The benefits of long read HiFi sequencing for metagenomic analysis

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

SeqOccIn

INRAE
la science pour la vie, l'humain, la terre

The benefits of long read HiFi sequencing for metagenomic analysis

Adrien Castinel

Genotoul GeT

http://get.genotoul.fr
@Get_Genotoul

http://bioinfo.genotoul.fr
@BioinfoGenotoul

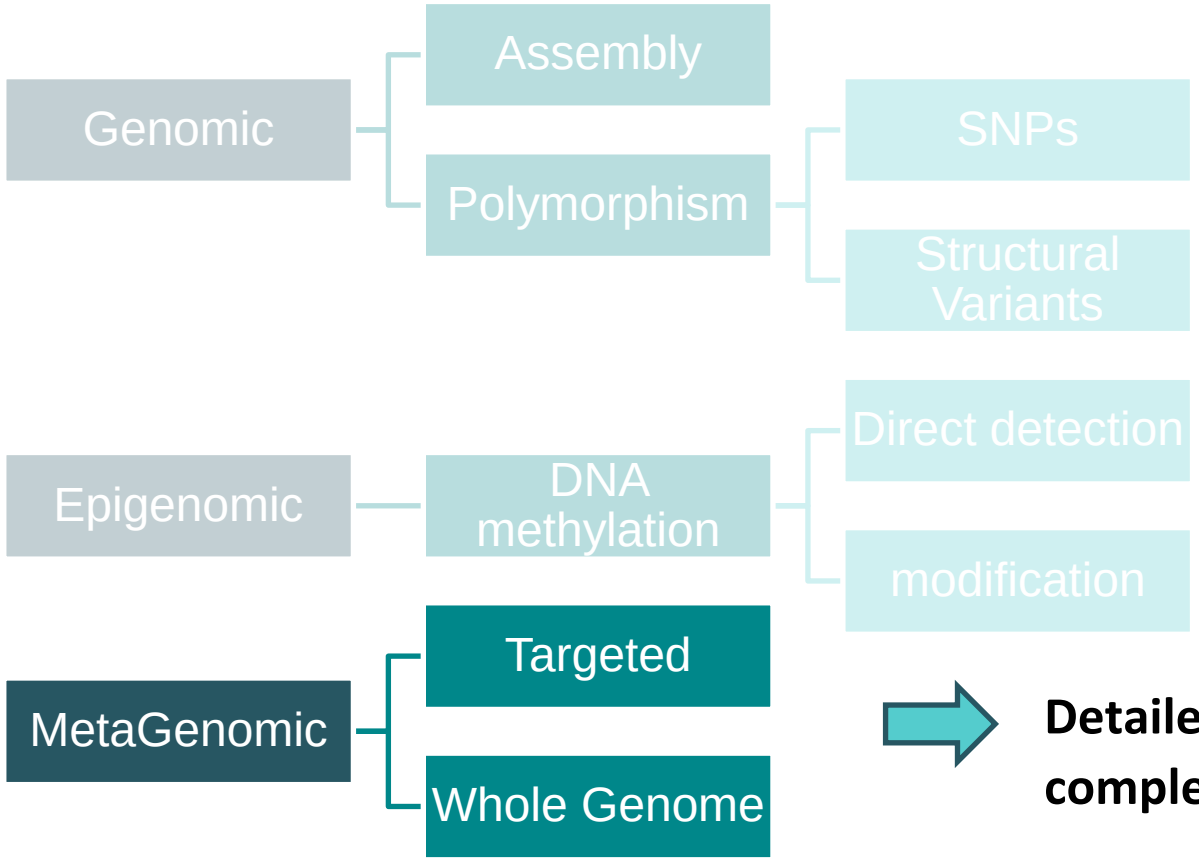
Genotoul Bioinfo

SeqOccIn

SeqOccIn

SeqOccIn Project: carried by GeT-PlaGe and Bioinfo Genotoul

Aim of the project:
Acquire expertise in the use of long read sequencing technology in 3 domains:



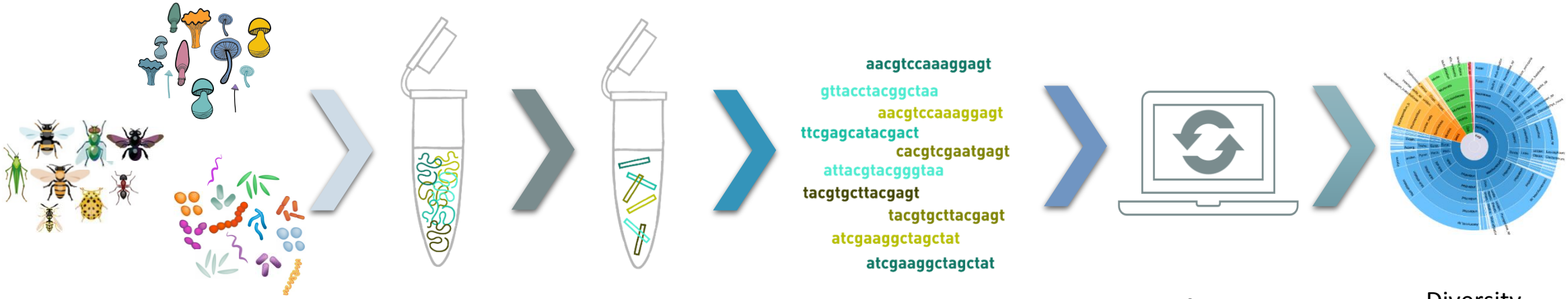
Detailed knowledge of communities in a complex environment

METABARCODING

SeqOccIn

What is metabarcoding?

- Some DNA fragments are highly conserved within a species and variable between species. These are the genetic markers or barcodes.
- Metabarcoding, by identifying barcodes through sequencing, allows blind identification of all species present in a sample at once.

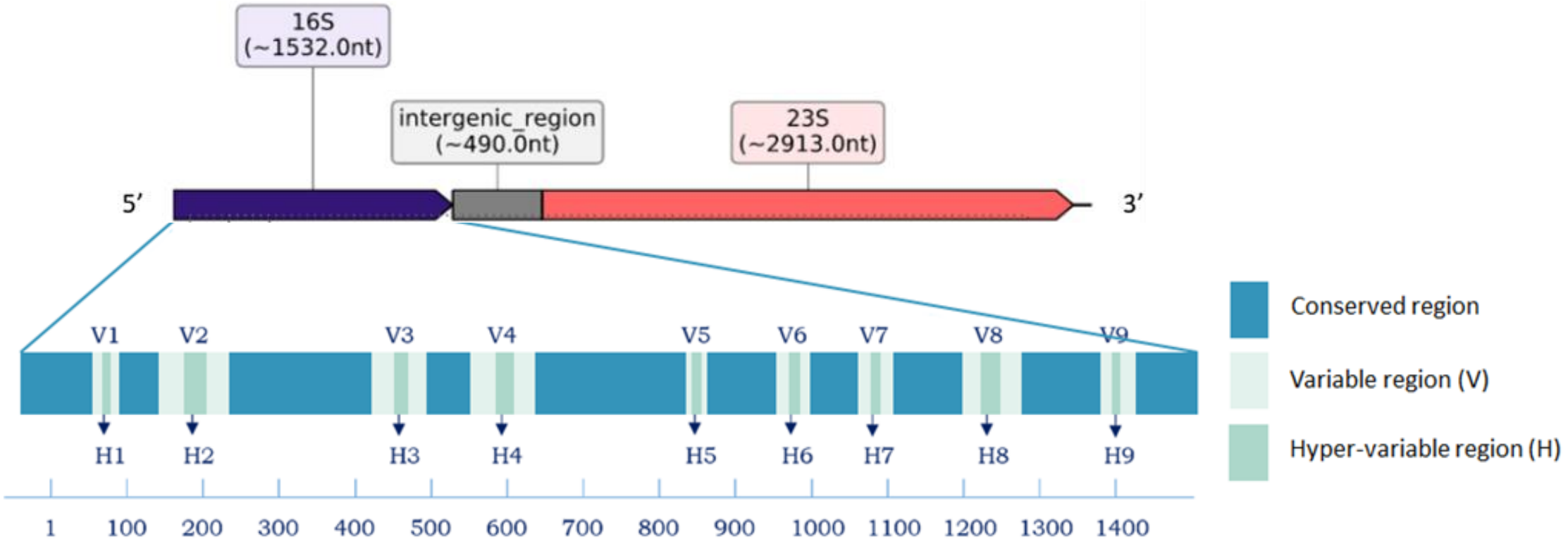


Community analysis | DNA extraction | Barcode amplification | Amplicon sequencing | Bioinformatics process | Diversity characterisation

Genotoul GeT Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité République Française
L'Europe s'engage l'Occitanie agit
Union Européenne
Région Occitanie
Projet cofinancé par le Fonds Européen de Développement Régional

