Implementation of new methods for metagenomics analysis using PacBio HiFi sequencing

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

SeqOccIn

Sequencing Occitanie Innovation

Implementation of new methods for metagenomics analysis using PacBio HiFi sequencing

Adrien Castinel[1], Jean Mainguy[2], Sylvie Combes[3], Carole Iampietro[1], Christine Gaspin[2], Denis Milan[1,3], Cécile Donnadieu[1], Claire Hoede[2], Géraldine Pascal[3] and Olivier Bouchez[1]

[1] INRAE, GeT-PlaGe, Genotoul – INRAE – 31326 Castanet-Tolosan, France (doi: 10.15454/1.5572370921303193E12), France
[2] INRAE, UR 875 Unité de Mathématique et Informatique Appliquées, PF Bioinfo Genotoul – Institut National de la Recherche Agronomique : UR875 – 31320 Castanet Tolosan, (doi: 10.15454/1.5572369328961167E12), France
[3] GenPhySE, Université de Toulouse, INRAE, ENVT – INRAE – 31326 Castanet-Tolosan, France

The variable regions of 16S rRNA gene is widely used to profile microbial communities, however the limited read lengths (e.g 460pb for the V3-V4 region) combined with the complexity of microbial samples makes it difficult to accurately identify bacterial strains and their abundance. The arrival **PacBio Sequel II sequencer** allow us to generate longer reads named "**HiFi reads**" using the circular consensus sequencing (CCS) mode, with an accuracy similar to the one observed for short reads sequencers.

In the framework of this project, we **implemented two protocols** on the Pacbio Sequel II sequencer in order to evaluate the **taxonomic resolution** of longer **target regions**, such as, the **full-length 16S** (1.5kb) and **16S-23S genes** (operon 4.5kb) and compared them to 16S rRNA region.

DNA extraction

For metabarcoding study, an extraction method able to produce **DNA fragment up to 10Kb** while maintaining a good **species representativeness** is necessary. That is why we tested 5 different protocols. **Three methods out of five produced fragments of sufficient size and quality** (red box, Fig1)**.** We selected them to assess the representativeness of the bacterial communities (Fig2). Among the three kits tested, **the DNA Miniprep kit** showed less biases compared to the two others. We decided to select this extraction kit for the following experiments.

	Innuprep Stool DNA Kit	Mag-Bind Stool DNA Kit	DNA Miniprep Kit	Quick DNA fecal/soil Kit	QIAamp Powerfecal Pro Kit
DNA size (Kb)	5,5	2	5,5	5,2	7

Fig. 1: Five methods evaluated on a **microbial mock community** (ZymoBIOMICS D6300). Average size obtained on a Fragment Analyzer (Agilent).

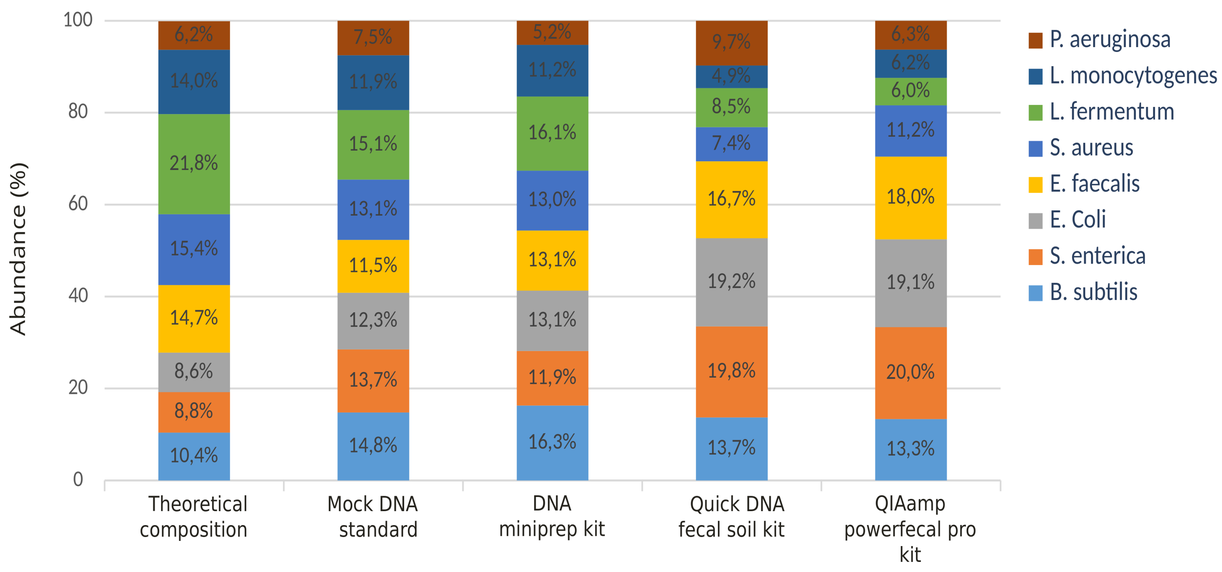

Fig. 2: The DNA Miniprep kit showed less biases compared to the two other kits. The 16S rRNA gene sequenced on Illumina MiSeq allowed us to evaluate the percentage of each species found in the microbial mock community and to compare it with the expected theoretical percentage.

