



**HAL**  
open science

## Avancées réalisées en métabarcoding et en métagénomique dans le cadre du projet SeqOccIn grâce au séquençage longs fragments

Denis Milan, Jean Mainguy, Adrien Castinel, Olivier Bouchez, Sylvie Combes,  
Claire Hoede, Géraldine Pascal

### ► To cite this version:

Denis Milan, Jean Mainguy, Adrien Castinel, Olivier Bouchez, Sylvie Combes, et al.. Avancées réalisées en métabarcoding et en métagénomique dans le cadre du projet SeqOccIn grâce au séquençage longs fragments. Colloque INRAE Genomics 2022, May 2022, Orléans, France. , 2022. hal-03776011

**HAL Id: hal-03776011**

**<https://hal.science/hal-03776011>**

Submitted on 13 Sep 2022

**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.

## Avancées réalisées en métabarcoding et en métagénomique dans le cadre du projet SeqOccln grâce au séquençage longs fragments

Jean Mainguy, Adrien Castinel, Olivier Bouchez,  
Sylvie Combes, Claire Hoede, Géraldine Pascal

*Denis Milan*

# Un topo à partir de deux topos complémentaires

Présentés lors de l'Assemblée Générale de SeqOocln

## Métabarcoding





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



**Strengths and weaknesses of metabarcoding long read.**

Axe 3: métagénomique  
Jean Mainguy, Adrien Castinel, Olivier Bouchez, Sylvie Combes, Claire Hoede, Géraldine Pascal


<http://get.genotoul.fr>  
 @Get\_Genotoul


<http://bioinfo.genotoul.fr>  
 @BioinfoGenotoul



## Métagénomique






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

**Contribution of long reads in bioinformatics for "whole genome" metagenomic analysis.**

Jean Mainguy, Adrien Castinel, Olivier Bouchez, Sylvie Combes, Géraldine Pascal, Claire Hoede


<http://get.genotoul.fr>  
 @Get\_Genotoul


<http://bioinfo.genotoul.fr>  
 @BioinfoGenotoul

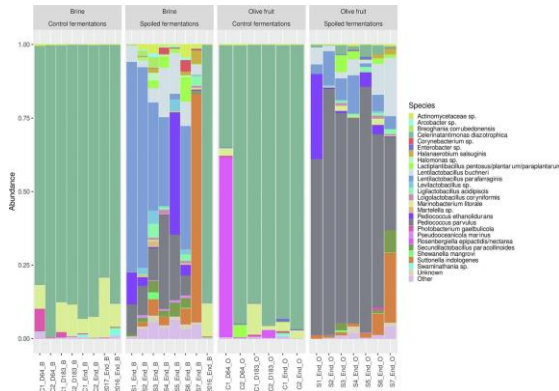
# Metabarcoding

# Metabarcoding & Metagenomic

## Microbiota DNA

### Amplicon sequencing

### Whole genome sequencing



Penland *et al.*, 2021

Fig. 2: Bacterial gene functional capacity enrichment in rhizosphere and soil under drought.

From: Genome-resolved metagenomics reveals role of iron metabolism in drought-induced rhizosphere microbiome dynamics

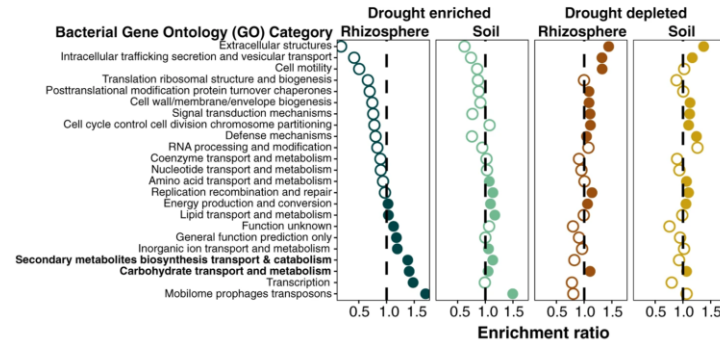
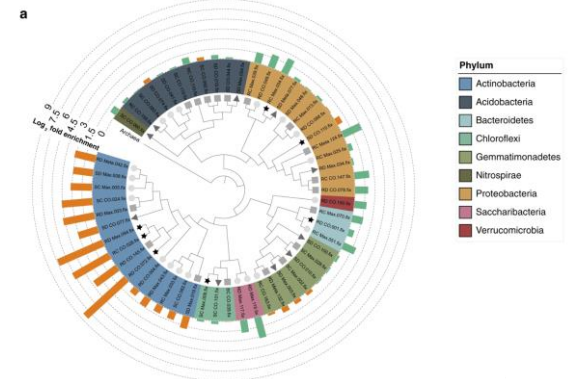


Fig. 3: Phylogenetic tree of the 55 MAGs.

From: Genome-resolved metagenomics reveals role of iron metabolism in drought-induced rhizosphere microbiome dynamics



