A Bos taurus sequencing methods benchmark for assembly, haplotyping, and variant calling

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

Bos taurus sequencing methods benchmark for assembly, haplotyping, and variant calling

SeqOccIn — Sequencing Occitanie Innovation

Camille Eché [1], Carole Iampietro [1], Clément Birbes [2], Andreea Dréau [2], Claire Kuchly [1], Arnaud Di Franco[2], Christophe Klopp [2], Thomas Faraut [3], Sarah Djebali [3], Adrien Castinel[1] Matthias Zytnicki [4], Erwan Denis[1], Mekki Boussaha [5], Cécile Grohs [5], Didier Boichard [5], Christine Gaspin [2], Denis Milan [1], Cécile Donnadieu [1]

[1] INRAE, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France. [2] Plateforme Bio-informatique Genotoul, Mathématiques et Informatique Appliquées de Toulouse, INRAE, Castanet-Tolosan, France. [3] GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet-Tolosan Cedex, F-31326, France. [4] Université Fédérale de Toulouse, INRAE, MIAT, 31326, Castanet-Tolosan, France. [5] Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France.

Background and objectives

Inspired by the production of reference datasets in the Genome in a Bottle project, we sequenced a Charolais heifer using different technologies: Illumina paired-end, Oxford Nanopore, Pacific Biosciences (HiFi and CLR), 10X Genomics linked-reads, and Hi-C. To generate haplotypic assemblies, we also sequenced both parents with short reads. From these data, we constructed two haplotyped trio high quality reference genomes and a consensus assembly, using current software packages. **The assemblies generated using PacBio HiFi reach a size of 3.2 Gb,** which is significantly larger than the 2.7 Gb ARS-UCD1.2 reference. The BUSCO score of the consensus assembly reaches a completeness of 95.8% completeness among highly conserved mammalian genes. **We also identified 35,866 structural variants greater than 50 base pairs**. This assembly is a contribution to the bovine pangenome for the Charolais breed. These datasets will prove to be useful, allowing the community to gain additional insight into sequencing technologies for applications such as SNP, indel or structural variant calling, and de novo assembly.

More information in our data paper ➔ https://doi.org/10.1038/s41597-023-02249-1

Sequencing technologies and pipelines used for the bovine trio

1 Sample	2 Sequencing technology	3 Library type	4 Dataset
Father	Oxford Nanopore Illumina	Ligation sequencing gDNA 10X Genomics Chromium Hi-C	a,b,c
Mother	Oxford Nanopore Illumina PacBio	Ligation sequencing gDNA 10X Genomics Chromium Hi-C Circular Long Read	a,b,c
Heifer	Oxford Nanopore Illumina PacBio	Ligation sequencing gDNA 10X Genomics Chromium PCR Free 2x250pb Hi-C Circular Long Read Consensus Long Read	a, b, c, d

Figure 1: Technologies used for the study

The parents and the heifer were sequenced using Oxford Nanopore Technologies on GridION and PromethION, Chromium 10X and the Hi-C method on MiSeq, HiSeq or NovaSeq6000. The heifer was additionally sequenced using Illumina 2x250pb on NovaSeq6000 and PacBio Sequel II (CLR and CCS (i.e. HiFi reads) mode). For the Trio approach, parent reads (2x150pb) from 10X Genomics Chromium data were used.

