SeqOccIn
Sequencing Occitanie Innovation

A Long Read project to find optimal technologic combinations for genome assembly and variability, epigenetic marks detection and metagenomic analysis

Iampietro C [1], Eché C [1], Castinel A [1], Serre RF [1], Klopp C [2,3,5], Denis E [1], Bouchez O [1], Kuchly C [1], Vandecasteele C [1], Broha A [1], Therville R [1], Di-Franco A [2,4], Djebali Quelen S [2,4], Dréau A [2,5], Hoede C [2,5], Korovina A [1], Birbes C [1], Laborie D [2], Mainguy J [2], Noirot C [2], Salin G [1], Terzian P [2], Trotard MS [2], Boichard D [6], Boussaha M [6], Grohs C [6], Charcosset A [7], Belcram H [8], Joets J [8], Combes S [4], Pascal G [4], Pitel F [4], Leroux S [4], Riquet J [4], Demars J [4], Tosser-Klopp G [4], Vitte C [7], Ianuuccelli N [4], Lluch J[1], Lopez-Roques C[1], Faraut T [4], Zytnicky M [5], Gaspin C [2,5], Milan D [1,4], Donnadieu C [1]

[1] Plateforme GeT-PlaGe, [2] Plateforme Genotoul Bioinfo (MIAT) [3] Plateforme SIGEN@E (MIAT & GenPhySE), [4] UMR GenPhySE, [5] UR MIAT - INRA Toulouse, Castanet-Tolosan
[6] GABI - INRA Jouy en Josas [7] GQE Le Moulon, [8]ABI Le Moulon - INRA Versailles, Gif-sur-Yvette

The SeqOccIn project (Sequencing Occitanie Innovation), supported by Get-PlaGe and Genotoul Bioinfo platforms, was selected by the Occitanie Region as part of the call for projects "Regional Research and Innovation Platforms". This project should enable us to acquire expertise on the optimal combination of long fragment sequencing technologies and associated applications to better characterize complex genomes in agronomical field: from SNP and structural variations detection, to the production of a high quality assembly at a lower cost. The analysis of native DNA molecules without amplification will allow the detection of some epigenetic marks, and the study of long fragments will allow us to go further on barcoding approaches, as well as on the sequencing of whole metagenomes. Our ambition is to combine and study the contribution of different technologies for three research axes: genome variability analysis, epigenetic mark analysis and metagenome.

The project benefits from the contributions of public research units "Genetics, Physiology and Livestock Systems (GenPhySE)", "Mathematics and Applied Informatics (MIAT)", "Animal Genetics and Integrative Biology (GABI)" and Quantitative Genetics and Evolution (GQE) and also 15 private partners. The 3-years project (2019-2021) will finance 12 people and the acquisition of expertise through the production and analysis of data sets of interest for public and private.

