scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Genome assembly and structural variations of Guyuan cattle

Shuang Liu[1,2,4], Huixuan Yan[3,4], Yuan Liu[1,4], Fen Li[1], Xiaoting Xia[3], Dawei Wei[1], Bei Cai[1], Chuzhao Lei[3] ✉, Ningbo Chen[3] ✉ & Yun Ma[1] ✉

Guyuan cattle, a transitional breed between northern and central Chinese cattle, are known for their climbing ability and disease resistance. However, the genomic resources of indigenous Chinese cattle breeds, including Guyuan cattle, remain largely unexplored. We first assembled the Guyuan cattle genome, named Guyuan_Btau_1.0, using PacBio HiFi sequencing. Guyuan_Btau_1.0, with a total size of 2.86 Gb, was anchored to 30 chromosomes (29 autosomes plus one X), achieving a contig N50 of 85.27 Mb and a scaffold N50 of 107.67 Mb. Additionally, we sequenced the genomes of 10 Guyuan cattle using Oxford Nanopore sequencing, which detected 65,273 structural variations (SVs) with lengths of 40.37 Mb and 4,849 shared SVs with lengths of 3.15 Mb. Across the shared SVs, we identified, visualized, and validated a 1,290 bp deletion in the second intron of *IGF2BP2* gene in the genomes of 10 Guyuan cattle relative to ARS-UCD1.2. Overall, this study enriches the genetic resource database of indigenous Chinese cattle and provides new foundational data for the breeding and genetic improvement of Guyuan cattle.

Background & Summary

Guyuan cattle are distributed mainly in Guyuan City, which is located in the Liupan Mountain area of the southern Ningxia Hui Autonomous Region, China. These cattle are also called as Liupan Mountain cattle and represent one of the important indigenous cattle genetic resources in northwest China. The Guyuan area is located at the intersection of agricultural and nomadic civilizations and serves as an important corridor for genetic exchange between northern and central Chinese cattle populations. As such, Guyuan cattle represent a transitional population between northern and central Chinese cattle[1]. The Guyuan area is also characterized by semi-arid and semi-humid climates, along with abundant grassland resources, which provide a favorable ecological environment for the growth and reproduction of Guyuan cattle. Owing to this unique geographical location and long-term human and natural selection, Guyuan cattle have developed excellent traits, such as climbing ability, roughage tolerance, disease resistance, high reproductive performance, and superior meat quality. These traits make them an essential genetic resource for agriculture and animal husbandry in the Guyuan region.

We previously established a conservation population of Guyuan cattle in their native habitat and conducted whole-genome resequencing of these cattle at high depths. On the basis of SNP analysis, we have revealed the genetic composition of Guyuan cattle, finding that these cattle are taurine × indicine cattle, with 78.09% East Asian taurine and 20.26% East Asian indicine ancestries[2]. The assembly of population-specific genomes is crucial for exploring the genomic variations in Chinese local cattle breeds[3]. Furthermore, structural variations (SVs), including insertions (INSs), deletions (DELs), duplications (DUPs), inversions (INVs), and translocations (TRAs), represent key genomic alterations that profoundly influence phenotypic traits[4]. To further understand the genetic basis of adaptability and important economic traits in Guyuan cattle, it is necessary to construct a high-quality reference genome and conduct a comprehensive analysis of structural variations, thereby providing theoretical foundations and data for the conservation and breeding of Guyuan cattle.

We sequenced the genome of an adult male Guyuan cattle using PacBio HiFi sequencing technology. The base quality value (QV) of the PacBio HiFi data was 70.9979, indicating an accuracy exceeding 99.99%. We first performed a *de novo* genome assembly on the basis of PacBio HiFi reads, resulting in 725 contigs, a contig N50

[1]Key Laboratory of Molecular Cell Breeding of Ruminants in Ningxia Hui Autonomous Region, College of Animal Science and Technology, Ningxia University, Yinchuan, 750021, China. [2]School of Environmental Resources and Life Sciences, Ningxia Normal University, Guyuan, 756000, China. [3]Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling, 712100, China. [4]These authors contributed equally: Shuang Liu, Huixuan Yan, Yuan Liu. ✉e-mail: leichuzhao1118@nwafu.edu.cn; ningbochen@nwafu.edu.cn; mayun@nxu.edu.cn

Statistic	Value
QV of PacBio HiFi reads	70.9979
Number of contigs	725
Contig N50 (bp)	85,271,410 bp
Max contig (bp)	167,227,336 bp
Number of scaffolds	688
Scaffold N50 (bp)	107,667,937
Max scaffold (bp)	168,836,852

Table 1. Statistics of PacBio HiFi reads, contigs, and scaffolds.

of 85,271,410 bp, and a maximum contig of 167,227,336 bp (Table 1). Lengths of the contigs in ARS-UCD1.2 were 2,715,876,358 bp (Fig. 1a). We then performed a reference-guided scaffolding using ARS-UCD1.2 as the reference genome, resulting in 688 scaffolds, a scaffold N50 of 107,667,937 bp, and a maximum scaffold of 168,836,852 bp (Table 1). Finally, we assembled the chromosome-level genome of Guyuan cattle, named Guyuan_Btau_1.0, with a total size of 2,857,459,891 bp, which was anchored to 30 chromosomes (29 autosomes and one X) (Table 2). No gaps were detected across 16 chromosomes (4, 5, 7, 9, 14–16, 18, 19, 22–27, and 29) in Guyuan_Btau_1.0 (Table 2 and Fig. 1a). Lengths of the centromeres and telomeres in Guyuan_Btau_1.0 were 1,195,036 bp and 180,438 bp, respectively (Fig. 1a). In Guyuan_Btau_1.0, RNA transposons, including short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), and long terminal repeats (LTRs), along with DNA transposons, accounted for 38.04%, and total repetitive elements accounted for 47.09% (Fig. 1b). Cumulative lengths of the scaffolds in Guyuan_Btau_1.0 indicated that the assembly had high continuity (Fig. 1c). An analysis of single-copy orthologous genes in mammals revealed that 94.00% of BUSCO genes and 98.67% of Compleasm genes were present in Guyuan_Btau_1.0, indicating that the genome was nearly complete (Fig. 1d). The annotation analysis of the "transplant" from ARS-UCD1.2 to Guyuan_Btau_1.0 revealed that 20,497 out of 20,713 protein-coding genes (98.96%) were successfully transferred (Fig. 1d).

