Journal of Advanced Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Advanced Research

journal homepage: www.elsevier.com/locate/jare

Original Manuscript

Long read and preliminary pangenome analyses reveal breed-specific structural variations and novel sequences in Holstein and Jersey cattle

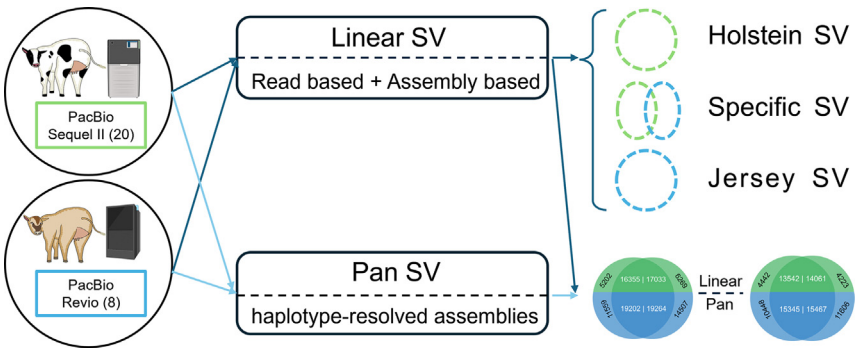
Yahui Gao [a,b,c,*], Liu Yang [b,c], Kristen Kuhn [d], Wenli Li [e], Geoffrey Zanton [e], Mary Bowman [b], Pengju Zhao [f], Yang Zhou [g], Lingzhao Fang [h], John B. Cole [i,j,k], Benjamin D. Rosen [b], Li Ma [c], Congjun Li [b], Ransom L. Baldwin VI [b], Curtis P. Van Tassell [b], Zhe Zhang [a], Timothy P.L. Smith [d,*], George E. Liu [b,*]

[a] State Key Laboratory of Swine and Poultry Breeding Industry, National Engineering Research Center for Breeding Swine Industry, Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, China
[b] Animal Genomics and Improvement Laboratory, Beltsville Agricultural Research Center, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD 20705, USA
[c] Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA
[d] USDA, ARS, U.S. Meat Animal Research Center (USMARC), Clay Center, NE, USA
[e] US Dairy Forage Research Center, USDA-ARS, Madison, WI, USA
[f] Hainan Institute, Zhejiang University, Yongyou Industry Park, Yazhou Bay Sci-Tech City, Sanya 572000, China
[g] Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, Huazhong Agricultural University, Wuhan 430070, China
[h] Quantitative Genetics and Genomics (QGG), Aarhus University, Aarhus, Denmark
[i] Council on Dairy Cattle Breeding, 4201 Northview Dr, Bowie, MD 20716, USA
[j] Department of Animal Sciences, Donald Henry Barron Reproductive and Perinatal Biology Research Program, and the Genetics Institute, University of Florida, Gainesville, FL 32611-0910, USA
[k] Department of Animal Science, North Carolina State University, Raleigh, NC 27695-7621, USA

HIGHLIGHTS

- Generated 56 haploid genomes from 28 dairy cattle using PacBio HiFi sequencing at 20× depth.
- First population-scale publication on Holstein and Jersey haploid assemblies.
- Built pangenome graphs for accurate SV genotyping in dairy cattle.

GRAPHICAL ABSTRACT



* Corresponding authors at: State Key Laboratory of Swine and Poultry Breeding Industry, National Engineering Research Center for Breeding Swine Industry, Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou, 510642, China (Y. Gao). Animal Genomics and Improvement Laboratory, Beltsville Agricultural Research Center, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD 20705, USA (Y. Gao). Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA (Y. Gao). USDA, ARS, U.S. Meat Animal Research Center (USMARC), Clay Center, NE, USA (T.P.L. Smith). Animal Genomics and Improvement Laboratory, Beltsville Agricultural Research Center, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD 20705, USA (G.E. Liu).

E-mail addresses: yahui.gao@scau.edu.cn (Y. Gao), yangqism@gmail.com (L. Yang), kristen.kuhn@usda.gov (K. Kuhn), wenli.li@usda.gov (W. Li), geoffrey.zanton@usda.gov (G. Zanton), mary.bowman@usda.gov (M. Bowman), zhaopengju2014@gmail.com (P. Zhao), yangzhou@mail.hzau.edu.cn (Y. Zhou), lingzhao.fang@qgg.au.dk (L. Fang), john.cole@uscdcb.com (J.B. Cole), ben.rosen@usda.gov (B.D. Rosen), lima@umd.edu (L. Ma), congjun.li@usda.gov (C. Li), ransom.baldwin@usda.gov (R.L. Baldwin VI), curt.vantassell@usda.gov (C.P. Van Tassell), zhezhang@scau.edu.cn (Z. Zhang), tim.smith2@usda.gov (T.P.L. Smith), George.Liu@usda.gov (G.E. Liu).

https://doi.org/10.1016/j.jare.2025.04.014
2090-1232/Published by Elsevier B.V. on behalf of Cairo University.

Please cite this article as: Y. Gao, L. Yang, K. Kuhn et al., Long read and preliminary pangenome analyses reveal breed-specific structural variations and novel sequences in Holstein and Jersey cattle, Journal of Advanced Research, https://doi.org/10.1016/j.jare.2025.04.014

ARTICLE INFO

Article history:
Received 13 October 2024
Revised 6 April 2025
Accepted 10 April 2025
Available online xxxx

Keywords:
Dairy cattle
Long read
Structural variation
Linear genome
Pangenome

ABSTRACT

Introduction: Most SV studies in livestock rely on short-read sequencing, posing challenges in accurately characterizing large genomic variants due to their limited read length.
Objectives: Our goal is to reveal structural variation and novel sequences specific to Holstein and Jersey cattle breeds using long-read and pan-genome analyses.
Methods: We sequenced 20 Holsteins and 8 Jersey cattle using PacBio HiFi to 20×, and integrated five read-based and one assembly-based SV caller to determine SVs.
Results: We assembled the 28 genomes averaging 3.25 Gb with a contig N50 of 69.36 Mb and using the ARS-UCD1.2 reference, we acquired Holstein/Jersey SV catalogs with 74,068/54,689 events spanning 202/135 Mb (7.43 %/4.97 % of the genome). SVs were enriched in less conserved, non-coding, and non-regulatory regions. Comparing Holsteins with differing feed efficiency (FE), SVs unique to high FE were linked to energy metabolism and olfactory receptors, while those specific to low FE were associated with material transport. We constructed Holstein/Jersey pangenome graphs with 148,598/105,875 nodes and 208,891/147,990 edges, representing 47,028/37,137 biallelic and multi-allelic events, and 63.75/42.34 Mb of novel sequence. We observed SV count saturation with 20 Holsteins, while adding Jerseys significantly increased the SV count, highlighting breed-specific SV events.
Conclusion: Our long-read data and SV catalogs are valuable resources, revealing that the cattle genome is more complex than previously thought.

Introduction

For thousands of years, cattle have supported human survival, first as hunted animals and, for the past 10,000 years, as livestock raised for meat, milk, and draft work [1]. Cattle, including *Bos taurus* and *Bos indicus*, exhibit significant genetic diversity impacting milk production, meat quality, disease resistance, and environmental adaptation. Understanding this variation will improve knowledge of cattle biology, provide new targets for breeding programs, and support efforts to improve livestock sustainability. Structural variations (SVs) refer to genomic differences that are not SNVs, including insertions, deletions, inversions, and translocations ranging from 50 bp to megabases between individuals [2]. "CNV" often refers to unbalanced SVs (>1 kb insertions and deletions). Significant progress has been made in understanding SVs in mammals [3–9]. While SNVs are more frequent, SVs involve more genomic sequences and have greater potential effects, including altering gene structure and dosage, gene regulation, and exposing recessive alleles [10]. In particular, segmental duplications (SDs) are significant catalysts for SV formation [9,11,12]. SVs are important for phenotypic variability and disease susceptibility. A human study showed that SNVs and SVs accounted for ~82 % and ~18 % of the total genetic variation in gene expression, respectively [13]. In humans, SVs are 3 × more likely to associate with GWAS signals than SNVs, and larger SVs (>20 kb) are up to 50 × more likely to affect gene expression compared to SNVs [7,14]. Although some SVs show linkage disequilibrium (LD) with flanking SNVs [15], many SVs are not easily tagged by SNVs and often fall in difficult-to-map repetitive regions (such as SDs) not well covered by SNP arrays [16,17]. Combining SV and SNV data in GWAS has linked SVs with human diseases and animal production traits [18–26].

Despite their significance, most SV studies in livestock rely on short-read sequencing, posing challenges in accurately characterizing large genomic variants due to their limited read length. Many studies reported CNV/SV discovery in cattle [25–35]. Notably, specific SVs in cattle have been linked to various traits. For example, color sidedness in cattle is determined by two alleles resulting from translocations involving the *KIT* gene on chromosomes 6 and 29 [36]. A large-scale study of CNVs in 336 cattle from 39 breeds identified 362 significant CNV regions (CNVRs) related to olfactory receptors, pathogen resistance, and productivity, highlighting their potential role in cattle adaptation [37]. Another study in over 500 bulls from 17 taurine breeds found 26,223 CNVRs covering 4.05 % of the genome, with results in an interactive database for detailed exploration [38]. However, few of these SVs are validated at the sequence level, and overlaps among these cattle SV datasets are low (< 40 %), indicating that better SV annotation is needed. Therefore, the precise characterization and comprehensive exploration of SVs in the cattle genome remains incomplete.

Detecting SVs from microarray or short-read sequencing suffers from low sensitivity (30–70 %) and up to 85 % false discovery [7,14,39–42]. CGH and SNP arrays have limitations that include a lack of information on variation structure, limited resolution, and an inability to detect balanced rearrangements. Short-read sequencing methods (RP, RD, SR, LA, hybrid) have constraints due to indirect inferences and the short length of reads [43], making it difficult to detect smaller or balanced events [6,44]. Aligning short reads to a reference assembly works well in confident (non-repetitive) regions, but is problematic in difficult repetitive regions like SDs enriched for SVs [4,9,11,12,45]. Therefore, long-read sequencing and high-throughput genotyping platforms are needed for accurate SV detection and analysis.