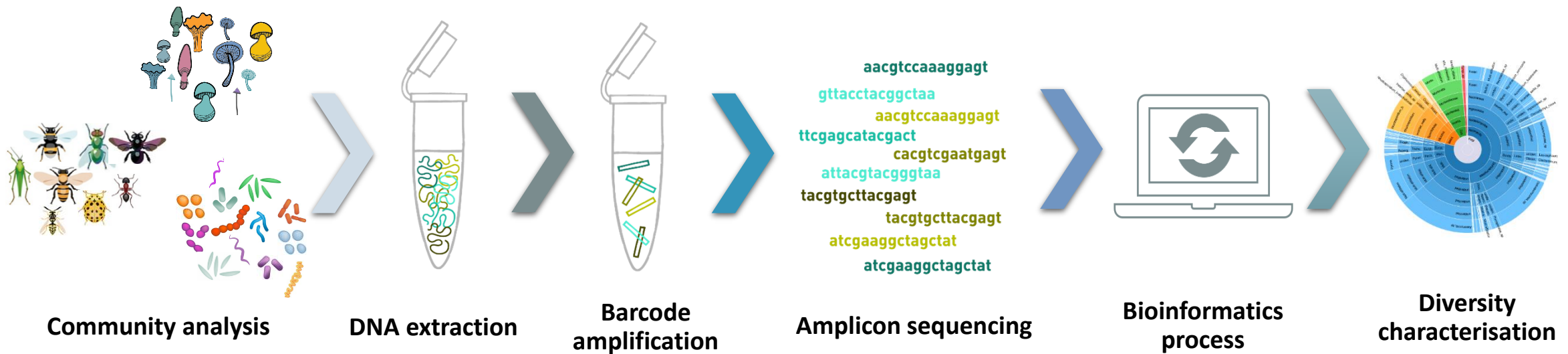
Xu *et al.*, 2021

**Metabarcoding**  
*Who's here ?*

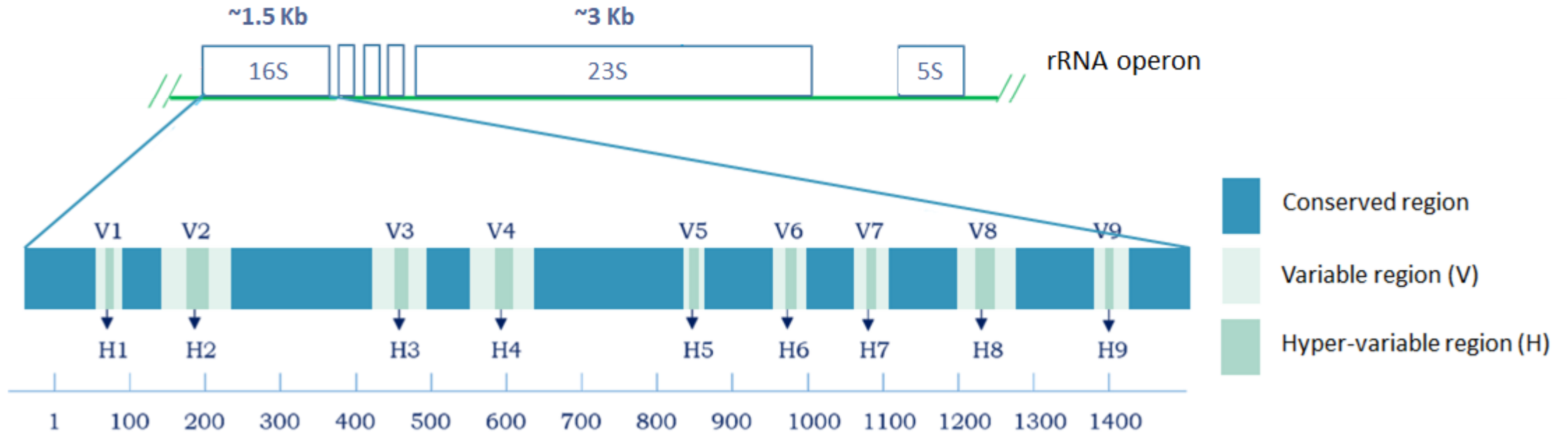
**Metagenomic**  
*What can they do ?*  
*Who can do what ?*

# 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.



# rRNA genes– metabarcoding using short reads or long reads



- A 16S region → with **short reads**
- The full 16S → with **long reads**
- The full 16S-23S operon → with **long reads**



# 16S analysis using short reads & long reads

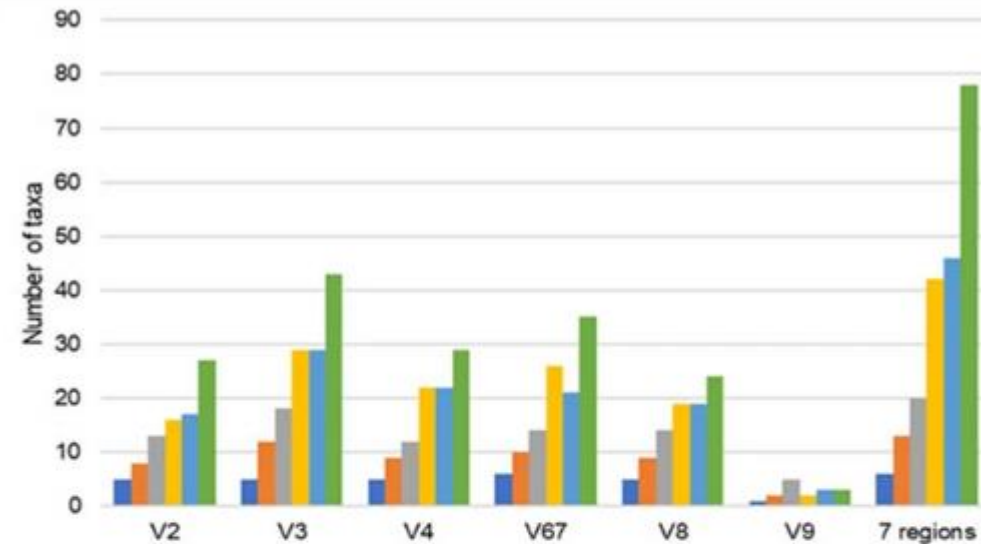
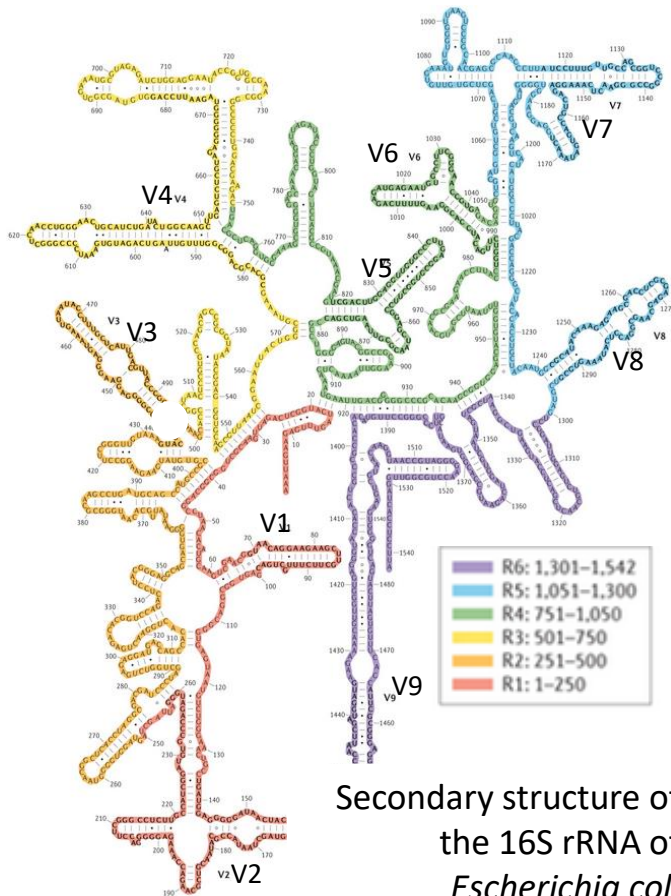
Original Paper | Published: 22 November 2020

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

Auttawit Sirichoat<sup>1</sup>, Nipaporn Sankuntaw<sup>1</sup>, Chulapan Engchanai<sup>1</sup>, Pranom Buppasini<sup>1</sup>, Kitichai Faksai<sup>1</sup>, Wiset Namwat<sup>1</sup>, Wasun Chantrattita<sup>1</sup> & Viraphong Lulitanond<sup>1</sup>✉

Archives of Microbiology 203, 1159–1166 (2021) | [Cite this article](#)

