Resource

Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history

Yang Zhou,[1,6] Lv Yang,[1,6] Xiaotao Han,[1] Jiazheng Han,[1] Yan Hu,[1] Fan Li,[1] Han Xia,[1] Lingwei Peng,[1] Clarissa Boschiero,[2] Benjamin D. Rosen,[2] Derek M. Bickhart,[3] Shujun Zhang,[1] Aizhen Guo,[4] Curtis P. Van Tassell,[2] Timothy P.L. Smith,[5] Liguo Yang,[1] and George E. Liu[2]

[1]*Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, Huazhong Agricultural University, Wuhan 430070, China;* [2]*Animal Genomics and Improvement Laboratory, BARC, USDA-ARS, Beltsville, Maryland 20705, USA;* [3]*Dairy Forage Research Center, ARS USDA, Madison, Wisconsin 53706, USA;* [4]*The State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University, Wuhan 430070, China;* [5]*U.S. Meat Animal Research Center, ARS USDA, Clay Center, Nebraska 68933, USA*

A cattle pangenome representation was created based on the genome sequences of 898 cattle representing 57 breeds. The pangenome identified 83 Mb of sequence not found in the cattle reference genome, representing 3.1% novel sequence compared with the 2.71-Gb reference. A catalog of structural variants developed from this cattle population identified 3.3 million deletions, 0.12 million inversions, and 0.18 million duplications. Estimates of breed ancestry and hybridization between cattle breeds using insertion/deletions as markers were similar to those produced by single nucleotide polymorphism–based analysis. Hundreds of deletions were observed to have stratification based on subspecies and breed. For example, an insertion of a Bov-tA1 repeat element was identified in the first intron of the *APPL2* gene and correlated with cattle breed geographic distribution. This insertion falls within a segment overlapping predicted enhancer and promoter regions of the gene, and could affect important traits such as immune response, olfactory functions, cell proliferation, and glucose metabolism in muscle. The results indicate that pangenomes are a valuable resource for studying diversity and evolutionary history, and help to delineate how domestication, trait-based breeding, and adaptive introgression have shaped the cattle genome.

[Supplemental material is available for this article.]

Cattle, as one of the most important livestock animals, contribute to human nutrition and agricultural economics throughout the Holocene epoch by providing milk, meat, hide, and draught force (Gilbert et al. 2018). Modern cattle were domesticated from wild auroch populations multiple times in distinct geographic locations. Two main independent domestication events occurred (Pitt et al. 2019), one in the Fertile Crescent ~10,000 yr ago, leading to humpless taurine cattle (*Bos taurus taurus*), and one in the Indus Valley ~8000 yr ago, leading to humped indicine cattle (*Bos taurus indicus*, also known as zebu cattle). The two lineages have been estimated to have diverged 210,000–350,000 yr ago, well before the domestication events (Loftus et al. 1994), indicating that the auroch populations from which they were derived were genetically distinct. The two subspecies are interfertile and have undergone historical and recent hybridization. A swift and widespread introgression of zebu from the Indus Valley has been suggested to have occurred ~4200 yr ago, possibly mediated by humans in response to a coincident multicentury drought (Verdugo et al. 2019). Multiple other migration waves between the subspecies (Papachristou et al. 2020), along with continued introgression from contemporary aurochs populations (Pitt et al. 2019) and other bovid species such as banteng and yak (Chen et al. 2018), have affected the genome of the species. Selection and adaptation to various climates and other environmental pressures such as altitude and endemic disease have further shaped and diversified the cattle genome, resulting in unusually high levels of diversity among modern cattle globally. Substantial differences remain between the phenomes and genomes of modern taurine and indicine cattle despite ancient and ongoing introgression (Bolormaa et al. 2013). The subspecies display distinct production phenotypes like milk yield, meat quality, stature, coat color, horns, and resistance to heat, drought, and disease (Whipple et al. 1990). Subspecies-specific alleles and differences in shared variant allele

[6]**These authors contributed equally to this work.**
Corresponding authors: George.Liu@usda.gov, yangzhou@mail.hzau.edu.cn, ylg@mail.hzau.edu.cn
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.276550.122.

frequencies may produce these divergent phenotypes and underlie unique sets of quantitative trait loci (QTLs).

The cattle reference genome is based on a single Hereford cow and has undergone multiple upgrades since the initial release (The Bovine Genome Sequencing and Analysis Consortium et al. 2009; Zimin et al. 2009). The current assembly, ARS-UCD1.2, spans 2.7 Gb and has 386 gaps, most of which are found on Chromosome X (Rosen et al. 2020). The reference has been widely used for cattle genetic studies and for genome-enabled selection across major breeds of cattle in both subspecies, although it does not fully capture the genetic diversity of global cattle breeds as shown by genome assemblies of Scottish Highland (ARS_UNL_Btau-highland_paternal_1.0_alt) and Simmental (ARS_Simm1.0) breeds (Rice et al. 2020; Heaton et al. 2021). Existing genotyping arrays are based on SNPs identified by mapping short reads to the Hereford reference genome (Matukumalli et al. 2009), leading to a reference bias and a lack of markers in breed-specific genome segments. The idea of a pangenome representation as a reference that incorporates genetic diversity across diverse populations within a species has been proposed (Tettelin et al. 2005). Pangenome representations for human populations (Sherman et al. 2019), pigs (Tian et al. 2020), and some plant species (Bayer et al. 2020) have been reported. Initial efforts to construct cattle pangenome representations have been recently reported (Crysnanto and Pausch 2020; Crysnanto et al. 2021). A bovine pangenome consortium was proposed in 2020 to coordinate international efforts to develop an inclusive representation of cattle and related bovid species based on the creation of high-quality complete genome assemblies of representatives of as many existing breeds as practical. Members of the consortium have applied the trio binning approach (Koren et al. 2018) to improve efficiency and have generated highly contiguous, haplotype-resolved assemblies for Angus (UOA_Angus_1) and Brahman (UOA_Brahman_1) breeds (Low et al. 2020), along with yak (Rice et al. 2020) and bison (Oppenheimer et al. 2021). In addition, nontrio high-quality assemblies of river buffalo (UOA_WB_1) (Low et al. 2019) and the Braunvieh cattle breed (unnamed) (Crysnanto et al. 2021) have been reported. Substantial differences between the genomes of the Angus (taurine) and Brahman (indicine) breeds were observed by alignment of their genomes, and indications of positive selection in genes with immune-related and fatty acid–related functions were observed (Low et al. 2020). In addition, alignment of all the cattle assemblies mentioned above identified an additional 70 Mb of genome sequence not present in the Hereford reference (Crysnanto et al. 2021), with between 3.3 and 4.4 Mb unique to each taurine assembly. The pangenome representation created from this small sample of breeds showed that variant calls made with the pangenome graph were more consistent than those based solely on the linear Hereford reference. However, the bulk of genome assemblies and genotyping resources represents taurine breeds, in particular breeds of European origin, and fails to adequately represent the indicine breeds being used in tropical and subtropical climates. The observation of breed-specific genome sequence in each breed and the level of diversity among global cattle adapted and selected in specific environments indicate the need to include a broader sampling of existing breeds to improve the pangenome representation. A pangenome incorporating threatened or historical breeds is also urgently needed to prioritize and support conservation efforts.

The discovery of structural variation (SV) revolutionized the understanding of the genomic landscape in many species (Eichler et al. 2007). This form of variation involves larger segments of the genome than the previously recognized microsatellite, single-nucleotide polymorphism (SNP), and short insertion/deletion (indel) variants. SVs can take the form of deletions, insertions, and duplications (commonly grouped under the term copy number variation [CNV]), as well as inversions and translocations, which have been observed to range from 50 bp to 5 Mb (Scherer et al. 2007). The large relative size of SV increases the likelihood that they might impact gene expression and function, such as changing gene dosage, interrupting coding sequence (CDS), or disturbing long-range gene regulation (Alkan et al. 2011; Sudmant et al. 2015b). The known mechanisms of SV formation include nonallelic homologous recombination, nonhomologous end joining (NHEJ), mobile element insertion (MEI), microhomology-mediated break-induced replication (MMBIR), and fork stalling and template switching (FoSTeS) during DNA replication (Zhang et al. 2009). Microhomology-mediated end joining (MMEJ), also known as alternative nonhomologous end-joining (Alt-NHEJ), is a backup pathway for repairing double-strand break (DSB) in DNA through the recombination of short stretches of microhomology (Ottaviani et al. 2014; Black et al. 2019).

Previous studies have shown that SV is present in the genomes of cattle (Bickhart and Liu 2014) and found associations between SV and phenotype (Liu et al. 2010; Bickhart et al. 2012; Cicconardi et al. 2013; Kommadath et al. 2019; Chen et al. 2020; Hu et al. 2020; Jang et al. 2021; Lee et al. 2021; Upadhyay et al. 2021). One example of an SV-affecting phenotype is a chromosomal translocation and subsequent duplication encompassing the *KIT* gene, leading to characteristic coat-color phenotypes in Belgian Blue and Brown Swiss cattle (Durkin et al. 2012). Another example is a 660-kb deletion found to be at high frequency in Nordic Red cattle, which is associated with antagonistic effects on fertility and milk production (Kadri et al. 2014). Cattle SV has been identified by mapping reads from various breeds to the Hereford reference genome, although detection can be more complicated than SNP. Complex SVs can be located in or near repetitive sequences like MEI, interfering with accurate read mapping and introducing ambiguity in breakpoint definition (Abel et al. 2020; Collins et al. 2020; Ho et al. 2020). Multiple solutions to this problem have been applied for detecting and genotyping SV, including read-pair (RP) or paired-end mapping (PEM) and read-depth (RD) or split-read (SR) analysis (Mills et al. 2011). The success of each algorithm depends on the SV type and size, and all are sensitive to the quality of the reference genome and the depth of sequence coverage. The RD approach is the most used and has successfully identified SV between *B. taurus* and *B. indicus* cattle (Low et al. 2020; Jang et al. 2021), although it has lower accuracy in defining SV boundaries and the use of a single strategy can introduce a high proportion of false positives (Handsaker et al. 2015). Combining strategies could significantly increase the sensitivity and specificity of SV detection.

Population analyses based on genomic SNP, indel, and microsatellite markers suggest that cattle can be divided into three major genetic groups (MacHugh et al. 1997; Bolormaa et al. 2013; Pérez O'Brien et al. 2014), including Asian indicine, Eurasian taurine, and African taurine. The genetic phylogeny reconstructed the historical migratory routes of cattle from their origins around the Fertile Crescent (Verdugo et al. 2019) and across East Asia (Chen et al. 2018). The construction of a representative pangenome for extant cattle would necessarily account for these three groups and represent all SVs. The 1000 Bull Genomes Project was initiated to provide a database for the imputation of genetic variants (mainly SNP and indel) in all cattle breeds. However, the approach in

that project creates a bias against SV as the regions with multiple mapped reads are generally ignored and thus excluded (Chen et al. 2017). Previous population genetics studies based on SV, including several of ours (Bickhart et al. 2016; Xu et al. 2016; Hu et al. 2020), are limited in sample size and breed representation, as well as applying suboptimal SV calling methods and typing platforms (Jang et al. 2021; Upadhyay et al. 2021). Therefore, the SV-based breed diversity in cattle is not well studied yet.

To address these needs, we applied a combination of multiple approaches of SV prediction to 898 cattle genomes representing 57 cattle breeds to assemble an enhanced SV catalog and construct a cattle pangenome based on unmapped reads in this study.

Results

SV catalog based on the ARS-UCDI.2 assembly for 898 WGS data of 57 cattle breeds

A bioinformatic pipeline (see Methods) based on the ARS-UCD1.2 assembly was developed to map 254 billion reads for 898 animals (average coverage of ~16×, ranging from 5.08× to 54.42×) of 57 well-known cattle breeds worldwide (Supplemental Table S1) and call SVs from the alignments. The pipeline combined four diverse SV-finding algorithms (Pindel, LUMPY, Breakdancer, and DELLY) (Chen et al. 2009; Ye et al. 2009; Rausch et al. 2012; Layer et al. 2014), maximizing sensitivity and precision by retaining only SV calls >50 bp where >80% of the predicted length was supported by at least two algorithms. A total of approximately 3.3 million deletion calls, approximately 0.18 million duplication calls, and approximately 0.12 million inversion calls were detected from these 898 animals (Supplemental Table S2). The SV calls derived by the pipeline from short reads for the Angus and Brahman samples previously described (Low et al. 2020) were compared with the results derived from the Pacific Biosciences (PacBio) long reads. Over 81% and 79% of SVs could be successfully validated by the PacBio long-read result generated by PBSV (https://github.com/PacificBiosciences/pbsv) and Sniffles (Sedlazeck et al. 2018), respectively. The total number of SVs from 100 iteratively random sampling suggested that our SV detection power became saturated either in the whole population (Fig. 1A) or in the five widely used cattle commercial breeds in Europe (Holstein, Angus, Hereford, Simmental, and Charolais in Supplemental Fig. S1).

