scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of Turpan wonder gecko (Teratoscincus roborowskii) reveals olfactory candidate genes

Ping Yang, Chao Wei, Hongxiong Chang, Xinyang Li & Jia Wang

***Teratoscincus roborowskii*, an endemic species to China, is of particular interest due to its unique biological characteristics. It is a strictly nocturnal reptile that inhabits extremely harsh environmen. In this study, we generated the first chromosome level genome assembly of *T. roborowskii* using PacBio circular consensus sequencing (CCS) combined with Hi-C scaffolding. The genome assembly at the contig level was mounted onto chromosomes, with approximately 2,077.46 Mb of sequence allocated to 18 chromosomes, accounting for 99.92% of the total length. Among the sequences mapped to the chromosomes, approximately 2,070.96 Mb have determined order and orientation, making up 99.69% of the total length of the mapped sequences, with a contig N50 and scaffold N50 values of 117.38 Mb and 156.21 Mb respectively, and contains 21,341 predicted protein-coding genes, of which 99.62% were successfully annotated against public database. This high-quality genome not only fills a critical gap in genomic resources for this species, but also provides a valuable foundation for future study on its evolutionary history, gene functions and conservation biology.**

Background & Summary

T. roborowskii belongs to the genus *Teratoscincus* within the family Sphaerodactylidae. It is endemic to China and is found exclusively in the Turpan region of Xinjiang Province[1]. The Turpan Basin is characterized by an arid continental climate and topography, resulting in extreme heat and dryness during the summer[2]. *T. roborowskii* has adapted to survive in this harsh environment and exhibits several intriguing biological traits worthy of scientific investigation. *T. roborowskii* is a strictly nocturnal species. Studies in mammals suggest that nocturnal species often compensate for limited vision through enhanced olfactory capabilities. For example, elevated olfactory receptor gene expression has evolved in many nocturnal mammals as an adaptation[3]. In addition to its nocturnal behavior, *T. roborowskii* also possesses a remarkable ability to regenerate its tail, making it an interesting model for studying both sensory adaptation and tissue regeneration[4].

To date, research on *T. roborowskii* has focused primarily on its activity rhythms and dietary ecology, with limited molecular or genomic data available due to the absence of high-quality genomic resources. While the family Sphaerodactylidae includes approximately 237 extant species, genome data have only been published for two — *Euleptes europaea*[5,6] and *Sphaerodactylus townsend*[7,8]. Therefore, generating a chromosome-level genome assembly for *T. roborowskii* with detailed annotations is crucial for advancing our understanding of the genetic mechanisms underpinning traits with family Sphaerodactylidae species, as well as for *T. roborowskii* specifically.

In this study, we employed PacBio circular consensus sequencing (CCS), Illumina short-read sequencing, and Hi-C (high-throughput chromosome cconformation capture) technologies to generate the first chromosome-level genome of *T. roborowskii*. High-molecular-weight DNA was extracted from muscle tissue, and libraries were constructed and sequenced using protocols optimized for each sequencing platform. After quality control and filtering of low-quality reads, The genomes was assessed using the LACHESIS software. Assembly completeness was assessed using CEGMA (98.03%) and BUSCO (98.27%), both confirming a high level of genome completeness. Overall, this high-quality genome provides an important foundation for future studies on the evolutionary biology, gene function, and conservation of *T. roborowskii* and related gecko species.

Xinjiang Key Laboratory for Ecological Adaptation and Evolution of Extreme Environment Biology, College of Life Sciences, Xinjiang Agricultural University, Urumqi, 830052, China. ✉e-mail: wangjia365@xjau.edu.cn

Second-generation reads back than				Third-generations of reads back than			
Total reads	Mapped reads	Mapped (%)	Coverage (≥20X) (%)	Total Reads	Mapped reads	Mapped (%)	Coverage(≥20X) (%)
1,103,036,572	1,093,210,061	99.11	99.05	6,104,947	6,100,660	99.93	98.61

Table 1. Summary statistics of read mapping from second- and third-generation sequencing data.