Axis 1

Genome variability analysis

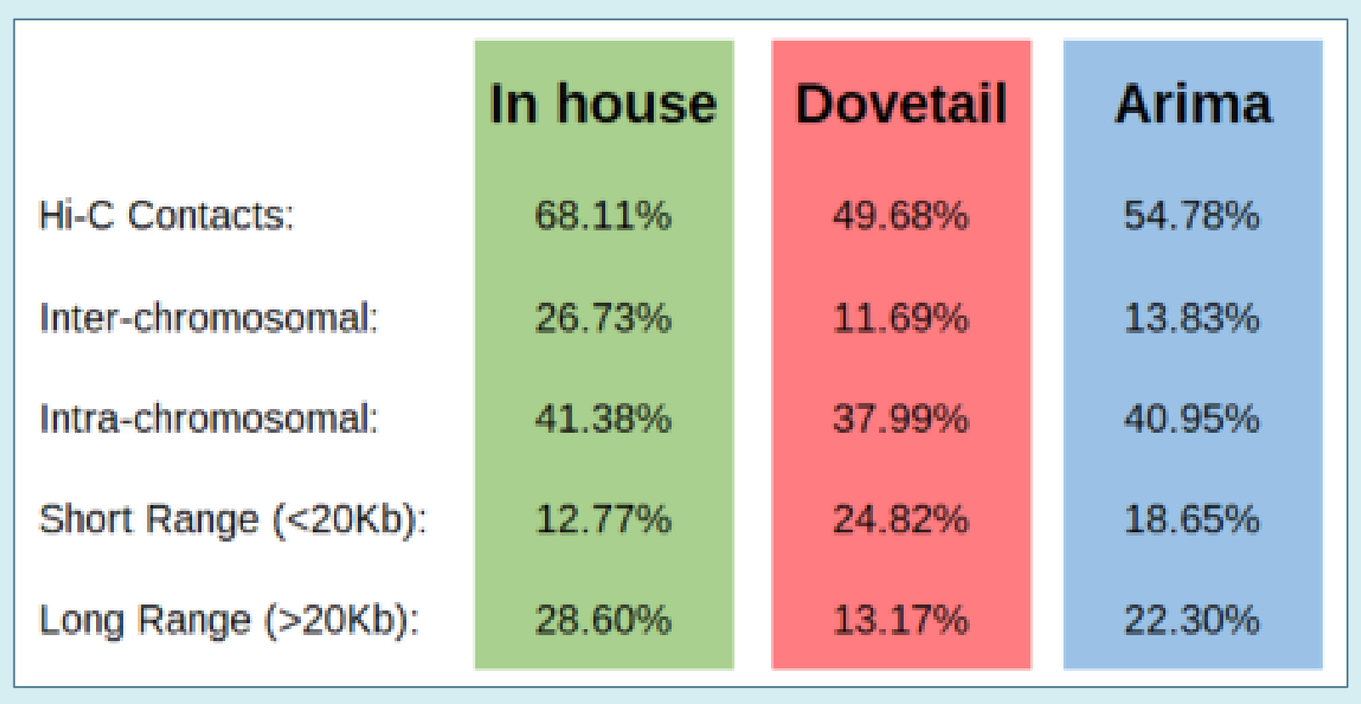
Validation phase: Study the potential and complementarity of different long read technologies, in addition to short read. Determine the best combination of technologies according to the objective pursued. Develop new methods for the bioinformatic analysis of these data.
After the first phase, we will validate the combinations taking into account the different applications on a larger number of individuals. Complementary developments will be carried out for genotyping targeted on regions of interest.

Technologies used:
- Illumina : mate pair and PCR free libraries
- 10X Genomics : Chromium libraries (« synthetic long read »)
- Oxford Nanopore longs reads
- Hi-C libraries
- Pacific Bioscience longs reads
