Supplemental Material

A Chinese indicine pangenome reveals a wealth of novel structural variants introgressed from other *Bos* species

Xuelei Dai[1#], Peipei Bian[1#], Dexiang Hu[1#], Funong Luo[1#], Yongzhen Huang[1#], Shaohua Jiao[1], Xihong Wang[1], Mian Gong[1], Ran Li[1], Yudong Cai[1], Jiayue Wen[1], Qimeng Yang[1], Weidong Deng[2], Hojjat Asadollahpour Nanaei[1,3], Yu Wang[1], Fei Wang[1], Zijing Zhang[4], Benjamin D. Rosen[5], Rasmus Heller[6*], Yu Jiang[1,7*]

[#]These authors contributed equally to this article.

[*]Correspondence:

Rasmus Heller, E-mail: rheller@bio.ku.dk

Yu Jiang, E-mail: yu.jiang@nwafu.edu.cn

This PDF file includes:

Supplemental Methods

Sample collection

Fresh blood samples were collected from 10 representatives of Chinese indicine cattle across different regions in southern China, which were used for Illumina and PacBio HiFi sequencing (Supplemental Table S20). We also collected fresh blood of 19 PiNan cattle [hybrids of Piedmontese (male) x Nanyang cattle (female)] from Nanyang County, Henan Province, China, which were leveraged for whole-genome and transcriptome resequencing by the Illumina platform (Supplemental Table S20). In addition, we downloaded 394 representative whole-genome resequencing data from species of the *Bos* genus for population analysis (including 2 African buffalo, 4 American bison, 5 European bison, 3 yak, 10 gayal, 4 gaur, 8 banteng, 2 ancient kouprey, 110 European taurine, 43 African taurine, 47 Asian taurine, 78 African indicine, 43 India-Pakistan indicine, and 35 Chinese indicine). Summary information and overview of above samples, including breeds, country of origin, geographical location, sequencing depth and contributors, are detailed in Supplemental Table S15, and Supplemental Table S16.

Short reads sequencing

In addition to the above, we extracted DNA from the same 10 Chinese indicine cattle and 19 PiNan cattle as above using a standard cetyltrimethylammonium bromide (CTAB) extraction protocol. This DNA was used to construct the sequence libraries according to the Illumina library preparation protocol, and were sequenced on the

Illumina HiSeq platform to generate 150 bp paired-end reads (Supplemental Table S1, Supplemental Table S20). The RNA of 19 PiNan cattle was extracted with a Qiagen RNeasy Mini Kit for transcriptome sequencing, and a total of 19 paired-end libraries were prepared and sequenced on an Illumina HiSeq platform with a 150 bp paired-end reads (Supplemental Table S20).

PacBio HiFi sequencing

We extracted high-quality gDNA from the fresh blood of the above 10 Chinese indicine to construct libraries using the SMRTbell Express Template Prep Kit 2.0 (PacBio) and fractionated on the SageELF (Sage Science, Beverly, MA) into narrow library fractions. Before sequencing, library fractions were bound to polymerase with the Sequel II Binding Kit 2.0 (PacBio) and then sequenced with the PacBio Sequel II Sequencing Kit 2.0 and 8M SMRT Cells on the Sequel II (PacBio) with 30-hour movie times for each sample. The HiFi reads was generated from raw data using the CCS algorithm (version 6.0.0, parameters: --minPasses 3 -- minPredictedAccuracy 0.99 –maxLength 21000). Generating CCS reads does not include or require alignment against a reference sequence, but does require at least two full-pass subreads from the insert (Supplemental Table S1).

***De novo* genome assembly and quality assessment**

Firstly, we used a fast haplotype-resolved *de novo* assembler, hifiasm (version 0.13-r308; with default parameters) (Cheng et al. 2021) to assemble the PacBio HiFi reads, which produces one primary contig (representing a mosaic of homologous haplotypes) and two sets of partially phased contigs (representing the two haplotypes in a diploid genome, though with occasional switch errors) for each sample. Next, we applied the reference-guided software RagTag (v2.0.1) (Alonge et al. 2022) to scaffold the contigs to chromosome level. The completeness of the assemblies was assessed with BUSCO (version 5.4.5), using the metaeuk backend (6.a5d39d9) and cetartiodactyla_odb10 database. The assembly base quality values (QV) were calculated with merqury (version 1.3) (Rhie et al. 2020), with *k*-mer databases constructed from short reads and HiFi reads using meryl (version 1.4) (https://github.com/marbl/meryl).

Although the individuals we sequenced and the reference genome are both female, theoretically the X Chromosome assembly should be better. However, as shown in Fig. 1e, the X Chromosome generally has more gaps than the autosomes. Despite this, compared to the reference genome, the number of gaps in our X Chromosome assembly is still lower. Previous studies have shown that transposable elements (TEs) content is likely the largest factor contributing to fragmented genome assemblies (Sotero-Caio et al. 2017) . Therefore, RepeatMasker was used to search known TEs by mapping sequences against the ruminant repeat library to annotate repeat sequences on each chromosome. The results revealed that the X Chromosome had a much higher proportion of TEs than the autosomes, especially containing more LINEs (Supplemental Fig. S23).

Detection of non-Hereford sequence

Minigraph (version 0.15) (Li et al. 2020) was used to integrate 21 genome assemblies into a multi-assembly graph. The Hereford-based linear reference genome (ARS-UCD1.2) was used as the backbone of the Chinese indicine multi-assembly graph, and 20 partially phased assemblies from the 10 Chinese indicine individuals were integrated into the multi-assembly graph with minigraph. To determine the support of the nodes, we aligned (with minigraph parameters "--cov -x asm") individual assemblies back to the multi-assembly graph. Nodes with non-zero coverage were then color labeled according to which assembly path traversed them. We used the bubble popping algorithm of gfatools (version 0.5) (Crysnanto et al. 2021) to derive structural variations from the multi-assembly graph, traversing all possible paths in the bubble, and retaining only paths with color-consistent labels. We then classify a path as a reference path if all nodes and edges are part of the Hereford-based reference assembly, and as non-reference otherwise (Crysnanto et al. 2021). We collected all non-Hereford alleles originating from bubbles (excluding complete deletions, paths without non-Hereford bases, and paths with length less than 100 bp) to obtain a comprehensive set of non-Hereford bases from the multi-assembly graph. The Brahman-backed pangenome was integrated using the same pipeline.

The previous pangenomes (Crysnanto et al. 2021; Leonard et al. 2022) and our multi-assembly graph were constructed using minigraph with ARS-UCD1.2 as the initial backbone of the graph structure. The 20 partially phased hifiasm assemblies of

10 Chinese indicine were integrated into a multi-assembly graph, which contained 160k non-reference nodes spanning 148.5 Mb in our study. In a previous study, five assemblies (Angus, Highland, Original Braunvieh, Brahman and their close relatives yak) were integrated into a multiassembly graph with minigraph. The multiassembly graph contained 59k nonreference nodes spanning 69 Mb bases (Crysnanto et al. 2021). In another previous study (Leonard et al. 2022), a pangenome constructed from five hifiasm assemblies added 82.5 Mb non-reference sequences across 88.5k nonreference nodes.

To exclude the non-reference nodes and variations caused by assembly errors and validate the non-reference nodes and variations, we mapped the PacBio HiFi reads from 10 Chinese indicine cattle to the multi-assembly graph using GraphAligner (version 1.0.13) (Rautiainen and Marschall 2020) with the command "GraphAligner -t 40 -x vg". We then calculated coverage (number of reads aligned) at each node and edge in the graph based on the graphical alignment format output from GraphAligner. 98.80% non-reference nodes and 98.49% of the structural variation breakpoints had support (more than one read in any one sample) from HiFi reads.

Whole-genome resequencing data of 84 cattle (16 African taurine, 16 African indicine, 16 Asian taurine, 8 European taurine, 8 Indian indicine, 8 gayal, 4 gaur, 8 banteng) were downloaded for analysing presence-absence of non-reference sequences. We appended the non-Hereford sequences as additional contigs to the ARS-UCD1.2 reference, making an extended reference genome. The reads were aligned to the extended ARS-UCD1.2 reference genome sequence. The presence and absence of each

novel sequence was then determined according to the sequence coverage and depth. Novel sequences with a depth greater than one-third of the whole-genome depth were identified as present.

Repeat analysis

Centromeric repeats were identified as Satellite/centr family by RepeatMasker (v4.1.1) (Chen 2004) using a modified Repbase database (release 20181026). Repeats with Smith-Waterman score < 20 and substitutions percentage > 40% were filtered out. Vertebrate telomeric repeats (TTAGGG) were identified within 10 kb of the chromosome end. According to a previous study (Leonard et al. 2022), the end-to-end was considered where the proximal end of a chromosome contains at least 50 kb centromeric repeat sequences and the distal end contains at least 500bp telomeric repeat sequences. Repeats of non-reference sequences were analyzed with a method combining homology comparison and *de novo* structure analyses. We applied RepeatMasker (v4.1.1) (Chen 2004) using the database of repetitive DNA elements from Repbase (release 20181026) to classify interspersed repeats in the longest allele sequence of each variation. We identified tandem repeats composed of a motif occurring twice or more by TRF (Tandem Repeats Finder) (Benson 1999).

Gene annotation and functional enrichment analysis of non-reference sequences

For gene annotation, *de novo* and homology-based approaches were used to annotate protein-coding genes. We collected all non-reference alleles (exclude paths with lengths less than 100 bp and without non-reference bases) and added 100 bp of flanking sequences to facilitate the mapping of short reads to the segment edges (Crysnanto et al. 2021). The filtered RNA-sequencing reads were mapped using HISAT2 (v2.1.0) (Pertea et al. 2016) to the extended reference which was constructed by appending non-reference sequences as unplaced contigs and putative novel transcripts were assembled *de novo* using StringTie (v1.3.4) (Pertea et al. 2016). For homology-based prediction, DIAMOND BLASTX (Buchfink et al. 2015) was used to search for non-reference transcripts sequences against the nr database of NCBI and the RefSeq protein sequences from mammals (GCF_002263795.1, GCA_003369695.2, GCF_000298355.1, GCF_000754665.1, GCF_003121395.1, GCF_001704415.1, GCF_002742125.1, GCF_000001405.39, GCF_000001635.26) to determine the potential presence of protein-coding genes with an E-value cutoff of 1×10–5. For *de novo* prediction, AUGUSTUS (Stanke et al. 2008) software was used to predict genes. Finally, we integrated all evidence as a catalog of the non-redundant genes residing in non-reference sequences. Annotated genes of novel sequences were analyzed for Kyoto Encyclopedia of Genes and Genomes (KEGG) terms and pathway enrichment using KOBAS (Bu et al. 2021).

Detection of high-confidence SVs

To obtain high-confidence SVs, we carried out a long-read mapping-based approach to call SVs with multiple tools and strict filtering steps. For each genome, we aligned PacBio HiFi reads to the cattle reference genome ARS-UCD1.2 (Rosen et al. 2020) by pbmm2 (https://github.com/PacificBiosciences/pbmm2) (version 1.4.0; with parameters: --sort --preset HiFi --rg "@RG\tID:SampleID"). Subsequently, pbsv v2.4.0 (https://github.com/PacificBiosciences/pbsv), SVIM v1.4.2 (Heller and Vingron 2019), Sniffles v1.0.12 (Sedlazeck et al. 2018), and cuteSV (Jiang et al. 2020) v1.0.10 were used to detect SVs in the aligned BAM files for each sample. pbsv was used with parameters “discover, call --ccs -m 50”, and we set the following parameter for SVIM: “alignment --min_mapq 20 --min_sv_size 30 --max_sv_size 100000 --minimum_depth 3 --sequence_alleles --insertion_sequences”, for the Sniffles with parameters: “ --genotype -s 3 --report-seq --report_str --ccs_reads”, and the cuteSV with parameters: “--max_cluster_bias_INS 1000 --diff_ratio_merging_INS 0.9 --max_cluster_bias_DEL 1000 --diff_ratio_merging_DEL 0.5 --genotype --min_mapq 20 --min_size 30 --max_size 100000”.

We merged the SV callsets of each individual derived from the above four SV callers for each sample. We applied the SURVIVOR (Jeffares et al. 2017) software to merge SVs for each SV type based on the variant position and length, and retained SVs detected by at least two callers for each individual. The max distance between breakpoints of SVs is equal to 10, and minimum size of SVs to be taken into account is equal to 50, hence parameters were: “merge $Sample.vcf_files_raw_calls.txt 10 2 1 0 0 50 $Sample.merged.vcf”. As suggested by benchmark analysis of LRS callers

(Dierckxsens et al. 2021), we chose results that have a priority of pbsv > SVIM > Sniffles > cuteSV. Finally, we merged the above high-quality SV call set for each sample using SURVIVOR with parameters: “merge vcf_files_calls.txt 50 1 1 0 0 50 Total_sample.merged.vcf”.

