



**HAL**  
open science

## Improvement of the *Bos taurus* Genome Assembly using New Sequencing Technologies

Camille Eché, Carole Iampietro, Clement Birbes, Andreea Dreau, Claire Kuchly, Arnaud Di Franco, Christophe Klopp, Thomas Faraut, Sarah Djebali, Adrien Castinel, et al.

### ► To cite this version:

Camille Eché, Carole Iampietro, Clement Birbes, Andreea Dreau, Claire Kuchly, et al.. Improvement of the *Bos taurus* Genome Assembly using New Sequencing Technologies. Discoveries Roadshow 2023 - PacBio, PacBio, May 2023, Paris, France. hal-04096124

**HAL Id: hal-04096124**

**<https://hal.inrae.fr/hal-04096124v1>**

Submitted on 12 May 2023

**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.



Distributed under a Creative Commons Attribution 4.0 International License

# Improvement of the *Bos taurus* Genome Assembly using New Sequencing Technologies

*Camille Eche, Carole Lampietro, Clement Birbes, Andreea Dreau, Claire Kuchly, Arnaud Di Franco, Christophe Klopp, Thomas Faraut, Sarah Djebali, Adrien Castinel, Matthias Zytnicki, Erwan Denis, Mekki Boussaha, Cecile Grohs, Didier Boichard, Christine Gaspin, Denis Milan, and Cecile Donnadieu*

## Acquire advanced expertise on the **new high throughput sequencing technologies available**

- Comparative potentials of technologies
- Identification of combinations of technologies to be implemented according to the objectives



