Scientific Data

https://doi.org/10.1038/s41597-026-06815-1

Article in Press

Telomere to telomere level genome assembly of the Yarkand hare (*Lepus yarkandensis*)

Received: 1 August 2025

Accepted: 3 February 2026

Cite this article as: Xu, M., Cui, Y., Kuang, H. *et al.* Telomere to telomere level genome assembly of the Yarkand hare (*Lepus yarkandensis*). *Sci Data* (2026). https://doi.org/10.1038/s41597-026-06815-1

Mengqi Xu, Yuge Cui, Hongcheng Kuang, Kai Wei & Wenjuan Shan

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Telomere to telomere level genome assembly of the Yarkand hare (*Lepus yarkandensis*)

Mengqi Xu, Yuge Cui, Hongcheng Kuang, Kai Wei*, Wenjuan Shan*

Xinjiang Key Laboratory of Biological Resources and Genetic Engineering, College of Life Science and Technology, Xinjiang University, Urumqi 830017, China.

*Address correspondence to Wenjuan Shan and Kai Wei. E-mail: swj@xju.edu.cn and kaiwei@xju.edu.cn

The Yarkand hare (*Lepus yarkandensis*) is endemic to the Tarim Basin in Xinjiang, China. It is a key species and a critical component of the Tarim Basin ecosystems. However, the lack of a reference genome has hindered evolutionary and genetic studies of this species. Here, we assembled a telomere-to-telomere (T2T) genome of the Yarkand hare (LepYark_1.0) using PacBio HiFi, Nanopore, and Hi-C sequencing. The assembled genome size is approximately 2.70 Gb, with a scaffold N50 of 126.86 Mb. About 94.88% of the assembled sequences could be anchored to 24 pseudo-chromosomes, with a BUSCO assessment indicating a completeness of 99.0%. Repetitive sequences comprise 46.38% of the genome, with short interspersed nuclear elements (SINEs) accounting for the largest proportion. Additionally, we identified 24 centromeres and 46 telomeres. 32,298 protein-coding genes were annotated using *de novo* prediction and transcriptome data, functionally annotating 85% of them. This genome assembly provides genomic resources for studies on conservation, adaptive evolution and the exploration of genetic basis related to important traits of the Yarkand hare.

Background & Summary

The Tarim Basin (Xinjiang, China) is characterized by an arid, low precipitation temperate continental climate, with significant diurnal temperature variations[1].The Yarkand hare (*Lepus yarkandensis*) belongs to the genus (*Lepus*) and is endemic to China's Tarim Basin and surrounding areas[2]. The Yarkand hare has a small, lightweight body and the sandy-colored fur that blends seamlessly with its habitat. In addition, it possesses exceptionally developed auditory organs, which are proportionally smaller than those of other hare species in Xinjiang[3]. Thus, this species has formed highly specialized characteristics adapted to their environment [4].

Research on the Yarkand hare primarily focuses on its adaptation to the extreme environmental conditions in the Tarim Basin. For example, measurements of digestive enzyme activity in the pancreas and intestines revealed that the activity of intestinal amylase and cellulase was higher in Yarkand hare than in domestic rabbits (*Oryctolagus cuniculus*) [5,6]. Techniques such as immunochemistry, aquaporin staining, real-time quantitative PCR, and protein immunoblotting were employed to detect mRNA and protein expression levels of Aquaporin (AQP) family genes[7]. Transcriptome sequencing of the Yarkand hare's and domestic rabbit's kidneys revealed the molecular mechanisms underlying renal water reabsorption, establishing a transcriptomic basis for Yarkand hare's adaptation to high salinity, elevated temperatures, and arid conditions [8].

Some environmental adaptive genes of the Yarkand hare were identified by whole-genome single-nucleotide polymorphism (SNP) analysis using Specific-Length Amplified Fragment Sequencing (SLAF-seq). And the functions of these candidate genes across diverse habitats, including altitudinal adaptation were annotated[9]. Subsequently, the mitochondrial genome was sequenced, and positively selected genes associated with adaptation to arid environments were identified through Ka/Ks analysis[10]. However, the absence of a complete reference genome for the Yarkand hare significantly hinders research on the species' adaptation to extreme aridity, and its genetic characteristics remain unclear. Reference genomes are available for only four *Lepus* species: the European hare (*L. europaeus*)[11], mountain hare (*L. timidus*)[12], woolly hare (*L. oiostolus*)[13], and Cape hare (*L. capensis*)[14]. This limited genomic representation hinders the investigation of evolutionary history and genetic diversity patterns of the genus *Lepus*. Therefore, developing a chromosome-level reference genome for the Yarkand hare is essential for future research.

Multiple mammalian telomere-to-telomere (T2T) genomes have been reported since the complete human (*Homo sapiens*) T2T genome (T2T-CHM13) was published[15]. Utilizing integrated sequencing technologies, researchers assembled a complete goose (*Anser cygnoides*) T2T genome, known as Taihu_goose_T2T_genome, providing critical foundations for genetic improvement and the utilization of trait inheritance mechanisms[16]. A gapless sheep (*Ovis aries*) T2T assembly (T2T sheep1.0) containing both sex chromosomes was also released, creating an essential genomic resource for ovine research. This assembly enabled the resequencing of 810 wild and domesticated sheep to identify genes associated with wool production[17]. The complete mouse (*Mus musculus*) T2T genome designated GRCm39, features detailed analysis of ribosomal DNA and centromeric repeats, demonstrating how reference genomes of model organisms can resolve gaps in homologous sequence knowledge[18]. Therefore, a T2T-level reference genome is indispensable for investigating both the genetics of environmental adaptation and the genome architecture of the Yarkand hare.