861 Accesses | 7 Citations | 3 Altmetric | [Metrics](#)

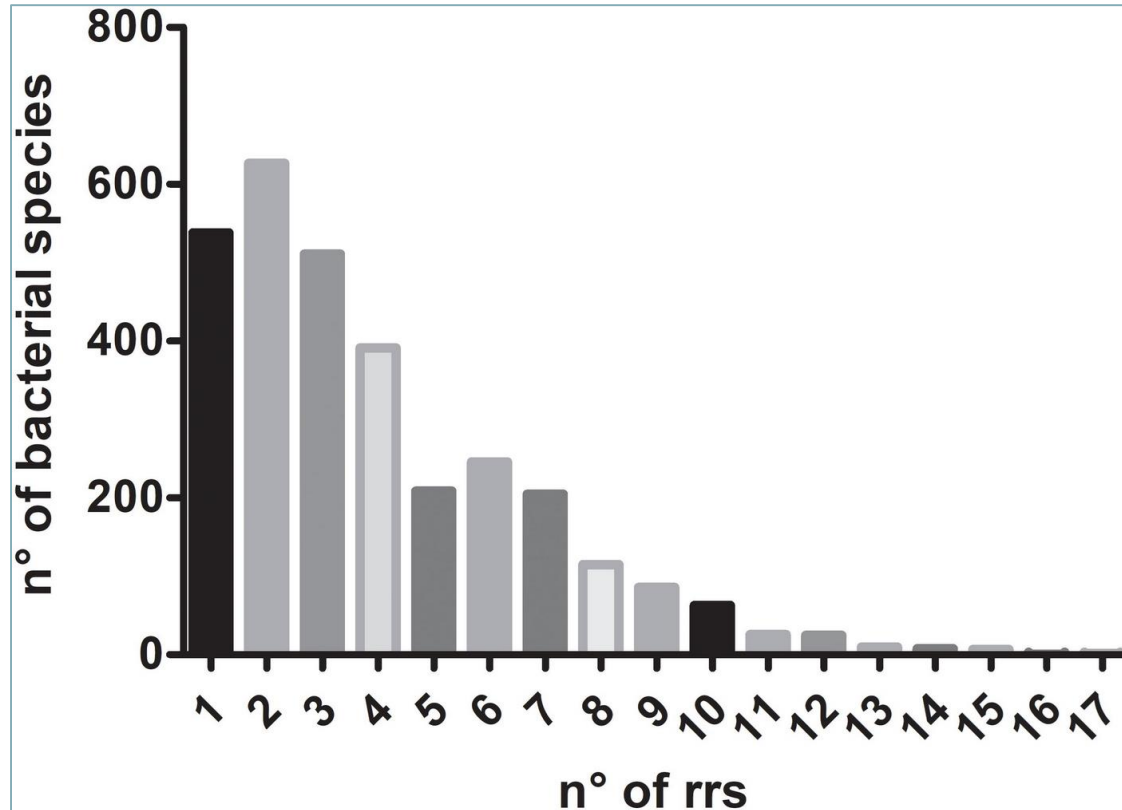


Phylum	5	5	5	6	5	1	6
Class	(83.3)	(83.3)	(83.3)	(100)	(83.3)	(16.7)	(100)
Order	8	12	9	10	9	2	13
Family	(61.5)	(92.3)	(69.2)	(76.9)	(69.2)	(15.4)	(100)
Genus	13	18	12	14	14	5	20
Species	(65.0)	(90.0)	(60.0)	(70.0)	(70.0)	(25.0)	(100)
Phylum	16	29	22	26	19	2	42
Class	(38.1)	(69.0)	(52.4)	(61.9)	(45.2)	(4.8)	(100)
Order	17	29	22	21	19	3	46
Family	(37.0)	(63.0)	(47.8)	(45.7)	(41.3)	(6.5)	(100)
Genus	27	43	29	35	24	3	78
Species	(34.6)	(55.1)	(37.2)	(44.9)	(30.8)	(3.8)	(100)

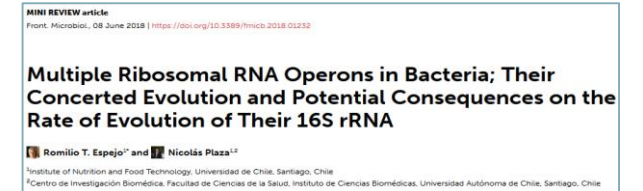
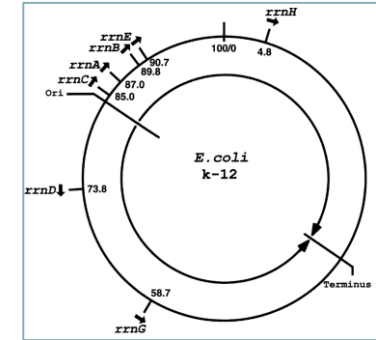
Number of taxa identified at each taxonomic level using individual and concatenated hypervariable



# 16S rRNA is not unique in all bacteria



Median of the number of 16S rRNA copies in 3,070 bacterial species according to data reported in *rrnDB* database - 2018



2022:

*Bacillus megaterium* entre 1 à 21 copies selon les souches (médiane à 13)

*Photobacterium damsela*e entre 15 et 21 copie selon les souches (médiane à 17)

# Biological matrices used in the project

## ✓ Bacteria mock (*ZymoBIOMICS*)



8 bacteria (3 Gram- et 5 Gram+)

Species	Theoretical Composition (%)				
	Genomic DNA	16S Only <sup>1</sup>	16S & 18S <sup>1</sup>	Genome Copy <sup>2</sup>	Cell Number <sup>3</sup>
<i>Pseudomonas aeruginosa</i>	12	4.2	3.6	6.1	6.1
<i>Escherichia coli</i>	12	10.1	8.9	8.5	8.5
<i>Salmonella enterica</i>	12	10.4	9.1	8.7	8.8
<i>Lactobacillus fermentum</i>	12	18.4	16.1	21.6	21.9
<i>Enterococcus faecalis</i>	12	9.9	8.7	14.6	14.6
<i>Staphylococcus aureus</i>	12	15.5	13.6	15.2	15.3
<i>Listeria monocytogenes</i>	12	14.1	12.4	13.9	13.9
<i>Bacillus subtilis</i>	12	17.4	15.3	10.3	10.3
<i>Saccharomyces cerevisiae</i>	2	NA	9.3	0.57	0.29
<i>Cryptococcus neoformans</i>	2	NA	3.3	0.37	0.18

## ✓ DNA mock (*ZymoBIOMICS*): artificial mix of DNA from individually extracted strain

## ✓ Pig feces (hard to extract, inhibitors)




32 feces sample from 16 individuals (GenPhySE) exposed or not to mycotoxine Fumonisin B1

- ✓ MiSeq 16S data available
- ✓ Illumina Shotgun data available

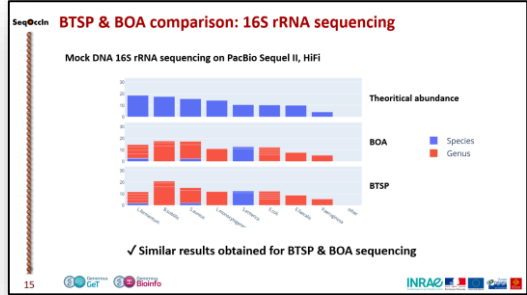
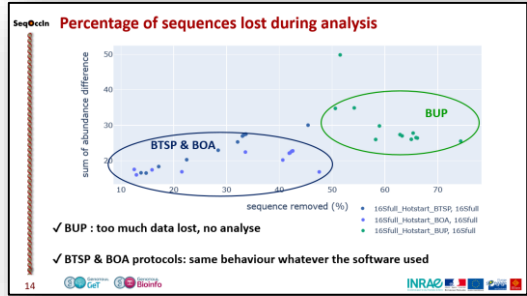
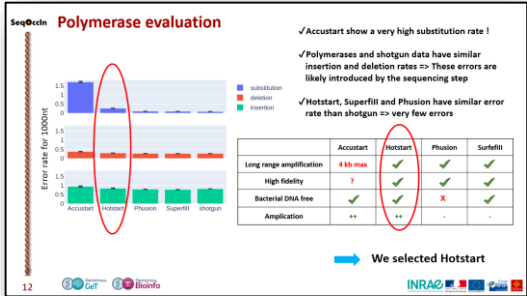