Most deletions (87.23%) were <5 kb, whereas some duplications (27.63%) and inversions (38.23%) ranged from 10 kb to 1 Mb (Fig. 1B). Nearly 75% of all detected SV regions were supported by at least two samples and covered 1.15 Gb of the autosomes of the ARS-UCD1.2 assembly. Most of the three types of detected SVs (73.5%~76.1%) had minor allele frequency (MAF) < 0.01, consistent with previous studies indicating that high SV diversity exists among different cattle individuals (Chen et al. 2018; Verdugo et al. 2019). But a small portion of SV regions (0.1%~0.6%) appeared in >50% of the 898 animals, including one sequence of 115,160 bp (Chr 6: 5,542,125–5,657,285) that appears to be deleted in 70.7% of the studied animals. SV cluster and desert regions across cattle populations were detected by binning the genome into nonoverlapping 10-kb windows and calculating a weighted number of supported samples for each window. A hidden Markov chain model located 428 SV desert regions (weighted frequency < 0.1, length > 500 kb) and 146 SV cluster regions (weighted frequency > 5, length > 500 kb) (Fig. 1C; Supplemental Fig. S2; Supplemental Table S3). The SV regions and their normalized frequencies (see "Data access") were provided as the UCSC Genome Browser customer tracks to facilitate future analyses, such as checking their relationship with respect to other genomic structures.

Detection and localization of missing sequences for the cattle pangenome

Insertions relative to the reference assembly represent potential population-specific genome segments. Putative nonreference genome sequence present in the population but missing from the Hereford reference genome was identified by assembling unmapped reads into 1,163,034 redundant contigs (≥1000 bp) for 898 samples (Supplemental Table S4). Approximately 83 Mb (83,650,473 bp) novel cattle genome DNA sequence distributed across 18,231 representative contigs was discovered after removing redundant contigs and contaminants such as microbial genomes, with the longest contig spanning 82,663 bp (Supplemental Table S5). Nearly 75% of this putative novel sequence (63,404,990 bp) is not present in the seven other existing highly contiguous cattle assemblies (Angus, Brahman, Brown Swiss, Jersey, Highland, Holstein, and Simmental), as alignment only identified contigs summing to 20 Mb (20,245,483 bp, 90% identity, length > 1000 bp) that failed to align to ARS-UCD1.2 but did align to one of the other assemblies. The Brahman assembly contributed the largest number of mapped contigs (848, 4.65% of all representative contigs, 90% identity and length coverage > 80% of supercontig) among the seven breeds and had the most mapped sequence (1,248,880 bp), likely because it represents a *B. indicus*–derived breed (Supplemental Fig. S3; Supplemental Table S6).

The representative novel contigs were localized on the ARS-UCD1.2 assembly using three strategies. First, each read that mapped to the contig was evaluated for whether the paired read also mapped to the contig. Where it did not, the paired read was aligned to the reference genome to provide information about the likely location. This procedure resolved the locations for 546 representative contigs directly. Second, for contigs that mapped to one of the seven other cattle assemblies, we extracted sequence 500 kb upstream of and downstream from the map position of the contig from that assembly and aligned that sequence to the reference. This procedure uniquely placed 449 representative contigs on the ARS-UCD1.2 assembly, with the longest placed sequence being observed in 859 animals and spanning 8562 bp. Finally, linkages were created between contigs and known mRNA sequences using EST, as described in the next paragraph. The contig location was defined by the mRNA-coding gene in the ARS-UCD1.2 assembly.

The potential of the novel sequence identified in the pangenome to represent parts of expressed genes was evaluated by determining if transcriptome sequence data that fail to map to the reference assembly might map to the representative nonreference contigs. Unmapped RNA-seq reads (2.28 billion reads) from 1415 previously reported transcriptome sequencing data sets representing 98 different tissues/cell lines (Supplemental Table S7) were aligned to the contigs, and 39,943 potential exons were identified in a small (<5%) proportion of the contigs. A substantial number (9410; 23.56%) of these putative exons were supported by reads mapping from two or more transcriptome data sets, with one exon supported by a maximum of 1087 samples. The set of novel exons with support from multiple data sets was dispersed in 1871 contigs and was further confirmed by alignment to the cattle EST database. We successfully verified the existence of 5861 exons supported by 11,441 cattle EST sequences through strict criteria (95% identity, 95% coverage, >30 bp in length). The EST

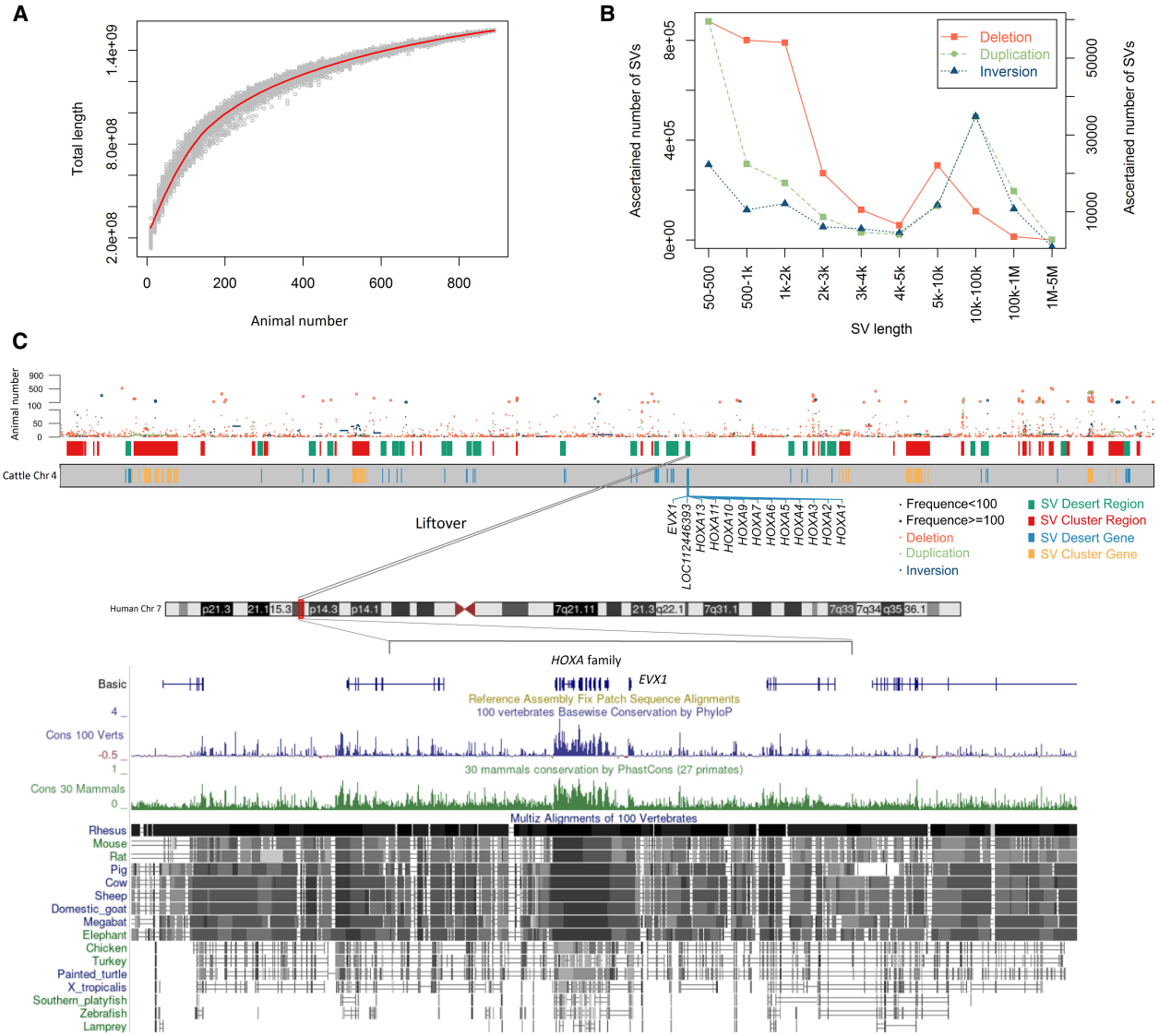

Figure 1. Statistics and chromosomal distribution of structural variations (SVs) in 898 cattle. (*A*) Simulations of the increase in SV number detected with the increase of animal number. No more substantial SV increase when the animal number is over 800, indicating that selected cattle were sufficient to capture the majority of bovine SV. The red line is the fit curve of data points in light green. (*B*) SV density of different sizes for each SV type; the *left y*-axis is for the deletion, and the *right y*-axis is for duplication and inversion. (*C*) SV distribution on cattle Chromosome 4 near the *HOXA* locus. (SV Desert Gene) Genes located in the SV desert region; (SV Cluster Gene) genes located in the SV cluster region.

sequences to which the exons were mapped were then aligned to the RefSeq database of cattle transcripts to assign the novel exons to genes, which successfully built linkages between 490 EST sequences and 188 RefGenes and indicated that the pipeline was able to reconstruct coding segments of expressed genes absent from the cattle reference genome. Overall, novel exon regions for 19 genes were conservatively identified (Supplemental Table S8). It is possible that some other putative exons are valid but either these putative exons were not present in the EST sequence or the transcript represented by the EST was not present in the database of known genes. The identification of EST and associated transcripts provided the third way to roughly localize additional 27 representative contigs on the ARS-UCD1.2 assembly with relatively low precision. Combining the read pair, alternate assembly, and EST strategies for determining the chromosomal position of novel genome segments, we localized 1007 contigs with a total of 2,622,333 bp (Fig. 2). By count, Chr X had 112 placed contigs, followed by Chr 1 (76), Chr 12 (73), and Chr 4 (61). By density per megabase, Chr 12 had 0.83 contig per megabase, followed by Chr X (0.81), Chr 23 (0.53), and Chr 4 (0.51). Overall, the pangenome assembly contributed large sequence insertions containing 279 genes, including 246 coding genes, 23 lncRNAs, and 10 pseudogenes (Fig. 2).

Breakpoint identification and footprints of deletion events

It is well known that short reads have compromised power to detect the other SVs (such as duplication and inversion, translocation, or other complex SV events) compared with deletions. For the rest of this project, we focused on deletions unless

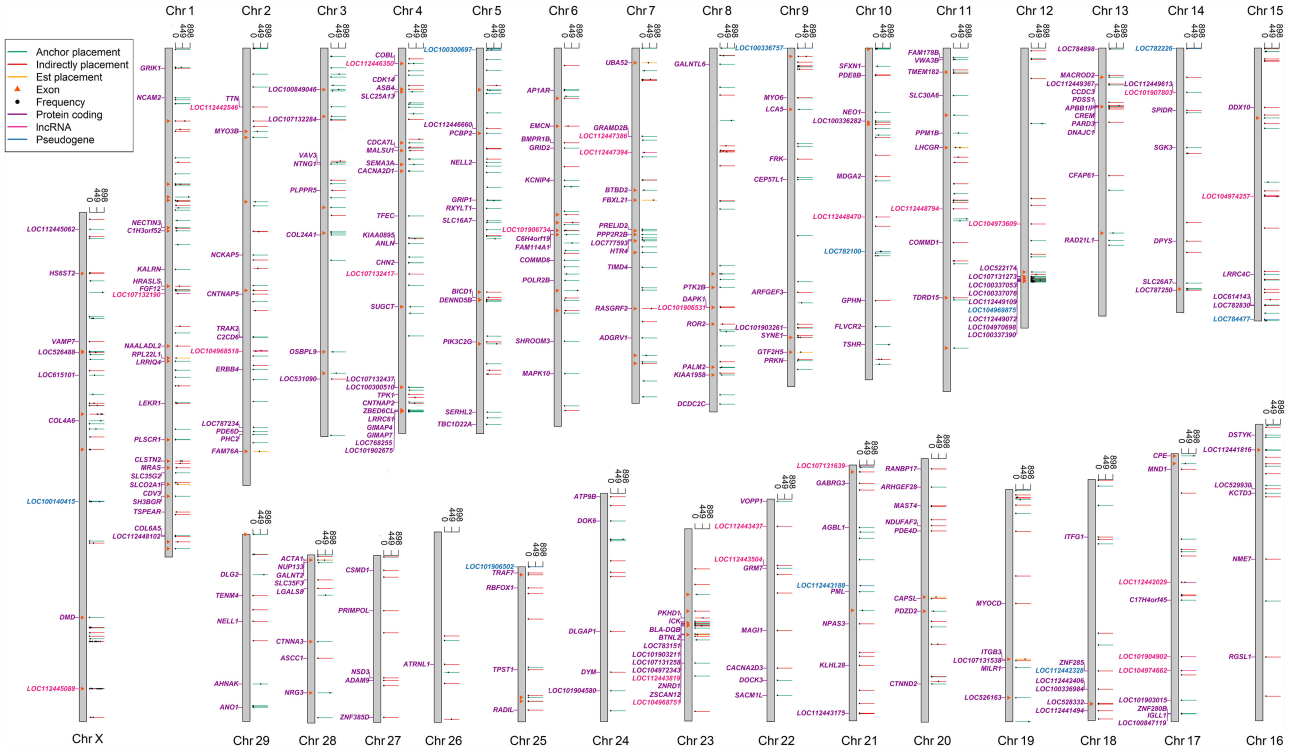

Figure 2. Cattle pangenome contig locations. Bars with three different colors represent contig mapped by three different strategies. The location of the black dot on the bar shows the frequency of the contig ($\log_2$(animal number)) in the 898 animals. The bars marked with triangles are the contigs with novel exons. Gene symbols with different colors represent protein-coding genes (purple), lncRNAs (pink), and pseudogenes (blue).

otherwise stated. The four algorithms used to detect SV often disagreed on the position of breakpoints, with only 22.87% of breakpoints being consistent between/among different algorithms for the same deletion. This ambiguity was addressed by the design and implementation of a new integrated pipeline using a target split reads strategy based on the four algorithms (for detail, see Methods) (Fig. 3A). The pipeline successfully identified 1.7 million deletion calls (51.5%) across 898 animals with breakpoints in 1-bp resolution corresponding to 92,520 unique events with an average length of 2467.92 bp. The pipeline was validated by design and testing of 80 target-PCR (eight deletions in 10 animals), which showed that all deletions and breakpoints were precisely consistent with the predictions (Supplemental Table S9; Supplemental Fig. S4). The pipeline could also classify the deletion into one of three categories according to mutational signatures at the breakpoints (Fig. 3A). These categories included type 1 with breakpoints adjacent to microhomology sequences (72,304/92,520, 78.15% of deletions), type 2 with breakpoints adjacent to small inserted sequences (11,648/92,520, 12.59% of deletions), and type 3 with breakpoints perfectly supported within the split reads (8558/92,520, 9.25% of deletions). The deletion breakpoints with adjacent microhomology sequences corresponding to type 1 may be characteristic of DSB repair through NHEJ or MMEJ and the signatures of FoSTeS/MMBIR (Supplemental Fig. S5). The deletion breakpoints with minor insertions are characteristic of DNA DSB repair through direct ligation by NHEJ, which includes type 2 (breakpoints adjoin a small inserted sequences) and probably part of type 3 (with the small inserted sequence deleted during the DSB repairing, Supplemental Fig. S5).

