De novo assembly of 90 bovines genomes using PacBio CLR application comparison and optimization

Camille Marcuzzo, Amandine Suin, Camille Eché, Andreea Dreau, Clément Birbes, Arnaud Di Franco, Christophe Klopp, Carole Iampietro, Thomas Faraut, Claire Kuchly, et al.

De novo assembly of 90 bovines genomes using PacBio CLR application comparison and optimization

Marcuzzo, Camille[1]; Suin, Amandine[1]; Eché, Camille[1]; Dréau, Andreea[2]; Birbes, Clement[2]; Di Franco, Arnaud[2]; Klopp, Christophe[2]; Iampietro, Carole[1]; Faraut, Thomas[5]; Kuchly, Claire[1]; Zytnicki, Matthias[2]; Fritz, Sébastien[3-4]; Boussaha, Mekki[3]; Grohs, Cécile[3]; Boichard, Didier[3]; Gaspin, Christine[2]; Milan, Denis[1 5]; Donnadieu, Cécile[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] Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France. [4] Allice, 75012 Paris, France [5] GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet-Tolosan Cedex, F-31326, France.

Background and objectives

GeT-PlaGe is a genomic core facility which provides access to technologies and expertise in genome sequencing to academic and private research teams. The SeqOccIn (Sequencing Occitanie Innovation) project, managed 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". The main objective is to acquire expertise on the optimal combination of long fragment sequencing technologies and associated applications to better characterize complex genomes in the agronomical field: from SNP and structural variation detection, to high quality assembly production.

These approaches have allowed to assemble a novel high quality bovine reference genome. We chose to identify single nucleotide polymorphisms and structural variants (insertions and deletions) by comparing multiple assemblies. A preliminary study done on a heifer (37160) showed that long-reads sequencing is the most suitable solution for structural variations detection. We sequenced 154 bulls from 14 breeds with PacBio Sequel II Continuous Long-Read (CLR). Here we present assembly as well as variation detection results for 90 bulls from 11 breeds (9 Abondances, 7 Aubracs, 4 Blondes d'Aquitaine, 3 Charolaises, 14 Holsteins, 11 Limousines, 19 Montbeliardes, 14 Normandes, 3 Rouges Flamandes, 2 Tarentaises and 4 Vosgiennes).