16S: a bacterial marker

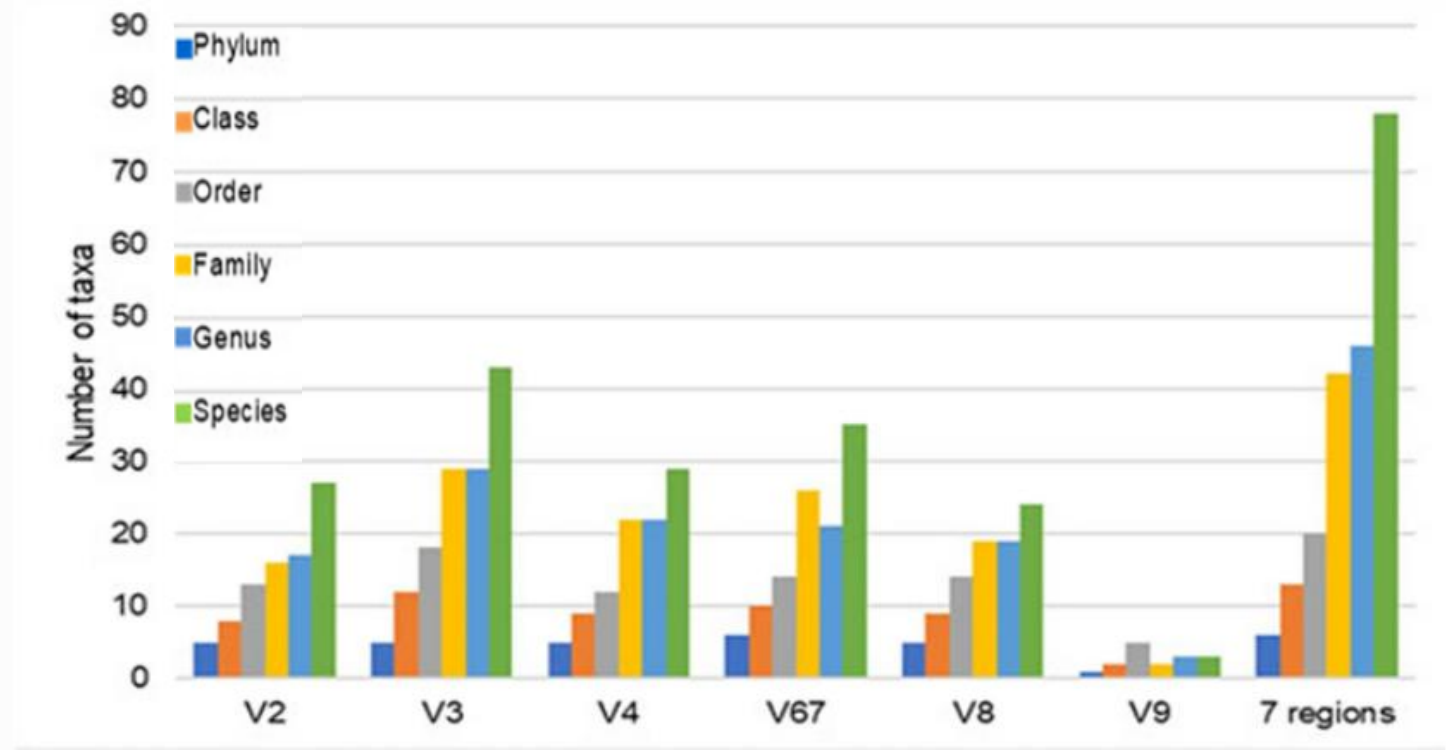
Ribosomal RNA: bacterial markers



- Gene encoding the 16S subunit of ribosomal RNA
- Highly conserved gene in all bacteria
- Highly conserved regions (for primers) interleaved with variable regions (bacteria identification)
- Choice of the regions depends on the community to be analyzed

Genotoul GeT Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité République Française
L'Europe s'engage L'Occitanie agit
Projet cofinancé par le Fonds Européen de Développement Régional

Comparison of different hypervariable regions of 16S rRNA



Number of taxa identified at each taxonomical using individual and concatenated hypervariable regions of the 16S rRNA gene.

- Analysis of a combination of sequences from V2 to V9 regions identified more taxa.

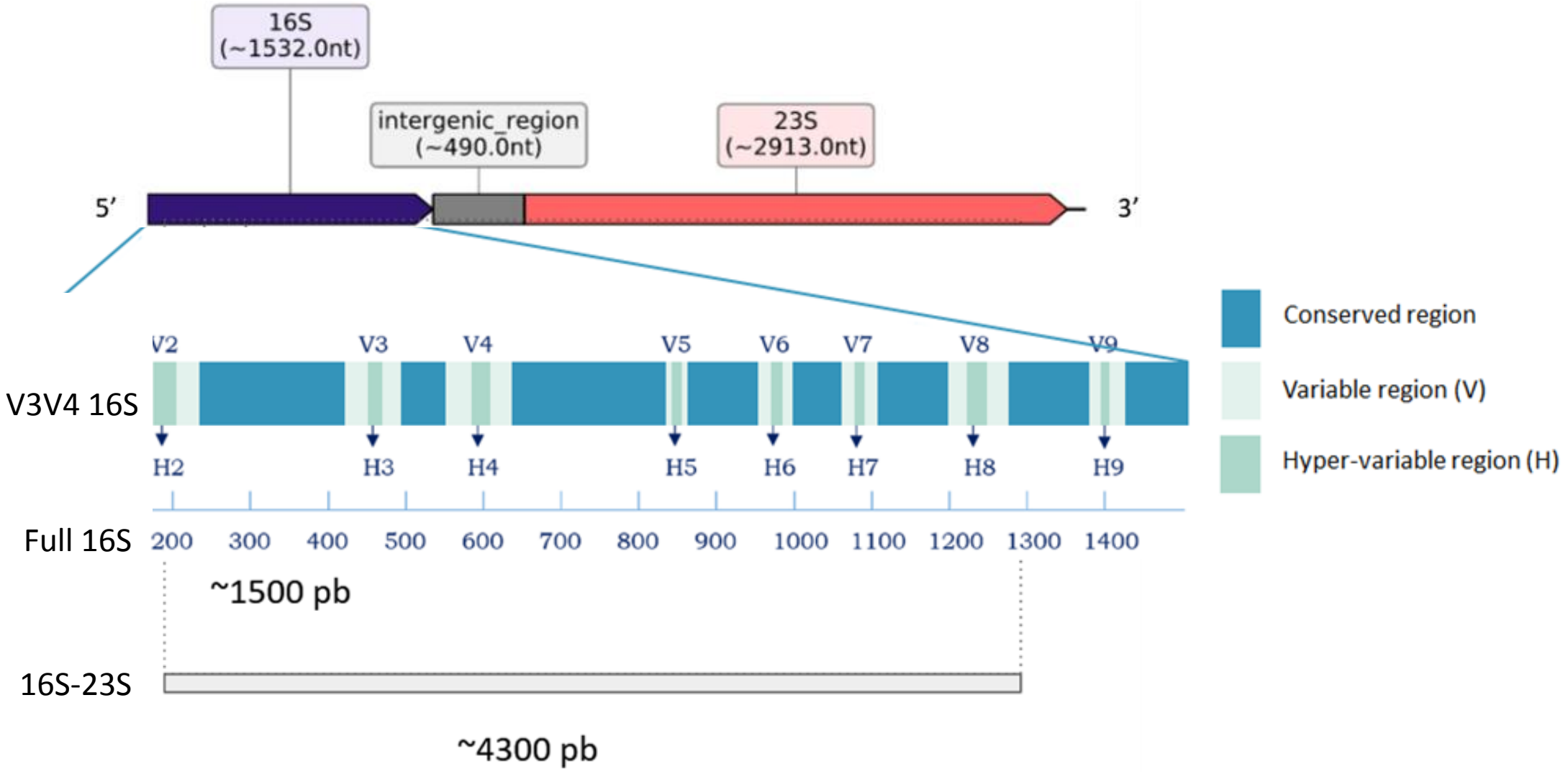
Original Paper | Published: 22 November 2020

Comparison of different hypervariable regions of 16S rRNA for taxonomic profiling of vaginal microbiota using next-generation sequencing

Auttawit Sirichoat, Nipapom Sankuntaw, Chulapan Engchanil, Pranom Buppasiri, Kiatichai Faksri, Wises Namwat, Wasun Chantratita & Viraphong Lulitanond

Genotoul GeT Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité République Française
L'Europe s'engage L'Occitanie agit
Projet cofinancé par le Fonds Européen de Développement Régional

rRNA genes - metabarcoding long read



Genotoul GeT
Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité République Française
L'Europe s'engage l'Occitanie agit
Projet cofinancé par le Fonds Européen de Développement Régional

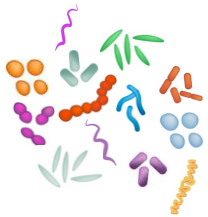
What would be the contribution of "long reads" in metabarcoding?

Focus on 16S / 16S23S markers

SeqOccIn

Known biaises in metabarcoding

Community analysis



- ✓ 16S copy number
- ✓ Horizontal gene transferts

DNA extraction

- ✓ DNA extraction kits/protocols

Barcode amplification

PCR
- ✓ Polymerase efficiency
- ✓ Polymerase contaminations
- ✓ non homogenous amplification

Amplicon sequencing

- ✓ Sequencing errors

Bioinformatics process

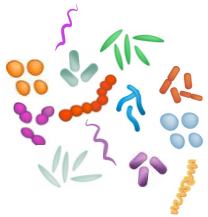
- ✓ Analysis softwares
- ✓ Databases exhaustivity

Genotoul GeT Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité République Française
L'Europe s'engage l'Occitanie agit
Projet cofinancé par le Fonds Européen de Développement Régional

SeqOccIn

Metabarcoding process for long read

Community analysis



DNA extraction

Barcode amplification

Amplicon sequencing

Bioinformatics process

- ✓ Long enough DNA to amplify 4,5Kb region.
- ✓ Low bias.