The majority (56,918/72,304; i.e., 78.72%) of type 1 deletions were enriched in 1- to 4-bp microhomology sequences, as shown by the red curve between two dashed lines in the upper panel of Figure 3B, consistent with previous studies in other species (Ottaviani et al. 2014). A total of 2235 (2.32% of 92,520) deletions were detected to have the same or near breakpoints as the underlying full-length MEIs have, and the full-length MEIs were defined to have at least 80% consensus sequence coverage; 1610 (72.03% of 2235) MEI deletions were with microhomology repeats, of which 1454 (90.31% of 1610) were with the length of microhomology repeats >5 bp. Deletions with 1- to 4-bp microhomology repeats showed much lower frequencies, whereas deletions with 6- to 20-bp microhomology repeats had significantly higher frequencies than other kinds of deletions (Fig. 3B, blue curve in upper panel). Deletions with 6- to 20-bp microhomology repeats were closer to the boundaries of MEIs than deletions with 1- to 4-bp microhomology repeats were (Fig. 3B, bottom panel). It is noted that most of these "deletions" that match MEIs were in fact MEI insertions that were inserted and thus present in the reference genome but not in the sequenced sample. This type of "deletion" event is different from the rare "true" MEI deletion event that precisely removed MEIs from a genome (van de Lagemaat et al. 2005). Those underlying MEI events were enriched in three peaks—~270 bp, ~1300 bp, ~8500 bp—that correspond to a list of bovine-specific MEIs, including BOV-A2, BTLTR families, and L1-BT, respectively (Fig. 3C). Comprehensive analysis of the microhomology repeat showed that each MEI type has its unique preferences for the microhomology repeat lengths and base components (Supplemental Figs. S6, S7), such as microhomology repeats of the BOV-A2 were enriched

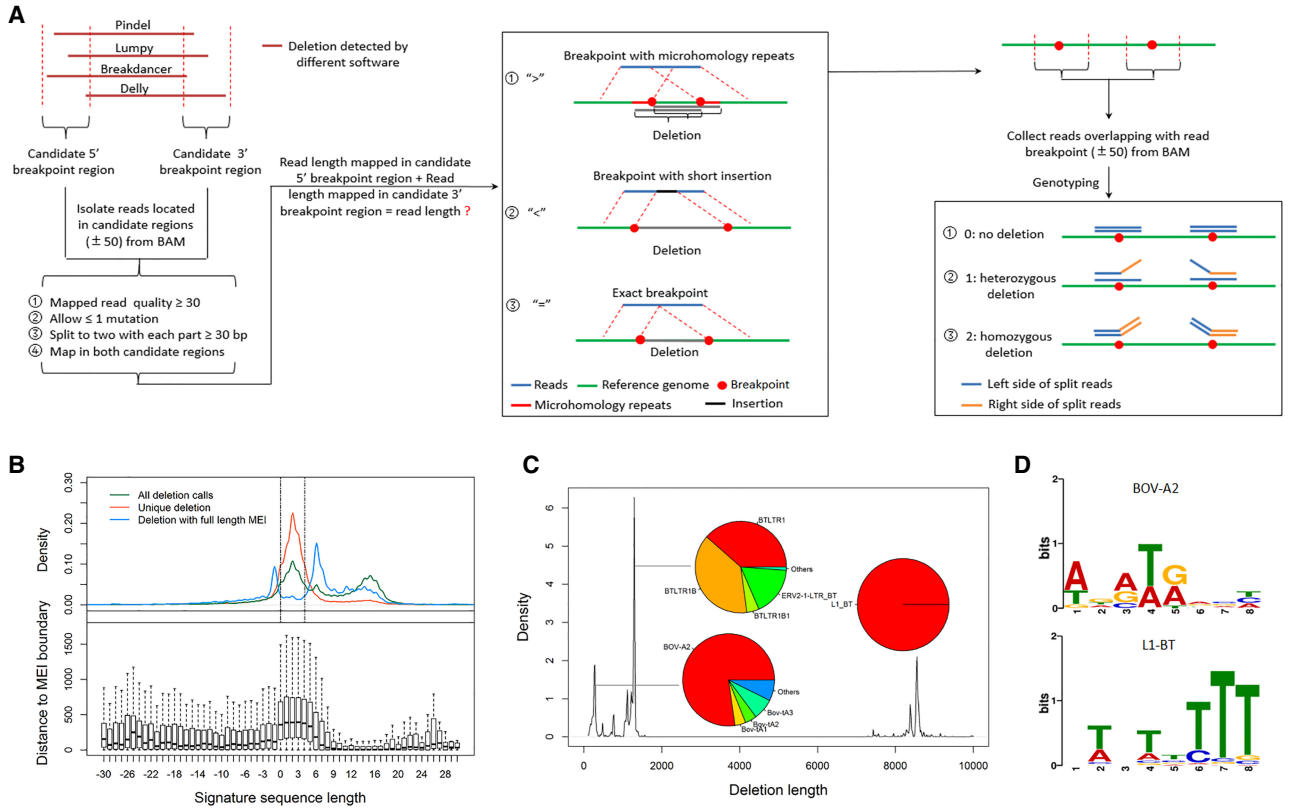

Figure 3. Characteristics of deletions with three breakpoint types. (*A*) Flow chart to detect and genotype deletions with three breakpoint types. (*B*) Distribution of deletion signatures; the negative values of the *x*-axis represent the lengths of the insertions, and the positive values of the *x*-axis correspond to the lengths of the microhomology repeats. (*C*) Length enrichment of the deletion with microhomology repeat. (*D*) Motif enriched by MEME for the microhomology repeat sequences of BOV-A2 and L1-BT.

for AT, whereas those L1-BT family members were mainly enriched for T (Fig. 3D).

Impact of cattle SVs on the functional genome

Potential effects of cattle SVs on genome function were examined by alignment with annotation of 27 distinct genome features, including genic/coding sequence and regulatory elements based on DNA methylation or histone modifications. Enrichment analysis (Fig. 4A) of these features revealed that a majority of SV were enriched (Log_2(Fold Enrichment) > 0) in nongenic gene deserts, intergenic regions, or pseudogenes and not predicted to disrupt genomic regions associated with gene function. Relatively few SV were identified within regulatory elements (e.g., promoters, enhancers, TSS-HMR, DMR) and fewer still in other genic regions. SVs that did appear in genic regions were less likely to be found in the CDS than in other genic regions (Fig. 4A), and among those in the CDS, fewer appeared in the genes with alternative splicing, and barely any appeared in the conserved exon for the gene with alternative splicing genes (Fig. 4B). The precisely mapped breakpoints produced by the SV pipeline supported the observation that rare deletions in CDS regions often peak in the lengths divisible by three, suggesting in-frame rather than out-of-frame, that is, frameshift mutations (Fig. 4C), reflecting selection pressure minimizing SV disruption of the coding function. Genes located in the SV cluster regions were involved in functions that vary among individuals, including many genes involved in the immunity and sensory perception of smell (Supplemental Table S10A,B). Several previously reported gene clusters, including pregnancy-associated glycoproteins (PAG), were detected (Gilbert et al. 2018). In contrast, genes located in the cattle SV desert regions were dispersed in the essential functions of life activities, including embryo development, nervous system, and others (Supplemental Table S10C,D). For example, the *HOXA* gene cluster that spatially and temporally controls embryo development is located in an SV desert region of Chr 4: 68,590,001–69,120,000. Moreover, like most SV desert regions, it showed a highly conserved sequence in multiple sequence alignments of 100 vertebrate species (Fig. 1C).

A database of functional genomic features (genes, promoters, enhancers, DMRs, and other chromatin states) affected by deletions was constructed (Supplemental Tables S11–S14). On average, 22 RefGenes were entirely deleted within a single animal. Within the populations, a total of 1489 nonredundent RefGenes (11.42% of 13,033 RefGenes) were affected and significantly enriched (FDR < 0.05) for immune and metabolic-related terms (Supplemental Table S15A,B). Analysis of deletions in regulatory elements showed that 338,674 regulatory elements were deleted entirely, supported by split reads in at least one animal.

Potential functional roles of deletions were also examined by comparing their chromosomal positions to cattle QTLs found in the public database that includes 161,781 QTLs for 680 different traits from 1049 publications (Hu et al. 2018). This analysis detected 13,850 QTLs/association sites, including genomic segments

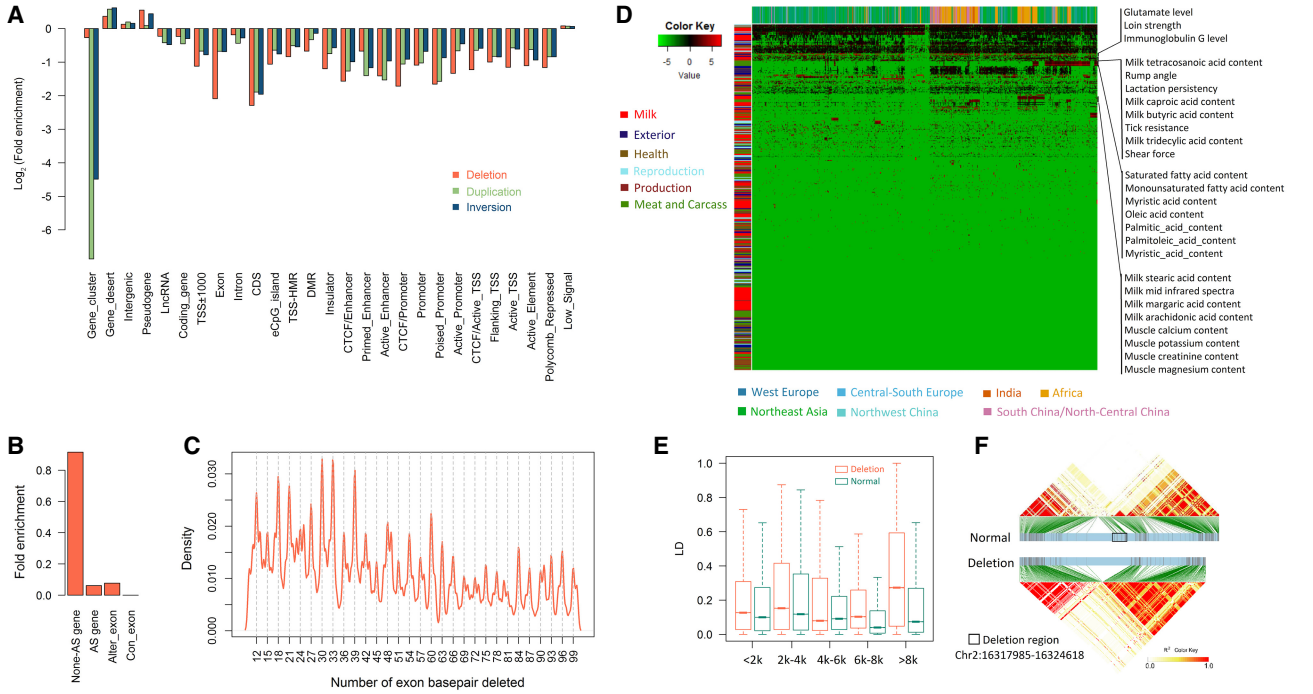

Figure 4. Impacts of the SV on the functional genomic features and QTLs. (*A*) Enrichment of SVs within various genomic features. (*B*) Enrichment of SVs within genes with alternative splicing events. (None_AS_gene) Genes with no alternative splicing event, (AS_gene) genes with alternative splicing events, (Alter_exon) alternatively spliced exons, (Con_exon) conserved exons for the genes with alternative splicing events. (*C*) Density distribution of the exon sequence length deleted. (*D*) Enrichment of deletions within different QTLs. The $\log_2$(fold enrichment) scores of deletions in each QTL for each animal were used to plot this heatmap. (*E*) SNP LD changes caused by deletions with different lengths. (*F*) The change of one LD block induced by one deletion.

affected by deletions. The QTLs that overlapped with deletion were most highly enriched ($\log_2$(fold enrichment) > 2 in at least 200 animals, 2000 bootstraps, $P < 0.05$) for immunoglobulin G level, loin strength, glutamate level, and other traits (Fig. 4D; Supplemental Table S16). In contrast, deletions were underrepresented in the QTLs related to body weight, body size, feed intake, height, etc. (Supplemental Table S16).

