Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A high-quality chromosome-level genome assembly and annotation of the Gansu Red Deer (Cervus elaphus kansuensis)

Hai Liu
Qinghai Normal University
Yun Fang
Xishuangbanna Tropical Botanical Garden
Xiaochan Bai
Qinghai Normal University
Zuoqing Yan
Forestry Station of Datong County
Qingling Wa
Forestry Station of Datong County
Nan Sun
Qinghai Normal University
Guogang Li
qhnulgg@126.com

Qinghai Normal University

data-descriptor

Keywords:

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11150129/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

The Gansu red deer (*Cervus elaphus kansuensis*) is an endemic cervid of China that is mainly distributed in the Qilian Mountains and is also known as the white-rumped deer. This subspecies is a nationally protected wild animal in China. However, the lack of a high-quality reference genome has limited further investigations into its genetic diversity, demographic history, and mechanisms of environmental adaptation. Here, we present a chromosome-level reference genome of the Gansu red deer generated using PacBio HiFi, Hi-C, and BGI short-read sequencing data. The assembled genome spans 3.11 Gb and consists of 72 contigs, with a contig N50 of 57.4 Mb and a scaffold N50 of 96.2 Mb. A total of 99.21% of the assembled sequences were anchored onto 34 pseudochromosomes. The assembly achieved 96.2% complete BUSCOs and a Merqury QV of 47.6, while 99.64% of the short reads were successfully mapped back to the assembled genome. Repetitive sequences accounted for 57.71% of the genome, and 20,349 protein-coding genes were predicted, with a BUSCO completeness of 99%. This chromosome-level reference genome provides a valuable genomic resource for future studies of genetic diversity, demographic history, adaptive evolution, and conservation of the Gansu red deer.

Background & Summary

The family Cervidae is one of the extant families of ruminant mammals and comprises deer species distributed across a wide range of habitats [1,2]. Deer are characterized by several distinctive biological features, among which the periodic growth, mineralization, casting, and regeneration of antlers is especially remarkable [3–5]. These features make cervids valuable models for studies of ruminant genome evolution, skeletal development, antler regeneration, ecological adaptation, and conservation biology [5–9].

The Gansu red deer (*Cervus elaphus kansuensis*, NCBI Taxonomy ID: 84700) [10,11] belongs to the family Cervidae and the genus *Cervus* and is a subspecies of the red deer (*Cervus elaphus*), with a distribution mainly in northwestern China [12,13]. Specifically, this subspecies occurs mainly in the Qilian Mountains and adjacent regions, including parts of Gansu, Qinghai, Ningxia, western Sichuan and eastern Xizang [12,14]. This subspecies inhabits high-elevation montane environments, including alpine grasslands, shrublands, coniferous forests, and valley forest–shrub habitats [15–17]. In Chinese literature, it is also known as “white-rumped deer” because of the conspicuous pale patch on the rump [18]. As a nationally protected wild animal in China [19], the Gansu red deer has conservation relevance and would benefit from high-quality genomic resources.

In recent years, genome assemblies have become available for several cervid species and subspecies, including red deer [20], sika deer [8,21], reindeer [22–24], milu (Père David’s deer) [25,26], muntjacs [27–30], and white-lipped deer [31]. These resources have improved our understanding of cervid evolution and ruminant comparative genomics. Nevertheless, genome quality and taxonomic coverage remain uneven across Cervidae [32], particularly at the subspecies level. A chromosome-level reference genome for the Gansu red deer would provide a useful genomic resource for comparative analyses within *Cervus*, for studies of

antler biology and environmental adaptation, and for future population genomic and conservation research.

Here, we report a chromosome-level genome assembly and annotation of the Gansu red deer generated using PacBio HiFi, Hi-C, BGI short-read, and RNA-seq data. The dataset includes raw sequencing reads, a chromosome-level genome assembly, gene annotation files, repeat annotation files, and mitochondrial genome resources. These data provide a reference framework for future studies of cervid genome evolution, red deer-red deer taxonomy, antler-related biology, and adaptation of deer populations in northwestern China.

Methods

Sample collection and sequencing

Muscle tissue samples were collected from a freshly deceased adult female Gansu red deer (2n = 68) in Datong County, Qinghai Province, China (100.97226287E, 37.39541892N) and immediately preserved in a liquid nitrogen tank at − 196°C (Fig. 1). The samples were subsequently transported under cryogenic conditions to the sequencing facility for DNA and RNA extraction.

High-molecular-weight genomic DNA was extracted from leaf tissue using a modified CTAB method, followed by RNase A treatment to remove residual RNA. DNA concentration and purity were assessed using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA) and a Qubit 3.0 fluorometer (Invitrogen, USA), and DNA integrity was evaluated by 1% agarose gel electrophoresis.

For PacBio HiFi sequencing, high-quality genomic DNA was fragmented to approximately 15 kb using a g-TUBE (Covaris, USA). A SMRTbell library was constructed using the SMRTbell Prep Kit 3.0 (Pacific Biosciences) according to the manufacturer's instructions. Large DNA fragments (> 15 kb) were selected using a BluePippin system (Sage Science, USA). The fragment-size distribution and concentration of the resulting library were evaluated using a Qsep100 system (BiOptic) and a Qubit 3.0 fluorometer, respectively. The qualified library was subsequently sequenced on the PacBio Revio platform to generate HiFi reads.

For short-read genome sequencing, genomic DNA was fragmented and size-selected, followed by end repair, dA-tailing, adapter ligation, PCR amplification, and library purification. The amplified products were subsequently circularized to generate single-stranded circular DNA libraries, which were sequenced on the DNBSEQ-T7 platform to generate paired-end reads.

For Hi-C sequencing, fresh tissue was ground in liquid nitrogen and crosslinked with formaldehyde. The crosslinked chromatin was digested with DpnII, and the resulting sticky ends were filled in with biotin-labeled nucleotides, followed by proximity ligation. After reversal of crosslinks, the DNA was purified and fragmented to approximately 300–500 bp. Biotin-labeled ligation products were enriched using

streptavidin magnetic beads and used for Hi-C library construction. The qualified Hi-C library was sequenced on the DNBSEQ-T7 platform.

In total, 120.3 Gb of PacBio HiFi reads, 134.6 Gb of short reads, and 402 Gb of Hi-C reads were generated. All DNA extraction, library construction, and sequencing procedures described above were performed by Haorui Genomics (Xi'an, China).