- ✓ Choose the most suitable and the least biased protocol for PacBio library preparation.

- ✓ HiFi read (>99,9% accuracy)

- ✓ Adapt bioinformatic tool for the long read

Genotoul GeT Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE
L'EUROPE S'ENGAGE L'OCCITANIE AGIT
Projet cofinancé par le Fonds Européen de Développement Régional

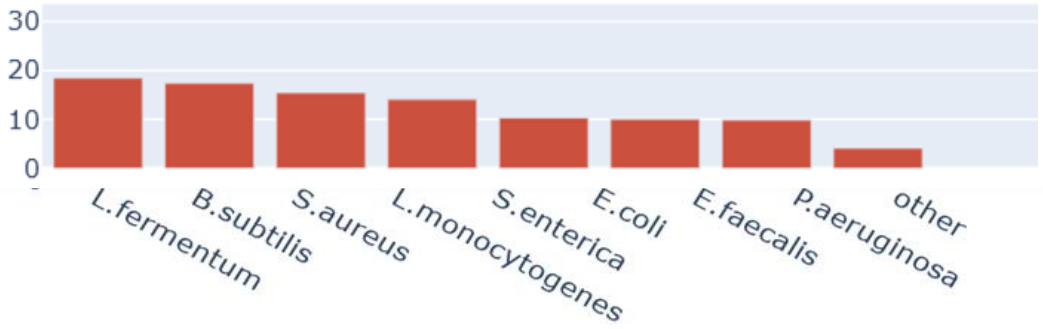
SeqOccIn

Biological matrix used in the project

Simple community

- **Bacterial mock** *(ZymoBIOMICS)*

8 bacterias (3 Gram- and 5 Gram+)



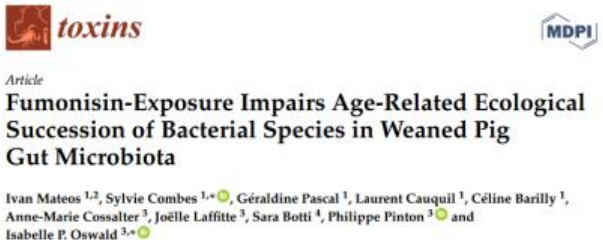
- **DNA mock** *(ZymoBIOMICS):* artificial mix of DNA from individualy extracted strain

Complex community

- **Pig faeces** (hard to extract, inhibitors)

32 faeces samples from 16 individuals (GenPhySE) exposed or not to mycotoxine Fumonisin B1

- Short read (V3-V4) data available


toxins MDPI

Article

Fumonisin-Exposure Impairs Age-Related Ecological Succession of Bacterial Species in Weaned Pig Gut Microbiota

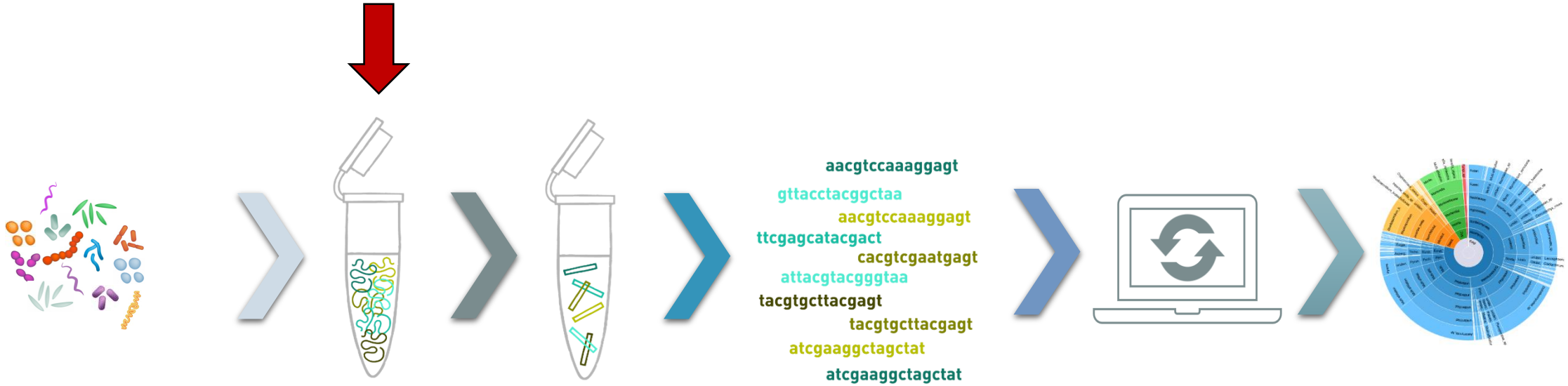
Ivan Mateos [1,2], Sylvie Combes [1,*], Géraldine Pascal [1], Laurent Cauquil [1], Céline Barilly [1], Anne-Marie Cossalter [3], Joëlle Laffitte [3], Sara Botti [4], Philippe Pinton [3] and Isabelle P. Oswald [3,*]

Genotoul GeT Genotoul Bioinfo

INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE
L'EUROPE S'ENGAGE L'OCCITANIE AGIT
Projet cofinancé par le Fonds Européen de Développement Régional

SeqOccIn

Metabarcoding process: DNA extraction



Community analysis

DNA extraction

Barcode amplification

Amplicon sequencing

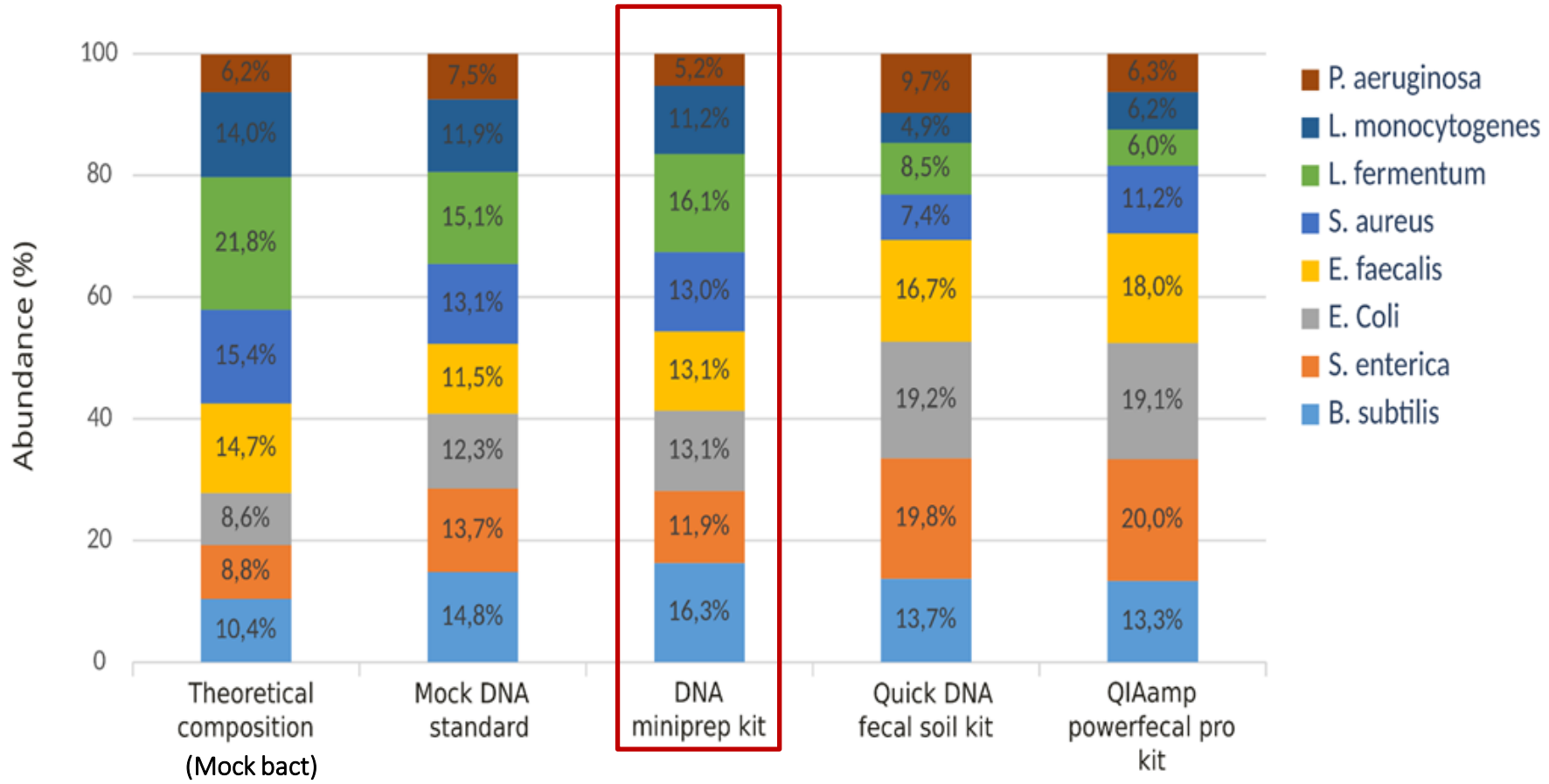
Bioinformatics process

Diversity characterisation

Genotoul GeT Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité République Française
L'Europe s'engage l'Occitanie agit
Projet cofinancé par le Fonds Européen de Développement Régional

