

RESEARCH

Open Access



Assembly of a pangenome uncovers novel non-reference unique insertion sequences in cattle highlighting their genetic diversity

Valentin Sorin^{1*} , Florian Besnard^{1,2} , Aurélien Capitan¹ , Cécile Grohs¹ , Maulana Mughitz Naji¹ , Clémentine Escoufflaire^{1,2} , Sébastien Fritz^{1,2} , Joanna Lledo³ , Camille Eché³ , Carole Iampietro³ , Cécile Donnadiou³ , Denis Milan^{3,4} , Laurence Drouilhet⁴ , Gwenola Tosser-Klopp⁴ , Didier Boichard¹ , Christophe Klopp⁵ , Marie-Pierre Sanchez¹ and Mekki Boussaha^{1*}

Abstract

Background The current cattle reference genome, derived from a single Hereford cow, does not capture the full spectrum of genetic diversity present within the species. Moreover, detecting structural variations (SVs ≥ 50 nucleotides long) remains challenging using only standard approaches of either short or long-read sequence approaches against a linear reference genome. Recent advances in long-read sequencing technologies and graph-based assembly now enable the construction of breed-specific pangenomes, revealing previously uncharacterized genomic regions that may contribute to important agricultural traits.

Results In this study we constructed a cattle pangenome graph using 16 high-quality haplotype-resolved genome assemblies originating from nine breeds representing the diversity of French cattle populations, and including yak (*Bos grunniens*) as a close outgroup species. Using a trio-based strategy combined with complementary sequencing technologies and bioinformatics methods, we identified and characterized 101,219 structural variations. Of these, 33,634 were classified as non-reference unique insertions (NRUIs), adding several megabases of novel genomic sequences absent from the current Hereford reference genome.

Analysis of the distribution of these NRUIs revealed significant genome-wide enrichment within QTL regions associated with milk production and morphological traits, suggesting their contribution to the genetic basis of economically relevant phenotypes. Furthermore, their functional annotation highlighted two NRUIs located within the intronic regions of *ARMH3* and *EPHA5*, both specific to the Normande breed and significantly associated with milk production and morphological traits, respectively.

Conclusions Our findings demonstrate the value of pangenome approaches to uncover functionally relevant SVs, particularly NRUIs, that are systematically not in the current reference genome. By linking these variants to economically important traits, our work underscores the need to incorporate breed diversity into future genomic analyses and reference-building efforts in cattle.

Keywords *ARMH3*, Cattle, *EPHA5*, Non-reference unique insertion, Pangenome, Structural variants

*Correspondence:

Valentin Sorin

valentin.sorin@inrae.fr

Mekki Boussaha

mekki.boussaha@inrae.fr

Full list of author information is available at the end of the article



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

The advent of long-read sequencing technologies that allow for the production of sequences for longer fragments has furthered our comprehension of the bovine genome. Over the past decade, significant efforts have been undertaken to produce de novo genome assemblies, enabling the development of pangenomic studies in cattle.

These initiatives have resulted in the creation of the international Bovine Pangenome Consortium (<https://bovinepangenome.github.io/>), whose primary objective is to construct a comprehensive pangenome of domestic cattle and closely related species to identify variants lost during domestication or breed differentiation [1].

Recent studies have confirmed the accuracy and robustness of cattle pangenomes, with high concordance in SV detection across platforms and good agreement among different graph-construction methods [2, 3]. These studies also highlight method-specific advantages and limitations that must be considered when interpreting variants derived from multiple assemblies.

In cattle, pangenomes have revealed tens of megabases of non-reference sequences and several functionally important SVs, including a complex variant upstream of *KIT* associated with the white-headed coat pattern [4], and an 8.2 kb deletion within *MATN3* associated with stature in Holstein [5]. Alongside these SVs, large panels of non-reference unique insertions (NRUIs), representing megabases of genomic sequences absent from the current bovine reference genome assembly, have been identified.

Collectively, these studies demonstrate that relying on a single linear reference assembly limits the characterization of cattle genetic diversity, particularly for uncovering novel sequences that may contain expressed novel genes and many previously unreported variants [6]. Pangenomes can also improve the precision of reference-guided alignment and variant detection compared to the current ARS-UCD1.2 reference assembly [7]. More recent pangenomic studies have further investigated NRUIs: pangenomes built from Indian dairy *Bos indicus* breeds [8], multiple French cattle breeds [5], and Holstein/Jersey populations [9], have reported between 17.7 and 63.75 Mb of NRUIs.

Together, these findings indicate that the organization of the cattle genome is more complex than previously believed and highlight the value of pangenome approaches to reveal functionally relevant SVs, particularly NRUIs, which contribute substantially to genomic diversity and the genetic architecture of key traits in cattle.

In the present study, we have constructed a cattle pangenome graph by integrating sixteen de novo haplotype-resolved genome assemblies originating from

six major French dairy and beef breeds (Aubrac, Blonde d'Aquitaine, Charolais, Holstein, Montbéliarde, and Normande), three regional French dairy breeds (Abondance, Tarentaise, and Vosgienne) and yak (*B. grunniens*). Using this pangenome, we constructed a comprehensive catalog of SVs and NRUIs, performed functional annotation, and explored their potential contributions to economically important traits in cattle.

Methods

Sample collection

In this study, we constructed 14 haplotype-resolved genome assemblies using short and long read sequences derived from five pure bred [Abondance (ABO), Aubrac (AUB), Blonde d'Aquitaine (BAQ), Tarentaise (TAR) and Vosgienne (VOS)] and two cross-bred or hybrid trios [Holstein×Normande (HN) and Yak×Montbéliarde (YM)]. To enable a comprehensive bovine pangenome analysis, we also included the current ARS-UCD1.2 Herford bovine reference genome [10] and two publicly available haplotype-resolved genome assemblies from the Charolaise breed [11].

DNA extraction and sequencing

Genomic DNA was extracted from blood, semen, or ear biopsies following the protocol “DNA isolation from cattle tissues: blood, semen or any kind of tissues, including ear biopsies” [12] using Puregene kits (Qiagen) [13]. DNA quantity and quality were assessed using a Nanodrop One spectrophotometer (ThermoFisher Scientific).

For each parent of the trios, DNA libraries were prepared with the TruSeq Nano DNA HT Sample Preparation Kit (Illumina, CA, USA) in accordance with the manufacturer's guidelines and then sequenced on an Illumina NovaSeq 6000 platform (Illumina, CA, USA) to produce 150-bp paired-end (PE) reads.

For the offspring, PacBio HiFi library preparation and sequencing were performed at the INRAE GeT-PlaGe core facility (Toulouse, France; <https://get.genotoul.fr/la-plateforme/get-plage/>) according to the manufacturer's instructions “Procedure & Checklist Preparing HiFi SMRTbell Libraries using SMRTbell Express Template Prep Kit 2.0” [14].

Construction of haplotype-resolved genome assemblies

Haplotype-resolved assemblies were generated using a trio binning approach, which partitions offspring long reads into haplotype-specific sets based on parental short reads [15]. Illumina PE short reads of the parents were used to create a kmer dictionary employing yak version 0.1 with the default parameters [16]. The offspring PacBio HiFi long reads were then assembled using Hifiasm version 0.19.8 [17] along with the parental kmer dictionaries

(parameters -1 for paternal specific kmers and -2 for maternal specific kmers) assign reads to their respective haplotypes. The resulting haplotype assemblies were scaffolded to near-chromosome level using RagTag scaffold tool (version v2.1.0) with default parameters [18] aligned to the ARS-UCD1.2 Hereford reference (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002263795.1/). Default RagTag settings were retained, as they generated well-balanced k-mer distributions that clearly differentiated sequence origins and produced chromosome-scale scaffolds collinear with the reference genome, ensuring high quality assemblies.

Validation of parental haplotype-resolved assembly

The quality of the haplotype-resolved genome assemblies was evaluated using three metrics: total assembly length, N50 score, and completeness assessment. Assembly completeness of the assemblies was assessed using completeness [19], which uses the miniprot protein-to-genome aligner in combination with conserved orthologous genes from BUSCO [20]. The assembly quality value (QV) was also quantified using Merqury version 1.3 [21]. In parallel, haplotypes were also aligned to the ARS-UCD1.2 reference genome using D-Genies 1.5 [22] with default parameters to visually inspect the dotplots for large-scale structural accuracy.

Construction of a high-quality cattle pangenome graph

All 17 assemblies, including the ARS-UCD1.2 Hereford reference genome, were used to build a cattle pangenome graph with Minigraph (version 0–21) [23].

First, genetic distances between each of the 16 assemblies and the reference genome were estimated using Mash [24]. The pangenome graph was then built progressively, using the ARS-UCD1.2 reference as the backbone and incorporating the remaining assemblies in order of increasing genetic “sample distance” (see Additional file 2: Table S1). To simplify the pangenome graph structure, only autosomes were included in the construction.

SV calling

The procedure for SV calling from the pangenome graph was conducted following the workflow described by Sorin et al. [5]. Briefly, allele information for each sample present in every bubble was retrieved by realigning each of the 16 assemblies back to the graph using the “-cxasm -call” option of Minigraph. Individual allele data were then merged into a single VCF file, following the guidelines set forth in the Minigraph-cookbook-v1 (<https://github.com/lh3/minigraph>) [23]. The resulting VCF file contains the final panel of SVs and records the presence or absence status of SVs in every sample.