Additionally, we sequenced the genomes of 10 Guyuan cattle using Oxford Nanopore sequencing technology. We detected 65,273 non-redundant and high-quality SVs, including 28,598 INSs (11,870,313 bp), 35,604 DELs (24,769,356 bp), 550 DUPs (1,971,172 bp), 259 INVs (1,761,980 bp), and 262 TRAs, with a total length of 40,372,821 bp (Table 3 and Fig. 2a). Repetitive elements accounted for 66.59%, 61.60%, 46.38%, and 48.28% of INSs, DELs, DUPs, and INVs, respectively (Table 3). Moreover, LINEs and LTRs were significantly enriched in INSs and DELs and not significantly enriched in DUPs, and INVs ($P < 0.05$) (Fig. 2b). The length distribution of total SVs revealed that INS events were generally short, whereas DEL events showed the greatest fluctuation in length distribution, and DUP and INV events maintained relatively constant frequencies for longer lengths (Fig. 2c). If an SV was detected in all individuals, we defined it as a shared SV. A total of 4,849 shared SVs were detected, including 1,916 INSs, 2,585 DELs, 7 DUPs, 43 INVs, and 25 TRAs, which accounted for 6.70%, 8.03%, 1.27%, 16.60%, and 9.54% of all INSs, DELs, DUPs, INVs, and TRAs, respectively (Fig. 2d). Total lengths of the shared SVs was 3,155,832 bp, consisting of 638,926 bp INSs, 1,849,622 bp DELs, 120,862 bp DUPs, and 546,422 bp INVs. 638,926 bp INSs, 1,849,622 bp DELs, 120,862 bp DUPs, and 546,422 bp INVs (Fig. 2d). A total of 1,064 protein-coding genes were overlapping with the shared SVs (Fig. 2e), including 692 genes overlapping with the shared INSs, 1,046 genes overlapping with the shared DELs, 3 genes overlapping with the shared DUPs, and 9 genes overlapping with the shared INVs (Fig. 2e).

Methods

Sample collection and genomic sequence. An adult male Guyuan cattle from Guyuan city (35.34°N, 105.28°E) was collected for genome assembly. DNA was extracted from blood using the CTAB method, followed by purification with the QIAGEN® Genomic Kit (Cat #13343, QIAGEN). PacBio HiFi libraries were prepared using the SMRTbell Express Template Prep Kit 2.1 (Pacific Biosciences, USA), size-selected with BluePippin, and sequenced on the PacBio Sequel II platform.

Additionally, 10 male Guyuan cattle were collected from the same region for detecting SVs (Fig. 3a). Genomic DNA was extracted from blood using the Nanobind CBB Big DNA Kit (Circulomics, USA), following the manufacturer's protocol optimized for Oxford Nanopore sequencing. All procedures were conducted with minimal mechanical stress to preserve DNA integrity. Oxford Nanopore libraries were prepared using the SQK-LSK109 Ligation Kit (Oxford Nanopore Technologies, UK) and sequenced on the PromethION platform with R9.4 flow cells. Base calling was performed using Guppy (v.5.1.13)[5].

Details of the total bases, average read length, and sequencing depth for PacBio HiFi and Oxford Nanopore sequencing data of 11 Guyuan cattle genomes are presented in Table 4.

Genome size estimation and genome assembly. Before genome assembly, the genome size of Guyuan cattle was estimated using the *K*-mer method. Counting *K*-mer and generating the *K*-mer frequency histogram were performed on PacBio HiFi reads with $K = 21$ using Meryl (v1.4.1)[6]. The genome size of Guyuan cattle was estimated as 3.12 Gb based on the *K*-mer ($K = 21$) histogram using GenomeScope (v2.0)[7]. The process of assembling the chromosome-level genome of Guyuan cattle is illustrated in Fig. 3b. First, PacBio HiFi reads were assembled using Hifiasm (v0.13-r308)[8], generating a set of primary contigs and two sets of haplotigs. To reduce redundancy and ensure continuity, only the primary contigs were retained for scaffolding. Next, contigs were anchored to the reference genome ARS-UCD1.2 using RagTag (v2.0.1)[9], which generated chromosome-level