The current cattle reference genome was derived from a female taurine Hereford cow L1 Dominette 01,449 [46]. However, it is widely acknowledged that a single reference genome cannot capture the full genomic diversity of an entire species. The separate domestications and selections of taurine (*Bos taurus taurus*) and indicine/Zebu cattle (*Bos taurus indicus*) highlight the limitations of ARS-UCD1.2, which does not adequately represent the extensive genetic variation within the *Bos taurus* species [47]. Aligning reads with non-reference alleles to a single reference genome often results in unmapped or misplaced reads [48]. This mapping bias leads to the underrepresentation of important genetic variants, especially rare alleles or large SVs that may play crucial roles in complex traits.

To overcome this bias, the concept of the pangenome provides a viable solution by including all genomic sequences within a species or specific phylogenetic group [49,50]. Early pangenome projects mainly focused on identifying novel sequences and integrating them into the reference genome [34,51,52], preserving the linear reference genome structure for compatibility with subsequent analyses. Recently, the development of graph-based genome structures has enabled the representation of all possible sequences

within a unified coordinate system, effectively mitigating mapping bias. Programs like vg [53], Minigraph [54], Minigraph-Cactus (MC) [55], and the PanGenome Graph Builder (PGGB) [56] were developed to construct pangenome graph assemblies [57]. Graph pangenomes, compared to their linear counterparts, improve read mapping rates and increase the detection sensitivity of variants, particularly SVs [52,58–61]. Like humans, cattle and other livestock are seeing increasing numbers of high-quality assembled pangenomes. Various graph pangenome projects have been initiated for farm animals, including cattle, sheep, and pigs [47,62–67]. For example, Pausch's lab adapted vg graphs and developed breed-specific and pan-genome reference graphs in cattle, showing their superior accuracy over traditional linear references and uncovering 70 Mb of novel sequences [58,68,69]. The Prendergast lab integrated 116 Mb of novel African cattle sequences into the reference assembly, improving read mapping rates and SV calling accuracy [52]. Leonard et al. showed structural variant-based pangenomes from haplotype-resolved assemblies were highly consistent across platforms and algorithms, creating multi-species super-pangenomes with good consensus [64,65]. Recently, they constructed a pangenome from 16 HiFi cattle assemblies and used it to identify SNVs and SVs [70]. After SV genotyping using short reads by PanGenie [71], they conducted a preliminary molQTL mapping with 117 testis transcriptome data, identifying 92 potential causal SV candidates. These studies collectively demonstrate the power of using variation-aware graph-based approaches in cattle genomics, providing a more accurate and comprehensive mapping of genetic variations compared to traditional linear references. A few T2T or near-complete assemblies were reported for Holstein cattle and goats, filling many gaps in the reference genome, particularly in immunogenomic regions [72,73]. Pangenomes have also been created for sheep, *Bos indicus* cattle, and yaks [47,62,63].

In our study, we PacBio HiFi-sequenced 20 Holstein and 8 Jersey cattle to high coverage, prepared high-quality SV catalogs and constructed their pangenome graphs. As expected, the SV catalogs generated from these samples significantly increased the SV count, highlighting breed-specific SV events. Compared to SNVs, SVs exhibited a unique contribution to genome diversity, underscoring their importance in understanding the genetic architecture of cattle.

Materials and methods

Compliance with ethics requirements

Ethical permission to collect blood samples from cattle was approved by the US Department of Agriculture, Agricultural Research Service, Beltsville Agricultural Research Center's Institutional Animal Care and Use Committee (Protocol 18-005).

Sample collection

Under the approval of the US Department of Agriculture, Agricultural Research Service, Beltsville Agricultural Research Center's Institutional Animal Care and Use Committee (Protocol 18-005), and the US Dairy Forage Research Center (A0050543-R04), we collected fresh blood samples from 20 Holstein and 8 Jersey dairy cattle. We isolated the high-molecular-weight (HMW) DNA from whole blood using the Nanobind/Circulomics UHMW Blood extraction protocol and sheared it to 20 kb mode size using a Diagenode Megaruptor 3. We checked DNA quantity on a Qubit Fluorometer with a dsDNA HS Assay kit (Thermo Fisher).

Library construction and PacBio sequencing

We examined the sizes of DNA fragments on a Fragment Analyzer (Agilent Technologies). To obtain long reads, we removed fragments below 15 kb in length on the BluePippin using the BLF-7510 cassette under the 0.75 % DF Marker S1 high-pass 15–20 kb protocol (Sage Science). We prepared SMRTbell libraries for sequencing according to the protocol 'Procedure-checklist-Preparing-whole-genome-and-metagenome-libraries-using-SMRTbell-prep-kit-3.0′. We bound the selected library fractions to polymerase with either the Sequel II Binding Kit 3.2 or the Revio Polymerase Kit and then sequenced them on Sequel II (for Holstein) or Revio (for Jersey) instruments (PacBio) with 30-h movie times for each sample. We sequenced samples to a minimum HiFi data amount of 60 Gb (20 × estimated genome coverage).

Data preprocessing

We generated the statistics including N50, Q20, and Q30 using SeqKit (v.0.10.1) [74] with the default parameters and randomly down-sampled full coverage datasets to lower coverage levels with the command "seqkit sample-p". We obtained the read length using bioawk (https://github.com/lh3/bioawk) with the default parameters. We removed the raw reads shorter than 1 kb using fastp (v.0.23.3) [75] with the parameter "-l 1000". We used the ggplot2 R package (v3.4.2) to visualize all the results.

SVs detection

Read-based approaches

We performed a read-based approach to call SVs with multiple tools and strict filtering steps. We aligned each dataset against the cattle reference genome ARS-UCD 1.2 [46] using pbmm2 v1.9.0 (https://github.com/PacificBiosciences/pbmm2) with parameters: –-preset HIFI –-sort –-rg '@RG\tID:myid\tSM:mysample' –-log-level INFO. We then removed the sequences with MAPQ less than 30 by SAMtools (v.1.12) [76] with the parameter "-q 30" and created new index files with the index function. Subsequently, we employed five SV callers to detect SVs, including pbsv (v.2.8.0) (https://github.com/PacificBiosciences/pbsv), SVIM (v.1.4.2) [77], Sniffles (v.2.0.7) [78], SVision (v.1.3.8) [79], and cuteSV (v.2.0.1) [80]. We applied these callers directly to aligned reads with default parameters except for cuteSV (v.2.0.1) [80], whose parameters were: –-max_cluster_bias_INS 1000 –-diff_ratio_merging_INS 0.9 –-max_cluster_bias_DEL 1000 –-diff_ratio_merging_DEL 0.5.

***Assembly-based** approach*

Firstly, we used Hifiasm (v.0.18.5) [81] to assemble the PacBio HiFi reads with default parameters and convert the GFA assembly graph to a FASTA file of all sequences using gfatools (v.0.4-r214) (https://github.com/lh3/gfatools). Next, we applied minimap2 (v.2.24-r1122) [82] to align the assembled files to the cattle reference genome [46] with default parameters. Then, we run SVIM-asm (v.1.0.2) [83] with parameters "svim-asm haploid –-min_sv_size 50 –-max_sv_size 200,000 –-tandem_duplications_as_insertions –-interspersed_duplications_as_insertions" for SV detection. We assessed the completeness of the assemblies with BUSCO (v.5.4.3) [84] and the single plus duplicated complete BUSCO gene counts were reported. We calculated the assembly base QVs with Inspector (v.1.0.2) [85]. Key metrics (N50, N90,

longest contigs, number of contigs, GC content, BUSCO scores) have been represented as snailplots.

High-confidence SVs

We merged the five SV callsets of each individual derived from the above read-based approaches for each sample. We used Jasmine (v.1.1.5) [86]to merge SVs based on the coordinates and length by running the command "jasmine file_list = vcf_list.txt out_file = combined.vcf max_dist = 1000 spec_len = 50 spec_reads = 1". As suggested by benchmark analysis of LRS callers [87,88], we chose results that have a priority of pbsv > SVIM > Sniffles > SVision > cuteSV. Using the VCF file generated by Jasmine, we conducted three steps to filter out lower-quality SVs. We removed SVs in the sex and unplaced chromosomes, as well as those shorter than 50 bp, and retained SVs detected by at least two callers for each individual. For the SVs detected by the assembly-based method, we also removed those in the sex and unplaced chromosomes, as well as those shorter than 50 bp for each individual. Consequently, we directly merged this clean SV dataset with the dataset from the read-based approaches above and obtained the final high-confidence SVs.

Functional relevance of SVs

SV annotation

We first performed SV annotation for the SV dataset using ANNOVAR (v.2020–6-8) [89]. We then classified SVs into 10 categories based on their coordinates with genomic features if they overlapped by at least 1 bp, including intergenic, intronic, exonic, ncRNA_intronic, downstream, upstream, ncRNA_exonic, UTR3, UTR5, and splicing. Similarly, to explore the repeats of the SV dataset, we annotated the repeat elements and SVs where their reciprocal overlap percentage was at least 0.8. The repeat information was downloaded from the UCSC genome browser's cattle RepeatMasker database (https://hgdownload.soe.ucsc.edu/goldenPath/bosTau9/database/rmsk.txt.gz). We utilized the intersect function of bedtools (v.2.30.0) [90] to find the overlapping regions.