Article

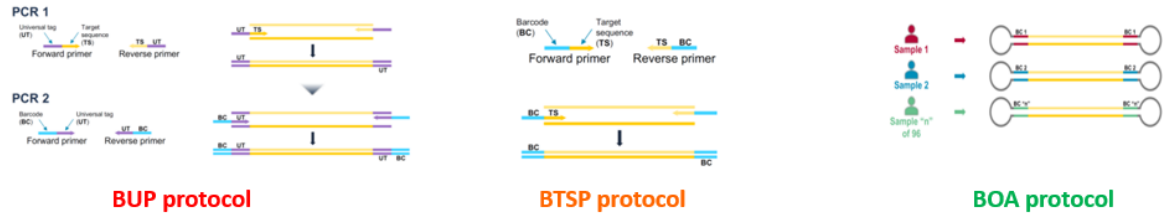
### Fumonisin-Exposure Impairs Age-Related Ecological Succession of Bacterial Species in Weaned Pig Gut Microbiota

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

# SeqOcln Protocol optimisation using PacBio Hifi reads

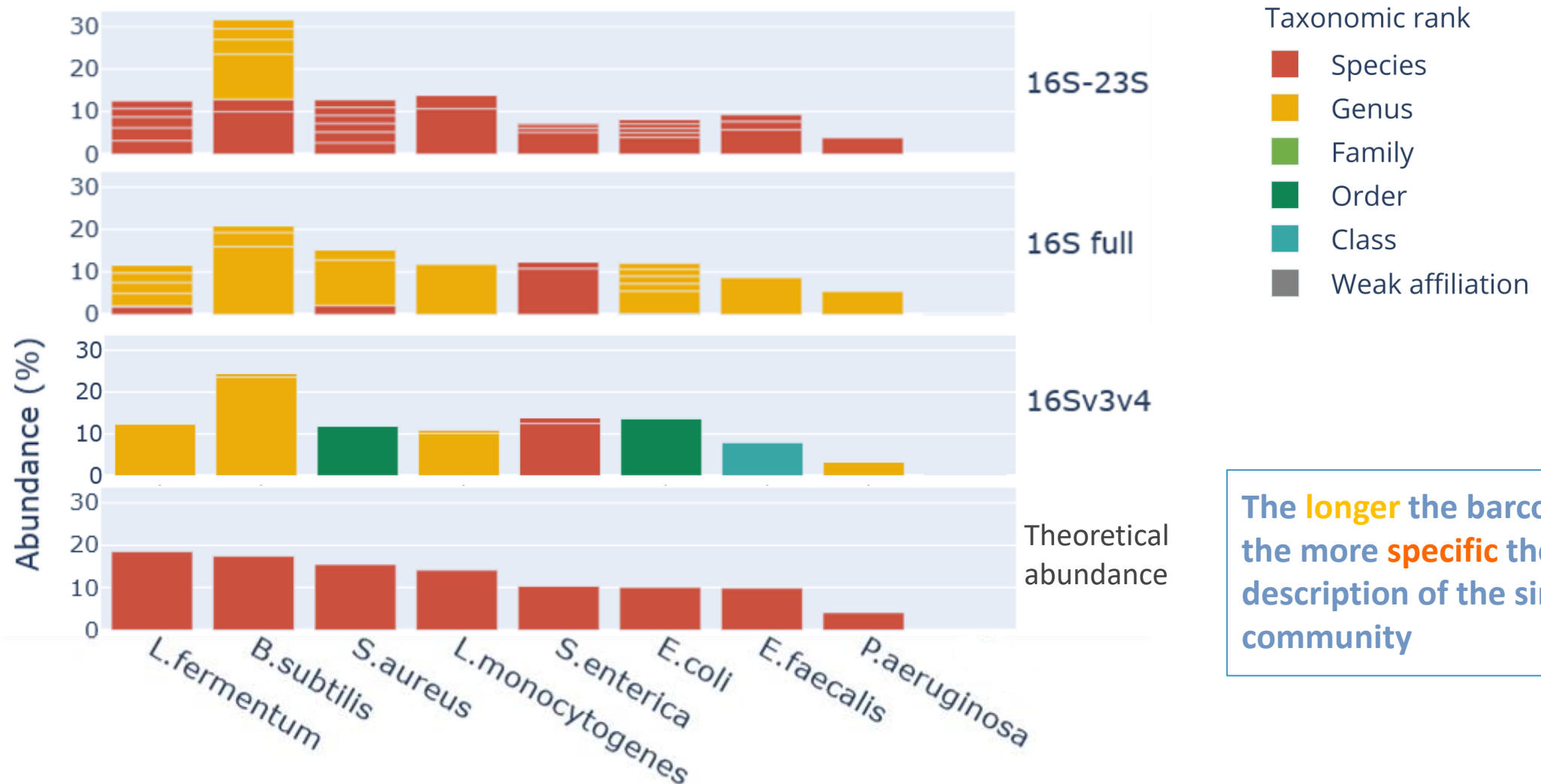


## SeqOcln PacBio protocols: conclusions



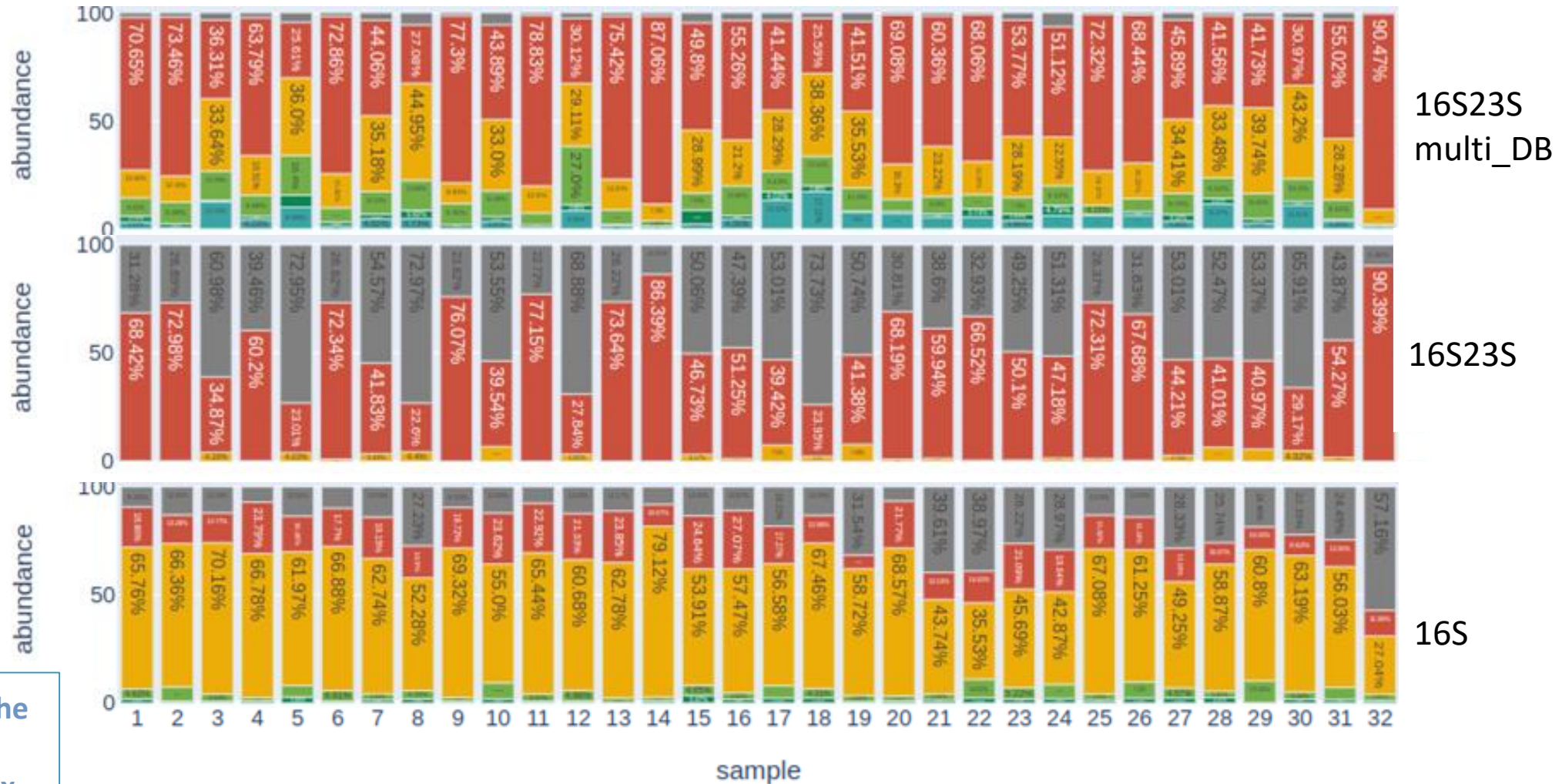
	<b>BUP protocol</b>	<b>BTSP protocol</b>	<b>BOA protocol</b>
	<b>X</b>		
	2 PCR PCR Development Primer design	Primer price PCR Development Primer design	1 PCR Primer design
		Usefull at a production scale on 1 target	1 PCR Lab process
PCR + sequencing (96 multiplex)	~70 €/sample	~70 €/sample + primers	~80 €/sample
Conclusions	Too many PCR biases	Primer design <u>complex</u> PCR optimisation not easy Lab process complicated	Protocol <u>easy to implement whatever the target</u>