- Bionano optical mapping
- Other future technologies….

Samples selected for the validation phase:
Cattle: selection of trios for haplotypes and structural variants identification.
Maize: lines for which there is already a quality assembly carried out by NRGene.

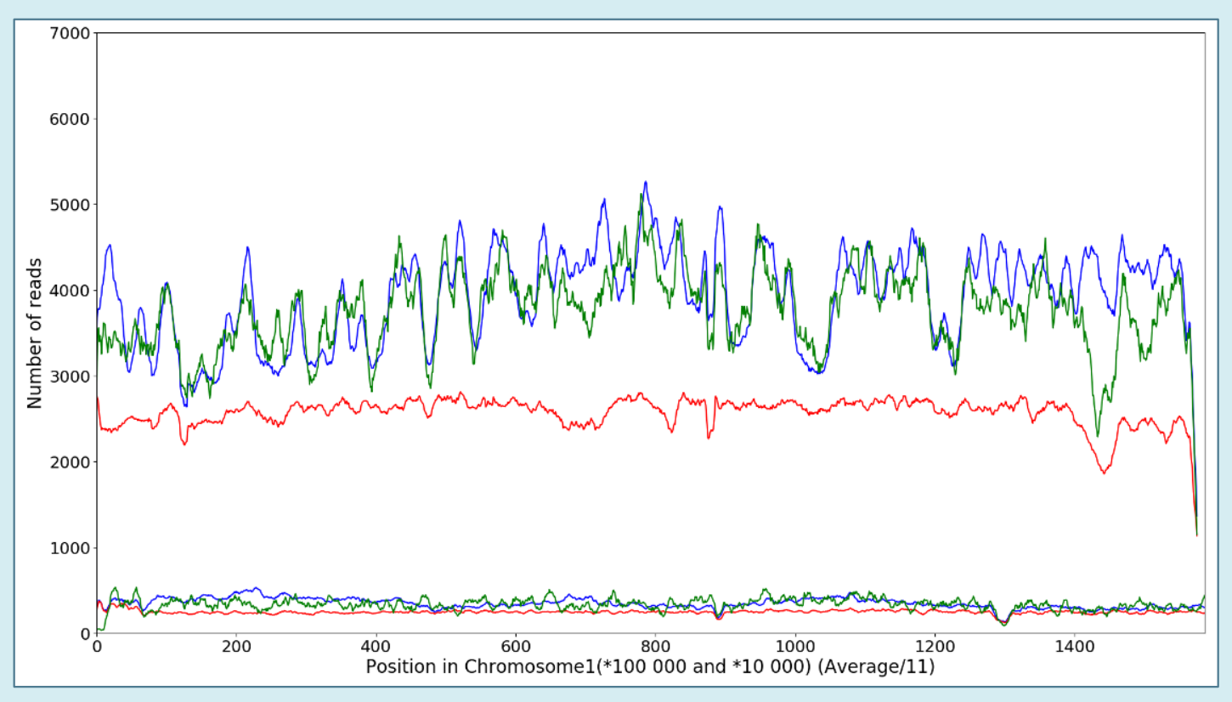
Scientific issues :
- Characterization and functioning of the genome**,** representation of genetic variability
- Identification of structural variants present in genomes: deletions, duplications, copy number variation, insertions, and chromosomal translocations.
- Detection of causal mutations
- Analysis of pangenome that is not present in the reference assembly
- Search for optimized methods for the complete assembly of heterozygous genomes using heterogeneous data

