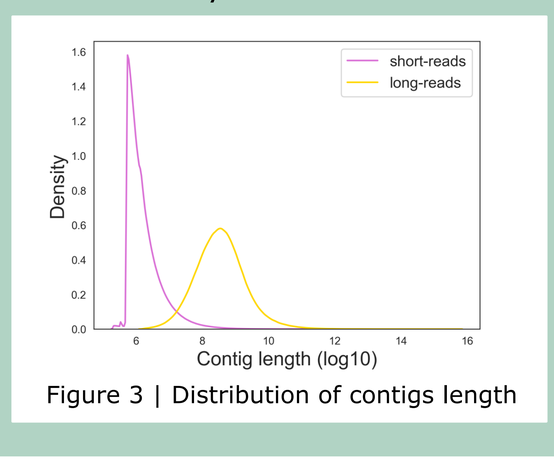

Figure 3 | Distribution of contigs length

Table 1 | Summary of raw and assembled datasets

Data	Metrics	Illumina NextSeq 2000	PacBio Sequel +Revio
Reads	Number	1.53B	11.12M
	Total Bases (Gb)	426.8	81.23
	Avg Length (bp)	151	7,252
	Max. Length (bp)		40,482
	N50 (bp)		7,881
Contigs	Contigs Number	31.07M	1.55M
	Total Bases (Gb)	19.14	11.59
	N50 (bp)	620	9,553

Despite Illumina yielding nearly twice the amount of PacBio data, assemblies are highly fragmented resulting in truncated genomic content (e.g. N50, Table 1). In contrast, PacBio reads allowed the production of a **more contiguous and complete genomic representation** of soil microbes.

In this study, we applied *de novo* WGS of DNA from an uncharacterized soil sample without *a priori* knowledge of genomic content. Metabarcoding being the gold standard for profiling microbial communities [4], we sequenced multiple metabarcodes to evaluate the taxonomic composition. All metabarcoding assays resulted in consistent inventories. Then, compared to metabarcoding data (e.g., 16S rRNA and ITS), our analysis shows that **PacBio sequencing offers the most robust and accurate WGS dataset** of the complex soil microbial genomic landscape (Fig 4 A, B).

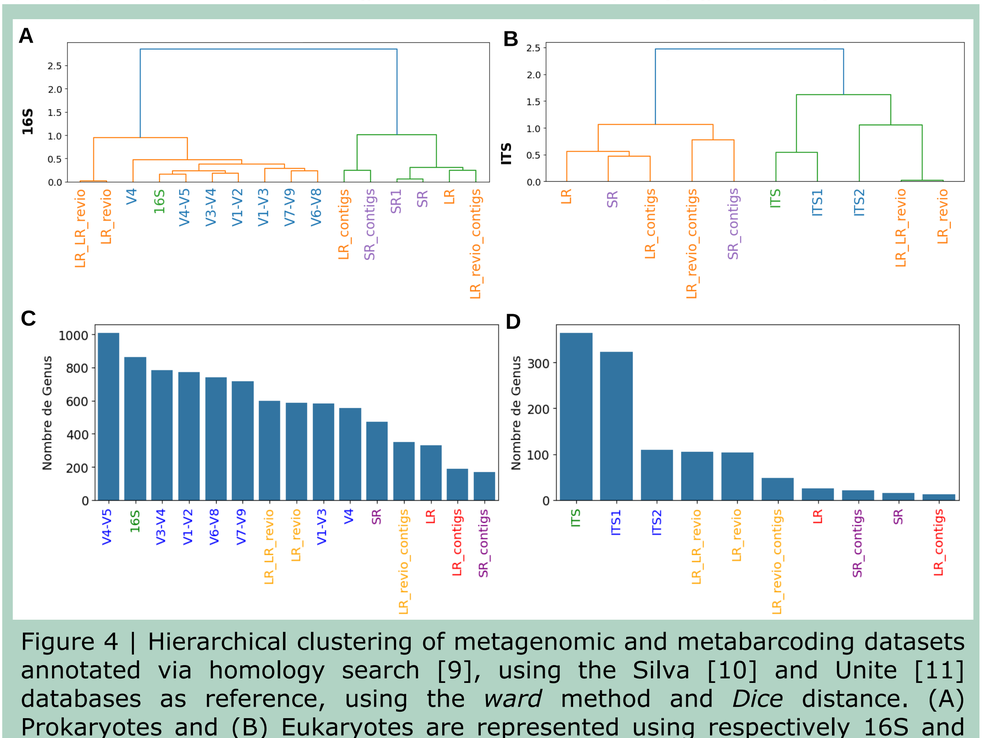

Figure 4 | Hierarchical clustering of metagenomic and metabarcoding datasets annotated via homology search [9], using the Silva [10] and Unite [11] databases as reference, using the *ward* method and *Dice* distance. (A) Prokaryotes and (B) Eukaryotes are represented using respectively 16S and ITS diversity. Then, the specific diversity was assessed on (C)16S RNA and (D) ITS for these Metabarcoding and Metagenomic taxonomic annotations.

Integrating SR and LR Enhances Binning Accuracy

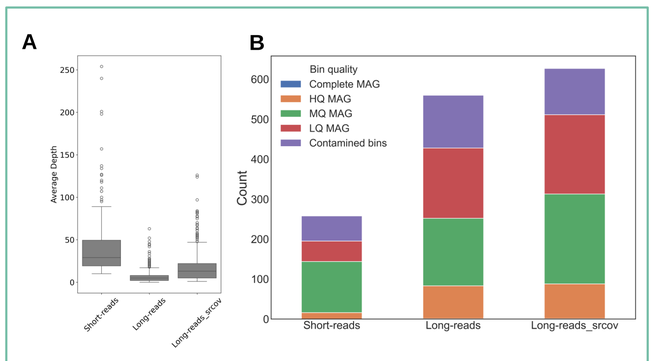

Figure 6 | (A) The distribution of contig depth is shown for self-mapping of reads and Illumina SR mapping. (B) Classification of bins by quality based on completeness (cn) and contamination (ct) scores: **Complete** (cn = 100, ct = 0), High Quality (**HQ**: cn ≥ 90, ct < 5), Medium Quality (**MQ**: cn ≤ 50, ct < 10), **Low Quality** (LQ: cn < 50, ct < 10), and **Contaminated** (ct > 10) [11]. (B) The distribution of contig depth is shown for self-mapping of reads and SR mapping (SR mapping dataset ending by 'srcov').