This study assembled a T2T-scale genome of the Yarkand hare by integrating Illumina, PacBio HiFi, and Nanopore sequencing data using Hi-C technology. The final genome assembly spanned 2.70 Gb and had an N50 of 126.86 Mb. Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness and Quality Value (QV) reached 99.0% and 67.81, respectively. The total assembly was assembled into 24 chromosomes. Genome annotation predicted 32,298 protein-coding genes. The availability of such complete and high-quality genome assemblies is essential for in-depth basic biological research. This study provides a valuable genomic resource for investigating the molecular mechanisms and evolutionary processes of the Yarkand hare.

Methods

Ethics Statement. The Animal Care and Use Committee of the College of Life Sciences and Technology, Xinjiang University, Urumqi, China, approved the experimental protocol used in this study (XJUAE-2023-020).

Materials. A female Yarkand hare, captured in the Aksu region of Xinjiang Uygur Autonomous Region (41.0°N, 82.86°E), was used for whole-genome sequencing. Leg skeletal muscle tissue was obtained, flash-frozen in liquid nitrogen, and used for genomic DNA extraction using the OMEGA Tissue DNA Kit (China, Shanghai). Additionally, samples from six fresh tissues (skeletal muscle, heart, renal cortex, renal medulla, colon, and small intestine), collected from the same individual, were flash-frozen in liquid nitrogen and used to assist in genomic annotation.

Library Establishment and Genome Sequencing. The Hi-C library was constructed using 1 μg of starting DNA. A separate Hi-C library was prepared using the MGIEasy Universal DNA Library Prep Kit V1.0 (CAT#1000005250, MGI) following the standard protocol. Clusters were generated by bridge amplification with MGIEasy DNA Clean Beads (CAT#1000005279, MGI) and sequenced on the DNBSEQ-T7RS platform. To anchor hybrid scaffolds to chromosomes, genomic DNA was extracted from fibroblasts derived from the same muscle tissue used for initial genomic DNA sequencing. This DNA was used to construct a Hi-C library. DNA purification was performed using the QIAamp DNA Mini Kit (QIAGEN)[19] according to the manufacturer's instructions. Additionally, a Hi-C library compatible with Illumina sequencing was prepared using the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB) according to the manufacturer's instructions. The final library was sequenced on the Illumina NovaSeq platform[20].

SMRTbell target size libraries were constructed for sequencing according to PacBio's standard protocol (Pacific Biosciences, CA, USA) using 15kb preparation solutions. Genomic DNA was cut using g-TUBEs (Covaris, USA) to achieve fragments of the desired size for library preparation. The fragments were ligated with hairpin adaptors for PacBio sequencing. Subsequently, the library was treated with nuclease using the SMRTbell Enzyme Cleanup Kit and purified with AMPure PB Beads[21]. Sequencing was performed on a PacBio Sequel II instrument using Sequencing Primer

V5 and the Sequel II Binding Kit 2.2 at the Genome Center of Biozeron, Shanghai, China.

ONT regular DNA was extracted using Genomic and BAC-long DNA kits following the manufacturer's guidelines. Target fragments in high-quality DNA were size-selected using a BluePippin fully automated nucleic acid size selection system, followed by damage repair and end-repair. DNA libraries (approximately 400 ng) were constructed and sequenced on the PromethION platform (Oxford Nanopore Technologies)[22] at the Genome Center of Biozeron.

Total RNA was extracted from the six tissues using TRIzol reagent (TIANGEN) on dry ice following the manufacturer's protocols. Approximately 300 ng of total RNA was reverse-transcribed into cDNA and amplified using the NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module and the Iso-Seq Express Oligo Kit. The resulting cDNA was purified using ProNex Beads. SMRTbell libraries were constructed using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences), which included damage repair, end repair, A-tailing, and ligation to sequencing adapters. Finally, the SMRTbell template was annealed to a sequencing primer, bound to polymerase, and sequenced on the PacBio Sequel II platform using the Sequel II Binding Kit 2.0 (Pacific Biosciences).

Approximately 296 Gb of Hi-C reads, 156 Gb of PacBio HiFi reads, 265 Gb of filtered Illumina short reads, and 293 Gb of Nanopore reads were obtained and used to assemble the Yarkand hare genome. Detailed statistics of the sequenced reads are provided in Supplementary Tables S1–S4.

Genome Survey and Assembly. The genome size, heterozygosity, and repetitive content were estimated using JCVI utility libraries v1.2.1[23] based on K-mer depth. The estimated genome size of the Yarkand hare is approximately 2.4 Gb, as determined from the 21-mer histogram (Fig. 1a, 1b). The genomic K-mer frequency distribution plot is shown in Fig 1. The K-mer genome complexity assessment results are presented in Supplementary Table S5.

A rapid T2T assembly was performed using Hifiasm v0.19.9-r616[24] with PacBio HiFi, Oxford Nanopore, and Hi-C reads. Genome scaffolds were anchored and ordered using ALLHiC v0.9.8[25]. Hi-C-assisted assembly leverages the principle that *cis* interactions (within the same chromosome) occur more frequently than *trans* interactions (between chromosomes), and that the interaction strength decreases with increasing linear genomic distance. Based on these principles, contigs or scaffolds were clustered, ordered, and oriented to obtain a chromosome-level genome assembly.