Recovery of non-reference unique insertion sequences (NRUIs)

NRUIs were extracted as outlined by Sorin et al. [5]. Briefly, three filters were applied: (1) nodes not present in any individual assembly paths were categorized as nested nodes and subsequently excluded from further analysis, (2) all nodes found in the reference path were also excluded, and (3) nodes corresponding to true bi-allelic insertions (defined as having a reference allele of zero length and an alternative allele exceeding 50 nucleotides) were retained as NRUIs.

Population structure analysis using NRUIs

In order to capture a wider range of variability present in the *Bos taurus* genome, we expanded our NRUIs analysis to incorporate an additional panel of NRUIs derived from a previous pangenome built with 64 de novo assemblies [5]. The final panel of NRUIs consisted of 45,136 NRUIs covering a total length of 48,202,473 nucleotides.

To assess NRUI distribution across a larger French bovine population, we used the Minigraph tool to realign a total of 181 de novo assemblies from 14 French cattle breeds and yak (comprising 165 publicly available de novo assemblies retrieved from projects PRJEB64023 and PRJEB59364, along with the 16 haplotype-resolved assemblies of this study). These assemblies were aligned onto both the pangenomes we developed in this study and the one previously created with 64 de novo assemblies [5].

From the pangenome alignments, a pathway was established for each assembly and the genotype of each animal was predicted for every NRUI in the merged panel. Genotypes for NRUIs common to both panels were extracted from the alignments on the current pangenome, while genotypes or NRUIs specific to the previous panel were obtained from the alignments on the previous pangenome. Subsequently, both sets of genotypes were merged into a single VCF file. A binary presence/absence matrix was generated for each animal and used for hierarchical clustering analysis using the hclust function of the R package [25].

Functional characterization of NRUIs

To investigate the potential functional impact of NRUIs on cattle phenotypes, two complementary analyses were performed.

First, cattle quantitative trait loci (QTL) data from the publicly accessible Animal QTL database [26] with NRUI genomic regions and enrichment analyses were conducted using the GALLO R package [27]. Traits with

an FDR-adjusted P -value < 0.001 were considered significantly enriched.

In the second study, the cattle gene annotation GTF file (Ensembl release 110; https://ftp.ensembl.org/pub/release-110/gtf/bos_taurus/Bos_taurus.ARS-UCD1.2.110.gtf.gz) was retrieved. NRUIs overlapping exons, coding sequences (CDS), introns, and untranslated regions (UTRs) were identified, and the corresponding genes were subjected to Gene Ontology (GO) enrichment analyses using the PANTHER Classification System [28].

Effect of selected NRUIs on recorded traits

The potential phenotypic impact of NRUIs was evaluated using a sequential approach. First, the SV catalogue was filtered to identify NRUIs that were breed-specific and located within annotated gene regions, including both coding and non-coding sequences.

These NRUIs were then further refined using functional annotations, that were obtained in the previous step, to retain only those within known genes associated with milk production and developmental processes. From this filtered set, two candidate NRUIs, specific to the Normande breed and located within the intronic regions of *ARMH3* and *EPHA5*, were selected for proof-of-concept analyses.

To investigate the impact of these two NRUIs on key cattle phenotypes, we applied a haplotype-based testing strategy using a large cohort of animals with available phenotypes and genotypes derived from various low- or medium-density SNP arrays. Long-read sequences for 28 Normande samples were retrieved from the European Nucleotide Archive (projectID PRJEB68295 and PRJEB59364), and realigned to the ARS-UCD1.2 reference genome using the PacBio pbmm2 aligner [29]. SVs were then called using the PacBio pbSV SV caller [30], allowing us to determine the true genotypes of the 28 samples for the two NRUIs of interest (see Additional file 2: Table S2).

After identifying NRUI carriers (cases) and non-carriers (controls) based on sequence-derived genotypes, we screened a panel of 241,413 Normande animals genotyped for routine genomic evaluation to identify haplotypes of markers that exhibit strong linkage

disequilibrium (LD) with the candidate NRUIs. Two filters were applied: (1) all cases had to possess the haplotype, and (2) the number of controls having this haplotype was minimized. These two filters substantially decreased the number of false positives (i.e., individuals carrying the haplotype but not the NRUI). Although a limited number of false negatives persisted (NRUI carriers classified as controls), the LD between the haplotypes and the candidate NRUIs was high, ensuring reliable subsequent analysis.

In order to evaluate the haplotype-based prediction strategy and to confirm the concordance between the haplotype status and true genotypes, the NRUI insertion positions were confirmed through PCR amplification and agarose gel electrophoresis. DNA from six animals were used to performed this cross-validation: two heterozygotes, two homozygous carriers, and two non-carrier. Primers were designed using primer3 version 4.1.0 [31–33] (Table 1) and synthesized by Eurofins Genomics (Ebersberg, Germany). They were specifically designed to span the NRUI breakpoints, producing two amplicons of different sizes corresponding to each allele. PCR amplification was performed using previously described protocol [5] and PCR products were analysed through gel electrophoresis on a 2% agarose gel. For each NRUI, non-carriers were expected to exhibit the shortest amplicon, homozygous carriers the largest amplicon, and heterozygotes both types of amplicons.

Sanger sequencing was also performed on DNA from one animal homozygous for the haplotype. The PCR product was sequenced by Eurofins Genomics (Ebersberg, Germany) to determine the exact genomic insertion location of each examined NRUI.

The effects of the haplotypes in heterozygous and homozygous states were estimated for 14 morphological and production traits routinely recorded for selection purposes as described in Besnard et al. [34]. To account for environmental factors influencing these phenotypes, we used yield deviation (YD) data, i.e., phenotypes adjusted for non-genetic effects as estimated in the national genomic evaluation models. YD are a by-product of the official genomic evaluations carried out by GenEval for French breeding organizations [35].

Table 1 Features of PCR primers

NRUI	Primer type	Primer sequence (5'→3')	0/0 amplicon size, bp	0/1 amplicon size, bp	1/1 amplicon size, bp
<i>ARMH3</i>	Forward	GCATTCAAGAAGTAGGAAAGGGA	303	303 and 356	356
	Reverse	TGGGCTCTCTAGTGGGATCT			
<i>EPHA5</i>	Forward	ATCTTCATTGTGTCATGCAAGATCTTTC	418	418 and 502	502
	Reverse	AAGGAAATCAAACAGTCAATTGTAAAG			

YD were analysed using the blupf90 family of programs, specifically renumf90 followed by blupf90. A mixed model was fitted with haplotype and year of birth as fixed effects, and an animal random polygenic effect. Overall, sample sizes ranged from 13,990 Normande cows for teat length and rear teat placement to 119,842 Normande cows for heifer non-return rate (see Additional file 2: Table S3).

Trait means between group of genotypes were compared using Student's *t*-test, and *P*-values were adjusted for multiple testing using the Benjamini–Hochberg correction in R software. To facilitate comparisons across traits, haplotype effects were expressed in genetic standard deviations (GSD), as estimated for the national genomic evaluations.

Prediction of novel functional features in NRUIs

To identify novel functional elements in NRUI sequences, yak-specific NRUIs were excluded and then repetitive sequences were masked using RepeatMasker v4.1.5 with the parameter option “-species cow” [36]. Ab initio gene structure prediction was then carried out on the masked sequences using the Augustus version 3.3.3 tool [37] with default parameters trained on the human genome. This facilitated the identification of NRUIs that contained potential gene structures. Amino acid sequences of the predicted genes were extracted using the “getAnnoFasta.pl” perl script from the Augustus package and subsequently compared against the cattle protein database to identify reference-like gene coding proteins.

In parallel, masked NRUI sequences were compared against a multi-species protein database comprising three Bovidae (*Bos taurus*, *Bos grunniens*, *Bos mutus*), 11 other mammalian species (*Capra hircus*, *Ovis aries*, *Bison bison bison*, *Camelus dromedaries*, *Equus caballus*, *Delphinapterus leucas*, *Balaenoptera musculus*, *Homo sapiens*, *Mus musculus*, *Physeter catodon*, and *Tursiops truncatus*), along with the curated protein databases of SwissProt (<https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/swissprot.gz>) and PDB (<https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/pdbaa.gz>). Comparisons were performed using BlastX implemented in the Diamond-version 0.9.30 tool [38] enabling the identification of both reference-like protein-coding genes and novel genes encoding proteins known in other species but not yet annotated in cattle.

Finally, the expression profiles of the predicted novel genes were assessed using a total of 71 raw RNA-seq data from blood and milk cells downloaded from the NCBI Short-Read Archive database (see Additional file 2: Table S4). This analysis enabled the validation of these predicted genes as truly expressed genes in the species under study. The raw RNAseq data were processed with

the nf-core/RNAseq pipeline (<https://nf-co.re/rnaseq/3.21.0/>), aligned to the NRUI sequences using STAR [39] and quantified using RSEM [40] along with the GTF gene annotation output file produced by Augustus.