SV annotation and hotspot identification

Annotation of the SV callset was performed using SnpEff (version 5.0e) (Cingolani et al. 2012) and ANNOVAR (Wang et al. 2010) software. According to a previous study (Ebert et al. 2021), we selected the middle position of each SV and used the “hotspotter” function of the primatR package for hotspot analysis with parameters: “bw=200000, pval = 1e-08, num.trial =2000”. Previously published long-read-based SV datasets by Crysnanto et al and Talenti et al (Crysnanto et al. 2021; Talenti et al. 2022) were also used for hotspot detection and comparison. We classified our inferred SV hotspot (n=206) into three catalogues: “terminal”- residing in the last 5 Mb of the chromosome end (n=61)”, known”- overlapping with previous study (n=26), and “novel” - unique for this study (n=119). We further investigated whether our SV hotspots overlap with protein-coding genes. For this, we extracted 20,271 unique protein-coding genes from the generic feature file of ARS-UCD1.2 (NCBI Assembly: GCF_002263795.1). To test for overrepresentation of SV hotspots in coding regions, we ran a permutation test using the regioneR package with 1000 iterations, the permTest function was used for calculation with the circularRandomizeRegions function. We obtained the repeat file (GCF_002263795.1_ARS-UCD1.2_rm.out.gz)

and gap file (GCF_002263795.1_ARS-UCD1.2_genomic_gaps.txt.gz) from NCBI. Then, we extracted 'Satellite/centr' entries from the repeat file and merged them with the gap regions to create a comprehensive mask file. As a comparison, we also performed the same permutation test with all SVs.

Reads mapping and variants calling

The short read data were filtered by fastp program (version 0.20.0) (Chen et al. 2018) with default parameters. All passed quality-filtered reads were aligned to the reference genome (ARS-UCD1.2) with BWA-MEM (version 0.7.17) (Li and Durbin 2009) using default parameters. After sorting the BAM files and removing PCR deduplicates using Picard (version 2.1.1) (http://broadinstitute.github.io/picard/) tools, SNPs were called using the Genome Analysis Toolkit (GATK) (version 2.1.1) (McKenna et al. 2010) and then we filtered raw single nucleotide polymorphisms (SNPs) with the parameters "QUAL <40.0, MQ < 25.0, MQ0> = 4 & ((MQ0/(1.0 * DP))> 0.1, -cluster 3 -window 10". The final obtained variants were filtered using BCFtools (version 1.1) (Danecek et al. 2021) with parameters '-i 'MAF>0.01 & F_MISSING<=0.1' -v.

Graph genotyping of structural variation

We downloaded a large amount of short-read sequence data (~13x coverage) from representative breeds from all over the world. All short-reads data were mapped to ARS-UCD1.2 cattle reference using BWA-MEM (version 0.7.17-r1188) (Li and

Durbin 2009) with default parameters. Using above SVs set by the long-reads mapping-based approach identified from HiFi sequencing data from 10 Chinese indicine cattle, we used Paragraph (Chen et al. 2019) (v2.4a; with default parameters) to genotype SVs in each individual. To ensure a high-quality SVs set for population genetics analysis, we excluded SVs that failed to be genotyped in >90% of the samples and an excess of heterozygotes deviated from Hard Weinberg equilibrium using VCFtools (Danecek et al. 2011). 41,622 SVs were removed after filtering, leaving 114,387 variants for downstream analysis.

Given the imprecision of SV breakpoints from the pangenome (Supplemental Fig. S12), we employed the graph-based Paragraph (Chen et al. 2019) for an SV dataset from read-based methods, leading to an underutilization of the constructed pangenome. Additionally, we implemented a new pipeline using Minigraph-Cactus (Hickey et al. 2023) for construction and PanGenie (Ebler et al. 2022) for genotyping. We also explored the use of vg (Hickey et al. 2020) for RNA mapping (vg mpmap). Although its output was not used in subsequent analysis, this pipeline provides valuable insight for future pangenome graph applications.

Ancient kouprey data processing pipeline

The sequencing data of two extinct kouprey (Sinding et al. 2021) were downloaded and processed using the paleomix pipeline (Schubert et al. 2014). Briefly, reads shorter than 25 bp were removed, and reads were then further trimmed for low-quality termini and adapter sequences using AdapterRemoval v2.2 (Schubert et al. 2016).

Then we used BWA v0.7.17 aln algorithm (Li and Durbin 2009) to map reads to the cattle reference genome ARS-UCD1.2, disabling seeds and excluding reads whose mapping qualities were lower than 25. Finally, PCR duplicates were removed using Picard Tools v2.18.7 (http://broadinstitute.github.io/picard/) MarkDuplicates command and insertion-deletion (indel) realignment was performed using GATK v3.8 (McKenna et al. 2010). To minimize bias from low coverage sequencing, we applied pseudo-haploid calls for the two ancient samples by randomly sampling alleles using ANGSD v0.935 (Korneliussen et al. 2014).

Identifying archaic introgressed fragments

To more accurately identify the archaic fragments of Chinese indicine, we constructed a high-quality SNP collection. For partially phased assemblies, we aligned each partially phased assembly to the reference genome (ARS-UCD1.2) using minimap2 (version 2.17-r941; with parameters: --paf-no-hit -a -x asm5 --cs -r2k) (Li 2018) and SAMtools (version 1.12; with default parameters) (Danecek et al. 2021), then the SNPs was called by bcftools mpileup (version 1.12; with parameters: -Q 20 -q 20 --annotate FORMAT/AD,FORMAT/DP) (Danecek et al. 2021) and bcftools call (with parameters: -m -v -O z). For the whole-genome resequencing data SNPs set, we used the method in the SNP calling part above. To reduce any bias from combining these two SNPs sets, we also added the resequencing data of the corresponding samples of the partially phased assembly; then Beagle (version 5.1; with default parameters) (Browning et al.

2018) was used to phase the SNPs; and according to the evaluation results of phasing accuracy below in the Methods section, we removed the resequencing Chinese indicine SNPs from the phased set. The SNP set of both partially phased assemblies and resequencing data were integrated into the final SNPs collection with BCFtools (version 1.12) (Danecek et al. 2021) and in-house scripts, which were used for subsequent analysis.

To infer archaic introgressed fragments in Chinese indicine, we used a two-state hidden Markov model (Skov et al. 2018) to remove the variation found in outgroup and then using the remaining variants to group the genome into regions of different variant density, which the introgressed regions have higher variant density than non-introgressed. According to the major update pipeline of this model (https://pypi.org/project/hmmix/), (1) we first found SNPs which are derived in 321 non-Chinese indicine domestic populations (Indian indicine, African indicine, African taurine, Asian taurine and European taurine as outgroup) (with the command: hmmix create_outgroup); (2) then using the number of variants in the outgroup to estimate the substitution rate as a proxy for mutation rate for each independently partially phased assembly (with parameter: hmmix mutation_rate); (3) and keep variants that are not found to be derived in the outgroup for each individual in Chinese indicine (with parameter: hmmix create_ingroup); (4) and we train the parameters for haploid data and then decode (with parameters: hmmix train -haploid; hmmix decode -haploid). The introgressed fragments were retained with a posterior probability of more than 90% of being archaic.

In addition, we also obtained the introgressed fragments of short-read corresponding to partially phased assemblies and compared the results of them from the same individual. The introgression content obtained by partially phased assemblies is slightly higher than that of short-read individuals, but their length is significantly longer (Supplemental Fig. S21). To corroborate this, we randomly selected 20 introgressed fragments and visualized the original read mapping and phasing genotypes through IGV software (Robinson et al. 2023), and found that the introgressed fragments of short-read based SNPs are much shorter than those of partially phased assemblies based SNPs because the former was interrupted by a switch error, meaning that introgressed fragments in complex regions cannot easily be found by short-read data. Combined with the evaluation results of phasing accuracy below in the Methods section further illustrates that partially phased assemblies have a significant advantage in haplotype continuity. Therefore, we used the introgressed fragments inferred from partially phased assemblies for subsequent analysis.

Finally, to determine the possible source of introgression for each archaic fragment, we constructed a tree topology for each fragment across *Bos* genus samples. We extracted the corresponding SNPs of other species of the *Bos* genus (the populations were African buffalo, Indian indicine, Asian taurine, European taurine, yak, American bison, European bison, ancient kouprey, banteng, gayal, and gaur) on each archaic introgressed fragment, where heterozygous SNPs sites were represented using the standard International Union of Pure and Applied Chemistry chemical nomenclature (IUPAC) codes (Cornish-Bowden 1985). We inferred the tree topology by using the

maximum-likelihood statistical method for each archaic fragment using the MEGA (Tamura et al. 2021) (version 11.0.10) (with parameters: Bootstrap 100 replicates; Model Tamura-Nei model) program. The Environment for Tree Exploration (ETE) package (V3) (Huerta-Cepas et al. 2016) was used to identify the closest branch for each archaic fragment as a possible donor.

Archaic introgression

To explore the introgression landscape of other cattle groups like European taurine and Indian indicine, we perform the *D* and f_4-ratio statistics across the combinations of trio populations using Dsuite software package (Malinsky et al. 2021). The significant introgression signal between indicine cattle (Indian indicine cattle or Chinese indicine cattle) and wild Asian *Bos* species (*Z* scores > 3) (Supplemental Table S22). According to the results of f_4-ratio measure, about 4.9% of the genome per Indian indicine cattle is introgressed from wild Asian *Bos* species using European taurine as background population, while about 15.1% of the per Chinese indicine genome was introgressed using European taurine as background population. When using the Indian indicine population as the background population, 10.7% of the genome per Chinese indicine was introgressed from wild Asian *Bos* species, which is similar to the average introgression proportion of 9.5% among the sequenced Chinese indicine cattle in our main text. 73.3% of the genome covered by archaic introgressed fragments were newly introgressed during the migration of Indian indicine cattle to southern China.

Phasing accuracy evaluation

In order to ensure that the sliding window size can contain a sufficient number of SNPs and that the window size is much larger than the length of the long reads (the average length of the long reads is ~15 kb), we defined sliding windows of 100 kb to evaluate the difference of between the long read and short read phasing. We calculated the relative phasing difference between data sets each window by using the common heterozygous phasing SNP of the long read and short read of the same individual. The results show that the average phasing difference of long read relative to short read is 0.0232. To further determine the respective phasing difference of long reads and short reads, we randomly selected 100 windows and manually checked them using the IGV software (Robinson et al. 2023) combined with the sequencing read mapping and phasing. Statistics show that the average difference of short reads is 0.0197, while no errors were found in long reads. hifiasm (since v0.15) can produce two sets of partially phased contigs, which represents the two haplotypes in a diploid genome, though with occasional switch error (Cheng et al. 2021). Therefore, the haplotype block size can be represented by the length of partially phased contigs, and the length distribution of these reveals very high continuity of haplotypes (Supplemental Fig. S22).

Incomplete Lineage Sorting

To verify that the haplotype fragments are introgressed, we calculated the probability of incomplete lineage. According to a previous paper (Huerta-Sanchez et al. 2014),

the expected length of a shared ancestral sequence is L = 1/ (r × t), where r is the recombination rate per generation per base pair and t is the branch length between Chinese indicine and introgressed source species since divergence. We assume a generation time of 6 years for domesticated cattle, and a median divergence time of 500 KYA between Chinese indicine and Banteng (or Kouprey, Gaur, Gayal) (Wu et al. 2018), and a recombination rate of 1.23 cM/Mb (Weng et al. 2014). The mean length of haplotypes shared through ILS is then 487.8 bp, and the probability of a shared 1 kb haplotype length deriving from ILS is 0.3926415, according to this formula: 1 – GammaCDF (m, shape = 2, rate = 1/L), where GammaCDF is Gamma distribution function (Supplemental Table S13).