Chromatin state and TF enrichment

To validate the activity of SVs, we conducted the chromatin state and TF enrichment analysis of SVs. We first downloaded 14 chromatin states predicted in eight major cattle tissues using ChIP-seq including four histone modification marks (i.e., H3K4me3, H3K4me1, H3K27ac, and H3K27me3) and chromatin accessibility (ATAC-seq) [91]. The 14 chromatin states were Active_TSS, CTCF/Active_TSS, Flanking_TSS, Promoter, Active_Promoter, CTCF/Promoter, Poised_Promoter, Active_Enhancer, CTCF/Enhancer, Primed_Enhancer, Active_Element, Polycomb_Repressed, Insulator, and Low_Signal. We then obtained the TF dataset from the AnimalTFDB4 [92] for further enrichment analysis. We carried out enrichment analysis for each SV type and each chromatin state using the fisher function of bedtools (v.2.30.0) [90].

Conservation score annotation

We annotated the SVs using a range of conservation metrics spanning GERP (genomic evolutionary rate profiling) [93], PhastCons [94], and phyloP [95]. GERP is a well-validated, evolutionary-constrained method that can identify and quantify deleterious mutations genome-wide, including at synonymous and noncoding sites [93]. Low GERP scores predict neutral mutations not under selective constraints. High GERP scores, on the other hand, predict deleterious mutations subject to strong purifying selection. Moderate GERP scores, however, are ambiguous and cannot reliably distinguish between neutral and deleterious mutations [93]. We began by downloading the GERP scores for cattle, calculated from a whole genome alignment of 91 mammalian genomes, from the Ensembl server (https://ftp.ensembl.org/pub/release-110/bed/ensembl-compara/91_mammals.gerp_constrained_element/gerp_constrained_elements.bos_taurus.bb). Then, we used bigBedToBed (v.1) (https://github.com/ENCODE-DCC/kentUtils/) to convert.bb file to.bed file. In light of the wide range in GERP scores, we segmented them into 11 levels (<10, [10, 20), [20, 30), [30, 40), [40, 50), [50, 60), [60, 70), [70, 80), [90, 100), >1000). We applied the Fisher function in bedtools (v.2.30.0) [90] to perform enrichment analysis for each SV type and GERP score.

In addition, we downloaded conservation tracks of the phastcons-100way and phyloP-447way from the UCSC genome browser (https://hgdownload.cse.ucsc.edu/goldenpath/hg38/phastCons100way/hg38.100way.phastCons; https://hgdownload.cse.ucsc.edu/goldenpath/hg38/phyloP447way/hg38.447way.phyloP) and converted wigFix format to bed format with wig2bed (v.6.3.3.12) (https://bedops.readthedocs.io/en/latest/). We proceeded to map the coordinates of the sites to their corresponding points on the cattle genome by liftOver [96] with the parameter "-minMatch = 0.8" and calculated the mean phastCons and phyloP score for each SV.

SV hotspot identification

Referencing an earlier study [97], we selected the midpoint of each SV and performed hotspot analysis using the 'hotspotter' function of the primatR package (v0.1.0) with the parameters "bw = 200000, pval = 1e-08, num.trial = 2000". For hotspot detection and comparison, we incorporated long-read SV data sets from previous publications [47,52,69]. In accordance with the prior study [47,97], we sorted our inferred SV hotspots into three groups: 'terminal', situated within the last 5 Mb of the chromosome; 'known', coinciding with hotspots identified in earlier research if they overlap by at least one bp; and 'novel', specific to this study.

Differential SVs identification

To detect differential SVs, we first merge the individual SVs within each group using Jasmine (v.1.1.5) [86] with its default settings. Next, we employed the intersect function of bedtools (v.2.30.0) [90] to distinguish between differential and shared SVs between comparison groups. We considered SVs to be differential if they had no overlap (the differential SVs were also treated as the group-specific SVs because they were present in all samples in one group and not present in any samples of the other group, for both Holstein high and low RFI groups and Holstein-Jersey comparisons) and shared when the reciprocal overlap exceeded or equaled 80 % (The shared SV was defined as one where at least one sample in each group has an overlapping SV). To test the significance of the SV count differences between the two groups or breeds, we employed two non-parametric statistical methods: permutation testing and the Chi-Square/Fisher's Exact test. These approaches are well-suited for small sample sizes and make no assumptions about data normality. In particular, for the HOL and JER comparison, we used a logistic regression model to account for sequencing platform effects prior to conducting the non-parametric analyses.

We conducted the chromatin state enrichment analysis of differential and shared SVs using the fisher function of bedtools (v.2.30.0) [90]. We utilized the Genomic Regions Enrichment of Annotations Tool (GREAT) (v4.0.4) [98] for gene ontology (GO) analysis to explore the function of SVs. After multiple comparison

corrections, we included significant GO terms, retaining results with BinomFdrQ ≤ 0.05 and RegionFoldEnrich ≥ 2.

Construction of cattle pangenome

We used the Minigraph pipeline to construct the pangenome graph for cattle, with the ARS-UCD1.2 serving as the backbone of the graph. Due to the presence of numerous small fragments in the ARS-UCD1.2 assembly, we only considered the sequences of autosomes.

Biallelic and multiallelic SV detection using a graph-based method

By slightly modifying a previously reported workflow [69], we identified biallelic and multiallelic SVs from the graph genome. First, with the current Hereford-based linear reference genome [46] as the backbone, we augmented the 40 or 16 partially phased assemblies into the graph individually to build the multiassembly graph for Holstein and Jersey with minigraph (v.0.20-r559) [54]. Then by using gfatools (v.0.4-r214) (https://github.com/lh3/gfatools), which is built on a bubble-popping algorithm, we extracted bubbles from the multiassembly graph. In the reference graph model of minigraph (v.0.20-r559) [54], each bubble indicates an SV, comprising the start and end nodes of reference sequences as well as the paths that connect them. We annotated these SVs using the same method outlined above. With the help of bedtools (v.2.30.0) [90], we compared the graph-based SVs to the linear-based SVs identified above, differentiating between differential and shared SVs. We considered SVs to be shared if they exhibited an 80 % or greater overlap.

Results

PacBio sequencing of Holstein and Jersey cattle samples

We chose the BARC (Beltsville Agricultural Research Center) and DFRC (Dairy Forage Research Center) herds to represent the diversity of the Holstein and Jersey populations. We generated long-read whole-genome sequence data from a group of 20 Holstein and 8 Jersey cattle, utilizing the PacBio Sequel II for the former and the Revio platform for the latter. Each cow blood sample was subjected to HiFi sequencing with 1–2 cells, aiming at an average sequencing depth of at least 20× (ranging from 27 to 54 ×). Each sample yielded long-read data of more than 60 billion bases, achieving a depth of over 20× (Fig. 1A and Table S1). The mean N50 read length was 19.75 Kb, with a range from 15.25 to 27.79 Kb (Fig. 1B and Table S1). When the read length reached around 20 kb, the cumulative throughput accounted for nearly half of the total. (Fig. 1C). This consistent trend across various sequencing runs indicated that a significant amount of the data yield was concentrated within the first 20 Kb. All samples exhibited a central tendency of GC content at 50 % across reads of disparate lengths (Fig. S1A, B). By filtering the raw data to exclude below 1 kb, we effectively selected high-quality sequences for more in-depth analysis (Fig. 1D).

Using pbmm2 (v.1.9.0) (https://github.com/PacificBiosciences/pbmm2), we successfully aligned 99 % of the reads to the ARS-UCD1.2 reference genome [46], exhibiting an average mapping intensity of around 98 % (Fig. 1E). For *de novo* assembly, we utilized the Hifiasm (v.0.18.5) assembler [81], resulting in the generation of one primary assembly and two partially phased contig assemblies (Hap1 and Hap2 in Table S2). The final total primary genome lengths of 3.26 and 3.16 Gb with average contig N50s were 72.98 Mb and 56.80 Mb for the Holstein and Jersey, respectively, (Fig. 1F, Fig. S1C, D). According to BUSCO (v.5.4.3) [84], the completeness of mammalian universal single-copy orthologs in the Holstein and Jersey genome assemblies was found to be 94.99 % and 95.88 %, respectively (Table S2), which was consistent with the 95.80 % seen in previously analyzed cattle genomes. Similarly, the duplication rates of 2.14 % and 2.29 % in the assemblies are consistent with the typical 2.00 % range observed in Hereford genomes. BUSCO scores for the partially phased assemblies were lower but still consistently > 90 % complete (Table S2). The quality value (QV) of the assemblies assessed by the Inspector (v.1.0.2) [85] indicated high QV values (average 45.51) for all the primary than phased assemblies (Table S2). According to Rhie et al. [99], these cattle assemblies were therefore considered as high-quality under the VGP-2020 standards.

Building and characterizing a catalog of SVs

Previous studies have shown that HiFi sequencing significantly increases SV discovery [47,64,81,100]. In this study, we detected four classes of canonical SVs (DEL, INS, DUP, and INV). To obtain reliable SVs, we used two strategies including the "read-based" and "assembly-based" strategies to detect SVs. For the "read-based" strategy, we selected five callers: pbsv (v.2.8.0) (https://github.com/PacificBiosciences/pbsv), SVIM (v.1.4.2) [77], Sniffles (v.2.0.7) [78], SVision (v.1.3.8) [79], and cuteSV (v.2.0.1) [80], all specifically designed for SV detection by long-read mapping-based approaches for each genome. For "assembly-based", we used SVIM-asm (v.1.0.2) [83] to call SVs (Fig. 2A). The counts of SVs obtained from different combinations of multiple callers ranged from 3,158 to 75,114 (Fig. 2B, C). For each sample, we applied three filtering steps to remove unreliable SVs (on unplaced contigs, <50 bp, and detected by only one read-based caller) (Fig. 2D). We then merged SVs through the union of SV sets from two approaches. Finally, we identified an average of 28,463 high-confidence SVs per sample, ranging from 26,365 to 29,739 (Fig. 2E, F and Fig. S2A, B). DELs and INSs were predominant, and each sample contained an average of 13,816 DELs (48.54 %), 14,222 INSs (49.97 %), 163 DUPs (0.57 %), and 262 INVs (0.92 %) (Fig. 2G). Especially, SVIM reported the most SVs (Fig. S2C). We observed that the median lengths of DELs and INSs were 140/140 bp and 128/132 bp, respectively, significantly shorter than those of INVs (2,174/1,820 bp) and DUPs (5,991/5,710 bp) (Fig. 2H, I). Upon merging all SVs discovered in each individual, we obtained 74,068 SVs for Holstein and 54,689 SVs for Jersey (Fig. S2D, Table S3). Most INVs and DUPs can be found in BTA5 and BTA10, 15 and 26, respectively (Fig. S2D). As the number of individuals increased, the increase in the number of detected SVs reached plateaus (Fig. 2J). However, an obvious jump occurred when transitioning from Holstein to Jersey breeds, indicating the widespread presence of breed-specific SV events (Fig. 2J).