Methods

Sample collection. Samples for genome sequencing were collected in Turpan City, Xinjiang Uygur Autonomous Region, China. Only individuals that died naturally during field investigations were used in this study. Fresh muscle tissue was obtained from an adult female *T. roborowskii* that had recently died. The tissue was immediately frozen in liquid nitrogen and cryopreserved. DNA was later extracted from the muscle tissue in the laboratory of the School of Life Sciences, Xinjiang Agricultural University[9], and subsequently sent to Beijing Baimaik for sequencing. All procedures were approved by the Ethics Committee of Xinjiang Agricultural University and complied with national and institutional guidelines for the care and use of animals.

Library construction and sequencing. For Illumina short-read sequencing, high-quality genomic DNA was fragmented into ~350 bp fragments using ultrasonic shearing (Covaris). The fragmented DNA underwent end repair, A-tailing, adapter ligation, size selection, and PCR amplification for library construction. Library quality and concentration were assessed using Qubit and Qseq. 400 systems to meet sequencing standards. Qualified libraries were immobilized on the flow cell via bridge PCR and sequenced on the Illumina platform (PE150) to generate paired-end reads.

For PacBio long-read sequencing, approximately 10 µg of high-molecular-weight genomic DNA (50 ng/µL, clear, without visible precipitation) was diluted in Elution Buffer to a final volume of 200 µL. DNA was sheared to ~20 kb using the Megaruptor system (Diagenode). Size distribution was assessed with an Agilent 2100 Bioanalyzer, and DNA concentration was measured by Qubit. SMRTbell libraries were then constructed using the PacBio Binding Kit (Pacific Biosciences), including the ligation of SMRTbell adapters, primer annealing, and polymerase binding. The final libraries were purified using AMPure PB beads and sequenced on the PacBio Sequel II platform.

For Hi-C library preparation, formaldehyde was used to cross-link proteins to DNA, stabilizing chromatin interactions through protein-mediated binding. This stabilizes the spatial proximity of chromatin segments, thereby preserving the native 3D genome architecture. The cross-linked DNA was digested with the restriction enzyme DpnII[10], generating sticky ends at the cleavage sites. These were subsequently end-repaired with biotin-labeled nucleotides to enable later purification and capture of ligated fragments. The DNA fragments containing chromatin interaction were then ligated, de-crosslinked, purified, and fragmented into 300–700 bp segments. Biotin-labeled fragments were isolated using streptavidin magnetic beads, and sequencing libraries were constructed. The quality and concentration of the libraries were evaluated using a Qubit fluorometer (v2.0) and an Agilent Bioanalyzer (v2100), while the effective Libraries that passed quality control were sequenced on the Illumina platform using paired-end 150 bp reads (PE150)[11].

In order to evaluate assembly completeness and sequencing coverage uniformity, Illumina short reads were aligned to the assembled genome using BWA (v0.7.17)[12], while PacBio HiFi long reads were mapped using Minimap2[13]. Key metrics such as mapping rate, genome coverage percentage, and sequencing depth distribution were analyzed. The second-generation (Illumina) data yielded 1,103,036,572 clean reads, with 99.05% of bases covered at a depth of ≥20×. The third-generation((HiFi) Reads produced 6,104,947 clean reads, with 98.61% of bases covered at a depth of ≥20× (Table 1).

Genome assembly. Raw data sequencing data were first filtered to remove low-quality reads, high-accuracy HiFi reads were then assembled using Hifiasm (v 0.19)[14] software. To begin with, we performed K-mer analysis using Illumina sequencing data and estimated the haploid genome size to be about 1.93 Gb with a heterozygosity of about 0.57% (Fig. 1B). Using Hifiasm (v 0.19)[14], we obtained an allelic assembly with a total genome size of 2.08 Gb and allele N50 of 117.38 Mb (Table 2). To identify and remove and contaminant sequences, the assembled genome was aligned against the NCBI NT database, as well as mitochondrial and plastid databases. Mitochondrial sequences and any detected contaminants were removed from the assembly HiC-based de-redundancy was subsequently performed to obtain the final genome assembly[14]. The completeness of the assembled genome was assessed using CEGMA pipeline[15], in combination with *tblastn*, GeneWise, and GeneID, to evaluate the presence of core conserved genes.