### ... in four axes:

**The Genome.** Genome assembly and variant detection

**The Epigenome.** Studies of epigenetic marks that regulate gene expression

**The Metagenomes.** In-depth knowledge of communities

**Data Management.** High molecular weight DNA extraction and evolution of the IT infrastructure



#### ***Bos taurus***

- 2.7 Gb genome size
- 30 Chromosomes
- 40 % Repeats

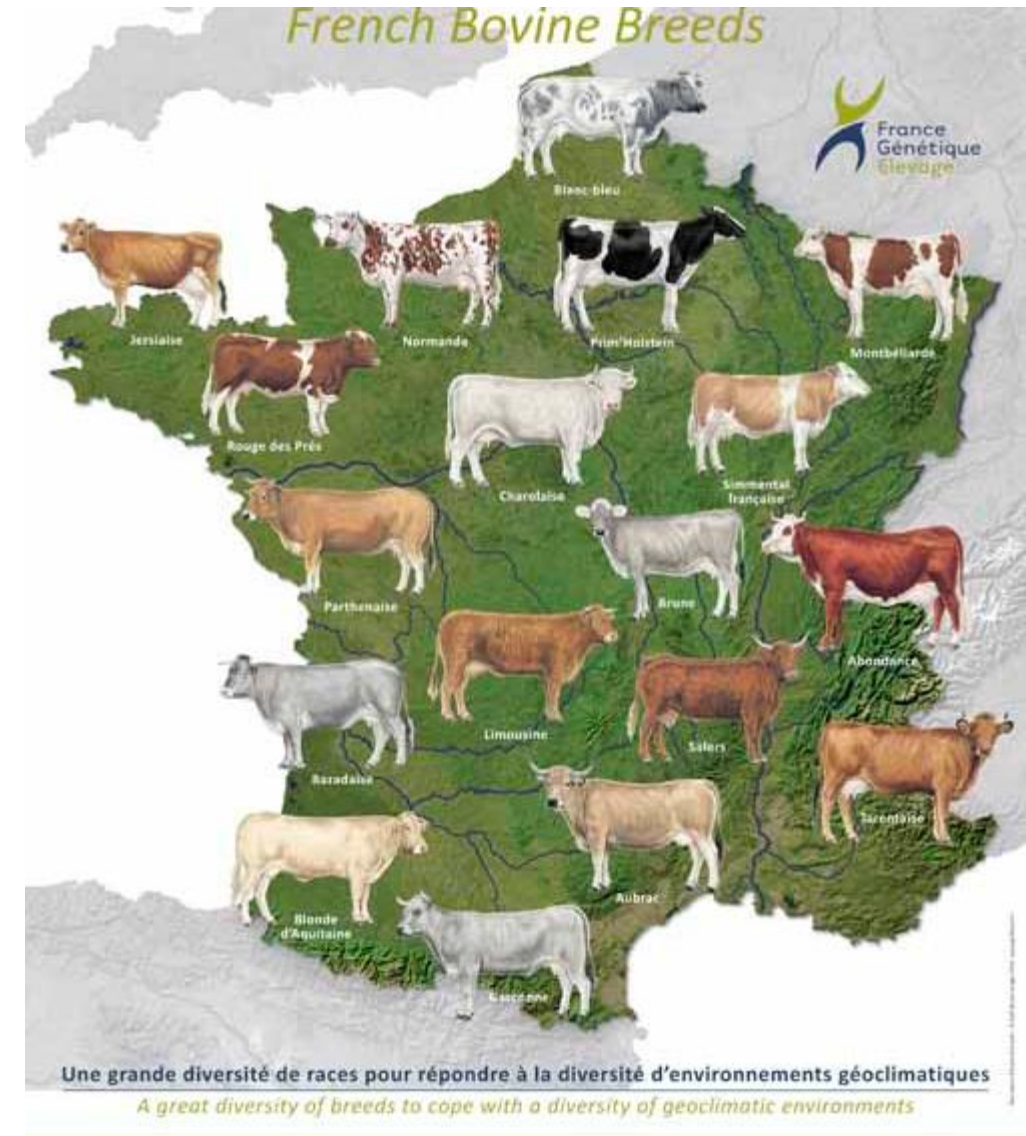


#### ***Zea mays***

- 2.4 Gb genome size
- 10 Chromosomes
- **quasi homozygous**
- **85 % Repeats**

# Why assemble a Charolais breed Bos taurus Genome?

- Leading suckling cattle breed in Europe (25% of the total cows)
- International extension
- Developed on grazing and extensive production systems
- Excellent maternal qualities with high growth potential and an excellent beef
- The specificities of its genome are poorly known






# The reference genome ARS-UCD1.2

- Submitted by USDA Ars on April 2018
- Line-bred Hereford cow who was selected for her high level of inbreeding.

ARS-UCD1.2		
Assemblage	Data type	CLR
	Quantity	80X
	Number of contigs	3 077
	Total size	2 700 000 000
	N50 contigs length	12 000 000
Scaffolding	Data type	Hi-C / Optical + Recombination map
	Quantity	84X Hi-C
	Number of scaffolds	2 211
	Total size	2 715 853 794
	N50 scaffolds length	103 308 737
	BUSCO	C:95.8%



