Strengths and limits of long read metabarcoding

Jean Mainguy, Adrien Castinel, Olivier Bouchez, Sylvie Combes, Carole Iampietro, Christine Gaspin, Denis Milan, Cécile Donnadieu, Claire Hoede, Géraldine Pascal

Strengths and limits of long read metabarcoding.

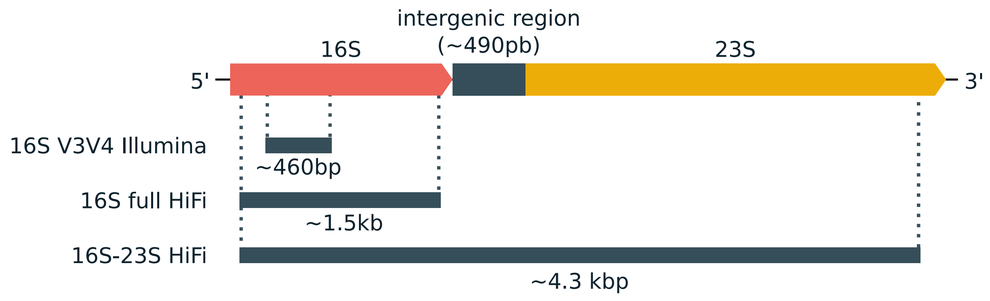
Jean MAINGUY[1 2], Adrien CASTINEL[3], Sylvie COMBES[4], Christine GASPIN[1 2], Denis MILAN[3 4], Cécile DONNADIEU[3], Carole IAMPIETRO[3], Olivier BOUCHEZ[3], Claire HOEDE[1 2] and Géraldine PASCAL[4]

1 Université Fédérale de Toulouse, INRAE, BioinfOmics, GenoToul Bioinformatics facility, 31326, Castanet-Tolosan, France
2 MIAT, PF Bioinfo GenoToul, Université de Toulouse, INRAE, Chemin de Borde Rouge, 31320 Castanet-Tolosan, France
3 GeT-PlaGe, US 1426, Genotoul, INRAE, 31320 Castanet-Tolosan, France
4 GenPhySE, Université de Toulouse, INRAE, INPT, ENVT, Chemin de Borde Rouge, 31320 Castanet Tolosan, France

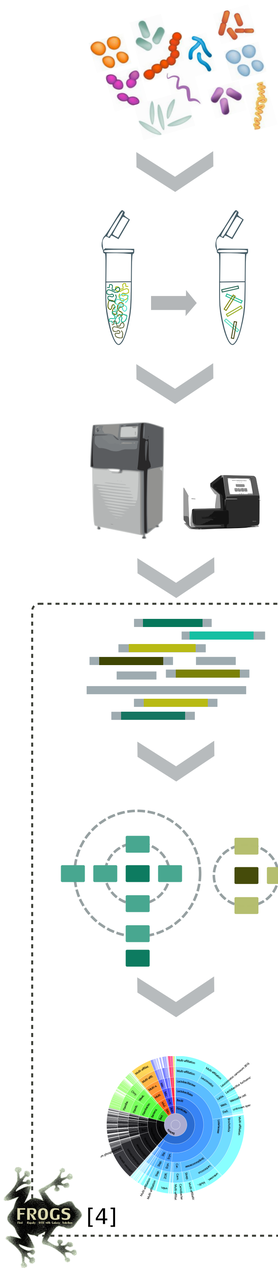
Contact: geraldine.pascal@inrae.fr

Introduction

Amplicon sequencing of the 16S rRNA gene fragments is commonly used to study microbial communities but often fails to assign microbes at the genus or species taxonomic level. **Long read** sequencing technologies can help to overcome this limitation by sequencing longer regions such as the full **16S rRNA gene** and the **16S-23S operon** region.
We conducted a comparative study to understand the limitations and strengths of PacBio HiFi long read sequencing technology for metabarcoding analyses.



Long and short reads metabarcoding method



Microbial community

DNA extraction and barcode amplification
In long read, DNA extraction needs to give long DNA fragments while maintaining a good species representativeness. We tested 5 different protocols on a mock Bacteria community and on fecal microbiota samples and empirically kept the ZymoBIOMICS DNA Miniprep kit as it showed less biased.

Amplicon sequencing
Long read metabarcoding can be sequenced by the Pacbio Sequel II using circular consensus sequencing mode to produce HiFi reads. Short reads are typically sequenced in 2x250pb using Illumina MiSeq.

Sequences preprocessing
Sequences are filtered by size and primers sequences are trimmed. In short read, paired-end reads need to be assemble to reconstitute the targeted sequence.

