Exploring and quantifying the soil genetic diversity captured by long and short-read shotgun metagenomic sequencing

Carole Belliardo, Samuel Mondy, Arthur Pere, Claire Lemaitre, Riccardo Vicedomini, Clémence Frioux, David James Sherman, Pierre Abad, Marc Bailly-Bechet, Etienne Danchin

▶ **To cite this version:**

Carole Belliardo, Samuel Mondy, Arthur Pere, Claire Lemaitre, Riccardo Vicedomini, et al.. Exploring and quantifying the soil genetic diversity captured by long and short-read shotgun metagenomic sequencing. Journées 2024 du programme Agroécologie et Numérique, Jan 2024, Rennes, France. , pp.1-1, 2024. hal-04423917

HAL Id: hal-04423917
https://hal.science/hal-04423917

Submitted on 29 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.

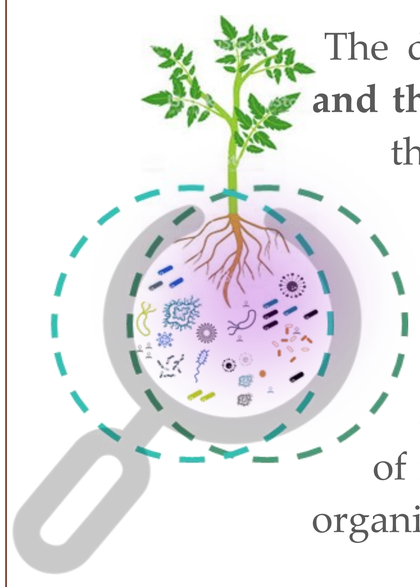
Distributed under a Creative Commons Attribution 4.0 International License

Exploring and quantifying the soil genetic diversity captured by long and short-read shotgun metagenomic sequencing

Carole Belliardo[1], Samuel Mondy[2], Arthur Pere[1], Claire Lemaitre[3], Riccardo Vicedomini[3], Clémence Frioux[4], David James Sherman[4], Pierre Abad[1], Marc Bailly-Bechet[1] and Etienne Danchin[1]

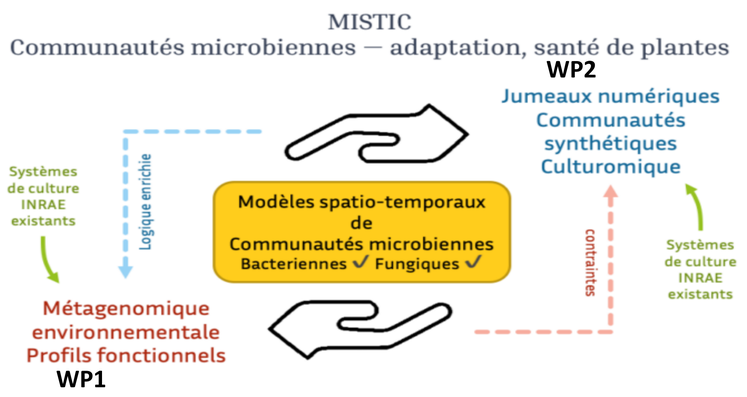
1 Institut Sophia Agrobiotech, Université Côte d'Azur, INRAE, CNRS, Sophia Antipolis, France
2 INRAE, Institut AGRO Dijon, Université de Bourgogne, 21065 DIJON, France
3 Univ Rennes, Inria, CNRS, IRISA, 35042, Rennes, France
4 Univ Bordeaux, Inria, 33405 Talence, Bordeaux, France

Introduction



The development of **agroecological practices** for **sustainable agriculture and the preservation of natural flora** requires a thorough understanding of these complex systems, including their **microbiomes** that remain the **darkest issue** [1]. The microbial community wield considerable influence over plants, exerting both beneficial and deleterious effects, while also possible **contributing to the plants' adaptive capacities** in the face of various biotic and abiotic factors (i.e. water availability and temperature fluctuations) [2]. To unlock these intricate networks of interactions, uncovering the **biological functions** supported by these organisms through **genome reconstruction** becomes indispensable.