Transcriptome sequencing

Total RNA was extracted using TRIzol reagent (Invitrogen, USA) according to the manufacturer's instructions. RNA purity and integrity were assessed using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA), a Qsep100 system (BiOptic), and 1.5% agarose gel electrophoresis. Poly(A)+ mRNA was enriched using oligo(dT)-coupled magnetic beads and used to construct sequencing libraries with the VAHTS Universal V10 RNA-seq Library Prep Kit for MGI (Vazyme, Nanjing, China). Unique indexes were incorporated into each library to distinguish individual samples. The resulting libraries were quantified using a Qubit 3.0 Fluorometer (Invitrogen, USA), and their fragment-size distributions were evaluated using a Qsep100 system. Qualified libraries were sequenced on the DNBSEQ-T7 platform to generate paired-end reads. In total, 6.1 Gb of RNA-seq reads were generated.

Mitochondrial genome assembly and Species taxonomy

The complete mitochondrial genome of the Gansu red deer was assembled from short reads using GetOrganelle (v1.7.7.1) [33]. Clean reads were used as input, and the assembly was performed under the animal mitochondrial mode with the parameters “-F animal_mt, -R 10, and -t 8”. The resulting assembly graph was inspected using Bandage (v0.8.1) [34] to confirm the circular structure of the mitochondrial genome. The circular mitochondrial genome was annotated using two complementary approaches. First, MitoZ (v3.6) [35] was used for automated annotation with the parameters “--clade Chordata” and “--genetic_code 2”, corresponding to the vertebrate mitochondrial genetic code. In parallel, MITOS2 [36] was run on the Galaxy [37] platform using the Vertebrate genetic code and the RefSeq89 Metazoa reference dataset. Finally, the annotation results from MitoZ and MITOS2 were compared, and the boundaries of protein-coding genes, rRNA genes, and tRNA genes were manually checked and curated to generate the final mitochondrial genome annotation. The mitochondrial genome map was visualized using OGDRAW (v1.3.1) [38] (Fig. S1).

To further assess the taxonomic placement of the sequenced individual, mitochondrial phylogenetic analyses were performed using three datasets: complete mitochondrial genomes, the 13 mitochondrial protein-coding genes (PCGs), and the cytochrome b (cytb) gene. For the complete-mitogenome analysis, representative *Cervus* taxa were selected to cover the traditionally recognized Chinese red deer subspecies. Eight red deer subspecies have been reported from China [39], of which seven were represented by available complete mitochondrial genome sequences; *Cervus elaphus wallichii* was

excluded because no complete mitogenome sequence was available (Table S1). For the 13-PCG analysis, taxon sampling followed the framework used in the previously published mitochondrial genome study of *Cervus elaphus kansuensis* [19], with the mitochondrial sequence generated in this study added to the dataset (Table S2). To increase taxon coverage, particularly the number of reference individuals with published subspecific assignments, an additional phylogenetic analysis was conducted using available cytb sequences (Table S3).

For all three datasets, nucleotide sequences were aligned using MAFFT (v7.525) [40] and trimmed using trimAl (v1.5) [41]. Maximum-likelihood phylogenetic trees were reconstructed using IQ-TREE (v3.1.3) [42], with the best-fitting nucleotide substitution models selected automatically by ModelFinder [43]. For the 13-PCG dataset, each gene was aligned and trimmed separately, and the resulting alignments were analyzed as a concatenated dataset with each gene initially defined as an independent partition. The optimal partitioning scheme and corresponding substitution models were determined using the partition-merging procedure implemented in IQ-TREE (v3.1.3) (MFP+MERGE). Branch support was evaluated using 1,000 ultrafast bootstrap replicates [44]. The ML trees were visualized using iTOL (v7.5.0) [45].

Across all three mitochondrial datasets, the sequenced individual consistently clustered within the Chinese red deer lineage and showed the closest phylogenetic affinity to reference sequences assigned to *C. e. kansuensis* (Fig. S2). In the 13-PCG phylogeny, the mitochondrial sequence generated in this study formed a strongly supported sister relationship with the published *C. e. kansuensis* mitogenome NC_039923, with an ultrafast bootstrap support of 100%. The expanded cytb analysis, which included multiple independently reported *C. e. kansuensis* sequences, likewise placed the study sample within the kansuensis-associated mitochondrial cluster. Although several deeper relationships among closely related Chinese red deer lineages received relatively low bootstrap support, the placement of the study sample relative to the *C. e. kansuensis* reference sequences was consistent across datasets.

The sampling locality in Datong County, Qinghai Province, is situated in the eastern Qilian Mountain region, within the traditionally recognized distribution range of Gansu red deer [12,19]. Previous studies have reported *C. e. kansuensis* from the Qilian Mountains and adjacent areas of Qinghai and Gansu, and recent mitochondrial analyses of *Cervus elaphus* populations from Qinghai and Gansu included a Datong population and assigned these populations to *C. e. kansuensis* [46]. Taken together, the mitochondrial phylogenetic evidence and the geographic origin of the specimen support its identification as Gansu red deer, traditionally referred to as *Cervus elaphus kansuensis*.

Genome size and ploidy assessment

Before the *de novo* assembly, a genome survey and ploidy assessment of the Gansu red deer were conducted to estimate genomic characteristics, including genome size, heterozygosity, and ploidy. A total of 134.6 Gb clean short reads were used for genome size and heterozygosity estimation (Table S4).

The 21-mers were counted using Jellyfish (v.2.3.0) [47] with parameters -C -m 21 -s 10G -t 16, and then the k-mer frequency histogram was generated using the histo subcommand. The resulting histogram was analyzed using GenomeScope (v2.0) [48] to estimate genome size, heterozygosity, and k-mer-based repetitive content. Based on the 21-mer distribution, the genome size of the Gansu red deer was estimated to be approximately 3.01 Gb, with an estimated heterozygosity of 0.611% (Fig. S3; Table S5). Genome ploidy was inferred using Smudgeplot (v0.4.0) [48], with k-mers counted using KMC (v3.1.1) [49] and filtered according to coverage cutoffs estimated from the k-mer histogram. The resulting Smudgeplot profile supported a diploid genome for the sequenced individual (Fig. S4).

