



**HAL**  
open science

## Major contribution to the bovine pangenome: whole genome sequences, SNPs, and structural variants of 154 bulls from 14 breeds

Amandine Suin, Camille Marcuzzo, Camille Ech , Andreea Dr au, Cl ment Birbes, Arnaud Di Franco, Christophe Klopp, Carole Iampietro, Thomas Faraut, Claire Kuchly, et al.

### ► To cite this version:

Amandine Suin, Camille Marcuzzo, Camille Ech , Andreea Dr au, Cl ment Birbes, et al.. Major contribution to the bovine pangenome: whole genome sequences, SNPs, and structural variants of 154 bulls from 14 breeds. *Plant and Genome San Diego 2024*, Jan 2024, San Diego, United States. hal-04432840

**HAL Id: hal-04432840**

**<https://hal.science/hal-04432840v1>**

Submitted on 1 Feb 2024

**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 destin e au d p t et   la diffusion de documents scientifiques de niveau recherche, publi s ou non,  manant des  tablissements d'enseignement et de recherche fran ais ou  trangers, des laboratoires publics ou priv s.



Distributed under a Creative Commons Attribution 4.0 International License



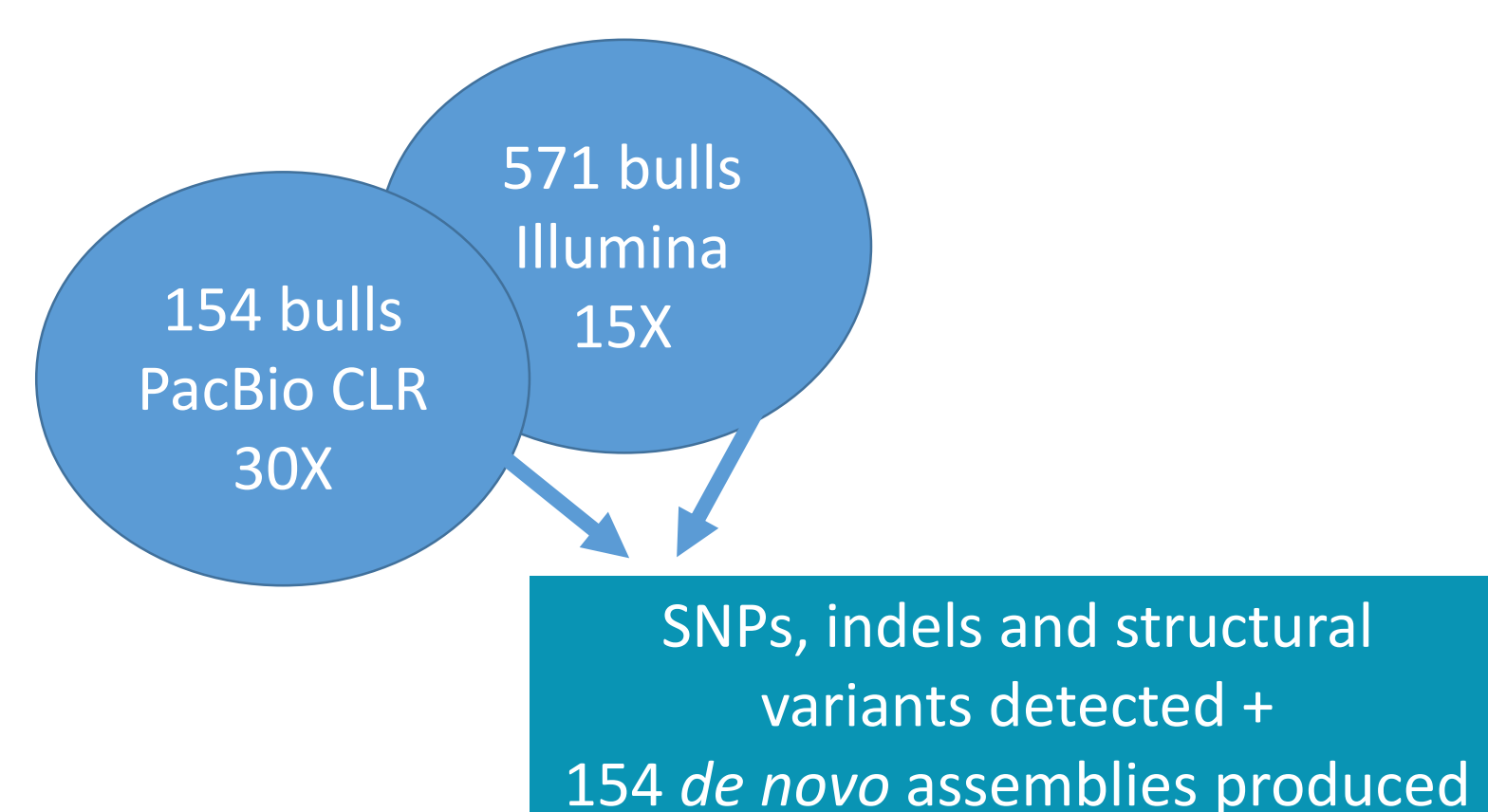
Marcuzzo, Camille<sup>1</sup>; Suin, Amandine<sup>1</sup>; Eché, Camille<sup>1</sup>; Dréau, Andreea<sup>2</sup>; Birbes, Clement<sup>2</sup>; Di Franco, Arnaud<sup>2</sup>; Klopp, Christophe<sup>2</sup>; Iampietro, Carole<sup>1</sup>; Faraut, Thomas<sup>5</sup>; Kuchly, Claire<sup>1</sup>; Zytnicki, Matthias<sup>2</sup>; Fritz, Sébastien<sup>3,4</sup>; Boussaha, Mekki<sup>3</sup>; Grohs, Cécile<sup>3</sup>; Boichard, Didier<sup>3</sup>; Gaspin, Christine<sup>2</sup>; Milan, Denis<sup>1,5</sup>; Donnadiou, Cécile<sup>1</sup>