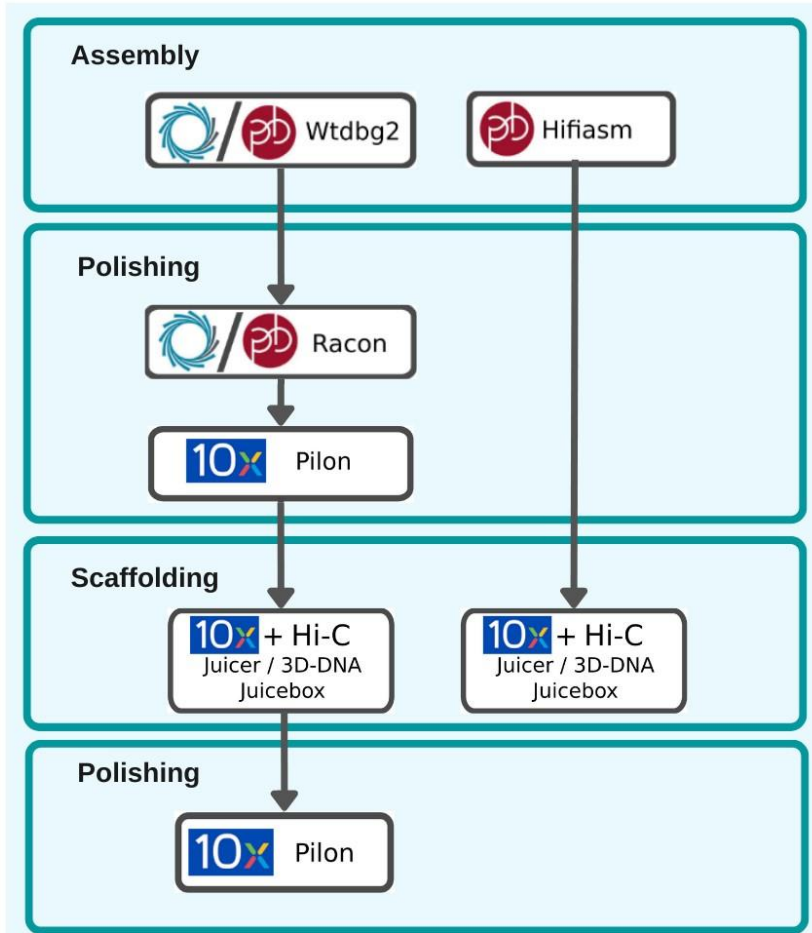
# Technologies used for the study

1 Sample	2 Sequencing technology	3 Library type
 <b>Father</b>	Oxford Nanopore Illumina	Ligation sequencing gDNA 10X Genomics Chromium Hi-C
 <b>Mother</b>	Oxford Nanopore Illumina PacBio	Ligation sequencing gDNA 10X Genomics Chromium Hi-C Circular Long Read
 <b>Heifer</b>	Oxford Nanopore Illumina PacBio	Ligation sequencing gDNA 10X Genomics Chromium PCR Free 2x250pb Hi-C Circular Long Read Consensus Long Read

Create reference datasets for

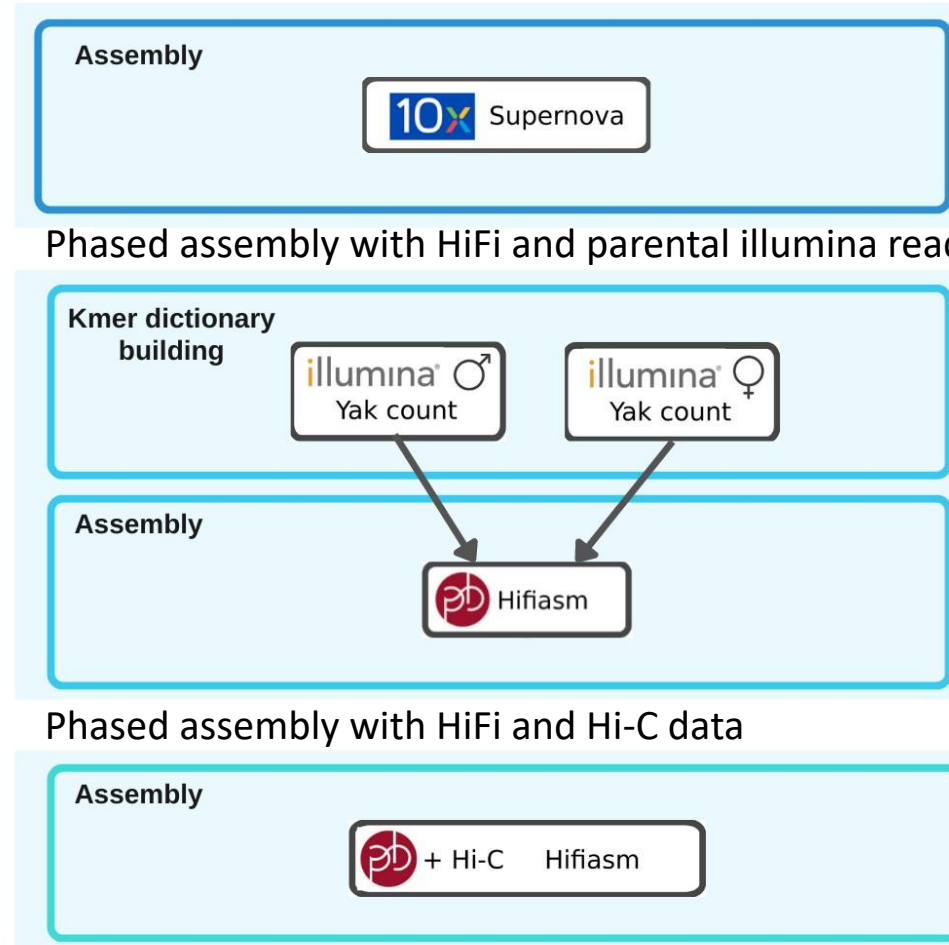
- Genome assembly
- Haplotyping
- Variability discovery