Results

Construction of haplotype-resolved genome assemblies

Sequencing of the seven cattle samples yielded a total of 793 Gb of raw PacBio HiFi long-read sequences, achieving an average sequence depth of 41X per genome. For each sample, two haplotype-resolved genome assemblies were constructed; resulting in the production of 14 assemblies with an average scaffold N50 value of 89.9 Mb and an average L50 value of 13 scaffolds (Table 2). The mean QV score was 61.5 and ranged from 54.5 for the VOS_hap1 up to 70.3 for Yak_hap1 (Table 2). Also, the mean assembled genome size was 3.03 Gb and ranged from 2.7 Gb for the ABO_hap1 assembly to 3.26 GB for the VOS_hap2 assembly. Together, these metrics indicate that the genome assemblies were highly continuous and displayed minimal fragmentation. The mean completeness of genome assemblies was 95.4%, with individual completeness scores varying from 86.7% to 98.8%, confirming their high overall quality. Finally, the alignment of these assemblies with D-Genies demonstrated a high degree of similarity and concordance with the chromosomes of the ARS-UCD1.2 reference genome assembly (see Additional file 1: Fig. S1–S8).

Construction of a bovine pangenome graph

Integration of the 16 haplotype-resolved assemblies using the autosomal sequences of the Hereford ARS-UCD1.2 reference genome as a backbone produced a bovine pangenome graph with 273,192 nodes linked by 388,858 edges, totalling 2,569,341,538 nucleotides. The core component of the pangenome consisted of 101,248 nodes spanning 2,435,570,613 bases, whereas the flexible region comprised 170,648 nodes corresponding to 132,903,963 nucleotides. A small subset of 1,296 nodes, representing 866,962 nucleotides, could not be traversed by any of the assemblies. These likely reflect limitations in accurately mapping certain complex graph structures rather than true biological absence, consistent with previous reports [5]. Such unresolved nodes primarily occurred within highly nested or structurally complex bubbles in the graph and were therefore excluded from downstream analyses.

Discovery of SVs from the pangenome graph

Analysis of the pangenome graph identified 101,219 SVs, of which 95,137 variants (94%) were classified as biallelic, while only 6,082 SVs (6%) were found to be multi-allelic. Biallelic SVs were predominantly

Table 2 Metrics of the 16 assemblies used to construct the pangenome graph

Trios*	Haplotypes	Size	N50	L50	QV	Compleasm
ABO	1	2,799,434,593	82,192,957	13	57.4	86.7
	2	2,980,096,468	91,257,097	13	57.8	89.5
AUB	1	3,033,007,949	88,081,180	14	66.7	91.7
	2	3,168,610,677	98,421,325	13	67.1	93.6
BAQ	1	2,701,070,261	88,084,713	12	65.1	87.0
	2	3,260,617,466	86,982,281	14	63.7	93.3
CHA	1	3,156,038,177	91,997,377	13	47.76	98.6
	2	3,113,499,445	90,512,730	13	47.25	98.2
HN	NMD_1	3,172,052,848	85,091,105	14	60.3	93.6
	HOL_2	3,092,355,700	93,697,923	13	61.6	93.5
TAR	1	2,951,668,414	84,449,537	13	55.4	92.4
	2	2,944,315,815	92,533,001	13	56.1	84.7
VOS	1	2,730,743,480	85,932,234	12	54.5	84.4
	2	3,264,165,777	86,842,039	14	55.8	93.4
YM	YAK_1	3,140,988,036	93,478,221	13	70.3	93.5
	MON_2	3,165,380,519	101,716,749	13	69.6	93.6

* ABO Abondance, AUB Aubrac, BAQ Blonde d'Aquitaine, CHA Charolais, HOL Holstein, MON Montbéliarde, NMD Normande, TAR Tarentaise, VOS Vosgienne

Table 3 Distribution of SVs by allele count and SV type

Mutations	Bi-allelic SVs		Multi-allelic SVs	Total
Insertions	43,096	33,855* 9,241**	609	43,705
Deletions	33,422	26,286* 7,136**	778	34,200
Others variants***	18,619		4,695	23,314
Total	95,137		6,082	101,219

* Alternative allele sequence length = 0 for deletion and reference allele sequence length = 0 for insertions; ** Alternative allele sequence length between 1 and 5 nucleotides for deletion and reference allele sequence length between 1 and 5 nucleotides for insertions; *** All others SVs not classified as insertions nor deletions

composed of insertions (45.3%) and deletions (35.1%), while the remaining SVs (19.6%) corresponded to sequence substitutions, defined by the presence of non-zero sequence lengths for both reference and alternate allele (Table 3).

To further characterise the SVs, we analysed the size distribution of insertions and deletions and observed very similar profiles. The vast majority of variants were small, ranging from 50 to 2,000 nucleotides (Fig. 1). Furthermore, both deletions and insertions exhibited distinct peaks at approximately 150–300 bp, 1.5 kb, and 8.5 kb, which are likely due to the activity of specific transposable element classes of SINES (mainly Bov_A2 at the 150–300 bp peak) and LINEs (mainly BovB at the 1.5 kb peak; and the L1_BT class at the 8.5 kb peak).

Recovery of NRUIs from the pangenome graph

Examination of true insertions from biallelic bubbles, excluding reference genome nodes, identified 33,634 non-reference unique insertions (NRUIs), totalling 34,405,480 bases. Of these, 16,614 NRUIs were specific to yak, spanning 14,663,556 bases, while the remaining 17,020 NRUIs were present in at least one *Bos taurus* breed, covering 19,741,924 bases.

Comparison of the 17,020 *Bos taurus* NRUIs with the panel constructed in our previous pangenome study based on 64 de novo genome assemblies revealed the identification of 9,481 common unique insertions (55.6%) representing 6,038,185 nucleotides (30.6%). Out of these, 94.4% (8,950 NRUIs covering 5,542,808 bases) showed identical insertion coordinates in both panels.

Because these two panels were developed from genetically diverse populations of French cattle, we merged these NRUIs to form a single panel, allowing us to capture a wider spectrum of genetic diversity found within the genome of French cattle breeds. The resulting merged NRUI panel contained 28,998 unique insertions spanning a total length of 33,742,494 nucleotides. This panel was subsequently used to genotype a broader population consisting of 181 animals originating from 14 French cattle breeds and then investigate their frequencies at the population level.

The vast majority of NRUIs had a low frequency within the studied population (Fig. 2A). Approximately 42% had frequencies of 5% or less. Conversely, 124 NRUIs have been observed in at least 99% of the animals but were absent from the reference genome. These likely reflects

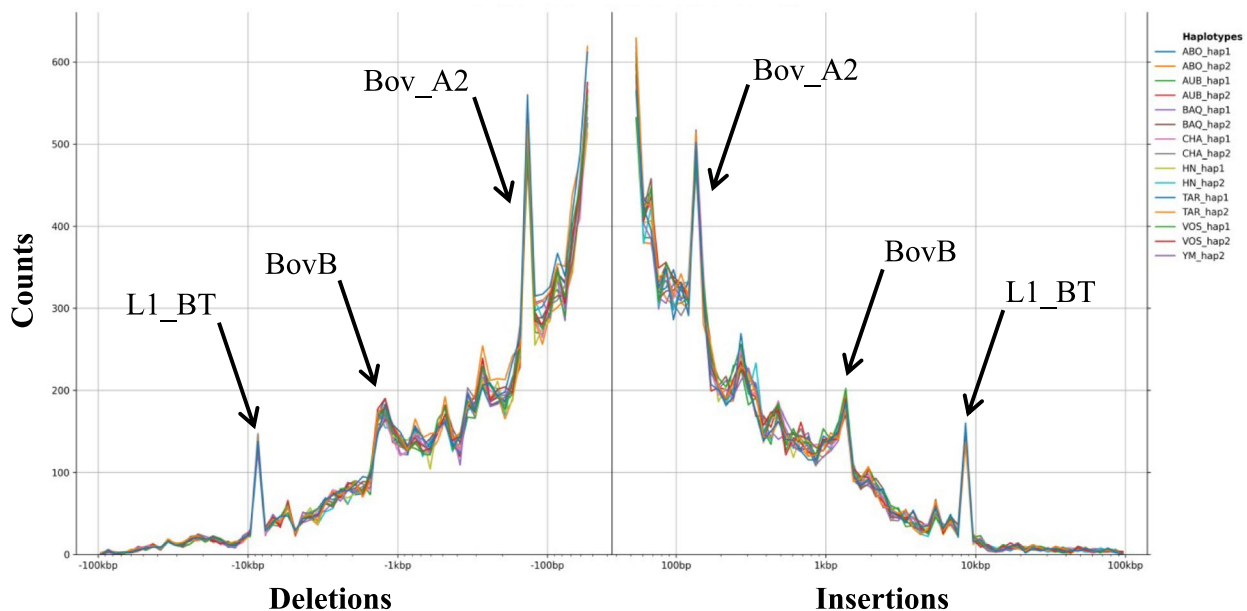


Fig. 1 Size distribution of INS and DEL. Size distribution of SVs classified as deletions (left) and insertions (right), identified using Minigraph across the 16 phased haplotypes. The arrows highlight peaks at 150–300 bp, 1.5 kb and 8.5 kb, which are likely driven by the activity of specific transposable element classes

deletions specific to the Hereford genome or to the animal (Dominette) used to construct the reference genome sequence. Furthermore, 5,311 NRUIs were found to be private as they were observed in only one breed (Fig. 2B). These private NRUIs covered a total of 10.4 Mb, with breed-specific totals ranging from 243.8 kb in Limousine to 1.4 Mb in Holstein.