SeqOccIn

DNA extraction: validation on short read (V3-V4 region)

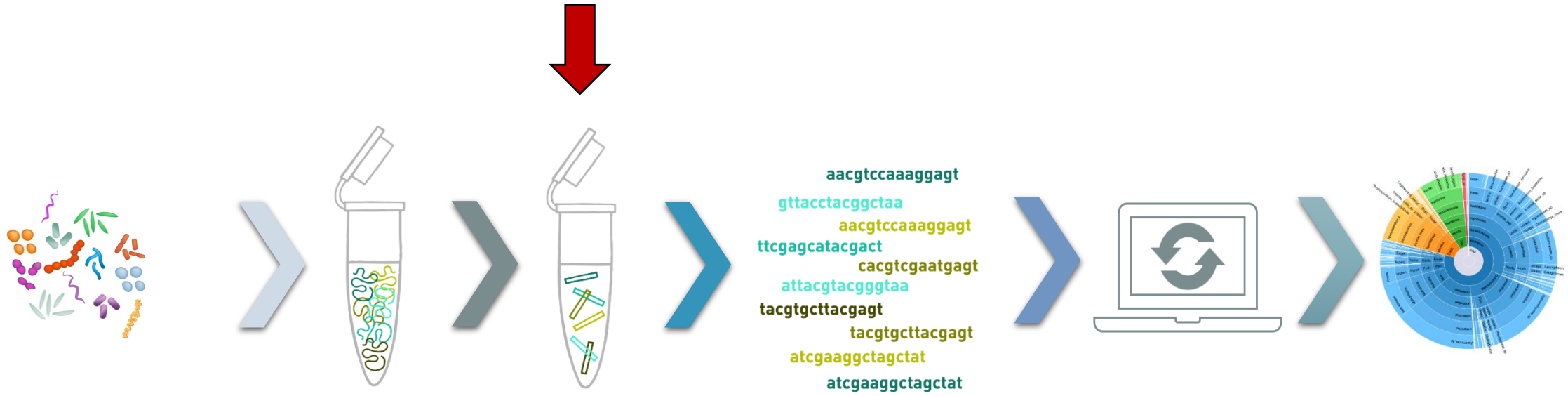


- **Slight biases when sequencing mock DNA**
- **Zymo DNA miniprep kit gave the best results, choosen for metabarcoding**

Genotoul GeT Genotoul Bioinfo

INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE
L'EUROPE S'ENGAGE L'OCCITANIE AGIT
Projet cofinancé par le Fonds Européen de Développement Régional

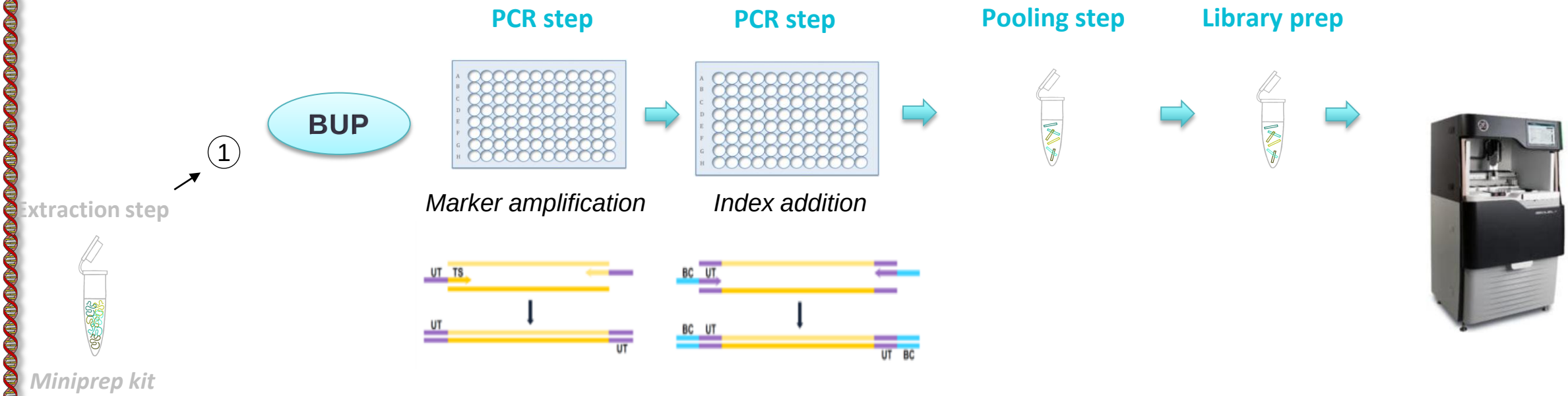
Metabarcoding process: barcode amplification



Community analysis ✓ DNA extraction ✓ **Barcode amplification** Amplicon sequencing Bioinformatics process Diversity characterisation

Genotoul GeT Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité République Française
L'Europe s'engage l'Occitanie agit
Projet cofinancé par le Fonds Européen de Développement Régional

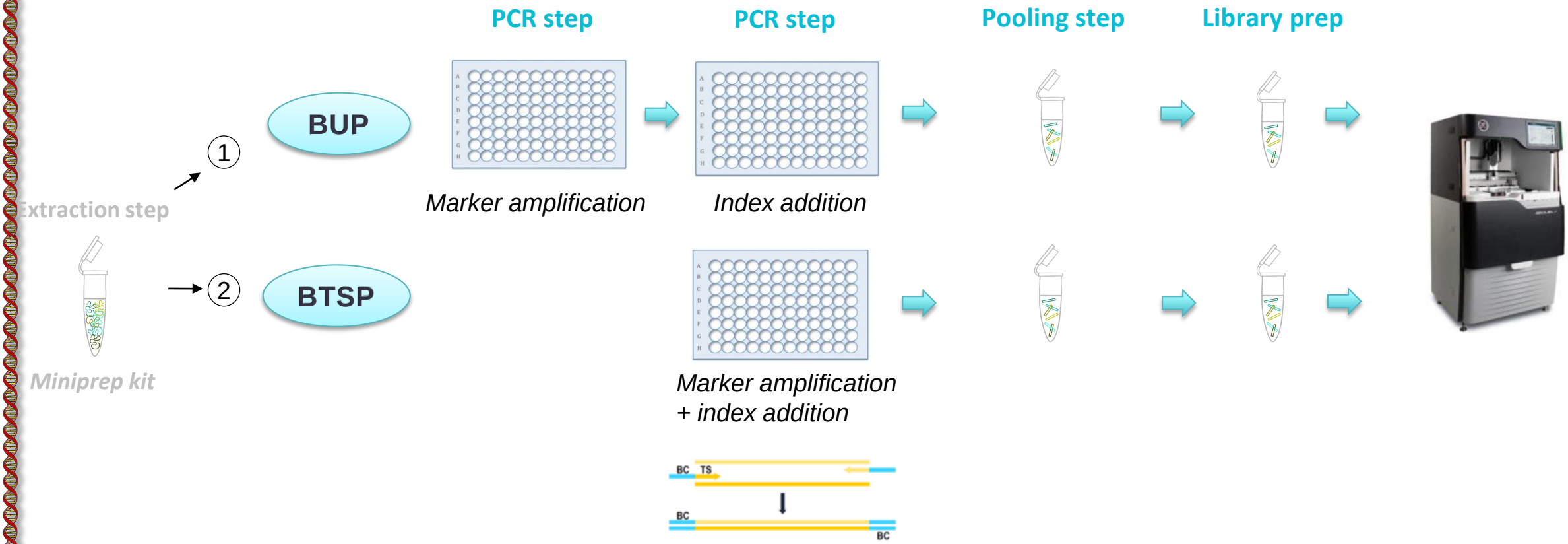
Comparison of PacBio protocols



- **Protocols have been tested with 16S and 16S-23S barcodes on mock and pig faeces samples**

https://www.pacb.com/wp-content/uploads/Multiplexed-Amplicon-Library-Preparation-Using-SMRTbell-Express-Template-Prep-Kit-2.0-%E2%80%93-Customer-Training.pdf