Some SV-containing QTL regions associated with traits like milk content (stearic acid, margaric acid, arachidonic acid)–related and muscle content (milk calcium, magnesium, potassium, creatinine contents)–related traits were highly enriched in the African cattle with *B. indicus* ancestry (Fig. 4D; Supplemental Table S16). We also detected that the LD between SNPs on the two sides of the deleted sequences increased, especially when the deletion length was >8 kb (Fig. 4E). For example, the deletion located in Chr 2: 16,317,985–16,324,618 increased the LD values between SNPs around it and enlarged the LD block on its right side (Fig. 4F).

Population structure derived from deletions in cattle

High-quality deletion events (i.e., deletions supported by at least two software with 80% overlap) were genotyped in each animal, and 719 animals had a call rate >90% (Fig. 3A). Genotype accuracy of the pipeline was evaluated using a family trio, for which 1747 common deletions (94.33%) were validated by pedigree information. A more global evaluation of genotype accuracy was performed using individual deletion calls as alleles of genetic markers to analyze population structure and compare the predicted structure to that inferred from SNP markers (Fig. 5A). Principal component analysis (PCA) of deletion genotypes successfully distinguished animals from different geographical regions (Fig. 5B) and mirrored the structure predicted from SNP. The first component divided all animals into *B. taurus* and *B. indicus*. The cattle from India and South China within *B. indicus* were clearly separated by the second component. The hybrid breeds were located between *B. taurus* and *B. indicus*.

Admixture analysis based on deletion genotypes also reflected the results of SNP-based analysis and confirmed the accuracy of the deletion identification and genotyping (Fig. 5C). At K = 2, deletion genotypes clearly distinguished the pure *B. taurus* and *B. indicus* and accurately estimated admixture components for the cattle breeds from North Central China and Africa. At K = 3, besides *B. indicus* (in green), *B. taurus* were labeled as three clusters according to the geographically ancestral components: European taurine (in red), Eurasian taurine (red and blue), and East Asian taurine (blue). At K = 7, all five geographically ancestral components (India indicine and South China indicine except for the three taurine breeds) were separated. The Hereford breed for the ARS-UCD1.2 assembly was also separated. This offered an excellent chance to explain the three-cross breeding breed Beefmaster, which carries about one-half Brahman ancestry (India indicine) and one-fourth each of Hereford (European taurine) and Shorthorn (European taurine in England) using deletions. Overall, our results showed, for the first time in cattle, that the deletion genotyped produced essentially the same population structure and ancestral components compared with the published results based on SNP genotypes (Chen et al. 2018; Kim et al. 2020), providing strong

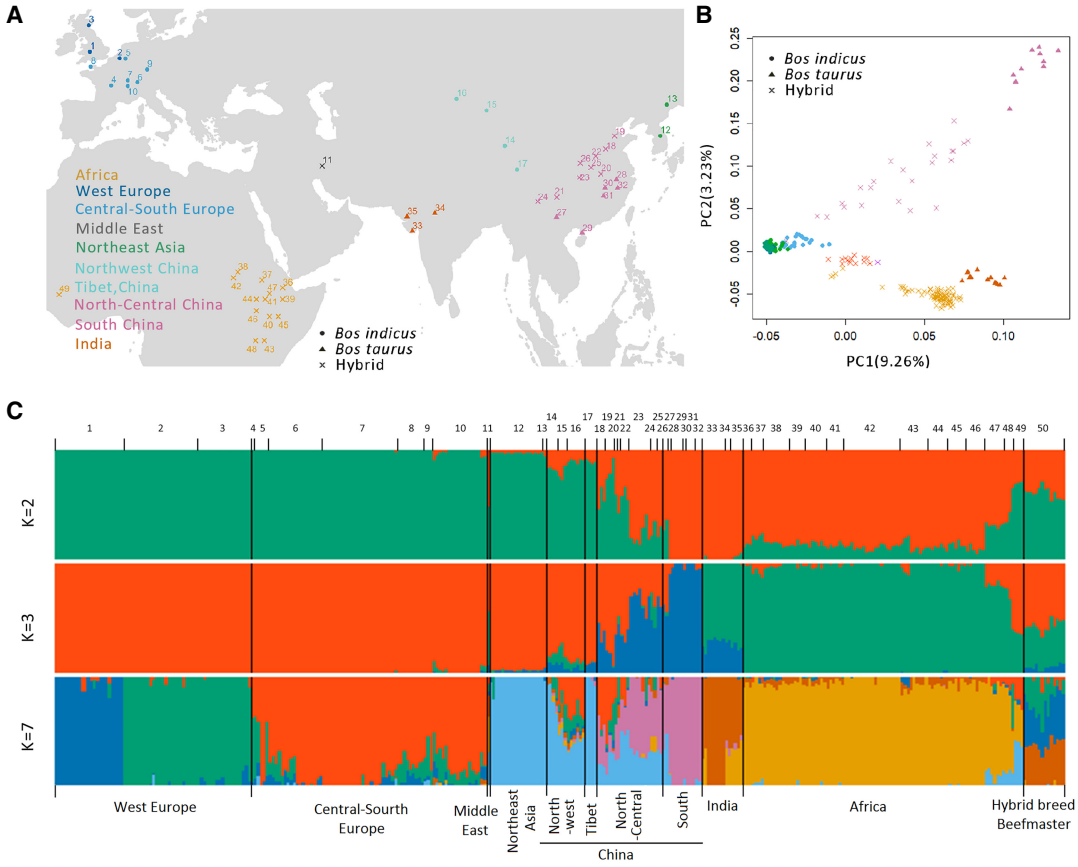

Figure 5. Population structure derived from deletions in cattle. (*A*) Geographic map of the origins of 50 cattle breeds after sample and deletion genotyping quality filtering (see Methods); the dot, triangle, and cross represent *B. indicus*, *B. taurus*, and Hybrid, separately. (*B*) Principle component analysis (PCA) showing PC1 and PC2 of 50 different cattle breeds. (*C*) ADMIXTURE analysis of 50 different cattle breeds. The numbers listed in the figure represent (1) Hereford, (2) Holstein, (3) Angus, (4) MaineAnjou, (5) Belgian Blue, (6) Simmental, (7) Limousin, (8) Jersey, (9) Gelbvieh, (10) Charolais, (11) Rashoki, (12) Hanwoo, (13) Yanbian, (14) Chaidamu Yellow, (15) Mongolian, (16) Kazakh, (17) Tibetan Yellow, (18) Luxi, (19) Bohai Black, (20) Dabieshan, (21) Weining, (22) Jiaxian Red, (23) Enshi, (24) Dengchuan, (25) Zaobei, (26) Xuanhan, (27) Wenshan, (28) Wannan, (29) Leiqiong, (30) Jinjiang, (31) Ji'an, (32) Guangfeng, (33) Nelore, (34) Gir, (35) Brahman, (36) Ogaden, (37) Mursi, (38) Kenana, (39) Horro, (40) Goffa, (41) Fogera, (42) Butana, (43) Boran, (44) Barka, (45) Arsi, (46) Afar, (47) Sheko, (48) Ankole, (49) N'Dama, and (50) Beefmaster.

support that the SV calling pipeline accurately identifies and genotypes this class of SV.

Selection signatures of deletions among different cattle populations

Deletion genotypes were used to identify signatures of selection within the population, beginning with an estimation of fixation index (F_{ST}) statistics to determine deletion frequency differences between *B. taurus* and *B. indicus* (Fig. 6A). There were 135 deletions and 551 genes within a 300-kb proximity to SV, corresponding to the top 1% of F_{ST} values, which implies potential selection signatures. Functional enrichment analysis of the genes with differing SV frequencies between cattle subspecies indicates that they are enriched for functions in resistance, heat stress, energy metabolism, and others (Supplemental Table S17), which might be related to the adaptive selection to the local environments between *B. taurus* and *B. indicus*. Two additional selection signature mappings using *di* (Akey et al. 2010) statistics were also applied to the population after subclassifying them by geographic ancestry components according to admixture results to explore the differences among different subpopulations and breeds (Supplemental Figs. S8, S9). The cattle subpopulations were first classified according to the admixture results, namely, geographically ancestral components (West Europe, Central-South Europe, Northeast Asia, South China, India, Africa). The top 1% deletions with the highest *di* values, representing candidate selection regions, included 625 deletions unique to each population, and 106 deletions were shared by two or more subpopulations (Fig. 6B). A detailed focus on breed differences among five widely used cattle commercial breeds in Europe (Holstein, Angus, Hereford, Simmental, and Charolais; each breed with over 70 animals) identified 446 deletions under positive selection in at least one breed (Fig. 6C). There were 2542 and 1442 genes within 300-kb windows overlapping SV with selection signatures according to each of the two *di*-based analyses. GO analyses showed consistent enrichment results with the positively selected genes at either the subpopulation or breed level. For example, positively selected genes were primarily enriched in the cellular response to oxygen-containing compound, leukocyte apoptotic process, and cellular response to lipid in India or *B. indicus* (INA), as well as the cellular response to starvation in Africa (AF), and cell death in response to oxidative stress in West Europe (WE) (Supplemental Fig. S10A). Positively selected genes were mainly enriched in regulation of transport and regulation of cellular location

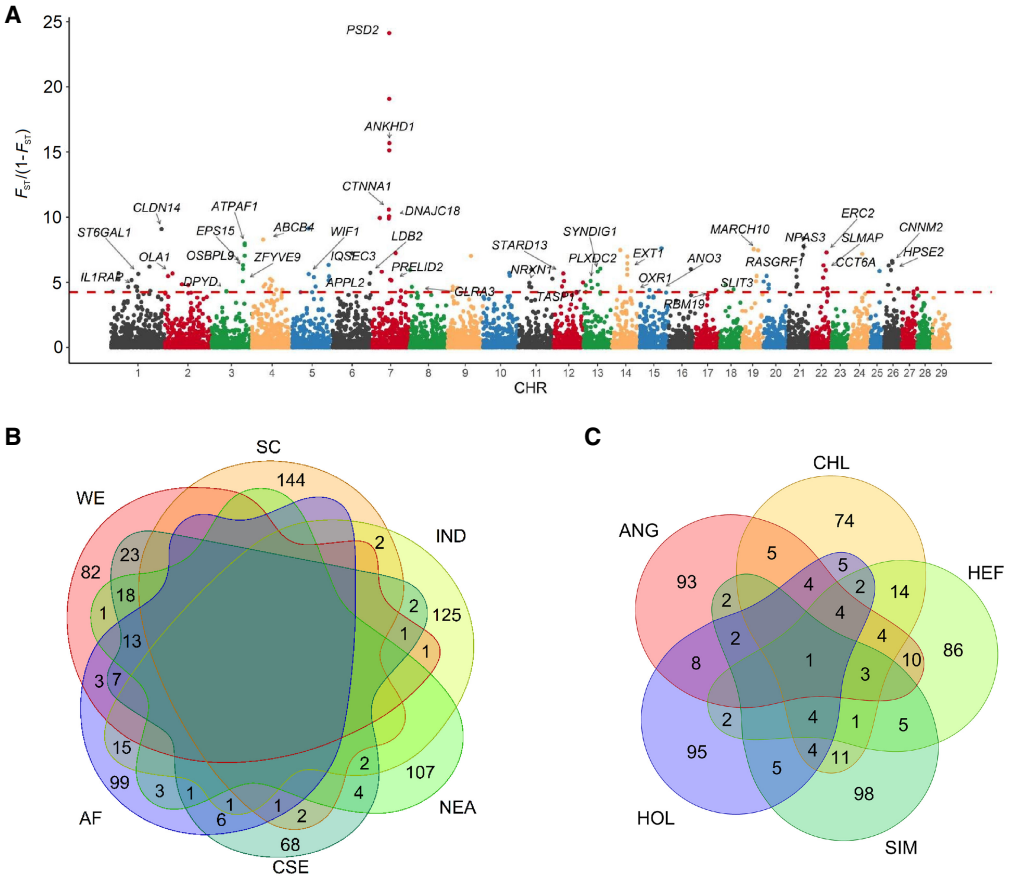

Figure 6. Selection signature analysis of different cattle populations and breeds. (*A*) F_{ST} analysis between *B. taurus* and *B. indicus*. (*B*) Venn diagram for shared versus population-specific selection events among six cattle populations of different geographic origins. (*C*) Venn diagram for shared versus breed-specific selection events among five widely used European commercial cattle breeds. (WE) West Europe, (CSE) Central-South Europe, (NEA) Northeast Asia, (NWC) Northwest China, (NCC) North-Central China, (SC) South China, (INA) India, (AF) Africa; (ANG) Angus, (CHL) Charolais, (HEF) Hereford, (SIM) Simmental, (HOL) Holstein.

in Holstein (HOL), along with positive regulation of calcium ion transport in Hereford (HEF) (Supplemental Fig. S10B). These results confirmed 54 genes, which had been reported under selection before (Supplemental Table S18), including one well-known gene (*KIT*) related to coat color in South China cattle.

The merged positively selected SV results of three mappings were evaluated for deletion type, identifying 35 with type 1 deletions falling between the ends of full-length MEIs (Supplemental Table S19). Examination of the 35 MEIs among 21 artiodactyl species assemblies ranging from family Tragulidae to Bovidae provided an estimate of the age of each element (Supplemental Fig. S11). There were 23 MEIs only observed in assemblies from the genus *Bos* (66%), whereas the other 12 MEIs predated the split from *Bos* and were found in water buffalo, bison, and yak. The frequencies of MEI deletion genotypes also showed high diversity among different cattle subpopulations or breeds (Supplemental Fig. S12). Nine coding genes and two noncoding genes with introns were impacted by MEIs (Supplemental Fig. S12; Supplemental Table S19).