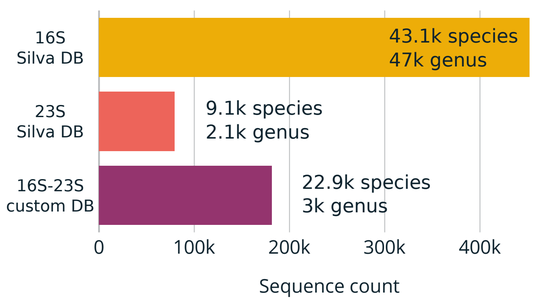
Denoising and clustering
Sequences are grouped by similarity creating OTUs, ASVs-like or ASVs. Amplification and sequencing errors are masqued. We tested several strategies to group long HiFi reads (dada2 [1], Swarm [2] and CD-HIT [3]). Dada2 and Swarm gave the best results.

Taxonomic affiliation and abundance table
Each group is affiliated to a taxon through a reference database. 16S V3-V4 and 16S full sequences was affiliated using 16S Silva database. The 16S-23S sequences were affiliated using a custom database (see box below). Sequenced that failed to be affiliated with this database were affiliated with the 16S and the 23S Silva databases.

References:
1. Callahan, Benjamin J., et al. "High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution." Nucleic acids research 47.18 (2019): e103-e103.
2. Mahé, Frédéric, et al. "Swarm: robust and fast clustering method for amplicon-based studies." PeerJ 2 (2014): e593.
3. Fu, Limin, et al. "CD-HIT: accelerated for clustering the next-generation sequencing data." Bioinformatics 28.23 (2012): 3150-3152.
4. Bernard, Maria, et al. "FROGS: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers." Briefings in Bioinformatics 22.6 (2021): bbab318.

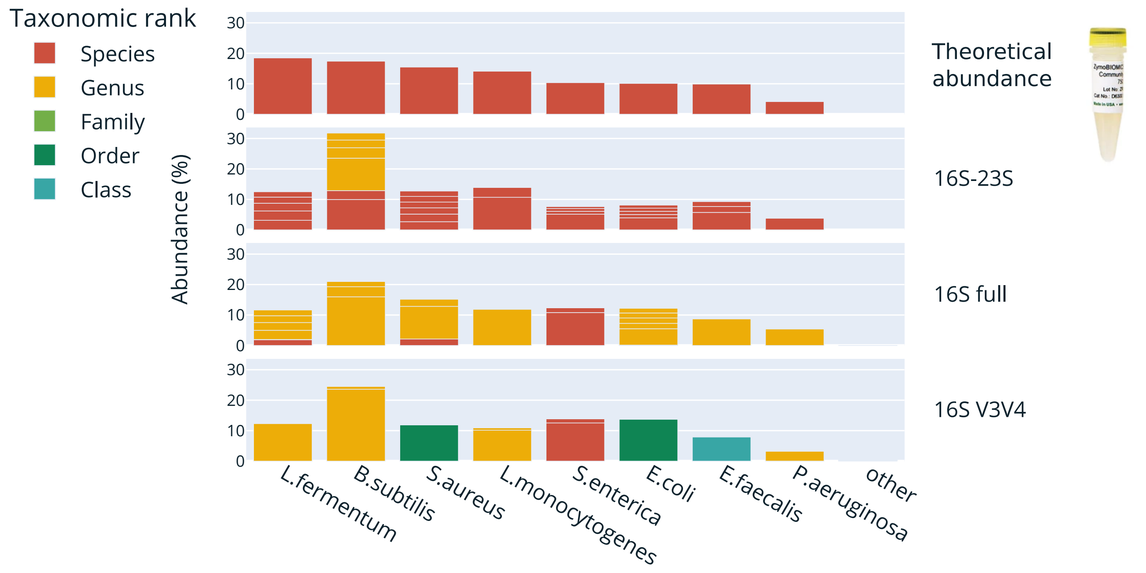
16S-23S custom reference database

To build our custom 16S-23S reference database, we first extracted the 16S-23S region from 25.2k Prokaryotic RefSeq genomes based on their annotations. A curation step was applied to remove eucaryotic rRNA sequences and flag suspicious sequences that migth be a contamination.