To assess the effects of coverage on the SV detection, we randomly down-sampled our full coverage data sets to lower coverage levels, i.e., 5, 10, 15, 18, 20, 25, and 28 × coverage. The read-length distributions of different coverages were similar (Fig. S3A), but with the increase of depth, the N50 length also increased (Fig. S3B). The number of SVs increased marginally when the depth was more than 10-fold (Fig. S3C). We then compared the SVs to the results based on the full coverage (Fig. S3D) and measured recall rates and false positive rates (Fig. S3E). Both the recall rates and false positive rates seem leveled at 10 × coverage with a recall rate of ~90.6 % and a false positive rate of ~9.4 %.

Functional relevance of SVs

We first grouped the SVs into singleton and non-singleton according to whether they appear in one individual or more than

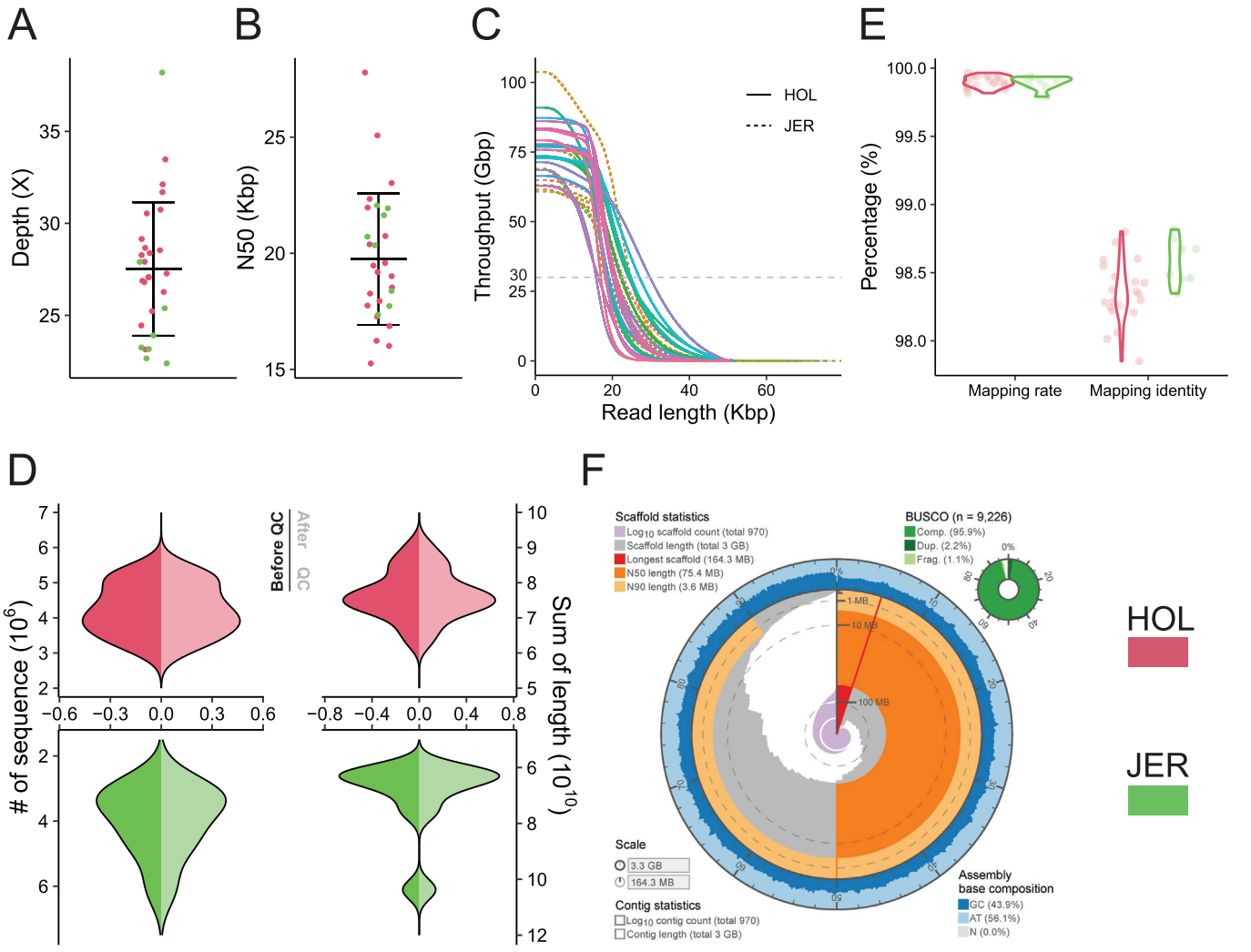


Fig. 1. Summary of the long-read sequencing data. A. The sequencing depth of long-read data obtained from 20 Holstein (pink) and eight Jersey (green) cattle via the PacBio Sequel II and Revio platforms, respectively. **B.** N50 statistics for long-read sequencing data of 28 samples. **C.** Cumulative distribution of total bases (Y-axis) over read length (X-axis) for 28 samples (solid line represents HOL, dashed line represents JER). The quantities of bases in reads 10 Kbp + were labeled (the grey dased line). **D.** Density distribution of sequence number and length before (right in each panel) and after (left in each panel) data quality control. **E.** The mapping rate and identity of long-read sequencing data of 28 samples. **F.** Snail plots of one Holstein genome assembly. Key metrics are shown for the genome such as the longest scaffold (red vertical line), N50 (orange track), N90 (light orange track), GC content (external blue track), and BUSCO scores (outer circular pie chart in green).

one individual. Non-singletons represented 78.16 % of the total identified SVs (Fig. 3A, Fig. S4A). To explore their potential functions, we annotated SVs based on their co-localization with genomic features using their genomic coordinates. A substantial percentage (65.44 %) of the SVs were in intergenic regions (Fig. 3B). We further observed that SVs showed enrichment with repetitive elements, with LINE/L1 elements being particularly dominant (Fig. 3C, Fig. S4B, C). In addition, we investigated the enrichment of SVs in 14 chromatin states (Active_TSS, CTCF/Active_TSS, Flanking_TSS, Promoter, Active_Promoter, CTCF/Promoter, Poised_Promoter, Active_Enhancer, CTCF/Enhancer, Primed_Enhancer, Active_Element, Polycomb_Repressed, Insulator, and Low_Signal) across eight tissues (adipose, cerebellum, cortex, hypothalamus, liver, lung, muscle, and spleen) [91]. We observed that INVs and DUPs had higher enrichment in poised_promoter and polycomb_repressed states of the cortex tissue (Fig. 3D). We further downloaded the transcription factors (TFs) information from the AnimalTFDB4 [92] for the enrichment and found that INVs had an enrichment with P53 (Fig. S4D-F). We also retrieved GERP information [93] and classified the scores into 11 levels (Fig. S4G). We found that DEL and INS were mainly enriched with low GERP, while INV and DUP were mainly enriched with high GERP (Fig. 3E). We further calculated the PhastCons value and phyloP values for each SV region [94,95], discovering that these values were approximately 0 for most SV regions (Fig. S4H-K).

The number distributions of DEL and INS of different sizes were similar (Fig. 3F). Most of all SVs obtained in this study overlap with previously published results [34], and this high overlap (75.62 %/74.95 %) was consistent across all SV classes (Fig. 3G). Our analysis revealed 169/156 SV hotspots spanning about 161/138 Mb of the genome (Table S4), which were nonrandomly distributed (Fig. 3H). Of these hotspots, 79/83 were within the last 5 Mb of chromosome arms. Except for the terminal regions of chromosomes, 46/41 hotspots overlap with hotspots identified in previously published long-read-based SV datasets [47,52,69], whereas the 44/32 remaining hotspots were novel (Fig. 3H).

Differential SVs between Holsteins with high and low RFI

Feed intake is one of the major expenses associated with milk production and animals that produce the same amount of milk while eating less feed are more efficient than other animals