Preliminary Results : testing Bos taurus on several HI-C protocols



	In house	Dovetail	Arima
Hi-C Contacts:	68.11%	49.68%	54.78%
Inter-chromosomal:	26.73%	11.69%	13.83%
Intra-chromosomal:	41.38%	37.99%	40.95%
Short Range (<20Kb):	12.77%	24.82%	18.65%
Long Range (>20Kb):	28.60%	13.17%	22.30%

Juicer Statistics



Reads distribution on chromosome 1 (130M paired reads)

Axis 2

Epigenetic marks analysis

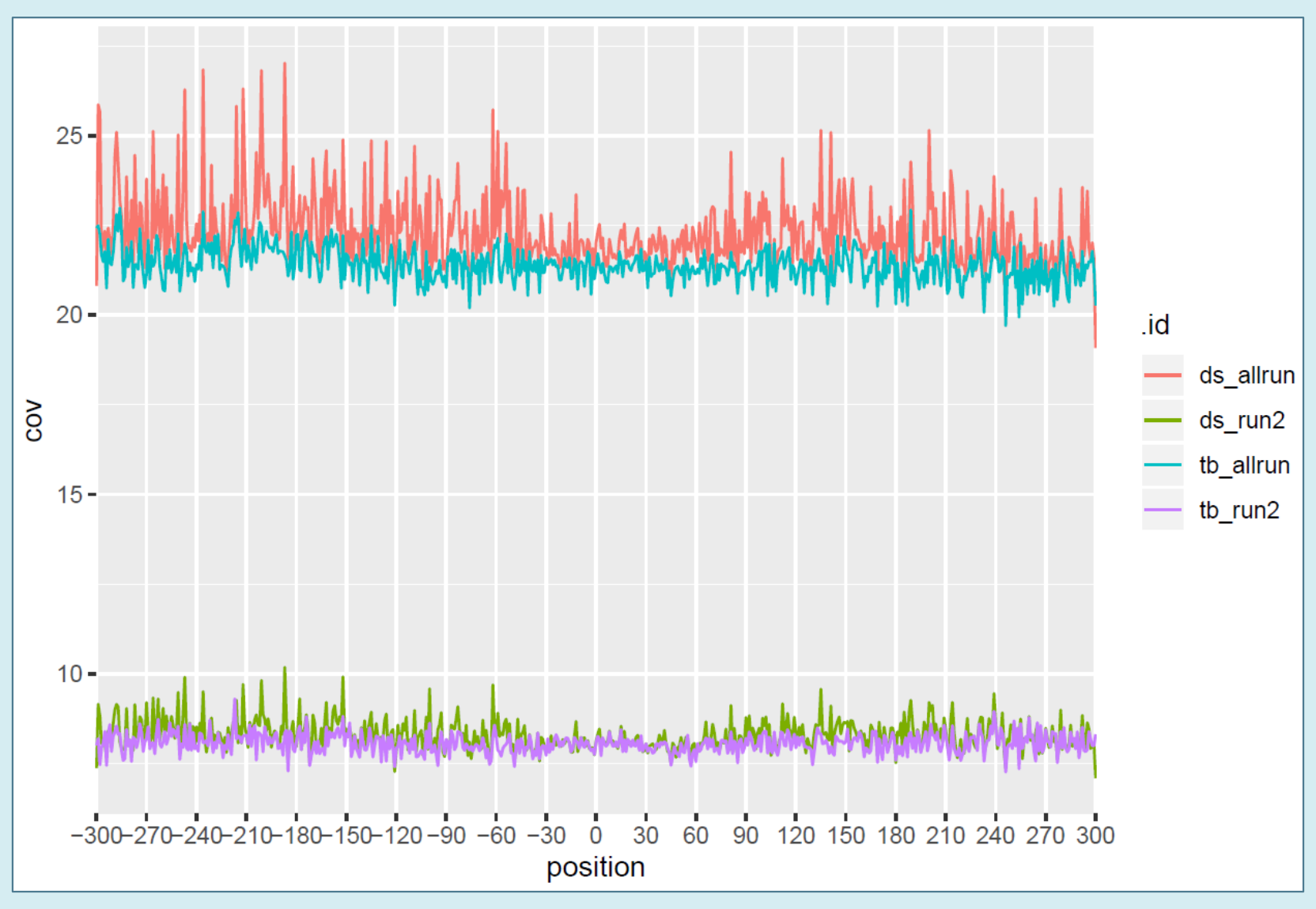
Validation phase: Detection of DNA methylation marks, 5mC, CPG block context, 6mA.
After the validation phase, we will identify an area of particular interest for which we will seek to analyze the methylation on a large number of individuals. We will continue to detect new markers as the models evolve.
Oxford Nanopore sequencers also allow direct RNA sequencing, and detection of 5mC methylations. This project will test and evaluate protocols for RNA methylation detection.

Technologies used:
- Illumina: after bisulphite treatment
- Oxford Nanopore: detection of methylated DNA markings

