scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Genome assembly and whole-genome resequencing study of Butuo Black sheep (*Ovis aries*)

Changsheng Zhong[1,4], Lianting Zhang[2,4], Weijia Song[1,4], Yingang Guo[1], Jinwang Liu[1], Zhangjia Baqian[3], Jinkang Wang[3], Ran Li[2] ✉ & Jianmin Su[1] ✉

Butuo Black sheep (BBS), an ancient indigenous Chinese breed, has co-evolved with the Yi people's semi-nomadic lifestyle and demonstrated exceptional adaptability to high-altitude migrations. However, a high-quality reference genome for BBS is still lacking. In this study, we established a high-quality chromosome-level genome assembly of BBS using PacBio HiFi sequencing. The final assembled genome size was approximately 2.95 Gb, with a contig N50 of 71.45 Mb and a scaffold N50 of 92.26 Mb. The genome assembly achieved a high Benchmarking Universal Single-Copy Orthologs (BUSCO) score of 95.9%, indicating its high completeness and quality. The de novo genome prediction revealed that repetitive sequences accounted for 47.74% of the genome, with long interspersed nuclear elements (LINEs) being the most abundant. Additionally, we present 50 BBS shotgun genomes sequenced using the Illumina HiSeq 2000 platform, with a mean coverage of 10.36×. The study generated approximately 1.2 terabytes of raw data, with 99.9% clean reads mapping successfully to the sheep reference genome at 99.6% coverage. This extensive dataset provides a valuable resource for studying genetic diversity and evolutionary patterns in BBS.

Background & Summary

Sheep (*Ovis aries*), as one of the world's most important livestock species, provide critical primary products like meat, milk, and wool, alongside useful secondary by-products[1]. As a Chinese indigenous breed adapted to Sichuan's high-mountain environments, Butuo Black sheep (BBS) have undergone millennia of natural and artificial selection, thereby developing specialized adaptations, including hypoxia tolerance, efficient energy metabolism, and cold resistance[2]. Although there have been several population genetic studies on BBS[2–5], a high-quality reference genome is still lacking, which hinders the in-depth understanding of the genetic basis and molecular breeding of this precious breed. Recent whole - genome studies in yaks (*Bos grunniens*) have demonstrated that combined natural and artificial selection promote the adaptation of livestock to harsh environmental conditions and identify relevant candidate genes[6]. As another livestock species adapted to southwestern China's high-mountain regions, whether BBS has undergone similar genomic selection pressures remains unclear, necessitating high-quality genomic data for validation.

Chromosome-level genome assembly and whole-genome resequencing to analyze genetic variation can provide fundamental data for uncovering the genetic diversity and adaptive evolution mechanisms of BBS, while also laying a foundation for future research aimed at decoding its genetic structure and environmental adaptation pathways. De novo genome assembly is a cornerstone of modern molecular research, enabling comprehensive exploration of genetic architecture and evolution[7]. To date, high-quality de novo assemblies of East Friesian sheep[8], Tibetan sheep[9], Rambouillet sheep[10], and Guide Black-Fur sheep[11] are publicly available. In this study, we constructed a chromosome-level genome assembly of BBS using approximately 106 Gb PacBio HiFi reads (Table S1; Figure S1), yielding a 2.95 Gb genome with 3,731 contigs (N50 = 71.45 Mb) and 3,610 scaffolds (N50 = 92.26 Mb) (Table 1). In addition, we annotated 29,169 protein-coding genes in this assembly, with an

[1]Key Laboratory of Livestock Biology, College of Veterinary Medicine, Northwest A&F University, Yangling, Shaanxi Province, 712100, China. [2]Key Laboratory of Livestock Biology, College of Animal Science and Technology, Northwest A&F University, Yangling, Shaanxi Province, 712100, China. [3]Butuo County Agriculture and rural Bureau, Liangshan Yi Autonomous Prefecture, Xichang, Sichuan Province, 615014, China. [4]These authors contributed equally: Changsheng Zhong, Lianting Zhang, Weijia Song. ✉e-mail: ran.li1986@hotmail.com; sujm@nwafu.edu.cn

Project	BBS Assembly Genome
Total length (bp)	2,950,585,589
Number of scaffolds	3610
N50 of scaffolds (bp)	92,258,836
Chromosome-scale scaffolds (bp)	2,639,684,734 (89.5%)
Number of contigs	3731
N50 of contigs (bp)	71,454,981
GC (%)	43.8
QV	68.2974
Busco evaluation	C:95.9%[S:93.5%,D:2.4%],F:1.1%,M:3.0%,n:9226

Table 1. Summary of the BBS genome assembly.

Class	Subclass	Number	Total length(bp)	% of genome
SINEs		2,035,566	297,150,218	11.26
	Alu/B1	0	0	0.00
	MIRs	390,638	55,872,612	2.12
LINEs		1,327,978	746,221,029	28.27
	LINE1	593,382	341,401,259	12.93
	LINE2	247,290	61,858,732	2.34
	L3/CR1	32,970	6,805,215	0.26
	RTE	453,261	335,992,918	12.73
LTR elements		410,413	125,945,621	4.77
	ERVL	75,864	29,223,068	1.11
	ERVL-MaLRs	122,651	39,590,106	1.50
	ERV_classI	83,712	35,755,823	1.35
	ERV_classII	110,706	17,308,681	0.66
DNA elements		290,203	57,228,587	2.17
	hAT-Charlie	163,790	30,350,574	1.15
	TcMar-Tigger	44,672	11,793,118	0.45
Unclassified		5,243	883,081	0.03
Total interspersed repeats			1,227,428,536	46.50
Small RNA		252,609	42,564,441	1.61
Satellites		3,000	5,182,868	0.20
Simple repeats		533,353	22,872,627	0.87
Low complexity		84,678	4,169,724	0.16

Table 2. Statistics of repetitive elements.

average of 29.1 exons per gene. The repetitive elements comprise 47.74% of the genome—predominantly LINEs (28.27%; Table 2). This resource supports sheep breeding and future genetic mechanism exploration.