Multiple technologies structural variant detection profiles

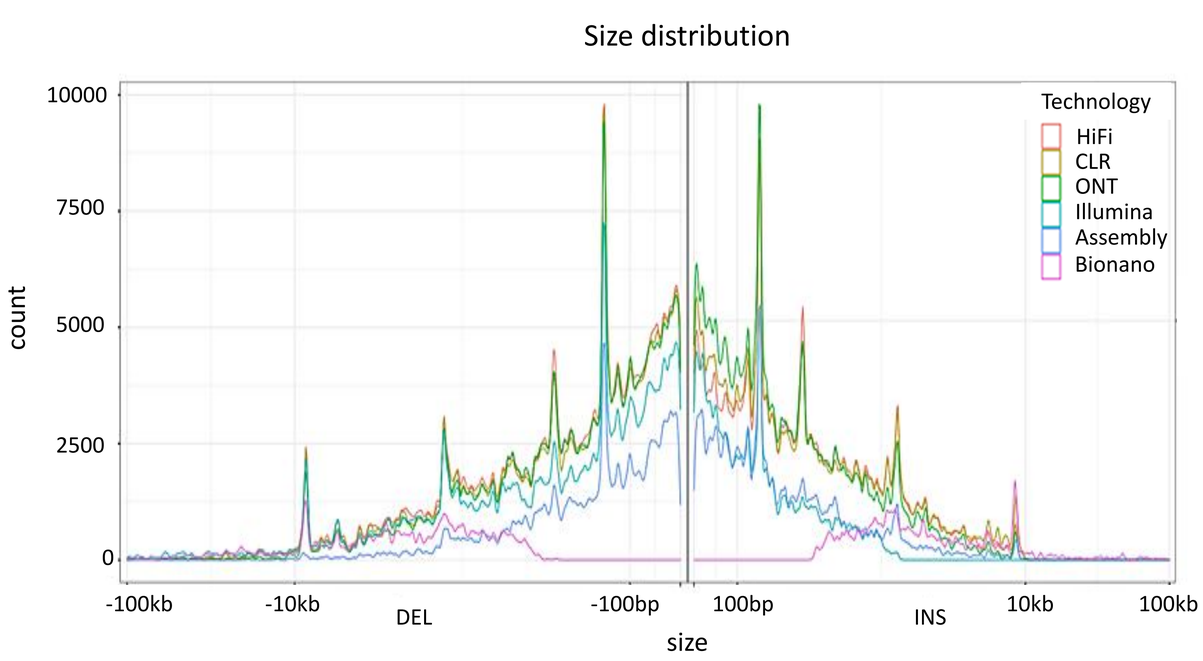

Figure 1: Heifer 37160 structural variation size distribution

Size (x-axis) and count (y-axis) of detected variants by different sequencing technology. This graph shows a symmetrical visualization of insertions and deletions. Peaks correspond to repeated elements families.

Structural variant detection profiles are homogeneous with a long-read technologies (HiFi, CLR, ONT). The Bionano technology does not allow the detection of small repeated elements.

Thanks to these results, a reference variant set was established with 27389 deletions and 29205 insertions.

Structural variations are longer than 50 bp. Whatever the size, few are detected above 10 kb.