# Results: affiliation of OTUs from metabarcodes with different targets on mock community





# Results: Affiliation of OTUs from metabarcodes with different targets on 32 feces samples



The longer the barcode, the more specific the description of the complex community

# Conclusion sur l'apport des Longs Fragments au métabarcoding

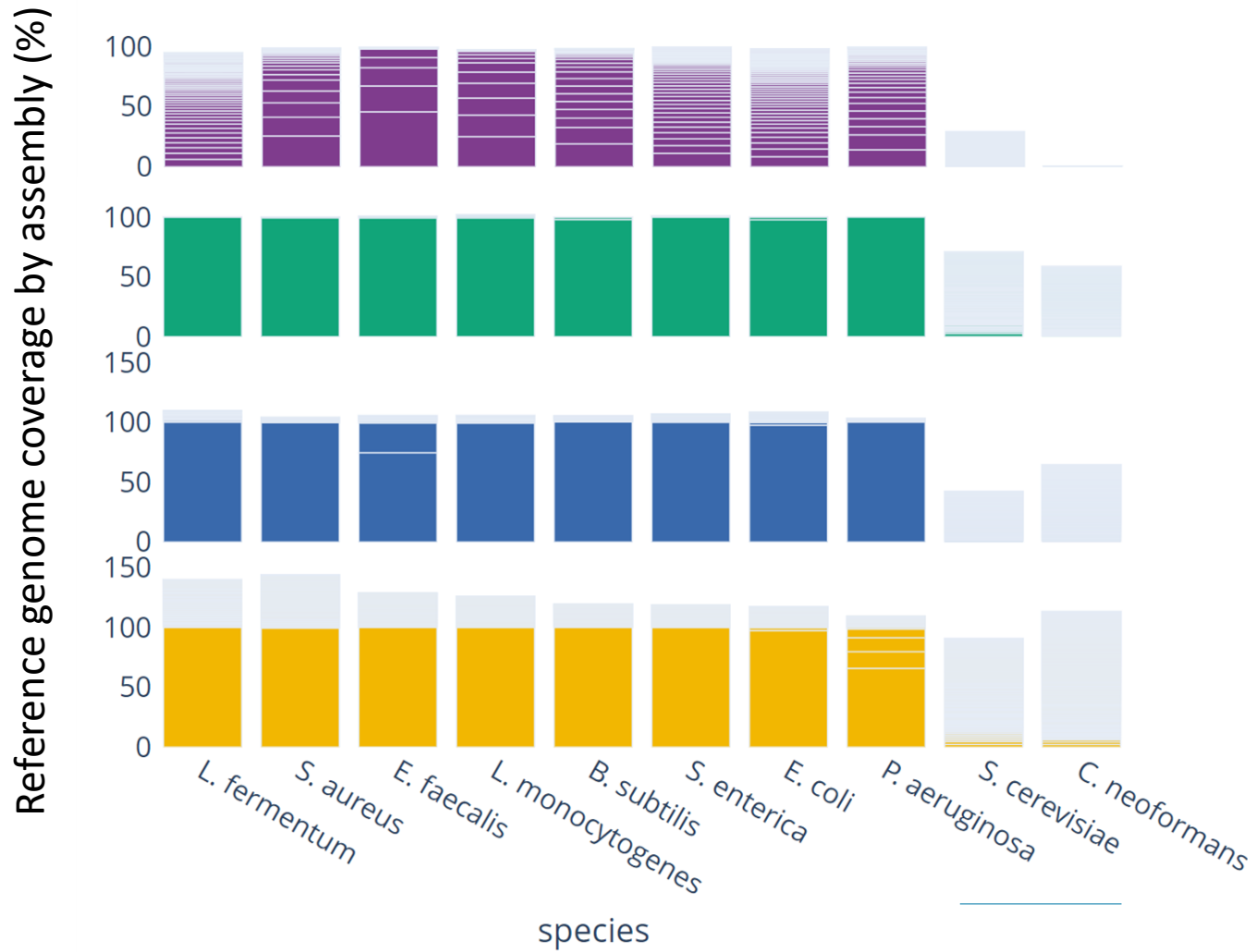
- L'utilisation du séquençage **Longs Fragments** donne une meilleure détermination des espèces présentes
  - 16 S assignation plus précise que V3-V4
  - 16S-23S a priori encore plus précis
- La **représentativité des bases de données** est critique
- L'utilisation de **plusieurs bases de données** peut être une solution
- Suggestion :
  - Faire du **16S-23S**, car ce n'est **pas plus cher** que 16S seul, et l'on peut toujours extraire le 16S seul dans un premier temps
- En terme de couts, c'est quand même plutôt **4 fois plus cher** que les courts fragments (10-20 k lectures par échantillon contre plutôt 30 k en V3-V4)