<sup>1</sup> INRAE, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France <sup>2</sup> Plateforme Bio-informatique Genotoul, Mathématiques et Informatique Appliquées de Toulouse, INRAE, Castanet-Tolosan, France. <sup>3</sup> Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France. <sup>4</sup> Allice, 75012 Paris, France <sup>5</sup> GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet-Tolosan Cedex, F-31326, France.

## Background and objectives

Currently, most bovine studies are based on the Dominette ARS-UCD1.2 cow, but newly discovered sequences are missing from this Hereford assembly. Anomaly detection chips are incomplete, and structural variation analysis using short reads remains difficult. This is the background against which the Sequencing Occitanie Innovation (SeqOCCln) project was born. The main objective is to acquire expertise on the optimal combination of long fragment sequencing technologies and related applications to better characterise complex genomes in the agronomic field. In our first publication, we contributed to the bovine pangenome by producing a high-quality haplotype assembly for the Charolais breed. We chose to investigate single nucleotide polymorphisms and structural variants (insertions and deletions) by comparing several assemblies. This recent study has shown that long-read sequencing is the most suitable solution for detecting structural variations. Using PacBio Sequel II Continuous Long-Read (CLR) and Illumina 15X, we sequenced 154 bulls from 14 breeds (Holstein, Montbéliarde, Normande, Brune, Simmental, Abondance, Tarentaise, Vosgienne, Blonde d'Aquitaine, Charolaise, Limousine, Aubrac, Flamande, Parthenaise). These datasets will provide a useful resource for the community to better understand sequencing technologies for applications such as the identification of SNPs, indels or structural variants, and *de novo* assembly.