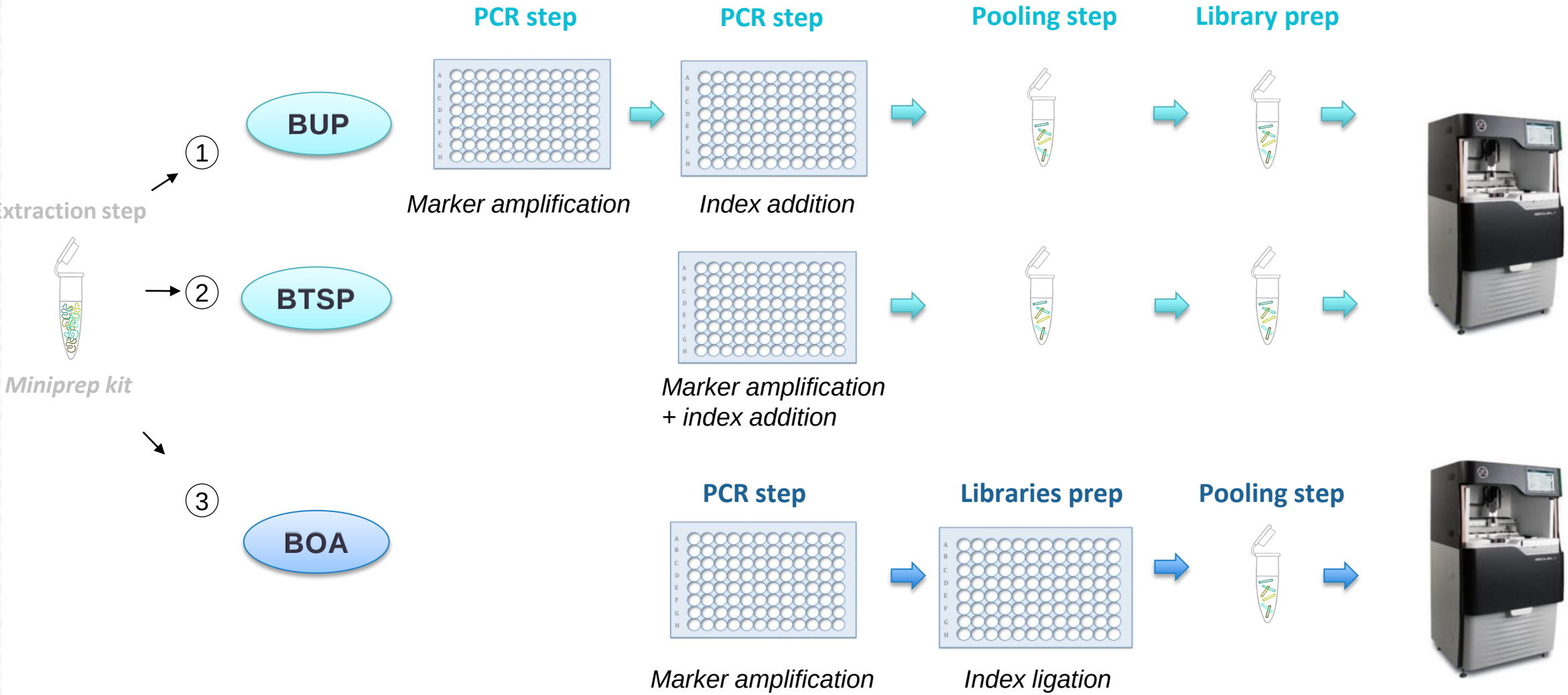
Comparison of PacBio protocols



- **Protocols have been tested with 16S and 16S-23S barcodes on mock and pig faeces samples**

https://www.pacb.com/wp-content/uploads/Multiplexed-Amplicon-Library-Preparation-Using-SMRTbell-Express-Template-Prep-Kit-2.0-%E2%80%93-Customer-Training.pdf

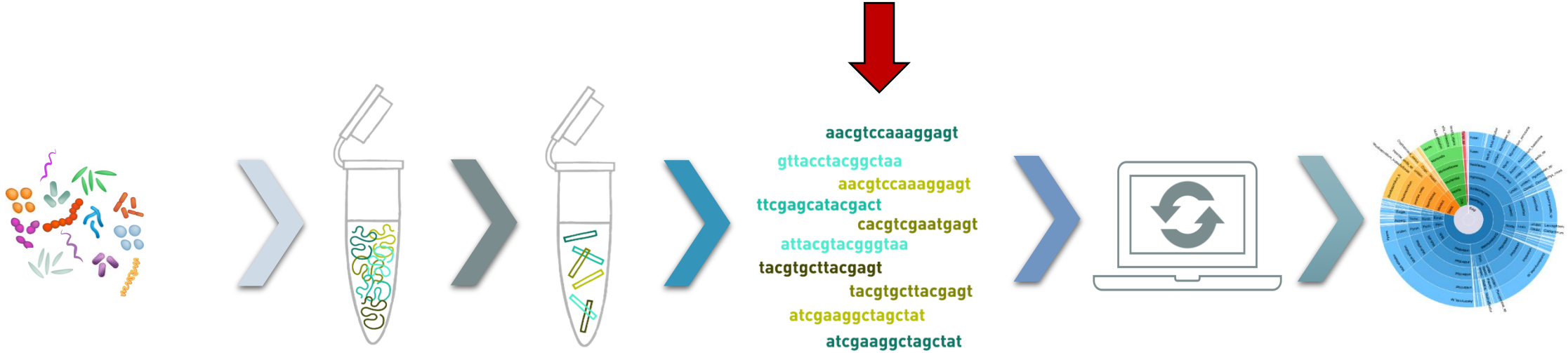
Comparison of PacBio protocols



- **Protocols have been tested with 16S and 16S-23S barcodes on mock and pig faeces samples**

https://www.pacb.com/wp-content/uploads/Multiplexed-Amplicon-Library-Preparation-Using-SMRTbell-Express-Template-Prep-Kit-2.0-%E2%80%93-Customer-Training.pdf

Metabarcoding process: amplicon sequencing



Community analysis ✓ DNA extraction ✓ Barcode amplification ✓ **Amplicon sequencing** Bioinformatics process Diversity characterisation

Genotoul GeT Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité République Française
L'Europe s'engage l'Occitanie agit
Projet cofinancé par le Fonds Européen de Développement Régional

SeqOccIn

Sequencing results for 16S and 16S-23S barcodes

Protocol	Subreads N50	HiFi Reads	Gb CCS	Quality
16S				
BUP	1,602	2 813 935	4,4	Q40
BTSP	1,569	3 139 687	4,8	Q40
BOA	1,568	2 834 523	4,4	Q42
16S-23S				
BUP	4,2	2 962 319	11,7	Q46
BTSP	4,204	3 173 610	12,7	Q45
BOA	4,325	2 534 890	10,1	Q43

- **Sequencing Ok!**

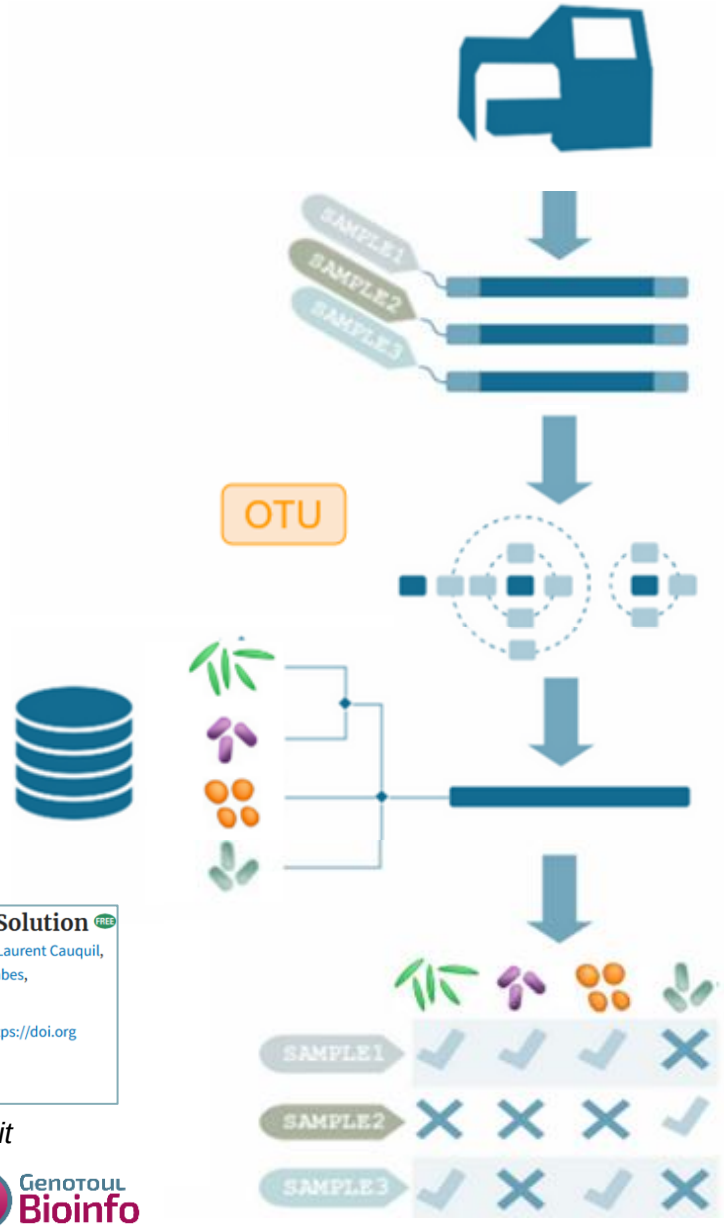
Genotoul GeT Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité République Française
L'Europe s'engage l'Occitanie agit
Projet cofinancé par le Fonds Européen de Développement Régional

Metabarcoding process: bioinformatics process

Community analysis ✓ DNA extraction ✓ Barcode amplification ✓ Amplicon sequencing ✓ **Bioinformatics process** **Diversity characterisation**

Genotoul GeT Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité République Française
L'Europe s'engage l'Occitanie agit
Projet cofinancé par le Fonds Européen de Développement Régional

Bioinformatics process for metabarcoding



Sequencing

Demultiplexing
Each index is assignes to its original sample

Denoising and clustering
Sequences are groupes by similarity. Amplification and sequencing errores are masked.