# Methods and pipelines used



Long reads assemblies from ONT and CLR PacBio data

## 10X Chromium assembly



nextflow

<https://forgemia.inra.fr/seqoccln/>

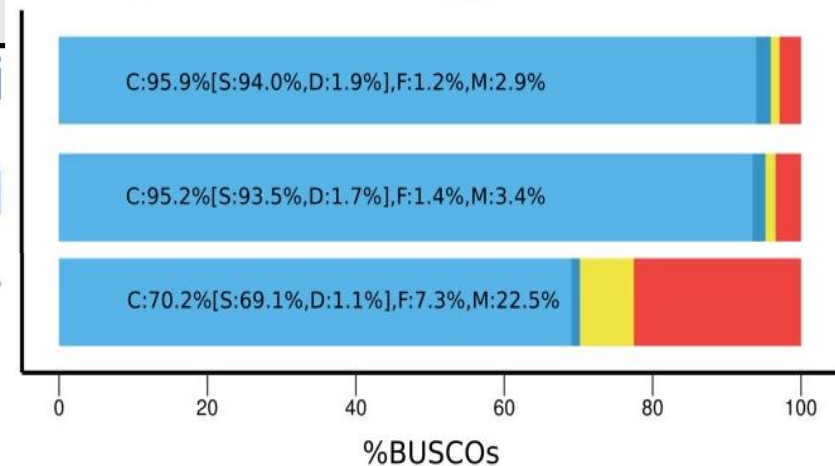
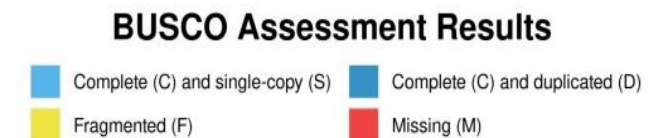
# Assembly metrics

	ARS-UCD1.2	ONT	HiFi
Assemblage	Data type	CLR	HiFi CCS
	Quantity	80X	58X
	Number of contigs	3 077	7 226
	Total size	2 700 000 000	2 701 288 401
	N50 contigs length	12 000 000	23 641 545
Scaffolding ( after polishing )	Data type	Hi-C / Optical + Recombination map	Hi-C / 10X
	Quantity	84X Hi-C	28X / 95X
	Number of scaffolds	2 211	4 600
	Total size	2 715 853 794	2 705 347 253
	N50 scaffolds length	103 308 737	100 959 810
	BUSCO	C:95.8%	C:95.2%