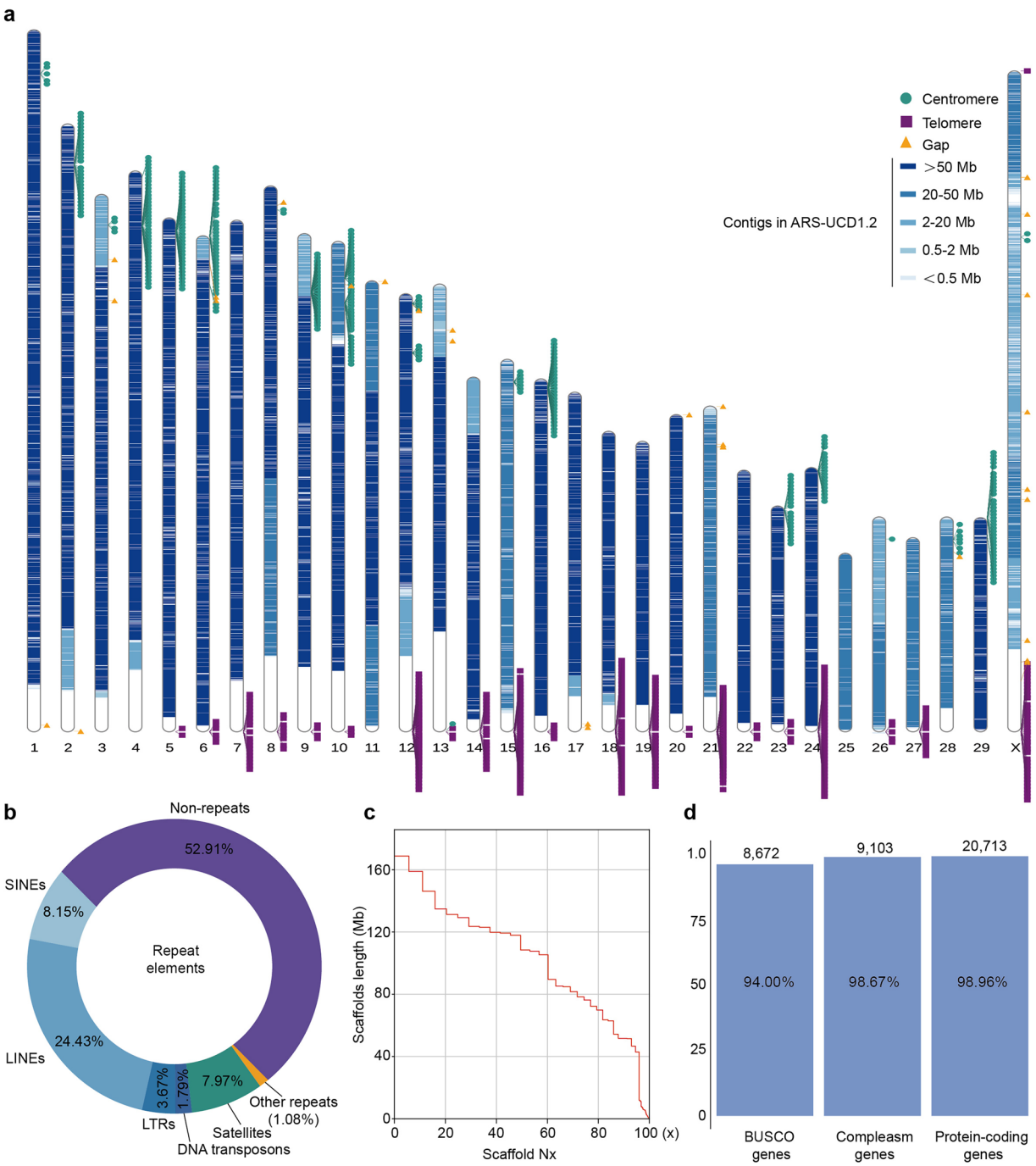

Fig. 1 A global picture of cattle genome assembly Guyuan_Btau_1.0. (**a**) Ideograph of Guyuan_Btau_1.0. (**b**) Composition ratio of repetitive elements in Guyuan_Btau_1.0. (**c**) Cumulative lengths of the scaffolds in Guyuan_Btau_1.0. The *x*-axis indicates a scaffold Nx and the *y*-axis indicates the length of a scaffold Nx. (**d**) Completeness assessment of Guyuan_Btau_1.0.

scaffolds. Finally, a BLASTX search was performed on the scaffolds, which contained 500 gaps, with the non-redundant protein database (NR) using Diamond (v0.9.24)[10]. Sequences from non-*Bovini* species were filtered out, resulting in the final chromosome-level genome assembly of Guyuan cattle. Ideograph of the cattle genome assembly was visualized using the R package Ideogram[11].

Identification of centromeres, telomeres, and gaps. Centromere sequences were identified on the basis of the methodology reported by Melters *et al.*[12]. Centromere sequences were identified through alignment of Guyuan_Btau_1.0 with centromeric satellite DNA reference sequences (M36668.1.fa and J00036.1.fa) using LASTZ (v1.04.19)[13]. To identify telomere sequences, 10 kb sequences at the ends of each chromosome in Guyuan_Btau_1.0 were extracted. Telomere sequences were extracted on the basis of the vertebrate repeat

Chromosome	Length (bp)	Number of gaps (bp)	Length of gaps (bp)
1	168836852	1	100
2	146187200	1	100
3	129237790	2	200
4	134868509	0	0
5	123578907	0	0
6	119274895	2	200
7	123039234	0	0
8	131281974	1	100
9	119798090	0	0
10	117936486	1	100
11	108440026	1	100
12	105351271	1	100
13	107667937	2	200
14	85271410	0	0
15	89559076	0	0
16	84856188	0	0
17	81662325	2	200
18	72254303	0	0
19	69837455	0	0
20	76261616	1	100
21	78319428	4	400
22	62884943	0	0
23	54276144	0	0
24	63536143	0	0
25	42893817	0	0
26	51651957	0	0
27	46642173	0	0
28	51642108	1	100
29	51475784	0	0
X	158935850	11	1100

Table 2. Statistics of 29 autosomes plus one X in Guyuan_Btau_1.0.

Total SVs	INSs	DELs	DUPs	INVs	TRAs
Number	28,598	35,604	550	259	262
Length (bp)	11,870,313	24,769,356	1,971,172	1,761,980	—
Repetitive elements	66.59%	61.60%	46.38%	48.28%	—

Table 3. Statistics of the total SVs detected in Guyuan cattle.

sequence (TTAGGG/CCCTAA) as described by Sharma *et al.*[14]. Gaps in the genome were identified through the extraction of continuous N-base regions from the Guyuan_Btau_1.0 assembly.

Annotation of repeat elements and protein-coding genes. The ruminant database Repbase (v20181026)[15] was used to annotate the repeat elements of Guyuan_Btau_1.0 using RepeatMasker (v4.0.5)[16], and the content of repeat elements in Guyuan_Btau_1.0 was subsequently calculated. The annotation of protein-coding genes was performed using Liftoff (v1.6.3)[17], through mapping the ARS-UCD1.2 annotation file (GCF_002263795.1_ARS-UCD1.2_genomic.gff) onto Guyuan_Btau_1.0. The parameter "--chroms" was used to specify the chromosome correspondence between ARS-UCD1.2 and Guyuan_Btau_1.0.

Quality assessment of genome assembly. Multiple methods were used to assess the quality of Guyuan_Btau_1.0 comprehensively. First, the continuity of the scaffolds in Guyuan_Btau_1.0 was assessed using QUAST (v5.2.0)[18]. Next, a *K*-mer ($K = 21$) database for the PacBio HiFi data was constructed using Meryl (v1.4.1)[6], and the base quality value (QV) was calculated using Merqury (v1.3)[6]. Finally, the completeness of Guyuan_Btau_1.0 was calculated using BUSCO (v5.4.5)[19] and Compleasm (v0.2.6)[20], by assessing its representation in the mammalian single-copy gene database (mammalia_odb10).