Impact of fragment size and run throughput on contigs N50

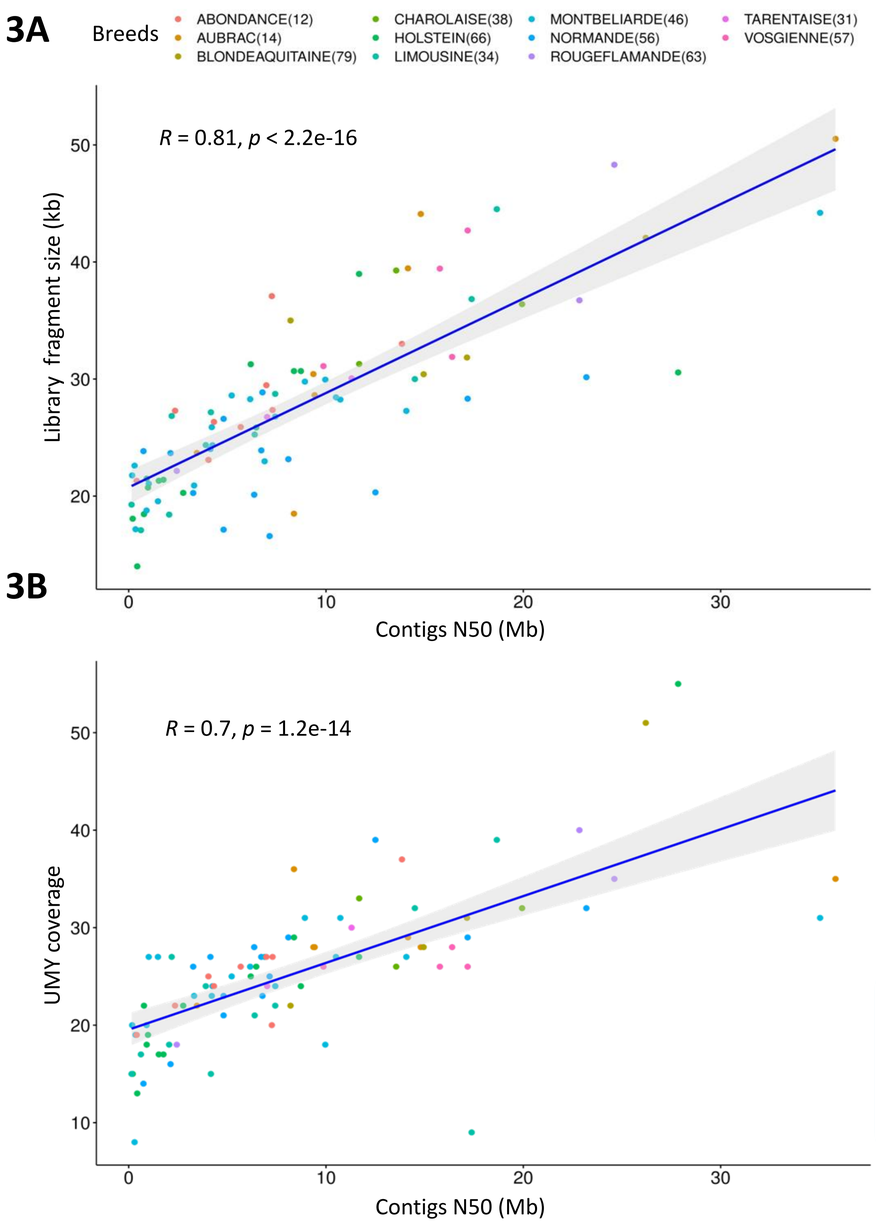

Figure 3: 90 bulls assembly metrics correlation graphs

3A: Library fragment size impact on contigs N50 (Mb).

3B: Impact of UMY (Unique Molecular Yield) coverage on contigs N50 (Mb).

Breeds have no impact on metrics

R: linear correlation coefficient ($0 < R < 1$).

p: p-value

Contigs N50: contigs ordered by decreasing length, N50 length of the shortest contig at 50 % of the total genome length.

Best CLR values produced :

- **2.6 Gb** of total assembly genome
- **130 Mb** of longest contig
- **35,8 Mb** of contig N50