This study also presents whole-genome resequencing data of 50 BBS sampled from Sichuan, China, yielding more than 45 million single-nucleotide polymorphisms (SNPs) aligned to the sheep reference genome ARS-UI_Ramb_v2.0 (GCF_016772045.1)[10]. High-quality metrics (Q20 >95%, mapping rate >99%) validate the accuracy and resolution of this dataset, which supports diverse applications: detecting positive selection signals, inferring demographic history, analyzing gene flow, identifying candidate genes for economic traits, evaluating local adaptation, discovering breed-specific variants, assessing genetic diversity, and developing SNP arrays for breeding and species identification[12]. Thus, these data serve as an ideal resource for population genetics, evolutionary biology, and molecular breeding of high-altitude sheep.

Methods

Ethics statement. All experimental procedures and animal use were approved by the Institutional Animal Care and Use Committee at Northwest A&F University (approval no. IACUC2024-0419).

Chromosome-level genome assembly of Butuo Black sheep (BBS). *Sample collection and sequencing.* A female BBS was randomly selected from Liangshan, Sichuan Province, China (27°43′N, 102°49′E) for assembly. The assembled sequence does not include the Y chromosome due to sampling from females. The animal and its parents were confirmed healthy with no observed genetic defects. Genomic DNA was extracted from blood using the TIANamp Genomic DNA kit (Tiangen, Beijing, China), following the manufacturer's protocol. DNA purity and concentration were evaluated using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific,

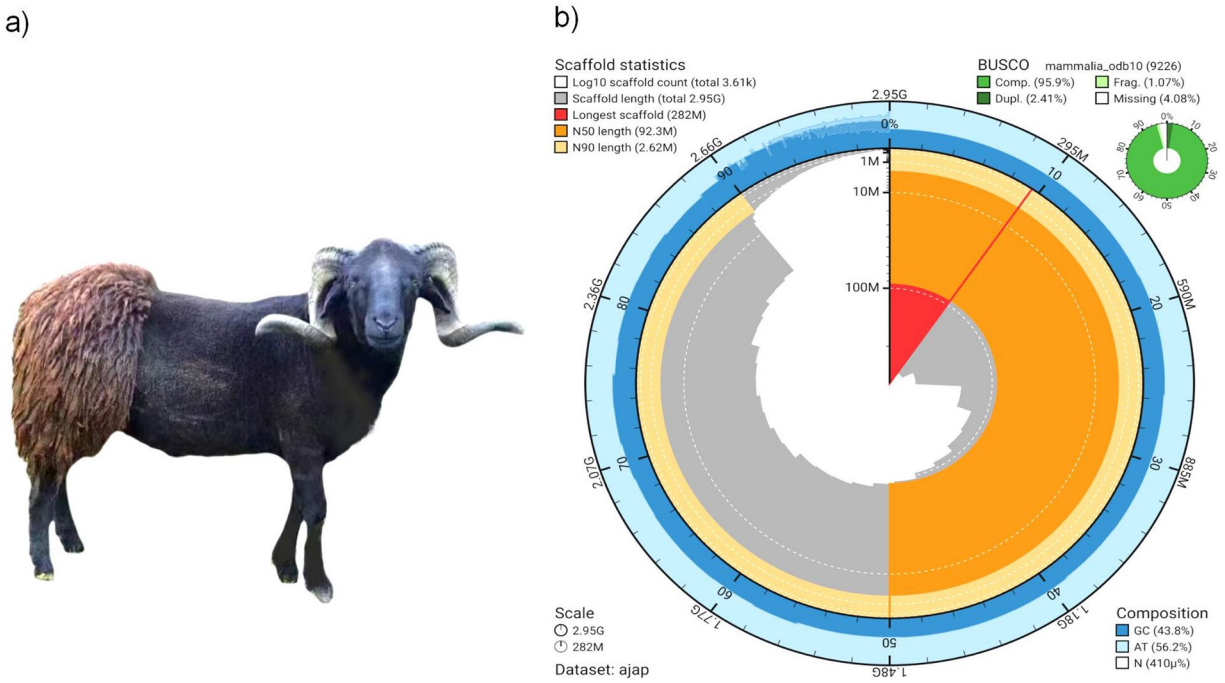

Fig. 1 Snail plot depicting the assembly quality of the Butuo Black sheep genome. **a**) Butuo Black sheep; **b**) snail plots of BBS genome features.

Waltham, USA) and Qubit 4.0 fluorometer (Thermo Fisher Scientific), with absorbance ratios (A_{260}/A_{280}) and fluorescence-based concentration measurements recorded for each sample.