Genome assembly and quality assessment

Based on 120.3 Gb of PacBio HiFi reads, the Gansu red deer genome was initially assembled using Hifiasm (v0.16.1) [50] with default parameters. The resulting contigs were polished through three rounds of Racon (v.1.5.0) [51] using long reads to enhance sequence accuracy. Further base-level refinement was performed with Nextpolish (v.1.4.0) [52], which utilized short reads under default parameters. To address redundancy from heterozygous regions, purge_dups (v1.2.5) [53] was applied with default parameters, yielding contig-level assemblies, generating a 3.367 Gb contig-level assembly comprising 200 contigs, with a contig N50 of 95.09 Mb. Hi-C reads (402 Gb) were used to scaffold the contigs into chromosome-scale sequences using HapHiC (v1.0.7) [54]. Briefly, Hi-C reads were mapped to the contig-level assembly, and valid Hi-C interaction pairs were used by HapHiC to cluster, order, and orient the contigs into chromosome groups. The resulting assembly was visually inspected using Hi-C contact maps in Juicebox (v2.17) [55], and apparent misjoins, incorrect orientations, and contig-ordering errors were manually corrected. The curated assembly was subsequently processed with 3D-DNA [56] to generate the final chromosome-level assembly. The final genome assembly was 3.112 Gb in length and consisted of 72 sequences, with a scaffold N50 of 96.25 Mb (Fig. 2). In total, 34 pseudochromosomes were constructed, anchoring 3.087 Gb of sequences and representing 99.21% of the assembled genome (Fig. 3). To further evaluate assembly quality, clean short reads were mapped back to the final assembly, resulting in a mapping rate of 99.64%. The completeness and integrity were assessed with BUSCO (v6.0.0) [57] against the cetartiodactyla_odb10 [58] and mammalia_odb10 database [58], yielding complete BUSCO scores of 96.2% and 96.3% (Table S6), respectively. The base-level quality value (QV) was estimated using merqury (v1.3) [59], resulting in a genome-wide QV of 47.68 (Table S7). Potential contamination in the final genome assembly was assessed using the BlobToolKit pipeline (v4.4.5) [60], implemented as a Snakemake workflow with Snakemake (v5.20.1) [61]. No obvious contamination was detected in the final assembly, and all 34 pseudochromosomes were taxonomically assigned to Chordata (Fig. S5).

Genome annotation

Repetitive elements in the Gansu red deer genome were annotated using a combined *de novo* and homology-based strategy. A species-specific *de novo* repeat library was constructed from the final genome assembly using RepeatModeler (v2.0.9) [62] and combined with Cervidae-associated repeat families derived from Dfam4.0 [63] and RepBase 20181026 [64]. The resulting custom library was used to identify and mask repetitive elements using RepeatMasker (v4.2.4) [65] with the RMBlast engine and the parameters “-xsmall -poly -gff -html”.

In total, 1.795 Gb of the Gansu red deer genome was annotated as repetitive sequences, accounting for 57.71% of the assembly. Interspersed repeats occupied 1.309 Gb, corresponding to 42.09% of the genome (Table S8). LINEs represented the predominant interspersed repeat class, accounting for 27.97% of the genome, followed by LTR elements (4.85%), SINEs (3.21%), DNA transposons (2.11%), and unclassified repeats (3.95%). Satellite sequences accounted for 14.93% of the genome, whereas simple repeats and low-complexity sequences accounted for 0.54% and 0.11%, respectively. Notably, satellite DNA accounted for 14.93% of the assembled genome. This proportion is broadly consistent with previous observations that satellite DNA can constitute a substantial fraction of cervid genomes [66]. Satellite DNA was estimated to account for 11.9% of the red deer (*Cervus elaphus*) genome [67], while individual satellite families have been reported to comprise 8–12% of ruminant genomes and up to 35% in some cervid species [66,68].

Gene structure annotation was performed using EGAPx (v0.5.2) [69], the publicly available implementation of the NCBI Eukaryotic Genome Annotation Pipeline. The final chromosome-level genome assembly of the Gansu red deer was used as input for the EGAPx workflow. EGAPx was executed with Nextflow (v23.10.1) [61] using the Singularity engine, and the workflow was run with Singularity (v3.11.0) [70] in an offline configuration. The EGAPx container image egapx_0.5.2.sif and the local support data package egapxsupportdata_20251017 were used for the annotation workflow. In total, 20,349 protein-coding genes were annotated in the Gansu red deer genome.

Functional annotation of the predicted protein-coding genes was performed using multiple complementary approaches. For each protein-coding gene, the longest protein isoform was retained for downstream functional annotation. Homology-based annotations were assigned by sequence similarity searches against the NCBI non-redundant protein (NR), Swiss-Prot, COG, and KOG databases. A total of 20,328 (99.90%), 20,096 (98.76%), 7,583 (37.26%), and 19,556 (96.10%) genes were annotated against the NR, Swiss-Prot, COG, and KOG databases, respectively. Orthology-based functional annotation was performed using eggNOG-mapper (v2.1.13) [71] with eggNOG database (v5.0.2) [72], resulting in 20,254 (99.53%) annotated genes. Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway information were extracted from the eggNOG-mapper annotation results, with 17,082 (83.95%) and 9,420 (46.29%) genes assigned GO and KEGG annotations, respectively. Protein domains and conserved functional signatures were further annotated independently using Pfam and InterProScan (v5.76-107.0) [73]. Pfam domains were identified in 19,357 (95.13%) genes, whereas 20,049 (98.53%) genes received InterProScan annotations. The annotation results from all databases and methods were

integrated at the gene level. In total, 20,347 of the 20,349 predicted protein-coding genes were assigned at least one functional annotation (Table S9, S10). Non-coding RNAs were annotated separately according to RNA type. Transfer RNA (tRNA) genes were predicted using tRNAscan-SE (v2.0.13) [74] in eukaryotic mode. tRNAscan-SE identified 266,176 tRNA-like loci, including 234,581 possible pseudogenes. The predictions were subsequently processed using EukHighConfidenceFilter with default thresholds, resulting in a final set of 689 high-confidence tRNA genes. Other structured non-coding RNAs were identified using Infernal (v1.1.5) [75] against the Rfam database (v15.1) [76]. The resulting Rfam hits were classified according to their corresponding RNA family types, yielding 875 rRNAs, 591 miRNAs, 2,471 snRNAs, 1,153 snoRNAs, and755 other non-coding RNAs.

The completeness of the predicted protein-coding gene set was assessed using BUSCO (v6.0.0) [57] in protein mode (Table S11). Based on the cetartiodactyla_odb10 database [58], the annotated protein set achieved a BUSCO completeness of 99.0%, with 0.2% fragmented and 0.8% missing BUSCOs. Using the mammalia_odb10 database [58], the protein set showed a BUSCO completeness of 99.4%, with 0.2% fragmented and 0.4% missing BUSCOs. The quality of the predicted protein-coding gene set was additionally assessed using OMAmer (v2.1.2) [77] and OMArk (0.5.0) [78]. The longest protein isoform of each protein-coding gene was searched against the LUCA OMAmer database using OMAmer with default parameters, and the resulting HOG assignments were evaluated using OMArk. Artiodactyla was selected as the ancestral lineage for completeness assessment (Fig S6), with 97.22% of conserved HOGs classified as single-copy, 1.88% as duplicated, and 0.90% as missing. OMArk further classified 98.57% of the predicted proteins as taxonomically consistent, with 0.25% remaining unknown. These results collectively support the high completeness and reliability of the predicted protein-coding gene set. Overall, the assembly continuity and annotation statistics of the Gansu red deer genome were broadly comparable to those of previously published red deer [20] and domestic cattle [79] reference genomes (Table 1).