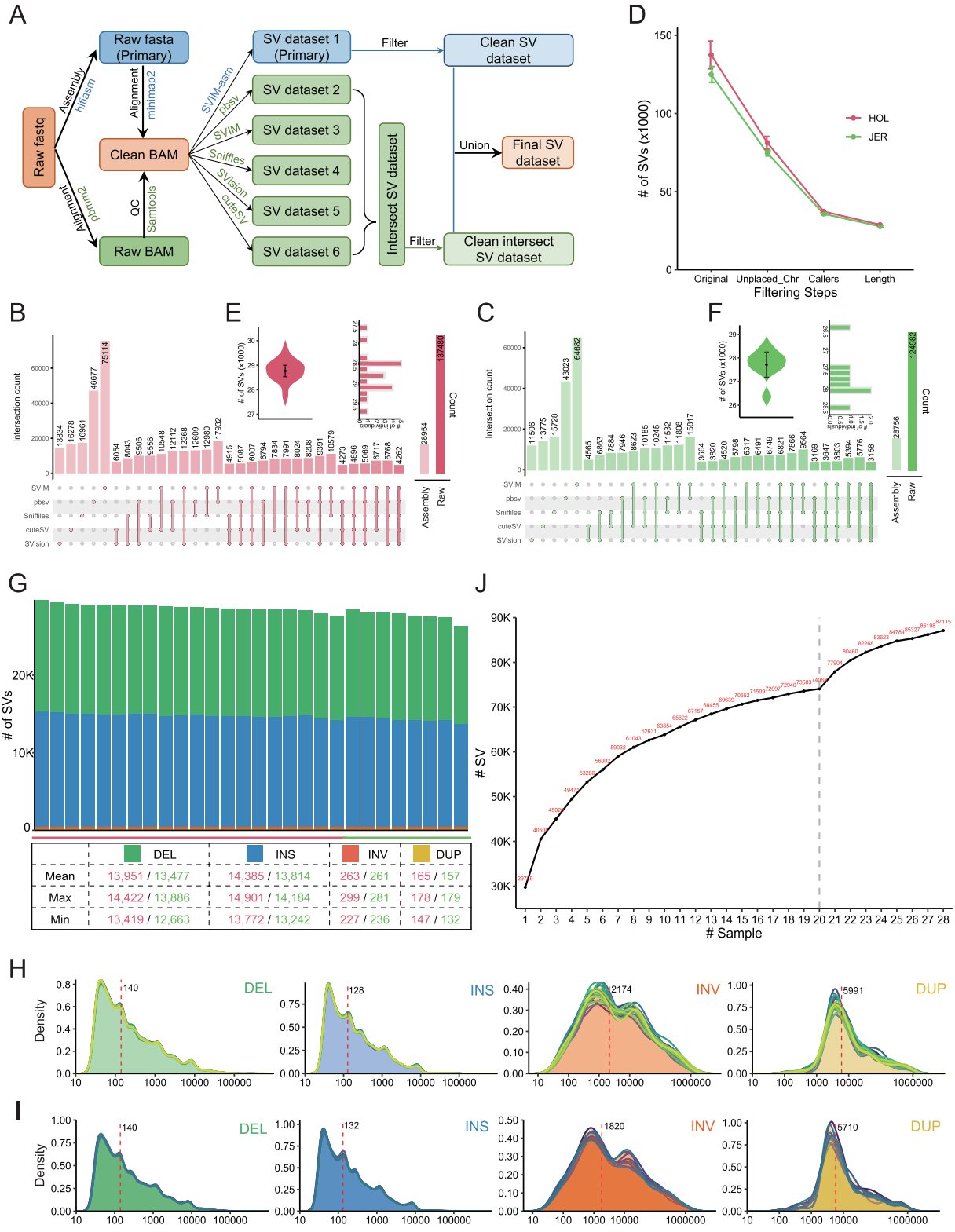
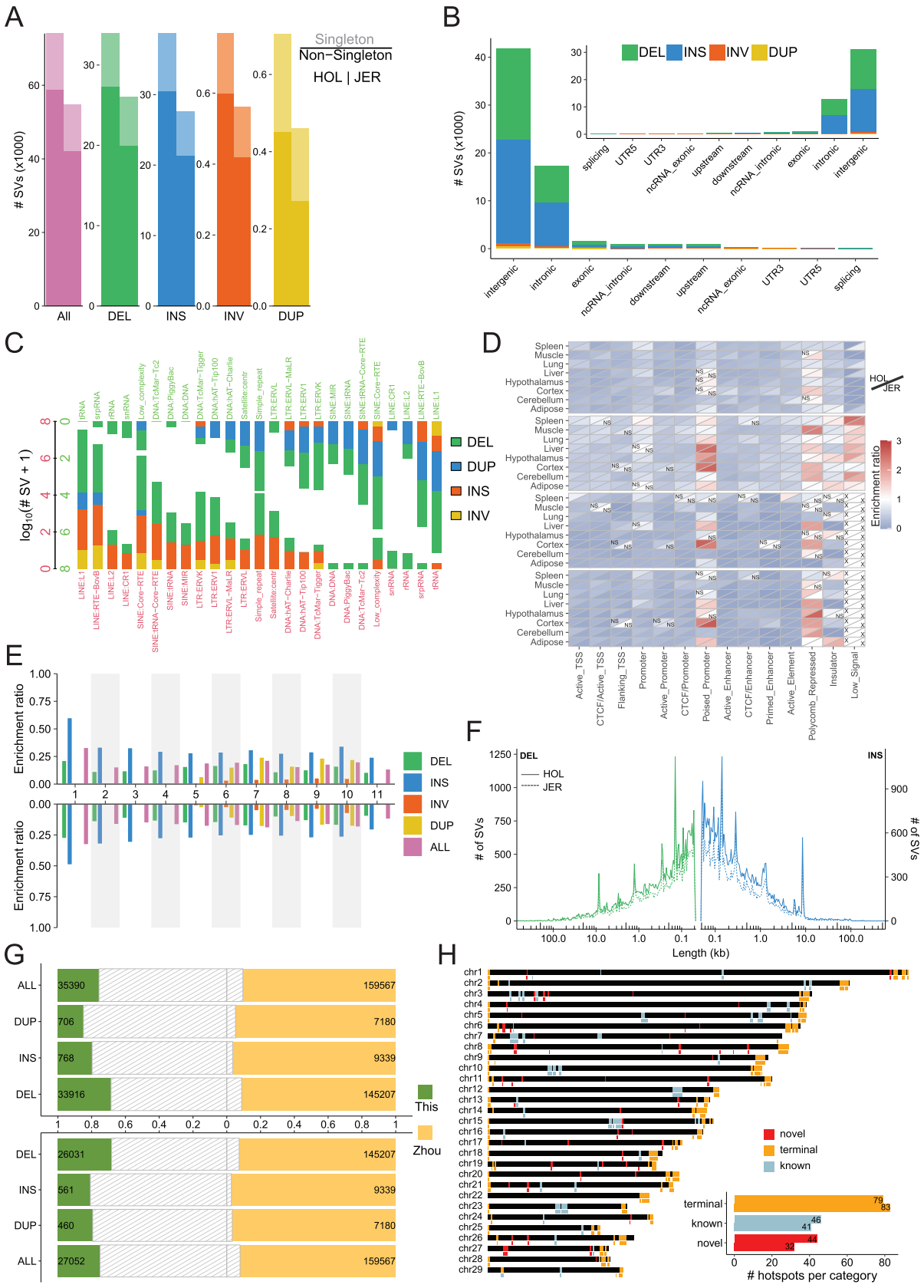


Fig. 2. Building two SV catalogs. A. The pipeline to get an SV dataset for each sample. **B and C.** UpSet plot shows the total number of SVs in the Holstein (pink) and Jersey (green) population identified by read-based methods (SVIM, pbsv, Sniffless, SVision, cuteSV) and the assembly-based method (SVIM-asm), along with their shared variants. **D.** The average number of SVs after each filtering step. "Original" represents unfiltered SVs, "Unplaced_Chr", "Callers", and "Length" represent SVs filtered according to the unknown chromosomes, supported callers (detected by at least two callers), and SV length (>=50 bp), respectively. All SVs are less than 1 Mb. **E and F.** The mean SV count per sample (left) and SV distribution of all individuals (right) for Holstein (pink) and Jersey (green) populations. **G.** SV counts of each type in each individual. **H and I.** Length distribution for each SV type in the Holstein (**H**) and Jersey population (**I**). The red line indicates the median length of each SV type. **J.** The relationship between the number of SVs and the sample size. The number above each dot represents the number of SVs detected using the corresponding number of samples. The grey dashed line serves as a boundary, with the first 20 belonging to the HOL and the remaining 8 to the JER.


A
SVs (x1000)
Singleton
Non−Singleton
HOL | JER
All
DEL
INS
INV
DUP
B
SVs (x1000)
DEL INS INV DUP
splicing
UTR5
UTR3
ncRNA_exonic
upstream
downstream
ncRNA_intronic
exonic
intronic
intergenic
C
log10(# SV + 1)
DEL
DUP
INS
INV
D
Spleen
Muscle
Lung
Liver
Hypothalamus
Cortex
Cerebellum
Adipose
HOL
JER
Enrichment ratio
Active_TSS
CTCF/Active_TSS
Flanking_TSS
Promoter
Active_Promoter
CTCF/Promoter
Poised_Promoter
Active_Enhancer
CTCF/Enhancer
Primed_Enhancer
Active_Element
Polycomb_Repressed
Insulator
Low_Signal
E
Enrichment ratio
DEL
INS
INV
DUP
ALL
F
DEL
INS
HOL
JER
# of SVs
Length (kb)
G
ALL
DUP
INS
DEL
35390
159567
706
7180
768
9339
33916
145207
This
Zhou
26031
145207
561
9339
460
7180
27052
159567
H
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chr23
chr24
chr25
chr26
chr27
chr28
chr29
novel
terminal
known
hotspots per category

[101]. Lactating dairy cows in the BARC herd have been selected for divergent residual feed intake (RFI) since 2012, and those data were available for all Holsteins used in this study. Individual RFIs were sorted from high to low and the top 5 and the bottom 5 were selected as comparison groups to explore the SV that affects RFI (Table S1). Notably, high RFI correlates with low feed efficiency (FE; animals eat more than expected), while low RFI correlates with high FE (animals eat less than expected). We defined the group-specific SVs, which were present in all samples in one group and not present in any samples of the other group for the Holstein high and low RFI group comparison. We found that most SVs were shared in common between the two groups, and the numbers of specific SVs in the high and low groups were 4,906 and 4,505 for high and low RFIs, respectively (Fig. 4A, Fig. S5A). Both permutation testing and Chi-Square/Fisher's Exact test yielded p-values below 0.05, indicating that the difference in the number of SVs between the two groups was significant. We investigated the enrichment of SVs in 14 chromatin states in eight tissues in both groups and except for the polycomb-repressed state, no SVs were enriched in other states (Fig. 4B). We then used the GREAT (v4.0.4) [98,102] to annotate specific SVs based on their coordinates. We found that the specific SVs in the low RFI group were mainly related to material transportation (Fig. 4C, Table S5), while the specific SVs in the high group were primarily related to energy metabolism and olfactory receptor functions (Fig. 4C, Table S6).