Long-read genome sequencing was performed using the PacBio Sequel II system (Pacific Biosciences, California, USA)[13]. PacBio libraries were constructed using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, California, USA) to generate libraries with an average insert size of ~15 kb following the manufacturer's protocol. Library fragments were fractionated into narrow size ranges using SageELF (Sage Science, Beverly, MA, USA). Each library was sequenced on the Sequel II platform (PacBio) with 2–3 8 M SMRT cells. Raw subreads were processed using the circular consensus sequencing (CCS) algorithm (SMRTLink v6.0.0)[14] with parameters: --minPasses 3 --minPredictedAccuracy 0.99 --maxLength 21,000, yielding 106 Gb of PacBio long high-fidelity (HiFi) reads (40.68 × coverage). After adapter trimming, clean PacBio long-reads with an N50 read length of 16.238 kb were obtained (Table S1; Figure S1).

Genome assembly and quality assessment. Contig-level assembly was performed using Hifiasm (v0.19.7)[15] with default parameters. Hifiasm is capable of generating a primary haplotype-collapsed, contig-level assembly and two partially phased contig-level assemblies based solely on an individual's HiFi reads. Contig sequences were extracted from gfa files generated by Hifiasm. RagTag[16] is a tool that is particularly useful for optimizing assemblies in the absence of genetic maps. It was designed for chromosome-level assembly by anchoring genome contigs to a reference genome. RagTag (v2.0.1) was employed to scaffold the primary assembly contigs of BBS to the chromosomal level using the reference genome ARS-UI_Ramb_v2.0 (GCF_016772045.1)[10]. The final assembly yielded a 2.95 Gb genome with a scaffold N50 of 92.26 Mb, where 89.46% of scaffolds were anchored to 27 chromosomes. A summary of the BBS genome assembly is shown in Table 1. Genomic feature maps were visualized using Snail Plots via the Blobtoolkit (v4.4.5)[17] package (Fig. 1).

The quality value (QV), a Phred-scale metric quantifying base-calling accuracy, is essential for assessing the precision of nucleotide identification in high-throughput sequencing pipelines[18]. In this study, the yak tool (v0.1-r56, https://github.com/lh3/yak) was employed to compute the QV, yielding a final value of 68.29, indicating a low error rate of approximately 15 errors per 100 million bases. The benchmarking universal single-copy orthologs (BUSCO v5.4.7)[19] analysis was performed to assess genome completeness, based on single-copy orthologs from the mammalia odb10 database[20]. Of the 9226 mammalian BUSCO groups (mammalia_odb10), 8,850 (95.9%) were fully identified in the assembled genome (Figure S2). Genome module benchmarking yielded the following: C = 95.9% (complete BUSCOs, including S = 93.5% single-copy and D = 2.4% duplicated), F = 1.1% (fragmented), M = 3.0% (missing).

To resolve residual gaps in the chromosomal scaffolds, we utilized quarTeT (v1.1.3)[21] for gap filling, leveraging paired-end reads to extend contigs across intergenic regions. The specific parameters used were "-GapFiller -g *fasta -t 30 -l 5000 -i 60". The distribution of interstitial gaps within each chromosome is visualized in Fig. 2, highlighting the effectiveness of this approach in improving assembly continuity.

Identification of telomeres and centromeres. Telomeric sequences in the BBS genome assembly were mapped using quarTeT (v1.0.3)[21] with the '-c animal' parameter, leveraging its TeloExplorer module specialized for