Table 1
Summary statistics of genome assemblies and annotations for the Gansu red deer and related species.

Assembly size (bp)	Gansu red deer	Red deer [a]	Domestic cattle [b]
	3,112,034,529	**2,886,603,524**	**2,770,686,120**
Contig N50 (bp)	57,407,273	68,741,660	26,402,946
Number of scaffolds	72	144	1,958
Longest scaffold (bp)	171,302,835	179,953,079	158,534,110
Scaffold N50 (bp)	96,251,848	83,473,711	103,308,737
Scaffold N90 (bp)	64,652,836	51,846,305	51,992,305
Number of pseudo-chromosomes	34	34	30
Number of Gaps	88	40	386
Number of protein-coding genes	20,349	22,941	21,667

[a] Genome assembly data for red deer (*Cervus elaphus*) was obtained from NCBI (GCF_910594005.1).
[b] Genome assembly data for domestic cattle (*Bos taurus*) was obtained from NCBI (GCF_002263795.3).

Data Records

The raw sequencing data generated in this study have been deposited in the Genome Sequence Archive (GSA) at the National Genomics Data Center (NGDC) [80], China National Center for Bioinformation (CNCB), under BioProject accession PRJCA073564 [81]. These data include PacBio HiFi reads, Hi-C reads, BGI paired-end short reads, and RNA-seq reads, with the corresponding BioSample and GSA accession numbrs SAMC8460867 [82] and CRA049869 [83], respectively. The chromosome-level genome assembly of the Gansu red deer has been deposited in the Genome Warehouse (GWH) [84] at NGDC under accession GWHKCMM00000000.1 [85]. Genome annotation files and predicted protein sequences have been deposited in Figshare [86].

Technical Validation

The quality of the final genome assembly was evaluated using multiple complementary approaches. Potential contamination was assessed using BlobToolKit pipeline (v4.4.5) [60] based on sequence composition, read coverage, and taxonomic assignment. The resulting BlobToolKit profiles showed that the vast majority of assembled sequences were assigned to the expected chordate lineage, with no evidence of substantial non-target contamination in the final assembly.

Genome completeness and consensus accuracy were evaluated using BUSCO (v6.0.0) [57] and Merqury (v1.3) [59]. BUSCO analyses against the cetartiodactyla_odb10 [58] and mammalia_odb10 [58] datasets recovered 96.2% and 96.3% complete orthologues, respectively, with 92.9–93.0% identified as single-copy BUSCOs and only 3.3% as duplicated. Fragmented BUSCOs accounted for 1.1% in both datasets, whereas 2.8% and 2.5% were missing (Table S6), respectively. Merqury yielded a consensus QV of 47.68 (Table S7), supporting the high completeness and accuracy of the final chromosome-level assembly.

The completeness and reliability of the protein-coding gene annotation were evaluated using complementary approaches. BUSCO (v6.0.0) [57] analysis in protein mode recovered 99.0% and 99.4% complete orthologues using the cetartiodactyla_odb10 [58] and mammalia_odb10 [58] datasets (Table S11), respectively, with only 0.2% fragmented BUSCOs in both datasets and 0.8% and 0.4% missing BUSCOs, respectively. Gene structural characteristics, including mRNA, CDS, exon, and intron lengths, were further compared with those of closely related cervid and bovine genomes (Fig. 4). The Gansu red deer showed broadly consistent distribution patterns across these structural features, with no obvious systematic deviation from the reference species, supporting the reliability of the gene models.

Annotation quality was additionally assessed using OMAmer (v2.1.2) [77] and OMArk (v0.5.0) [78]. Based on the Artiodactyla lineage, OMArk classified 97.22% of conserved HOGs as single-copy, 1.88% as duplicated, and only 0.90% as missing (Fig S6). Together, the BUSCO, comparative gene-structure, and OMArk assessments indicate a highly complete and taxonomically consistent protein-coding gene annotation.

Declarations

Data Availability

The raw sequencing data generated in this study have been deposited in the Genome Sequence Archive (GSA) at the National Genomics Data Center (NGDC) [80], China National Center for Bioinformation (CNCB), under BioProject accession PRJCA073564 [81]. These data include PacBio HiFi reads, Hi-C reads, BGI paired-end short reads, and RNA-seq reads, with the corresponding BioSample and GSA accession numbers SAMC8460867 [82] and CRA049869 [83], respectively. The chromosome-level genome assembly of the Gansu red deer has been deposited in the Genome Warehouse (GWH) [84] at NGDC under accession GWHKCMM00000000.1 [85]. Genome annotation files and predicted protein sequences have been deposited in Figshare [86].

Code Availability

No custom code was generated for this study. All analyses were performed using publicly available bioinformatics software according to the developers' manuals and protocols, with parameter settings reported in the Methods section.

Author Contributions:

HL: performed genome assembly and annotation, drafted and revised the manuscript; YF: revised the manuscript, provided methodological guidance on mitochondrial genome analysis; BX: performed genome survey analysis, and data deposition; YZ: contributed to sample collection; WQ: contributed to sample collection; NS: contributed to sample collection; GL: designed and supervised the research, revised the manuscript; All authors reviewed and approved the final manuscript.

Competing Interests:

The authors declare no conflicts of interest.

Acknowledgements:

Sequencing service was provided by Haorui Genomics (Xi'an, China).

Funding:

This work was supported by the National Natural Science Foundation of China (Grant No. 32560252, 32360133).

Ethics statement:

All experiments concerning animals were conducted in accordance with the guidelines of the China Council on Animal Care, and all protocols were approved by the Institutional Animal Care and Use Committee of Qinghai Normal University (Approval No. 2026-12).