Estimation of admixture time based on length distribution of introgressed fragments

We applied *MultiWave* 2.0 (Ni et al. 2019) to infer the optimal model without prior model assumptions or estimate parameters and admixture time, whereas based on the length distribution of introgressed fragments in admixed genomes. The program first applies a likelihood ratio test and an exhaustion method to choose an optimal admixture model based on the length distribution of introgressed fragments, and then uses an EM-algorithm to estimate the corresponding admixture times and proportions under the above optimal model. The length distributions for introgressed fragments were taken from the results of hmmix above. The results of *MultiWave* 2.0 showed

that the “HI model” fitted the Chinese indicine population well, with two discrete admixture events and 100% confidence intervals of the admixture dates for each donor population obtained from 1,000 bootstrapping repeats (Supplemental Table S14). Additionally, to ensure the reliability of the results we used MultiWaverX (Zhang et al. 2022) to infer the admixture time for each individual of Chinese indicine cattle with each donor *Bos* species, and the results were similar to the above (Supplemental Table S14). Although the proportion of introgression among sub-breeding of Chinese indicine cattle was significantly different (Supplemental Fig. S24c), the length distribution of introgressed fragments and admixture time from different donor species of all sub-breeding were similar (Supplemental Fig. S24a, b), which further indicates that introgression events occurred at the same time period (before the indicine cattle introduced to China).

Identification of population-stratified and introgressed SVs

In order to determine Chinese indicine cattle candidate population-stratified SVs, we calculated the fixation index (F_{ST}) of SVs between Chinese indicine cattle and Indian indicine cattle using VCFtools (Danecek et al. 2011) We then performed permutation tests 1000 times to calculate the empirical p-values. We finally applied screening to retain 16,584 SVs in the Chinese indicine cattle population-stratified that satisfy (1) $F_{ST} > 0.1$, (2) empirical p-value < 0.05, and (3) maximum missingness rate < 0.1 and minimum allele frequency > 0.01.

For the validation of introgressed SVs in Chinese indicine cattle population derived from other *Bos* species, we considered: (1) SVs uniquely shared between wild Asian *Bos* species and Chinese indicine cattle, but absent in other indicine and taurine cattle; (2) SVs that are positioned on the archaic introgressed fragments (see above). A total of 14,716 introgressed SVs were retained. Finally, the intersection of population-stratified and introgressed SVs of Chinese indicine cattle identified a total of 3,136 shared SVs (Supplemental Table S17). The pipeline and custom script used for the analyses are available on GitHub (https://github.com/Xuelei-Dai/Chinese_indicine_pan-genome_project_scripts_and_piplines).

Haplotype pattern and network analysis

To visualize the specific genotype pattern of the introgressed regions flanking the 6.3 kb insertion (13:63675338), we extracted the phased SNPs in the 20 kb region flanking the variant from 389 individuals and visualized their genotypic pattern in a heatmap (Fig. 5c). We also constructed haplotype networks for this region using R package PEGAS based on the pairwise differences (Paradis 2010) (Fig. 5e). We retained at most four individuals with consistent haplotypes in the same population and excluded removed SNPs with minor allele frequency < 0.05. In this introgressed region, we retained 173 SNPs in 118 samples from 14 *Bos* populations. In addition, by combining the introgressed SV in the "Identifying archaic introgressed fragments" section, we found haplotypes where insertions were clustered together and defined

two main haplotypes in Chinese indicine population as introgressed haplotype-A and non-introgressed haplotype-B.

Detecting expression-associated SVs

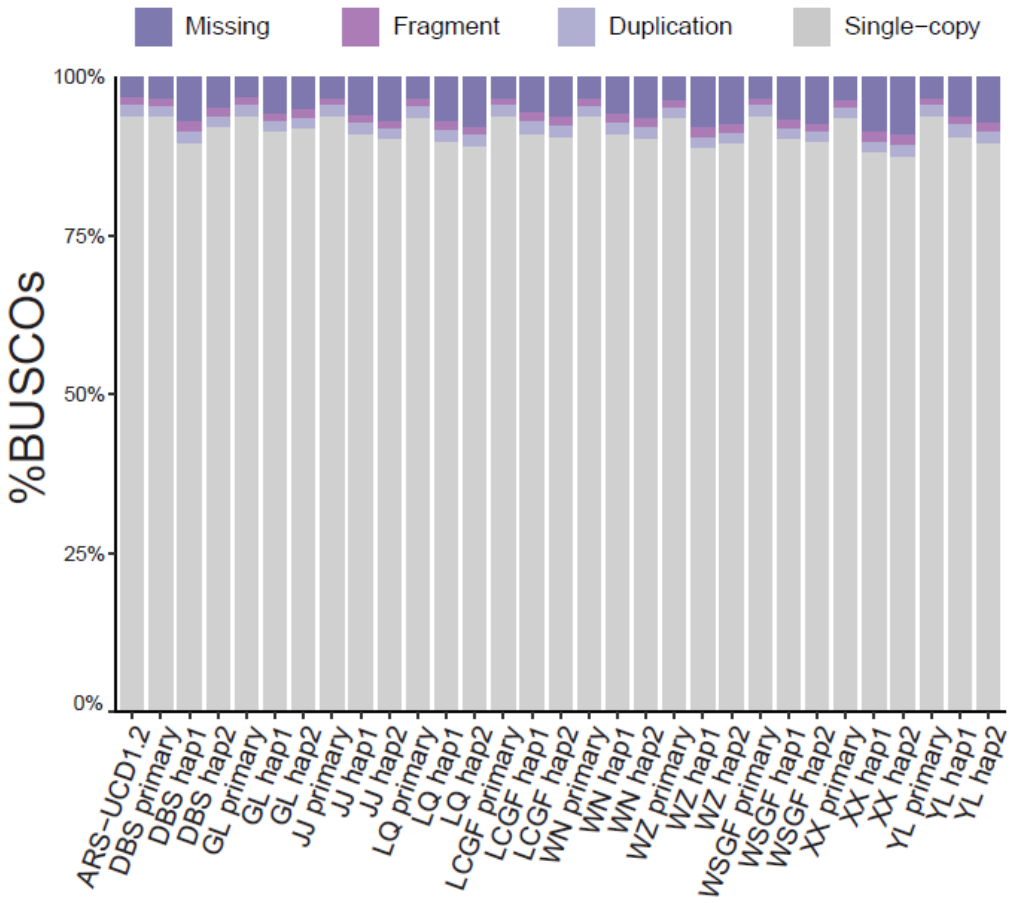
In order to reduce potential mapping bias caused by non-Hereford sequences, we constructed a custom genome by merging the non-Hereford sequences and the reference genome (ARS-UCD1.2). RNA sequencing reads of PiNan cattle were aligned to the custom genome by STAR (version 2.2.1) (Dobin et al. 2013). In an effort to minimize mapping bias, we also used the 'mappability filtering' module of WASP software to filter our BAM files (https://github.com/bmvdgeijn/WASP)(van de Geijn et al. 2015).

Then, the ASE genes were identified following the pipeline described in our previous study (Wang et al. 2022). Briefly, the ASE SNPs were identified from RNA-seq data using the following parameters: (1) at least 20 total reads; (2) allele ratio (allele reads count/total reads count)> 0.65 or <0.35; (3) significant allelic imbalances with FDR<0.05 (Chi-squared test). Those genes with at least two ASE SNPs on gene exons and found in at least three samples were defined as candidate ASE genes. For each SV-gene pair, the following conditions need to be met: (1) the SV is located 100 kb upstream and downstream of the gene; (2) the SV is heterozygous; (3) the ASE gene corresponds to the heterozygous state of the associated SV in at least three RNA-seq samples. Finally, we considered these SVs as plausibly expression-associated.

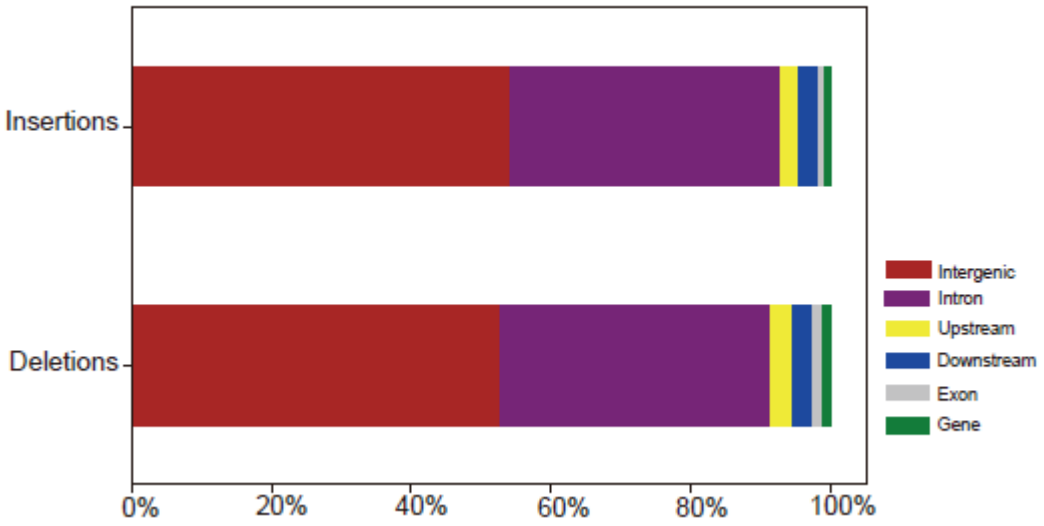
Gene-level transcript abundance was estimated as transcripts per million (TPM) using the StringTie (version 1.2.2) (Kovaka et al. 2019). We considered a gene as detected/transcribed when it had an expression value greater than 1 TPM in at least three samples. Local eQTL were calculated based on linear regression models using FastQTL (v2.0) (Ongen et al. 2016). The mapping window to detect local eQTL was defined as 1 Mbp up- and down-stream of the gene. To account for confounding effects, we incorporated gender and population structure (the top three principal components derived from PCA analysis of the 19 individuals) as covariates in our analysis. FastQTL was run in normal mode to calculate nominal P-values of all local gene-variant associations. Statistical significance was considered with $P < 0.05$ (Benjamin Hochberg adjusted).

In addition, in order to identify ASE of the novel genes, we merged the novel sequence and the reference genome, re-mapped the transcriptome data across 19 PiNan cattle to the newly merged reference genome, and focused on checking the expression of the novel genes. We found that 120 novel genes (including 41 complete gene models with more than 150 amino acids) exhibited expression levels (transcripts per million (TPM) > 1) in at least three samples. Furthermore, the enrichment analysis for these novel genes revealed that multiple GO terms are related to immune response (Supplemental Fig. S25).

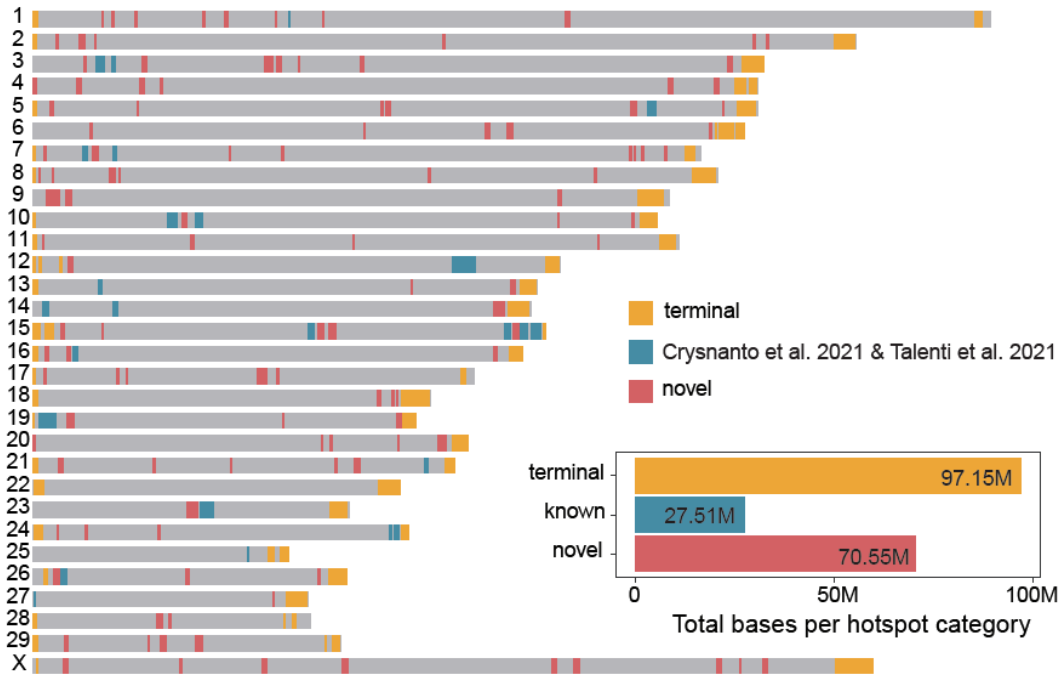
Supplemental Figures



Supplemental Fig. S1 The assembly completeness of Chinese indicine cattle.