HiFi

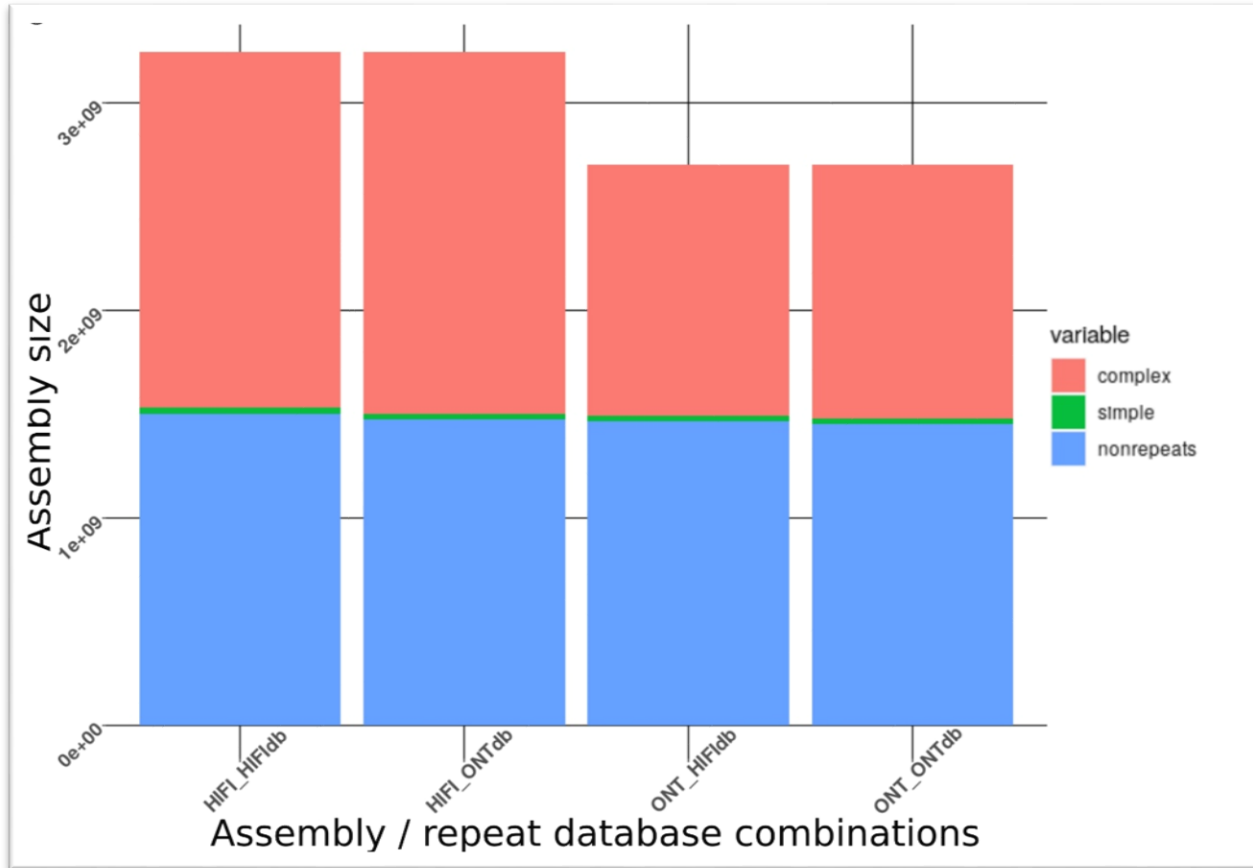
ONT polished

ONT

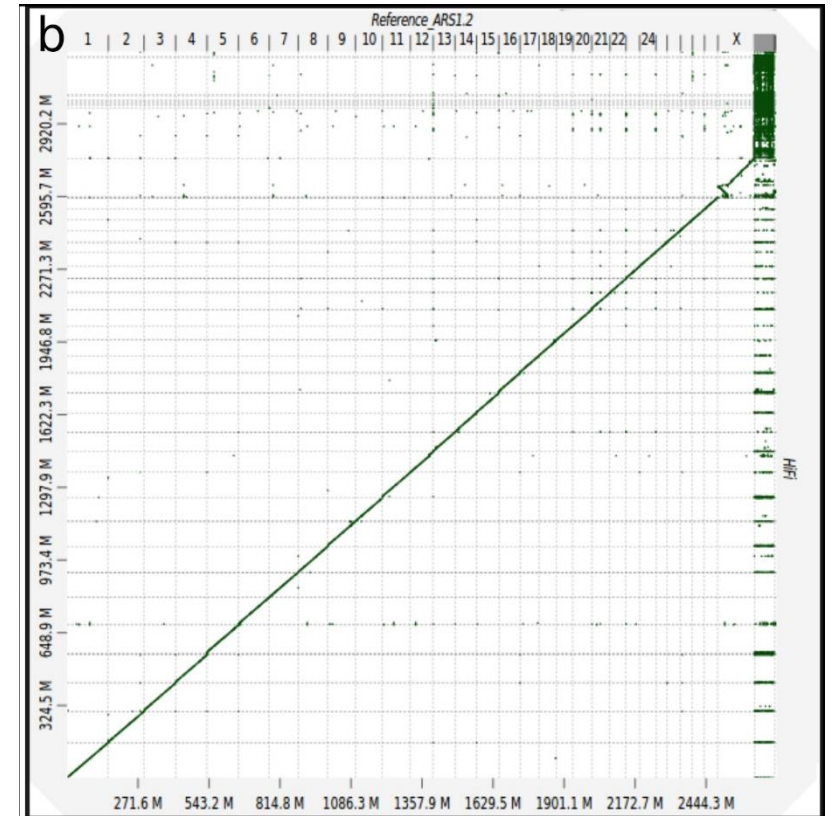




# Charolais reference vs ARS reference



RepeatMasker / RepeatModeler representation of HiFi assembly and ONT assembly.



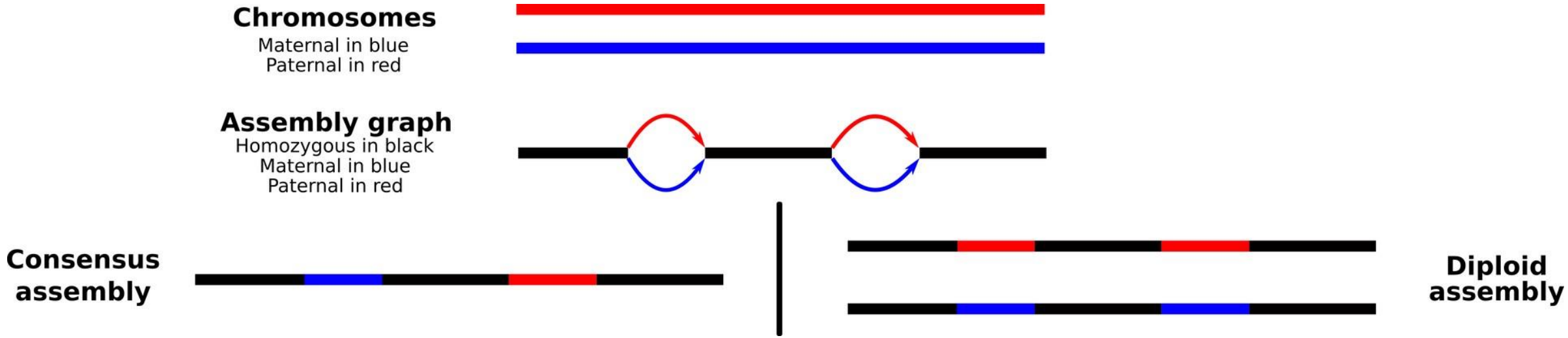
Dgenies Dot-Plot of HiFi Final assembly against *Bos taurus* reference ARS-UCD1.2

The additional information in the HiFi assembly is mainly complex duplications.

# Phased assembly



A diploid assembly is an assembly in which the maternal and paternal haplotypes are separated to create 2 sets of chromosomes.