References

1. Hackmann, T. J. & Spain, J. N. Invited review: ruminant ecology and evolution: perspectives useful to ruminant livestock research and production. *J Dairy Sci* 93, 1320–1334 (2010). https://doi.org/10.3168/jds.2009-2071
2. Gilbert, C., Ropiquet, A. & Hassanin, A. Mitochondrial and nuclear phylogenies of Cervidae (Mammalia, Ruminantia): Systematics, morphology, and biogeography. *Mol Phylogenet Evol* 40, 101–117 (2006). https://doi.org/10.1016/j.ympev.2006.02.017
3. Price, J. S., Allen, S., Faucheux, C., Althnaian, T. & Mount, J. G. Deer antlers: a zoological curiosity or the key to understanding organ regeneration in mammals? *J Anat* 207, 603–618 (2005). https://doi.org/10.1111/j.1469-7580.2005.00478.x
4. Li, C., Yang, F. & Sheppard, A. Adult stem cells and mammalian epimorphic regeneration-insights from studying annual renewal of deer antlers. *Curr Stem Cell Res Ther* 4, 237–251 (2009). https://doi.org/10.2174/157488809789057446
5. Wang, Y. *et al.* Genetic basis of ruminant headgear and rapid antler regeneration. *Science* 364 (2019). https://doi.org/10.1126/science.aav6335

6. Chen, L. *et al.* Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science* 364 (2019). https://doi.org/10.1126/science.aav6202
7. Qin, T. *et al.* A population of stem cells with strong regenerative potential discovered in deer antlers. *Science* 379, 840–847 (2023). https://doi.org/10.1126/science.add0488
8. Xing, X. *et al.* The First High-quality Reference Genome of Sika Deer Provides Insights into High-tannin Adaptation. *Genomics Proteomics Bioinformatics* 21, 203–215 (2023). https://doi.org/10.1016/j.gpb.2022.05.008
9. Zhong, Z. *et al.* Phylogenomic analysis of Cervidae provides insights into antler origin and evolution. *Zool Res* 47, 289–302 (2026). https://doi.org/10.24272/j.issn.2095-8137.2025.143
10. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* 2020 (2020). https://doi.org/10.1093/database/baaa062
11. NCBI. *Taxonomy Cervus elaphus kansuensis Taxonomy Browser.*, %3Chttps://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=84700%3E (2026).
12. Tang, L., Wen, X., Zhang, R. & Xing, X. Current Situation and Utilization of Velvet Deer Germplasm Resources in China. *Animals (Basel)* 12 (2022). https://doi.org/10.3390/ani12243529
13. Hu, P. *et al.* Genome-wide study on genetic diversity and phylogeny of five species in the genus Cervus. *BMC Genomics* 20, 384 (2019). https://doi.org/10.1186/s12864-019-5785-z
14. Zhang, Z. *et al.* Integrated omics analysis reveals the alteration of gut microbiota and fecal metabolites in Cervus elaphus kansuensis. *Appl Microbiol Biotechnol* 108, 125 (2024). https://doi.org/10.1007/s00253-023-12841-5
15. FU, J. & Ji, R. Evaluation on trampling of grazed Gansu wapiti (*Cervus elaphus kansuensis* Po cock) and its effects on soil property in winter grazing land. *Acta Ecologica Sinica* 3, 486–495 (2003).
16. Fu, J., Li, G. & Yang, F. Grazing behavior of Gansu wapiti (*Cervus elaphus kansuensis*) in summer & winter on the alpine grasslands of Qilianshan Mountain. *Acta Ecologica Sinica* 23, 1807–1815 (2003).
17. Meijuan, S. *et al.* Long-term grazing of *Cervus elaphus kansuensis* restrains expansion of Salix cupularis shrubs in the Eastern Qilian Mountains. *Global Ecology and Conservation* 33, e01987 (2022). https://doi.org/https://doi.org/10.1016/j.gecco.2021.e01987
18. Chen, Y. *et al.* Integrative Analyses of Antler Cartilage Transcriptome and Proteome of Gansu Red Deer (*Cervus elaphus kansuensis*) at Different Growth Stages. *Animals (Basel)* 12 (2022). https://doi.org/10.3390/ani12070934
19. Liu, H. *et al.* The complete mitochondrial genome of *Cervus elaphus kansuensis* (Artiodactyla: Cervidae) and its phylogenetic analysis. *Mitochondrial DNA Part B* 4, 1720–1722 (2019). https://doi.org/10.1080/23802359.2019.1607588
20. Pemberton, J., Johnston, S. E. & Fletcher, T. J. The genome sequence of the red deer, *Cervus elaphus* Linnaeus 1758. *Wellcome Open Research* 6 (2021). https://doi.org/10.12688/wellcomeopenres.17493.1

21. Yang, Q. *et al.* High-quality chromosome-scale genome assembly and annotation of Taohongling Sika deer (*Cervus nippon kopschi*). *Sci Data* 12, 1839 (2025). https://doi.org/10.1038/s41597-025-06122-1
22. Li, Z. *et al.* Draft genome of the reindeer (*Rangifer tarandus*). *Gigascience* 6, 1–5 (2017). https://doi.org/10.1093/gigascience/gix102
23. Weldenegodguad, M. *et al.* Genome sequence and comparative analysis of reindeer (*Rangifer tarandus*) in northern Eurasia. *Sci Rep* 10, 8980 (2020). https://doi.org/10.1038/s41598-020-65487-y
24. Tørresen, O. K. *et al.* Chromosome-level genome assembly of Norwegian wild alpine reindeer (*Rangifer tarandus tarandus*). *J Hered* 117, 557–565 (2026). https://doi.org/10.1093/jhered/esaf094
25. Zhang, C. *et al.* Draft genome of the milu (*Elaphurus davidianus*). *Gigascience* 7 (2018). https://doi.org/10.1093/gigascience/gix130
26. Zhu, L. *et al.* Endangered Père David's deer genome provides insights into population recovering. *Evol Appl* 11, 2040–2053 (2018). https://doi.org/10.1111/eva.12705
27. Mudd, A. B., Bredeson, J. V., Baum, R., Hockemeyer, D. & Rokhsar, D. S. Analysis of muntjac deer genome and chromatin architecture reveals rapid karyotype evolution. *Commun Biol* 3, 480 (2020). https://doi.org/10.1038/s42003-020-1096-9
28. Yin, Y. *et al.* Molecular mechanisms and topological consequences of drastic chromosomal rearrangements of muntjac deer. *Nat Commun* 12, 6858 (2021). https://doi.org/10.1038/s41467-021-27091-0
29. Cui, Y. *et al.* A haplotype-resolved and chromosome-scale genome assembly of Hainan muntjac (*Muntiacus nigripes*). *Sci Data* 11, 1395 (2024). https://doi.org/10.1038/s41597-024-04167-2
30. Supapannachart, P. *et al.* Genome Assembly of the Threatened Fea's Muntjac (*Muntiacus feae*) Reveals Adaptive Evolution and Comparative Population Genomics of Fea's and Red Muntjacs. *Animal Genetics* 57, e70093 (2026). https://doi.org/https://doi.org/10.1002/age.70093
31. Yang, Y. *et al.* A high-quality Chromosome-level reference genome assembly of white-lipped deer (*Przewalskium albirostris*). *Sci Data* 12, 727 (2025). https://doi.org/10.1038/s41597-025-04796-1
32. Tang, L., Dong, S. & Xing, X. Comparative Genomics Reveal Phylogenetic Relationship and Chromosomal Evolutionary Events of Eight Cervidae Species. *Animals (Basel)* 14 (2024). https://doi.org/10.3390/ani14071063
33. Jin, J. J. *et al.* GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol* 21, 241 (2020). https://doi.org/10.1186/s13059-020-02154-5
34. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 31, 3350–3352 (2015). https://doi.org/10.1093/bioinformatics/btv383
35. Meng, G., Li, Y., Yang, C. & Liu, S. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Res* 47, e63 (2019). https://doi.org/10.1093/nar/gkz173