Differential SVs between Holstein and Jersey

Holstein and Jersey cattle are two of the most popular dairy breeds, each with distinct characteristics. Jersey cows are smaller in build and their milk contains more protein, more fat, and tastes far richer and creamier than Holstein milk. Based on the above two SV datasets obtained from Holstein and Jersey cattle, we explored the differential SV that may influence their characteristics. We defined the breed-specific SVs, which were present in all samples in one breed and not present in any samples of the other breed for the Holstein-Jersey breed comparison. We found that most SVs were shared in common between the two breeds, and the numbers of specific SVs in the Holstein and Jersey cattle were 19,235 and 7,603 respectively (Fig. 4D, Fig. S5B). Both permutation testing and Chi-Square/Fisher's Exact test yielded p-values below 0.05, indicating that the difference in the number of SVs between the two breeds was significant. We investigated the enrichment of SVs in 15 chromatin states in eight tissues in both breeds and except for the polycomb-repressed state, no SVs were enriched in other states (Fig. 4E). The Holstein-specific SVs primarily related to cellular signaling, transport, regulation, and metabolic processes, including lipid and fatty acid handling, sensory perception, gene expression modulation, cytoskeletal dynamics, and neuronal communication. They collectively highlighted key biological functions essential for maintaining cellular homeostasis, response to stimuli, and overall organismal health (Fig. 4F, Table S7). The Jersey-specific SVs mainly involved immune defense, metabolism of specific compounds, regulation of protein modifications, cell migration, neural and cardiac development, and enzyme functions, highlighting essential biological and physiological processes (Fig. 4F).

Cattle pangenome construction

We then constructed genome graphs of Holstein and Jersey from partially phased contig assemblies to further explore SVs. Generally, the three indicators, including N50, BUSCO scores, and QV, were all greater in the primary genome than the other two haplotypes (Fig. S6A, B, and Table S2). The Hereford-based linear reference genome formed the backbone of the bovine multiassembly graph. We augmented the two graphs with the 40 and 16 additional assemblies, added to increase the Mash distance from the reference (Fig. S6C, D). Generally, the pangenome size increased as additional genomes were added (Fig. 5A). The two resulting multiassembly graphs contained 119,031 and 78,922 nonreference nodes spanning 915 and 697 Mb with 32.91 % and 25.27 % of the resulting pangenome being flexible (i.e., not shared by all assemblies) (Fig. 5B, Table S9). The average number of non-reference segments obtained from both breeds was 16,306 and 113,781, corresponding to 23.23 Mb and 19.88 Mb (Fig. S6E, Table S10). Finally, 63.75 and 42.34 Mb were added to the reference respectively (Fig. S6E) and the pangenome's size amounted to 2.78/2.76 Gb. Additionally, the non-reference length shared by 41 or 17 assemblies was the longest (Fig. S6F, Table S11). We estimated the complexity of the pangenome by calculating the ratio of edges to nodes (edges/nodes). In general, the complexity of the graph structures remained consistent across all chromosomes. However, chromosomes 23 and 29 exhibited notably lower complexity compared to the other chromosomes (Fig. S6G, H). This disparity underscores the inadequacy of the current cattle reference genome, which is solely derived from a Hereford cow, in capturing the extensive genetic variation within the cattle species.

SV discovery from the bovine genome graph

In total, our constructed biallelic SV panel encompassed 18,739/15,722 DELs, 23,083/18,070 INSs and 5,206/3,345 multiallelic SVs for Holsteins and Jerseys, respectively (Fig. 5C, Fig. S6I, J,

Fig. 3. Functional relevance of SVs. A. Frequency distribution of all SVs and the four types of SVs. In each of the five small panels, the left side represents Holstein and the right side represents Jersey. Dark color indicates 'Non-Singleton' and light color indicates 'Singleton'. **B.** Counts of the four types of SVs across different genomic regions, where the larger panel displays Holstein and the smaller shows Jersey. **C.** Annotation of four types of SV based on different repeat types, with the upper panel for Holstein and the lower panel for Jersey. This figure is a stacked graph, and due to the significant differences among the four SV types overlapping with repeats (e.g., in the Hol group (pink), the overlap counts of INV, INS, DUP, and DEL are 9, 176, 7, and 2534 respectively), we used a log10 transformation for better visualization. **D.** Enrichment between SVs and chromatin states. Within each small cell, the antidiagonal line serves as the boundary. The left triangle represents the enrichment results for Holstein, and the right triangle represents those for Jersey. An "X" marks the absence of an enrichment signal, while "NS" signifies that enrichment was detected but failed to reach significance ($p > 0.05$). Cells that display neither "X" nor "NS" indicate that significant enrichment was achieved. **E.** Enrichment between SVs and all GERP, with the upper panel for Holstein and the lower panel for Jersey. **F.** Length distribution of deletions and insertions in Holstein and Jersey population. **G.** The SV overlap between the current study (left of 0 point) and Zhou et al.'s study (right of 0 point). Striped areas show the proportion of overlapping SVs relative to the total SV count in either study (calculated as the overlapping SV count divided by the total SV count at the far ends of each bar). Green sections represent SVs unique to the current study, and yellow sections indicate SVs unique to Zhou et al.'s study. Numbers next to each bar represent the counts of SV in each dataset. The upper panel displays Holstein and the lower panel displays Jersey. **H.** Genome-wide distribution of SV hotspots is divided into three categories: last 5 Mbp of chromosomes (yellow), overlapping (light blue), and previously unidentified (red) when compared with SV analysis of Crysnanto et al. (2021), Talenti et al. (2022) and Dai et al. (2023). The bar plot (inset) shows the total number of hotspots in each group. In each chromosome, the upper panel represents Holstein, while the lower panel represents Jersey.

Fig. 4. Differential SVs between different comparison groups. A. The number of different types of SV unique to the high-RFI and the low-RFI group within the Holstein population. **B.** Enrichment of high- and low-RFI group-specific SVs across 14 chromatin states. **C.** The top 20 significant pathways enriched by genes from high- (blue) and low-RFI (brown) group-specific SVs. The x-axis displays the value of $-\log_{10}$(FDR). **D.** The number of different types of SV unique to the Holstein and Jersey population. The definitions of unique/group-specific and shared are the same as (**A**). **E.** Enrichment between breed-specific SVs and 14 chromatin states. **F.** The top 20 significant pathways enriched by genes from Holstein- (pink) and Jersey-specific (green) SVs. The x-axis displays the value of $-\log_{10}$(FDR).

and Table S12). They collectively spanned 2.12 %/1.66 % of the cattle genome, with DEL accounting for 0.73 %/0.63 %, INSs for 0.90 %/0.71 %, and multiallelic SVs for 0.50 %/0.33 %, summing to a length of 57.57/45.16 Mb (19.73/17.10 Mb for DEL, 24.40/19.22 Mb for INS, and 13.45/8.84 Mb for multiallelic SVs). The number distributions of DEL and INS of different sizes were similar (Fig. 5D). In terms of length, multiallelic INS exhibited the longest mean length of 3,061 bp, while DEL had the shortest mean length of 1,053 bp (Fig. 5C). We observed that a substantial percentage (64.69 %/64.82 %) of the SVs were in intergenic regions (Fig. 5E). Our investigation of the relationship between SVs and repetitive elements unveiled significant intersections, particularly