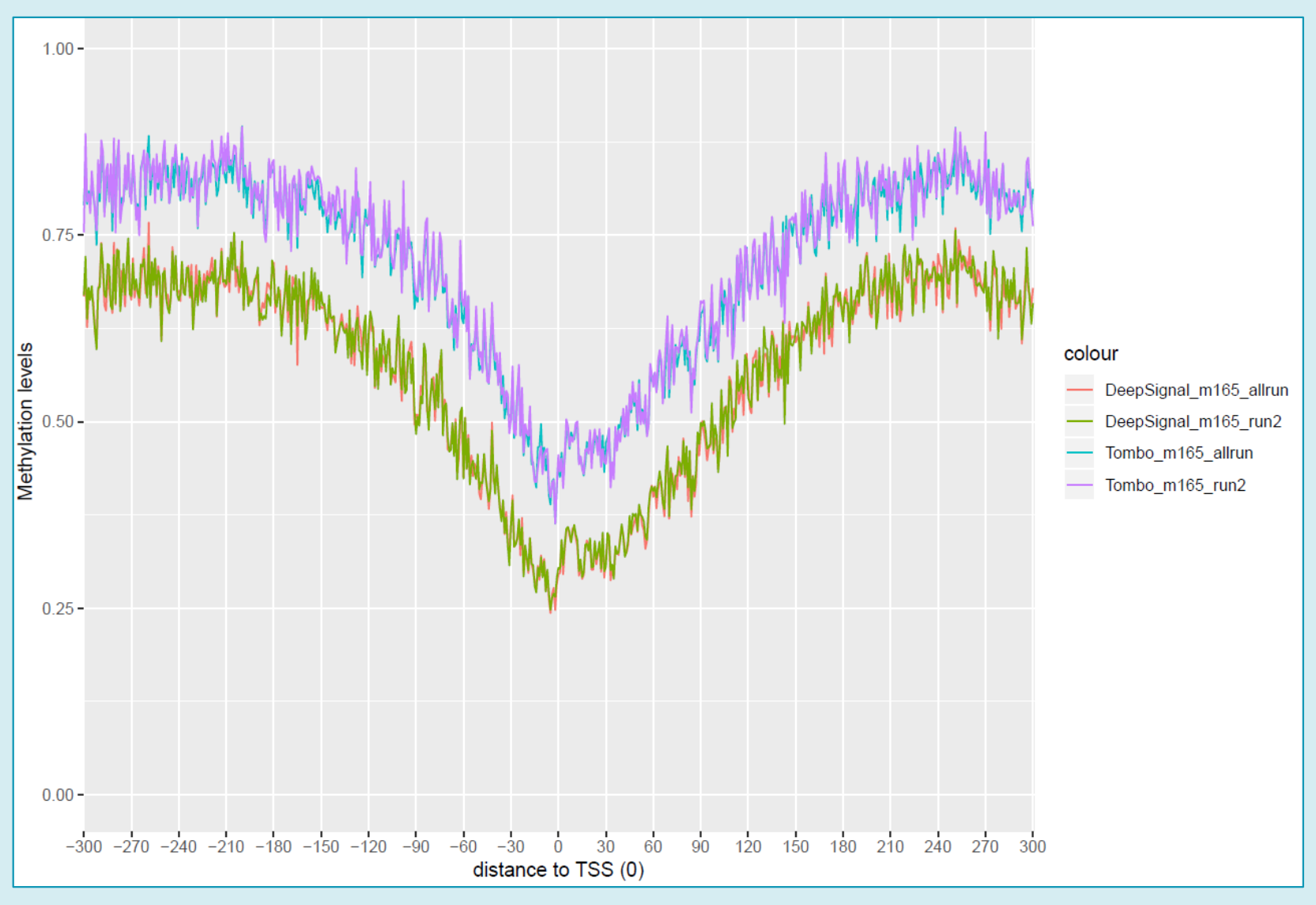
Samples selected for the validation phase :
Quail: its small genome will be used as reference genome for the study of the methylation marks 5mC and 6mA.
Pork: study of methylation marks and comparison on a slightly larger set (proof of principle)

Scientific issues :
- Is there a genetic heredity in quail?
- Study of parental genetic fingerprinting in pork: identifying and characterizing fingerprinting genes in order to be use in selection