SV detection and genotyping. The process of detecting high-quality and non-redundant SVs of Guyuan cattle is illustrated in Fig. 3c. First, reads for Oxford Nanopore sequencing were mapped to ARS-UCD1.2 using NGMLR (v0.2.7)[21], and the parameter "--ont" was set. Next, SVs for each individual were detected using three

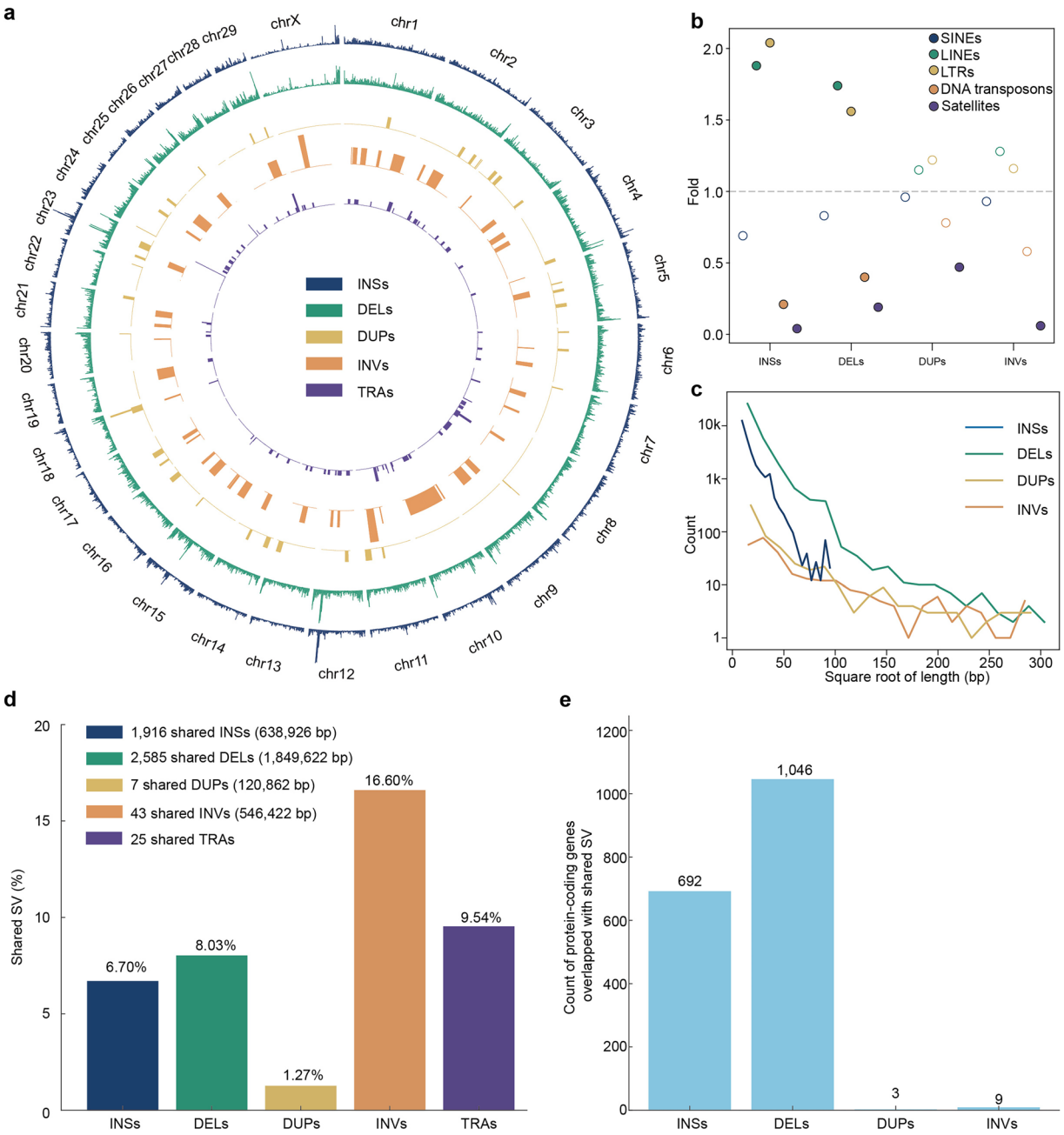

Fig. 2 A global picture of SVs in Guyuan cattle. (**a**) Circular layout of the total SVs distribution in 500-kb windows for each SV type. (**b**) Enrichment of repeat elements in each SV type against the whole genome. Folds were calculated by the proportion of repeat length in each SV type divided by the proportion of total repeat length in the genome. Significance was determined by the Chi-squared Test and adjusted by the Bonferroni methods with threshold adjust $P < 0.05$. (**c**) Size distributions of the total SVs. The *x*-axis is transformed by square root and *y*-axis is transformed by log10. (**d**) Count, ratio, and length of the shared SVs. (**e**) Count of protein-coding genes overlapping with the shared SVs.

methods: SVIM (v1.2.0)[22], CuteSV (v1.2.3)[23], and Sniffles (v2.2)[24]. For merging SVs in each individual, the parameter "--10 2 1 0 0 50" was set in SURVIVOR (v1.0.7)[25] to retain SVs supported by at least two methods. For merging SVs in all individuals, the parameter "--50 1 1 0 0 50" was set in SURVIVOR (v1.0.7)[25] to generate a merged VCF file. Genotypes of the SVs in all individuals of were calculated using Sniffles (v2.2)[24]. Finally, redundant SVs from the merged VCF file were again merged using Truvari (v4.0.0)[26], and filtering out SVs with a genotype of 0/0 in all individuals, resulting in a high-quality and non-redundant SVs. Circular layout of the total SVs distribution in 500-kb windows for each SV type was visualized using Circos[27].