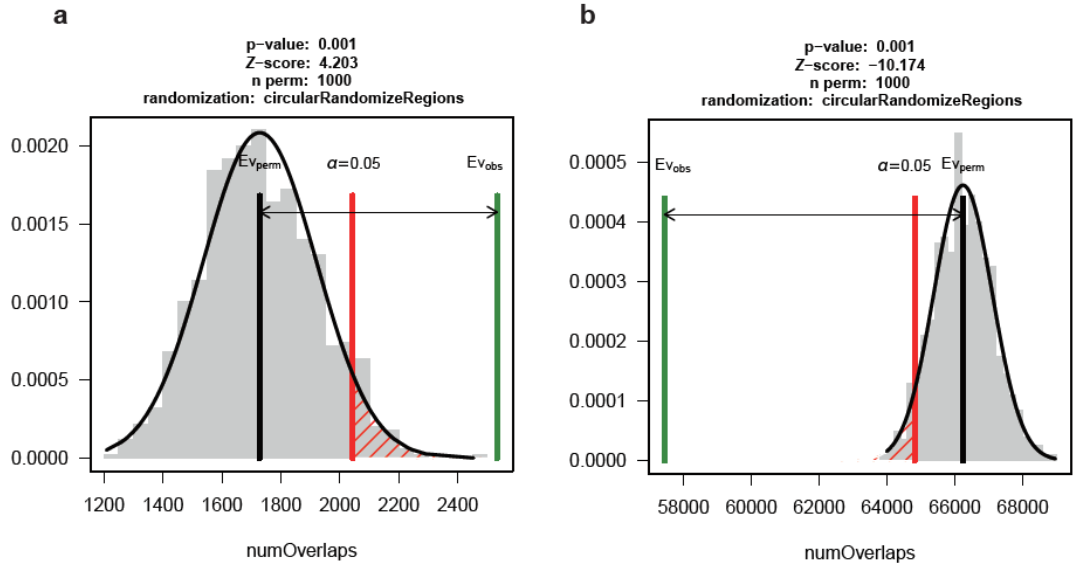


Supplemental Fig. S2 Genomic annotation of SVs calls set by SnpEff. Stacked bar chart showing the percentages of SVs overlapping with different genomic features.

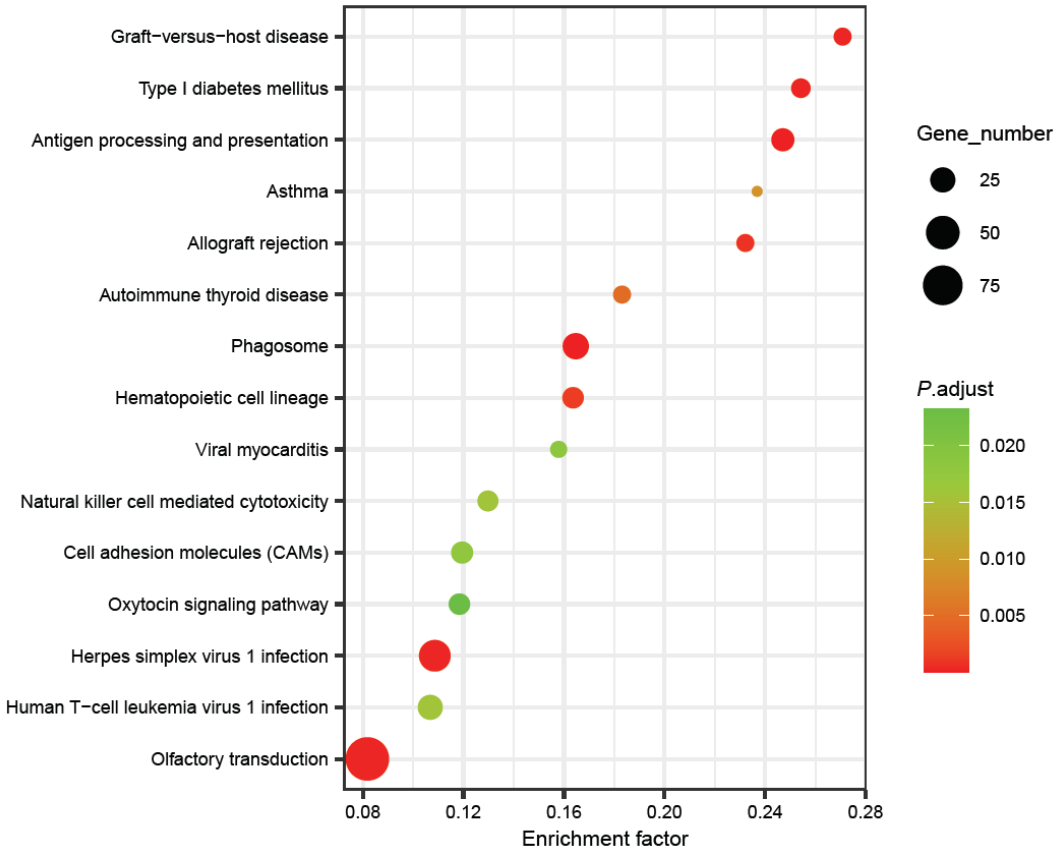


Supplemental Fig. S3 Genome-wide distribution of our SV hotspots. SV hotspots are classified into three categories: 'terminal' - hotspots are within the last 5 Mb of

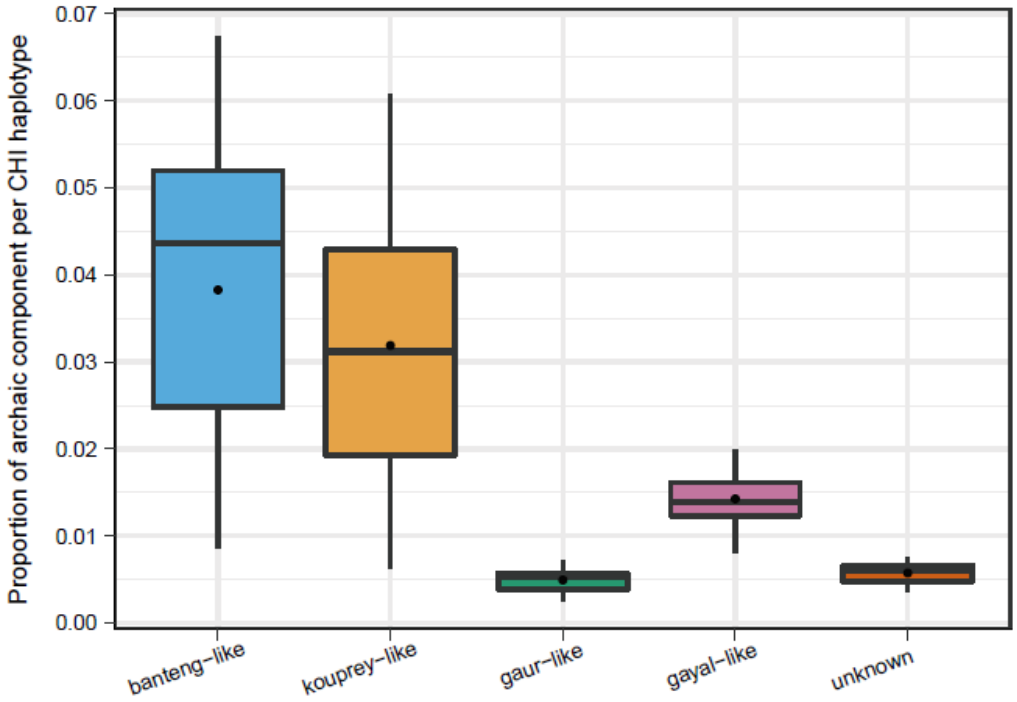
chromosome arms, ‘known’ - overlapping with hotspots identified in previously published SV datasets by Crysnanto et al and Talenti et al, and ‘novel’ - unique for this study. The inset shows the total sequence length in each hotspot category.



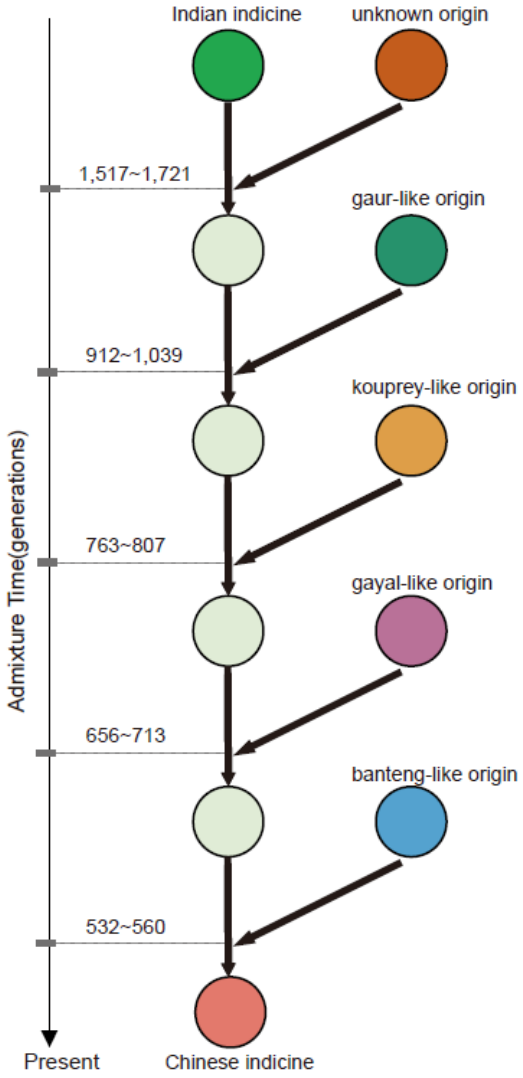
Supplemental Fig. S4. The permutation result of protein-coding genes intersected with 206 SV hotspots (a) and all SVs (b). Gray regions represent the distribution of the evaluation of the randomized regions, the green line indicates the evaluation of the original region set, and the red line indicates the significance limit.



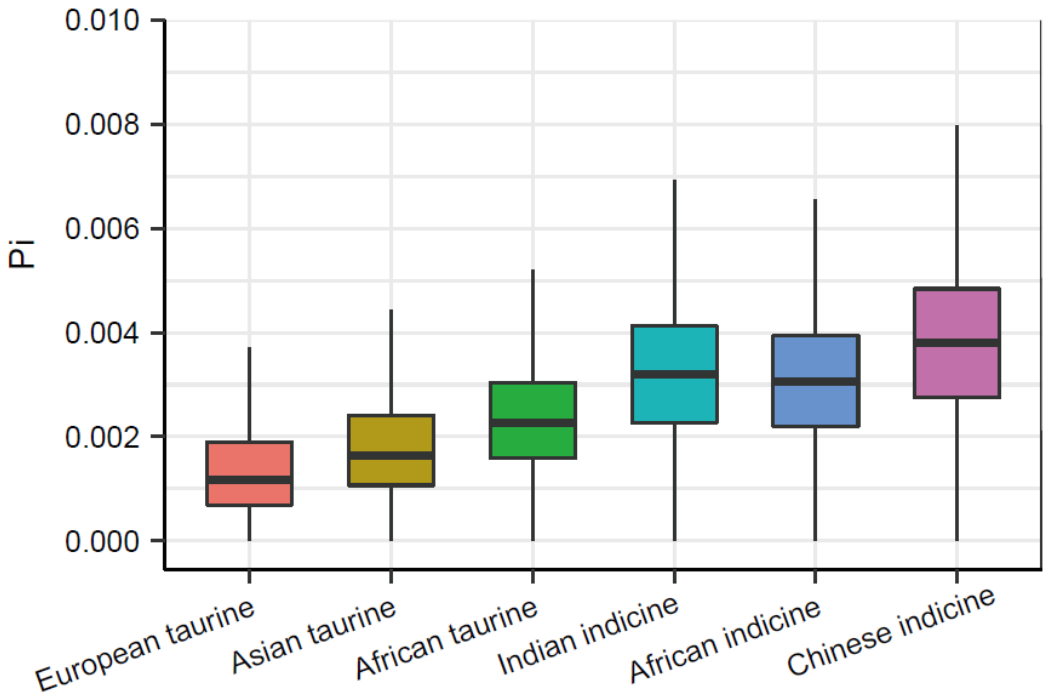
Supplemental Fig. S5 KEGG pathway analysis of the hotspots-genes accomplished by KOBAS. Enrich pathways with corrected *P*-values ≤ 0.05 are shown in the figure.