Functional characterization of NRUIs

The merged panel of NRUIs was subsequently used to characterize their functional properties by performing GO enrichment analysis for genes whose exonic, intronic, or regulatory regions overlapped with NRUIs, and QTL enrichment and association analyses to explore their impact on economically important cattle traits.

Most NRUIs (69.1%) were identified in intergenic regions. The remaining 8,951 NRUIs (30.9%) were located within 4,348 unique cattle gene regions. Among these, 8,641 NRUIs (29.8%) were found within intronic regions, while only 212 NRUIs (0.73%) were identified within exons of 201 genes and 98 NRUIs (0.34%) were located within the 5' or 3' UTR regions of 91 unique genes.

Gene enrichment analyses using Panther classification [28] revealed that 1,712 genes intersecting with NRUIs were unclassified, lacking both Panther molecular function and protein class annotation. Among the classified genes, several GO molecular function terms (see Additional file 2: Table S5) were found to be significantly over-represented, namely binding (GO:0005488),

catalytic activity (GO:0003824), transporter activity (GO:0005215), molecular function regulator activity (GO:0098772), and transcription regulator activity (GO:0140110; Fig. 3A). The three most significant protein classification categories corresponded to metabolite interconversion enzymes (PC00262), protein modifying enzyme (PC00260), and transporter class (PC00227; Fig. 3B) (see Additional file 2: Table S6). Finally, pathway analysis identified Wnt signalling pathway (P00057), and Gonadotropin-releasing hormone receptor pathway (P06664) as the two most over-represented pathways (Fig. 3C) (see Additional file 2: Table S7).

QTL enrichment analyses revealed that NRUIs were significantly enriched in QTL regions associated with 24 economically important traits in cattle (Fig. 4). The majority of NRUIs overlapping QTL were related to milk production (59.5%), with the most significant associations observed for milk fat percentage (26.0%), followed by milk protein percentage (21.3%), and milk yield (16.3%).

Impact of NRUIs on cattle key phenotypes

Combined gene annotation efforts along with QTL enrichment and analyses of the distribution of allele frequencies within breeds enabled us to identify 1,660 breed-specific NRUIs located within candidate genes. Of particular interest, two NRUIs have been found to be specific yet polymorphic (MAF=9%) within the Normande breed, and were located within the intronic regions of two genes of interest, namely *ARMH3* and *EPHA5*.



Fig. 2 Distribution of NRUIs within the cattle breeds. **A** Distribution of NRUI frequencies at the population level. **B** Distribution of the private NRUIs. ABO= Abondance, AUB= Aubrac, BAQ= Blonde d'Aquitaine, CHA= Charolais, HOL= Holstein, MON= Montbéliarde, NMD= Normande, TAR= Tarentaise, VOS= Vosgienne

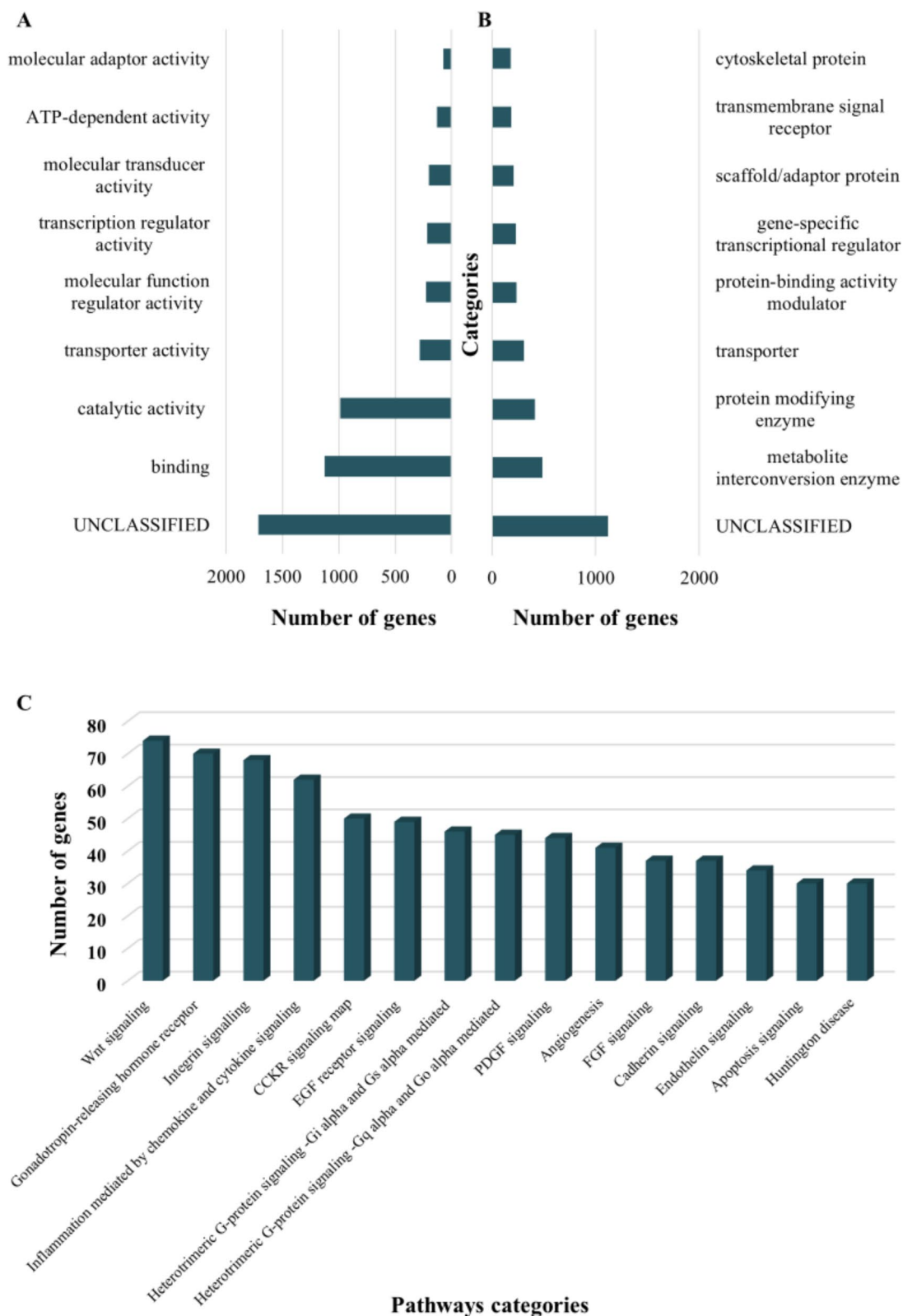


Fig. 3 Functional implications of SVs. **A** Molecular enrichment analysis of the NRULs panel. **B** Protein classes enriched in the NRULs panel. **C** Representation of the 15 most enriched pathways in the NRULs panel

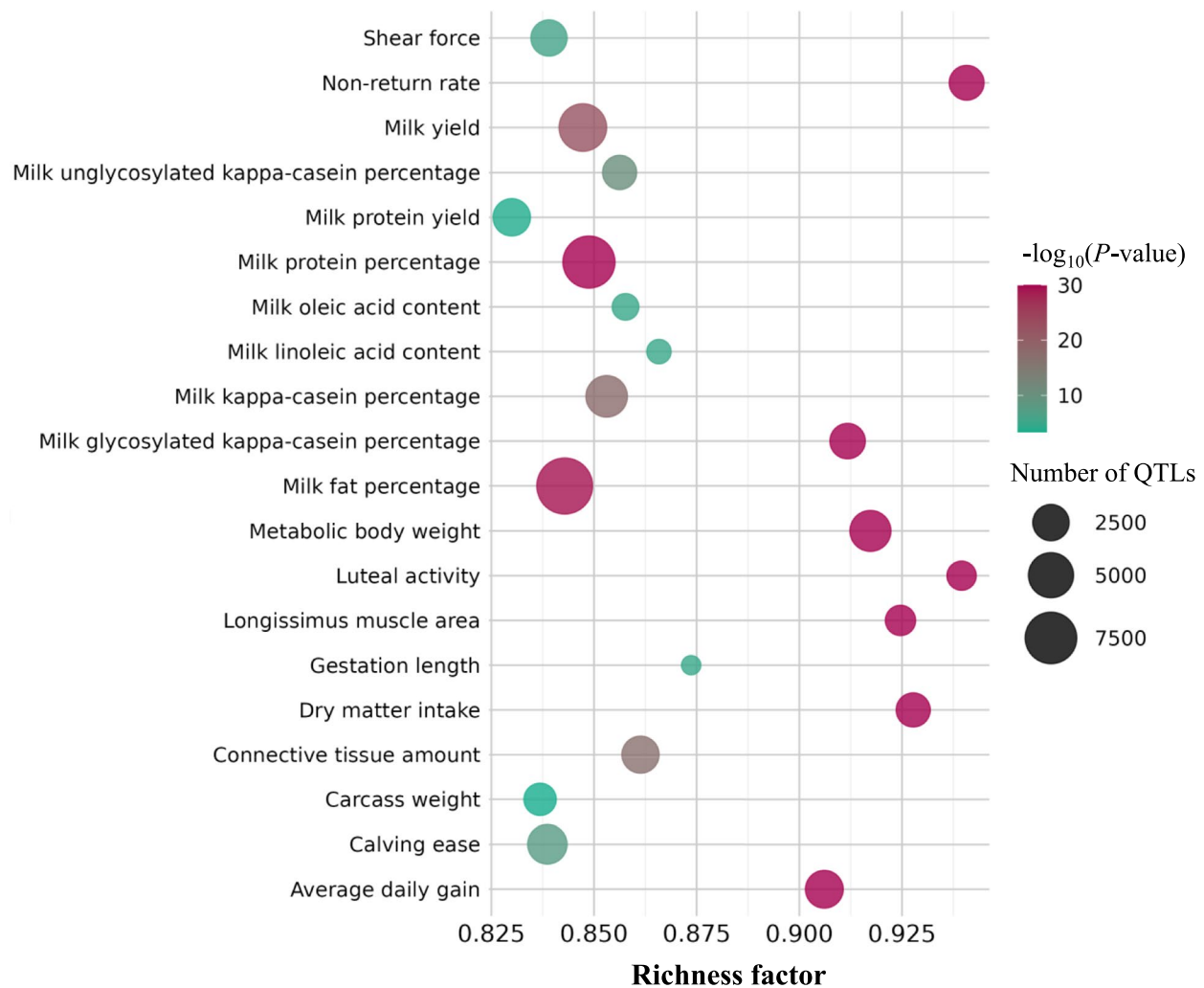


Fig. 4 QTL enrichment analysis of NRUIs. Bubble plot of the enrichment analysis of the top 20 QTLs impacted by NRUIs