After completing the genome assembly, we performed gap filling with TGS-GapCloser v1.2.1 (https://github.com/BGI-Qingdao/TGS-GapCloser.git), using error-corrected ONT data. The basic default parameters were: tgsgapcloser --scaff scaffold.fasta --reads ont_corr.fasta --output output --ne > pipe.log. Manual telomere extension and individual gap filling were performed.

The final assembled genome size was 2.84 Gb, with 2.70 Gb of the sequence anchored to 24 autosomes. The scaffold N50 was 126.86 Mb (Table 1, Fig. 1c). Chromosome lengths and related statistics are presented in Table 1, and details of the chromosomes based on the genome assembly are shown in Table 2.

Centromere and Telomere Candidates Prediction. Telomere candidates were predicted using seqtk v1.4[26] based on the telomeric repeat sequence "AACCCT." Centromeres were predicted using quarTeT v1.2[27], which utilized raw HiFi and Hi-C sequencing data and the genome assembly. Our analysis predicted that the Yarkand hare chromosomes possessed 24 centromeres and 46 telomeres (Fig. 2 and Table 1). Details of the centromere sequence are provided in Supplementary Table S5.

Genome Quality Control. Second- and third-generation sequencing data were used to build a 21-mer library with Merfin v1.0 to determine the QV of the genome assembly[28]. Subsequently, Merqury v1.3[29] evaluated the QV of the genome using K-mers, as it can be assessed without a reference genome or database. The overall QV for the HiFi-based assembly was 67.8106, and the overall error rate was 1.65553e–07. Gene completeness was assessed using BUSCO with Compleasm v5.3.2[30]. This analysis predicted gene profiles present in the assembly using single-copy homologs in the OrthoDB database (glires_odb12)[31], thereby assessing the completeness of the assembled genome. Detailed results are shown in Table 2.

Genome assembly	
Sequence Number	24
Total Length(bp)	2,695,446,921

N50 length (bp)	126,863,262
N90 length (bp)	67,642,775
Max. length	187,498,641
Min. length	37,355,584
GC Content (%)	43.59 %
N rate (%)	0%
Anchored rate (%)	94.88
QV	67.81
BUSCO evaluation	C: 99.0% [S:96.7%, D:2.3%], F:0.3%, M:0.7%, n: 12556

Table 1. Statistics of the Assembled Genome.

#ID	Length	N %	GC_content	Telomere				Centromere	
				5'telomere_start	5'telomere_end	3'telomere_start	3'telomere_end	Start	End
Chr 1	187,498,641	0	41.82%	1	2,230	187,498,227	187,498,641	186535258	187479635
Chr 2	179,041,502	0	41.77%	1	984	179,039,302	179,041,502	1	5445519
Chr 3	173,315,621	0	42.03%	1	308	173,307,126	173,315,621	141386085	156791811
Chr 4	171,042,741	0	43.28%	1	7,431	171,036,546	171,042,741	102929562	105450978
Chr 5	155,715,540	0	45.42%	1	1,864	155,715,124	155,715,540	1	2584200
Chr 6	144,382,476	0	42.33%	1	1,092	144,381,520	144,382,476	94774254	96105810
Chr 7	137,418,895	0	41.09%	1	1,488	137,416,116	137,418,895	134703161	137388984
Chr 8	136,778,091	0	40.72%	1	1,312	136,776,955	136,778,091	90706622	97364953
Chr 9	126,863,262	0	44.35%	1	2,595	126,859,960	126,863,262	50632240	62182970
Chr 10	126,574,618	0	45.07%	1	1,209	126,573,138	126,574,618	54635295	61855743
Chr 11	123,190,878	0	43.59%	1	474	123,190,535	123,190,878	106391634	123184868
Chr 12	108,959,477	0	43.70%	1	756	108,958,178	108,959,477	105050227	105565452

Chr 13	108,422,781	0	44.72%	1	3,447	108,418,157	108,422,781	88035726	93173592
Chr 14	107,850,124	0	43.36%	1	5,305	-	-	103174396	107714801
Chr 15	101,085,814	0	43.99%	1	2,831	101,084,451	101,085,814	99811169	100245317
Chr 16	98,871,952	0	42.31%	1	5,296	98,871,304	98,871,952	32502514	36763560
Chr 17	92,863,345	0	41.63%	1	1,152	92,862,046	92,863,345	89949193	92860594
Chr 18	86,993,330	0	45.30%	1	380	86,992,636	86,993,330	4194891	5499178
Chr 19	67,642,775	0	48.24%	1	641	67,640,957	67,642,775	45546410	65092632
Chr 20	67,438,080	0	44.90%	1	479	-	-	7022875	11997147
Chr 21	63,882,881	0	48.76%	1	3,237	63,881,472	63,882,881	38236103	40595669
Chr 22	51,322,076	0	45.82%	1	1,191	51,305,781	51,322,076	40514778	40852393
Chr 23	40,936,437	0	50.18%	1	5,557	40,936,019	40,936,437	15495581	17446938
Chr 24	37,355,584	0	50.17%	1	1,671	37,354,202	37,355,584	30170514	37190389