Genes harboring SVs

Two examples (*APPL2* and *ATPAF1*) were shown in Figure 7 and Supplemental Figure S13. A deletion was found in the intronic region of ATP Synthase Mitochondrial F1 Complex Assembly Factor 1 (*ATPAF1*), which was associated with energy metabolism (Supplemental Fig. S13). One Bov-tA1 element insertion was studied in detail in the first intron of the *APPL2* gene in the Hereford reference assembly of Chr 5: 68,751,179–68,751,347. This insertion was also observed throughout the Hippotraginae, Caprinae, Alcelaphinae, Cephalophinae, Antilopinae, and Bovinae-Bubalus genomes but was absent in bison and yak (Supplemental Fig. S12). The presence of Bov-tA1 insertion in Cervidae suggests that its origin was in the last common ancestor ~27 million yr ago and survived in all current lineages, except for the linage leading to bison and yak, where it was probably lost (Fig. 7A). Although only insertion genotype (0/0) was observed in all European *taurus*, it was not seen in India or *B. indicus* (INA) sampled. Small portions of heterozygous deletions (0/1) were observed mostly in the *B. taurus* Northeast Asia and Northwest China groups. Furthermore, the frequency of Bov-tA1 deletion in *APPL2* trended higher as the proportion of *B. indicus* haplotypes increased (Fig. 7B).

This Bov-tA1 indel in the first intron of *APPL2* was validated through the IGV visualization (Fig. 7C; Thorvaldsdóttir et al. 2013). The APPL2 protein is an adaptor protein that regulates immune response and olfactory functions and can mediate growth factor–induced cell proliferation and glucose metabolism in muscle (Miaczynska et al. 2004; Cheng et al. 2014). A combination of whole-genome DNA methylation and chromosome state data

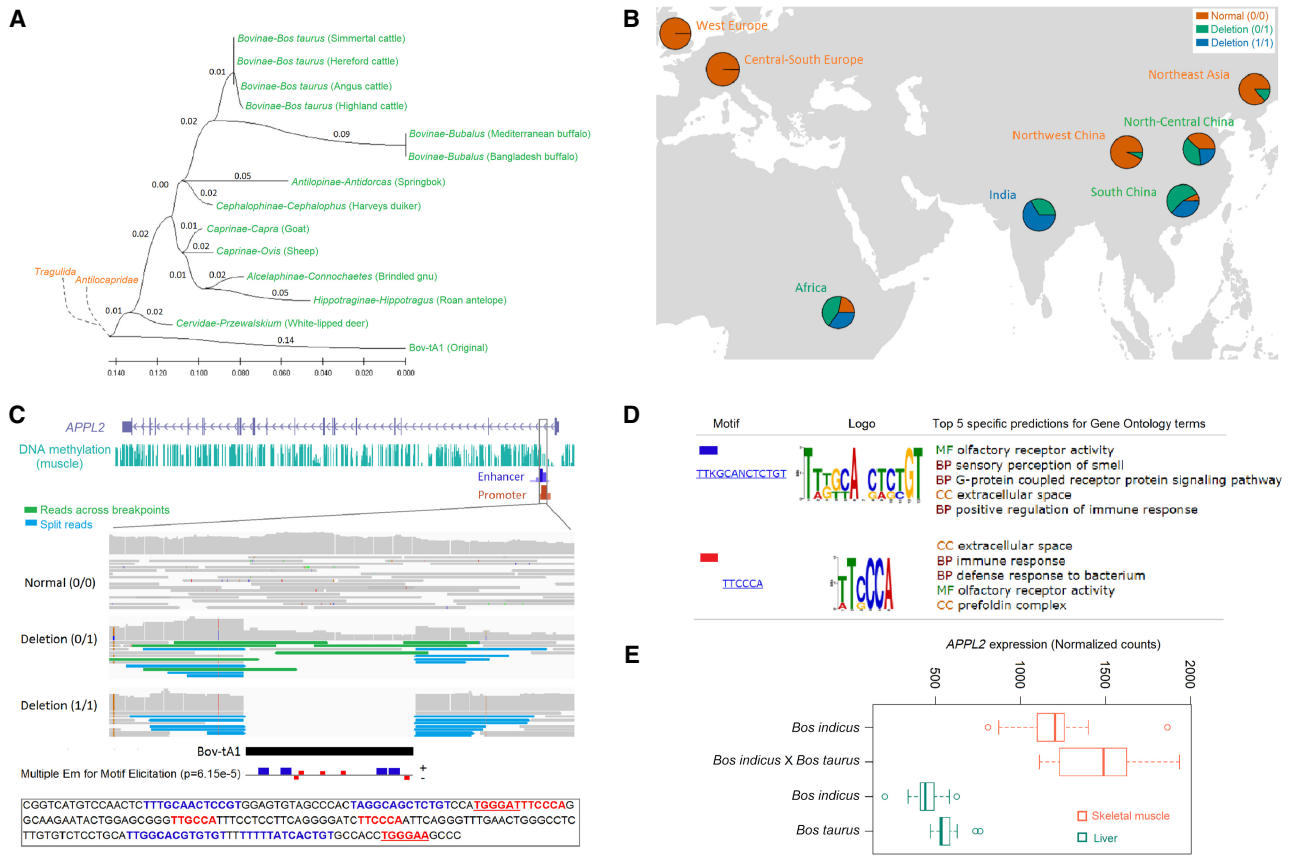

Figure 7. Evolution and function analysis of the Bov-tA1 insertion/deletion in *APPL2*. (*A*) The phylogenic tree of Bov-tA1 in *APPL2* in species within ruminants. (*B*) Pie plot of genotype frequencies for Bov-tA1 deletion in *APPL2* in cattle according to the geographical distribution. (*C*) Bov-tA1 location and sequence analyses. (*D*) GO enrichment analysis of the motifs in Bov-tA1 sequences in *APPL2*. (*E*) Boxplot of *APPL2* expression levels in skeletal muscle and liver using RNA sequencing data. A number of 15 samples were randomly selected for each population.

revealed that the Bov-tA1 indel was adjacent to a TSS-HMR and overlapped with enhancer and promoter regions of *APPL2* (Fig. 7C; Supplemental Table S20). Two kinds of transcription factor binding motifs (TTKGCNACTCTGT and TTCCCA) in the inserted Bov-tA1 sequence were predicted using MEME software, which are related to immune, olfactory, and smell response, corresponding to the *APPL2* known functions (Fig. 7D). RNA sequencing data for muscle and liver tissues indicated that the expression of *APPL2* in *B. taurus* was higher than in *B. indicus* (Fig. 7E). These results suggest that the Bov-tA1 insertion might introduce transcription factor binding motifs that increase the expression of *APPL2*, which might further respond to the different adaptions between *B. taurus* and *B. indicus* in the immune responses, olfactory functions, or muscle development.

Discussion

Quantitative genetics has had a large impact on improved milk and meat production traits in cattle during the past 100 yr. Selective (natural and human-imposed) and nonselective (demographic events and introgression) forces have driven changes within the cattle genome. Their combined effects have created extensive phenotypic diversity and genetic adaptation to local environments across the globe within the modern cattle breeds. Genomics has improved animal health, production, and well-being by shortening the generation interval, identifying genetic markers, and illustrating molecular mechanisms underlying economic traits (Rexroad et al. 2019).

The catalog of SV has been greatly expanded by the development of long-read sequencing technologies, but the cost of this approach has precluded application at the population scale in livestock, and global SV in cattle remains poorly characterized. Short-read sequencing presents challenges in accurate identification of SVs but is currently the most practical way to assess SV diversity, characterize retention and loss of genomic sequence during domestication, and evaluate the effects of selection on SV. The pipeline we developed combines different bioinformatic strategies to improve the accuracy and sensitivity of short-read-based SV detection, using strict filtering criteria for SR, RP, and RD approaches to reduce false-positive rates. This pipeline was applied to 898 cattle samples from 57 breeds representing four cattle subpopulations: European taurine, African taurine, Asian indicine, and African indicine. The pipeline generates individual SV genotypes for each animal, which were used to search for population-specific frequency differentiation that represents a signature of local adaptation. A cattle pangenome was derived by an assembly of unmapped reads to identify nonreference genomic segments.

Missing sequences

The earlier human pangenome studies reported 4- to 40-Mb nonreference sequences (Li et al. 2010; Hehir-Kwa et al. 2016;

Huddleston et al. 2017). A recent human study among African individuals further highlighted the importance of pangenome (Sherman et al. 2019) by reporting ~300 Mb (~10%) of novel DNA sequences missing from the human reference genome. A similar result was published for pigs, with 72.5 Mb of novel sequence added to the pangenome (Tian et al. 2020). Our study obtained similar results in cattle, identifying ~83 Mb or ~3.1% more DNA beyond the current cattle reference ARS-UCD1.2 assembly from our global cattle populations.

Pangenome

The indicine genomes (like Brahman and Nelore) had the highest unique, nonrepetitive sequence contributions, most likely because the cattle reference genome was derived from the European taurine animal. The function of these sequences is not known; however, we detected 16 intact high-confidence protein-coding genes, and the rest appear to be intergenic. Also, the functions of genes in nonreference sequence were similar to those of genes found near *Bos indicus*–specific SV regions (Supplemental Table S8), as previously described (Low et al. 2020). This novel sequence will provide a future foundation for capturing global cattle diversity in population genetics and evolutionary genomics studies.

Improved SV detection to enhance the SV catalog

The problem of lack of diversity has been satisfactorily addressed for SNP and indels because they are easy to map and genotype, thus compensated by the databases of known SNPs like 1000 Bull Genomes Project (Daetwyler et al. 2014). However, this is not the case for SV events because they are challenging to detect or map accurately and are even more difficult to assign states or assemble their sequences.

This study combined four strategies into an integrated SV prediction pipeline and identified more than 3.58 million SV events across 898 cattle genomes; 72.21% of the total SVs we identified overlapped with the SVs reported before (Bickhart et al. 2016; Mielczarek et al. 2018; Kommadath et al. 2019; Jang et al. 2021). These results show that our samples from diverse cattle breeds successfully reflected a large proportion of previously reported SVs as well as novel SVs. It is noted that these public WGS data have various coverage, ranging from 5.08× to 54.42×, with an average coverage of 16×. A simulation in humans showed that at 30× coverage, almost all deletions were detectable, whereas at 15×, 10×, and 8× coverage, 90%, 75%, and 70% of the deletions remained detectable (Yang 2020). Therefore, we estimated that increasing the coverage from 16× to 30× would lead to 10% more deletions detected per sample. Thus, we believe it will not significantly change our major conclusions based on deletions.

It was observed that Asian and African cattle contained more SVs relative to the reference assembly, whereas European cattle had a lower incidence, likely because the UCD-ARS1.2 reference assembly was derived from the Hereford breed of European origin (Supplemental Fig. S14). The differences in SV counts are consistent with population demography, as previously suggested (Upadhyay et al. 2021). More deletions were observed compared with duplications and inversions, agreeing with previous studies (Bickhart et al. 2016; Xu et al. 2016), possibly resulting from bias against detection of duplications in the RP and RD strategies owing to the small insert size or weak signal, respectively. However, it is also possible that this apparent bias is owing to a prevalence of nonallelic homologous recombination, which produces more deletions than duplications (Turner et al. 2008).

SV cluster and SV desert regions were identified among the cattle populations. A significant underrepresentation of SVs in genic regions and regulatory elements was found, implying negative selection against SV in these functional regions. These new insights into cattle genome biology are valuable for understanding the effects SVs have on gene function, with the prospect of identifying important novel alleles that can be used to improve cattle.

Deletion mechanisms and breakpoint signatures

This study presents the results of one of the most extensive SV studies within a livestock species and carefully assesses the breakpoint signature and associated microhomology repeat sequence signatures. MEI complicates SV detection, but true MEI events were accurately predicted by our pipeline (see validation results). At least 17% of the SV calls >50 bp are associated with MEIs, contributing significantly to genomic variation. This leads to our hypothesis that certain MEIs may play an important role in gene regulation based on their distribution near genes, consistent with other studies (Moran et al. 1999; Gilbert et al. 2002). The abundance of SVs with high sequence similarity to known MEIs suggests that some SVs are products of MEI activity. The occurrence of MEIs in the upstream regions of genes, where promoters and regulatory motifs reside, indicates that SVs may be critical agents for gene expression pleiotropy that is often observed in stress-responsive genes. Studies in human genomes have found SV and MEIs associated with aberrant expression of nearby genes (Chiang et al. 2017). Our study revealed that the deletions with microhomology repeat of 6–20 bp showed more consistent boundaries with the full-length MEIs than any other deletions. Three peaks for the deletions with microhomology repeats of 6–20 bp corresponded to bovine-specific MEIs, which further supported that the MEI could cause cattle genome variation. Our observation of deletions with microhomology repeat of 1–4 bp was in line with two possible signatures of DSB repairing mechanisms. These included the deletions with microhomology repeat lengths between 1 and 4 bp were presumably largely caused by MMEJ as previously reported (Ottaviani et al. 2014). In comparison, the short insertions adjoining the deletion breakpoints were likely a possible signature of NHEJ. These results supply new hints for studying the formation mechanisms of cattle genome deletion.

Population genetics of deletions

This study effectively overcame some of the current problems for the SV study, that is, complexity for genotyping and inconsistent breakpoint mapping for different individuals. A previous study indicated that East Asian cattle populations are mainly composed of three distinct ancestries: an earlier East Asian taurine ancestry 3900 yr ago, a later introduced Eurasian taurine ancestry, and a distinct Chinese indicine ancestry that diverged from Indian indicine ~36,600–49,600 yr ago (Chen et al. 2018). Adaptive introgression from indicine cattle into Italian cattle breeds with a white coat color was also reported (Barbato et al. 2020). Our genotyping and evolution results using the deletions generally agreed with the previous SNP-based studies, providing additional validation of the identified SVs. Our population analyses generally divided the animals into taurine and indicine. African indicine showed high levels of shared genetic variation with Asian indicine but not with African taurine. Our SV-based findings confirmed both evolutionary histories of cattle in East Asia and Central Italy, revealing the importance of introgression in adaptation of cattle to new environmental challenges.