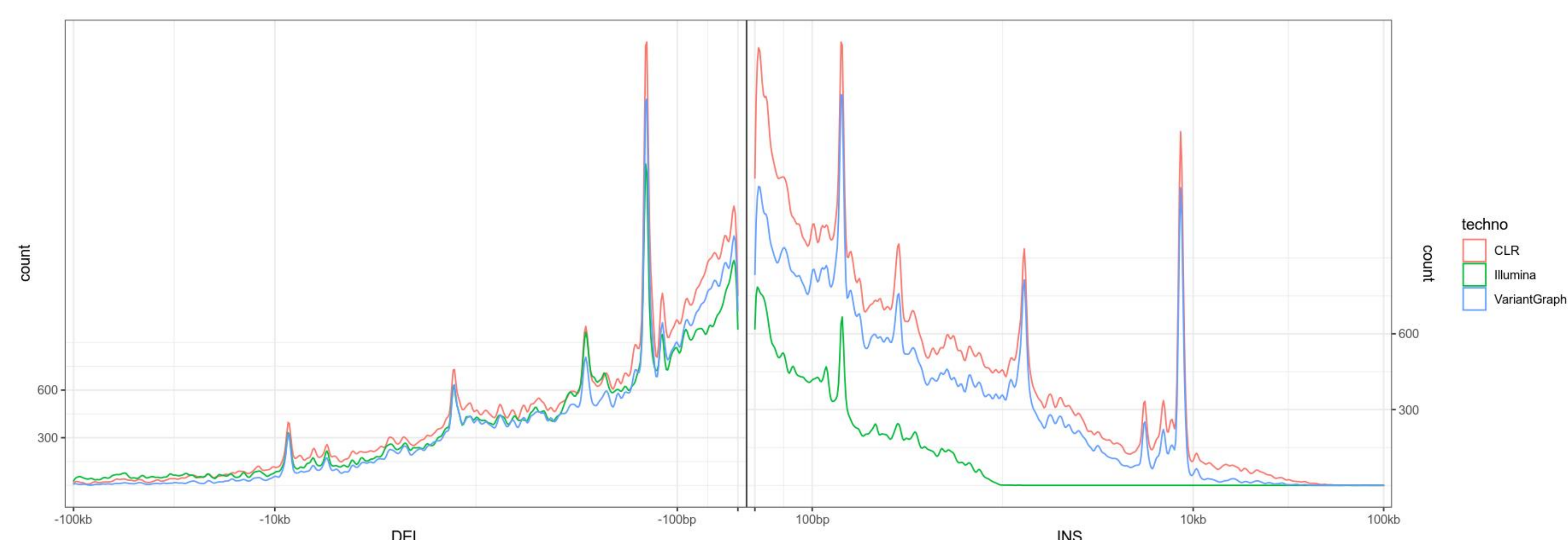
## Sequencing strategy



**Figure 1 : Sequencing strategy**

In order to demonstrate the added value of long reads for the detection of variants and more specifically structural variants, a set of 154 bulls were sequenced using both PacBio CLR long read and Illumina short read technologies. 154 *de novo* assemblies were also produced based on the CLR sequence data.

## Structural variation size distribution



**Figure 2 : Recovering the structural variability landscape using variation graphs**

Size distribution of variants detected by long (red) and short (green) read sequencing technologies. Sharp peaks correspond to the structural variability generated by transposable elements. Short reads fail to detect a large proportion of variants, specifically insertion variants, that were detected on the same population of 154 bulls using long reads. If the corresponding variants are inserted in a variation graph, the short reads can successfully recover a large proportion of variants and associated genotypes on the same set of 154 animals (blue). Genotypes are recovered using paragraph. **Variation graphs together with a priori knowledge of variants segregating in a population can be successfully used in combination with short reads to recover a large spectrum of structural variation diversity in a population.**

## Bioinformatics pipelines for SNPs and SVs detection

Pipeline w/ 4 steps	A.		B.		
	Illumina SR		PacBio CLR	PacBio HIFI	Nanopore
alignment	BWA mem		pbmm2 (rapper of minimap)		minimap
SNPs+ Indels calling	GATK		Longshot		Deepvariant
SV calling	Lumpy + Delly + Pindel		Sniffles2 + CuteSV + pbSV		
annotation	SnEff/VEP		SnEff/VEP		
	C.				
	PacBio CLR	Nanopore			
Assembly	<i>de novo</i> (wtDBG2)				

**Table 1 : Pipelines strategy**

A. Use of the ARS-UCD1.2 bovine reference genome in all analyses B. Development of a pipeline for long reads data. C. Development of a pipeline for the construction of bovine genome assemblies.