Table 2. Assembly Sequence Length Statistics

Functional Gene Annotation. We employed a combination of *de novo*, homology-based, and transcriptome-based evidence for gene prediction in the Yarkand hare genome. The specific steps were as follows:(1) *De novo* gene prediction was performed using AUGUSTUS v3.2.3[32]. (2) Homology-based prediction utilized protein sequences from eight reference species: *L.europaeus* (GCA_033115175.1), *Peromyscus eremicus* (GCF_949786415.1), *Oryctolagus cuniculus* (GCA_000003625.1), *Ochotona princeps* (GCF_030435755.1), *Homo sapiens* (GCA_000001405.29), *Mus musculus* (GCF_000001635.27), and *Canis lupus familiaris* (GCF_011100685.1). Protein sequences were first aligned to the draft genome assembly using tblastn. The resulting alignments were refined and used to determine the exact gene structures (coding regions and introns) with Genewise v2.4.1[33].(3) Transcriptome-based prediction involved aligning RNA-seq reads to the genome assembly using TopHat v2.1.1[34], followed by transcript assembly using Trinity v2.11.0[35].(4) Evidence integration: The gene sets generated from the above three methods were integrated into a consensus set of protein-coding genes using EVidenceModeler (EVM) v1.1.1[36].(5) Gene set evaluation: The completeness of the final predicted gene set was assessed using BUSCO v6.0.0[37] with the metazoa_odb10 dataset. The predicted protein sequences were functionally annotated by comparing them against the Nr (Non-Redundant Protein Sequence Database)[38], Swiss-Prot (UniProt consortium)[39], KEGG (Kyoto Encyclopedia of Genes and Genomes)[40], and GO (Gene Ontology)[41] databases using DIAMOND BLAST v2.8.14[42]. Functional gene statistics are presented in Table 3 and Fig. 1d.

Gene	
Number of intron in gene	170879
Number of single exon gene	7056
mean exons per gene	6.9
mean introns in per gene	5.9
Total gene length	1022599349
Total exon length	36768105
Total intron length per gene	982717859
mean gene length	35563.73
mean cds length	1278.71
mean exon length	184.18
mean intron in cds length	5750.96
Longest gene	479223
Longest cds	113091
Longest exon	14832
Longest intron into cds part	332014
Shortest gene	201

Table 3． Summary of genes identified in the assembled genome.

Genomic Component Analysis. Repeat sequences were annotated using both *ab initio* prediction and homology-based methods. For *ab initio* prediction, a library of repeat sequences was created using RepeatModeler v2.0.5[43], while unclassified repeat sequences were annotated using

TEclass2[44]. Repeat sequences, annotated using RepeatMasker v4.1.6[45], revealed that 46.38% (1,250,041,673 bp) of the assembled genome comprised repetitive sequences, with Long Interspersed Nuclear Elements (LINEs) being the predominant type (Table 4). Non-coding RNA identification was done using tRNAscan-SE v2.0.7[46] to predict tRNAs, several reference tRNAs were extracted and compared and filtered (including: GCA_010411085.1, GCA_000003625.1, GCA_016077325.2, GCA_030254825.1, GCA_030435755.1, GCA_033115175.1, GCA_036321535.1, GCA_903992535.2), followed by RNAmmer v1.2[47] for rRNA prediction. Finally, Rfam v15.0[48] was used for comparative annotation against the Rfam database, using its CMsearch program[49] with default parameters to identify the final small RNAs (sRNAs), small nuclear RNAs (snRNAs), and microRNAs (miRNAs). The predicted non-coding RNA data are presented in Table 5.

Type	The number of TE-related sequence fragments (#)	Total Length (bp)	In Genome (%)	Average length(bp)
LTR	574,355	131,721,827	4.8868	235
DNA	769,647	85,115,707	3.1578	123
LINE	984,034	397,414,876	14.7439	413
SINE	2,200,778	518,351,016	19.2306	249
RC	18,924	1,989,152	0.0738	109
scRNA	9	544	0	60
Unknown	447,544	143,437,218	5.3215	398
Total	4,995,291	1,250,041,673	46.376	273

table 4. Scattered repeat sequence results statistics

	Type	Number#	Avg_Len	Total_Len	% in Genome
	tRNA	497	74	36,986	0.0014
rRNA_de	5S	568	111	63,279	
	5.8S	0	0	0	
	18S	7	1,868	13,079	0.0049
	28S	7	7,818	54,729	
rRNA_ho	sRNA	3	67	201	0
	snRNA	1,040	126	131,240	0.0049
	miRNA	412	84	34,941	0.0013

Table 5. ncRNA result statistics

Genome Synteny Analysis. Synteny analysis using the MCScanX[50] module in TBtools v2.309[51] revealed significant genetic conservation between the domestic rabbit[52] (*Oryctolagus cuniculus*; genome assembly UM_NZW_1.0) and the European hare (*L. europaeus*; genome assembly mLepTim1.pri; Fig. 4). The results indicated that chromosome eight corresponds to the X chromosome. Notably, the Yarkand and European hares exhibited high chromosomal homology and identical chromosome-numbering systems. The Yarkand hare's chromosome numbering diverged from the domestic rabbit despite overall high genomic homology. These results suggest that the Yarkand and European hares retained greater karyotype conservation during chromosome evolution.

Data Record

The assembled genome has been deposited at GenBank under accession number GCA_047496845.1[53].

Data on gene functional and repeat annotation have been deposited at figshare[54].

The genomic Illumina sequencing data were deposited in the SRA at NCBI SRR36906164[55].

The genomic PacBio sequencing data were deposited in the SRA at NCBI SRR36906168[55].