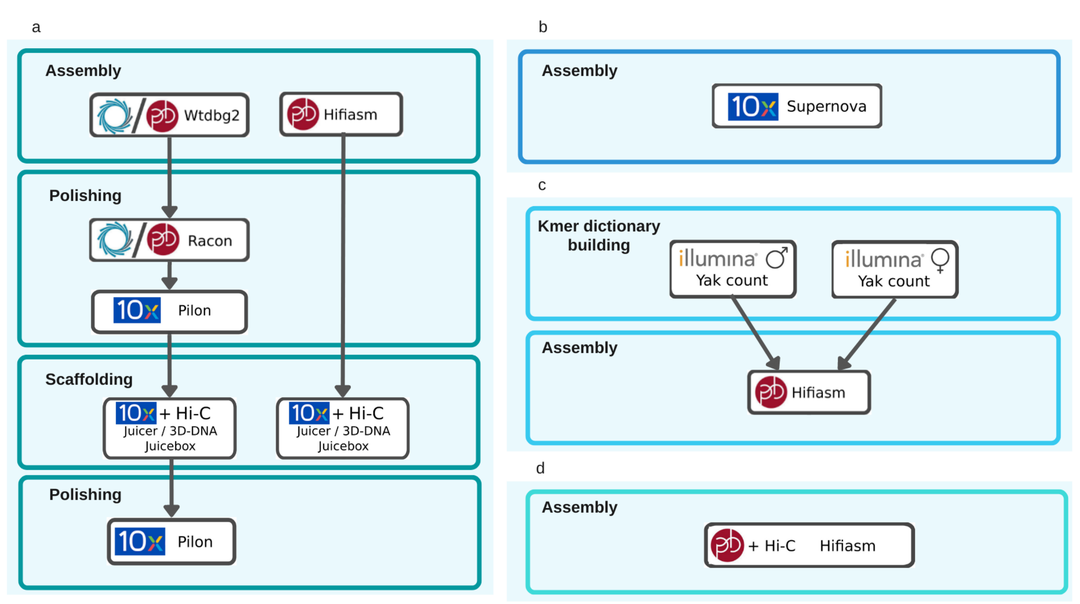

Figure 2: Details of the 5 pipelines used to produce assemblies

a-Long reads assemblies from Oxford Nanopore Technologies and Pacific Biosciences followed by a polishing step for erroneous assemblies and scaffolding step. **b-**10XChromium assembly and scaffolding with Supernova. **c-Phased** assembly with HiFi and parental Illumina reads. **d-**Phased assembly with HiFi and Hi-C data.

Comparison of heifer consensus and haplotyped assemblies

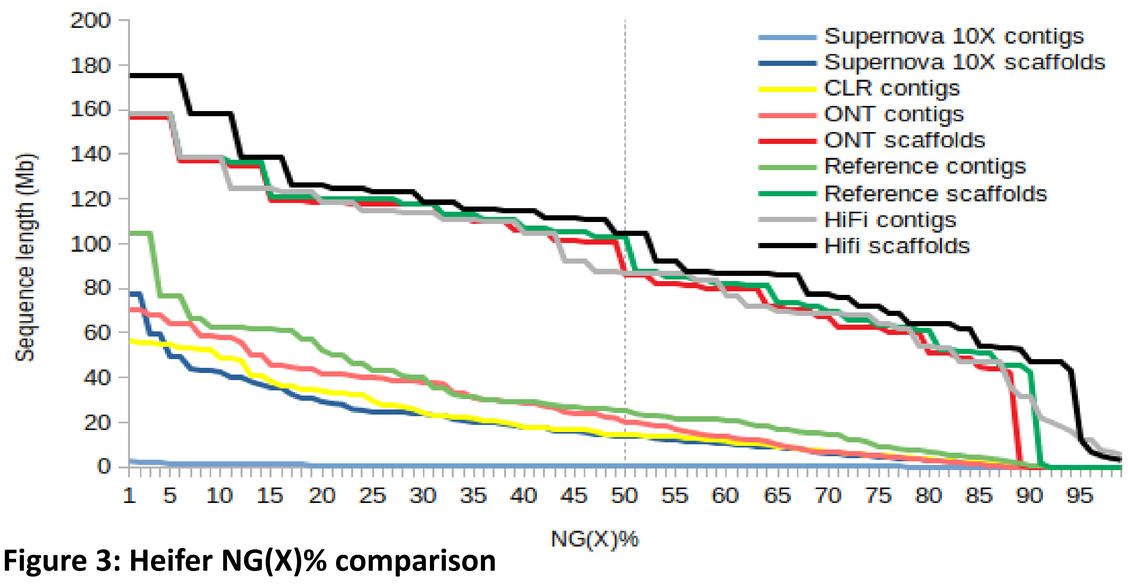

Figure 3: Heifer NG(X)% comparison

Comparison and evolution of the heifer assemblies size metrics after different steps in different assemblies pipelines.

	ARS-UCD1.2	Heifer w/ Parents short read data		Heifer w/ Heifer Hi-C data	
		Hap1 (c)	Hap2 (c)	Hap1 (d)	Hap2 (d)
Pipeline used					
Data type	CLR	CLR + Trio	CLR + Trio	CCS + Hi-C	CCS + Hi-C
Number of contigs	3 077	2 871	2 300	2 658	2 136
Total size (Gb)	2.7	3.16	3.11	3.08	3.18
N50 contigs length	12 000 000	71 619 842	69 165 538	80 106 842	71 644 334
BUSCO	95.7%	95.8%	95.3%	95.8%	95.7%
Phasing ratio	**	97.3%	96.7%	62.6%	60.5%
Contigs phasing ratio	**	97.5%	96.9%	84.6%	85.6%

Table 1 : Summary of heifer phased produced assemblies

Details of pipeline used in this study are shown in Figure 2. CLR (80X), CCS (40X) and Hi-C (28X) Illumina (30X)* BUSCO analysis was performed on polished contigs. ** Reference is not haplotyped