Selection signatures

Population differentiation of SV may contribute to the phenotypic variation between populations (Redon et al. 2006). SVs were confirmed to be less likely to occur in the exon regions, consistent with the drastic effects this could have on gene expression and function. The harmful or lethal SVs will have more chances to be selectively eliminated, especially when disrupting CDS with out-of-frame mutations. Genes with the exon overlapped with the SV were found to be highly enriched in the immune function, which is supported by many research results that the immune gene was highly diverse and complex among individuals (Redon et al. 2006; Sudmant et al. 2015a,b; Ebert et al. 2021). Chr 15 and Chr 23 have drawn the attention of the SV studies because they are enriched for olfactory receptor (OR) genes and major histocompatibility complex (MHC) genes. One hundred forty-six regions in different chromosomes that were enriched SVs were detected in the cattle genome. Some of them were caused by the high variable gene families among animals, such as ZNF and Defensins, beta (Liu et al. 2010).

Other interesting genes harboring SVs include *APPL2* and *ATPAF1*. ATPAF1 protein is involved in ATP synthesis, a crucial energy metabolism process that has been intensively selected during the adaptation to geographical locations and climate variations. One SNP-based study of dairy and beef cattle comparisons also associated a selection signature region located in the vicinity of *ATPAF1* (Zhao et al. 2015). This gene is also observed to be highly expressed in the adipose tissue of cattle. A later gene expression profile reported that *ATPAF1* is up-regulated in Holstein compared with Nelore oocytes (Ticianelli et al. 2017). It can be speculated that deletion of this gene might be related to local environment adaptation through breed differences in the energy metabolism of cattle. In summary, our study provided proof of concept for using SVs as important markers in evolutionary studies and breeding programs of local adaptive cattle. Our data also revealed the role of introgression in shaping the landscape of SVs and supplied vital information to promote the understanding of adaptation and phenotype differences between taurine and indicine cattle on the SV level.

Limitations and future directions

The improvements in SV calling accuracy and resolution we present remain limited in sensitivity and specificity owing to the low coverage (<10×) for half of the samples, the reliance on short reads, and the use of a reference genome assembly of a single animal. For example, short reads are ineffective for detecting large duplications and inversions, so caution is needed before their results are fully used. More SVs remain to be identified as it is challenging to map short reads to repetitive/complex regions of the genome. Identifying SVs in such regions can be achieved using long-read sequencing technologies, such as the PacBio or Oxford Nanopore platforms (Logsdon et al. 2020). Using the human 1000 Genomes Project, they estimated that long-read sequencing data provide fivefold higher sensitivity for genetic variants compared with short-read sequencing data (Huddleston et al. 2017). High-quality reference genomes and third-generation sequencing technologies and platforms hold great promises to improve SV detection and studies (Ebert et al. 2021; Logsdon et al. 2021). A single reference genome represents only one set of haplotypes in a single individual, such that nonreference alleles and haplotypes are often not represented. Future studies will need to improve reference genome quality further (Miga et al. 2020; Cheng et al. 2021; Garg et al. 2021) and incorporate as many as possible high-quality genome assemblies into a graph representation of the pangenome to alleviate these problems and enhance the genome-wide SV detection and genotyping in cattle. It became clear that graph pangenomes already have advantages in many applications, including SV calling (Miga and Wang 2021).

In conclusion, an enhanced SV catalog was successfully generated using short reads with improved accuracy and resolution despite difficulties and limitations associated with accurate identification of SV. This SV catalog represents a vital public resource across the diverse cattle breeds and provides new insights for studying the possible SV roles in cattle. SV analysis offers an opportunity to uncover genomic architecture and identify the change of gene content during domestication, breeding, and improvement, helping the selection of future cattle breeds with desired traits. SV-based GWAS will provide a powerful complement to SNP-based GWAS for identifying functional variants of economically or evolutionary important traits.

Methods

Data set collection and generation

The 898 WGS data of 57 cattle breeds were retrieved from the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra/) and our previous studies. The accession ID for each data set is in Supplemental Table S1. Moreover, 10 new WGS data were generated for Enshi cattle from South-Center China. The Enshi cattle, a hybrid breed between *B. taurus* and *B. indicus*, were also used to validate the deletion boundary and genotyping results. Genomic DNA was isolated from the blood of Enshi cattle collected during routine veterinary treatments. These DNA samples were used to construct WGS libraries following the Illumina protocols. Paired-end libraries with a 500-bp insert size were prepared and sequenced using the HiSeq 2500 platform. The PacBio long reads and WGS short reads used for assembling the Angus and Braham genome (accession ID in NCBI BioProject database [https://www.ncbi.nlm.nih.gov/bioproject/]: PRJNA432857) were downloaded to evaluate the accuracy of SV calling. In addition, 1415 RNA sequencing data were downloaded from the NCBI public database (Supplemental Table S7). Within them, 1141 were from 10 out of the 57 breeds, including *B. taurus* originated in Europe (Angus, Hereford, Simmental, Belgian Blue, Holstein), *B. taurus* from Asia (Hanwoo, Kazakh, Xinjiang brown), *B. indicus* (Brahman), and African cattle (Nguni).

Read mapping and SV detection

The adapters and low-quality reads were filtered away using the NGS QC Toolkit (v.2.3.3) with the parameters "-p 8 -l 70 -s 20 -z g." The clean reads were then mapped on the latest cattle reference genome assembly (ARS-UCD1.2, generated from a Hereford cow: L1 Dominette 01449; ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/263/795/GCF_002263 795.1_ARS-UCD1.2) using BWA-MEM (v.0.7.17) with the default parameters (Li 2013). Duplicated reads were removed using sambamba (v.0.6.7) (Tarasov et al. 2015). As described previously (Hu et al. 2020), four SV detection tools were selected, including Pindel (v.0.2.5), LUMPY (v.0.2.13), DELLY (v.0.7.9), and BreakDancer (v.1.4.5), to improve sensitivity and support of SV predictions, especially for large variants. Only SVs with a length between 50 bp and 5 Mb were kept for subsequent analyses. To detect the SV desert and cluster region, the reference genome was binned into nonoverlapping 10-kb windows. A weighted number of supported samples for each

window was calculated. The window with a weighted number of samples over five (~0.5% frequency) was marked as "1" (confident regions with SV); otherwise, it was marked as "0." Then, we applied the Viterbi algorithm in a hidden Markov model in the R package (HMM) to detect regions with continuous "1" (SV cluster regions) or continuous "0" (SV desert regions).

Assembly of missing sequences

The PopIns2 workflow (Krannich et al. 2022) was applied to detect the nonreference sequence of the data set from 898 animals. First, the assemble submodule was used to identify reads without high-quality alignment to the reference genome using default parameters. At the same time, the assemble submodule was used to filter reads with low quality and assemble left reads into contigs for each animal. The FCS-genome software (https://hub.docker.com/r/ncbi/cgr-fcs-genome) was then used to eliminate the potential contaminants in contigs, according to its default recommendations. Moreover, only contigs >1000 bp were considered for further analysis. The merge submodule in the PopIns2 workflow was then used to merge the clean contigs into supercontigs (representative contigs). To confirm the supercontigs without contamination, we ran FCS-genome again and double-checked by BLAST against the NCBI Nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide/). The contigs matching with bacterial and viral sequences (identity >90% and length >25%) were removed from the final representative contigs. Then, all contigs of different individuals were merged by PopIns2 and generated 18,231 supercontigs with 83 Mb in length.

Exon identification

The adapters and low-quality reads were filtered away for the RNA sequencing data using the NGS QC Toolkit (v.2.3.3) with the parameters "-p 8 -l 70 -s 20 -z g." The clean reads for the RNA sequencing data were initially mapped on the latest cattle genome reference assembly (ARS-UCD1.2) using HISAT2 (v.2.1.0) with the default parameters (Kim et al. 2019). Unmapped reads were extracted using SAMtools (v 1.9) (Li et al. 2009) and remapped on the nonredundant missing sequencing data set using HISAT2 (v.2.1.0). Exons were identified using StringTie (v.2.1.4) (Pertea et al. 2015). The cattle EST sequences from the NCBI dbEST database (https://www.ncbi.nlm.nih.gov/genbank/dbest/) were used to validate the novel exons. Novel exon sequences were blasted to the EST sequences using BLAST+ (v.2.11.0) (Altschul et al. 1990). Only the exons >30 bp with >95% identity and 95% coverage by the EST sequences were defined as a perfect match. The EST sequences were mapped against the known cattle mRNA sequences to decide to which genes they belong. The ESTs with >98% sequence identity with known mRNA sequences were defined as successful linkages.

Contig placing on the ARS-UCD1.2 assembly

Three methods were used to place contigs on the cattle reference genome. First, we applied the contigmap and place submodules in PopIns2 to place the final representative contigs on the ARS-UCD1.2 assembly. Generally, read pairs with one read aligning on the reference genome and another read unmapping on the reference genome were used to generate candidate positions for supercontigs. Moreover, split alignment of reads was also used to provide exact positions at a single-base resolution. This strategy will place the supercontigs as three kinds: two-end-placed, one-end-placed, and unplaced. To be more confident about the placement, we removed supercontigs with a distance placed >500 bp for the two ends of supercontigs and one-end-placed supercontigs without anchor information. Second, the unplaced contigs were mapped against assemblies of additional seven cattle breeds—Angus (UOA_Angus_1), Brahman (UOA_Brahman_1), Simmental (ARS_Simm1.0), Jersey (ARS-LIC_NZ_ Jersey), Holstein (ARS-LIC_NZ_Holstein-Friesian_1), Brown Swiss (Brown_Swiss_cow), and Highland cattle (ARS_UNL_Btau-highland_paternal_1.0_alt)—to take advantage of the synteny sequence similarity with the ARS-UCD1.2. The contigs with 90% identity and 80% coverage were recognized as successful matches. All contigs that mapped to multiple locations were removed to avoid mismatches caused by multiple alignments. The upstream and downstream 500-kb sequences of the matched positions were defined as containing potential synteny sequences between ARS-UCD1.2 and other cattle breed assemblies. The FASTA sequences were isolated using the BEDTools getfasta option and blasted to the ARS-UCD1.2 assembly sequences. Only the unique alignment >5000 bp, 95% identity, and the placed distance between upstream and downstream 500-kb sequences <2000 bp were considered as a final placement. Third, linkages were created between contigs and known mRNA sequences using EST. The contig location was defined by the mRNA-coding gene in the ARS-UCD1.2 assembly.

Pipelines for deletion breakpoint identification and genotyping

A pipeline was designed to identify the exact deletion boundaries based on split reads within the target region to overcome differences among the deletion boundaries identified by four different software (Fig. 3A). First, the target regions were defined as the largest intervals of the boundaries reported by different SV tools. The reads located in each target region were evaluated and refiltered using the following thresholds: mapped reads quality of 30 or more, only allow up to one mutation by considering the diversity of different cattle individuals and breeds, split reads separated into two parts that mapped to two candidate regions with at least 30 bp. Deletions were separated into three categories according to the mapped reads: breakpoints with microhomology repeats, breakpoints with short insertions, and those that perfectly support the deletion breakpoints by split reads. Thus, the exact boundaries of the deletion were defined. The number of reads adjacent to or across the deletion boundaries was first evaluated to genotype the deletion for each animal. Only the deletion boundaries supported by at least five reads were considered for genotyping. The deletions with both normal reads and split reads were genotyped as heterozygous (0/1). Otherwise, the deletions with only normal or split reads were genotyped as normal (0/0) or total deletion (1/1), respectively.

Validation of the deletion breakpoints and genotypes

Eight deletion regions were randomly selected and validated using specific PCR for the 10 Enshi cattle with WGS data. In detail, primers (Supplemental Table S21) were designed and searched throughout the ARS-UCD1.2 assembly sequences to confirm the unique amplification of the target region. The PCR amplification was performed in 50 μL reaction volume using the manufacturer's Taq DNA polymerase protocol (Taq PCR master mix kit, Qiagen). The genomic DNA was PCR-amplified on a Bio-Rad MyIQ thermocycler. The PCR-amplified products were run in 1.5% agarose gel, and the target bands were cut to perform Sanger sequencing. The final sequences were blasted against the ARS-UCD1.2 assembly and checked for the deletion boundaries by visualization.

One animal of Holstein and its parents were sequenced. The deletions of the three animals were identified and genotyped according to our pipeline. Totally, 2807 deletions were detected, and 1747 deletions were successfully genotyped in all three

animals. Genotypes of each deletion were compared. If the genotype could be traced back to sire and dam according to the Mendelian inheritance, the deletion was recognized as a successful genotyping.