Taxonomic affiliation
Each group is affiliated with a taxon through a reference database.

Abundance table
Detect and count the taxa present in each sample.

FROGS: Find, Rapidly, OTUs with Galaxy Solution

Frédéric Escudié, Lucas Auer, Maria Bernard, Mahendra Mariadassou, Laurent Cauquil, Katia Vidal, Sarah Maman, Guillermina Hernandez-Raquet, Sylvie Combes, Géraldine Pascal ✉ Author Notes

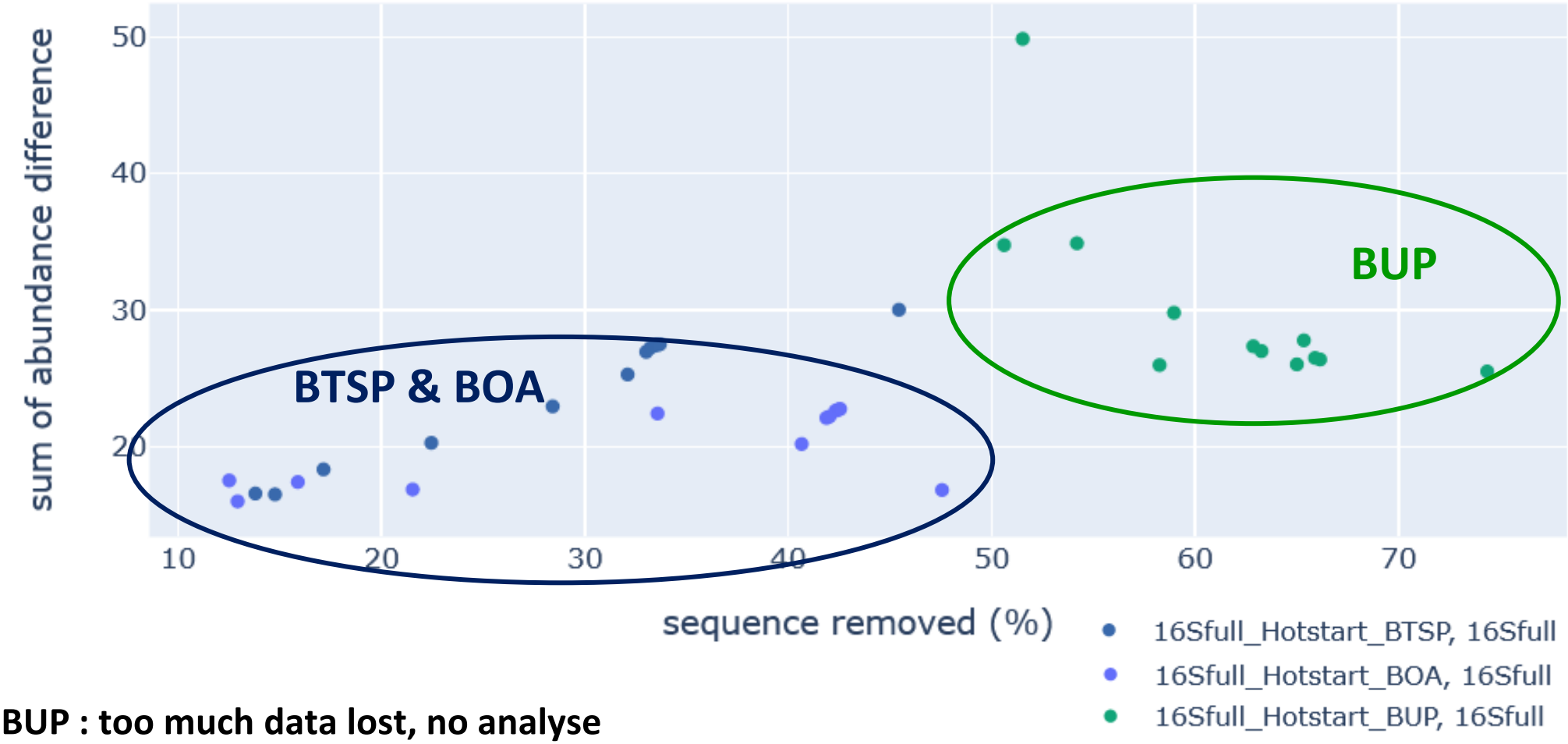
Bioinformatics, Volume 34, Issue 8, 15 April 2018, Pages 1287–1294, https://doi.org/10.1093/bioinformatics/btx791

Published: 07 December 2017 **Article history**

*OTU : Operational Taxonomic Unit

Projet cofinancé par le Fonds Européen de Développement Régional

% of sequences lost during analysis

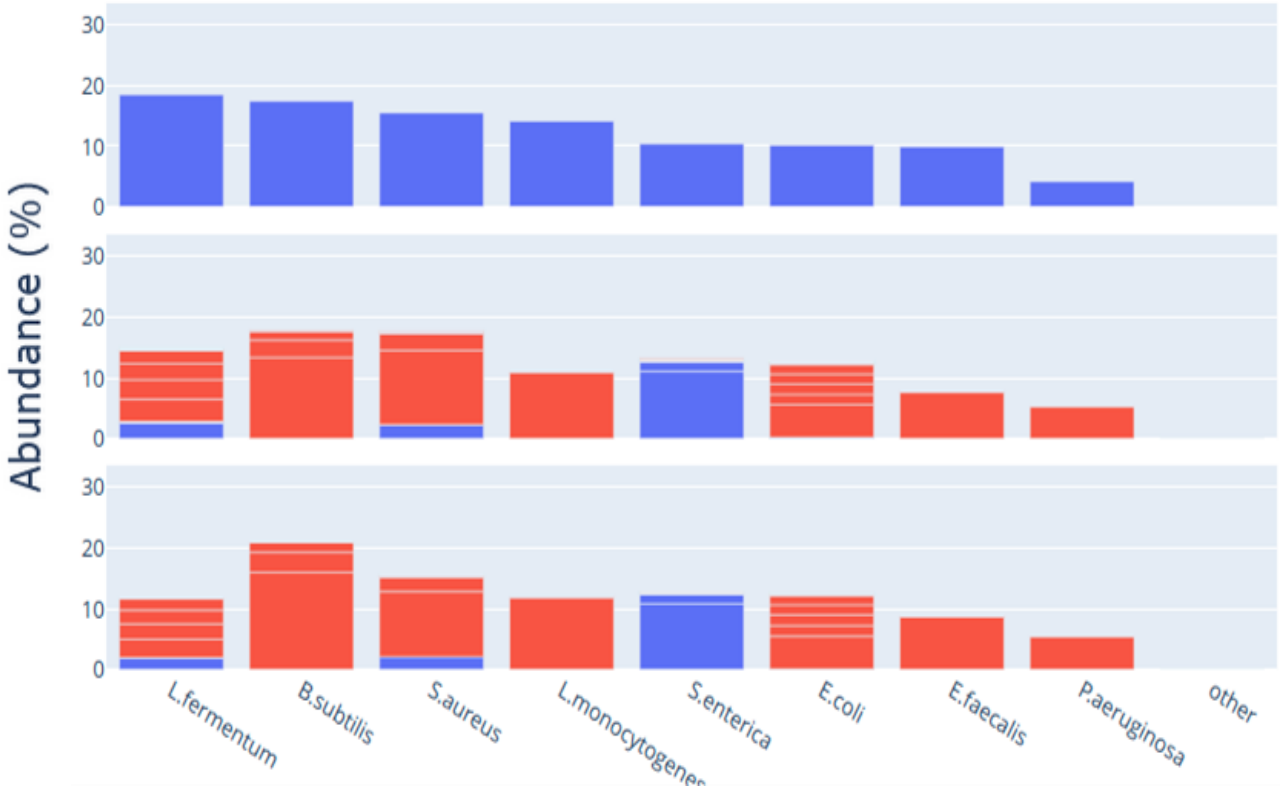


- **BUP : too much data lost, no analyse**
- **BTSP & BOA protocols: same behaviour whatever the software used**

Genotoul GeT Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Projet cofinancé par le Fonds Européen de Développement Régional

BTSP & BOA comparison: 16S rRNA sequencing

Mock DNA 16S rRNA sequencing on PacBio Sequel II, HiFi



Theoritical abundance
(Mock bact)