Structural variation and SNPs

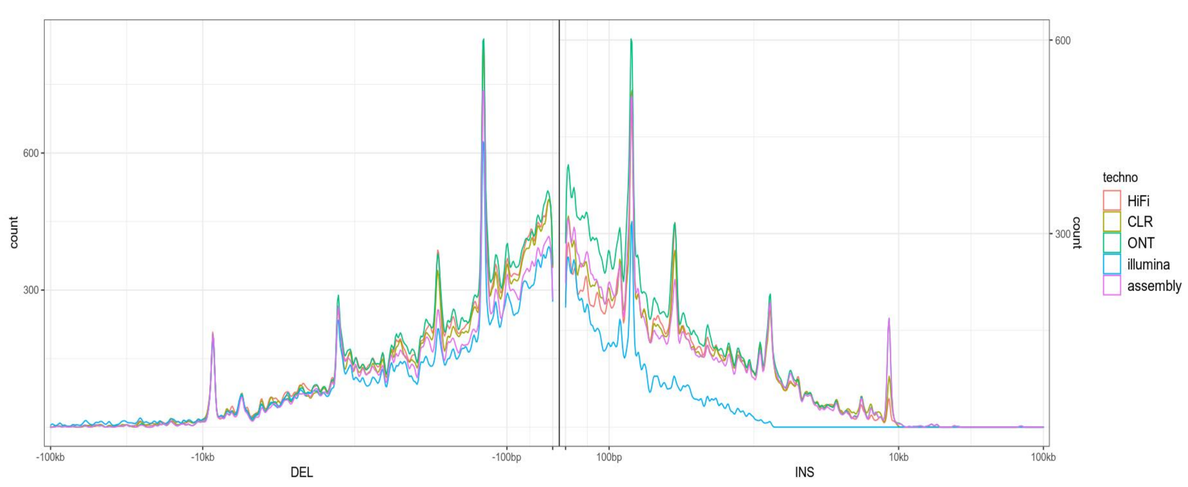

Figure 4: Structural variant length distribution

Length distribution of DEL (left) and INS (right) variants, detected by the different sequencing technologies on the Charolais heifer. Sharp peaks correspond to the structural variability generated by transposable elements (SINE, LTR and LINE). All the long reads technologies exhibit roughly the same distribution underlining the fact that they are comparable in their ability to detect DEL and INS variants (see also Fig 5). This holds also with short reads for DEL variants. In contrast, there is a significant difference between long and short read technologies in their ability to detect INS variants, short reads failing to detect a large proportion of INS variants, the larger the size the more pronounced the trend. This is an expected consequence of the limited size of reads that prevents to recover long insertions. **Long read technology seems imperative for an accurate detection of the full spectrum of structural variants.**

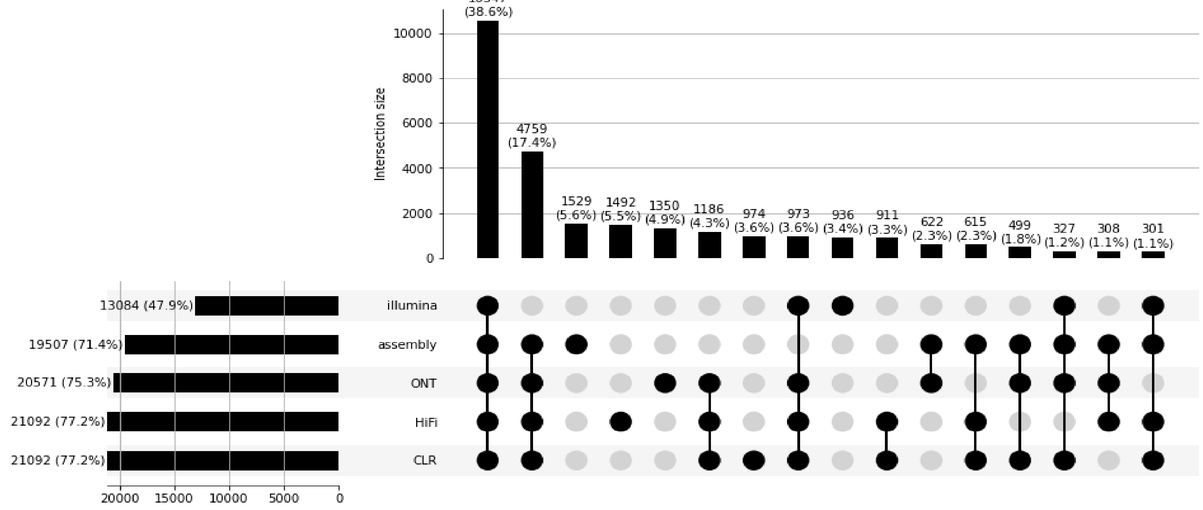

Figure 5: Upset plot of the identified deletions and insertions

The majority of variants are identified by all the long read technologies as well as for the assembly comparison based variant detection. This upset plot again highlights the added value of long reads for structural variant detection.

High-quality phased and consensus genome assemblies for Charolais heifer with PacBio, Hi-C and parents data !
Very promising SV detection with long reads data !

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. It's 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 in the project.

Projet cofinancé par le Fonds Européen de Développement Régional