Functional elements, QTL annotation, and gene enrichment analyses

Genome annotation files, including genic, lncRNA, pseudogene, intron, exon, and other features, were downloaded from the NCBI ftp website (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/263/795/GCF_002263795.1_ARS-UCD1.2/). DNA methylation functional regions, including DMR, TSS-HMR, and tissue-specific HMR, were generated from our previous study using 29 whole-genome bisulfite sequencing (WGBS) data sets for 16 tissues and 47 corresponding RNA-seq data sets (Zhou et al. 2020). The 14 chromatin state predictions were downloaded from http://farm.cse.ucdavis.edu/~ckern/Nature_Communications_2020/, which included regulatory elements for eight different cattle tissues (Kern et al. 2021). Cattle QTL information was downloaded from the animal QTL database (https://www.animalgenome.org/cgi-bin/QTLdb/index). We evaluated fold enrichment of deletion regions in QTLs for each animal. The 2000 bootstraps were used to estimate the 95% confidence intervals to exclude the enrichment by chance. Gene functional annotation analyses were performed using the online DAVID software (https://david.ncifcrf.gov/). The Fisher's exact test was conducted to measure gene enrichment in annotation terms (0.05).

SNP detection and LD analysis

The Genome Analysis Toolkit (GATK, v.4.1.9.0) software was used to detect SNPs for 47 Holstein cattle (McKenna et al. 2010). The known SNPs from the 1000 Bull Genomes Project was used as reference to call SNP (Hayes and Daetwyler 2019). The GVCF files was first created individually using HaplotypeCaller in the GATK. All GVCF files were merged using "CombineGVCFs" in GATK. The following thresholds were used to avoid possible false-positive calls by VariantFiltration: --cluster-size 3 --cluster-window-size 10 --filter-expression QUAL < 30 || QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0.

The 47 Holstein cattle were classified according to their genotyping information of deletions. For each deletion, only the Holstein cattle genotyped as homozygote (normal two copies or total deletion) were considered for further LD of SNPs analysis, and the population for each genotype included at least five animals. The SNPs located within 50 kb upstream of and downstream from the deletions were used to calculate the R^2 of the linkage disequilibrium using LDBlockShow v.1.40 (Dong et al. 2021).

Population structure and select signature analysis

Multiple steps were applied to filter away deletion or animal called missing values to obtain deletion loci information and include all represented breeds; 71,240 genotyped deletions for 339 animals were obtained as input. To prepare samples for population analyses, two rounds of filtering were performed. First, samples and deletions with low call rates were removed. Then the call rates for left samples and deletions were recalculated. Only samples and deletions with a call rate >90% were kept for further analysis. Finally, 14,942 deletions in 339 animals of 50 breeds were obtained with their average call rates ~95%. The PCA was performed using -pac option of the PLINK software v.1.9 (Purcell et al. 2007). ADMIXTURE was run for each possible group (K = 2 to 13) with 200 bootstraps using Admixture v.1.3.0 (Alexander et al. 2009).

The fixation index (F_{ST}) was used to identify the deletion under selection between *B. taurus* and *B. indicus*. The animals of both *B. taurus* and *B. indicus* were selected according to the PCA and admixture analysis results. The *B. taurus* population includes the European taurine, Eurasian taurine, and East Asian taurine. To increase the representative and the number of the animals in the *B. indicus* population, the animals with more than 3/4 ancestry were also included in the F_{ST} analysis. The locus-specific divergence in allele frequencies for each cattle population/breed was measured by performing the *di* statistics based on unbiased estimates of pairwise F_{ST}. The *di* value was calculated according to the following formulation:

$$di = \sum_{j \neq i} \frac{F_{ST}^{ij} - E[F_{ST}^{ij}]}{sd[F_{ST}^{ij}]},$$

where the $E[F_{ST}^{ij}]$ and $sd[F_{ST}^{ij}]$ represented the expected value and SD of F_{ST} between groups *i* and *j*. The deletions with the top 1% of the F_{ST} and *di* values were defined as the candidates under selection.

Evolution and function analyses of the Bov-tA1 insertion in *APPL2*

The genome assemblies were downloaded for Cervidae–*Przewalskium albirostris* (white-lipped deer), Antilocapridae–*Antilocapra americana* (pronghorn), Tragulidae–*Tragulus kanchil* (even-toed ungulates), Bovinae–*Bos_taurus* (Angus_cattle), Bovinae–*Bubalus* (Mediterranean_buffalo), Caprinae–*Capra* (goat), Bovinae–*Bos_taurus* (Highland_cattle), Caprinae–*Ovis* (sheep), Bovinae–*Bos_taurus* (Simmental_cattle), Antilopinae–*Antidorcas* (springbok), Alcelaphinae–*Connochaetes* (brindled gnu), Cephalophinae–*Cephalophus* (Harvey's duiker), Hippotraginae–*Hippotragus* (roan antelope), Cervidae–*Przewalskium* (white-lipped_deer), Bovinae–*Bubalus* (Bangladesh_buffalo), and Bovinae–*Bos_taurus* (Hereford_cattle) from the NCBI Assembly database (https://www.ncbi.nlm.nih.gov/assembly/). The sequence of Bov-tA1 with upstream and downstream 1000 bp in the *APPL2* gene was used to blast against each assembly. The assembly covering the whole length of the query sequence with >90% identity was considered as containing the Bov-tA1 insertion, which was further confirmed by visualization. The Bov-tA1 sequences were isolated from each assembly and aligned with each other using ClustalW. The aligned Bov-tA1 sequences from each assembly were used to construct the maximum likelihood tree using MEGA software (Tamura et al. 2021). The MEME online tool (https://meme-suite.org/meme/) was used to search and enrich related Gene Ontology terms of motifs in the Bov-tA1 sequence. Sixty RNA sequencing data were randomly downloaded for skeletal muscle and liver in *B. taurus*, *B. indicus*, and the hybrid cattle to compare the expression of *APPL2* with or without the Bov-tA1 sequence between them (Supplemental Table S7). RNA sequencing data were processed following the HISAT2 (v.2.1.0) and StringTie (v.2.1.4) pipeline. DESeq2 (v.1.30.1) was used to compare the gene expression of different populations through normalization by considering different sequencing depths and libraries (Love et al. 2014).

Data access

The pipelines for deletion breakpoint identification and genotyping are available at GitHub (https://github.com/yangzhou-bio-lib/SV-information) and Zenodo (https://sandbox.zenodo.org/record/1080619). SV regions and their normalized frequencies have been deposited as customer track files for the UCSC Genome Browser available at GitHub (https://github.com/yangzhou-bio-lib/SV-information/tree/main/SV-information).
The cattle pangenome contigs and the detailed information for

contigs placed in the reference genome are available from GitHub (https://github.com/yangzhou-bio-lib/cattle_pangenome_storage) and Zenodo (https://sandbox.zenodo.org/record/1084439#.YvDXCy1h2v4). The newly generated raw WGS data from this study were submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA6917 41 and PRJNA762638. All relevant codes are also provided as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (31902148) and the Natural Science Foundation of Hubei Province of China (2021CFB463). G.E.L. is supported in part by Agriculture and Food Research Initiative (AFRI) grant numbers 2019-67015-29321 and 2021-67015-33409 from the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA). This research used resources provided by the SCINet project of the USDA Agricultural Research Service (ARS) project number 0500-00093-001-00-D.

Author contributions: All authors have read and approved the manuscript. Y.Z., L.Y., and G.E.L. conceived and designed the experiments. Y.Z., L.Y., X.T.H., Y.H., J.Z.H., and L.W.P. performed in silico prediction and computational analyses. F.L. and H.X. performed PCR confirmation. A.Z.G., S.J.Z., C.B., D.M.B., B.D.R., C.P.V.T., T.P.L.S., and L.G.Y. collected samples and generated genome sequencing data. Y.Z., L.Y., and G.E.L. wrote the paper.

References

Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, Layer RM, Neale BM, Salerno WJ, Reeves C, et al. 2020. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583:** 83–89. doi:10.1038/s41586-020-2371-0

Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, Nicholas TJ, Neff MW. 2010. Tracking footprints of artificial selection in the dog genome. *Proc Natl Acad Sci* **107:** 1160–1165. doi:10.1073/pnas.0909918107

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19:** 1655–1664. doi:10.1101/gr.094052.109

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12:** 363–376. doi:10.1038/nrg2958

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410. doi:10.1016/S0022-2836(05)80360-2

Barbato M, Hailer F, Upadhyay M, Del Corvo M, Colli L, Negrini R, Kim E-S, Crooijmans RPMA, Sonstegard T, Ajmone-Marsan P. 2020. Adaptive introgression from indicine cattle into white cattle breeds from central Italy. *Sci Rep* **10:** 1279. doi:10.1038/s41598-020-57880-4

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. 2020. Plant pan-genomes are the new reference. *Nat Plants* **6:** 914–920. doi:10.1038/s41477-020-0733-0

Bickhart DM, Liu GE. 2014. The challenges and importance of structural variation detection in livestock. *Front Genet* **5:** 37. doi:10.3389/fgene.2014.00037

Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, Song J, Schnabel RD, Ventura M, Taylor JF, et al. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res* **22:** 778–790. doi:10.1101/gr.133967.111

Bickhart DM, Xu L, Hutchison JL, Cole JB, Null DJ, Schroeder SG, Song J, Garcia JF, Sonstegard TS, Van Tassell CP, et al. 2016. Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. *DNA Res* **23:** 253–262. doi:10.1093/dnares/dsw013

Black SJ, Ozdemir AY, Kashkina E, Kent T, Rusanov T, Ristic D, Shin Y, Suma A, Hoang T, Chandramouly G, et al. 2019. Molecular basis of microhomology-mediated end-joining by purified full-length Polθ. *Nat Commun* **10:** 4423. doi:10.1038/s41467-019-12272-9

Bolormaa S, Pryce JE, Kemper KE, Hayes BJ, Zhang Y, Tier B, Barendse W, Reverter A, Goddard ME. 2013. Detection of quantitative trait loci in *Bos indicus* and *Bos taurus* cattle using genome-wide association studies. *Genet Sel Evol* **45:** 43. doi:10.1186/1297-9686-45-43

The Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, Adelson DL, Eichler EE, Elnitski L, et al. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324:** 522–528. doi:10.1126/science.1169588

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6:** 677–681. doi:10.1038/nmeth.1363

Chen L, Chamberlain AJ, Reich CM, Daetwyler HD, Hayes BJ. 2017. Detection and validation of structural variations in bovine whole-genome sequence data. *Genet Sel Evol* **49:** 13. doi:10.1186/s12711-017-0286-5

Chen N, Cai Y, Chen Q, Li R, Wang K, Huang Y, Hu S, Huang S, Zhang H, Zheng Z, et al. 2018. Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nat Commun* **9:** 2337. doi:10.1038/s41467-018-04737-0

Chen N, Fu W, Zhao J, Shen J, Chen Q, Zheng Z, Chen H, Sonstegard TS, Lei C, Jiang Y. 2020. BGVD: an integrated database for bovine sequencing variations and selective signatures. *Genomics Proteomics Bioinformatics* **18:** 186–193. doi:10.1016/j.gpb.2019.03.007

Cheng KK, Zhu W, Chen B, Wang Y, Wu D, Sweeney G, Wang B, Lam KS, Xu A. 2014. The adaptor protein APPL2 inhibits insulin-stimulated glucose uptake by interacting with TBC1D1 in skeletal muscle. *Diabetes* **63:** 3748–3758. doi:10.2337/db14-0337

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18:** 170–175. doi:10.1038/s41592-020-01056-5

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Montgomery SB, et al. 2017. The impact of structural variation on human gene expression. *Nat Genet* **49:** 692–699. doi:10.1038/ng.3834

Cicconardi F, Chillemi G, Tramontano A, Marchitelli C, Valentini A, Ajmone-Marsan P, Nardone A. 2013. Massive screening of copy number population-scale variation in *Bos taurus* genome. *BMC Genomics* **14:** 124. doi:10.1186/1471-2164-14-124

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* **581:** 444–451. doi:10.1038/s41586-020-2287-8

Crysnanto D, Pausch H. 2020. Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. *Genome Biol* **21:** 184. doi:10.1186/s13059-020-02105-0

Crysnanto D, Leonard AS, Fang Z-H, Pausch H. 2021. Novel functional sequences uncovered through a bovine multiassembly graph. *Proc Natl Acad Sci* **118:** e2101056118. doi:10.1073/pnas.2101056118

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, Liao X, Djari A, Rodriguez SC, Grohs C, et al. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* **46:** 858–865. doi:10.1038/ng.3034

Dong SS, He WM, Ji JJ, Zhang C, Guo Y, Yang TL. 2021. LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Brief Bioinform* **22:** bbaa227. doi:10.1093/bib/bbaa227

Durkin K, Coppieters W, Drögemüller C, Ahariz N, Cambisano N, Druet T, Fasquelle C, Haile A, Horin P, Huang L, et al. 2012. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature* **482:** 81–84. doi:10.1038/nature10757

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372:** eabf7117. doi:10.1126/science.abf7117

Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, Carter NP, Church DM, Felsenfeld A, Guyer M, Lee C, et al. 2007. Completing the map of human genetic variation. *Nature* **447:** 161–165. doi:10.1038/nature05761

Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, Mac S, Peluso P, Hatas E, Ghurye J, et al. 2021. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat Biotechnol* **39:** 309–312. doi:10.1038/s41587-020-0711-0

Gilbert N, Lutz-Prigge S, Moran JV. 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110:** 315–325. doi:10.1016/S0092-8674(02)00828-0

Gilbert M, Nicolas G, Cinardi G, Van Boeckel TP, Vanwambeke SO, Wint GRW, Robinson TP. 2018. Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. *Sci Data* **5:** 180227. doi:10.1038/sdata.2018.227

Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA. 2015. Large multiallelic copy number variations in humans. *Nat Genet* **47:** 296–303. doi:10.1038/ng.3200

Hayes BJ, Daetwyler HD. 2019. 1000 Bull Genomes Project to map simple and complex genetic traits in cattle: applications and outcomes. *Annu Rev Anim Biosci* **7:** 89–102. doi:10.1146/annurev-animal-020518-115024