Preliminary Results : Detection CpG - Bos taurus - Chrmosome 1



Average coverage around TSS



Average methylation ratio around TSS

Axis 3

Metagenomes analysis

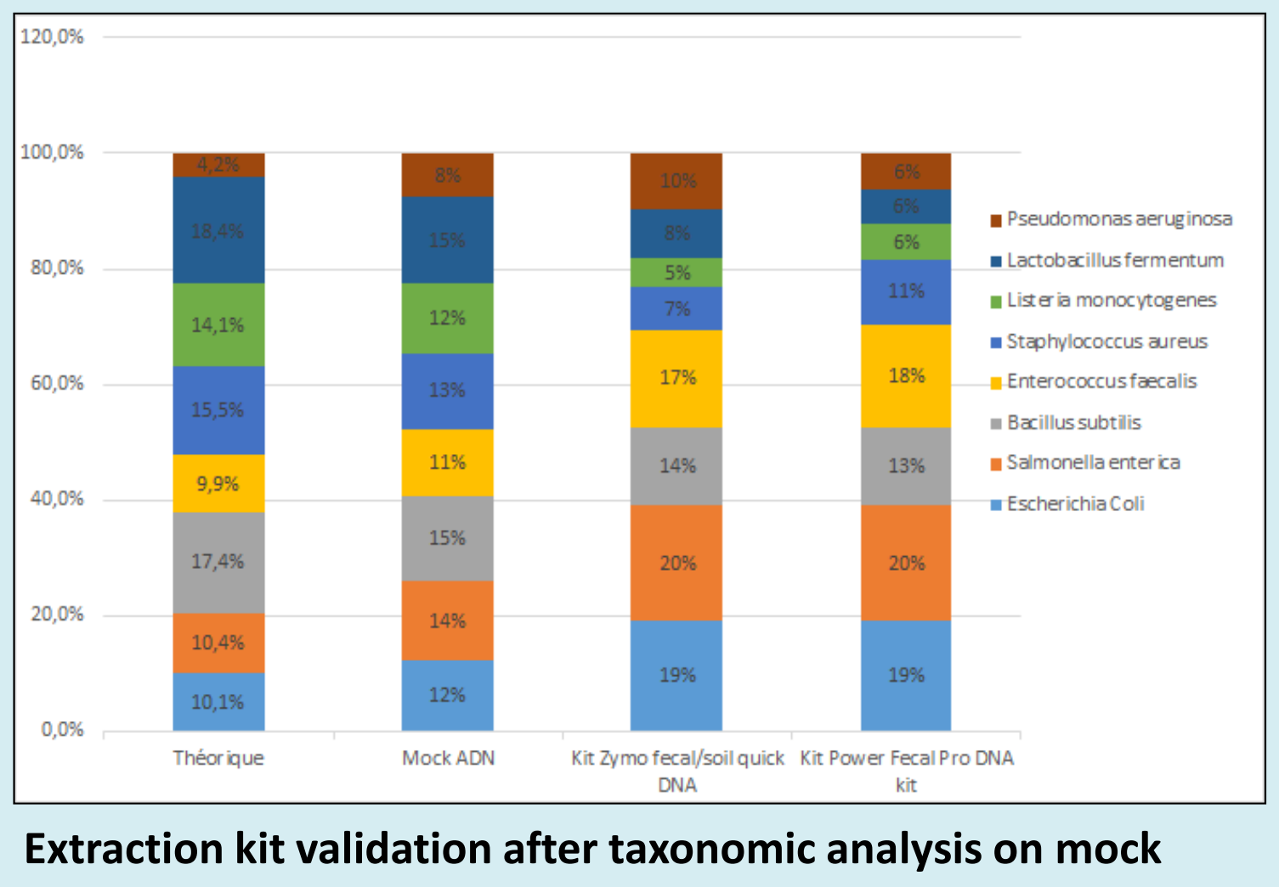
Validation phase: Development of new protocols for the targeted approach by barcoding & Whole Meta Genome, development of new methods for the bioinformatic analysis of these data
Initially, an in silico approach will be carried out in order to identify new biodiversity markers that can be used in different kingdoms: 16S, 18S fragments, rpoB, ITS. The taxonomic resolution of the selected biomarker will be tested in real condition.
In parallel, we will test different protocols in order to achieve the best assemblies and the best identification of the strains present in the environmental samples.

Technologies used:
- Illumina
- Oxford Nanopore longs reads
- 10X Genomics
- Pacific Bioscience longs reads
- Librairies Hi-C

Samples selected for the validation phase :
Faeces and Rumen

Scientific issues :
- Who's there? How much is there?
- Who's doing what? What community dynamics?

Preliminary Results :

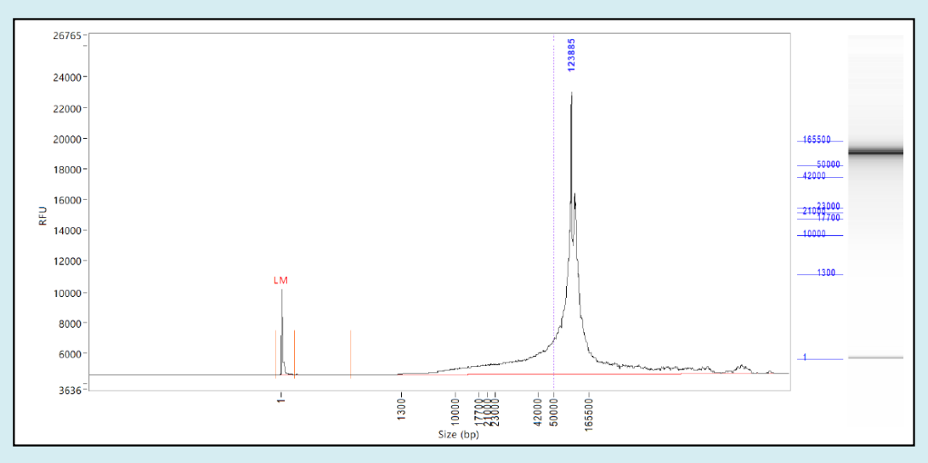


Extraction kit validation after taxonomic analysis on mock community using Miseq sequencing of 16S V3-V4 region