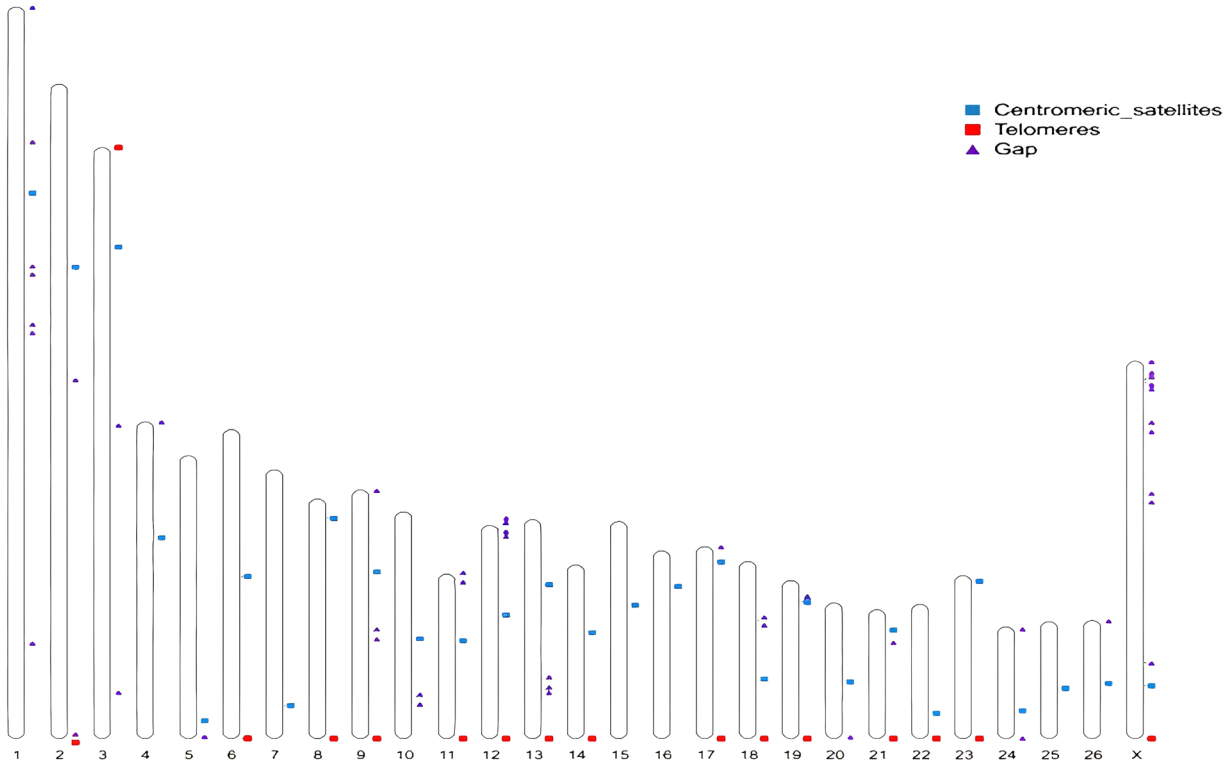

Fig. 2 Distribution map of centromeres, telomeres, and gaps on the assembled Butuo Black sheep genome.

telomere prediction. The quarTeT pipeline integrates four functional components: AssemblyMapper for scaffold alignment, GapFiller for intergenic extension, TeloExplorer for telomere repeat detection, and CentroMiner for centromeric region annotation. The canonical animal telomere repeat monomer "TTAGGG/CCCTAA" was identified as the primary telomeric unit. For centromere annotation, the Centromics toolkit (https://github.com/ShuaiNIEgithub/Centromics) was applied to detect repetitive sequence clusters and histone modification signatures. This analysis identified 16 telomeres (one per chromosome end) and 27 centromeres corresponding to the diploid chromosome number, with detailed statistics provided in Table S2.

Annotation of repetitive sequences and gene annotation. RepeatMasker[22], a widely used tool for repetitive sequence detection, accurately identifies and masks repetitive elements in genomes through similarity alignment with reference databases, representing a homology-based prediction and annotation approach. Using RepeatMasker (v4.1.5) in conjunction with the Repbase database[23], we conducted detection of DNA transposable elements (TEs). As shown in Table 2, dispersed repeats constituted 46.5% of the total length of the BBS genome, amounting to approximately 1,227,428,536 base pairs (bp). Among these repeats, retrotransposons accounted for 872,166,650 bp (33.0%), while DNA transposons represented 57,228,587 bp (2.18%). Notably, long interspersed nuclear elements (LINEs) comprised 28.3% of the Butuo sheep genome sequences, making them the most abundant repetitive sequence class. Long Terminal Repeats (LTRs) constituted 4.77% of the genome, and small interspersed nuclear elements (SINEs) represented 11.3%.

The Liftoff software (v1.6.3)[24] was utilized for annotating protein-coding genes in the BBS genome and their corresponding mRNAs. The annotation pipeline identified 29,169 protein-coding genes, yielding 71,601 RNA transcripts composed of 851,089 exons and 751,841 coding sequences (CDS) (Figure S3). These annotations were systematically validated against orthologous gene models to ensure accuracy and completeness. Our assembly displayed comparable completeness, continuity, and base accuracy compared to other public sheep genome assemblies (GCF_016772045.1; GCA_033439445.1; GCA_040259355.1) with higher sequencing depth (Figure S4; Table S3).

Whole genome sequences of 50 Butuo Black sheep. *Sample collection, DNA extraction, and quality control.* Genomic DNA was extracted from 50 BBS blood samples using the phenol–chloroform method[25]. Rigorous QC was performed: DNA degradation and contamination were assessed on 1% agarose gels[26], the A_{260}/A_{280} ratio was determined with a NanoDrop UV spectrometer to assess DNA purity, and the DNA concentration was measured with a Qubit® DNA assay kit on a Qubit® 3.0 fluorometer (Invitrogen, USA). Only DNA fragments over 0.2 μg, without degradation or contamination and with an A_{260}/A_{280} ratio of 1.8–2.0, were used for library construction. DNA libraries (350 bp) for Illumina sequencing were constructed for each accession following the manufacturer's specifications. After library construction, sequencing with 150 bp read lengths was carried out