36. Bernt, M. *et al.* MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol Phylogenet Evol* 69, 313–319 (2013). https://doi.org/10.1016/j.ympev.2012.08.023
37. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res* 50, W345–w351 (2022). https://doi.org/10.1093/nar/gkac247
38. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res* 47, W59–w64 (2019). https://doi.org/10.1093/nar/gkz238
39. Tumur, A., Abliz, D. & Halik, M. Habitat dynamics and its influence on the genetic diversity of Tarim red deer (*Cervus elaphus yarkandensis*) Xayar population of Xinjiang, China. *Quaternary International* 311, 140–145 (2013). https://doi.org/https://doi.org/10.1016/j.quaint.2013.07.007
40. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30, 772–780 (2013). https://doi.org/10.1093/molbev/mst010
41. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973 (2009). https://doi.org/10.1093/bioinformatics/btp348
42. Wong, T. K. F. *et al.* IQ-TREE 3: phylogenomic inference software using complex evolutionary models. *Mol Biol Evol* 43 (2026). https://doi.org/10.1093/molbev/msag117
43. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14, 587–589 (2017). https://doi.org/10.1038/nmeth.4285
44. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* 35, 518–522 (2018). https://doi.org/10.1093/molbev/msx281
45. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res* 52, W78–w82 (2024). https://doi.org/10.1093/nar/gkae268
46. Dong, S. *et al.* Mitochondrial PCGs Provide Novel Insights into Subspecies Classification, Codon Usage and Selection of *Cervus canadensis* Distributed in Qinghai and Gansu, China. *Animals (Basel)* 15 (2025). https://doi.org/10.3390/ani15101486
47. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770 (2011). https://doi.org/10.1093/bioinformatics/btr011
48. 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). https://doi.org/10.1038/s41467-020-14998-3
49. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 33, 2759–2761 (2017). https://doi.org/10.1093/bioinformatics/btx304
50. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 18, 170–175 (2021).

https://doi.org/10.1038/s41592-020-01056-5

51. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 27, 737–746 (2017). https://doi.org/10.1101/gr.214270.116
52. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* 36, 2253–2255 (2020). https://doi.org/10.1093/bioinformatics/btz891
53. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* 36, 2896–2898 (2020). https://doi.org/10.1093/bioinformatics/btaa025
54. Zeng, X. *et al.* Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat Plants* 10, 1184–1200 (2024). https://doi.org/10.1038/s41477-024-01755-3
55. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* 3, 99–101 (2016). https://doi.org/10.1016/j.cels.2015.07.012
56. Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* 356, 92–95 (2017). https://doi.org/10.1126/science.aal3327
57. Tegenfeldt, F. *et al.* OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res* 53, D516–d522 (2025). https://doi.org/10.1093/nar/gkae987
58. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* 38, 4647–4654 (2021). https://doi.org/10.1093/molbev/msab199
59. 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). https://doi.org/10.1186/s13059-020-02134-9
60. Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. BlobToolKit - Interactive Quality Assessment of Genome Assemblies. *G3 (Bethesda)* 10, 1361–1374 (2020). https://doi.org/10.1534/g3.119.400908
61. Molder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res* 10, 33 (2021). https://doi.org/10.12688/f1000research.29032.3
62. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* 117, 9451–9457 (2020). https://doi.org/10.1073/pnas.1921046117
63. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* 12, 2 (2021). https://doi.org/10.1186/s13100-020-00230-y
64. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 6, 11 (2015). https://doi.org/10.1186/s13100-015-0041-9

65. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics Chap. 4*, 4.10.11-14.10.14 (2009). https://doi.org/10.1002/0471250953.bi0410s25
66. Vozdova, M. *et al.* Sequence Analysis and FISH Mapping of Four Satellite DNA Families among Cervidae. *Genes (Basel)* 11 (2020). https://doi.org/10.3390/genes11050584
67. Zhang, Y., Chu, J., Cheng, H. & Li, H. De novo reconstruction of satellite repeat units from sequence data. *Genome Res* 33, 1994–2001 (2023). https://doi.org/10.1101/gr.278005.123
68. Blake, R. D., Wang, J. Z. & Beauregard, L. Repetitive sequence families in *Alces alces* americana. *J Mol Evol* 44, 509–520 (1997). https://doi.org/10.1007/pl00006175
69. Thibaud-Nissen, F. *et al.* P8008 The NCBI Eukaryotic Genome Annotation Pipeline. *Journal of Animal Science* 94, 184–184 (2016). https://doi.org/10.2527/jas2016.94supplement4184x
70. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLoS One* 12, e0177459 (2017). https://doi.org/10.1371/journal.pone.0177459
71. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* 38, 5825–5829 (2021). https://doi.org/10.1093/molbev/msab293
72. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 47, D309–d314 (2019). https://doi.org/10.1093/nar/gky1085
73. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240 (2014). https://doi.org/10.1093/bioinformatics/btu031
74. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* 49, 9077–9096 (2021). https://doi.org/10.1093/nar/gkab688
75. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935 (2013). https://doi.org/10.1093/bioinformatics/btt509
76. Ontiveros-Palacios, N. *et al.* Rfam 15: RNA families database in 2025. *Nucleic Acids Res* 53, D258–d267 (2025). https://doi.org/10.1093/nar/gkae1023
77. Rossier, V., Warwick Vesztrocy, A., Robinson-Rechavi, M. & Dessimoz, C. OMAmer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches. *Bioinformatics* 37, 2866–2873 (2021). https://doi.org/10.1093/bioinformatics/btab219
78. Nevers, Y. *et al.* Quality assessment of gene repertoire annotations with OMArk. *Nat Biotechnol* 43, 124–133 (2025). https://doi.org/10.1038/s41587-024-02147-w
79. Rosen, B. D. *et al.* De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience* 9, giaa021 (2020). https://doi.org/10.1093/gigascience/giaa021

80. Partners, C. N. M. a. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2026. *Nucleic Acids Res* 54, D28–d47 (2026). https://doi.org/10.1093/nar/gkaf1172
81. *NGDC BioProject accession*, https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA073564.
82. *NGDC BioSample accession*, https://ngdc.cncb.ac.cn/biosample/browse/SAMC8460867.
83. *NGDC GSA*, https://ngdc.cncb.ac.cn/gsa/browse/CRA049869.
84. Ma, Y. *et al.* The Updated Genome Warehouse: Enhancing Data Value, Security, and Usability to Address Data Expansion. *Genomics, Proteomics & Bioinformatics* 23, qzaf010 (2025). https://doi.org/10.1093/gpbjnl/qzaf010
85. *NGDC GWH*, https://ngdc.cncb.ac.cn/gwh/Assembly/137728/show.
86. Liu, H. A high-quality chromosome-level genome assembly and annotation of the Gansu Red Deer (*Cervus elaphus kansuensis*). *figshare* (2026). https://doi.org/https://doi.org/10.6084/m9.figshare.33582463.v2

Figures



Figure 1

A female Gansu red deer in the Datong County, from the same population of the sequenced individual.