BOA

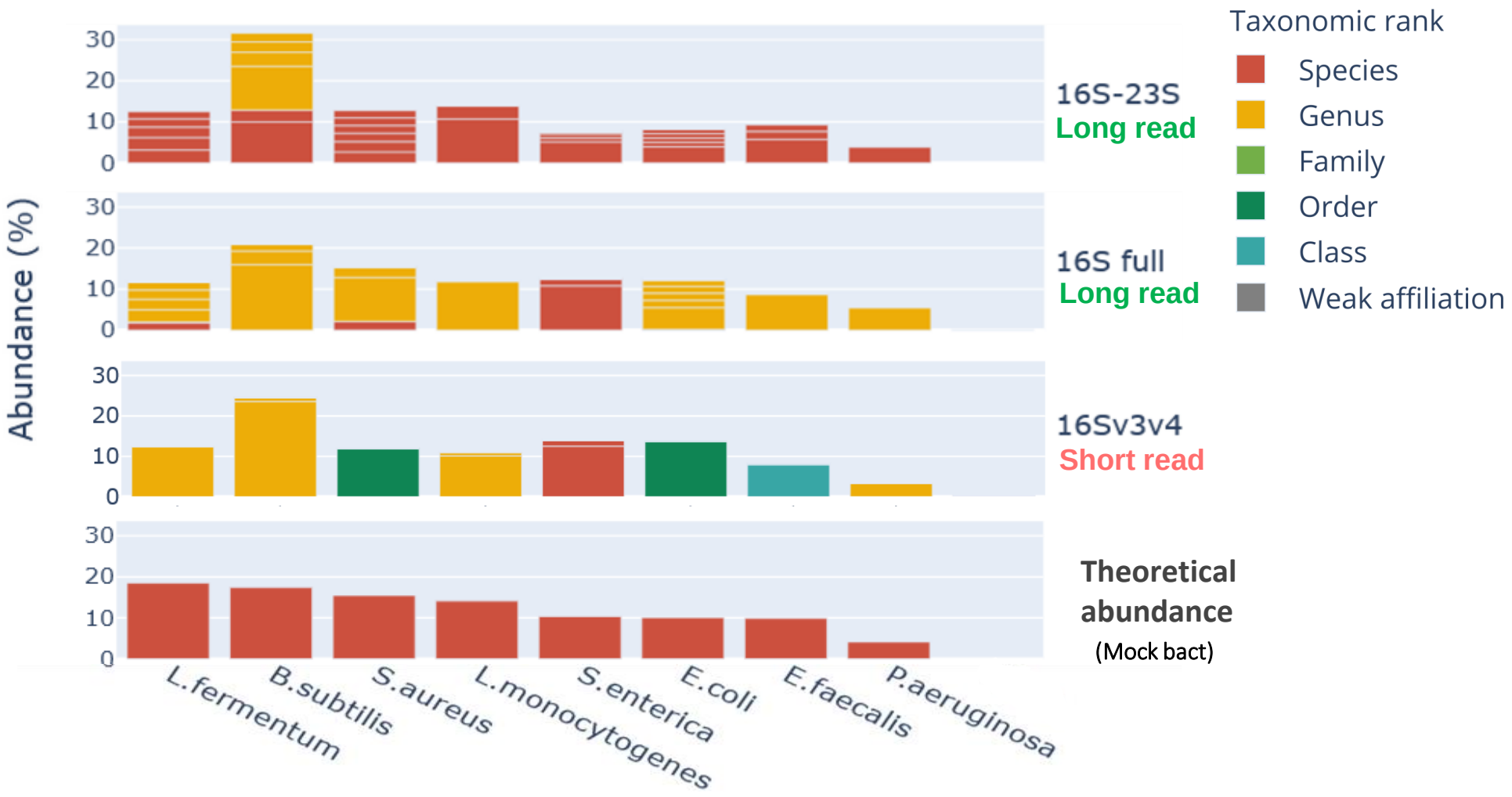
BTSP

Taxonomic rank
- Species
- Genus

➢ **Similar results obtained for BTSP & BOA sequencing**

SeqOccIn

Results: affiliation of OTUs from metabarcodings with different targets - mock community



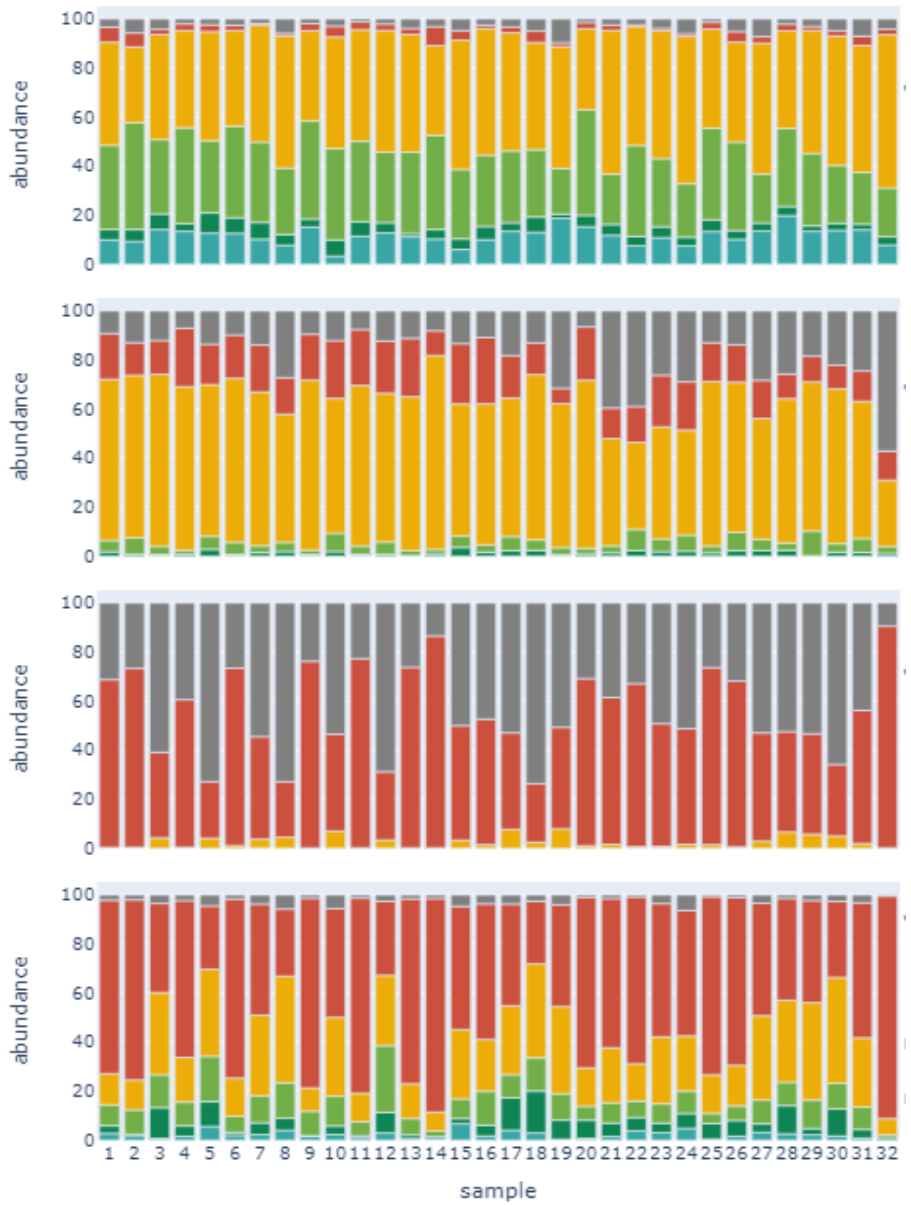
- **A longer barcode allows to describe more precisely a simple community.**

Projet cofinancé par le Fonds Européen de Développement Régional

SeqOccIn

Results: Affiliation of OTUs from metabarcodings with different targets - 32 pig faeces samples

- Weak affiliation consist of sequences that have no affiliation hit above 99 % identity and 99% coverage.
- The multi DB approach of the 16S-23S allows to greatly reduce the proportion of weak affiliations
- A longer barcode allows to describe more precisely a complex community.



Genotoul GeT Genotoul Bioinfo

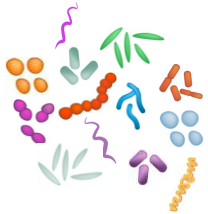
INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité République Française
L'Europe s'engage l'Occitanie agit
Projet cofinancé par le Fonds Européen de Développement Régional

SeqOccIn

Conclusions

- Implemented an **optimized process to performed long read metabarcoding** on 16S and 16S-23S with PacBio HiFi technology :

Community analysis



Pig faeces

DNA extraction

Zymo DNA miniprep kit

Barcode amplification

BOA and BTSP protocols on 16S and 16S-23S barcodes

Amplicon sequencing

PacBio HiFi technology

Bioinformatics process

FROGS
Adapted to long read

- **16S-23S region is more informative** than full length 16S
- Problem remains with the database exhaustivity that require time to be completed

Genotoul GeT Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité République Française
L'Europe s'engage l'Occitanie agit
Projet cofinancé par le Fonds Européen de Développement Régional