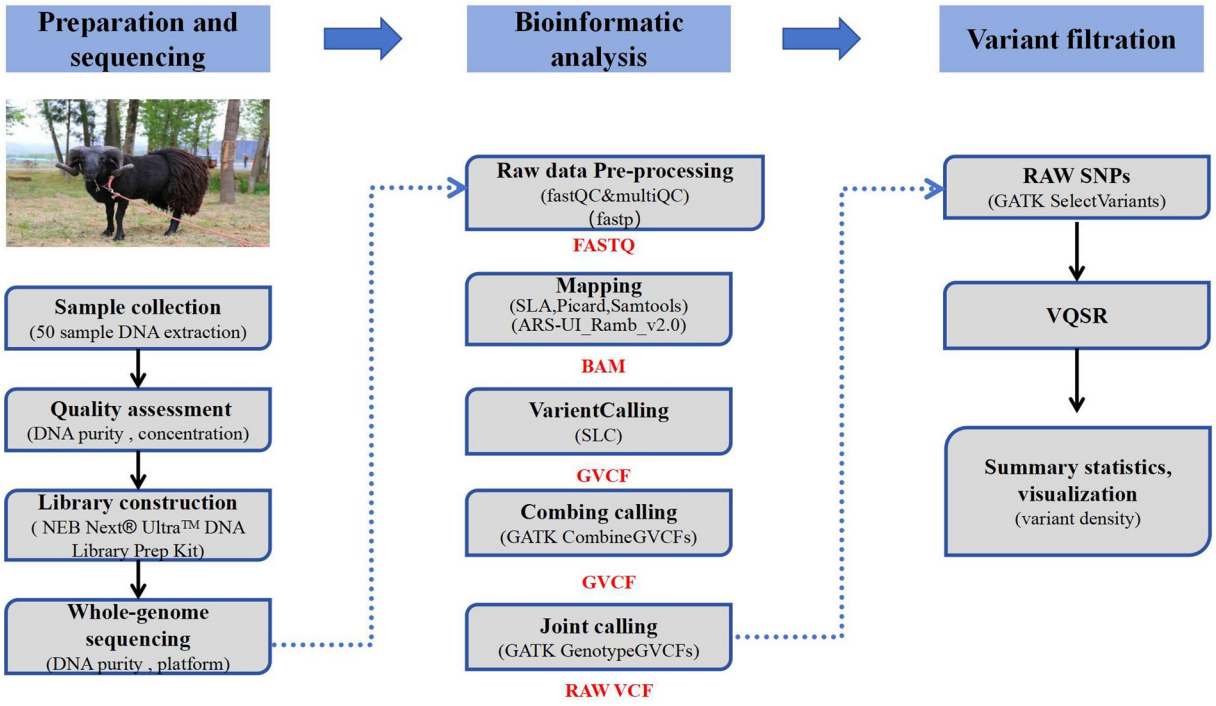

Fig. 3 The overall workflow of the quality control procedure and parameters used across all the stages of DNA sequencing (data pre-processing, variant discovery, and callset refinement).

on an Illumina NextSeq 2000 platform by a commercial service (Biomarker Technologies, Beijing, China). More information on the sequencing method is shown in Table S4.

Sequence data pre-processing and mapping. The raw reads were processed using fastp (v0.20.0)[27] with default parameters for adapter sequence identification and removal, as well as filtering and trimming of low-quality reads, resulting in clean reads that had passed QC. Before performing SNP calling, it was necessary to map the reads to the reference genome. In this study, the sheep reference genome ARS-UI_Ramb_v2.0 (GCF_016772045.1)[10] was adopted, which consists of 26 autosomes and X chromosome. BaseNumber (Sailegene Co., Ltd., Beijing), a commercial GPU-accelerated software, was primarily used for sequence alignment and SNP calling (https://github.com/WCH-IRD/BaseNumber)[28]. The SLA (Saile aligner) module applied the same algorithm as BWA-MEM, which was used to map the short reads in the fastq files with default parameters (-R '@RG\tID:'RGID'\tLB:'{LBID}'\tSM:'SMID'\tPL:'{PL}/reference/{fastq1}/${fastq2}), resulting in preliminary BAM files. Subsequently, SAMtools (v1.9)[29] was employed to sort the reads in the BAM files. Then, the Picard (v2.20.1) MarkDuplicates tool (http://broadinstitute.github.io/picard/) was used to mark and remove the duplicate and redundant reads generated during the PCR process for each individual, yielding the final sorted and non-redundant BAM files. Base quality score recalibration (BQSR) was performed using the Genome Analysis Toolkit (GATK, v4.4.0.0), with input comprising duplicate-marked BAM files and an internal variant panel of 83,386,953 SNPs from 1167 individuals as employed in our previous study[30].

SNP calling and filtering. We then used the SLC module of BaseNumber to perform individual SNP calling, generating gVCF files. Subsequently, the CombineGVCFs and GenotypeGVCFs modules of GATK (v4.4.0.0) were applied to merge individual gVCF files and perform joint genotyping across the population. Variant filtering was performed by applying the variant quality score recalibration (VQSR) approach in GATK (v4.4.0.0) using the same internal panel of 83,386,953 SNPs as described at the BQSR stage. During the VQSR step, the following annotations or context statistics were considered: read depth (DP), variant quality by depth (QD), root mean square of mapping quality (MQ), mapping quality rank sum test statistic (MQRankSum), read position rank sum test statistic (ReadPosRankSum), and strand bias statistics (FS and SOR). A tranche sensitivity threshold of 99% was applied for filtering variants. As the final quality control of the called variants, any SNPs with a missing genotype rate of more than 20% across the samples were filtered out using VCFtools (option–max-missing 0.8). The workflow from sampling to variant filtering is shown in Fig. 3.