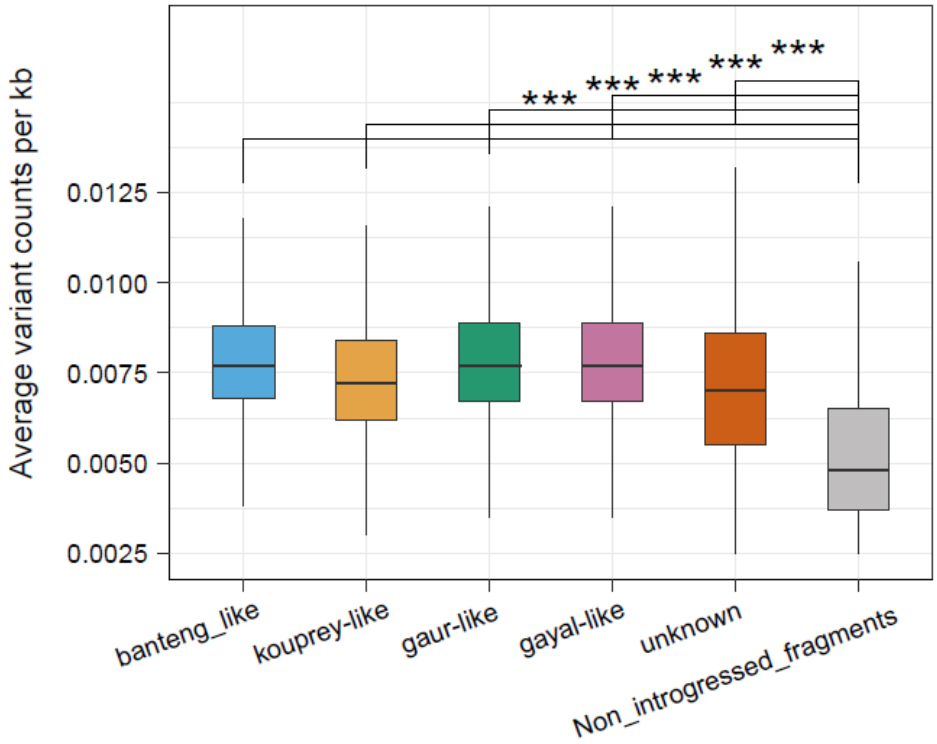
Supplemental Fig. S6 Average archaic proportion from different origins in each Chinese indicine (CHI) cattle haplotype. The mean proportion for each origin is shown with black dots.



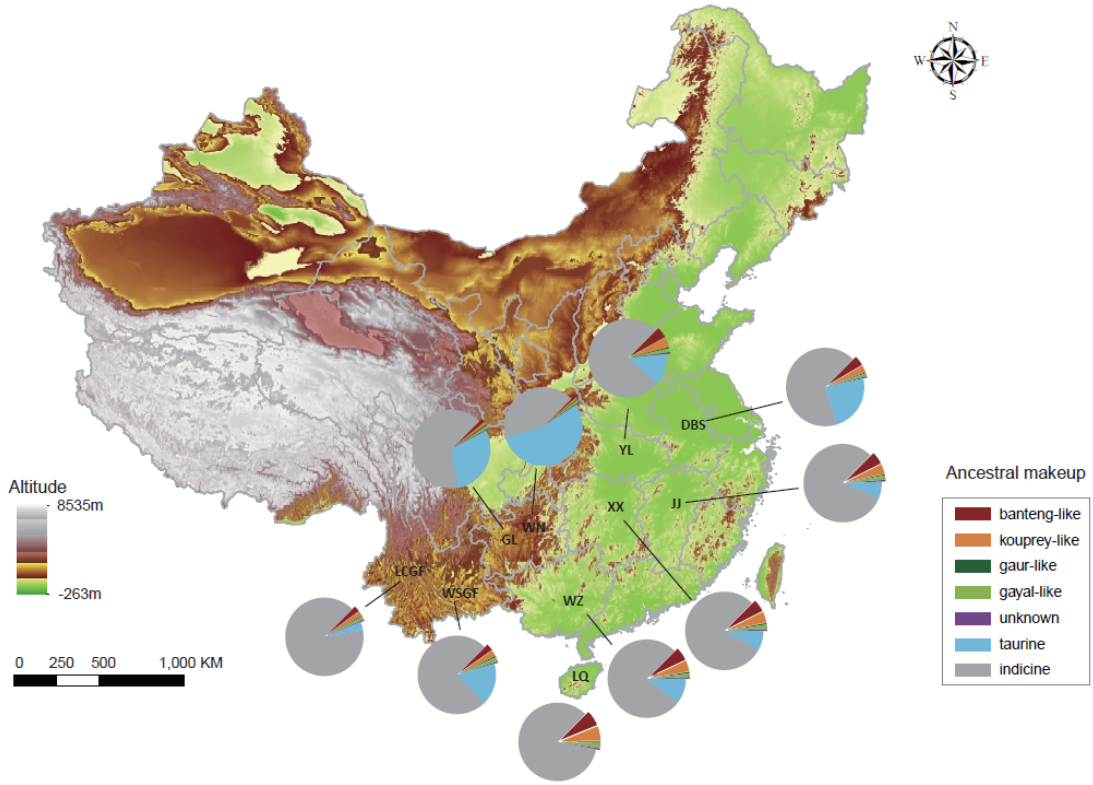
Supplemental Fig. S7 Schematic diagram of admixture topology and admixture time of multiple *Bos* species in Chinese indicine cattle based on admixture history results.



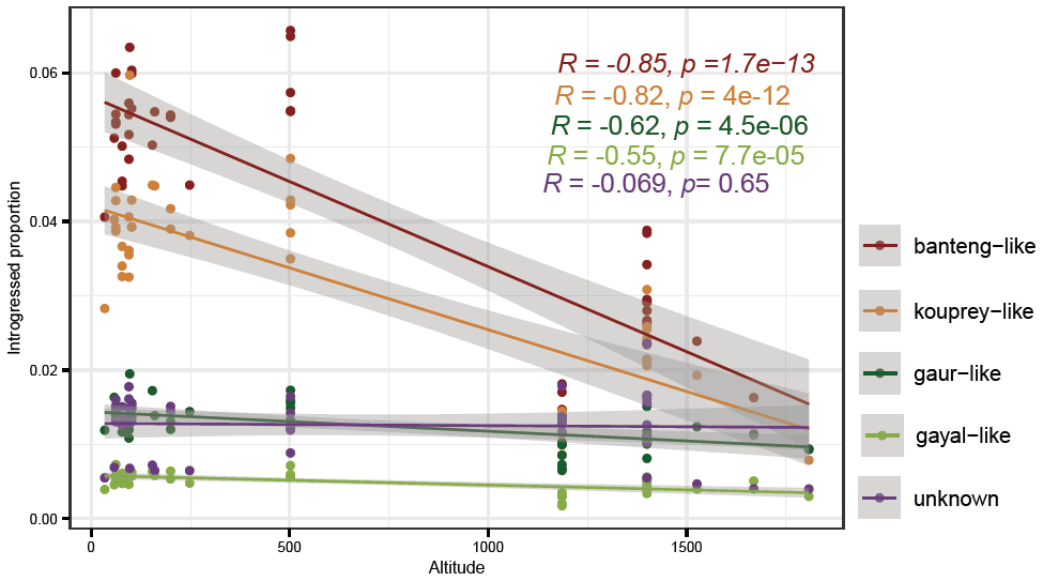
Supplemental Fig. S8 Genome-wide nucleotide diversity (Pi) of different populations.



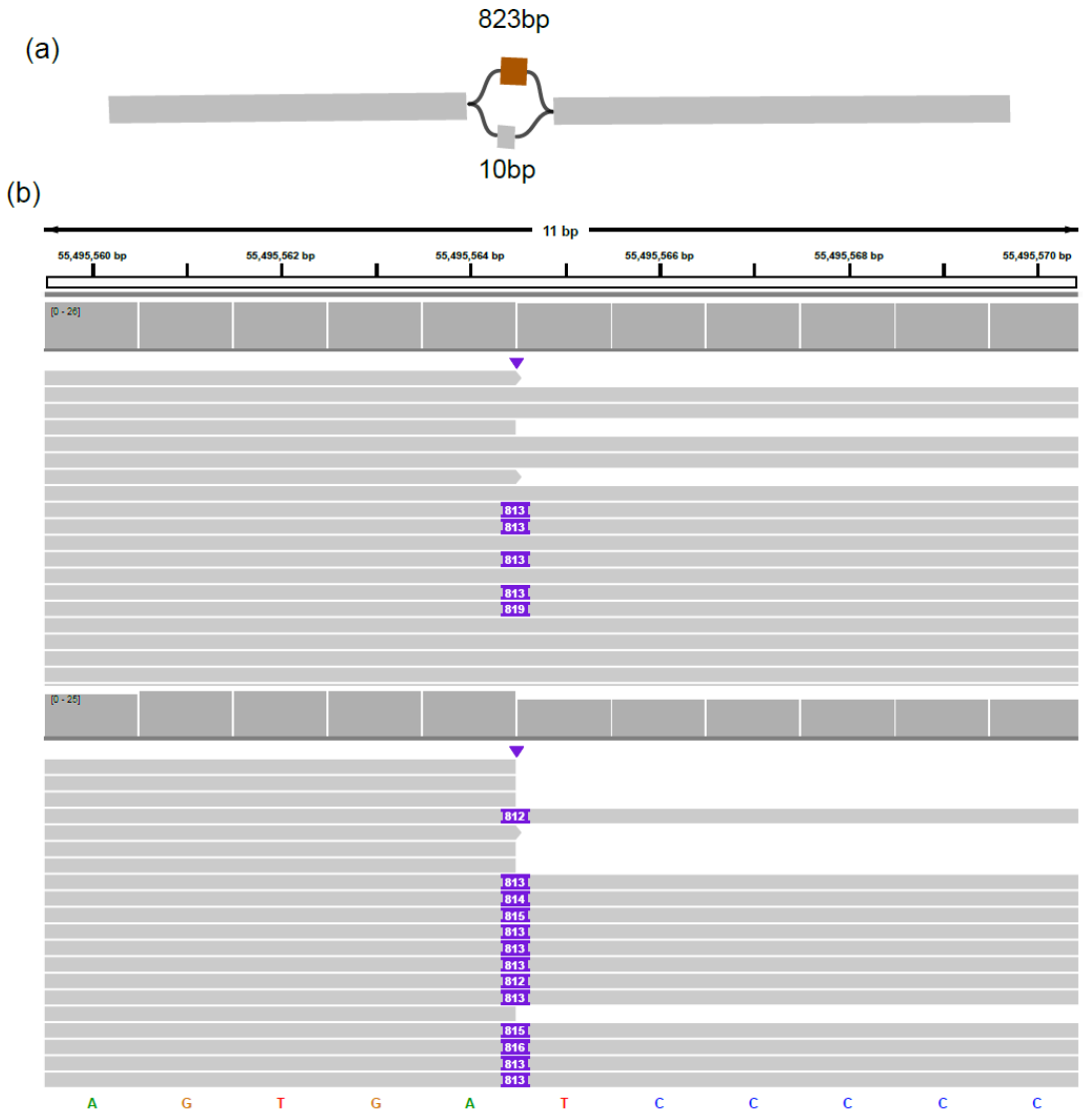
Supplemental Fig. S9 Average variant counts for banteng-like, kouprey-like, gaur-like, gayal-like, and unknown archaic fragments as well as non-introgressed fragments.



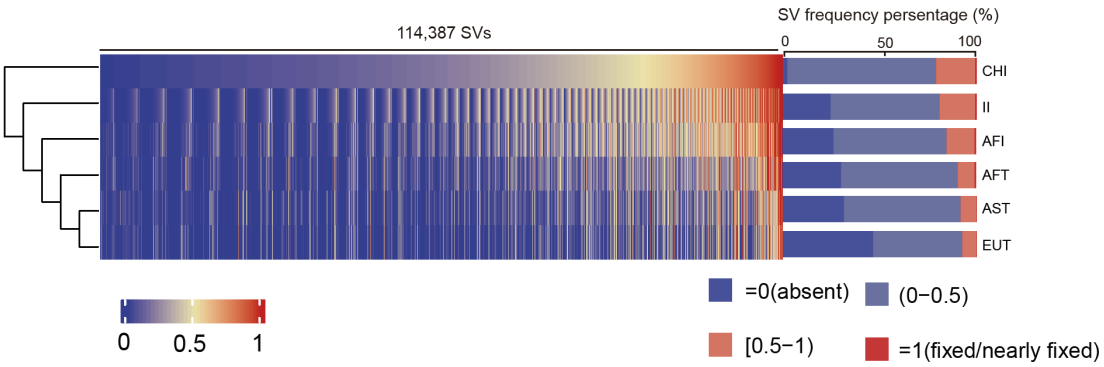
Supplemental Fig. S10 Approximate geographical locations and genetic makeup of 10 Chinese indicine cattle assemblies. Colors ranging from blue to red on the map represent the altitude from low to high.