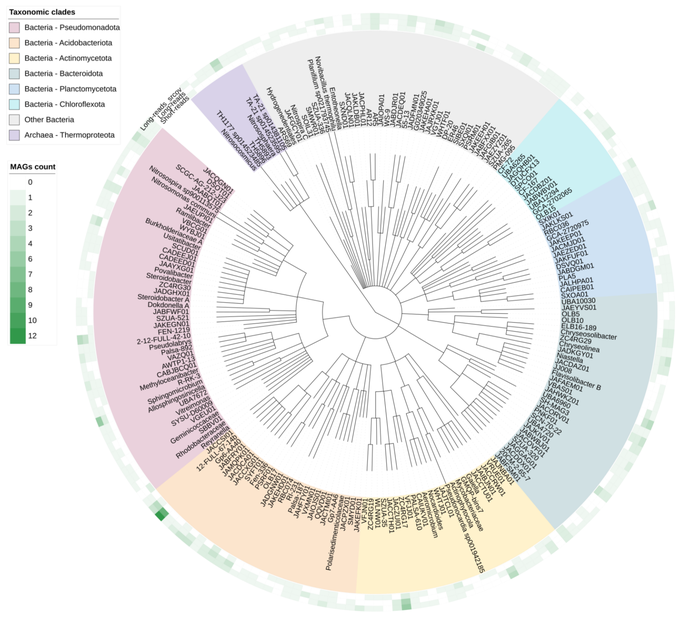

Figure 7 | Phylogenetic tree of taxonomic annotations for MAGs, generated using GTDB-tk [8]. The occurrence of each taxon across datasets is visualized as a heatmap.

Metabarcoding analysis suggests our sample contains ca. 1000 diverse prokaryotic genera and 384 eukaryotic genera (Fig. 4C, D). Although only around 100 of these genera are present across all datasets, **the most significant overlap between WGS and metabarcoding data is observed in the LR** dataset (Fig. 5).

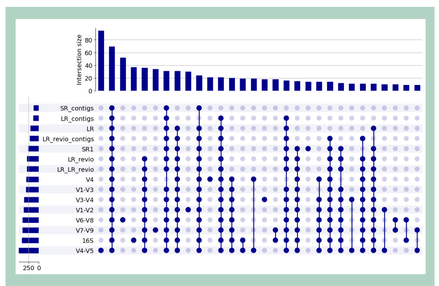

Figure 5 | Comparison of shared taxa originating from 16S taxonomic annotations.

Binning aims to cluster contigs from the same genome. Bins with at least 50% completeness are considered MAGs, taxonomically assigned using the accurate GTDB-tk method [8].

This processing relies on two signals: molecular composition and species abundance, calculated from contig depth via mapping reads. LR self-mapping produces lower depth values with limited variability, and SR mapping significantly enhances LR contig depth (Fig. 6A). Improving abundance signal enables better discrimination of biological origins (Fig. 6B). SR mapping on Sequel+Revio LR contigs yields the highest number of bins and MAGs, facilitating the discovery of novel taxa such as genus *SXQI01* specific to this dataset. It also improves resolution by separating contigs from closely related species, as observed with bins associated with genus 12-FULL-67-14b (Fig.7).

Conclusion and Perspectives

- **Enhanced MAG Reconstruction with PacBio HiFi Long Reads**: HiFi-LR significantly improve the reconstruction of MAGs, producing genomes with higher completeness and contiguity.
- **Increased coverage information leads to improved binning accuracy**: Integrating coverage data from long and short-read technologies substantially enhances binning performance, improving the resolution of microbial genomes.
- **Comprehensive Soil Microbial Diversity Representation**: long-reads provide the best representation of soil diversity for WGS; however, the number of reconstructed MAGs remains below the true species diversity in complex environments like soil assessed by metabarcoding inventories. Increasing sequencing depth through additional SMRT cells could enable the reconstruction of genomes from less abundant species.

MISTIC project

[1] Thompson, L. R., et al., 2017, Nature.
[2] Hug, L. A., et al., 2016, Nature.
[3] Fierer N. Nat. Rev. Microbiol. 2017.
[4] Stevens, B.M. et al., 2023, Scientific Report.
[5] Benoit G. et al.,2024, Nature.
[6] Dinghua Li et al., 2015, Bioinformatics.
[7] Christian M.K. et al., 2018, Nat Microbiology.
[8] Chaumeil P-A et al., 2019, Bioinformatics.
[9] Steinegger M. Et al.,2017, Nat. Biotech.
[10] Quast C, et al., 2013, Nucl. Acids Res.
[11] Abarenkov K et al., 2023, Nucl. Acids Res.
[12] Bowers, R., et al., 2017, Nat Biotechnol.

We thank the Sophia Agrobiotech bioinformatics platform, Genosol, Gentyane and PGTB platforms for their contributions. This work was funded by ANR and France2030 under the MISTIC project (*ANR-22-PEAE-0011*) and supported by Université Côte d'Azur through the Maison de la Modélisation, Simulation et Interaction contribution (*ANR-15-IDEX-01*).