➡ *Shotgun metagenomics sequencing allowing to bulk sequence environmental microbiomes could decipher biological functions of uncultured microorganisms*



Short-read (SR) shotgun metagenomics

- ✓ Provided interesting **insights into microbiome gene diversity**
- ✗ Not delivering comprehensive microbial genome reconstructions

Metagenome-assembled genomes (MAGs) from SR
= highly fragmented assemblies and incomplete gene sets with over 90% contigs < 1kB, and thus unusable for gene prediction [2].

Highly accurate long-read HiFi Soil Metagenomes

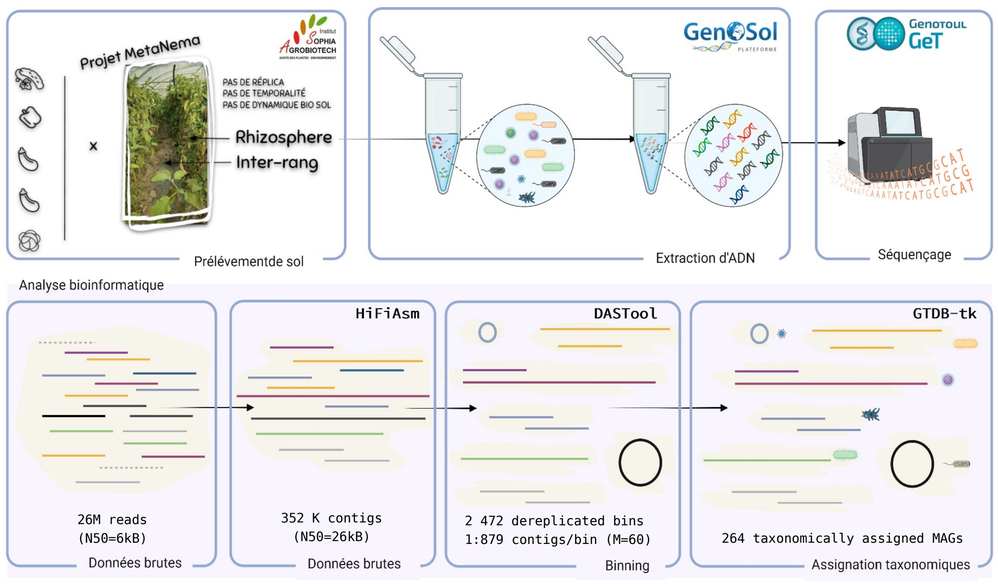


Fig. 1 Experimental design and bioinformatic analysis of the HiFi long-read metagenomic study

Hifi LR sequencing improve the Metagenome-assembled genomes reconstrcution

- We previously produced HiFi long-reads (N50: 6kb) from tunnel culture soil metagenomes (Fig. 1).
- Our HiFi LR surpassing the contig length of publicly available short-read metagenomes (N50: 1kb) [2].
- Dozens of MAGs were obtained after assembly, binning and dereplication (MAG was defined as bin with completeness >50% & nb. contigs <100), encompassing bacterial, archaeal, and viral genomes from terrestrial environments (Fig. 2).
- Our long-read metagenomes suggest a strong endemism of species within our local environment and across various plots. Only 6 MAGs have been assigned at the species level, with 2 different species names. The only one of the two that has been isolated in the laboratory is *Flavobacterium soyae t*hat has been isolated and described exclusively in the rhizosphere until today [3].

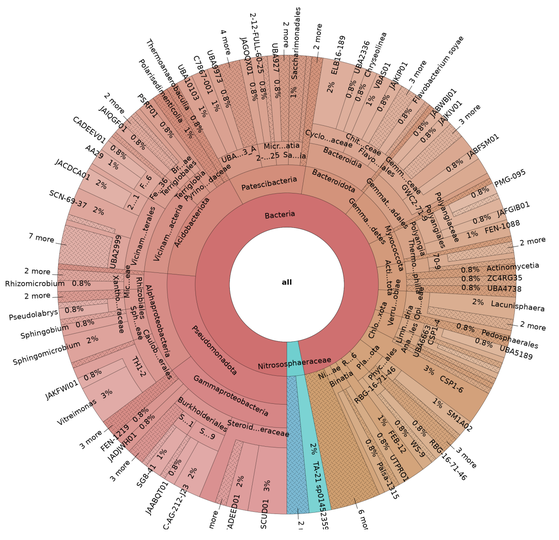


Fig. 2 Krona representation of HiFi LR MAGs taxonomic assignation using GTDB-tk

Hifi Long-reads versus short-reads Metagenomics: comparaison of diversity representation

One of same soil was submitted two times to ultra deep sequencing (NextSeq 2000 2x150pB) producing a total of 2x991 million short-reads.