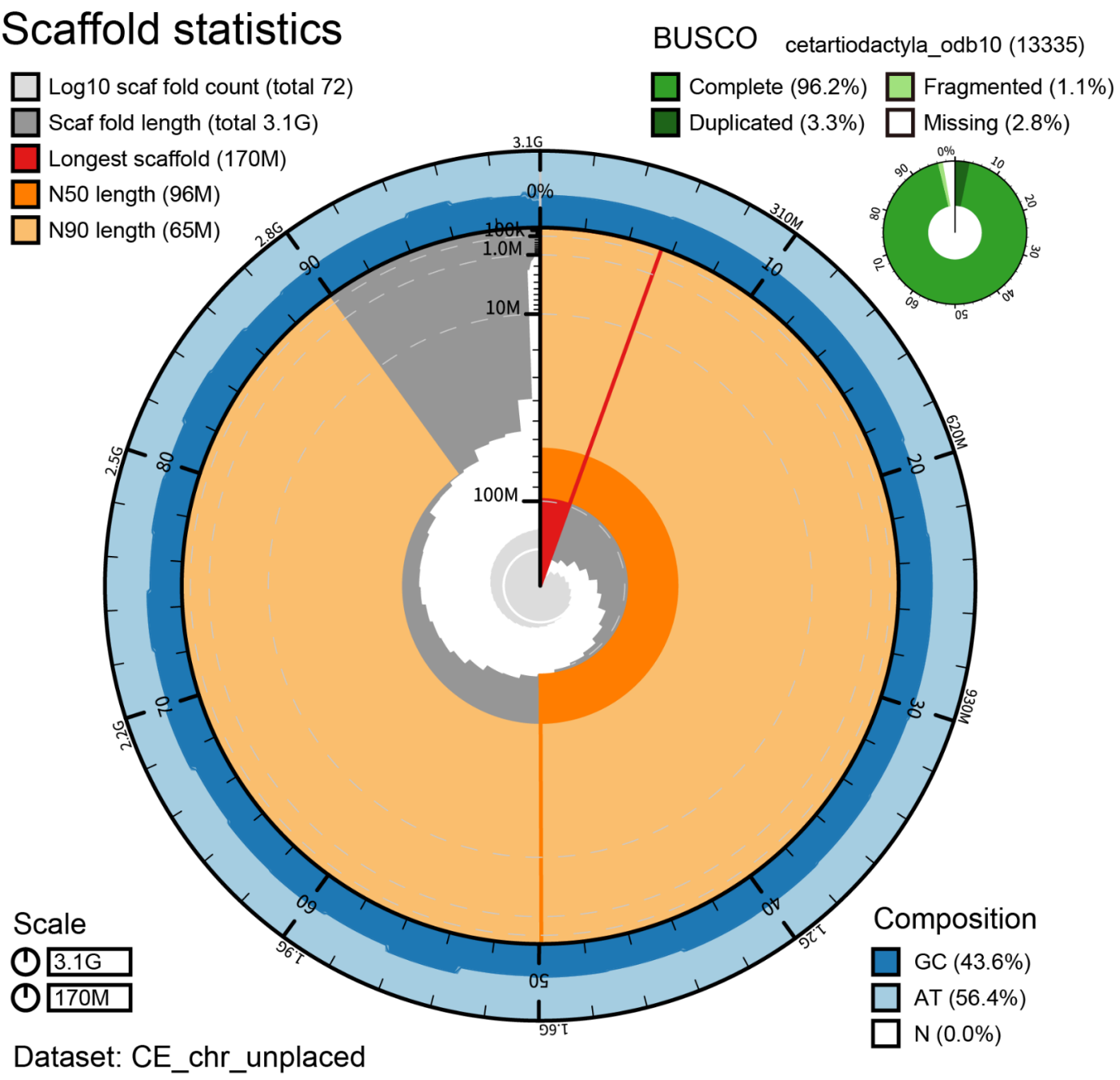


Figure 2

Assembly statistics and completeness assessment of the chromosome-level Gansu red deer genome. The BlobToolKit snail plot summarizes scaffold length distribution, assembly size, N50 and N90

statistics, nucleotide composition, and genome completeness assessed using BUSCO with the cetartiodactyla_odb10 lineage dataset.