To achieve chromosome-level assembly, Hi-C reads were mapped to the contig-level genome using BWA (v0.7.17)[12]. Invalid data, such as self-ligation, no-ligation, and other invalid data including start sequences, PCR interaction pairs—including self-ligation, non-ligation, start-site artifacts, PCR duplicates, random fragmentation, and extreme-length fragments—were filtered out. Lachesis software was then employed to cluster, order, and orient the contigs based on the interaction signals captured by Hi-C data, ultimately producing a high-quality, chromosome-level genome assembly.

Leveraging Hi-C data, approximately 2,077.46 Mb of the genome was successfully anchored to 18 chromosomes, representing 99.92% of the total assembly. Of these anchored sequences, about 2,070.96 Mb (99.69%) were both ordered and oriented (Fig. 1C). These results are consistent with previous karyotype analysis of the species[16]. The genomic features of *T. roborowskii* are shown in a circle diagram (Fig. 1D). The results of synteny analysis showed that the proportion of collinear gene pairs between *T. roborowskii* and *E. europaea* was

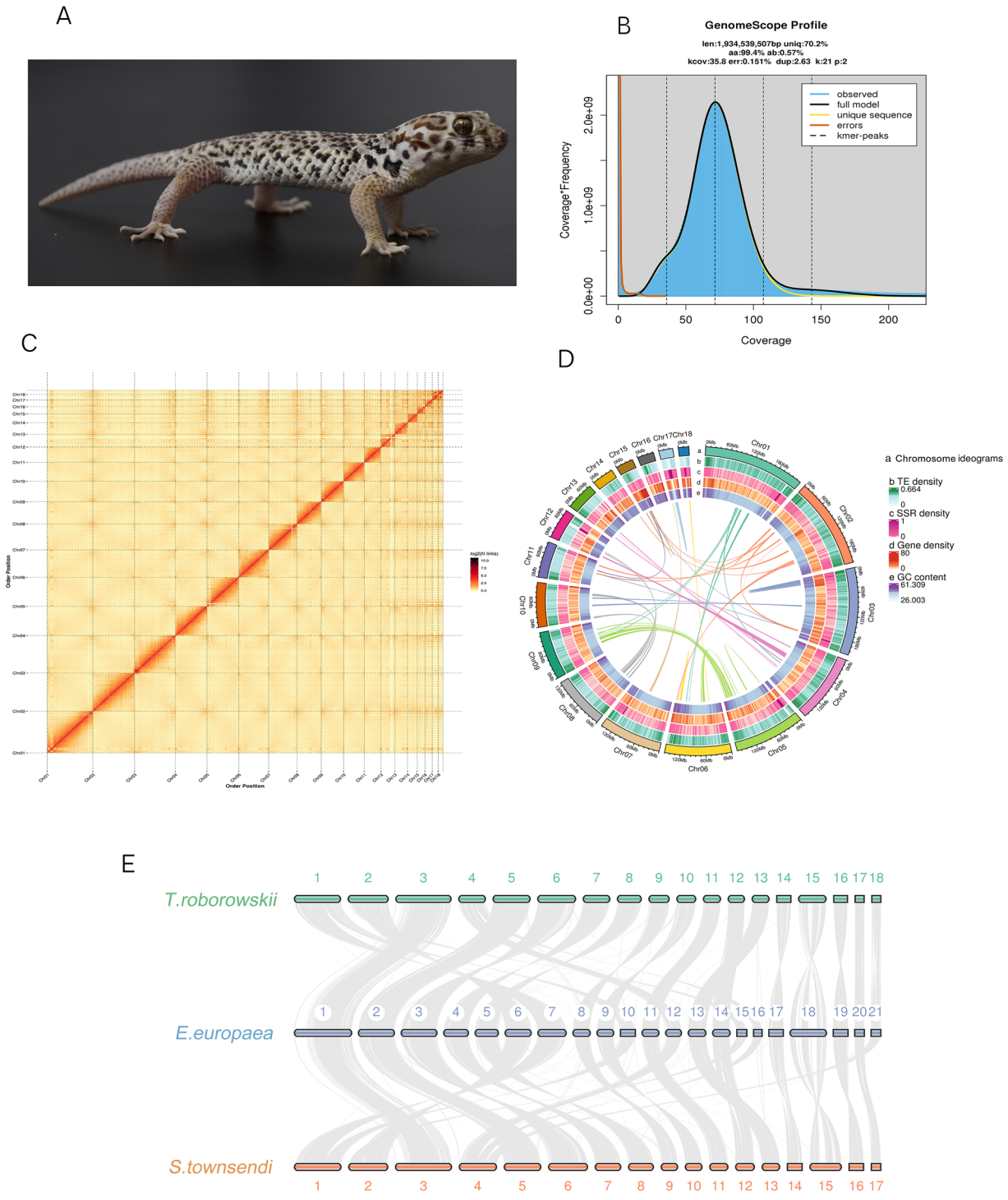

Fig. 1 Genome assembly and genome-wide mapping of *T. roborowskii*. (**A**) Photograph of *T. roborowskii*. (**B**) K-mer analysis of the *T. roborowskii* genome. The x-axis indicates the k-mer depth (coverage), and the y-axis represents the product of coverage and frequency. Based on diploid genome fitting, the first peak appears at a depth of 35.8, and the estimated haploid genome size is approximately 1.93 Gb. (**C**) Hi-C interaction heatmap of the *T. roborowskii* genome. The 18 blocks represent pseudochromosomes. Color intensity indicates contact frequency, with red indicating