The genomic Nanopore sequencing data were deposited in the SRA at NCBI SRR36906166[55], SRR36906167.

The Hi-C sequencing data were deposited in the SRA at NCBI SRR36906169[55].

The RNA sequencing data were deposited in the SRA at NCBI SRR36906165[55].

Data Availability

The assembled genome data that support the findings of this study are openly available in NCBI at https://identifiers.org/ncbi/insdc.gca:GCA_047496845.1 [53].

The annotation data that support the findings of this study are openly available in figshare at https://doi.org/10.6084/m9.figshare.29369999.v1 [54]

The Illumina sequencing data that support the findings of this study are openly available in NCBI

at https://identifiers.org/ncbi/insdc.sra:SRR36906164 [55]
The PacBio sequencing data that support the findings of this study are openly available in NCBI at https://identifiers.org/ncbi/insdc.sra:SRR36906168 [55]
The Nanopore sequencing data that support the findings of this study are openly available in NCBI at https://identifiers.org/ncbi/insdc.sra:SRR36906166 [55], https://identifiers.org/ncbi/insdc.sra:SRR36906167[55]
The Hi-C sequencing data that support the findings of this study are openly available in NCBI at https://identifiers.org/ncbi/insdc.sra:SRR36906169 [55]
The RNA sequencing data that support the findings of this study are openly available in NCBI at https://identifiers.org/ncbi/insdc.sra:SRR36906165 [55]

Technical Validation

Multiple methods were employed to verify the accuracy and integrity of the Yarkand hare genome. First, the Hi-C heatmap demonstrated high consistency across all chromosomes, reflecting precise sequencing, ordering, and orientation of contigs in the assembly (Fig. 1c). Second, 24 centromeres were mapped to the 24 chromosomes, and 46 telomeres were identified. Third, Illumina sequencing data were aligned to the genome using BWA v0.7.17[56], achieving a 98.30% mapping rate. Fourth, alignment using minimap2 v2.24[57] demonstrated that 99.10% of ONT reads and 99.67% of HiFi reads mapped to the T2T assembly. Finally, the QV was 67.81 (Table 2), with genome assembly integrity and protein-coding gene completeness assessed using BUSCO and the glires_odb12 single-copy ortholog set from OrthoDB, identifying 94.8% of the 12,556 conserved orthologs (Table 1). We compared the genomic assembly data from this study with those from other lagomorph species, as detailed in Table 6, the comparison results of T2T genome assemblies with other species are presented in Supplementary Table S6. In conclusion, the assembled Yarkand hare T2T genome is highly complete and accurate. This high-quality genome provides a robust foundation for investigating the species' evolutionary and adaptive mechanisms in its arid environment.

species	genome size	Scaffold N50	Busco	Number of protein-	Repeat Sequence	Scaffold number	Sequencing technology

				coding genes	Coverage		
Lepus yarkandensis	2.70G	126.86 Mb	99.0%	32,298	46.38%	24	Illumina, PacBio HiFi, Nanopore, Hi-C
Lepus oiostolus	2.80G	64.25 Mb	96.2%	22,295	49.84%	24	Illumina, Hi-C
Lepus capensis	2.90G	124.44 Mb	98.2%	13, 868	46.13%	24	Illumina, PacBio HiFi, Hi-C
Oryctolagus cuniculus	2.88G	148.90 Mb	98.3%	22,674	47.09%	23	PacBio HiFi, ONT ultra-long, Hi-C
Ochotona princeps	2.33G	75.8 Mb	92.4%	21,186	26.22%	9,350	Illumina, Hi-C

Table 6. Comparison of assembly metrics, including *Lepus oiostolus*[13], *Lepus capensis*[14], *Oryctolagus cuniculus*[58] and *Ochotona princeps*[59].

Code availability

All software used in this work is in the public domain, with parameters being clearly described in Methods. If no detail parameters were mentioned for a software, default parameters were used as suggested by developer.

Reference

1. Sweet-Jones, et al. Genotyping and Whole-Genome Resequencing of Welsh Sheep Breeds Reveal Candidate Genes and Variants for Adaptation to Local Environment and Socioeconomic Traits. *Frontiers in genetics* **12**, 612492 (2021).
2. Tian, Haowen et al. Population genetic diversity and environmental adaptation of Tamarix hispida in the Tarim Basin, arid Northwestern China. *Heredity* **133**, 298-307 (2024).
3. Shan, W.J., et al. Genetic consequences of postglacial colonization by the endemic Yarkand hare (*Lepus yarkandensis*) of the arid Tarim Basin. *Chinese Science Bulletin* **56**, 1370-1382 (2011).
4. Weiss, Bruno, et al. Unraveling a Lignocellulose-Decomposing Bacterial Consortium from Soil Associated with Dry Sugarcane Straw by Genomic-Centered Metagenomics. *Microorganisms* **9**, 5995 (2021)
5. Wang, Jieru et al. Corrigendum: Genetic diversity, population structure, and selective signature of sheep in the northeastern Tarim Basin. *Frontiers in genetics* **14**, 1336294 (2023).
6. Hui, X.H. & Zhao, M.F. Analysis of characteristics of *Lepus yarkandensis* adapting to desert ecology. *Contemporary Animal Husbandry* **15**, 43-44 (2013).