# Whole metagenomics



# Mocks assembly are very good with HiFi reads

Mock bact



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

Illumina : 68M PE150

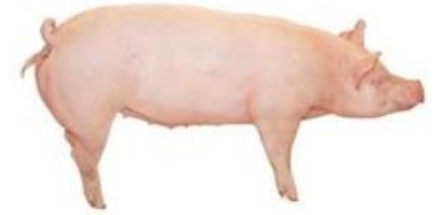
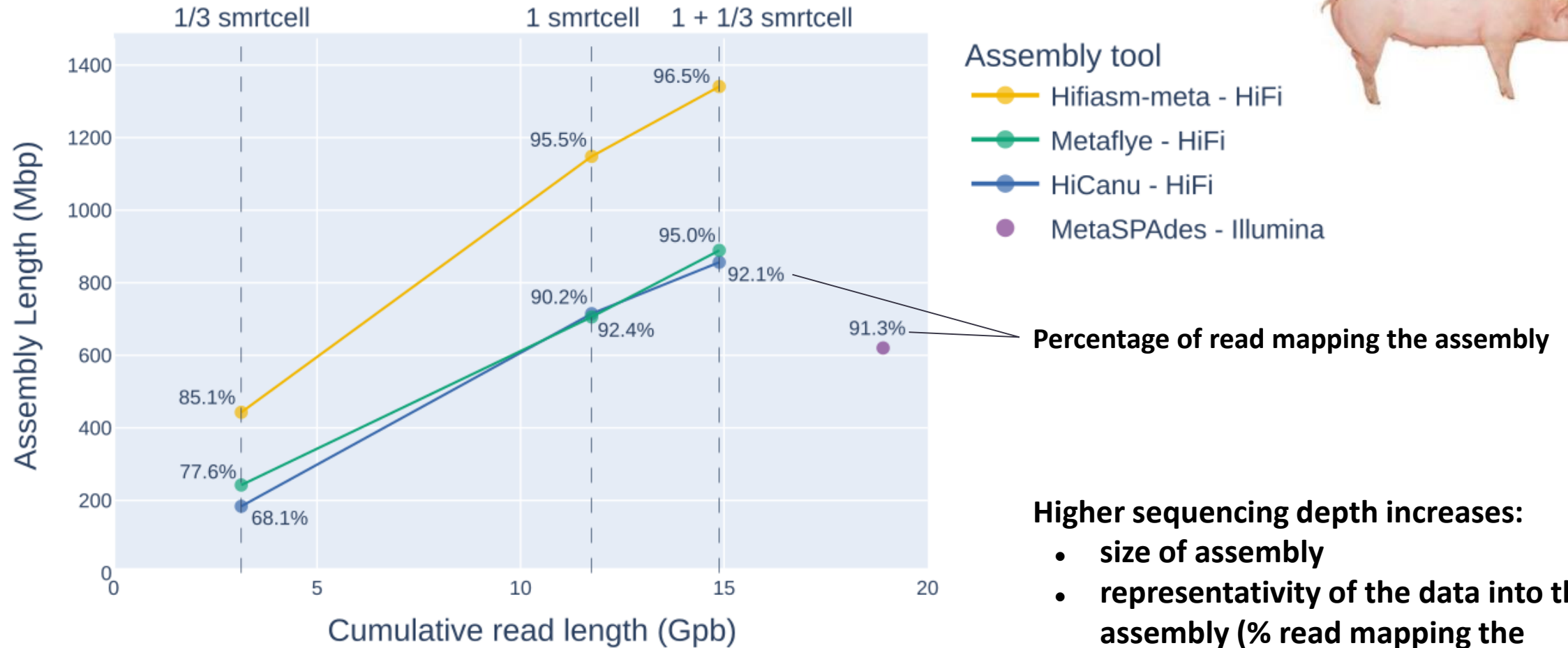
Metaflye

HiCanu

Hifiasm-meta

# Assembly size increases with higher sequencing depth

Feces sample 8



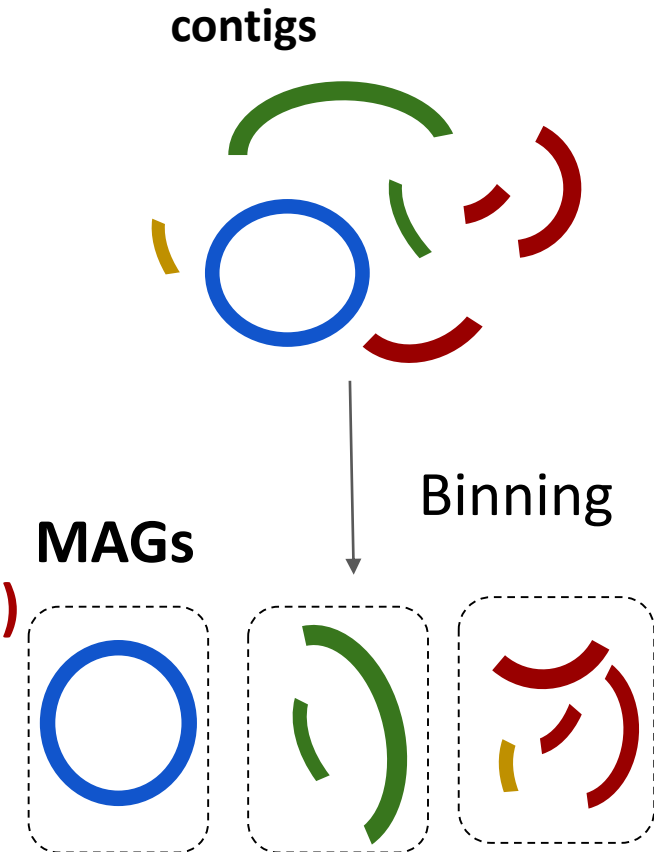
Higher sequencing depth increases:

- size of assembly
- representativity of the data into the assembly (% read mapping the assembly)