The first NRUI is 53 nucleotides in length and is inserted on chromosome 26 at position 22,672,107 within intron 22 (out of 25) of the *ARMH3* gene. The second NRUI is 84 nucleotides in length and is located on chromosome 6 at position 81,086,501 within intron 3 (out of 16) of the *EPHA5* gene. In this study, we investigated their potential impact on several economically important traits in cattle by first identifying haplotypes in complete LD with the candidate NRUIs. To assess the reliability of these haplotypes, we compared their inferred status in six selected animals (2 homozygous reference, 2 heterozygous and 2 homozygous alternate) with the true genotypes determined by PCR amplification of genomic DNA and validated the exact genomic positions of the NRUIs by Sanger sequencing (see Additional file 3). The perfect concordance between haplotype status and PCR-derived genotypes supports that these haplotypes are in strong LD with the two variants and that carriers of the

haplotypes are also carriers of the NRUIs. Moreover, visual inspection of the NRUIs and their flanking regions in the pangenome were performed using BandageNG.

Following this, we performed a Student's *t*-test to investigate potential associations between these two NRUIs and significant phenotypes in cattle using a panel of Normande cows ranging from 13,990 to 119,842 depending of the trait (see Additional file 2: Table S3). The NRUI in the intronic regions of *ARMH3*, showed a clear dosage-dependent pattern for stature, with heterozygotes increasing height by 0.25 GSD and homozygotes by 0.50 GSD (Fig. 5A). Additive effects were also detected for chest depth (0.18 GSD) and udder depth (0.25 GSD) (Fig. 5B), as well as milk protein (−0.17 GSD) and fat content (−0.12 GSD) (Fig. 5C and D). No significant effect on milk yield was observed (see Additional file 2: Table S3).

For the NRUI located with the *EPHA5* intronic regions, a strong additive effect was observed on milk

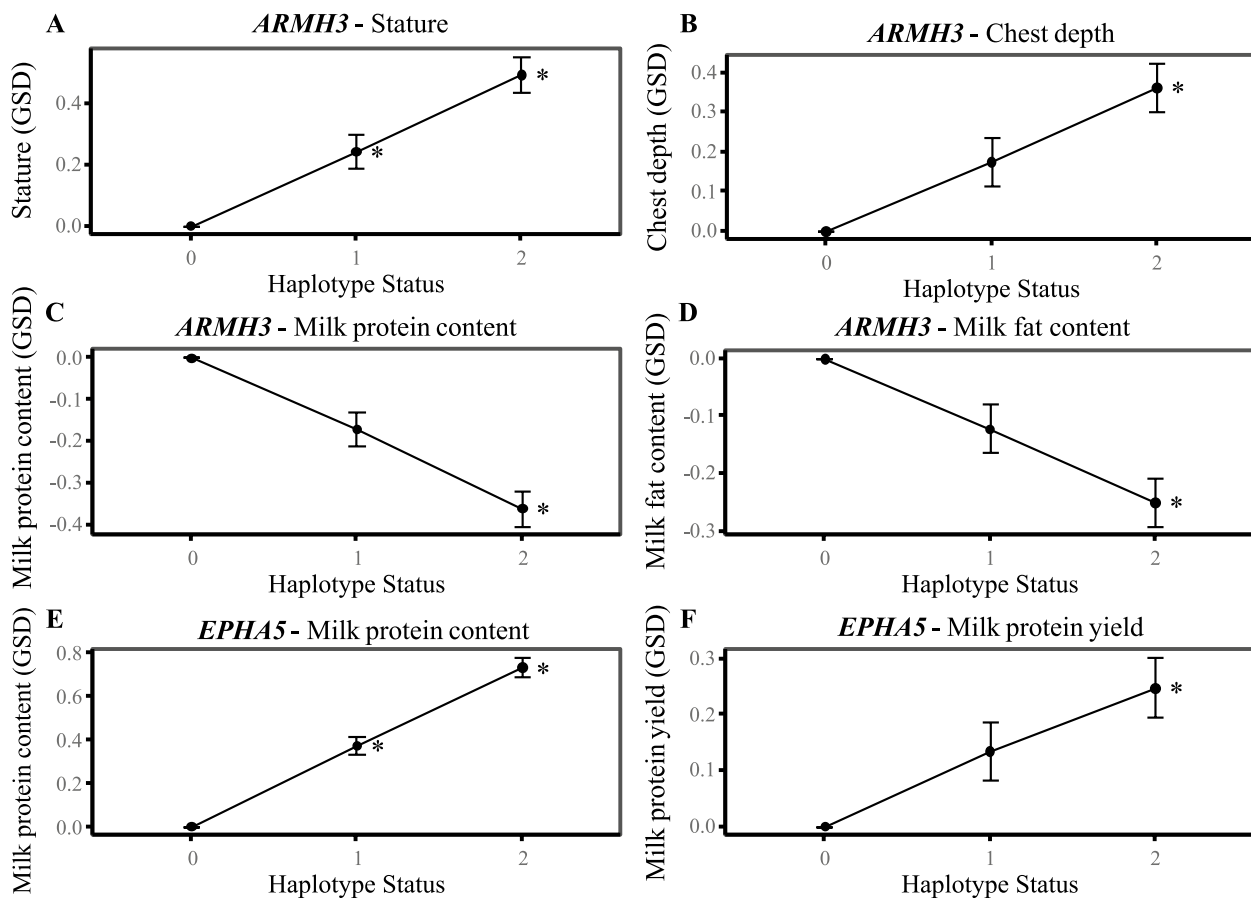


Fig. 5 Effects of NRUIs on cattle phenotypes. **A** Effect of *ARMH3* NRUI on stature trait. **B** Effect of *ARMH3* NRUI on chest depth trait. **C** Effect of *ARMH3* NRUI on milk protein content trait. **D** Effect of *ARMH3* NRUI on milk fat content trait. **E** Effect of *EPHA5* NRUI on milk protein content trait. **F** Effect of *EPHA5* NRUI on milk protein yield trait. GSD: genetic standard deviations, 0: homozygous for reference allele, 1: heterozygous, and 2: homozygous for the NRUI allele

composition traits. The effect on milk protein content was particularly pronounced, reaching 0.37 GSD at the homozygous state (Fig. 5E). We also observed a moderate but consistent additive effect on milk protein yield (0.13 GSD), which was only significant at the homozygous state (Fig. 5F). A smaller yet positive effect was also detected on milk fat content (−0.14 GSD at the homozygous state) (see Additional file 2: Table S3).

Prediction of novel functional elements in NRUIs

Repeat annotation of NRUIs revealed that 45.34% (15.3 Mb) were repeated sequences (see Additional file 2: Table S8). Most of these were interspersed repeats (39.9%), which primarily consist of long interspersed nuclear elements (LINEs, 36.5%), long terminal repeats (LTR, 2.25%), DNA elements (0.7%), and short interspersed nuclear elements (SINEs, 0.5%).

Functional annotation of NRUIs using Augustus gene structure prediction identified a total of 923 NRUIs that

encompassed 970 novel genes, of which 429 were predicted to be complete gene models with an average size of 310 amino acids and ranging from 67 up to 1,182 amino acids (see Additional file 2: Table S9). For these complete gene models, a transcription start site, a start codon, at least one exon, one stop codon, and a transcription termination site were predicted. Among the 429 complete genes, 173 demonstrated homology to known bovine proteins (bovine-like genes), whereas 251 genes showed homology with proteins from at least one species other than *Bos taurus* (true novel genes). Five complete genes had no detectable homology (see Additional file 2: Table S11). Among the five genes, two were identified as being expressed in at least one RNA-Seq sample, qualifying them as true novel genes. The first gene displayed a transcripts per kilobase million (TPM) of 19.91, while the second gene was expressed across five RNA-Seq samples, with a TPM varying from 5.43 to 11.73 (see Additional file 2: Table S11). In contrast, the remaining three genes did not exhibit any expression profiles in the RNA-Seq samples. This lack of

expression may indicate either false predictions by Augustus or that these genes are not expressed in the tissues from which the 71 RNA samples were derived.