Data records and accession numbers. The 50 Butuo Black sheep pair-end raw sequencing data sets (in fastq.gz format) are available at NCBI under sequence read archive (SRA) accession numbers PRJNA1194142 (https://identifiers.org/ncbi/insdc.sra:SRP552853)[31]. The VCF file is available in the European Variation Archive (EVA) through the accession number for the project, PRJEB87694 (https://identifiers.org/ena.embl:PRJEB87694)[32]. Raw data from the long-reads has been deposited in the NCBI

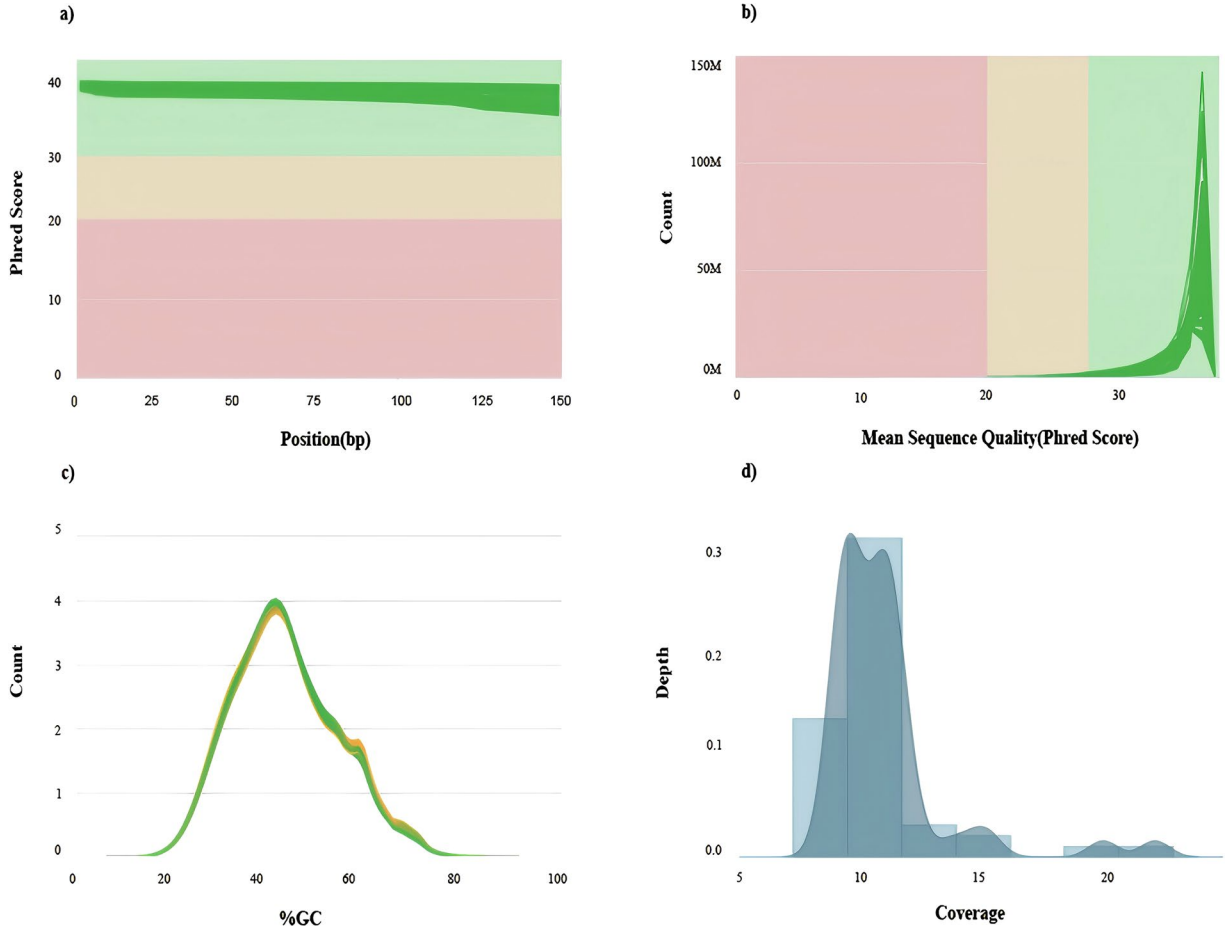

Fig. 4 Quality assessment of sequencing data and genome assembly. **a**) Mean quality scores across each base position; **b**) mean quality scores per read; **c**) GC content of reads; **d**) statistics of coverage (x) of genome assembly.

Sequence Read Archive database with accession number SRP553502 (https://identifiers.org/ncbi/insdc.sra:SRP553502)[33]. The genome assembly have been deposited in the GenBank database under the accession number JBPCZR000000000 (https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_050916685.1)[34]. The genome annotations were deposited in the Figshare database (https://doi.org/10.6084/m9.figshare.28600841)[35].

Technical Validation

Raw sequence reads underwent rigorous quality control to ensure suitability for downstream analyses. Key validation metrics included base quality, nucleotide composition, and coverage statistics (Fig. 4a–d).

Sequencing output and quality. A total of 1667 Gb of clean data was generated from 50 BBS samples using the Illumina HiSeq platform, with an average of 33.34 Gb, and the data per sample ranged from 27.07 Gb to 65.46 Gb. Over 97% of reads achieved a Phred quality score ≥ Q20 (99% base accuracy), and 92.36% ≥ Q30 criteria (99.9% accuracy), confirming high-fidelity sequencing (Fig. 4a,b; Figure S5).

Genomic coverage and depth. Recently, Jiang *et al*. suggested 4 × as the lowest boundary and 10 × as an ideal depth for achieving greater than 99% genome coverage in pigs[36]. The average estimated coverage for each of the 50 BBS samples was above the threshold with an average depth of 10.36× (Table S4; Fig. 4d). This moderate-depth WGS strategy balanced cost-effectiveness and variant detection reliability, enabling robust genetic analysis of population dynamics and trait-associated variants.