high interaction and yellow indicating low interaction. (**D**) Genome features of *T. roborowskii*. (a) Chromosome ideograms; (b) Transposable element (TE) density; (c) Simple sequence repeat (SSR) density; (d) Gene density; (e) GC content. (**E**) Synteny analysis of *T. roborowskii* with E. europaea and S. townsendi.

73%, and between *T. roborowskii* and *S. townsendi* was 70.19% (Fig. 1E). These values were calculated using MCScanX[17] by dividing the number of collinear gene pairs by the total number of orthologous gene pairs identified. The analysis was performed with the following parameters: MATCH_SCORE = 50, MATCH_SIZE = 5,

Characteristics	*T. roborowskii*
Genome Size (GB)	2.08
Number of contigs	183
Number of chromosomes	18
Scaffold N50 length (Mb)	156.21
contig N50 length (Mb)	117.38
BUSCO completeness (%)	98.27

Table 2. Genome assembly statistics of *T. roborowskii*.

Type	Number	Length	Rate(%)
/Maverick	1,627	311,946	0.02
ClassII/Merlin	392	26,725	0.00
ClassII/Mutator	957	69,757	0.00
ClassII/P	109	7,623	0.00
ClassII/PIF-Harbinger	1,470	95,987	0.00
ClassII/PiggyBac	123	6,169	0.00
ClassII/Sola	4	270	0.00
ClassII/Tc1-Mariner	8,149	2,218,749	0.11
ClassII/Unknown	308,578	67,206,140	3.23
ClassII/Unknown?	1	200	0.00
ClassII/Zator	1	40	0.00
ClassII/Zisupton	1,030	54,323	0.00
ClassII/hAT	50,441	13,001,320	0.63
Unknown	1,463	528,739	0.03
srpRNA	10	1,962	0.00
Total	3,562,004	798,100,421	38.39

Table 3. Summary of repetitive sequence annotation results.

Type	Number	Length	Rate (%)
Microsatellite (1–9 bp units)	752,434	14,010,780	0.67
Minisatellite (10–99 bp units)	220,968	33,471,342	1.61
Satellite (>=100 bp units)	18,658	39,557,317	1.90
Total	992,060	87,039,439	4.19

Table 4. Statistical information of the joint repeat sequences.