Expression validation using 71 RNA-seq datasets demonstrated that 802 of the 970 genes (82.6%) were expressed in at least one of the RNA-Seq samples (see Additional file 2: Table S10). Among these, 290 genes (~36.2%) were expressed in over 90% of the samples, suggesting their potentially significant role in cellular processes, while 155 genes (19.3%) were expressed in fewer than 10% of the samples, indicating potential involvement in specific or rare biological processes.

Discussion

In comparison to linear reference genome sequences, graph-based pangenomes facilitate a more comprehensive detection and representation of genomic sequences and variations, resulting in improved read alignment rates with the reference and enhancing the ability to capture missing heritability within populations.

In this study, we constructed a cattle pangenome graph using 16 haplotype-resolved genome assemblies originating from 8 French cattle breeds (Abondance, Aubrac, Blonde d'Aquitaine, Charolais, Holstein, Montbéliarde, Normande, Tarentaise, and Vosgienne) and a closely related species, yak. We focused solely on autosomal sequences since the haplotype-resolved assemblies depict either paternal or maternal haplotypes, thereby excluding sex chromosomes (X and Y), mitochondrial DNA, as well as the unplaced contig chromosomal sequences.

The resulting bovine pangenome graph encompassed 2.57 Gb distributed across 273,192 nodes interconnected by 388,858 edges. The core genome represented 94.7% (2.44 Gb) of the graph whereas the flexible region accounted for only 5.1% (133 Mb) of the pangenome. These results were very similar to those previously reported by two independent pangenome studies that also considered only autosomes and used Minigraph to build the bovine pangenome [3, 6]. However, our current pangenome exhibited a significantly lower number of nodes (273,192 compared to 507,822) and a smaller overall size (2.57 vs. 2.86 Gb) than our previous pangenome [5]. The core (90%) and flexible (10%) genomes were also slightly lower than those reported in the same study probably due to the exclusion of the X chromosome and a lower number of assemblies (16 vs. 64 individuals). Furthermore, the study by Dai et al. [41], which used 20 pseudo-phased HiFi assemblies from Chinese indicine cattle, reports even smaller core genome proportions (78%) and larger flexible regions (22%), reflecting their greater genetic divergence from taurine breeds.

Systematic analysis of the flexible pangenome regions identified 101,219 SVs, of which ~92% were bi-allelic,

predominantly insertions and deletions (80%). The remaining SVs, with non-zero reference and alternative allele lengths, were categorised as sequence substitutions.

Comparison with our previous SV panel from 64 assemblies revealed the identification of common 28,488 SVs (28%) with identical coordinates (same start and end positions), and 32,507 SVs (32%) showing at least 90% reciprocal overlap. Size distributions of insertions and deletions were consistent with two previous studies [5, 11], showing most variants were small (50 nucleotides to 2 kb) and exhibiting peaks at 150–300 bp, 1.5 kb, and 8.5 kb, corresponding to transposable elements. Our results corroborate previous findings which highlighted how transposable elements contribute to structural variation and demonstrated that deletion and insertion events significantly contribute to genomic differences in other species [42, 43].

Several SVs detected in our panel are known to be associated with phenotypic traits in cattle. For example, in the Normande breed, we detected an 8.4 kb LINE1 insertion at the ASIP locus on BTA13, which encodes the AGOUTI signalling protein that plays a role in mammalian pigmentation [44]. Additionally, segmental duplications upstream of the *KIT* gene were observed in Abondance, Normande, Montbéliarde, and Simmental breeds, consistent with depigmentation in white-headed breeds [4, 5].

By systematically analysing the SV panel, we further identified 33,634 NRUIs covering a total of 34.4 Mb, of which 14.7 Mb were specific to the yak genome. The remaining 19.7 Mb were identified in at least one *Bos taurus* breed. Approximately 56% of these *B. taurus* NRUI sequences were shared with our earlier panel, which was constructed using a cattle pangenome developed from 64 de novo assemblies. This led us to combine panels in order to investigate the spectrum of genetic diversity within French cattle breeds. The hierarchical classification based on the genotyping of these NRUIs across a panel of 181 animals from 14 French dairy and beef breeds revealed that all samples were grouped into 14 clusters corresponding to their respective breeds of origin (see Additional file 4). These results are of particular interest as they demonstrate the potential of these NRUIs to provide a very clear description of the breed structure. All samples belonging to the same breed are grouped together, without exception. Furthermore, this classification has allowed us to pinpoint four primary clusters of breeds. The first group corresponds to the Holstein breed, which originated in northern Europe and was more recently reintroduced from North America. The second group corresponds to the Normande breed, which hails from the Channel area and is significantly influenced by Great Britain. The third group encompasses

six breeds (Brown Swiss, Montbéliard, Simmental, Vosgienne, Tarentaise, and Abondance), all of which originated from the eastern Alpine regions. It is important to note, however, that the classification of the Montbéliard breed is further from what was expected. Lastly, the final group comprises three beef breeds (Limousine, Aubrac, and Blonde d'Aquitaine) that originate from the southwestern region of France and are related to certain Spanish breeds. Additionally, it is worth mentioning that the Charolais breed, which is geographically situated in Central France, is classified between the Alpine and South-West groups, highlighting their dual influence.

Combined annotation efforts along with QTL enrichment and analyses of the distribution of NRUIs between breeds enabled us to pinpoint several NRUIs located within introns of candidate genes. Two Normande-specific NRUIs were identified in *ARMH3* (BTA26:22,672,107; 53 bp, intron 22/25) and *EPHA5* (BTA6:81,086,501; 53 bp, intron 3/16). The *ARMH3* NRUI significantly increases morphological traits (stature, skeletal development, and chest depth) while reducing milk protein and fat contents without having any effect on milk yields. Although effect on milk yield was not significant, we observed an additive effect of 0.09 GSD (at the heterozygous state) and 0.15 (at the homozygous states). Several studies have already reported significant associations between variants within the *ARMH3* gene and fertility traits in Spring-calved Dairy Cows in New Zealand [45], milk fatty acids in Dual-Purpose Belgian Blue cows [46], and with carcass merit traits in a multibreed population of beef cattle [47]. Also, a genome analysis of the distribution of ROH patterns in Ayrshire cows revealed the identification of several candidate genes, including *ARMH3*, associated with milk production traits, that are under selection pressure [48].

The *ARMH3* gene has been found as implicated in the regulation and structural organization of the Golgi apparatus [49, 50]. The Golgi complex plays a central role in protein maturation, lipid processing and vesicular export [51]. Because these functions represent a major role for both milk protein and lipid secretion [52] and skeletal tissue development in animals [53], the Golgi is directly involved in resources allocation between milk production and growth. Consequently, a variation within *ARMH3* gene could modulate the normal function of the Golgi apparatus, therefore influencing the allocation of metabolic resources.

The *EPHA5* NRUI BTA6 at position 81,086,501 exerted a strong influence on milk protein content without affecting milk yield. Since the casein gene cluster is located around 85 Mb on the same chromosome, we assessed the linkage disequilibrium (LD) between the NRUI status and the genotypes of the casein markers included in the EuroGMD genotyping array [54]. This analysis aimed

to determine whether the observed effect on the NRUI could be explained by casein gene variants. Across 6 markers in casein genes tested, we observed low to moderate LD values (R^2 ranging from 0.0002 to 0.6442, with an average of 0.1864), suggesting that the detected effect most likely originates from the NRUI in *EPHA5*, rather from the casein region. These results are in agreement with previous studies in cattle and sheep, which reported notable correlations between variants in *EPHA5* and both milk and fat production in dairy cattle adapted to tropical climates [55], as well as in fat deposition in domestic sheep breeds originating from Africa and Eurasia [56]. Another study using iHS analysis has pinpointed several genes, including *EPHA5*, currently under selection. This gene is known to play roles in neural development and synaptic functions, possibly contributing to adaptive traits in diverse environments [57].

Conclusions

This study demonstrates the potential of the pangenome graph approach in comprehensively identifying structural variants, with a particular focus on non-reference unique insertions (NRUIs), providing new insights into genetic diversity within cattle breeds. Our results highlight the potential of pangenomes to enhance reference genomes and uncover genetic variations that are missed using linear assemblies. Importantly, the NRUIs characterized here show strong associations with breed-specific phenotypic traits, providing valuable resources for future research in cattle genetics and breeding programs. It would be very interesting in future work to convert these NRUIs into potentially virtual SNPs, as previously described [58], to add them to the custom EuroGMD SNP chip for genotyping, and to be able to genotype several hundred thousand animals with phenotypes. This data will then be used to explore potential links with phenotypes of interest on a larger scale.

Abbreviations

ABO	Abondance breed
AUB	Aubrac breed
BAQ	Blonde d'Aquitaine breed
CHA	Charolais breed
GSD	Genetic standard deviation
GWAS	Genome-wide association study
HOL	Holstein breed
MON	Montbéliarde breed
NMD	Normande breed
NRUI	Non-reference unique insertion
QTL	Quantitative trait loci
SNP	Single nucleotide polymorphism
SV	Structural variations
TAR	Tarentaise breed
TPM	Transcripts per kilobase million
VOS	Vosgienne breed
YD	Yield deviation

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40104-026-01373-3>.