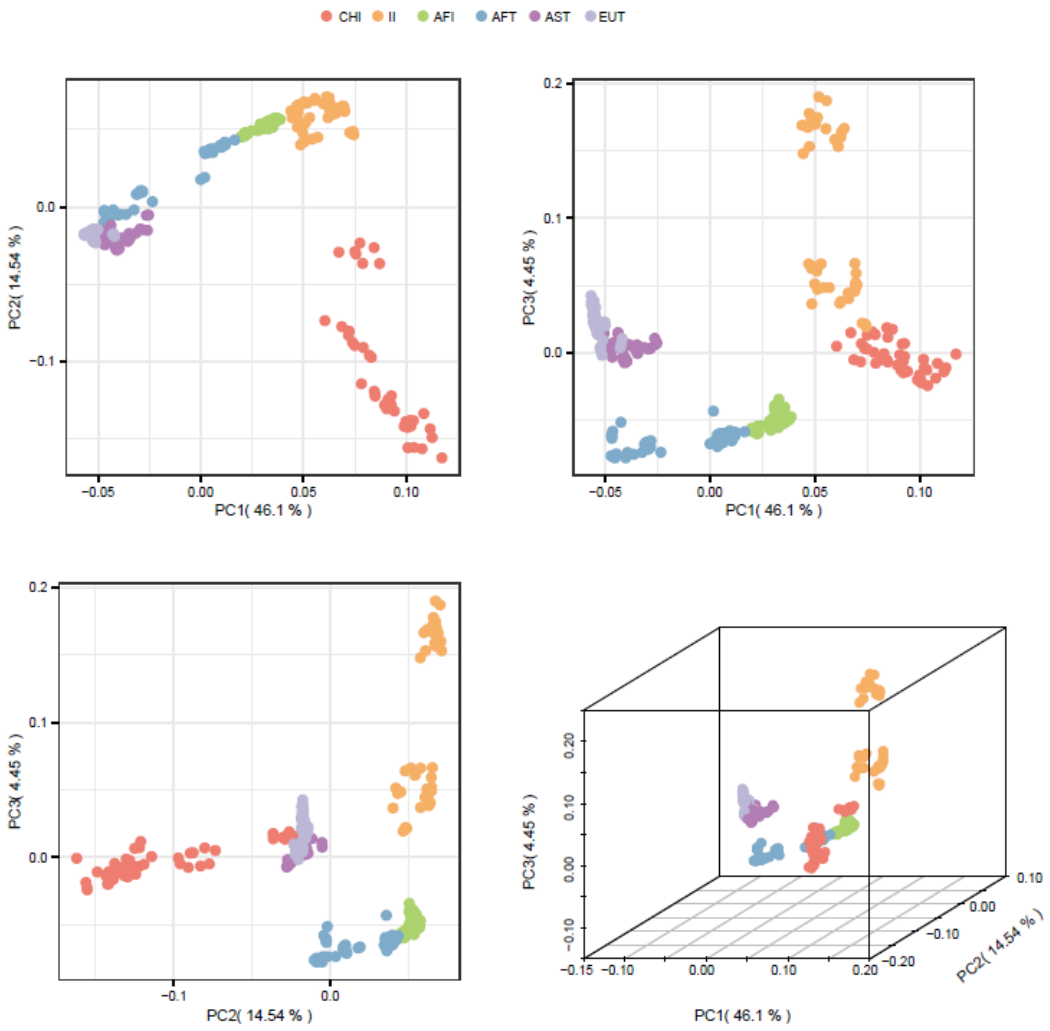
Supplemental Fig. S11 The relationship between the mean introgressed proportion and altitude for each Chinese indicine cattle, the gray interval represents the 95% confidence interval.



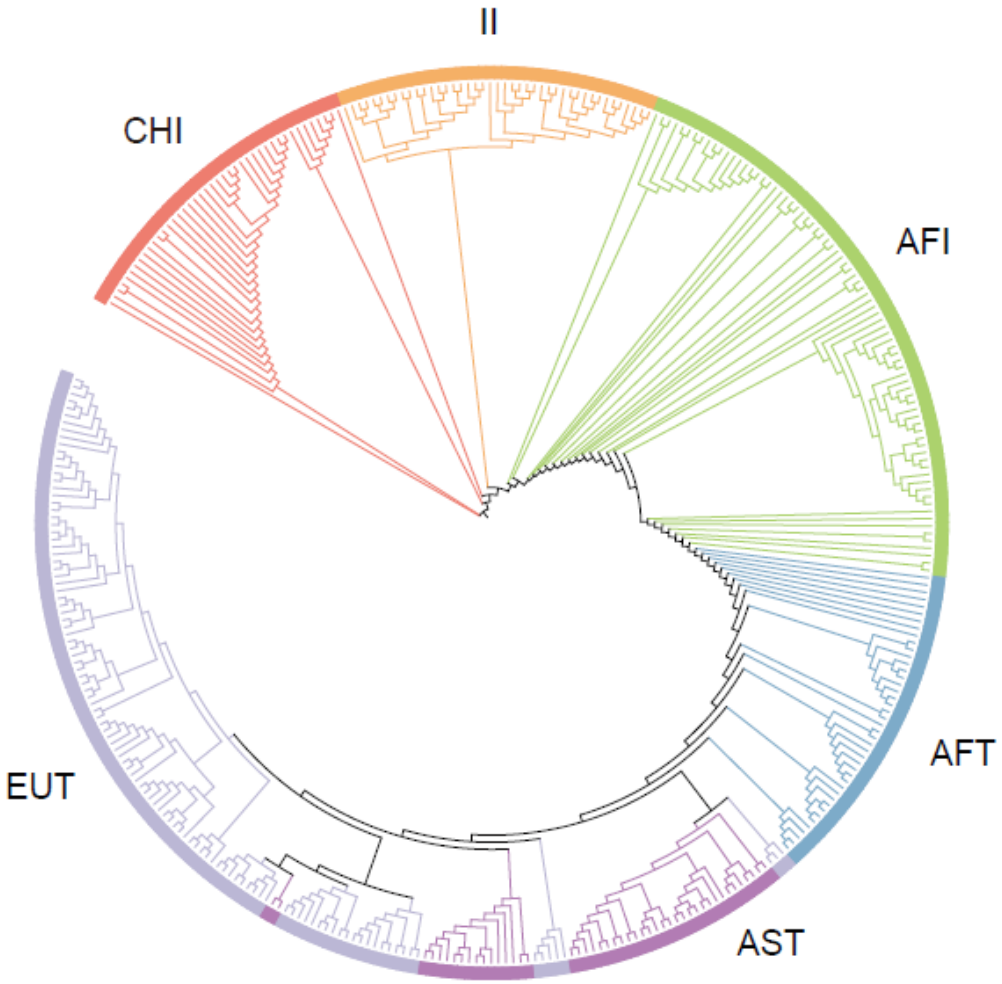
Supplemental Fig. S12. The insertion (INS.35432) was found by minigraph (a) and read-based alignment (b).



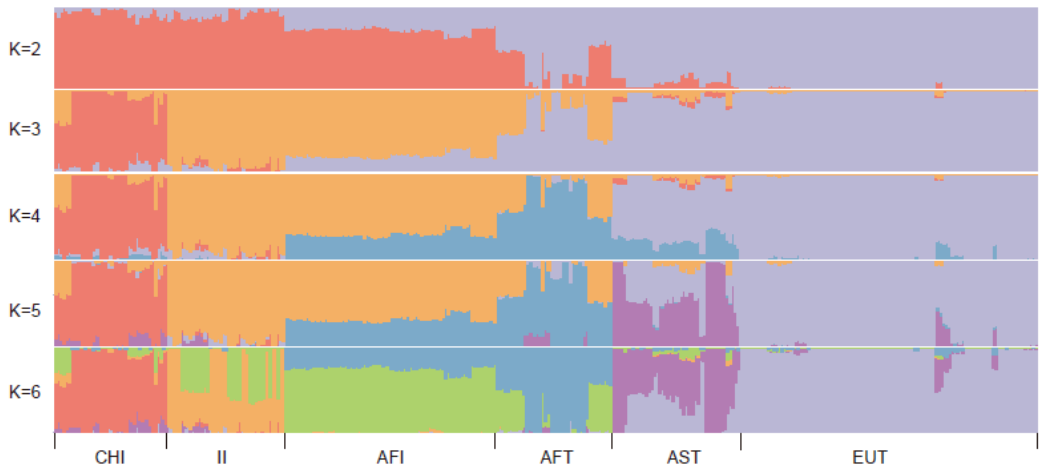
Supplemental Fig. S13 114,387 high-quality SVs were used for population genetic analysis. Heatmap showed the frequency distribution of 114,387 SVs in each population. The bar plot at the right of the heatmap showed the percentage of the SV frequency of the corresponding population. CHI, Chinese indicine; II, Indian indicine; AFI, African indicine; AFT, African taurine; AST, Asian taurine; EUT, European taurine.



Supplemental Fig. S14 Principal components analysis (PCA) of 114,387 SVs separated all samples into six different populations.



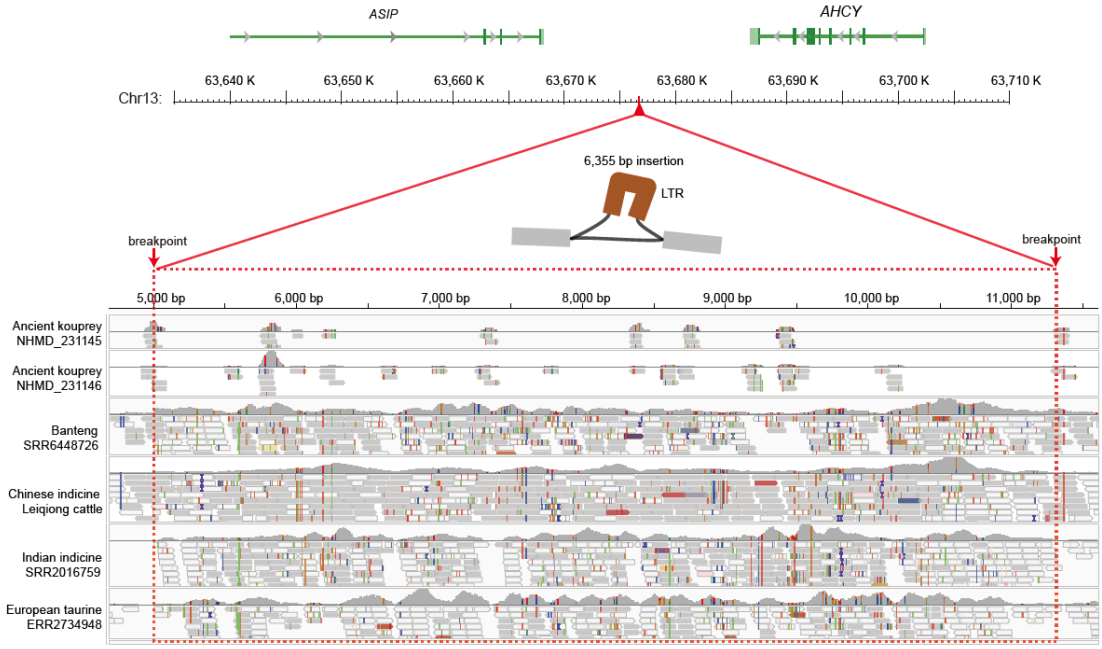
Supplemental Fig. S15 Phylogenetic tree based on deletion and insertion.



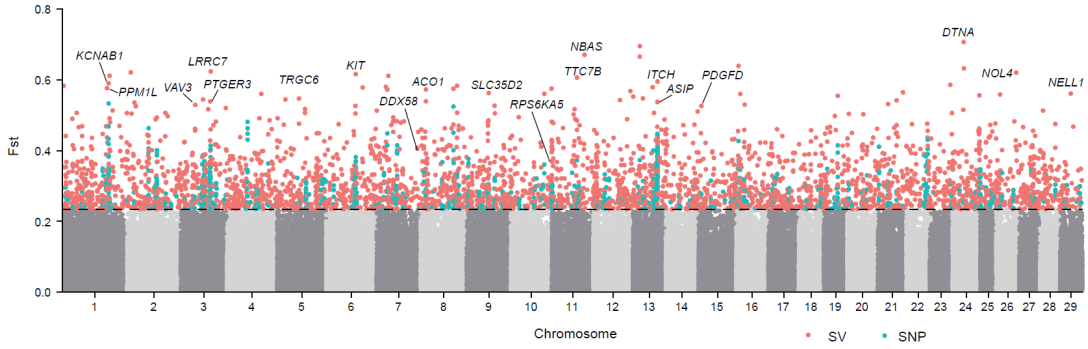
Supplemental Fig. S16 Admixture analysis of deletion and insertion with k = 2 to k = 6.