Transverse axis

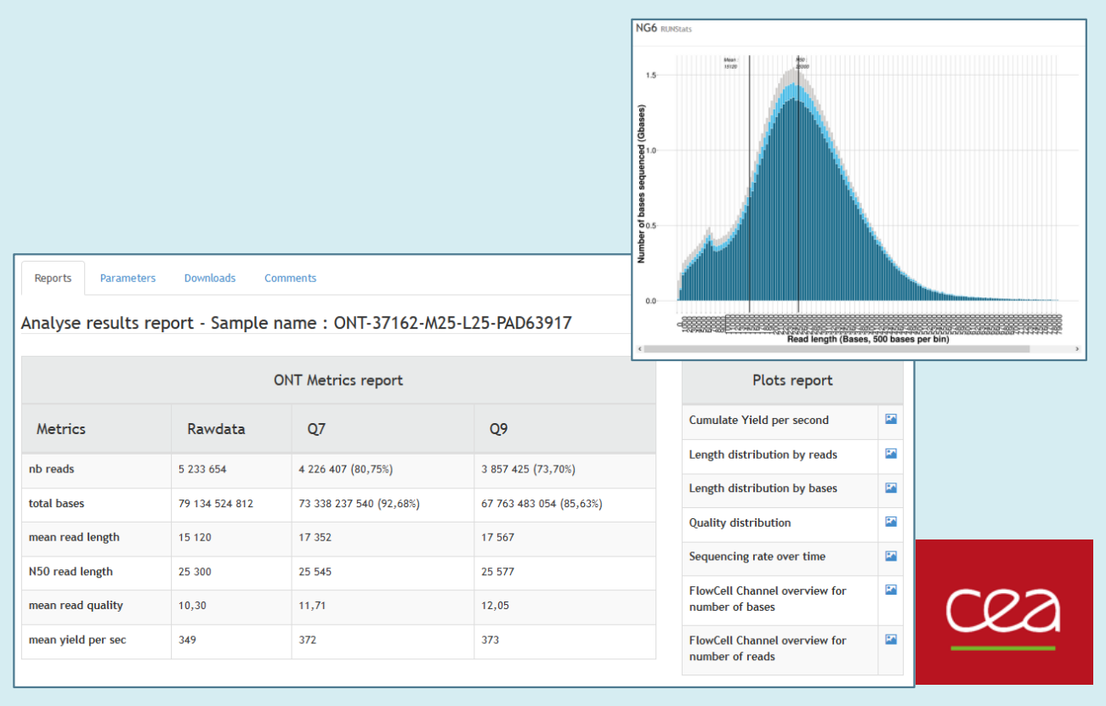
High molecular weight DNA extraction & Evolution of the IT infrastructure

Preparation of high molecular weight DNA
The high molecular weight DNA preparation is an essential step in the project realization. This project allows the evaluation of different extraction protocols on different tissues with the final objective of automating the preparation of high molecular weight DNA. The collaboration with Adelis could lead to the development of an automaton.



Femto results on Bos taurus Promega extraction

IT developments
The evolution of platform infrastructure goes hand in hand with the increasing amounts of data generated by the latest generations of sequencers. This project allows the implementation of a data management and sharing plan (open data) as well as new tools for: sample traceability, data storage, processing and availability.


Analyse results report - Sample name : ONT-37162-M25-L25-PAD63917
ONT metrics report
Plots report

Results visualization on our web interface NG6

Private partners

25 private partners are involved in the project: the actors of bovine and caprine breeding are grouped with **Apis-Gene**, the pork breeders with **Alliance R&D**, for sheep and poultry species **CNBL**, **SYSAAF** and **Adisseo** and **Lallemand** for metagenomics. The project also includes a plant component with the private partners of the **AMAIZING** project (**Biogemma, MaïsAdour, KWS, Caussade Semences, Limagrain, RAGT, Euralis, Syngenta**). **Adelis** is in partnership with the SeqOccIn project to develop a high molecular weight DNA extraction automated system.

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
INRAE la science pour la vie, l'humain, la terre
UNION EUROPÉENNE
L'Europe s'engage en Occitanie
La Région Occitanie Pyrénées - Méditerranée