SeqOccIn

INRAE
la science pour la vie, l'humain, la terre

Contribution of long reads for whole metagenome analysis

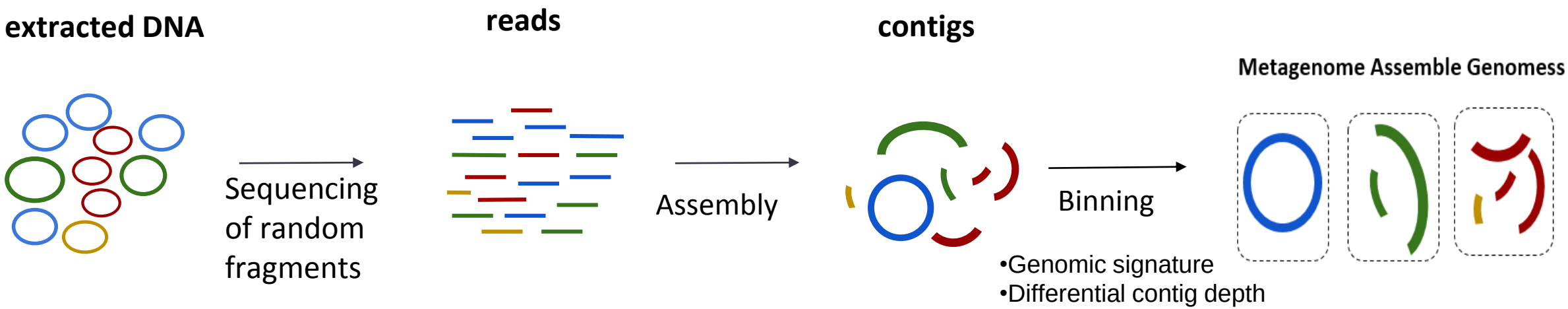
Genotoul GeT

http://get.genotoul.fr
@Get_Genotoul

http://bioinfo.genotoul.fr
@BioinfoGenotoul

Genotoul Bioinfo

Assembly: principle and tools used



- The Genotoul bio-info platform has developed a metagenomic shotgun pipeline for the analysis of Illumina and PacBio HiFi sequencing data. The PacBio HiFi part was developed during SeqOccIn project.

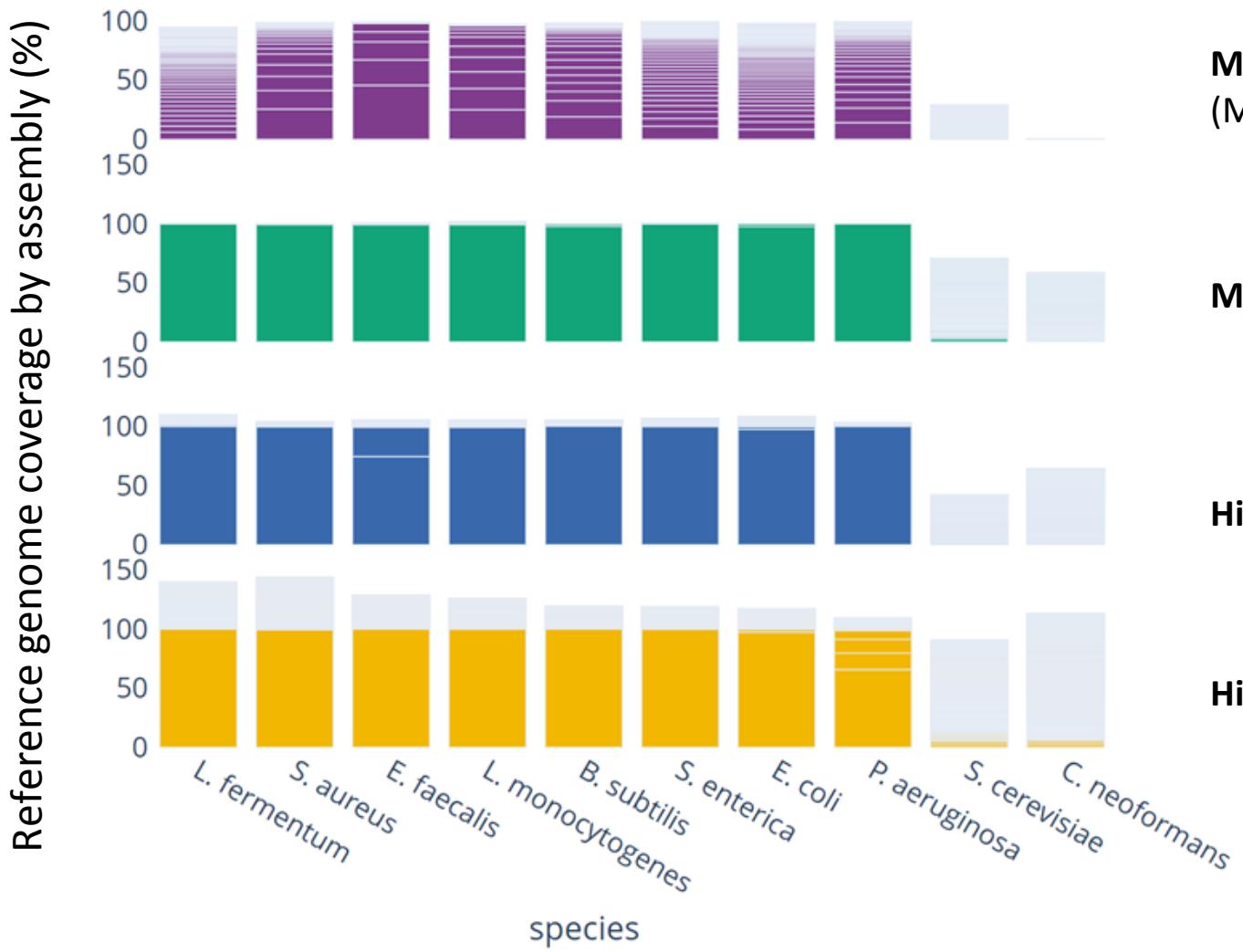
https://forgemia.inra.fr/genotoul-bioinfo/metagwgs

- This pipeline allows to perform assembly, taxonomic affiliation of contigs, functional annotation and binning of contigs.

Projet cofinancé par le Fonds Européen de Développement Régional

Mocks assembly are very good with HiFi reads

Mock bact



MetaSPAdes Illumina
(Maxime Manno, get-plage, get-IT)

Metaflye

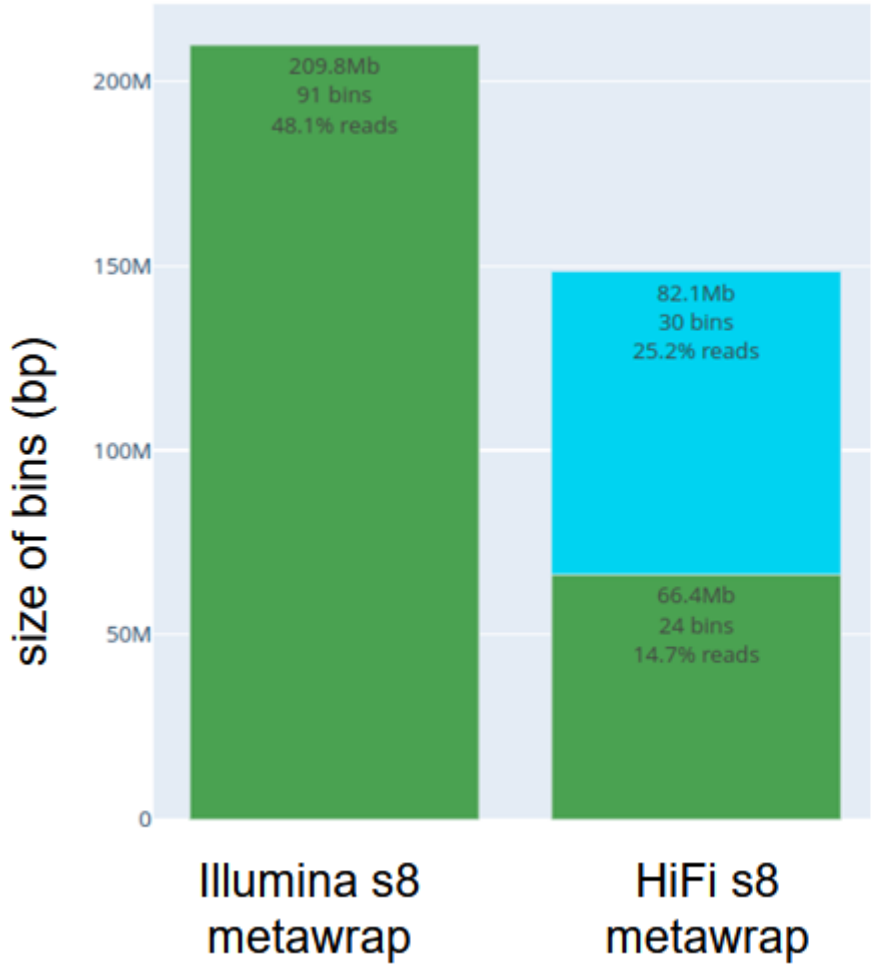
HiCanu

Hifiasm-meta

Illumina: 68M reads pairs (150pbs each)

Genotoul GeT
Genotoul Bioinfo
INRAE la science pour la vie, l'humain, la terre
Projet cofinancé par le Fonds Européen de Développement Régional

Binning of HiFi assemblies



High-quality
High quality and rna complete

High quality bin:

- >90% completeness
- <5% contamination

RNA complete:
Presence of the 23S, 16S and 5S rRNA genes and at least 18 tRNAs.

Less High quality bin with HiFi…
But 30 high quality bins are rna complete !

Thanks to

SeqOccIn

Coordination

Cécile Donnadieu
Christine Gaspin
Carole Iampietro
Denis Milan

Jean Mainguy
Olivier Bouchez
Sylvie Combes
Claire Hoede
Géraldine Pascal

Projet cofinancé par le Fonds Européen de Développement Régional