7. Zhang J, et al. Higher expression levels of aquaporin (AQP)1 and AQP5 in the lungs of arid-desert living Lepus yarkandensis. *J Anim Physiol Anim Nutr (Berl)* **104**, 1186-1195 (2020).
8. Luo S, et al. Expression Regulation of Water Reabsorption Genes and Transcription Factors in the Kidneys of *Lepus yarkandensis*. *Front Physiol* **13**, 856427 (2022).
9. Li, Z., et al. Selective sweep analysis of the adaptability of the Yarkand hare (*Lepus yarkandensis*) to hot arid environments using SLAF-seq. *Animal genetics* **55**, 681-686 (2024).
10. Wang R, Tursun M, Shan W. Complete Mitogenomes of Xinjiang Hares and Their Selective Pressure Considerations. *Int J Mol Sci* **25**, 11925 (2024).
11. Michell, C, et al. High quality genome assembly of the brown hare (*Lepus europaeus*) with chromosome-level scaffolding. *Peer Community Journal* **4**, e26 (2024).
12. Marques JP, et al. An Annotated Draft Genome of the Mountain Hare (*Lepus timidus*). *Genome Biol Evol* **12**, 3656-3662 (2020).
13. Feng, S., et al. Chromosome-scale genome assembly of *Lepus oiostolus* (*Lepus*, Leporidae). *Scientific data* **11**, 183 (2024).
14. Dong, X., et al. A chromosome-level genome assembly of Cape hare (*Lepus capensis*). *Scientific data* **11**, 1081 (2024).
15. Nurk, S., et al. The complete sequence of a human genome. *Science (New York, N.Y.)* **376**, 44-53 (2022).
16. Zhao, H., et al. Telomere-to-telomere genome assembly of the goose Anser cygnoides. *Scientific data* **11**, 741 (2024).
17. Luo, L. Y., et al. Telomere-to-telomere sheep genome assembly identifies variants associated with wool fineness. *Nature genetics* **57**, 218-230 (2025).
18. Liu, J., et al. The complete telomere-to-telomere sequence of a mouse genome. *Science (New York, N.Y.)* **386**, 1141–1146 (2024).
19. Pavlova, A. S, et al Genomic DNA extraction protocol using DNeasy Blood & Tissue Kit (QIAGEN) optimized for Gram-Negative bacteria. https://doi.org/10.17504/protocols.io.paadiae (2018).
20. Modi A, et al. The Illumina Sequencing Protocol and the NovaSeq 6000 System. *Methods Mol Biol* **2242**, 15-42 (2021).
21. Duckworth AT, et al. Profiling DNA Ligase Substrate Specificity with a Pacific Biosciences Single-Molecule Real-Time Sequencing Assay. *Curr Protoc* **3(3)**, e690 (2023).
22. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics* **14**, 265-279 (2016).
23. Tang, H., et al. JCVI: A versatile toolkit for comparative genomics analysis. *iMeta* **3**, e211 (2024).
24. Cheng, H.Y., et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nature Biotechnology* **40**, 1332-1335 (2022).
25. Zhang, X.T., et al. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants* **5**, 833-845 (2019).
26. Shen, W., et al. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PloS one* **11**, e0163962 (2016).
27. Lin Y, et al. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res* **10**, 127 (2023).
28. Formenti, G., et al. Merfin: improved variant filtering, assembly evaluation and polishing via-mer validation. *Nature Methods*

19, 696-704 (2022).

29. Rhie, A., et al. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245 (2020).
30. Huang Neng & Heng Li. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics (Oxford, England)* **39**, 10 (2023).
31. Tegenfeldt, Fredrik et al. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic acids research* **53**, D516-D522 (2025).
32. Stanke, M., et al. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic acids research* **32**, 309-312 (2004).
33. Birney, E., Clamp, M., & Durbin, R. GeneWise and Genomewise. *Genome research* **14**, 988-995 (2004).
34. Kim, D., & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome biology* **12**, R72 (2011).
35. Cabau, C., et al. Compacting and correcting Trinity and Oases RNA-Seq *de novo* assemblies. *PeerJ* **5**, e2988 (2017).
36. Haas, B. J., et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, R7 (2008).
37. Simão, F. A., et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)* **31**, 3210-3212 (2015).
38. Pruitt, K. D., et al. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **35**, D61-65 (2007).
39. Boutet, E., et al. UniProtKB/Swiss-Prot. *Methods in molecular biology (Clifton, N.J.)* **406**, 89-112 (2007).
40. Kanehisa, M., & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30 (2000).
41. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic acids research* **47**, 330-338 (2019).
42. Persson E, Sonnhammer ELL. InParanoid-DIAMOND: faster orthology analysis with the InParanoid algorithm. *Bioinformatics* **38**, 2918-2919 (2022).
43. Flynn, J. M., et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451-9457 (2020).
44. Abrusán, G., et al. TEclass-a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatic*s **25**, 1329-1330 (2009).
45. Tarailo-Graovac, M., & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **Chapter 4**, 4.10.1-4.10.14 (2009).
46. Chan, P. P., & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods in molecular biology (Clifton, N.J.)* **1962**, 1-14 (2019).
47. Lagesen, K., et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research* **35**, 3100–3108 (2007).
48. Kalvari, I., et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic acids research* **46**, 335-342 (2018).
49. Cui, X., et al. CMsearch: simultaneous exploration of protein sequence space and structure space improves not only protein homology detection but also protein structure prediction. *Bioinformatics (Oxford, England)* **32**, 332-340 (2016).
50. Wang Y, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**. e49 (2012).