From	To	Name	From	To	Class
8	295	ERV1-1-LTR_BT	1	289	ERV/ERV1
319	364	ERV1-1B-LTR_BT	331	381	ERV/ERV1
365	5275	ERV1-1-I_BT	1	4975	ERV/ERV1
5276	6115	ERV1-1-LTR_BT	1	837	ERV/ERV1
6139	6184	ERV1-1B-LTR_BT	331	381	ERV/ERV1
6185	6355	ERV1-1-I_BT	4805	4975	ERV/ERV1

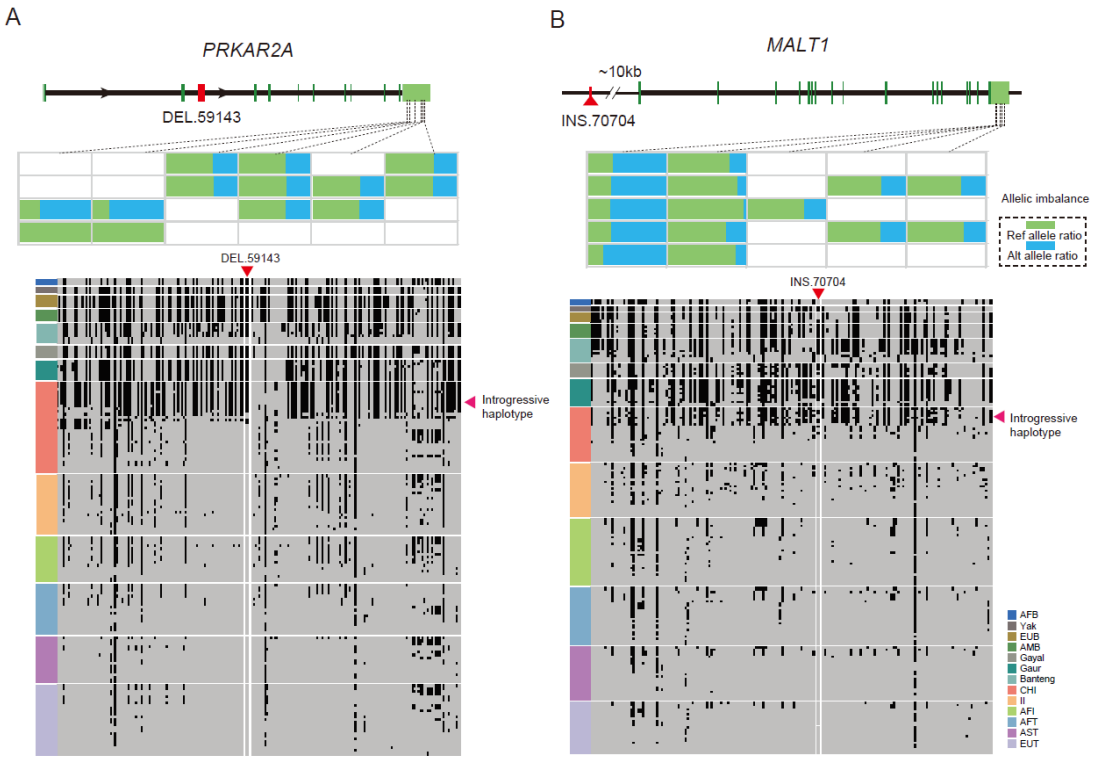
Supplemental Fig. S17 Annotation of the 6.3 kb insertion sequence using the Repbase database.



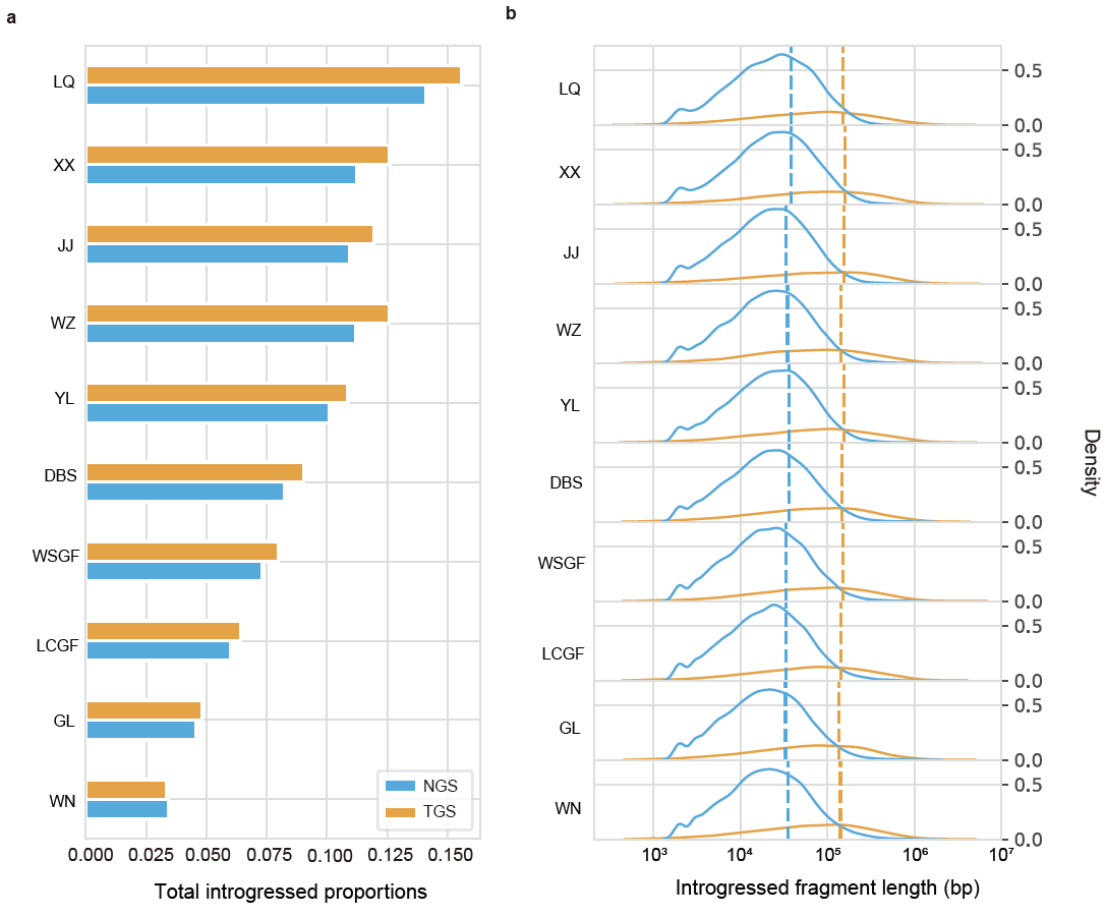
Supplemental Fig. S18 The genomic position of a 6.3 kb insertion downstream of *ASIP* (13:63675338) in the cattle reference genome ARS-UCD1.2 and the IGV screenshots for read alignments at this insertion. According to the presence or absence of spanning reads at the breakpoint, this insertion is shared among ancient kouprey, banteng, and Chinese indicine cattle (Leiqiong cattle).



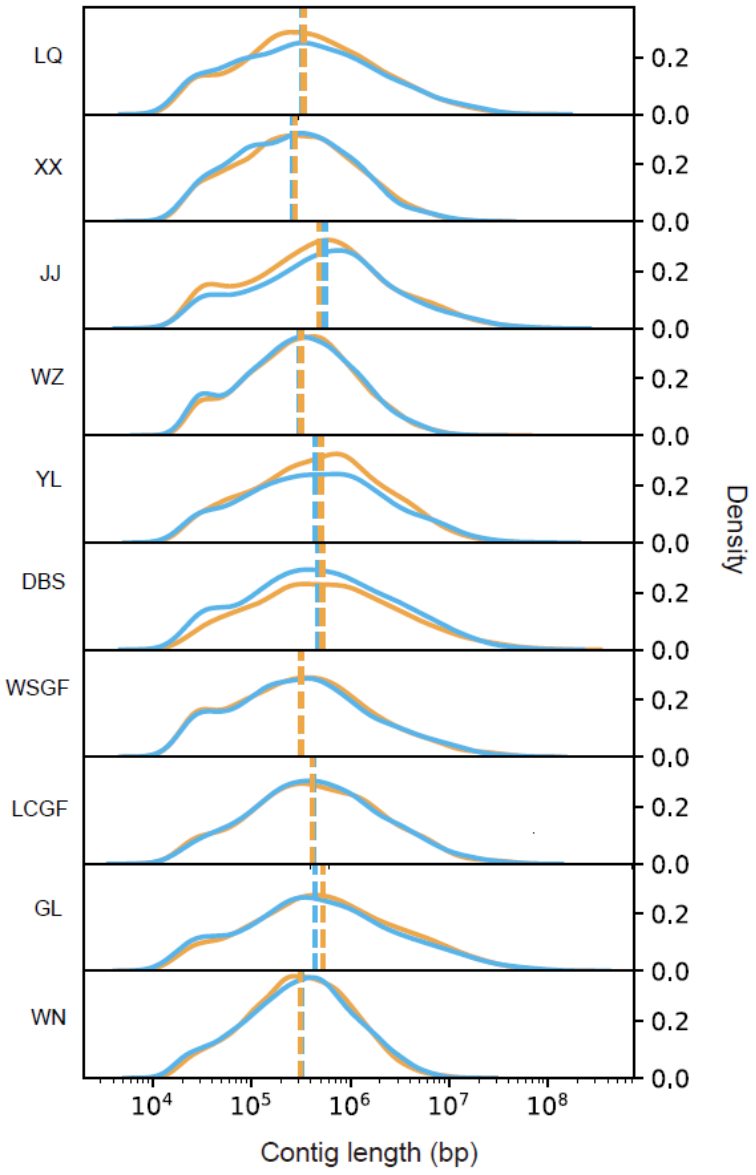
Supplemental Fig. S19 Manhattan plot of 72,049 F_{ST} values of SVs and 99,492 F_{ST} values of sliding windows of SNPs. The horizontal dashed line of SVs (the 1% right tail of the empirical F_{ST} distribution) was identified for Chinese indicine cattle and Indian indicine cattle.



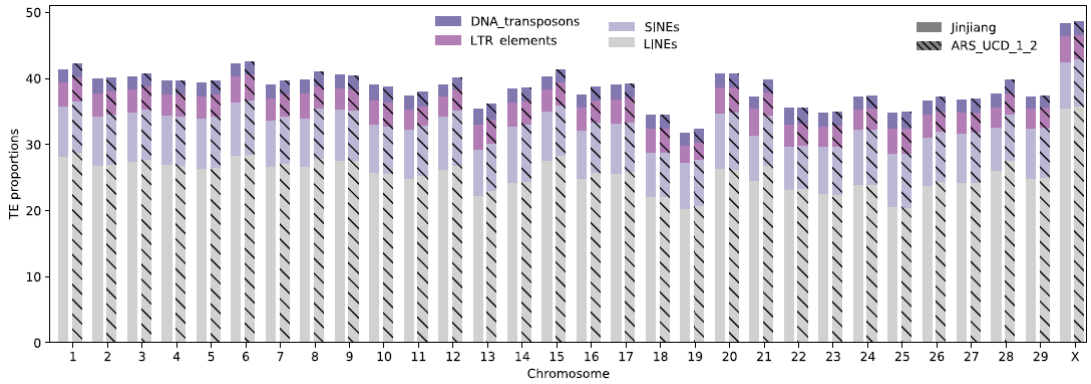
Supplemental Fig. S20 Examples of expression-associated SVs identified by allelic specific expression (ASE) mapping, the allelic imbalance (middle panel), and the haplotype structure (bottom panel) of expression-associated SVs and its flanking SNP in each cattle population. Each column represents one ASE SNP in exons and each row represents one RNA-seq sample. Abbreviations: AFB, African buffalo; EUB, European bison; AMB, American bison; CHI, Chinese indicine; II, Indian indicine; AFI, African indicine; AFT, African taurine; AST, Asian taurine; EUT, European taurine.