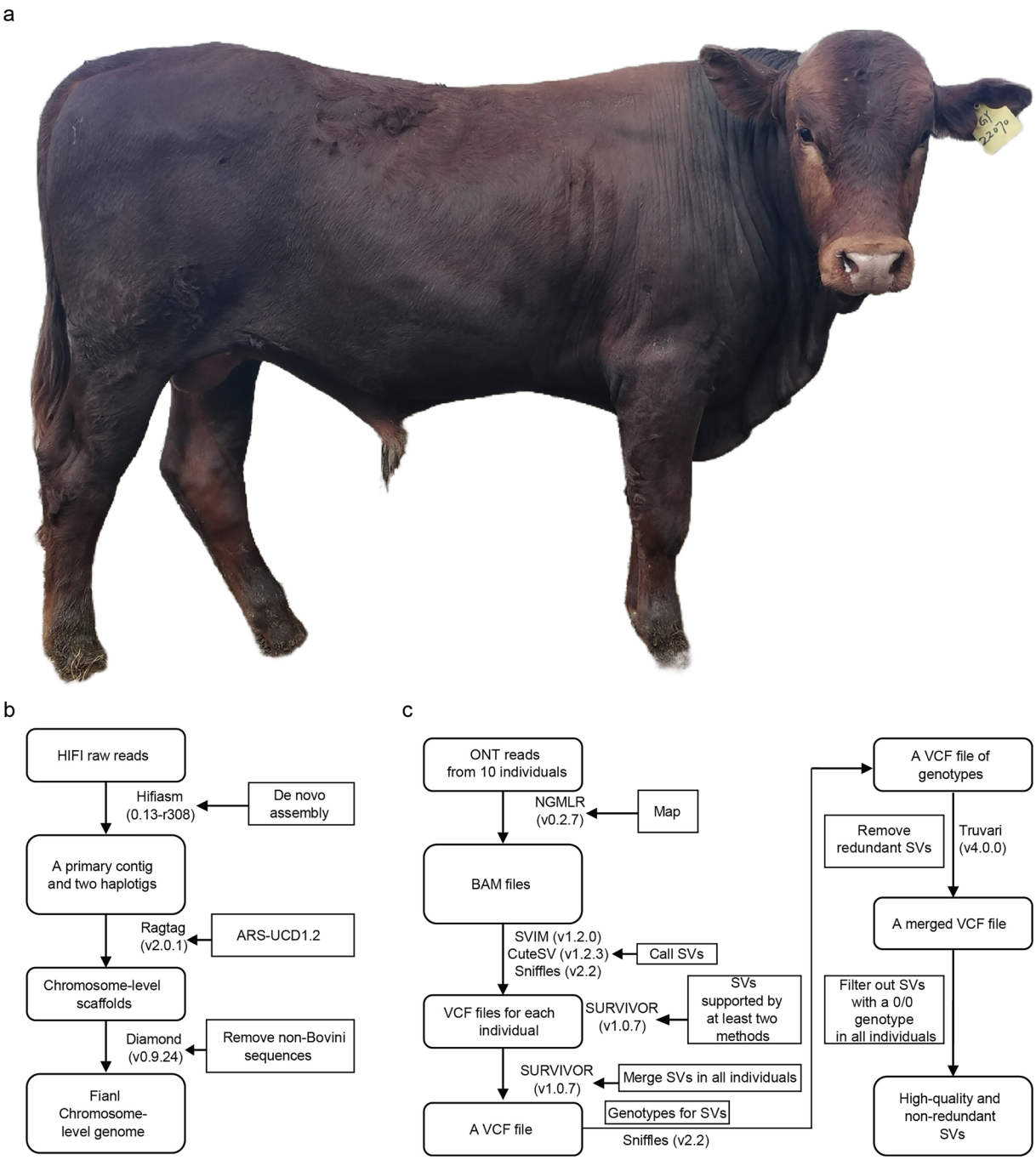

Fig. 3 Process of assembling genome and detecting SVs of Guyuan cattle. (**a**) One of the Guyuan cattle for sequencing. (**b**) Process of assembling the chromosome-level genome of Guyuan cattle. (**c**) Process of detecting non-redundant and high-quality SVs in Guyuan cattle.

Data Records

The raw PacBio HiFi sequencing data for Guyuan_Btau_1.0 is available in the NCBI Sequence Read Archive (SRA) under the SRA ID: SRR33819369[28], which is associated with the BioProject accession number PRJNA1200197. Guyuan_Btau_1.0 has been deposited at NCBI GenBank under the accession number JBKSDG000000000[29], which is also associated with the BioProject accession number PRJNA1200197.

The raw Oxford Nanopore sequencing data for 10 Guyuan cattle is available in the SRA under the SRA IDs: SRR31845301[30], SRR31845302[31], SRR31845303[32], SRR31845304[33], SRR31845305[34], SRR31845306[35], SRR31845307[36], SRR31845308[37], SRR31845309[38] and SRR31845310[39], which is associated with the BioProject accession number PRJNA1200939. All detected SVs from the genomes of 10 Guyuan cattle is available in the European Variation Archive (EVA) under the EVA ID: ERZ28495138, which is associated with the BioProject accession number PRJEB98009[40].

Data	Sample ID	Total bases (Gb)	Average read length (bp)	Sequencing depth
PacBio HiFi	GY22063	91.15	15,864	32.64
Oxford Nanopore	GY22062	31.26	15,038	10.58
Oxford Nanopore	GY22064	30.32	15,384	9.24
Oxford Nanopore	GY22065	30.45	15,724	9.50
Oxford Nanopore	GY22066	34.23	16,509	10.52
Oxford Nanopore	GY22067	34.12	17,788	9.95
Oxford Nanopore	GY22068	34.83	17,792	9.95
Oxford Nanopore	GY22069	31.93	17,876	12.13
Oxford Nanopore	GY22070	39.52	18,530	10.56
Oxford Nanopore	GY22071	32.44	19,857	9.85
Oxford Nanopore	GY22072	32.66	20,519	9.46

Table 4. The detailed information on sequencing data of Guyuan cattle.