Additional file 1: Fig. S1. Abundance haplotype 1 (A) and 2 (B). Fig. S2. Aubrac haplotype 1 (A) and 2 (B). Fig. S3. Blonde d'Aquitaine haplotype 1 (A) and 2 (B). Fig. S4. Charolais haplotype 1 (A) and 2 (B). Fig. S5. Holstein-Normande haplotype 1 (A) and 2 (B). Fig. S6. Tarentaise haplotype 1 (A) and 2 (B). Fig. S7. Vosgienne haplotype 1 (A) and 2 (B). Fig. S8. Yak-Montbéliarde haplotype 1 (A) and 2 (B).

Additional file 2: Table S1. The estimated mash distance of haplotype-resolved assemblies to the reference genome and the order of integrating the assemblies during pangenome construction. Table S2. The predicted genotype of each sample to the two NRUIs derived from alignment of LR sequences to the genome followed SV calling. Table S3. Effect of the two NRUIs on cattle key traits. Table S4. The list of public RNA-Seq data used for expression validation of the novel genes predicted by Augustus. Table S5. Significant enrichment of molecular function terms of cattle genes harboring NRUIs using the Panther enrichment system. Table S6. Significant enrichment of protein class terms of cattle genes harboring NRUIs using the Panther enrichment system. Table S7. Significant enrichment of pathway terms of cattle genes harboring NRUIs using the Panther enrichment system. Table S8. Repeats annotated by RepeatMasker. Table S9. The list of novel genes predicted by Augustus. Table S10. The average TPM estimated from RNA-Seq sequence alignment to the Augustus-predicted novel gene sequences. Table S11. The average TPM estimated from RNA-Seq sequence alignment to the 5 genes without detectable homology.

Additional file 3: Fig. S9. (A) Bandage visualisation of the NRUI identified within the *ARMH3* gene. (B) Gel electrophoresis of 5 samples for *ARMH3* NRUI: one homozygous reference sample (40515), two heterozygous samples (40516 and 40522), and two homozygous samples for the NRUI (40517 and 40529). (C) Sanger sequencing of the *ARMH3* NRUI sequence and its breakpoints. Fig. S10. (A) Bandage visualisation of the NRUI located within the *EPHA5* gene. (B) Gel electrophoresis of 6 samples for *EPHA5* NRUI: two homozygous reference samples (40497 and 46499), two heterozygous samples (40516 and 40522), and two homozygous samples for the NRUI (40515 and 40524). (C) Sanger sequencing of the *EPHA5* NRUI sequence and its breakpoints.

Additional file 4: Hierarchical clustering of the 181 de novo assemblies based on the NRUI PAV-matrix. PAV presence/absence variation. Diagram showing the clustering of the 181 assemblies according to the 14 breeds for the 28,998 NRUIs.

Acknowledgements

We are grateful to the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, <https://doi.org/10.15454/1.5572369328961167E12>) for providing computing and storage resources. We thank Claire Kuchly and Caroline Vernet for their help to submit the de novo genome assembly sequences in the ENA database. GeT-PlaGe is a member of France Génomique national infrastructure, funded as part of the "Investissements d'Avenir" program managed by the Agence Nationale de la Recherche (contract ANR-10-INBS-09). Finally, we would like to thank the Eliaze breeding companies that supplied the doses of the sequenced bulls.

Authors' contributions

VS, MB, MPS, GTK, DB and LD conceived the study; VS conducted the analysis and wrote the original draft; MB, MPS, GTK and LD contributed to writing the manuscript; CG, SF and CE collected samples and realized extraction; CG supervised PCR and sequencing validation of NRUIs; AC and MMN helped with LR bioinformatics analysis and SV genotyping; MMN performed hierarchical cluster classification; SF and FB conducted haplotype construction and statistical analyses; CE, JL, CI and CD and DM conceived the experimental design and supervised the technical aspects of the sequencing activity; CK constructed the haplotype-resolved assemblies; All authors reviewed the manuscript; All authors read and approved the final manuscript.

Funding

VS is recipient of a PhD grant from INRAE.

This work was conducted in the H2020 Rumigen project and in the SeqOccln project, which was funded by the Occitanie region, FEDER, and APIS-GENE.

Data availability

The whole genome sequencing raw datasets used in the current study to construct haplotype-resolved assemblies are available in the European Nucleotide Archive (ENA) under projects accession numbers (1) PRJEB64022 and PRJEB101071 for the Abundance, Tarentaise and Vosgienne trios; and (2) PRJEB64023 and PRJEB68296 for Aubrac, Blonde d'Aquitaine, Holstein x Normand and Yak x Montbéliarde Trios. The new haplotype-resolved phased assemblies constructed in this study and used to build the graph are available at: <https://doi.org/10.57745/D4DJYW>. The two Charolais phased assemblies are available at <https://doi.org/10.57745/40STPR>. All other assemblies used in this study for genotyping are available in the ENA under project accession numbers PRJEB68295 and PRJEB59364. The pangenome graph, all paths, the SV VCF file and the panel of NRUIs are available at: <https://doi.org/10.57745/OXASZE>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹ Université Paris Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy en Josas, France. ² Eliaze, 149 Rue de Bercy, 75012 Paris, France. ³ INRAE, GeT-PlaGeGenotoul, 31326 Castanet-Tolosan, France. ⁴ GenPhySE, Université de Toulouse, INRAE, ENVT, 31326 Castanet-Tolosan, France. ⁵ Genotoul BioinfoBioInfoMics, MIAT UR875, Sigene, INRAE, 31326 Castanet-Tolosan, France.

Received: 5 December 2025 Accepted: 30 January 2026

Published online: 09 March 2026

References

- Smith TPL, Bickhart DM, Boichard D, Chamberlain AJ, Djikeng A, Jiang Y, et al. The bovine pangenome consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species. *Genome Biol.* 2023;24:139. <https://doi.org/10.1186/s13059-023-02975-0>.
- Leonard AS, Crysanto D, Fang ZH, Heaton MP, Vander Ley BL, Herrera C, et al. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nat Commun.* 2022;13:3012. <https://doi.org/10.1038/s41467-022-30680-2>.
- Leonard AS, Crysanto D, Mapel XM, Bhati M, Pausch H. Graph construction method impacts variation representation and analyses in a bovine super-pangenome. *Genome Biol.* 2023;24:124. <https://doi.org/10.1186/s13059-023-02969-y>.
- Milia S, Leonard AS, Mapel XM, Bernal Ulloa SM, Drögemüller C, Pausch H. Taurine pangenome uncovers a segmental duplication upstream of KIT associated with depigmentation in white-headed cattle. *Genome Res.* 2025;35:1041–52. <https://doi.org/10.1101/gr.279064.124>.
- Sorin V, Naji MM, Birbes C, Grohs C, Escoufflaire C, Fritz S, et al. Application of a French cattle pangenome, from structural variant discovery to association studies on key phenotypes. *Genet Sel Evol.* 2025;57:61. <https://doi.org/10.1186/s12711-025-01012-x>.
- Crysanto D, Leonard AS, Fang ZH, Pausch H. Novel functional sequences uncovered through a bovine multiassembly graph. *Proc Natl Acad Sci USA.* 2021;118:e2101056118. <https://doi.org/10.1073/pnas.2101056118>.