GC content consistency. GC content across all libraries showed stable distribution (average 42%), with no significant bias observed in nucleotide composition (Fig. 4c). Quality control via fastQC (v0.11.5)[37] and multiQC (v1.8)[38] confirmed uniform base quality and metrics meeting preprocessing standards for variant calling and downstream bioinformatics pipelines. After QC and consolidating individual sample VCF files, the joint genotyping analysis yielded 45 million SNPs (Table 3; Fig. 5).

Chromosome	Chromosome length	SNP count	SNP density (count/kb)
1	278,617,202	3,966,305	14.24
2	250,202,058	3,431,661	13.72
3	226,089,100	3,049,038	13.49
4	121,578,099	1,715,446	14.11
5	108,220,788	1,482,799	13.69
6	118,469,697	1,776,225	14.99
7	101,274,418	1,406,956	13.89
8	91,792,871	1,277,541	13.92
9	95,179,658	1,397,483	14.68
10	86,459,471	1,381,393	15.97
11	62,547,497	778,169	12.43
12	80,403,655	1,110,762	13.82
13	83,511,835	1,039,031	12.44
14	66,516,657	892,104	13.39
15	82,538,637	1,203,094	14.57
16	71,897,364	1,045,003	14.53
17	73,167,223	1,038,199	14.18
18	67,984,460	979,107	14.40
19	60,561,550	794,417	13.11
20	51,451,717	805,713	15.64
21	47,514,194	692,225	14.57
22	51,512,491	749,304	14.55
23	62,440,644	912,397	14.60
24	42,630,236	586,662	13.74
25	44,863,754	681,250	15.17
26	45,052,359	667,668	14.80
X	143,171,725	1,110,639	7.76
Total	2,615,649,360	35,970,591	376.4

Table 3. Summary statistics of SNPs in the VCF file for each chromosome.

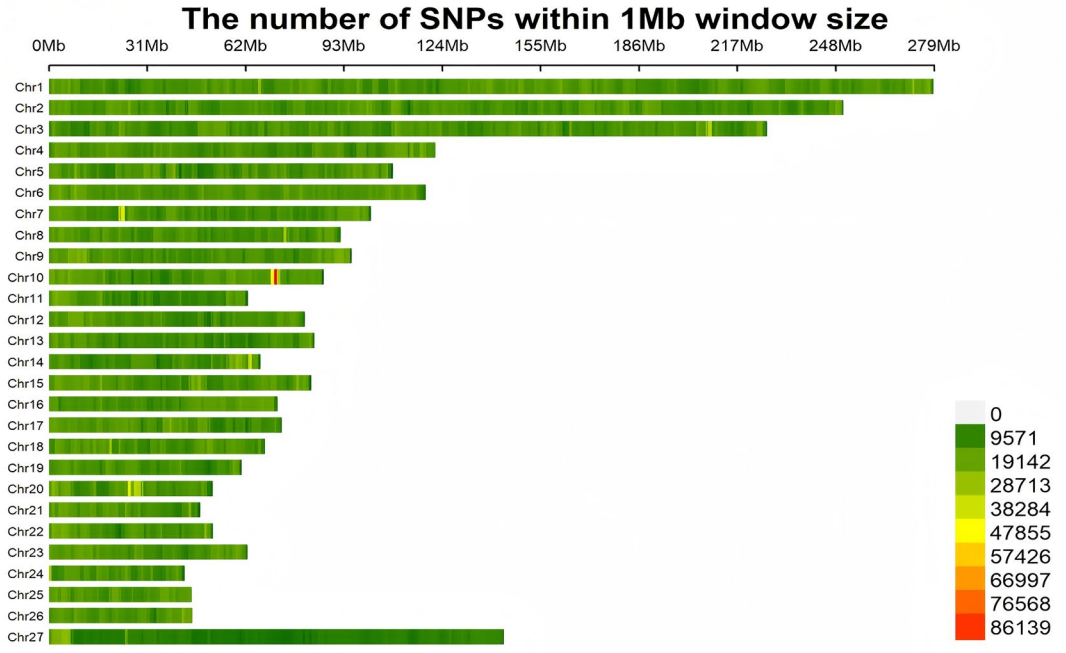

Fig. 5 Chromosome-wise SNPs distribution heatmap across the Butuo Black sheep genomes based on 45 Mb SNPs.

Code availability

Data analyses were primarily conducted using standard bioinformatics tools on the Linux operating system. We provide detailed information about the versions and code parameters in Figshare (https://doi.org/10.6084/m9.figshare.29224493)[39].

Received: 29 January 2025; Accepted: 11 August 2025;
Published online: 31 October 2025