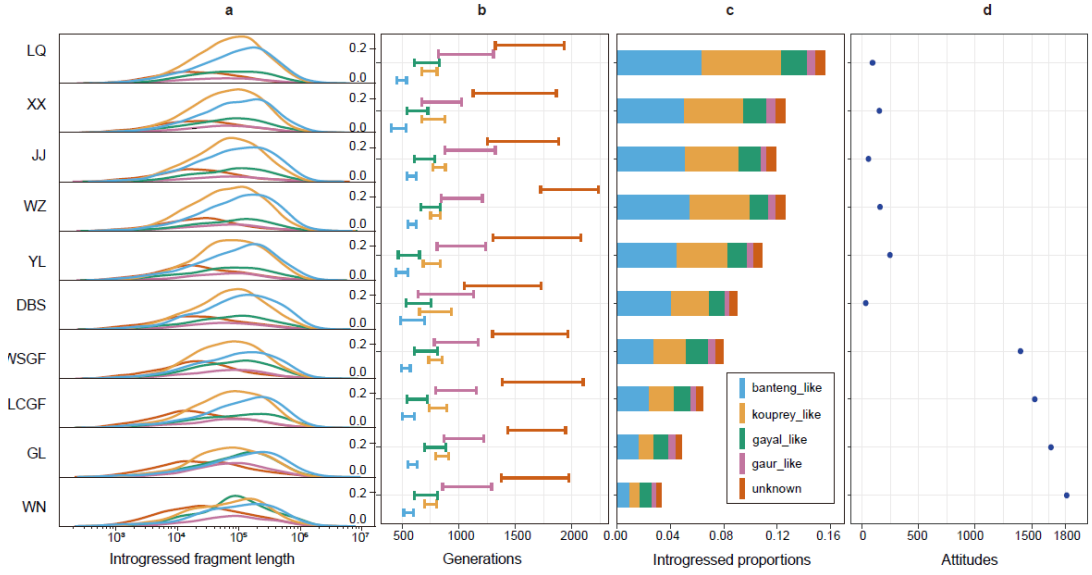
Supplemental Fig. S21 Comparison of introgression-covered proportions (a) and the fragment length distribution (b) between short-read sequencing (NGS) samples and long-read sequencing (TGS) samples. The dashed line indicates the mean fragment lengths. Breed abbreviations: LQ, Leiqiong Cattle; XX, Xiangxi Cattle; JJ, Jinjiang Cattle; WZ, Weizhou Cattle; YL, Yiling Cattle; DBS, Dabieshan Cattle; WSGF, Wenshangaofeng Cattle; LCGF, Lincanggaofeng Cattle; GL, Guanling Cattle; WN, Weining Cattle.



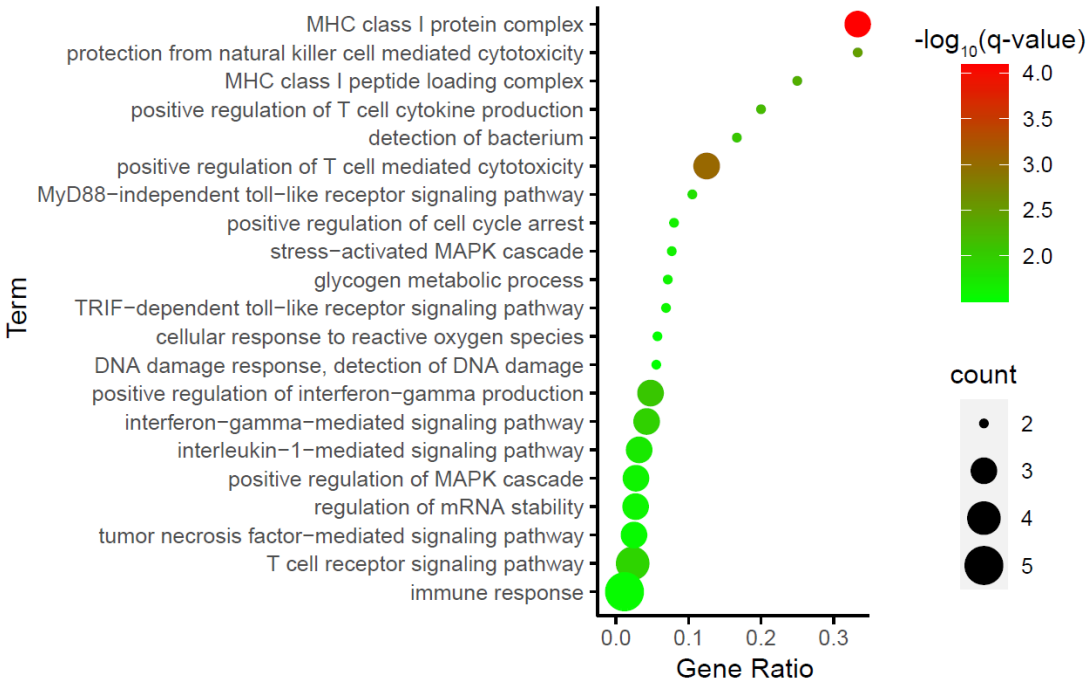
Supplemental Fig. S22 Length distribution of contigs in 20 Chinese indicine haplotype assemblies. The dashed line indicates the median contig lengths of the corresponding haplotype.



Supplemental Fig. S23. The transposable elements content for each chromosome of Jinjiang cattle and ARS-UCD1.2.



Supplemental Fig. S24. (a) Length distribution of introgressed fragments in 10 Chinese indicine cattle, whose introgression-covered proportions are shown in (c). (b) Estimated admixture time intervals with different introgressed sources in 10 Chinese indicine cattle. (d) The habitat altitude of each Chinese indicine cattle.



Supplemental Fig. S25 Gene ontology (GO) enrichment analysis. The significantly enriched GO ($-\log_{10}$ (q-value)) terms of the novel genes in cellular components, molecular function, and biological processes.

Supplemental Tables

Supplemental Table S1 The statistics of sequencing sample.

Supplemental Table S2 The Genome Assembly Statistics.

Supplemental Table S3 Genome accuracy statistics.

Supplemental Table S4 The centromeric sequences statistics.

Supplemental Table S5 The telomeric sequences (TTAGGG) statistics within 10 kb of chromosome end.

Supplemental Table S6 The end-to-end of chromosome statistics for primary

assemblies.

Supplemental Table S7 Number of gap statistics.

Supplemental Table S8 Presence and absence of non-reference sequences in *Bos* species.

Supplemental Table S9 The statistics of bubble breakpoints overlapping with Hereford genome coding sequences.

Supplemental Table S10 The statistics of predicted genes.

Supplemental Table S11 SV hotspots identified in this study.

Supplemental Table S12 The sample information of 321 non-Chinese indicine cattle.

Supplemental Table S13 The archaic fragments detected by hmmix and the probability of ILS.

Supplemental Table S14 The results of admixture history inference.

Supplemental Table S15 Summary information of the cattle breeds.

Supplemental Table S16 Overview of sample information and statistics.

Supplemental Table S17 The frequencies of Chinese indicine cattle population-stratified and introgressed SVs.

Supplemental Table S18 Candidate expression-associated SVs by SV-ASE mapping.

Supplemental Table S19 Candidate expression-associated SVs by eQTL mapping.

Supplemental Table S20 The *D* and related statistics across the combinations of trio population.

Supplemental Table S21 The PacBio HiFi, Illumina HiSeq data of Chinese indicine cattle.

Supplemental Table S22 The *de novo* assemblies data of Chinese indicine cattle.

Supplemental Data

The multi-assembly graph and the VCF file of the SV catalog are available online (https://zenodo.org/record/7607407#.ZAFVs3ZBzQM).

References:

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol* **23**: 258.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**: 573-580.

Browning BL, Zhou Y, Browning SR. 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet* **103**: 338-348.

Bu D, Luo H, Huo P, Wang Z, Zhang S, He Z, Wu Y, Zhao L, Liu J, Guo J. 2021. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic acids research* **49**: W317-W325.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**: 59-60.

Chen N. 2004. Using Repeat Masker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **5**: 4.10. 11-14.10. 14.

Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, Kirsche M, Bentley DR, Schatz MC, Sedlazeck FJ et al. 2019. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol* **20**: 291.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884-i890.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods* **18**: 170-175.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**: 80-92.

Cornish-Bowden A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* **13**: 3021-3030.

Crysnanto D, Leonard AS, Fang Z-H, Pausch H. 2021. Novel functional sequences uncovered through a bovine multiassembly graph. *Proc Natl Acad Sci U S A* **118**: e2101056118.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE,

Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156-2158.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**: giab008.

Dierckxsens N, Li T, Vermeesch JR, Xie Z. 2021. A benchmark of structural variation detection by long reads through a realistic simulated model. *Genome biology* **22**: 1-16.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

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.

Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, Mao Y, Korbel JO, Eichler EE, Zody MC et al. 2022. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat Genet* **54**: 518-525.

Heller D, Vingron M. 2019. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**: 2907-2915.

Hickey G, Heller D, Monlong J, Sibbesen JA, Siren J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B. 2020. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol* **21**: 35.

Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, Human Pangenome Reference C, Marschall T, Li H, Paten B. 2023. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol* doi:10.1038/s41587-023-01793-w.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution* **33**: 1635-1638.

Huerta-Sanchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**: 194-197.

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bahler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* **8**: 14061.

Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. 2020. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* **21**: 189.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**: 356.

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2.

Genome Biol **20**: 278.

Leonard AS, Crysnanto D, Fang Z-H, Heaton MP, Vander Ley BL, Herrera C, Bollwein H, Bickhart DM, Kuhn KL, Smith TP. 2022. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nature Communications* **13**: 1-13.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li H, Feng X, Chu C. 2020. The design and construction of reference pangenome graphs with minigraph. *Genome Biol* **21**: 265.

Malinsky M, Matschiner M, Svardal H. 2021. Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Mol Ecol Resour* **21**: 584-595.

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.

Ni X, Yuan K, Liu C, Feng Q, Tian L, Ma Z, Xu S. 2019. MultiWaver 2.0: modeling discrete and continuous gene flow to reconstruct complex population admixtures. *Eur J Hum Genet* **27**: 133-139.

Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. 2016. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**: 1479-1485.

Paradis E. 2010. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**: 419-420.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature protocols* **11**: 1650-1667.

Rautiainen M, Marschall T. 2020. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol* **21**: 253.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**: 245.

Robinson JT, Thorvaldsdottir H, Turner D, Mesirov JP. 2023. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics* **39**.

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.

Schubert M, Ermini L, Sarkissian CD, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernandez R, Kircher M, McCue M. 2014. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature protocols* **9**: 1056-1082.

Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter

trimming, identification, and read merging. *BMC research notes* **9**: 1-7.

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.

Sinding M-HS, Ciucani MM, Ramos-Madrigal J, Carmagnini A, Rasmussen JA, Feng S, Chen G, Vieira FG, Mattiangeli V, Ganjoo RK. 2021. Kouprey (Bos sauveli) genomes unveil polytomic origin of wild Asian Bos. *Iscience* **24**: 103226.

Skov L, Hui R, Shchur V, Hobolth A, Scally A, Schierup MH, Durbin R. 2018. Detecting archaic introgression using an unadmixed outgroup. *PLoS genetics* **14**: e1007641.

Sotero-Caio CG, Platt RN, 2nd, Suh A, Ray DA. 2017. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol Evol* **9**: 161-177.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**: 637-644.

Talenti A, Powell J, Hemmink JD, Cook EA, Wragg D, Jayaraman S, Paxton E, Ezeasor C, Obishakin E, Agusi E. 2022. A cattle graph genome incorporating global breed diversity. *Nature communications* **13**: 1-14.

Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol Biol Evol* **38**: 3022-3027.

van de Geijn B, McVicker G, Gilad Y, Pritchard JK. 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**: 1061-1063.

Wang F, Shao J, He S, Guo Y, Pan X, Wang Y, Nanaei HA, Chen L, Li R, Xu H et al. 2022. Allele-specific expression and splicing provide insight into the phenotypic differences between thin- and fat-tailed sheep breeds. *J Genet Genomics* **49**: 583-586.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**: e164.

Weng ZQ, Saatchi M, Schnabel RD, Taylor JF, Garrick DJ. 2014. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. *Genet Sel Evol* **46**: 34.

Wu DD, Ding XD, Wang S, Wojcik JM, Zhang Y, Tokarska M, Li Y, Wang MS, Faruque O, Nielsen R et al. 2018. Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nat Ecol Evol* **2**: 1139-1145.

Zhang R, Ni X, Yuan K, Pan Y, Xu S. 2022. MultiWaverX: modeling latent sex-biased admixture history. *Brief Bioinform* **23**.