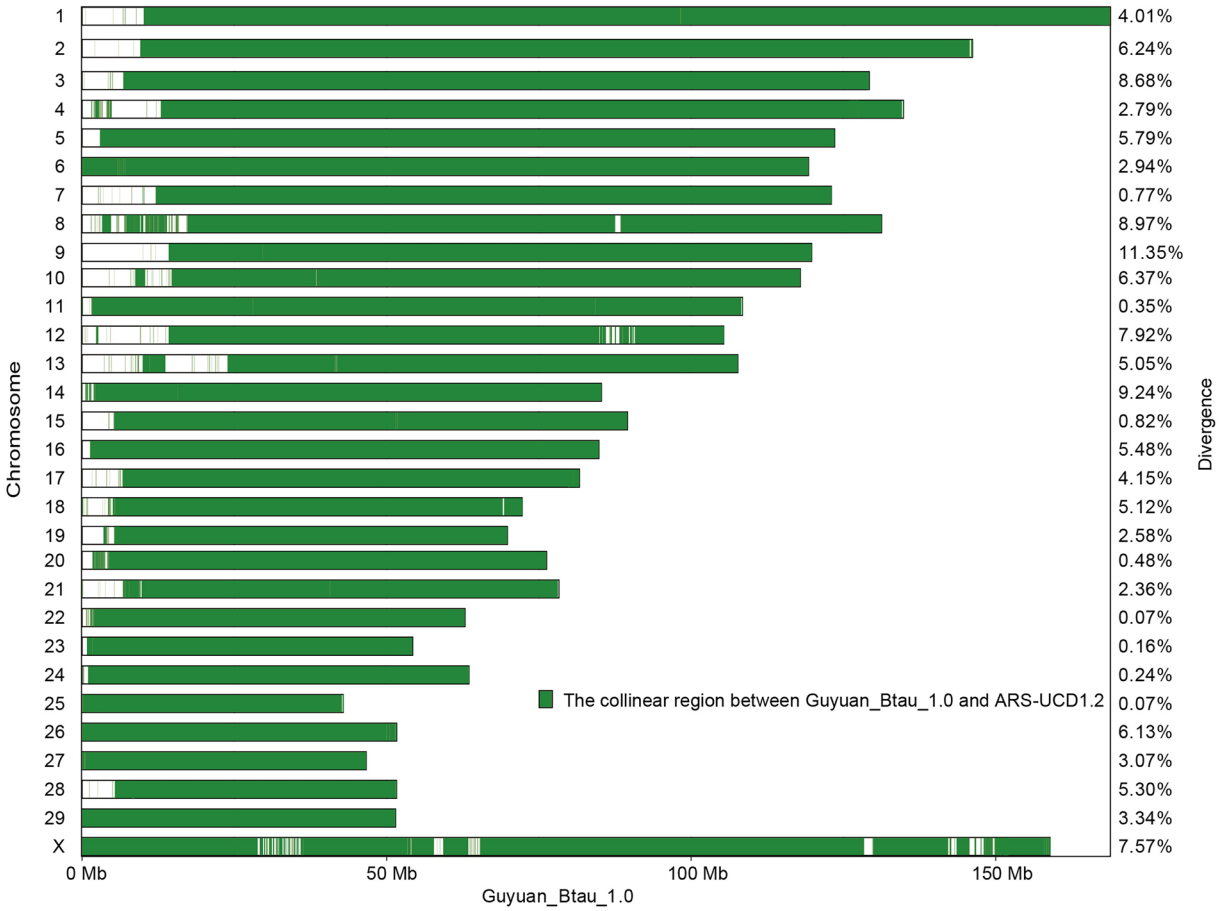

Fig. 4 Genome collinearity between Guyuan_Btau_1.0 and ARS-UCD1.2.

Genome	Genome size (Gb)	Contig N50 (Mb)	Scaffold N50 (Mp)	Number of scaffolds	Repetitive elements
Guyuan_Btau_1.0	2.86	85.27	107.67	688	47.09%
Mongolian_v1 (Genome Biology, 2023)	2.64	47.79	104.05	400	46.60%
Hainan_v1 (Genome Biology, 2023)	2.65	44.35	104.22	301	46.56%
NCBA_BosT1.0 (Nature Communications, 2023)	2.71	74.01	74.72	154	47.30%
ARS-UCD1.2 (GigaScience, 2020)	2.72	12.00	63.00	2,511	45.73%

Table 5. The statistics of Guyuan_Btau_1.0 and recent genomes.

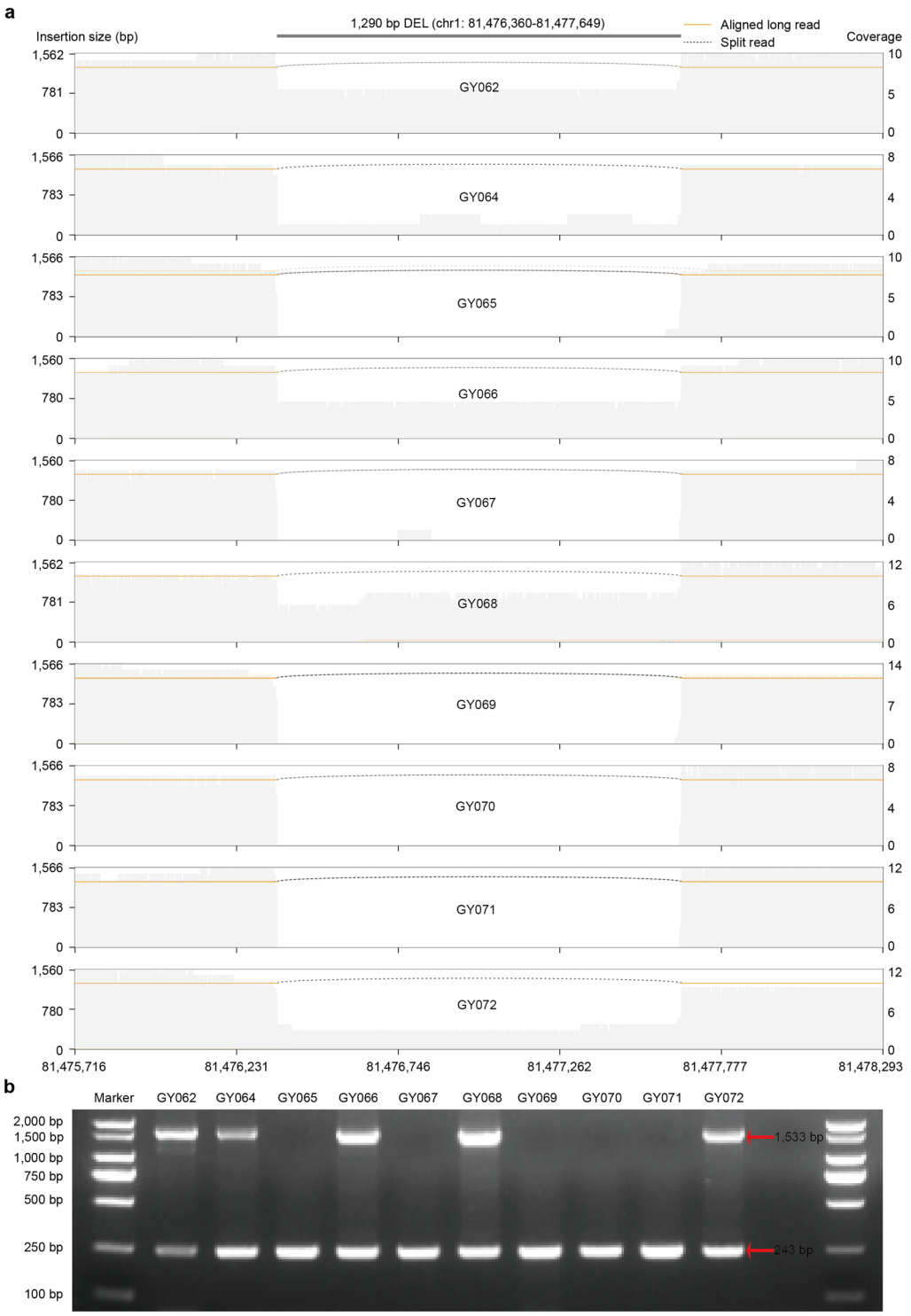

Fig. 5 A 1,290 bp deletion in the second intron of *IGF2BP2* gene, relative to ARS-UCD1.2, was identified, visualized, and validated in the genomes of 10 Guyuan cattle. (**a**) The 1,290 bp deletion was identified and visualized using Samplot based on Oxford Nanopore sequencing data, incorporating coverage depth, aligned long reads, and split reads. (**b**) The 1,290 bp deletion was validated by PCR from genomic DNA using primers flanking the deletion region.