- Different methods exist to create a diploid assembly:
- 1 Method using distant connection to separate haplotypes: Hi-C
  - 1 Method resolving haplotype based on parental K-mers: Yak

# Phased assembly



	ARS-UCD1.2	HiFi hifiasm parent hap1	HiFi hifiasm parent hap2	HiFi hifiasm Hi-C hap1	HiFi hifiasm Hi-C hap2
Data type	CLR	CCS + Trio	CCS + Trio	CCS + Hi-C	CCS + Hi-C
Quantity	80X	40X	40X	40X + 28X	40X + 28X
Number of contigs	3 077	2 871	2 300	2 685	2 136
Total size	2 700 000 000	3 156 028 877	3 113 483 345	3 0177 978 241	3 184 033 110
N50 contigs length	12 000 000	71 619 842	69 165 538	80 106 842	71 644 334
BUSCO	C:95.8%	C:95.8%	C:95.3%	C:95.8%	C:95.7%
Phasing ratio	*	97.3%	96.7%	62.6%	60.5%
Contigs phasing ratio	*	97.5%	96.9%	84.6%	85.6%

Hi-C separation is less efficient than parental separation but works fine on a contig level.

# Assembly production

229 High quality assemblies generated

## - Sus scrofa :

- 5 CLR assemblies
- 4 HiFi assemblies

## - Bos Taurus :

- 6 ONT assemblies
- 9 CLR assemblies
- 154 CLR assemblies for variant
- 5 HiFi phased assemblies

## - Capra Hircus :

- 7 CLR assemblies
- 1 HiFi phased assembly

## - Ovis aries :

- 10 CLR assemblies
- 1 HiFi phased assembly

## - Coturnix Japonica :

- 1 HiFi assembly
- 1 ONT assembly

## - Zea Mays :

- 25 HiFi assemblies

# Conclusions and upcoming work

## ➤ *Improvement of the *Bos taurus* genome*

- ✓ One high quality consensus genome GCA\_947034695.1 significantly larger than ARS-UCD1.2 -
- ✓ Two haplotyped trio haplotyped trio high quality reference genome
- ✓ Contribution to the bovine pangenome for the Charolais breed

## ➤ *Production of whole genome LONG READ sequences for ~150 animals corresponding to several breeds*

- ✓ Study **structural variations** at the whole genome level
- ✓ Construction of several genome assemblies (corresponding to several breeds) → **study the pangenome**

## ➤ *Production of both long and short reads for several trios*

- ✓ Construction of genome references using long read data
- ✓ Construction of haplotype graphs using the trio-binning approach
  - Construct several breed specific haplotype graphs
  - **study the pangenome**

Breed	Number of animals
HOL	25
MON	25
NOR	25
BSW	5
SIM	5
ABO	10
TAR	5
VOS	4
BLA	10
CHA	10
LIM	10
AUB	10
FLA	3
PAR	3
<b>Total</b>	<b>150</b>

Trios	Breeds
<b>Trio 1</b>	<b>CHA</b>
<b>Trio 2</b>	<b>CHA</b>
Trio 3	HOL x NMD
Trio 4	YAK x MON
Trio 5	AUB
Trio 6	BAQ
<b>Trio 7</b>	<b>ABO</b>
<b>Trio 8</b>	<b>TAR</b>
<b>Trio 9</b>	<b>VOS</b>

# Conclusion

nextflow

forgemia.inra.fr/seqoccln

SCIENTIFIC  
DATA

Data paper accepted



data.gouv.fr

<https://entrepot.recherche.data.gouv.fr/dataverse/seqoccln>



<https://github.com/GeTPlaGe/SeqOccln>



<https://www.ebi.ac.uk/ena/browser/view/PRJEB60075>

**Thanks !**

# Project partners

## SeqOccln

### Coordination

Cécile Donnadiou  
Christine Gaspin  
Carole Iampietro  
Denis Milan

### Axe1 Génomique

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

### Axe2 Epigénétique

Remy Félix Serres  
Paul Terzian  
Celine Vandecasteele  
Christophe Klopp

### Axe3 Métagénomique

Adrien Castinel  
Jean Mainguy  
Olivier Bouchez  
Géraldine Pascal  
Claire Hoede

### Axe 4 Transversal

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



### GenPhySE

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

### MIAT

Matthias Zytnicki

### GABI

Didier Boichard  
Mekki Boussaha  
Sébastien Fritz  
Cécile Grohs  
Aurelien Capitan

### Le Moulon

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