## Contribution to the Bovine Pangenome: First results for 3 breeds

	Breed	CHAROLAISE	TARENDAISE	VOSGIENNE	
Pan-genome	No. of assemblies	6	5	4	
	Graph length	2 524 730 836	2 551 978 622	2 554 024 324	
	Reference length		2 489 385 779		
	Core genome length	2 317 557 075	2 363 900 584	2 330 940 534	
	Length of flexible part of graph	<b>35 345 057</b>	<b>62 592 843</b>	<b>64 638 545</b>	
SVs	Bi-allelic deletions	5 475	4 019	3 851	
	Bi-allelic inserts	6 840	5 144	5 140	
	Multi-allelic insertions	787	121	217	
	Sequence substitution	AltInsertion Bi-al	6 823	10 445	10 241
		AltDeletion Bi-al	6 045	10 875	9 964
		AltDeletion Multi-al	10 390	4 630	6 223
	<b>Total SVs</b>	<b>36 360</b>	<b>35 234</b>	<b>35 636</b>	

**Table 4: Multi-assembly graph build**

The multi-assembly graph was constructed using minigraph tool and the ARS-UCD1.2 as the initial backbone of the graph structure. *De novo* assemblies were iteratively added into the graph and SVs were identified using gfatools with default parameters.

## New useful variants detected

	SNVs	SVs
<b>Total variants</b>	<b>34 252 085</b> 28 931 309 SNPs 5 320 776 InDels	<b>87 216</b> 58 571 deletions 14 836 duplications 13 809 inversions
<b>New candidate variants</b>	<b>1 548</b>	<b>1 219</b>
<b>New variants added to EuroGMD</b>	<b>1 342</b>	<b>874</b>

**Table 2 : Identification of new candidate variants**

2,767 new candidate variants (1,548 SNVs + 1219 SVs) with strong annotation and expected effect on phenotypes (genetic abnormalities or QTL). 2,216 (1,342 SNVs + 874 SVs) new candidate variants have been added to the EuroGMD chip. **These variants will help us to perform genotyping on a larger scale and to monitor these mutations in all breeds.**

## New causal mutation identified

BTA	Annotation	VEP SIFT	Gene	Phenotype
19	missense	deleterious(0)	<i>ITGB4</i>	junctional epidermolysis bullosa

**Table 3 : Example of prediction result : Epidermolysis bullosa junctionalis**

Based on the severity and certainty of these predictions, this mutation has been counter-selected. Several other mutations are monitored by genotyping on chip (detection and phenotyping of homozygotes).

**Huge dataset useful for the bovine community !**  
*De novo* assemblies production, Building an exhaustive catalog of bovine polymorphisms from whole genome sequences, Provide a systematic diagnosis on all sequenced bulls

The contribution of a phased assembly is clearly greater than a consensus assembly for variant detection (no loss of information)  
Haplotypes are obtained and are very useful for pangenomes

These long-read data have made these publications possible!

**Large-scale detection and characterization of interchromosomal rearrangements in normozoospermic bulls using massive genotype and phenotype data sets.** Jeanlin Jourdain & al DOI: 10.1101/gr.277787.123

**Integrin alpha 6 homozygous splice-site mutation causes a new form of junctional epidermolysis bullosa in Charolais cattle.** Boussaha, M., Boulling, A., Wolgust, V. et al <https://doi.org/10.1186/s12711-023-00814-1>