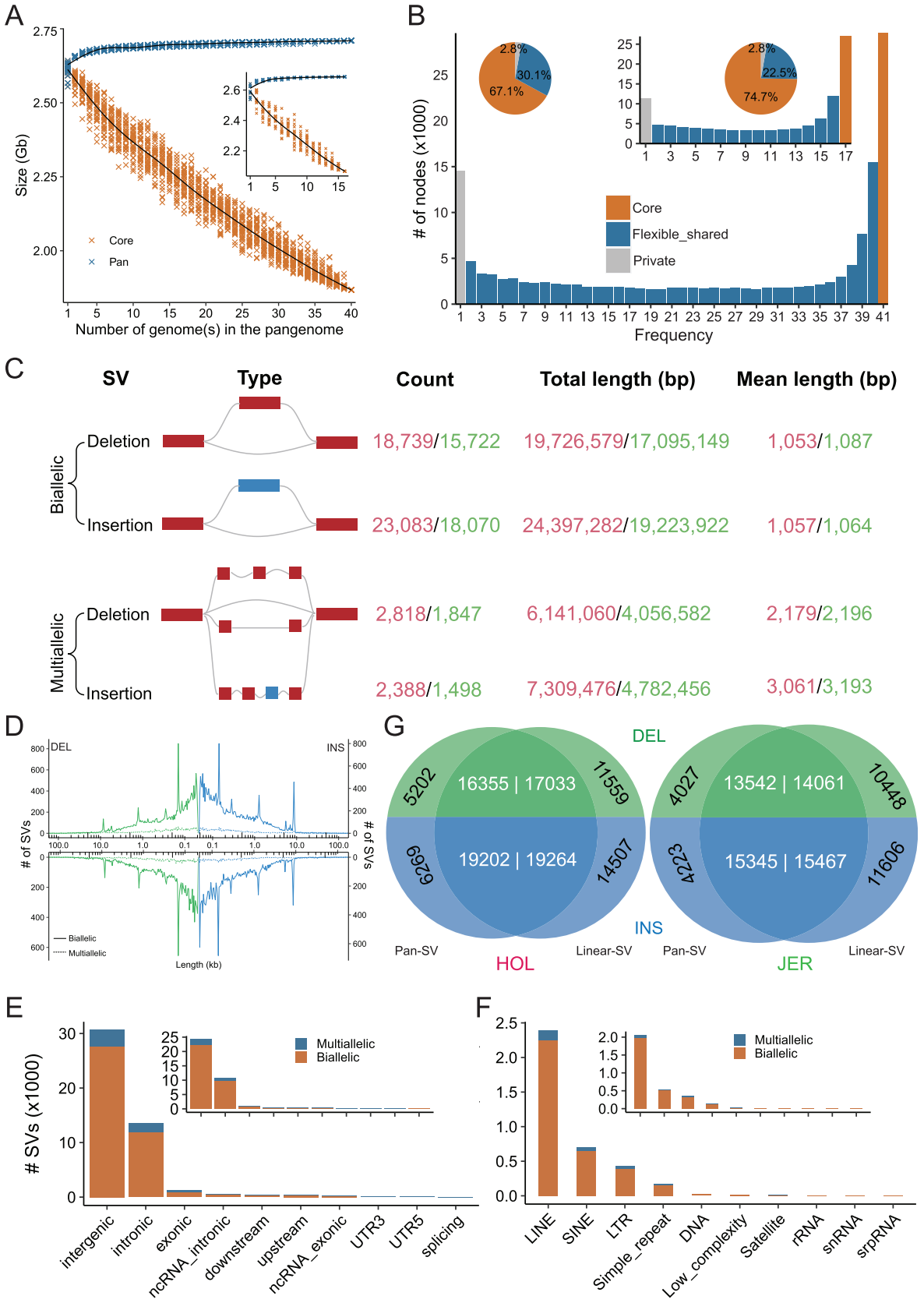


Fig. 5. SV discovery from the bovine genome graph. A. Variation of genome size in the pan-genome and core genome along with an additional cattle genome. The larger panel is for Holstein and the smaller one is for Jersey. **B.** Compositions of the pan-genome and individual genomes. The histogram shows the number of nodes in the 41 genomes with different frequencies. The pie shows the proportion of the nodes marked by each composition. The larger panel is for Holstein and the smaller one is for Jersey. **C.** Illustration and number of different SV types. The red lines indicate the reference sequence, and the blue lines represent the non-reference sequence. **D.** Length distribution of deletions and insertions. The upper panel is for Holstein and the lower panel is for Jersey. **E.** Counts of four types of SV in different genomic regions. The larger panel is for Holstein and the smaller one is for Jersey. **F.** Annotation of four types of SV with different repeat types. The larger panel is for Holstein and the smaller one is for Jersey. **G.** Count of SV in the linear SV call set and Pan SV call set and their overlap. In each Venn diagram, the left panel represents the Pan SV results, while the right panel represents the linear SV results. The upper (green) section corresponds to deletion results, and the lower (blue) section corresponds to insertion results.

with a large number of SVs overlapping with LINE (Fig. 5F). Compared with SVs obtained based on linear methods, it is observed that two-thirds of SVs overlap between the two data sets (Fig. 5G).

Discussion

In this study, we utilized 28 dairy cattle PacBio HiFi sequencing datasets with an average depth of 20 × to assemble 56 haploid genomes. To our knowledge, this is the first population-scale publication on Holstein and Jersey haploid assemblies, addressing a critical gap in dairy cattle genetic breeding research. The 56 haploid assemblies demonstrate genome continuity, completeness, and base accuracy comparable to the current cattle reference genome, ARS-UCD1.2. We constructed cattle pangenome graphs using these assemblies, enabling accurate SV genotyping from dairy cattle samples. As a high-quality SV panel was previously unavailable for dairy cattle, this study addressed this gap by constructing a robust SV catalog comprising 462,152 SVs derived from 56 haploid assemblies. This comprehensive SV panel introduces many novel SVs and achieves chromosome-wide phasing, made possible by haplotype-resolved genomes. The abundance of SVs in our panel surpasses previously published Holstein and Jersey SV datasets due to our use of haplotype-resolved genomes and the limitations of prior datasets that relied predominantly on short reads. Unlike previous studies that focused mainly on large SVs and unmapped sequences in a few local breeds, our approach comprehensively considers both large and small variations, enriching the pangenomic landscape for the two most important dairy breeds. Our findings highlight the value of long reads and pangenome graphs, enhancing alignment quality and reducing mapping errors. This refined mapping not only reduces reference bias but also improves the precision of downstream analyses. Our SV panel includes a diverse array of multi-allele SVs, which are complex and not previously explored in depth. As cattle long-read data availability continues to rise, we are committed to expanding and refining this panel to construct a comprehensive SV panel for the cattle species. However, the increased complexity of the genome graphs also poses challenges to read mapping efficiency, warranting further exploration to balance alignment efficiency and minimize reference genome bias. Additionally, while long-read sequencing is preferred for SV calling, its large-scale application on livestock remains financially challenging.

Our study reveals that analyzing a few dozen individuals per breed appears to reach a saturation plateau in SV counts, highlighting the importance of including proper sample size per breed and a broader range of breeds to fully represent cattle genetic diversity. At least a few dozen individuals per breed are essential to construct representative breed-specific pangenome graphs. Such extensive representation will allow us to effectively compare the advantages and disadvantages of breed-specific versus global pangenome graphs, similar to recent studies in humans [59–61]. Additionally, the SV catalog generated from this study will significantly aid in future imputation, enhancing the value of existing short-read data by enabling reprocessing through advanced tools like PanGenie [71]. Notably, olfactory receptors (ORs) have been independently observed multiple times using microarray, short-read, and now long-read sequencing technologies, as well as pangenomes. This repeated detection across different methodologies warrants further investigation to better understand these regions and their implications in cattle genomics, especially for feed efficiency.

The overlap of SVs between the pangenome graph and the linear SV panel in our study (67 %) is lower than the 96 % overlap reported for cattle by Leonard et al. (2023) [65]. This difference could stem from several factors. First, differences in the populations analyzed—such as genetic diversity, breed representation, and sample size—can influence the number and type of SVs detected. Second, variations in sequencing technologies and coverage levels used between the two studies may have affected the sensitivity and accuracy of SV detection. Third, the bioinformatic pipelines, including graph construction, SV calling algorithms, and overlap criteria, likely play a significant role in the observed differences. Finally, the evolutionary dynamics of SVs, including population-specific or rare variants, may also contribute to the discrepancy. These factors highlight the importance of standardizing methodologies for SV detection and comparison to ensure greater consistency across studies.

In summary, our findings highlight the necessity of including diverse breeds and multiple individuals per breed to capture the full spectrum of cattle genetic variation. The SV catalogs and comprehensive pangenome graphs will be valuable resources for improving imputation accuracy and augmenting the utility of short-read sequencing data, thus paving the way for more fruitful and accurate SV-based genetic evaluation, GWAS, and genomic analyses in cattle. Although our SV panel-enabled genotyping strategy overcomes challenges in identifying breed-specific SVs, it only included two dairy breeds. Additionally, the Jersey breed was represented by only 8 individuals, necessitating more samples. As more cattle are sequenced using long reads by the Bovine Pangenome Consortium, Bovine Long Read Consortium, and Ruminant Telomere 2 Telomere (RT2T) Project, this limitation will be overcome, paving the way for a more comprehensive understanding of breed-specific SVs in the future.

Conclusion

Our cattle SV panel represents a significant advancement in unraveling the intricate SV landscape within Holstein and Jersey genomes and their interplay with genetic diversity. This synergy facilitates comprehensive investigations into the genetic basis of various phenotypic traits, informing breeding strategies and advancing our understanding of cattle biology.

Compliance with ethics requirements

Ethical permission to collect blood samples from cattle was approved by the US Department of Agriculture, Agricultural Research Service, Beltsville Agricultural Research Center's Institutional Animal Care and Use Committee (Protocol 18-005).

Availability of data and material

The data that support the results of this research are available within the article and its Supplementary Information files. All raw data analyzed in this study are publicly available from the SRA (https://www.ncbi.nlm.nih.gov/sra/) database under accessions PRJNA1113979 and PRJNA1129520. Details of PacBio HiFi sequencing can be found in Supplementary Tables 1 and 2.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Reuben Anderson, Donald Carbaugh, Christina Clover, Cecelia Niland, and Sara McQueeney for technical assistance and sample collection. We thank the Council on Dairy Cattle Breeding (Bowie, MD, USA) for genotype, phenotype, and pedigree data,

Interbull (Uppsala, Sweden) for global trait evaluations, and the anonymous reviewers for many helpful comments. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture (USDA). The USDA is an equal opportunity provider and employer. This work was supported by AFRI grant numbers 2019-67015-29321 and 2021-67015-33409 from the USDA National Institute of Food and Agriculture (NIFA) Animal Genome Programs to GEL. LY is supported in part by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 32302699).

Authors' contributions

GEL, TPLS, and ZZ conceived and supervised the study. YG and LY carried out the analyses. YG and GEL drafted the manuscript. WL, GZ, MB, CL, and RLBIV collected the cattle samples. KK isolated HMW DNA, prepared libraries, and performed HiFi sequencing and data processing. PZ, YZ, LF, and JBC collected the genomics data. BDR, LM, and CPVT participated in the discussion of the results. All authors read and approved the final manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jare.2025.04.014.