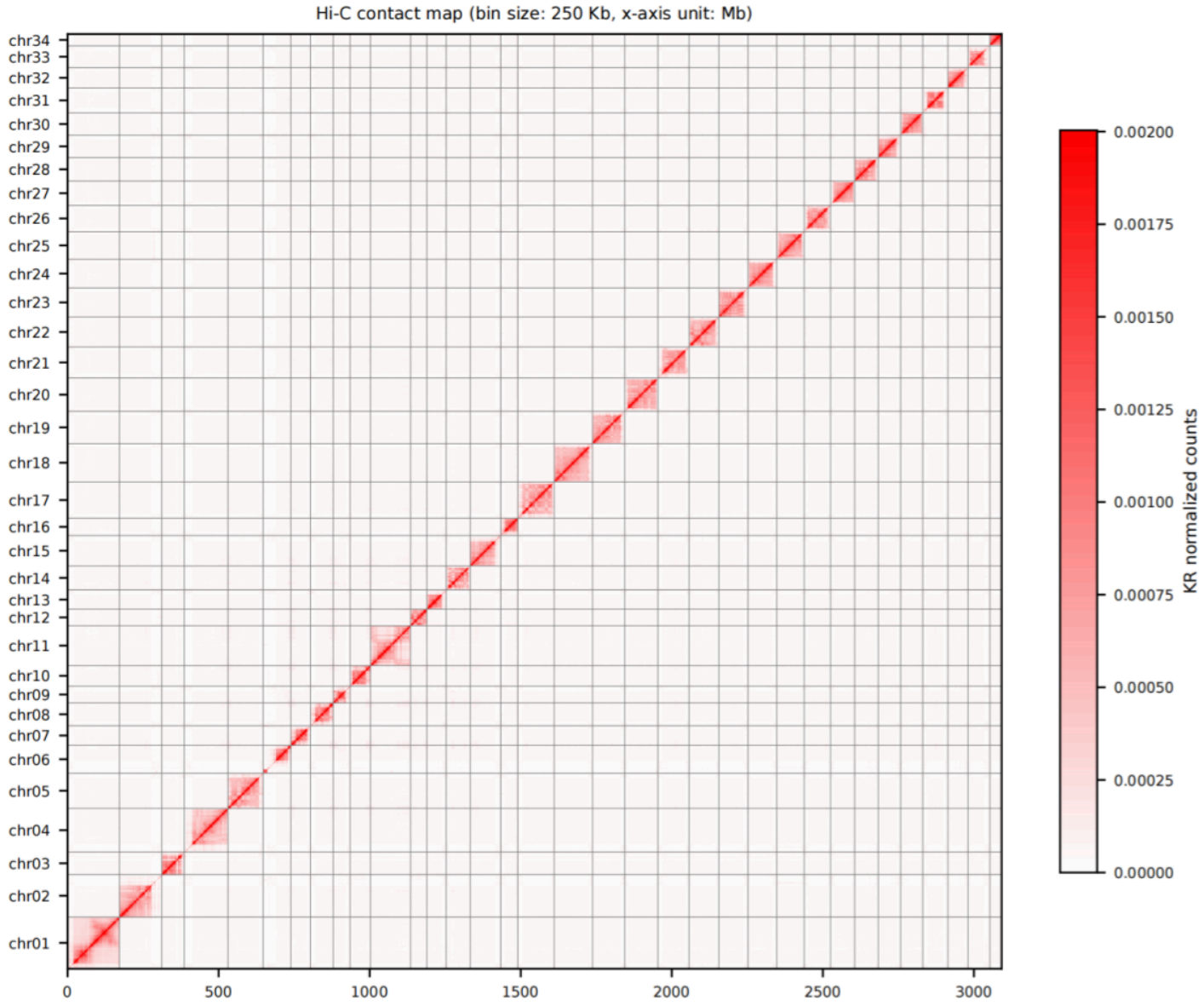


Figure 3

Hi-C contact map of the chromosome-level Gansu red deer genome assembly. Hi-C interaction frequencies among the 34 pseudochromosomes are shown at a bin size of 250 kb. The x-axis represents cumulative genomic coordinates in megabases (Mb), and the color intensity indicates KR-normalized Hi-C contact counts.

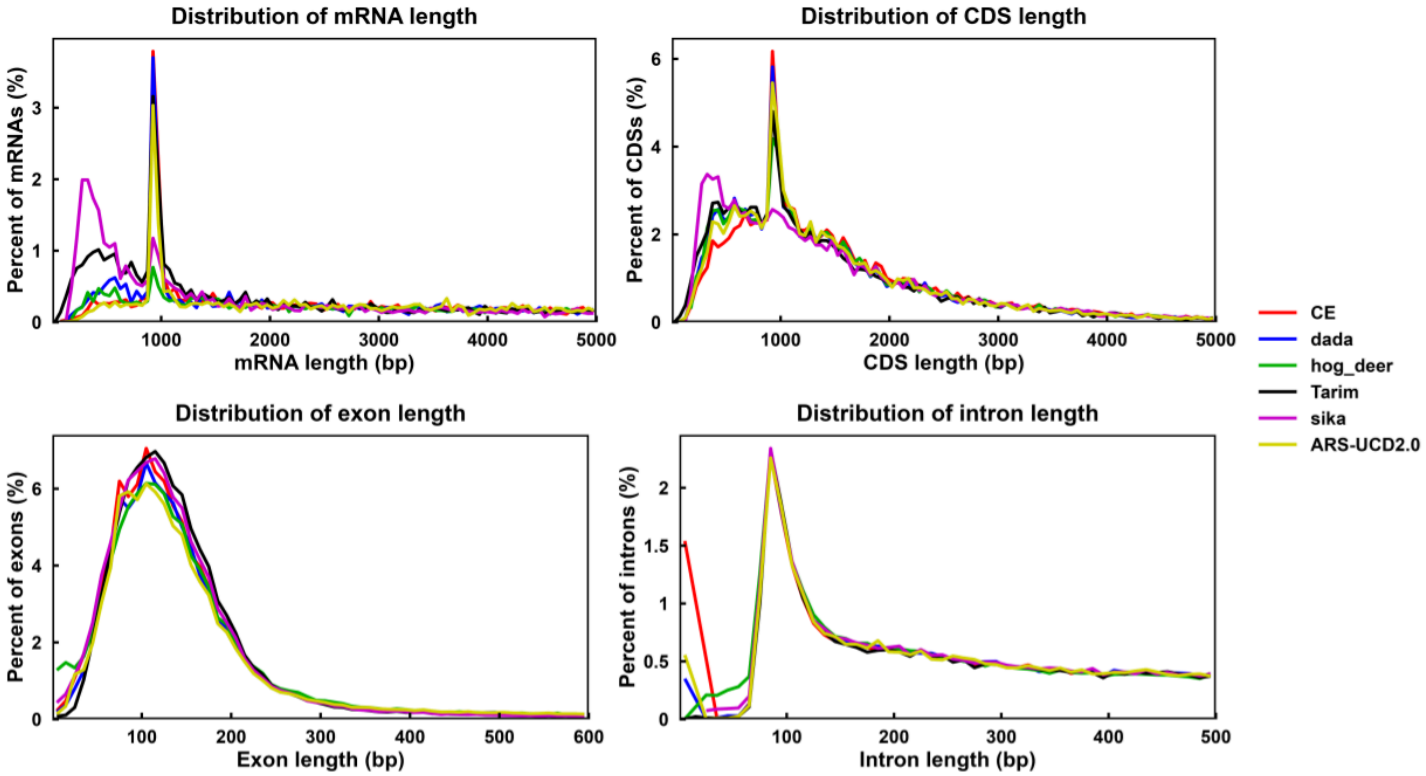


Figure 4

Comparison of gene structure length distributions among Gansu red deer and related species. Distributions of mRNA (a), CDS (b), exon (c), and intron (d) lengths were compared among the Gansu red deer and five reference species. Length distributions were summarized using a window size of 50 bp for mRNA and CDS and 10 bp for exon and intron.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CEAdditionalfile1.docx
- CEAdditionalfile2.xlsx