Metabarcoding protocols

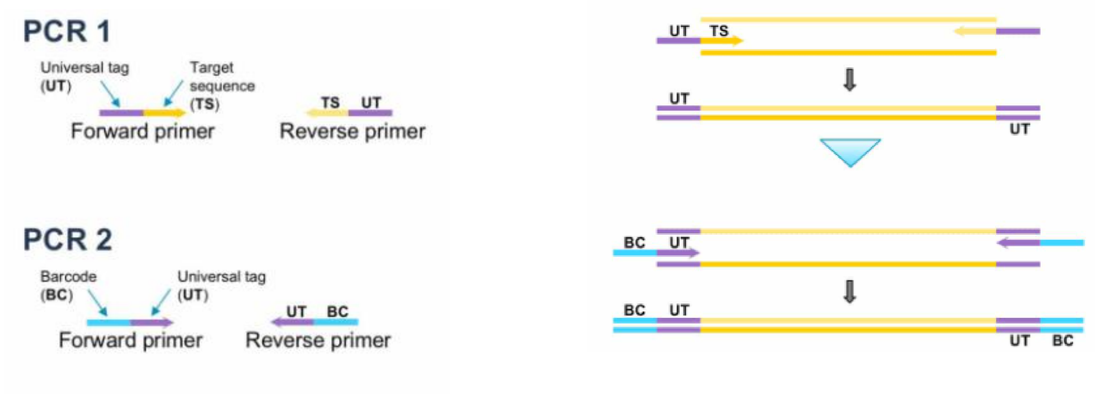

Fig. 3: PacBio Barcoded Universal Primers protocol (BUP).
PacBio barcodes are added to amplicon through **a 2 step PCR** method using Universal Sequence-tagged target-specific primers and Universal Sequence-tagged barcoded primers.

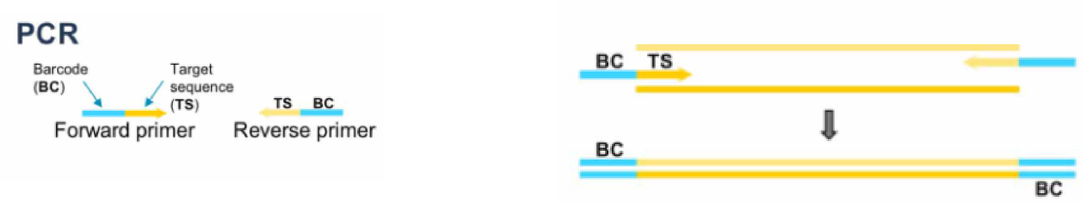

Fig. 4: PacBio Barcoded Target Specific Primer protocol (BTSP).
Target specific primers are tailed with PacBio barcodes to produce barcoded amplicon in a **single step PCR**.

https://www.pacb.com/smrt-science/smrt-sequencing/multiplexing/

PacBio protocols comparison: BUP vs BTSP

Mock DNA was extracted using DNA Miniprep kit. The full-length 16S region was amplified following BUP or BTSP protocol, while we only used BUP protocol for 16-23S region amplification. The three libraries were sequenced on PacBio with using HiFi sequencing.
Our work shows that the BTSP protocol gives better results than the BUP protocol. The different clustering methods did not improved significantly the results.