51. Chen C, et al. TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. *Mol Plant* **16**, 1733-1742 (2023).
52. Bai, Y, et al. Improving the genome assembly of rabbits with long-read sequencing. *Genomics* **113**, 3216–3223 (2021).
53. MengQi Xu, et al. *Genbank* https://identifiers.org/ncbi/insdc.gca:GCA_047496845.1 (2025)
54. MengQi Xu, et al. Telomere to telomere level genome assembly of the Yarkand hare (Lepus yarkandensis). *figshare* https://doi.org/10.6084/m9.figshare.29369999.v1 (2025)
55. NCBI Sequence Read Archive https://identifiers.org/ncbi/insdc.sra:SRR36906164 (2026)
56. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **1303**, 3997v2 [q-bio.GN] (2013).
57. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
58. Chen, Xuequn et al. Near telomere to telomere genome assembly of Chinese yellow rabbit (*Oryctolagus cuniculus*). *Scientific data* **12**,11786 (2025).
59. Sjodin, Bryson M F et al. Chromosome-Level Reference Genome Assembly for the American Pika (*Ochotona princeps*). *The Journal of heredity* **112**, 549-557 (2021).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32260116) and the Central Guidance for Local Projects of Xinjiang (ZYYD2024ZY04). Thanks to Associate Professor Linqiang Zhong of Xinjiang University for providing the photo of the Yarkand hare (Fig. 1a) and assisting in the sample collection.

Author contributions

In this study, Mengqi Xu was responsible for data collation and paper writing, Yuge Cui and Hongcheng Kuang were responsible for sample collection and collation, Kai Wei was responsible for technical guidance and paper revision, and Wenjuan Shan was responsible for outline writing, project management, and funding acquisition.

Competing interests

The authors declare no competing interests.

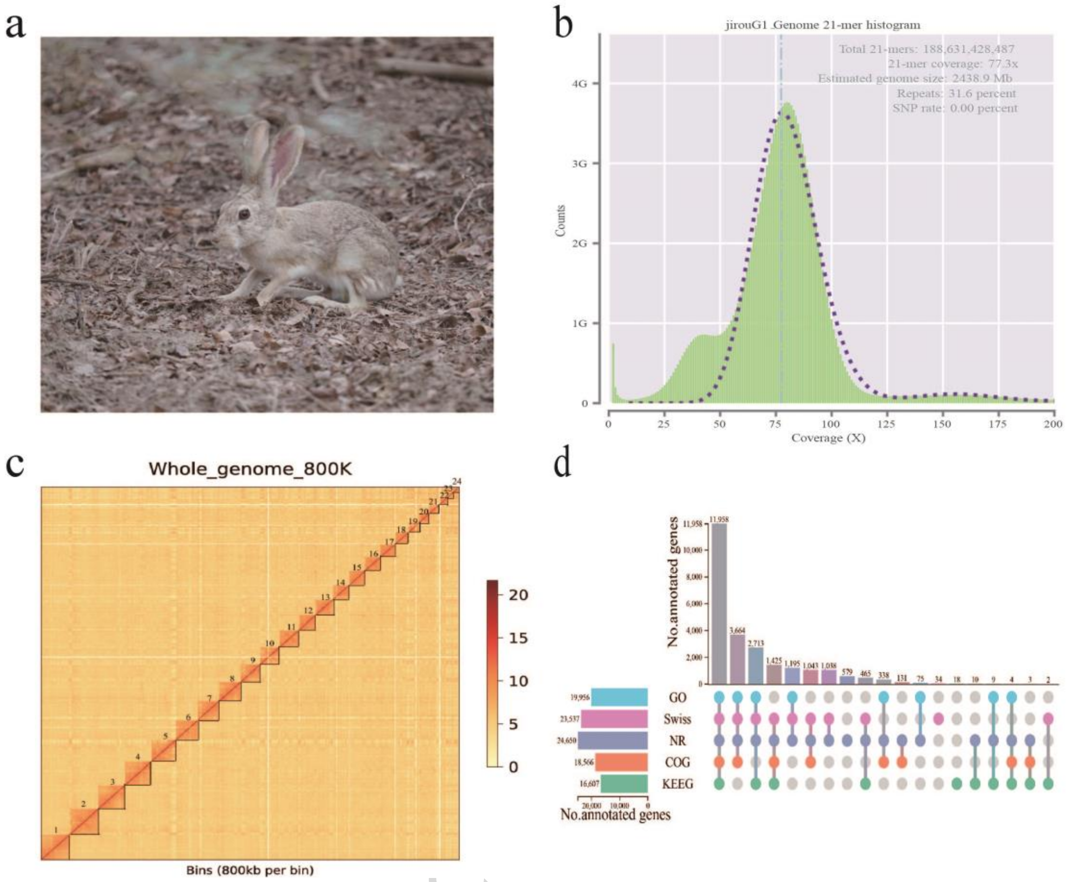

Fig. 1 Genome survey, heat map of the Hi-C contact matrix, and functional annotation of protein-coding genes in the Yarkand hare genome. (a) The Yarkand hare. (b) K-mer analysis of the Yarkand hare genome. (c) Heat map of the Hi-C contact matrix of the Yarkand hare genome assembly. (d) The upset bar plot shows functional annotations of protein-coding genes in the Yarkand hare genome. The left vertical bar represents the number of annotated genes, and the right vertical bar indicates the number of shared genes across the five databases.