References

[1] Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 2009;324:522–8.
[2] Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, et al. Challenges and standards in integrating surveys of structural variation. *Nat Genet* 2007;39:S7–S.
[3] Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. *Nature* 2006;444:444–54.
[4] Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. *Nature* 2010;464:704–12.
[5] She X, Cheng Z, Zollner S, Church DM, Eichler EE. Mouse segmental duplication and copy number variation. *Nat Genet* 2008;40:909–14.
[6] Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. *Nature* 2011;470:59–65.
[7] Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature* 2015;526:75–81.
[8] Chaisson M, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* 2019;10:1784.
[9] Vollger MR, Dishuck PC, Harvey WT, DeWitt WS, Guitart X, Goldberg ME, et al. Increased mutation and gene conversion within human segmental duplications. *Nature* 2023;617:325–34.
[10] Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* 2009;10:451–81.
[11] Marques-Bonet T, Girirajan S, Eichler EE. The origins and impact of primate segmental duplications. Trends Genet 2009;25:443–54.
[12] Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, et al. Segmental duplications and their variation in a complete human genome. Science 2022;376:eabj6965.
[13] Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 2007;315:848–53.
[14] Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 2015;517:608–11.
[15] McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 2008;40:1166–74.
[16] Estivill X, Armengol L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet* 2007;3:1787–99.
[17] Dwarshuis N, Kalra D, McDaniel J, Sanio P, Jerez PA, Jadhav B, et al., The GIAB genomic stratifications resource for human reference genomes. bioRxiv 2023:2023.10.27.563846.
[18] Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong association of de novo copy number mutations with autism. Science 2007;316:445–9.
[19] Cook EJ, Scherer SW. Copy-number variations associated with neuropsychiatric conditions. *Nature* 2008;455:919–23.
[20] Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, et al. Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* 2009;459:987–91.
[21] Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'Er I, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* 2009;460:753–7.
[22] Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, et al. Common variants conferring risk of schizophrenia. *Nature* 2009;460:744–7.
[23] Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 2009;459:569–73.
[24] Bochukova EG, Huang N, Keogh J, Henning E, Purmann C, Blaszczyk K, et al. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* 2010;463:666–70.
[25] Xu L, Cole JB, Bickhart DM, Hou Y, Song J, VanRaden PM, et al. Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. *BMC Genomics* 2014;15:683.
[26] Zhou Y, Utsunomiya YT, Xu L, Hay EH, Bickhart DM, Alexandre PA, et al. Genome-wide CNV analysis reveals variants associated with growth traits in Bos indicus. *BMC Genomics* 2016;17:419.
[27] Fadista J, Thomsen B, Holm LE, Bendixen C. Copy number variation in the bovine genome. *BMC Genomics* 2010;11:284.
[28] Bae JS, Cheong HS, Kim LH, NamGung S, Park TJ, Chun JY, et al. Identification of copy number variations and common deletion polymorphisms in cattle. *BMC Genomics* 2010;11:232.
[29] Cicconardi F, Chillemi G, Tramontano A, Marchitelli C, Valentini A, Ajmone-Marsan P, et al. Massive screening of copy number population-scale variation in Bos taurus genome. *BMC Genomics* 2013;14:124.
[30] Keel BN, Keele JW, Snelling WM. Genome-wide copy number variation in the bovine genome detected using low coverage sequence of popular beef breeds. *Anim Genet* 2017;48:141–50.
[31] Bickhart DM, Xu L, Hutchison JL, Cole JB, Null DJ, Schroeder SG, et al. Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. DNA Res 2016;23:253–62.
[32] Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, et al. Analysis of copy number variations among diverse cattle breeds. *Genome Res* 2010;20:693–703.
[33] Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, et al. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res* 2012;22:778–90.
[34] Zhou Y, Yang L, Han X, Han J, Hu Y, Li F, et al. Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history. *Genome Res* 2022;32:1585–601.
[35] Nguyen TV, Vander JC, Wang J, Daetwyler HD, Xiang R, Goddard ME, et al. In it for the long run: perspectives on exploiting long-read sequencing in livestock for population scale studies of structural variants. Genet Sel Evol 2023;55:9.
[36] Durkin K, Coppieters W, Drogemuller C, Ahariz N, Cambisano N, Druet T, et al. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature* 2012;482:81–4.
[37] Jang J, Terefe E, Kim K, Lee YH, Belay G, Tijjani A, et al. Population differentiated copy number variation of Bos taurus, Bos indicus and their African hybrids. *BMC Genomics* 2021;22:531.
[38] Kommadath A, Grant JR, Krivushin K, Butty AM, Baes CF, Carthy TR, et al. A large interactive visual database of copy number variants discovered in taurine cattle. *GigaScience* 2019;8.
[39] Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* 2015;12:780–6.
[40] Huddleston J, Chaisson M, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* 2017;27:677–85.
[41] Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* 2020;583:83–9.
[42] Collins RL, Brand H, Karczewski KJ, Zhao X, Alfoldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. *Nature* 2020;581:444–51.
[43] Ye K, Hall G, Ning Z. Structural variation detection from next generation sequencing. *J Next Gener Sequenc Appl* 2015;01:007.
[44] Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large multiallelic copy number variations in humans. *Nat Genet* 2015;47:296–303.
[45] Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, et al. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* 2005;77:78–88.

[46] 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.
[47] 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;33:1284–98.
[48] Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. *Genome Res* 2017;27:665–76.
[49] Miga KH, Wang T. The Need for a Human Pangenome Reference Sequence. *Annu Rev Genomics Hum Genet* 2021;22:81–102.
[50] Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* 2022;604:437–46.
[51] Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet* 2019;51:30–5.
[52] Talenti A, Powell J, Hemmink JD, Cook E, Wragg D, Jayaraman S, et al. A cattle graph genome incorporating global breed diversity. *Nat Commun* 2022;13:910.
[53] Hickey G, Heller D, Monlong J, Sibbesen JA, Siren J, Eizenga J, et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol* 2020;21:35.
[54] Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol* 2020;21:265.
[55] Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol* 2024;42:663–73.
[56] Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, et al., Building pangenome graphs. bioRxiv 2023.
[57] Andreace F, Lechat P, Dufresne Y, Chikhi R. Comparing methods for constructing and representing human pangenome graphs. *Genome Biol* 2023;24:274.
[58] Crysnanto D, Pausch H. Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. *Genome Biol* 2020;21:184.
[59] Kaminow B, Ballouz S, Gillis J, Dobin A. Pan-human consensus genome significantly improves the accuracy of RNA-seq analyses. *Genome Res* 2022;32:738–49.
[60] Tetikol HS, Turgut D, Narci K, Budak G, Kalay O, Arslan E, et al. Pan-African genome demonstrates how population-specific genome graphs improve high-throughput sequencing data analysis. *Nat Commun* 2022;13:4384.
[61] Groza C, Schwendinger-Schreck C, Cheung WA, Farrow EG, Thiffault I, Lake J, et al. Pangenome graphs improve the analysis of structural variants in rare genetic diseases. *Nat Commun* 2024;15:657.
[62] Li R, Gong M, Zhang X, Wang F, Liu Z, Zhang L, et al. A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. *Genome Res* 2023;33:463–77.
[63] Lan D, Fu W, Ji W, Mipam TD, Xiong X, Ying S, et al. Pangenome and multi-tissue gene atlas provide new insights into the domestication and highland adaptation of yaks. *J Anim Sci Biotechnol* 2024;15:64.
[64] Leonard AS, Crysnanto D, Fang ZH, Heaton MP, Vander LB, Herrera C, et al. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nat Commun* 2022;13:3012.
[65] Leonard AS, Crysnanto 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.
[66] Tian X, Li R, Fu W, Li Y, Wang X, Li M, et al. Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. Sci China Life Sci 2020;63:750–63.
[67] Smith T, 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.
[68] Crysnanto D, Wurmser C, Pausch H. Accurate sequence variant genotyping in cattle using variation-aware genome graphs. Genet Sel Evol 2019;51:21.
[69] Crysnanto D, Leonard AS, Fang ZH, Pausch H. Novel functional sequences uncovered through a bovine multiassembly graph. *Proc Natl Acad Sci U S A* 2021;118.
[70] Leonard AS, Mapel XM, Pausch H. Pangenome-genotyped structural variation improves molecular phenotype mapping in cattle. *Genome Res* 2024;34:300–9.
[71] Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat Genet* 2022;54:518–25.
[72] Li TT, Xia T, Wu JQ, Hong H, Sun ZL, Wang M, et al. De novo genome assembly depicts the immune genomic characteristics of cattle. *Nat Commun* 2023;14:6601.
[73] Wu H, Luo L, Zhang Y, Zhang C, Huang J, Mo D, et al., Telomere-to-telomere genome assembly of a male goat reveals novel variants associated with cashmere traits. bioRxiv 2024:2024.03.03.582909.
[74] Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* 2016;11:e0163962.
[75] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34:i884–90.
[76] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–9.
[77] Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. Bioinformatics 2019;35:2907–15.
[78] Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* 2024.
[79] Lin J, Wang S, Audano PA, Meng D, Flores JI, Kosters W, et al. SVision: a deep learning approach to resolve complex structural variants. *Nat Methods* 2022;19:1230–3.
[80] Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* 2020;21:189.
[81] 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.
[82] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34:3094–100.
[83] Heller D, Vingron M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. Bioinformatics 2021;36:5519–21.
[84] Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;31:3210–2.
[85] Chen Y, Zhang Y, Wang AY, Gao M, Chong Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol* 2021;22:312.
[86] Kirsche M, Prabhu G, Sherman R, Ni B, Battle A, Aganezov S, et al. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat Methods* 2023;20:408–17.
[87] Liu YH, Luo C, Golding SG, Ioffe JB, Zhou XM. Tradeoffs in alignment and assembly-based methods for structural variant detection with long-read sequencing data. *Nat Commun* 2024;15:2447.
[88] Lin J, Jia P, Wang S, Kosters W, Ye K. Comparison and benchmark of structural variants detected from long read and long-read assembly. *Brief Bioinform* 2023;24.
[89] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.
[90] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2.
[91] Kern C, Wang Y, Xu X, Pan Z, Halstead M, Chanthavixay G, et al. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nat Commun* 2021;12:1821.
[92] Shen WK, Chen SY, Gan ZQ, Zhang YZ, Yue T, Chen MM, et al. AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Res* 2023;51:D39–45.
[93] Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 2005;15:901–13.
[94] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;15:1034–50.
[95] Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 2010;20:110–21.
[96] Nassar LR, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res* 2023;51:D1188–95.
[97] Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 2021;372.
[98] McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 2010;28:495–501.
[99] Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 2021;592:737–46.
[100] Vollger MR, Logsdon GA, Audano PA, Sulovari A, Porubsky D, Peluso P, et al. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann Hum Genet* 2020;84:125–40.
[101] Loor JJ, Bionaz M, Invernizzi G. Systems biology and animal nutrition: insights from the dairy cow during growth and the lactation cycle, systems biology and livestock. Science 2011:215–45.
[102] Tanigawa Y, Dyer ES, Bejerano G. WhichTF is functionally important in your open chromatin data? *PLoS Comput Biol* 2022;18:e1010378.