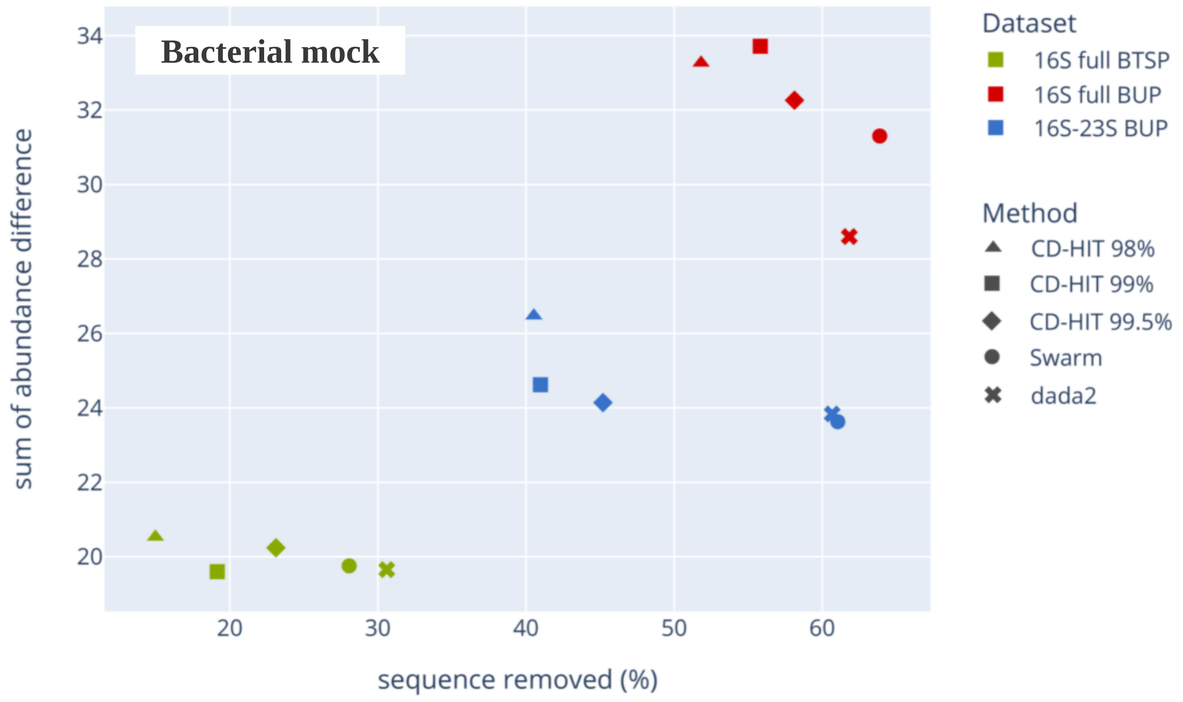

Fig. 5: Three datasets analyzed using FROGS* pipeline. In order to obtain the best analysis with FROGS, different clustering methods were applied on each dataset (CD-HIT, swarm and DADA2) and analyzed according to 2 metrics: the percentage of sequences removed during the analysis (x axis) and the sum of the difference between the observed (sum of all OTUs corresponding to the species) and expected abundances (y axis).

* F.Escudié. *FROGS: Find, Rapidly, OTUs with Galaxy Solution. Bioinformatics. 2018 Apr 15;34(8):1287-1294. doi: 10.1093/bioinformatics/btx791. PMID: 29228191.*

Taxonomic resolution

Based on those results, we were able to compare the taxonomic resolution between the targeted regions full-length 16S (PacBio sequencing), the 16S-23S operon (PacBio sequencing) and the 16S rRNA V3-V4 (Illumina sequencing). We found that the **16S-23S** was much **more discriminating** than the **full-length 16S** allowing species affiliation for almost all species.

Fig. 6: Abundances obtained using 3 different targeted regions on a Microbial mock. Each block in a bar represents an OTU.

Conclusions and perspectives

For metabarcoding, we showed that the extraction method selected combined with BTSP protocol gave the best results for the full-length 16S sequencing. The BUP protocol made possible the 16S-23S sequencing and showed that it is a more discriminating region than the full-length 16S region. The next step would be to test the BTSP protocol on the 16S-23S region and implement the PacBio overhang adapters protocol as well, to make a full comparison between the protocols that are available for metabarcoding analysis.

In parallel, we developed the metagenomics shotgun protocol on the PacBio Sequel II sequencer. The first analysis on the mock revealed that HiFi reads produce high quality assemblies but imply a higher cost than short read sequencing.

All the results, presented here, have been obtained within the framework of the "Sequencing Occitanie Innovation" project, which aims to acquire expertise on the study of microbial complex environments to better characterize species identification and their functions.

More about the SeqOccIn program : https://get.genotoul.fr/seqoccin/

Three majors topics are treated in this project: genome variability analysis, epigenetic mark analysis and metagenome analysis. SeqOccIn projet is carried out by Get-PlaGe and Genotoul Bioinfo platforms and financed by FEDER funds (Programme Opérationnel FEDER-FSE_Midi-Pyrénées et Garonne 2014-2020). The project benefits from the contributions of INRAE research units GenPhySE, MIAT, GABI, GQE. 25 private partners are involved, among them, Lallemand, for the metagenomics part of the project.