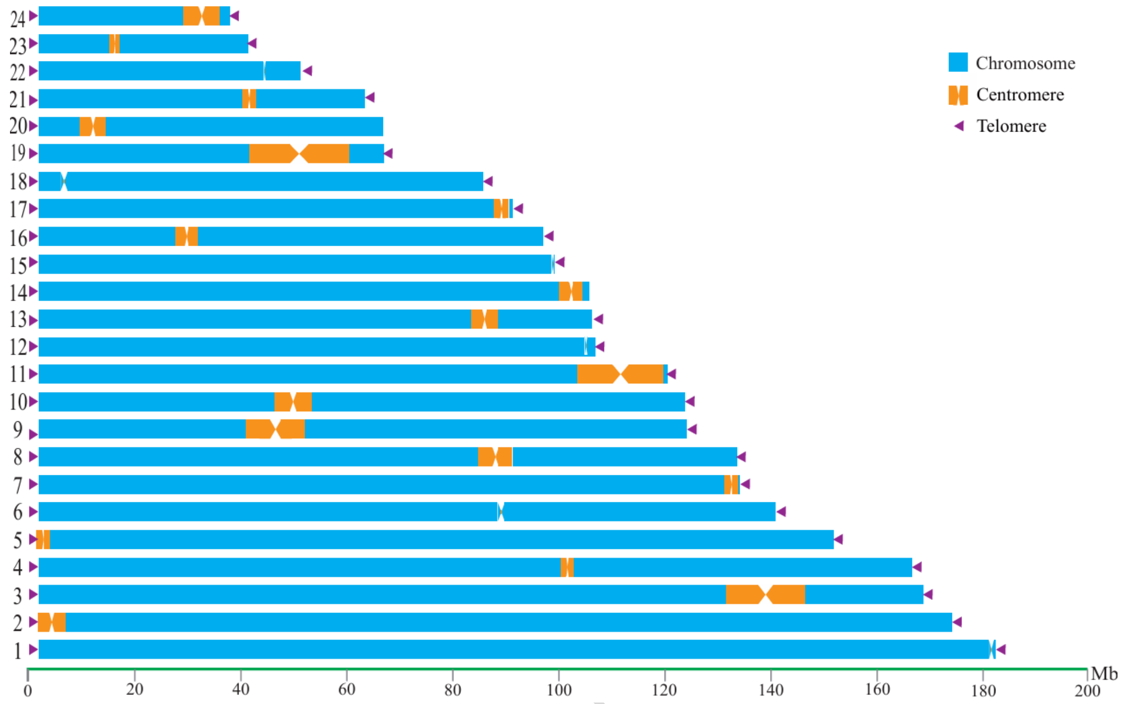

Fig. 2 An overview of the T2T gap-free reference genome of the Yarkand hare. Orange areas at both chromosome ends represent the telomere regions, and the gully area within each chromosome represents the centromere region.

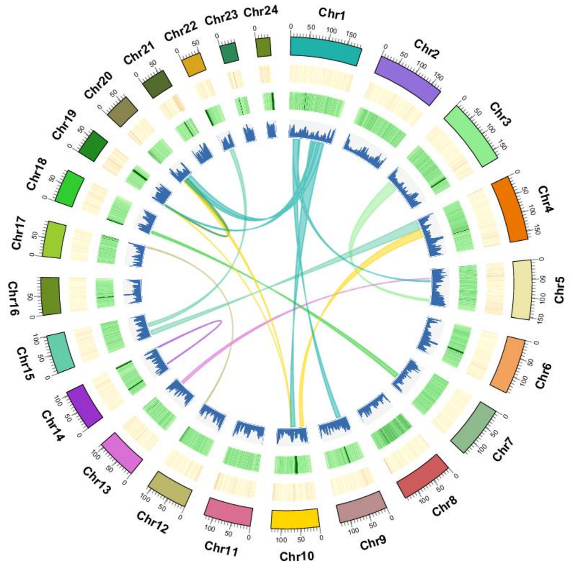

Fig. 3 Collinearity of the Yarkand hare assembled genome. From the outside in: first circle, chromosomes; second circle, gene density (window = 100 kb); third circle, REPEAT density

(window = 100 kb); fourth circle, GC content (window = 100 kb); inside, covariance within the genome obtained using MCScanX gene block analysis.

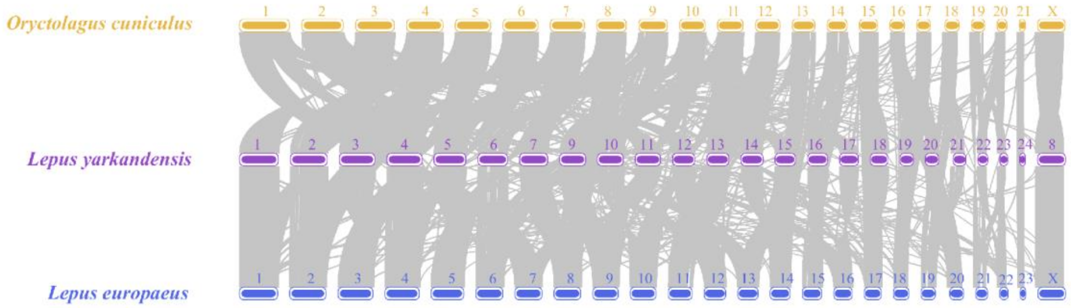

Fig. 4 Synteny analysis comparing the genome assembly of the Yarkand hare to the domestic rabbit and the European hare. Homologous genes are represented by grey lines connecting chromosomes.