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

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

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

SeqOccIn — Sequencing Occitanie Innovation

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

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 (SeqOccIn) 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

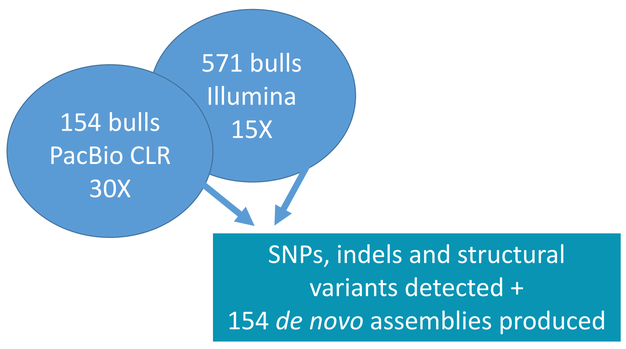

Figure 1 : Sequencing strategy

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

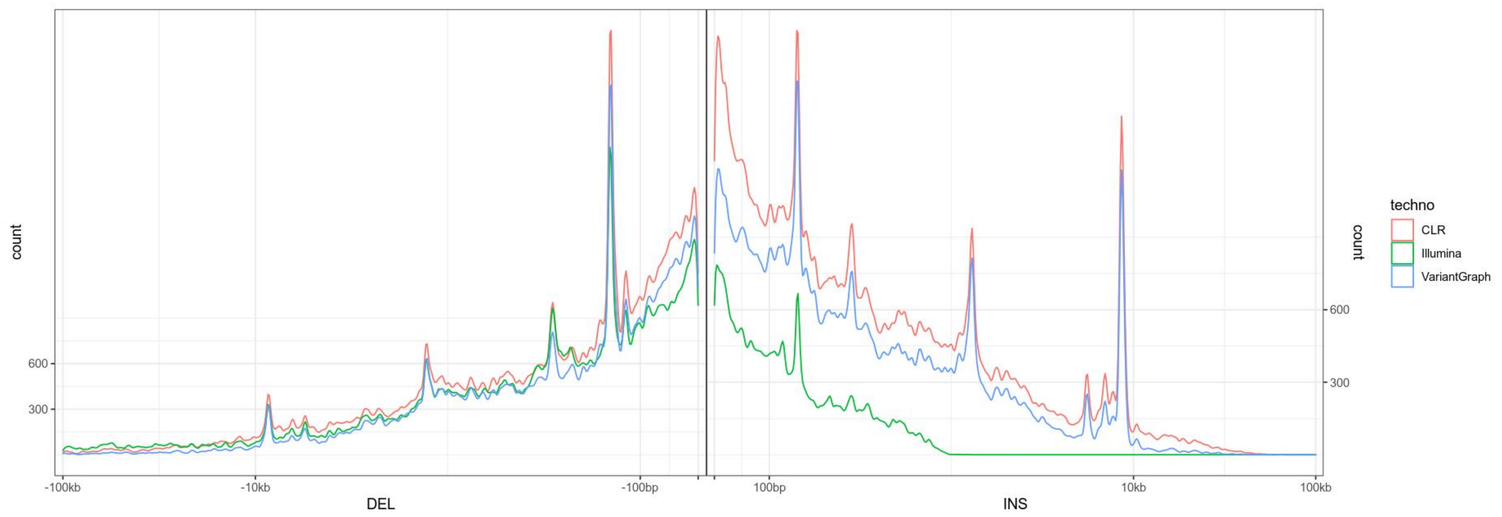

Figure 2 : Recovering the structural variability landscape using variation graphs

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. Illumina SR	B. PacBio CLR	B. PacBio HIFI	B. Nanopore
alignment	BWA mem	pbmm2 (rapper of minimap)	minimap	
SNPs+ Indels calling	GATK	Longshot	Deepvariant	
SV calling	Lumpy + Delly + Pindel	Sniffles2 + CuteSV + pbSV		
annotation	SnpEff/VEP	SnpEff/VEP		

	C. PacBio CLR	C. Nanopore
Assembly	*de novo* (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	TARENTAISE	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		**35 345 057**	**62 592 843**	**64 638 545**
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
		Total SVs	**36 360**	**35 234**	**35 636**

New usefull variants detected

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

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)	***ITGB4***	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).

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.

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

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