PacBio library preparation workflow

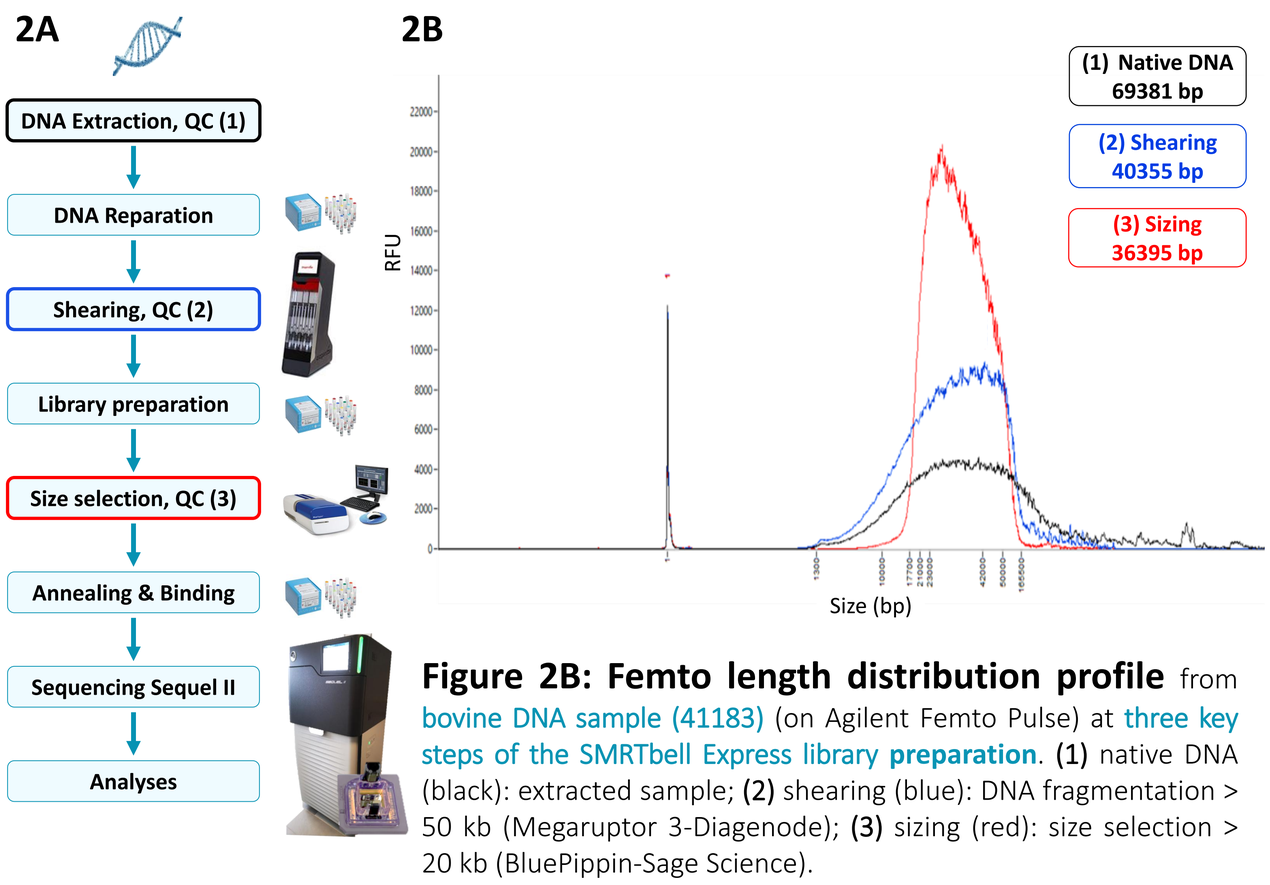

Figure 2B: Femto length distribution profile from bovine DNA sample (41183) (on Agilent Femto Pulse) at three key steps of the SMRTbell Express library **preparation**. **(1)** native DNA (black): extracted sample; **(2)** shearing (blue): DNA fragmentation > 50 kb (Megaruptor 3-Diagenode); **(3)** sizing (red): size selection > 20 kb (BluePippin-Sage Science).

RFU: Relative Fluorescence Units.

Figure 2A: PacBio Workflow

DNA Extraction kit: Qiasymphony; Quality control (QC): Femto-Qubit-Nanodrop; Reparation: SMRTbell Damage Repair Kit SPV3; Shearing: Megaruptor 1 or 3; Library preparation: SMRTbell Express Template Prep Kit 2.0, Size Selection: BluePippin; Sequencing: Sequel II Binding kit 2.2, Sequel II Internal Control complex 1.0, Sequel II Sequencing kit 2.0, SMRTcell 8M Tray, Sequencing Primer V4 or V5.

De novo assembly metrics

Metrics \ Bulls	*Acceptable metrics*				*Good metrics*			
	40539	40618	40533	41230	40514	40689	41183	41188
Final library size (kb)	21	19	21	19	30	39	36	43
Subread N50 (kb)	15	12	16	13	20	20	25	26
Total coverage	32X	40X	27X	27X	60X	40X	46X	36X
Coverage UMY	19X	20X	17X	**15X**	32X	27X	32X	26X
Total Size (Gb)	2.6	2.6	2.6	**2.4**	2.6	2.6	2.6	2.6
Number of contigs	9042	9337	6918	**30438**	4021	4108	4349	4389
Longest contigs (Mb)	5.5	7.4	10.9	4.9	66.5	42.3	73.1	81.4
NG50 (Mb)	0.9	0.9	1.7	0.1	21	11.2	19.7	16.9

Table 1: Variable bull assembly statistics

Among the 90 bull assembly metrics, 8 are presented here. They were chosen because they illustrate two distinct groups of acceptable and good assembly metrics. They were assembled with Wtdbg2 V2.3 using sq option adapted for Sequel II CLR data. This assembler needs few resources to produce assemblies with good metrics. The total assembly sizes are very close to the reference bovine genome: ARS-UCD1.2 (2.7 Gb).

NG50: is the same as N50 statistic except that it is 50 % of the reference genome size.

In order to have good assembly metrics and enable pangenomic structural variations detection:
→ we set PacBio Single Molecule Real-Time (SMRT) Unique Molecular Yield (UMY) at 20X
→ we built libraries with a size longer than 25 kb

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