Additionally, all raw sequencing data have been submitted to the National Genomics Data Center (NGDC). The PacBio HiFi sequencing data for Guyuan_Btau_1.0 is available in the NGDC Genome Sequence Archive (GSA) under the GSA ID: CRR1884109[41], which is associated with the BioProject accession number PRJCA041014. The Oxford Nanopore data for 10 Guyuan cattle are available in the GSA under the GSA IDs:

CRR1883663[42], CRR1883664[43], CRR1883665[44], CRR1883666[45], CRR1883667[46], CRR1883668[47], CRR1883669[48], CRR1883670[49], CRR1883671[50], and CRR1883672[51], which is associated with the BioProject accession number PRJCA040984.

The annotation of Guyuan_Btau_1.0 (file: Guyuan_Btau_1.0_protein_coding_genes.gff), and the dataset of 1,064 protein-coding genes overlapping with shared SVs in Guyuan cattle (file: Shared_SVs_overlapping_protein_coding_genes.xlsx) are publicly available in the Figshare database[52].

Technical Validation

First, we identified the collinear sequences between Guyuan_Btau_1.0 and ARS-UCD1.2 using Minimap2[53], and visualized the collinear regions and calculated the mean divergence level of each chromosome sequence using the R package pafr (https://github.com/dwinter/pafr) (Fig. 4c). As a result, extensive collinearity was observed across most chromosomes, with several regions showing structural rearrangements. The mean divergence levels varied among chromosomes, ranging from 0.07% to 11.35%, with chromosomes 9 and 14 exhibiting the highest sequence divergence. Compared with other recent cattle genomes, such as those of Mongolian cattle (Mongolian_v1)[3], Hainan cattle (Hainan_v1)[3], Holstein cattle (NCBA_BosT1.0)[54], and Hereford cattle (ARS-UCD1.2)[55], Guyuan_Btau_1.0 demonstrated excellent assembly quality (Table 5). With the largest genome size (2.86 Gb) and the highest contig N50 (85.27 Mb) and scaffold N50 (107.67 Mb), Guyuan_Btau_1.0 was excellent in both continuity and completeness. It contained 688 scaffolds, more than Mongolian_v1 (400 scaffolds) and Hainan_v1 (301 scaffolds), but fewer than ARS-UCD1.2 (2,511 scaffolds). Composition ratio of repetitive elements in Guyuan_Btau_1.0 was 47.09%, which was comparable to other assemblies and consistent with the repeat-rich nature of the cattle genome. These results suggest that Guyuan_Btau_1.0 is a high-quality reference resource for cattle genomic studies.

Additionally, across the shared SVs, we identified, visualized, and validated a 1,290 bp deletion in the genomes of 10 Guyuan cattle, relative to ARS-UCD1.2. In addition, we identified, visualized, and validated a 1,290 bp deletion in the genomes of 10 Guyuan cattle, relative to ARS-UCD1.2. The 1,290 bp deletion was identified and visualized using Samplot (v1.0.3)[56] based on Oxford Nanopore sequencing data, incorporating coverage depth, aligned long read, and split read (Fig. 5a). The 1,290 bp deletion is located in the region 81,476,360–81,477,649 on chromosome 1, which is located within the second intron of *IGF2BP2* gene in ARS-UCD1.2. Identification and visualization of the 1,289 bp deletion revealed that upstream and downstream regions of the deleted region showed normal and continuous alignment coverage, whereas the deleted region showed a complete absence or markedly reduced aligned long reads and split reads, further confirming the accuracy of the deletion call. Based on identification and visualization of the 1,290 bp deletion, genotype analysis revealed that the frequency of heterozygous genotype (0/1) was 0.5 and homozygous genotype (1/1) was 0.5 among the genomes of 10 Guyuan cattle. The 1,290 deletion was further validated by PCR using genomic DNA and primers flanking the deletion region (Forward primer: 5′-AGCAGGCTTTGTGGTGTATT-3′; Reverse primer: 5′-GGGAGTGTTCCAAATGTAAT-3′), confirming presence of the 1,290 deletion in the genomes of 10 Guyuan cattle (Fig. 5b). The deletion size and individual genotypes validated by PCR were fully consistent with those identified and visualized using Samplot.

Code availability

No specific code was used in this study. The data analyses adhered to the manuals and protocols offered by the creators of the corresponding bioinformatics tools, the parameter settings of which were outlined in the methods section.

Data availability

The raw sequencing data, including PacBio HiFi reads for genome assembly and Oxford Nanopore reads for structural variations of Guyuan cattle, are publicly available. The PacBio HiFi reads are available in the NCBI SRA under accession SRR3381936928, and Guyuan_Btau_1.0 is available in GenBank under accession JBKSDG00000000029. Oxford Nanopore reads from 10 Guyuan cattle are available in the NCBI SRA under accessions SRR31845301-SRR31845310, and the corresponding structural variants are available in the EVA under accession ERZ28495138. The annotation of Guyuan_Btau_1.0 and the dataset of 1,064 protein-coding genes overlapping shared SVs are available in Figshare (https://doi.org/10.6084/m9.figshare.28161233.v1).

Received: 22 January 2025; Accepted: 13 October 2025;
Published online: 25 November 2025

References

1. Liu, S. *et al*. Analysis On Genetic Background and Body Size Indexes and Beef-Purpose Index of Guyuan Cattle. *Acta Veterinaria Et Zootechnica Sinica*. **54**, 2376–2388 (2023).
2. Liu, S. *et al*. Admixture and Selection Offer Insights for the Conservation and Breeding of Guyuan Cattle. *Bmc Biol.* **23**, 128 (2025).
3. Xia, X. *et al*. Structural Variation and Introgression From Wild Populations in East Asian Cattle Genomes Confer Adaptation to Local Environment. *Genome Biol.* **24**, 211 (2023).
4. Dai, X. *et al*. A Chinese Indicine Pangenome Reveals a Wealth of Novel Structural Variants Introgressed From Other Bos Species. *Genome Res.* **33**, 1284–1298 (2023).
5. Sherathiya, V. N., Schaid, M. D., Seiler, J. L., Lopez, G. C. & Lerner, T. N. Guppy, a Python Toolbox for the Analysis of Fiber Photometry Data. *Sci. Rep.* **11**, 24212 (2021).
6. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies. *Genome Biol.* **21**, 245 (2020).
7. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. Genomescope 2.0 and Smudgeplot for Reference-Free Profiling of Polyploid Genomes. *Nat. Commun.* **11**, 1432 (2020).

8. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable Telomere-to-Telomere Assembly for Diploid and Polyploid Genomes with Double Graph. *Nat. Methods.* **21**, 967–970 (2024).
9. Alonge, M. *et al*. Automated Assembly Scaffolding Using Ragtag Elevates a New Tomato System for High-Throughput Genome Editing. *Genome Biol.* **23**, 258 (2022).
10. Buchfink, B., Xie, C. & Huson, D. H. Fast and Sensitive Protein Alignment Using Diamond. *Nat. Methods.* **12**, 59–60 (2015).
11. Hao, Z. *et al*. Rideogram: Drawing Svg Graphics to Visualize and Map Genome-Wide Data On the Idiograms. *Peerj Comput. Sci.* **6**, e251 (2020).
12. Melters, D. P. *et al*. Comparative Analysis of Tandem Repeats From Hundreds of Species Reveals Unique Insights Into Centromere Evolution. *Genome Biol.* **14**, R10 (2013).
13. Harris, R. S. *Improved Pairwise Alignment of Genomic Dna*, The Pennsylvania State University, 2007).
14. Sharma, V. R. & Sheardy, R. D. The Human Telomere Sequence, (Ttaggg)4, in the Absence and Presence of Cosolutes: A Spectroscopic Investigation. *Molecules.* **19**, 595–608 (2014).
15. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes. *Mob. Dna.* **6**, 11 (2015).
16. Tarailo-Graovac, M. & Chen, N. Using Repeatmasker to Identify Repetitive Elements in Genomic Sequences. *Curr Protoc Bioinformatics.* Chapter 4, 4–10 (2009).
17. Shumate, A. & Salzberg, S. L. Liftoff: Accurate Mapping of Gene Annotations. *Bioinformatics.* **37**, 1639–1643 (2021).
18. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. Quast: Quality Assessment Tool for Genome Assemblies. *Bioinformatics.* **29**, 1072–1075 (2013).
19. Seppey, M., Manni, M. & Zdobnov, E. M. Busco: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol.* **1962**, 227–245 (2019).
20. Huang, N. & Li, H. Compleasm: A Faster and More Accurate Reimplementation of Busco. *Bioinformatics*. **39** (2023).
21. Sedlazeck, F. J. *et al*. Accurate Detection of Complex Structural Variations Using Single-Molecule Sequencing. *Nat. Methods.* **15**, 461–468 (2018).
22. Heller, D. & Vingron, M. Svim: Structural Variant Identification Using Mapped Long Reads. *Bioinformatics.* **35**, 2907–2915 (2019).
23. Jiang, T., Liu, S., Cao, S. & Wang, Y. Structural Variant Detection From Long-Read Sequencing Data with Cutesv. *Methods Mol Biol.* **2493**, 137–151 (2022).
24. Smolka, M. *et al*. Detection of Mosaic and Population-Level Structural Variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).
25. Jeffares, D. C. *et al*. Transient Structural Variations Have Strong Effects On Quantitative Traits and Reproductive Isolation in Fission Yeast. *Nat. Commun.* **8**, 14061 (2017).
26. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: Refined Structural Variant Comparison Preserves Allelic Diversity. *Genome Biol.* **23**, 271 (2022).
27. Krzywinski, M. *et al*. Circos: An Information Aesthetic for Comparative Genomics. *Genome Res.* **19**, 1639–1645 (2009).
28. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33819369 (2025).
29. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBKSDG000000000 (2025).
30. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31845301 (2025).
31. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31845302 (2025).
32. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31845303 (2025).
33. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31845304 (2025).
34. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31845305 (2025).
35. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31845306 (2025).
36. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31845307 (2025).
37. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31845308 (2025).
38. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31845309 (2025).
39. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31845310 (2025).
40. *EMBL-EBI EVA* https://identifiers.org/ena.embl:ERP180447 (2025).
41. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRR1884109 (2025).
42. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRR1883663 (2025).
43. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRR1883664 (2025).
44. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRR1883665 (2025).
45. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRR1883666 (2025).
46. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRR1883667 (2025).
47. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRR1883668 (2025).
48. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRR1883669 (2025).
49. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRR1883670 (2025).
50. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRR1883671 (2025).
51. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRR1883672 (2025).
52. Liu, S. & Ma, Y. Genome Assembly and Structural Variations of Guyuan Cattle. *Figshare* https://doi.org/10.6084/m9.figshare.28161233.v1 (2025).
53. Li, H. New Strategies to Improve Minimap2 Alignment Accuracy. *Bioinformatics.* **37**, 4572–4574 (2021).
54. Li, T. T. *et al*. De Novo Genome Assembly Depicts the Immune Genomic Characteristics of Cattle. *Nat. Commun.* **14**, 6601 (2023).
55. Rosen, B. D. et al. De Novo Assembly of the Cattle Reference Genome with Single-Molecule Sequencing. *Gigascience*. **9** (2020).
56. Belyeu, J. R. *et al*. Samplot: A Platform for Structural Variant Visual Validation and Automated Filtering. *Genome Biol.* **22**, 161 (2021).

Acknowledgements

This work was supported by the National Natural Science Foundation of China [U22A20506, 32472871, and 32341054], the Key Research and Development Program of Ningxia Projects [2024BBF01007 and 2023BCF01006], and the earmarked fund for the China Agriculture Research System of MOF and MARA [CARS-37]. We thank the Key Laboratory of Molecular Cell Breeding of Ruminants in Ningxia Hui Autonomous Region for providing computing resources. We thank the High-Performance Computing (HPC) of Northwest A&F University (NWAFU) for providing computing resources.

Author contributions

C.-Z.L., N.-B.C. and Y.M. conceived and designed the study. S.L., Y.L. and F.L. collected the data. S.L., H.-Y.Y., Y.L., F.L., X.-T.X., D.-W.W., B.C. and N.-B.C. performed bioinformatics analysis, and visualized the results. S.L., H.-Y.Y. and Y.L. wrote the original draft. C.-Z.L., N.-B.C. and Y.M. reviewed and edited the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to C.L., N.C. or Y.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.