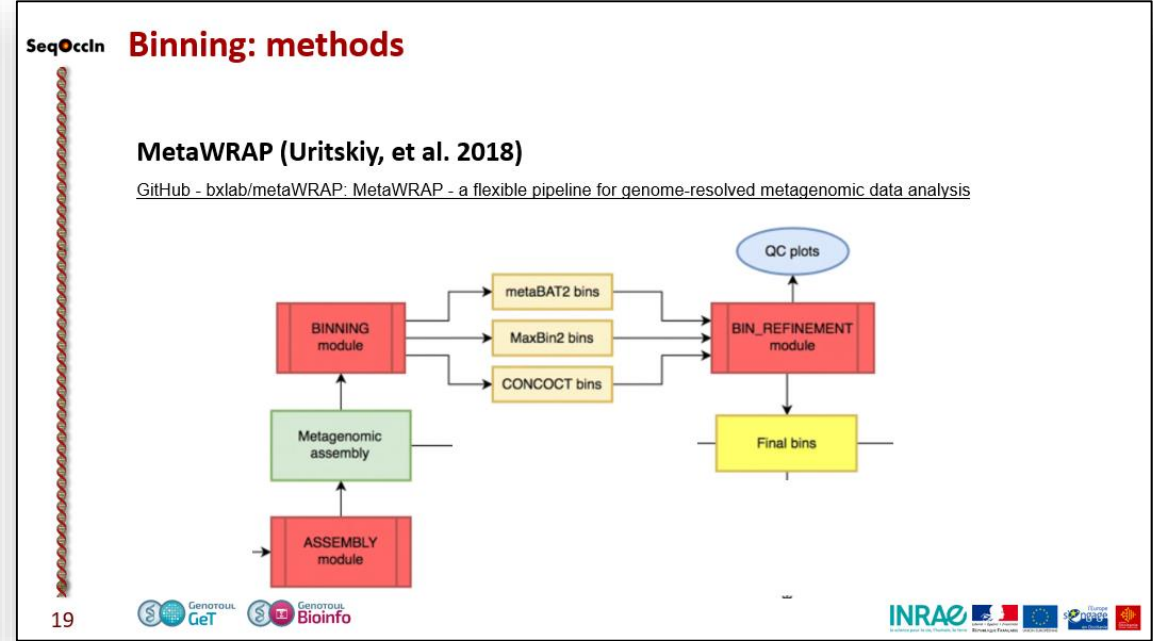
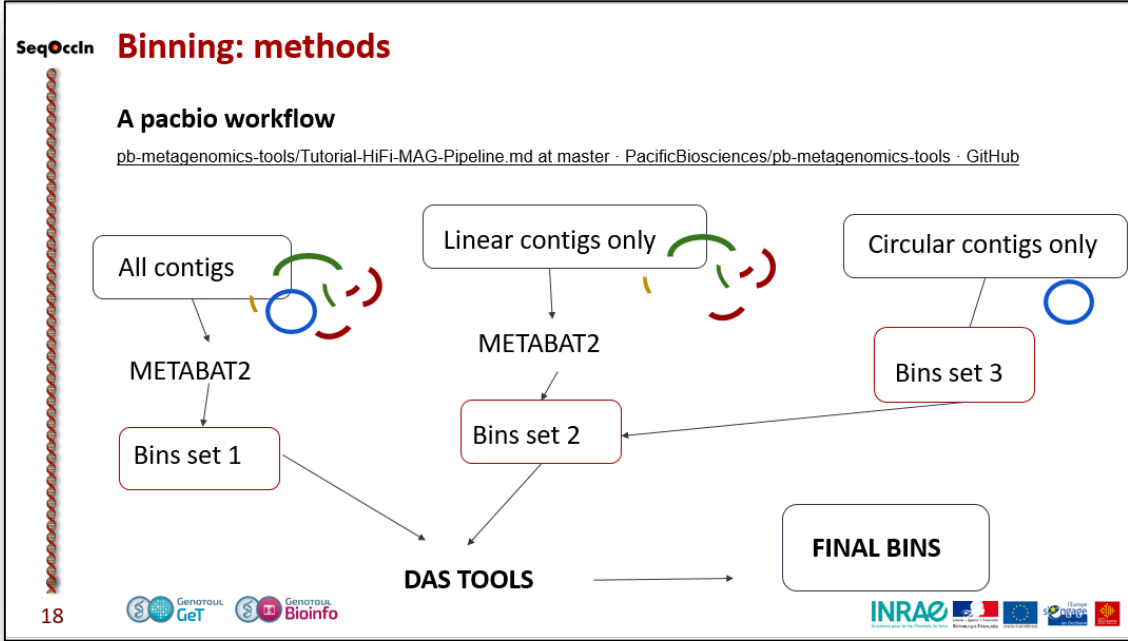
# Main steps of the analysis

*If an assembly of reads is produced*

- Quality check / cleaning / search for potential contamination
- Assembly → contigs
- Genes annotation
- Quantification via reads mapping
- Databank choice (generalist or specific)
- Taxonomic profiles from contigs
- Functional profiles
- **Binning → MAGs (Metagenome-Assembled Genome)**
- Bins taxonomic affiliation