Heaton MP, Smith TPL, Bickhart DM, Vander Ley BL, Kuehn LA, Oppenheimer J, Shafer WR, Schuetze FT, Stroud B, McClure JC, et al. 2021. A reference genome assembly of Simmental cattle, *Bos taurus taurus*. *J Hered* **112:** 184–191. doi:10.1093/jhered/esab002

Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R, et al. 2016. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun* **7:** 12989. doi:10.1038/ncomms12989

Ho SS, Urban AE, Mills RE. 2020. Structural variation in the sequencing era. *Nat Rev Genet* **21:** 171–189. doi:10.1038/s41576-019-0180-9

Hu Z-L, Park CA, Reecy JM. 2018. Building a livestock genetic and genomic information knowledgebase through integrative developments of animal QTLdb and CorrDB. *Nucleic Acids Res* **47:** D701–D710. doi:10.1093/nar/gky1084

Hu Y, Xia H, Li M, Xu C, Ye X, Su R, Zhang M, Nash O, Sonstegard TS, Yang L, et al. 2020. Comparative analyses of copy number variations between *Bos taurus* and *Bos indicus*. *BMC Genomics* **21:** 682. doi:10.1186/s12864-020-07097-6

Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al. 2017. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* **27:** 677–685. doi:10.1101/gr.214007.116

Jang J, Terefe E, Kim K, Lee YH, Belay G, Tijjani A, Han J-L, Hanotte O, Kim H. 2021. Population differentiated copy number variation of *Bos taurus*, *Bos indicus* and their African hybrids. *BMC Genomics* **22:** 531. doi:10.1186/s12864-021-07808-7

Kadri NK, Sahana G, Charlier C, Iso-Touru T, Guldbrandtsen B, Karim L, Nielsen US, Panitz F, Aamand GP, Schulman N, et al. 2014. A 660-kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. *PLoS Genet* **10:** e1004049. doi:10.1371/journal.pgen.1004049

Kern C, Wang Y, Xu X, Pan Z, Halstead M, Chanthavixay G, Saelao P, Waters S, Xiang R, Chamberlain A, et al. 2021. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nat Commun* **12:** 1821. doi:10.1038/s41467-021-22100-8

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37:** 907–915. doi:10.1038/s41587-019-0201-4

Kim K, Kwon T, Dessie T, Yoo D, Mwai OA, Jang J, Sung S, Lee S, Salim B, Jung J, et al. 2020. The mosaic genome of indigenous African cattle as a unique genetic resource for African pastoralism. *Nat Genet* **52:** 1099–1110. doi:10.1038/s41588-020-0694-2

Kommadath A, Grant JR, Krivushin K, Butty AM, Baes CF, Carthy TR, Berry DP, Stothard P. 2019. A large interactive visual database of copy number variants discovered in taurine cattle. *Gigascience* **8:** giz073. doi:10.1093/gigascience/giz073

Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. 2018. *De novo* assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol* **36:** 1174–1182. doi:10.1038/nbt.4277

Krannich T, White WTJ, Niehus S, Holley G, Halldórsson BV, Kehr B. 2022. Population-scale detection of non-reference sequence variants using colored de Bruijn graphs. *Bioinformatics* **38:** 604–611. doi:10.1093/bioinformatics/btab749

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15:** R84. doi:10.1186/gb-2014-15-6-r84

Lee YL, Takeda H, Costa Monteiro Moreira G, Karim L, Mullaart E, Coppieters W, Appeltant R, Veerkamp RF, Groenen MAM, Georges M, et al. 2021. A 12 kb multi-allelic copy number variation encompassing a GC gene enhancer is associated with mastitis resistance in dairy cattle. *PLoS Genet* **17:** e1009331. doi:10.1371/journal.pgen.1009331

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN].

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352

Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J, et al. 2010. Building the sequence map of the human pan-genome. *Nat Biotechnol* **28:** 57–63. doi:10.1038/nbt.1596

Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, Mitra A, Alexander LJ, Coutinho LL, Dell'aquila ME, et al. 2010. Analysis of copy number variations among diverse cattle breeds. *Genome Res* **20:** 693–703. doi:10.1101/gr.105403.110

Loftus RT, MacHugh DE, Bradley DG, Sharp PM, Cunningham P. 1994. Evidence for two independent domestications of cattle. *Proc Natl Acad Sci* **91:** 2757–2761. doi:10.1073/pnas.91.7.2757

Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21:** 597–614. doi:10.1038/s41576-020-0236-x

Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC, Rhie A, et al. 2021. The structure, function and evolution of a complete human chromosome 8. *Nature* **593:** 101–107. doi:10.1038/s41586-021-03420-7

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550. doi:10.1186/s13059-014-0550-8

Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, Thibaud-Nissen F, Murphy TD, Young R, Lefevre L, et al. 2019. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat Commun* **10:** 260. doi:10.1038/s41467-018-08260-0

Low WY, Tearle R, Liu R, Koren S, Rhie A, Bickhart DM, Rosen BD, Kronenberg ZN, Kingan SB, Tseng E, et al. 2020. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nat Commun* **11:** 2071. doi:10.1038/s41467-020-15848-y

MacHugh DE, Shriver MD, Loftus RT, Cunningham P, Bradley DG. 1997. Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics* **146:** 1071–1086. doi:10.1093/genetics/146.3.1071

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS, et al. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* **4:** e5350. doi:10.1371/journal.pone.0005350

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20:** 1297–1303. doi:10.1101/gr.107524.110

Miaczynska M, Christoforidis S, Giner A, Shevchenko A, Uttenweiler-Joseph S, Habermann B, Wilm M, Parton RG, Zerial M. 2004. APPL proteins link Rab5 to nuclear signal transduction via an endosomal compartment. *Cell* **116:** 445–456. doi:10.1016/S0092-8674(04)00117-5

Mielczarek M, Fr?szczak M, Nicolazzi E, Williams JL, Szyda J. 2018. Landscape of copy number variations in *Bos taurus*: individual- and inter-breed variability. *BMC Genomics* **19:** 410. doi:10.1186/s12864-018-4815-6

Miga KH, Wang T. 2021. The need for a human pangenome reference sequence. *Annu Rev Genomics Hum Genet* **22:** 81–102. doi:10.1146/annurev-genom-120120-081921

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585:** 79–84. doi:10.1038/s41586-020-2547-7

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470:** 59–65. doi:10.1038/nature09708

Moran JV, DeBerardinis RJ, Kazazian HH Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* **283:** 1530–1534. doi:10.1126/science.283.5407.1530

Oppenheimer J, Rosen BD, Heaton MP, Vander Ley BL, Shafer WR, Schuetze FT, Stroud B, Kuehn LA, McClure JC, Barfield JP, et al. 2021. A reference genome assembly of American bison, *Bison bison bison*. *J Hered* **112:** 174–183. doi:10.1093/jhered/esab003

Ottaviani D, LeCain M, Sheer D. 2014. The role of microhomology in genomic structural variation. *Trends Genet* **30:** 85–94. doi:10.1016/j.tig.2014.01.001

Papachristou D, Koutsouli P, Laliotis GP, Kunz E, Upadhyay M, Seichter D, Russ I, Gjoko B, Kostaras N, Bizelis I, et al. 2020. Genomic diversity and population structure of the indigenous Greek and Cypriot cattle populations. *Genet Sel Evol* **52:** 43. doi:10.1186/s12711-020-00560-8

Pérez O'Brien AM, Mészáros G, Utsunomiya YT, Sonstegard TS, Garcia JF, Van Tassell CP, Carvalheiro R, da Silva MVB, Sölkner J. 2014. Linkage disequilibrium levels in *Bos indicus* and *Bos taurus* cattle using medium

and high density SNP chip data and different minor allele frequency distributions. *Livest Sci* **166:** 121–132. doi:10.1016/j.livsci.2014.05.007
Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33:** 290–295. doi:10.1038/nbt.3122
Pitt D, Sevane N, Nicolazzi EL, MacHugh DE, Park SDE, Colli L, Martinez R, Bruford MW, Orozco-terWengel P. 2019. Domestication of cattle: two or three events? *Evol Appl* **12:** 123–136. doi:10.1111/eva.12674
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559–575. doi:10.1086/519795
Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28:** i333–i339. doi:10.1093/bioinformatics/bts378
Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444–454. doi:10.1038/nature05329
Rexroad C, Vallet J, Matukumalli LK, Reecy J, Bickhart D, Blackburn H, Boggess M, Cheng H, Clutter A, Cockett N, et al. 2019. Genome to phenome: improving animal health, production, and well-being: a new USDA blueprint for animal genome research 2018–2027. *Front Genet* **10:** 327. doi:10.3389/fgene.2019.00327
Rice ES, Koren S, Rhie A, Heaton MP, Kalbfleisch TS, Hardy T, Hackett PH, Bickhart DM, Rosen BD, Ley BV, et al. 2020. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *Gigascience* **9:** giaa029. doi:10.1093/gigascience/giaa029
Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C, et al. 2020. *De novo* assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* **9:** giaa021. doi:10.1093/gigascience/giaa021
Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, Feuk L. 2007. Challenges and standards in integrating surveys of structural variation. *Nat Genet* **39:** S7–S15. doi:10.1038/ng2093
Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15:** 461–468. doi:10.1038/s41592-018-0001-7
Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S, Vergara C, Ortega VE, et al. 2019. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet* **51:** 30–35. doi:10.1038/s41588-018-0273-y
Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. 2015a. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349:** aab3761. doi:10.1126/science.aab3761
Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang FM, et al. 2015b. An integrated map of structural variation in 2,504 human genomes. *Nature* **526:** 75–81. doi:10.1038/nature15394
Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis version 11. *Mol Biol Evol* **38:** 3022–3027. doi:10.1093/molbev/msab120
Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31:** 2032–2034. doi:10.1093/bioinformatics/btv098
Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci* **102:** 13950–13955. doi:10.1073/pnas.0506758102
Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14:** 178–192. doi:10.1093/bib/bbs017
Tian X, Li R, Fu W, Li Y, Wang X, Li M, Du D, Tang Q, Cai Y, Long Y, et al. 2020. Building a sequence map of the pig pan-genome from multiple *de novo* assemblies and Hi-C data. *Sci China Life Sci* **63:** 750–763. doi:10.1007/s11427-019-9551-7
Ticianelli JS, Emanuelli IP, Satrapa RA, Castilho ACS, Loureiro B, Sudano MJ, Fontes PK, Pinto RFP, Razza EM, Surjus RS, et al. 2017. Gene expression profile in heat-shocked Holstein and Nelore oocytes and cumulus cells. *Reprod Fertil Dev* **29:** 1787–1802. doi:10.1071/RD16154
Turner DJ, Miretti M, Rajan D, Fiegler H, Carter NP, Blayney ML, Beck S, Hurles ME. 2008. Germline rates of *de novo* meiotic deletions and duplications causing several genomic disorders. *Nat Genet* **40:** 90–95. doi:10.1038/ng.2007.40
Upadhyay M, Derks MFL, Andersson G, Medugorac I, Groenen MAM, Crooijmans RPMA. 2021. Introgression contributes to distribution of structural variations in cattle. *Genomics* **113:** 3092–3102. doi:10.1016/j.ygeno.2021.07.005
van de Lagemaat LN, Gagnier L, Medstrand P, Mager DL. 2005. Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. *Genome Res* **15:** 1243–1249. doi:10.1101/gr.3910705
Verdugo MP, Mullin VE, Scheu A, Mattiangeli V, Daly KG, Maisano Delser P, Hare AJ, Burger J, Collins MJ, Kehati R, et al. 2019. Ancient cattle genomics, origins, and rapid turnover in the fertile crescent. *Science* **365:** 173–176. doi:10.1126/science.aav1002
Whipple G, Koohmaraie M, Dikeman ME, Crouse JD, Hunt MC, Klemm RD. 1990. Evaluation of attributes that affect longissimus muscle tenderness in *Bos taurus* and *Bos indicus* cattle. *J Anim Sci* **68:** 2716–2728. doi:10.2527/1990.6892716x
Xu L, Hou Y, Bickhart DM, Zhou Y, Hay e, Song J, Sonstegard TS, Van Tassell CP, Liu GE. 2016. Population-genetic properties of differentiated copy number variations in cattle. *Sci Rep* **6:** 23161. doi:10.1038/srep23161
Yang L. 2020. A practical guide for structural variation detection in the human genome. *Curr Protoc Hum Genet* **107:** e103. doi:10.1002/cphg.103
Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25:** 2865–2871. doi:10.1093/bioinformatics/btp394
Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10:** 451–481. doi:10.1146/annurev.genom.9.081307.164217
Zhao F, McParland S, Kearney F, Du L, Berry DP. 2015. Detection of selection signatures in dairy and beef cattle using high-density genomic information. *Genet Sel Evol* **47:** 49. doi:10.1186/s12711-015-0127-3
Zhou Y, Liu S, Hu Y, Fang L, Gao Y, Xia H, Schroeder SG, Rosen BD, Connor EE, Li C-J, et al. 2020. Comparative whole genome DNA methylation profiling across cattle tissues reveals global and tissue-specific methylation patterns. *BMC Biol* **18:** 85. doi:10.1186/s12915-020-00793-5
Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, et al. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol* **10:** R42. doi:10.1186/gb-2009-10-4-r42

Received January 2, 2022; accepted in revised form July 21, 2022.

Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history

Yang Zhou, Lv Yang, Xiaotao Han, et al.

Genome Res. 2022 32: 1585-1601 originally published online August 17, 2022
Access the most recent version at doi:10.1101/gr.276550.122