References

1. Lv, F. H. *et al*. Whole-Genome Resequencing of Worldwide Wild and Domestic Sheep Elucidates Genetic Diversity, Introgression, and Agronomically Important Loci. *Molecular Biology and Evolution*. **39**, msab353 (2022).
2. Zhou, M. *et al*. Genetic diversity and population structure of sheep (Ovis aries) in Sichuan, China. *PLoS One*. **16**, e0257974 (2021).
3. Huang, Z. *et al*. Uncovering the genetic diversity and adaptability of Butuo Black Sheep through whole-genome re-sequencing. *PLoS One*. **19**, e0303419 (2024).
4. Wang Ting, W. T. & Wu DengJun, W. D. Study on genetic diversity of sheep populations in Sichuan by microsatellite markers. *Chinese Journal of Animal Science*. **4**, 36–41 (2018).
5. Zhang, M. *et al*. Transcriptomic and Metabolomic Analyses Reveal Molecular Regulatory Networks for Pigmentation Deposition in Sheep. *International Journal of Molecular Sciences*. **25**, 8248 (2024).
6. Ahmad, S. F. *et al*. Dissecting genomes of multiple yak populations: unveiling ancestry and high-altitude adaptation through whole-genome resequencing analysis. *BMC Genomics*. **26**, 214 (2025).
7. You, X. *et al*. A near complete genome assembly of the East Friesian sheep genome. *Scientific Data*. **11**, 762 (2024).
8. Li, R. *et al*. A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. *Genome Research*. **33**, 463–477 (2023).
9. Li, X. *et al*. Genomic analyses of wild argali, domestic sheep, and their hybrids provide insights into chromosome evolution, phenotypic variation, and germplasm innovation. *Genome Research*. **32**, 1669–1684 (2022).
10. Davenport, K. M. *et al*. An improved ovine reference genome assembly to facilitate in-depth functional annotation of the sheep genome. *Gigascience*. **11** (2022).
11. Lu, Z. *et al*. Chromosome-level genome assembly of Guide Black-Fur sheep (Ovis aries). *Scientific Data*. **11**, 711 (2024).
12. Tan, X. *et al*. Whole-genome variants dataset of 209 local chickens from China. *Scientific Data*. **11**, 169 (2024).
13. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics*. **13**, 278–289 (2015).
14. Chin, C. S. *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*. **10**, 563–569 (2013).
15. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*. **18**, 170–175 (2021).
16. Alonge, M. *et al*. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*. **20**, 224 (2019).
17. Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. BlobToolKit - Interactive Quality Assessment of Genome Assemblies. *G3 (Bethesda)*. **10**, 1361–1374 (2020).
18. Jiao, X. *et al*. A Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived from the PacBio RS. *Journal of Data Mining in Genomics and Proteomics*. **4** (2013).
19. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. **31**, 3210–3212 (2015).
20. Waterhouse, R. M. *et al*. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*. **35**, 543–548 (2018).
21. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture Research*. **10**, uhad127 (2023).
22. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics*. **21**(Suppl 1), i351–358 (2005).
23. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*. **6**, 11 (2015).
24. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics*. **37**, 1639–1643 (2021).
25. Di Pietro, F., Ortenzi, F., Tilio, M., Concetti, F. & Napolioni, V. Genomic DNA extraction from whole blood stored from 15- to 30-years at −20 °C by rapid phenol-chloroform protocol: a useful tool for genetic epidemiology studies. *Molecular and Cellular Probes*. **25**, 44–48 (2011).
26. Green, M. R. & Sambrook, J. Recovery of DNA from Agarose Gels Using Glass Beads. *Cold Spring Harbor Protocols*. **2019** (2019).
27. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. **34**, i884–i890 (2018).
28. Liu, H., Zhang, Q., Ai, F., Bu, F. & Yuan, H. Benchmarking germline variant calling performance of a GPU-accelerated tool on whole-genome sequencing datasets. *Available at Research Square* PREPRINT (Version 1) https://doi.org/10.21203/rs.3.rs-4318731/v1 (2024).
29. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).
30. Cheng, H. *et al*. Long divergent haplotypes introgressed from wild sheep are associated with distinct morphological and adaptive characteristics in domestic sheep. *PLoS Genetics*. **19**, e1010615 (2023).
31. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP552853 (2024).
32. *European Nucleotide Archive* https://identifiers.org/ena.embl:PRJEB87694 (2025).
33. *NCBI Sequence Read Archive*, https://www.ncbi.nlm.nih.gov/sra/SRP553502 (2024).
34. Zhong, C. S. Genome assembly and whole-genome resequencing study of Butuo Black sheep. *GenBank*. https://identifiers.org/ncbi/insdc.gca:GCA_050916685.1 (2025).
35. Zhong, C. S. Genome assembly and whole-genome resequencing study of Butuo Black sheep (Ovis aries). *figshare*. https://doi.org/10.6084/m9.figshare.28600841 (2025).
36. Jiang, Y. F., Jiang, Y., Wang, S., Zhang, Q. & Ding, X. Optimal sequencing depth design for whole genome re-sequencing in pigs. *BMC Bioinformatics*. **20**, 556 (2019).
37. Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*. **7**, 1338 (2018).
38. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. **32**, 3047–3048 (2016).
39. Zhong, C. S. Genome assembly code availability. *figshare* https://doi.org/10.6084/m9.figshare.29224493 (2025).

Acknowledgements

This work was financially supported by The Exploration of Genetic Resources of Butuo Black Sheep (K4050124028) and Shaanxi Livestock and Poultry Breeding Double-chain Fusion Key Project (2022GD-TSLD-46-0401). We would like to thank the High-performance Computing Platform of Northwest A&F University for computing support.

Author contributions

J.M.S. and R.L. conceived the study. Y.G.G., Z.J.B.Q., and J.K.W. collected the samples. J.W.L. extracted the genomic DNA. L.T.Z. performed genome assembly and bioinformatics analyses. C.S.Z. and W.J.S. wrote and edited the manuscript, with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05805-z.

Correspondence and requests for materials should be addressed to R.L. or J.S.

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.