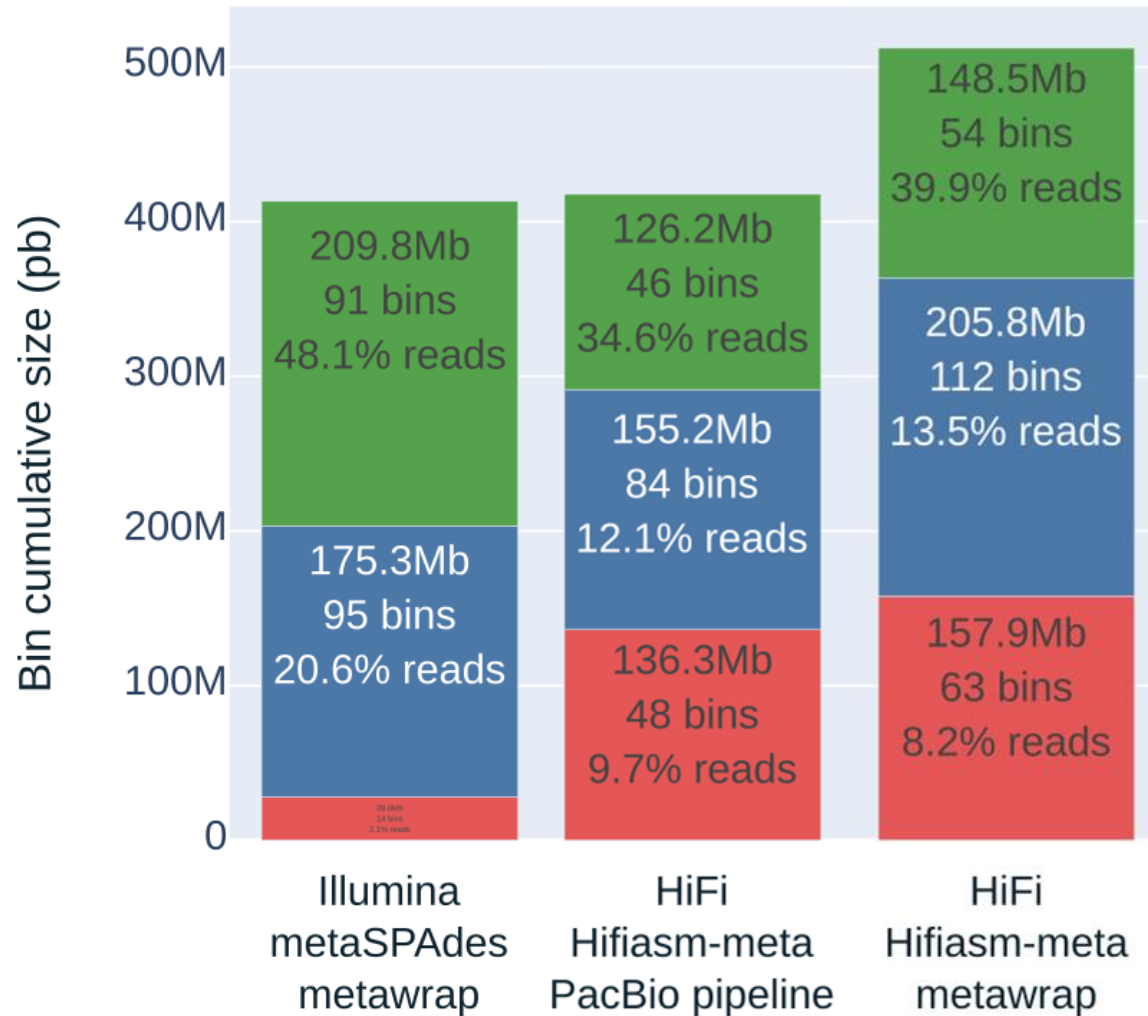


# SeqOocln Two Binning methods used



# Contigs' binning on sample 8 assemblies

Thanks to Andreea Dreau for her help to try to understand results



## Bin quality

- High-quality
- Medium-quality
- High-contamination

Completeness	Contamination
> 90 %	< 5%
> 50 %	< 10%
	> 10 %

**We need to dig further to understand**

# Overall conclusion on the use of HiFi reads in metagenomics

## HiFi reads:

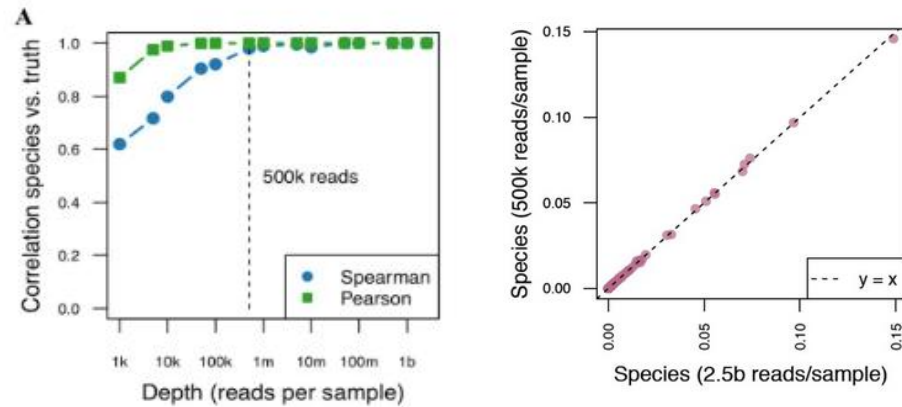
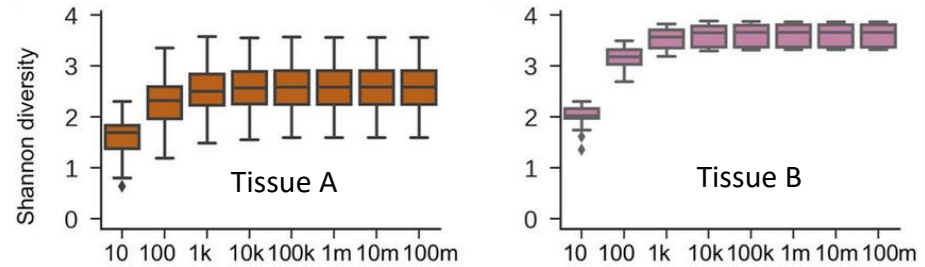
- Allow more contiguous assemblies than short reads.
- Are able to span repeated regions.
- Induce a higher sequencing cost, but provide a high quality capture of the sample diversity.
- On the mock, the taxonomic profile is similar to what is expected.
- The binning with the currently available tools requires improvements.
- *En Illumina plutôt 100 M lectures, en Pacbio plutôt 1 SMRT pour 4M lectures*

**Two slides to end**





Shallow sequencing = sequence sample at a very low coverage



METHODS AND PROTOCOLS  
 Novel Systems Biology Techniques  
 November/December 2018 Volume 3 Issue 6 e00069-18  
<https://doi.org/10.1128/mSystems.00069-18>

## Evaluating the Information Content of Shallow Shotgun Metagenomics

Benjamin Hillmann<sup>a</sup>, Gabriel A. Al-Ghalith<sup>b</sup>, Robin R. Shields-Cutler<sup>c</sup>, Qiyun Zhu<sup>d</sup>, Daryl M. Gohl<sup>e</sup>, Kenneth B. Beckman<sup>a</sup>, Rob Knight<sup>d,f,g</sup>, Dan Knights<sup>a,b,c</sup>

Shallow sequencing seems very promising:

- ✓ Cost
- ✓ Robustness

# Metagenomics Data produced for SeqOccIn project

## Metabarcoding PacBio HiFi:

- ✓ 9 SMRT cells
- ✓ 65 Gb HiFi data

## Shotgun PacBio HiFi:

- ✓ 14 SMRT cells
- ✓ 185 Gb HiFi data

## Shallow Illumina NovaSeq 6000:

- ✓ 1 SP lane
- ✓ 145 Gb data

## Holopig project (WGS HiFi):

- ✓ 12 SMRT cells (19 to be done)

## Magics project (WGS HiFi):

- ✓ 8 SMRTcells (to be done)

## Porcmicrob (Metabarcoding):

- ✓ 360 samples (to be done)



PacBio



illumina

# Partenaires du projet

## SeqOccln

### Coordination

Cécile Donnadiou  
Christine Gaspin  
Carole Iampietro  
Denis Milan



### Axe1 Génomique

Clément Birbes  
Arnaud Di-Franco  
Andreea Dreau  
Camille Eché  
Thomas Faraut  
Carole Iampietro  
Christophe Klopp  
Claire Kuchly  
Camille Marcuzzo  
Amandine Suin  
Matthias Zytnicki

### GenPhySE

Julie Demars  
Cédric Cabau  
Sylvie Combes  
Patrice Dehais  
Thomas Faraut  
Katia Feve  
Nathalie Iannucelli  
Sophie Leroux  
Géraldine Pascal  
Frédérique Pitel



### Axe2 Epigénétique

Remy Félix Serres  
Paul Terzian  
Celine Vandecasteele  
Christophe Klopp

### MIAT

Matthias Zytnicki



### Axe3 Métagénomique

Adrien Castinel  
Jean Mainguy  
Olivier Bouchez  
Géraldine Pascal  
Claire Hoede

### GABI

Didier Boichard  
Mekki Boussaha  
Sébastien Fritz  
Cécile Grohs



### Axe 4 Transversal

Amandine Broha  
Abdias-Archimede Towe-Patipe  
Erwan Denis  
Romain Therville  
Didier Laborie  
Céline Noirod  
Gérald Salin  
Marie-Stéphane Trotard

### Le Moulon

Alain Charcosset  
Johann Joets  
Delphine Madur  
Stéphane Nicolas  
Rémi Séraphin  
Clémentine Vitte



# Remerciements

## SeqOccln

### Coordination

Cécile Donnadiu  
Christine Gaspin  
Carole Iampietro  
Denis Milan

### Plus spécifiquement impliqués dans ces travaux

Adrien Castinel  
Olivier Bouchez  
Carole Iampietro



Jean Mainguy  
Claire Hoede



Sylvie Combes  
Géraldine Pascal