7. Talenti A, Powell J, Hemmink JD, Cook EAJ, Wragg D, Jayaraman S, et al. A cattle graph genome incorporating global breed diversity. *Nat Commun*. 2022;13:910. <https://doi.org/10.1038/s41467-022-28605-0>.
8. Azam S, Sahu A, Pandey NK, Neupane M, Tassell CP Van. Advancing the Indian cattle pangenome : characterizing non-reference sequences in *Bos indicus*. *J Anim Sci Biotechnol*. 2025;1:21. <https://doi.org/10.1186/s40104-024-01133-1>.
9. Gao Y, Yang L, Kuhn K, Li W, Zanton G, Bowman M, et al. Long read and preliminary pangenome analyses reveal breed-specific structural variations and novel sequences in Holstein and Jersey cattle. *J Adv Res*. 2025;79:137–50. <https://doi.org/10.1016/j.jare.2025.04.014>.
10. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience*. 2020;9:giaa021. <https://doi.org/10.1093/gigascience/giaa021>.
11. Eché C, Iampietro C, Birbes C, Dréau A, Kuchly C, Di Franco A, et al. A *Bos taurus* sequencing methods benchmark for assembly, haplotyping, and variant calling. *Sci Data*. 2023;10:369. <https://doi.org/10.1038/s41597-023-02249-1>.
12. Grohs C. DNA isolation from cattle tissues: blood, semen or any kind of tissues, including ear biopsies. *Protocols.io*. 2024. <https://doi.org/10.17504/protocols.io.kqdg3xojeg25/v1>.
13. Denis E, Grohs C, Donnadieu C, Iampietro C. Validated DNA isolation method ensuring successful long-read sequencing of cattle semen genome. *PLoS ONE*. 2024;19:e0308011. <https://doi.org/10.1371/journal.pone.0308011>.
14. PacBio. Procedure & Checklist – Preparing HiFi SMRTbell® Libraries using the SMRTbell Express Template Prep Kit 2.0. <https://www.pacb.com/wp-content/uploads/HiFi-Library-Preparation-Using-SMRTbell-Express-TPK-2.0-Customer-Training.pdf>, accessed January 6 2026.
15. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol*. 2018. <https://doi.org/10.1038/nbt.4277>.
16. BIOCONDA. Yet another k-mer analyzer. <https://github.com/lh3/yak>. Accessed 2 Dec 2025.
17. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;18:170–5. <https://doi.org/10.1038/s41592-020-01056-5>.
18. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol*. 2022;23:258. <https://doi.org/10.1186/s13059-022-02823-7>.
19. Huang N, Li H. Compleasm: a faster and more accurate reimplement of BUSCO. *Bioinformatics*. 2023;39:btad595. <https://doi.org/10.1093/bioinformatics/btad595>.
20. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol*. 2021;38:4647–54. <https://doi.org/10.1093/molbev/msab199>.
21. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;21:245. <https://doi.org/10.1186/s13059-020-02134-9>.
22. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*. 2018. <https://doi.org/10.7717/peerj.4958>.
23. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs. *Genome Biol*. 2020;21:265. <https://doi.org/10.1186/s13059-020-02168-z>.
24. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. 2016;17:132. <https://doi.org/10.1186/s13059-016-0997-x>.
25. Hierarchical clustering in R: dendrograms with hclust. <https://rdocumentation.org/packages/stats/versions/3.6.2>. 2019. Accessed 2 Dec 2025.
26. Hu ZL, Park CA, Reecy JM. Bringing the Animal QTLdb and CorDB into the future: meeting new challenges and providing updated services. *Nucleic Acids Res*. 2022;50:D956–61. <https://doi.org/10.1093/nar/gkab1116>.
27. Fonseca PAS, Suárez-Vega A, Marras G, Cánovas Á. GALLO: an R package for genome annotation and integration of multiple data sources in livestock for positional candidate loci. *Gigascience*. 2020;9:giaa149. <https://doi.org/10.1093/gigascience/giaa149>.
28. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L-P, Mi H. PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci*. 2022;31:8–22. <https://doi.org/10.1002/pro.4218>.
29. BioSciences P. pbmm2 – A minimap2 frontend for PacBio native data formats. <https://github.com/PacificBiosciences/pbmm2>. Accessed 2 Dec 2025.
30. BioSciences P. pbsv – Structural variant calling and analysis tools. <https://github.com/PacificBiosciences/pbsv>. Accessed 2 Dec 2025.
31. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3-new capabilities and interfaces. *Nucleic Acids Res*. 2012;40:e115. <https://doi.org/10.1093/nar/gks596>.
32. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. *Bioinformatics*. 2007;23:1289–91. <https://doi.org/10.1093/bioinformatics/btm091>.
33. Koressaar T, Lepamets M, Kaplinski L, Raime K, Andreson R, Remm M. Primer3_masker: integrating masking of template sequence with primer design software Bioinformatics. 2018;34:1937–8. <https://doi.org/10.1093/bioinformatics/bty036>.
34. Besnard F, Guintard A, Grohs C, Guzylack-Piriou L, Cano M, Escoufflaire C, et al. Massive detection of cryptic recessive genetic defects in dairy cattle mining millions of life histories. *Genome Biol*. 2024;25:248. <https://doi.org/10.1186/s13059-024-03384-7>.
35. Geneval. Evaluations génétiques des animaux d'élevage. <https://geneval.fr/>. Accessed 2 Dec 2025.
36. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*. 2009;25:4.10.1-4.10.14. <https://doi.org/10.1002/0471250953.bi0410s25>.
37. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res*. 2006;34:435–9. <https://doi.org/10.1093/nar/gkl200>.
38. Buchfink B, Xie C, Huson D. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2014;12:59–60. <https://doi.org/10.1038/nmeth.3176>.
39. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
40. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323. <https://doi.org/10.1186/1471-2105-12-323>.
41. Dai X, Bian P, Hu D, Luo F, Huang Y, Jiao S, et al. A Chinese indicine pangenome reveals a wealth of novel structural variants introgressed from other Bos species. *Genome Res*. 2023;33:1284–97. <https://doi.org/10.1101/gr.277481.122>.
42. Munasinghe M, Read A, Stitzer MC, Song B, Menard CC, Ma KY, et al. Combined analysis of transposable elements and structural variation in maize genomes reveals genome contraction outpaces expansion. *PLoS Genet*. 2023;19:e1011086. <https://doi.org/10.1371/journal.pgen.1011086>.
43. Mun S, Kim S, Lee W, Kang K, Meyer TJ, Han BG, et al. A study of transposable element-associated structural variations (TASVs) using a de novo-assembled Korean genome. *Exp Mol Med*. 2021;53:615–30. <https://doi.org/10.1038/s12276-021-00586-y>.
44. Girardot M, Guibert S, Laforet MP, Gallard Y, Larroque H, Oulmouden A. The insertion of a full-length *Bos taurus* LINE element is responsible for a transcriptional deregulation of the Normande Agouti gene. *Pigment Cell Res*. 2006;19:346–55. <https://doi.org/10.1111/j.1600-0749.2006.00312.x>.
45. Jayawardana JMDR, Lopez-Villalobos N, McNaughton LR, Hickson RE. Genomic regions associated with milk composition and fertility traits in spring-calving dairy cows in New Zealand. *Genes*. 2023;14:860. <https://doi.org/10.3390/genes14040860>.
46. Atashi H, Chen Y, Wilmot H, Vanderick S, Hubin X, Soyeurt H, et al. Single-step genome-wide association for selected milk fatty acids in dual-purpose Belgian Blue cows. *J Dairy Sci*. 2023;106:6299–315. <https://doi.org/10.3168/jds.2022-22432>.
47. Li J, Wang Y, Mukiibi R, Karisa B, Plastow GS, Li C. Integrative analyses of genomic and metabolomic data reveal genetic mechanisms associated with carcass merit traits in beef cattle. *Sci Rep*. 2022;12:3389. <https://doi.org/10.1038/s41598-022-06567-z>.
48. Ryabova AE, Pozovnikova MV, Dementieva NV, Shcherbakov YS, Tulinova OV, Romanova EA, et al. Analysis of the genetic diversity of Ayrshire cattle in Russia. *Message 2. Genome analysis based on data on the distribution of ROH patterns in Ayrshire cows*. *Ecol Genet*. 2023;21:235–48. <https://doi.org/10.17816/ecogen568871>.

49. Fang R, Jiang Q, Jia X, Jiang Z. ARMH3-mediated recruitment of PI4KB directs Golgi-to-endosome trafficking and activation of the antiviral effector STING. *Immunity*. 2023;56:500–515.e6. <https://doi.org/10.1016/j.immuni.2023.02.004>.
50. Ishida M, Golding AE, Keren-Kaplan T, Li Y, Balla T, Bonifacino JS. ARMH3 is an ARL5 effector that promotes PI4KB-catalyzed PI4P synthesis at the trans-Golgi network. *Nat Commun*. 2024;15:10168. <https://doi.org/10.1038/s41467-024-54410-y>.
51. Teasdale RD, Jackson MR. Signal-mediated sorting of membrane proteins between the endoplasmic reticulum and the Golgi apparatus. *Annu Rev Cell Dev Biol*. 1996;12:27–54. <https://doi.org/10.1146/annurev.cellbio.12.1.27>.
52. Keenan TW, Saacke RG, Patton S. Prolactin, the Golgi apparatus, and milk secretion: brief interpretive review. *J Dairy Sci*. 1970;53:1349–52. [https://doi.org/10.3168/jds.S0022-0302\(70\)86396-2](https://doi.org/10.3168/jds.S0022-0302(70)86396-2).
53. Wilson KM, Jagger AM, Walker M, Seinkmane E, Fox JM, Kröger R, et al. Glycans modify mesenchymal stem cell differentiation to impact on the function of resulting osteoblasts. *J Cell Sci*. 2018;131:jcs209452. <https://doi.org/10.1242/jcs.209452>.
54. Sanchez MP, Fritz S, Patry C, Delacroix-Buchet A, Boichard D. Frequencies of milk protein variants and haplotypes estimated from genotypes of more than 1 million bulls and cows of 12 French cattle breeds. *J Dairy Sci*. 2020;103:9124–41. <https://doi.org/10.3168/jds.2020-18492>.
55. Laodim T, Koonawootrittriron S, Elzo MA, Suwanasopee T, Jattawa D, Sarakul M. Genetic factors influencing milk and fat yields in tropically adapted dairy cattle: insights from quantitative trait loci analysis and gene associations. *Anim Biosci*. 2024;37:576–90. <https://doi.org/10.5713/ab.23.0246>.
56. Mastrangelo S, Bahbahani H, Moioli B, Ahbara A, Al Abri M, Almathen F, et al. Novel and known signals of selection for fat deposition in domestic sheep breeds from Africa and Eurasia. *PLoS ONE*. 2019;14:e0209632. <https://doi.org/10.1371/journal.pone.0209632>.
57. Akinsola OM, Musa AA, Muansangi L, Singh SP, Mukherjee S, Mukherjee A. Genomic insights into adaptation and inbreeding among Sub-Saharan African cattle from pastoral and agropastoral systems. *Front Genet*. 2024;15:1430291. <https://doi.org/10.3389/fgene.2024.1430291>.
58. Boussaha M, Esquerré D, Barbieri J, Djari A, Pinton A, Letaief R, et al. Genome-wide study of structural variants in bovine Holstein, Montbéliarde and Normande dairy breeds. *PLoS ONE*. 2015;10:e0135931. <https://doi.org/10.1371/journal.pone.0135931>.