- The assembly resulting in 31M of contigs (N50:620pB)
- Gene predictions on SR generated mainly truncated proteins, significantly shorted than those from LR.

SR sequencing seems to confirme a high endemism of soil microbial species in agricultural plots.

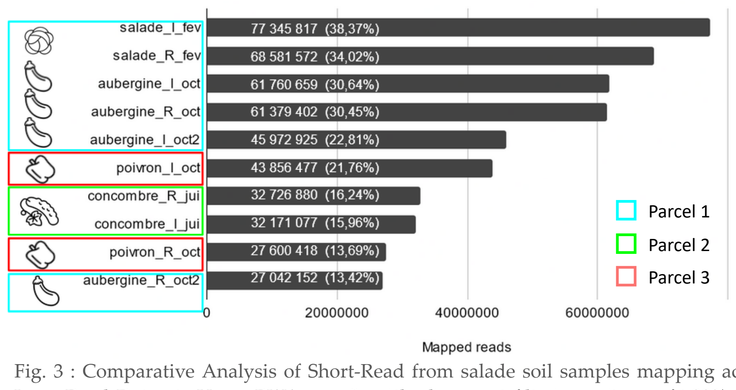


Fig. 3 : Comparative Analysis of Short-Read from salade soil samples mapping across 10 Long-Read Datasets Using BWA-mem2, with alignment filtering criteria of ≥90% identity and alignment length >130 bp.

Classic taxonomic assignment methods [5], based on reference database, fail in classifying soil metagenomic data for both LR and SR (> 70% data unclassified).

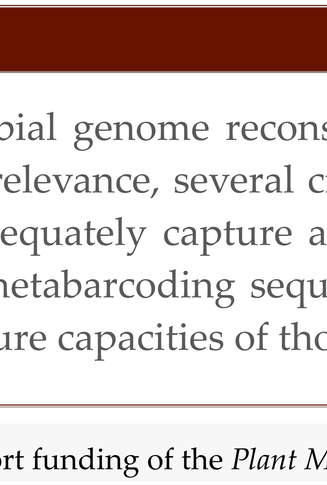


Fig. 4 : Workflow for virtual metabarcoding analysis, from the identification of ribosomal DNA in both SR and LR datasets using MMseq2 toolkit [6] and SILVA database as reference to the OTUs comparison though a ven diagram. Subsequently, barcodes are clustered at a 97% identity threshold using the cd-hit software [7]. Finally, OTU comparison is performed through a Venn diagram.

No sequencing technology seems to have successfully captured all of the soil's biodiversity yet

To assessed the genetic diversity respectively captured by long-reads and short-reads, we applied subsampling and k -mer counting to each dataset.

- The total k-mer diversity is higher in SR than LR (Fig. 5)
- The diversity distance observed via k-mer between LR and SR are consistent with mapping and taxonomic analyses (Fig3, Fig5).
- For both LR and SR, no inflexion was observed on k-mer diversity suggesting we did not saturate sequencing.

Fig. 5 : The rarefaction k-mer diversity assessed on 10 subsampling using DSK software.

Perspectives

HiFi LR sequencing holds a significant potential in microbial genome reconstruction and gives access to their metabolic pathways. To ensure downstream analysis relevance, several critical considerations need to be addressed such as the sequencing depth required to adequately capture and reconstruct a meaningful representation of the real soil microbial biodiversity. The metabarcoding sequencing for soil samples with LR and SR available data could give an accurate idea of capture capacities of those sequencing technologies.

References

1. Fierer, N., 2017. Nat Rev Microbiol
2. Hoysted G. & al., 2018, Curr. Op. in Pl. Bio
3. Hui Z. & al., 2022 .
4. Belliardo, C., et al., 2022, Scientific Data
5. Wood. & al., *2019,* BMC Biology
6. Steinegger & al., *2018,* Nature com.
7. Limin F. & al., 2012. Bioinformatics

MISTIC

Acknowledgments Grateful appreciation to the *ANR* and *France2030* for support funding of the *Plant Microbiome and ITC (MISTIC)* project.