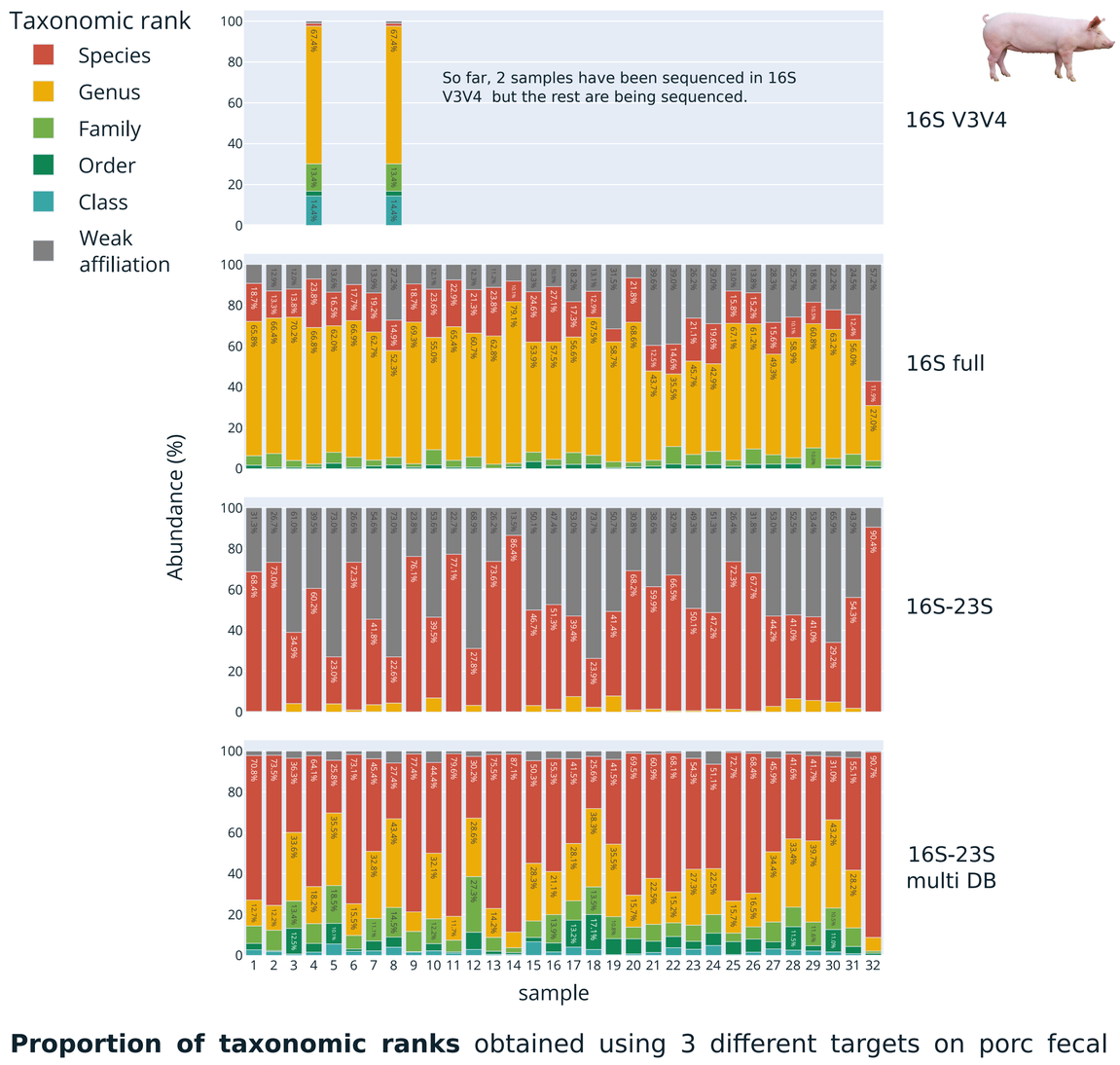
Number of sequences in our custom 16S-23S database compared to 16S and 23S Silva databases. The 16S Silva database is far more comprehensive than our 16S-23S database. But, a **multi DB** approach consisting of using the 3 databases allows to obtain a better affiliation of the biological data.

Mock community analysis



Abundance and affiliation obtained with the 16S-23S, the 16S full and the 16S V3V4 metabarcoding on a Bacteria mock comunity (Zymobiomics) containing 8 Bacteria species. **Longer the target region, the more precise the affiliations.** Each block in a bar represents a cluster.

Porc fecal microbiota analysis



Proportion of taxonomic ranks obtained using 3 different targets on porc fecal microbiota samples*. Weak affiliation are sequences that have no affiliation hit above 99% identity and 99% coverage. In **16S-23S multi DB**, unaffiliated sequences with the 16S-23S database have a **second round of affiliation** using the 16S and the 23S Silva databases. The multi DB approach allows to **greatly reduce** the proportion of **weak affiliations**. As for the mock community, **longer target regions increase the proportion of species affiliations.**

* We thank Philippe Pinton for providing the samples for analysis.

Conclusion

Long read sequencing in metabarcoding **improves the specificity** of sample characterization.

Long read sequencing remains **more expensive** and requires special **care in DNA extraction** to preserve species representativeness while giving long fragments.

In the case of 16S-23S, the use of **complementary databases** (Silva 16S and 23S databases) in addition to a full 16S-23S database is a complicating factor, but it helps to **increase the reliability** of the affiliations.

Projet cofinancé par le Fonds Européen de Développement Régional