GAP_PENALTY = −1, OVERLAP_WINDOW = 5, E_VALUE = 1e-05, and MAX_GAPS = 5. These results suggest that *T. roborowskii* is more closely related to *E. europaea* at the genomic level.

Repeat sequence annotation and prediction. For the annotation of repetitive sequences, we first performed ab into predictions using RepeatModeler2 (v2.0.1)[18], which primarily incorporates two programs: RECON (v1.0.8)[19] and RepeatScout (v1.0.6)[20]. Classification of the prediction elements was conducted using RepeatClassifier, aided by the Dfam (v 3.5) database for annotation of known repeat families. To specifically annotate long terminal repeats (LTRs), we utilized LTR_retriever (2.9.0)[21], which integrates the results of LTRharvest (v1.5.10)[22] and LTR_FINDER (v1.07)[23] for comprehensive ab initio LTR prediction. The ab initio results were then merged with entries from known repeat database, and redundant sequences were removed to construct a species-specific repeat library. Subsequently, we employed RepeatMasker (v4.1.2)[24] to identify transposon sequences (TEs) across the genome using the constructed custom repeat library. This result in the detection of approximately 798,100,421 bp of TE sequences, accounting for 38.39% of the genome (Table 3).

Tandem repeats were predicted using the MIcroSAtellite identification tool (MISA v2.1)[25] and Tandem Repeat Finder (TRF, version 409, parameters: 2 7 7 80 10 50 500 -d -h)[26]. About a total of ~87,039,439 bp of tandem repeat sequences were identified, representing approximately 4.19% of the genome (Table 4).

Gene prediction. Gene prediction using a combination of three approaches: ab initio, homology-based, and transcriptome-based methods. For ab initio prediction, Augustus (v3.1.0)[27] and SNAP (2006-07-28)[28] were employed. Homology-based predictions were carried out using GeMoMa (v1.7)[29], leveraging gene models from closely related species. Transcriptome-based RNA-seq data from second-generation sequencing, which were assembled using two distinct pipelines. In the first pipeline, raw reads were aligned with Hisat (v2.1.0)[30] and

Method	Software	Species	Gene number
Ab initio	Augustus		43,273
Homology-based	SANP		115,995
	GeMoMa	*Euleptes europaea*	17,685
		Lactobacillus agilis	17,434
		Podarcis muralis	17,364
		Sphaerodactylus townsendi	17,794
		Zootoca vivipara	18,025
RNAseq	GeneMarkS-T		15,581
	PASA		14,365
Integration	EVM		21,341

Table 5. Gene Prediction Results Statistics.

Anno Database	Annotated Number	Annotated Ratio
GO Annotation	20,014	93.78
KEGG Annotation	19,581	91.75
KOG Annotation	15,939	74.69
Pfam Annotation	20,496	96.04
Swissprot Annotation	13,847	64.88
TrEMBL Annotation	21,248	99.56
eggnog Annotation	19,789	92.73
Nr Annotation	21,213	99.40
All Annotated	21,259	99.62

Table 6. Gene Functional Annotation Statistics.

assembled using Stringtie (v2.1.4)[31], followed by gene structure prediction with GeneMarkS-T (v5.1)[32]. In the second pipeline, transcriptome assembly was conducted using Trinity (v2.11)[33] and gene models were predicted with PASA (v2.4.1)[34]. The gene models predicted by all three methods were then integrated using EVidenceModeler (EVM v1.1.1)[35], and further refined using PASA (v2.4.1)[34]. As a result, a total of 21,341 protein-coding genes were predicted in *T. roborowskii* genome (Table 5).

Gene function annotation. The predicted gene sequences were functionally annotated by aligning them against multiple public databases, including NR[36], eggNOG[37], GO[38], KEGG[39], TrEMBL[40], KOG[41], SWISS-PROT[40] and Pfam[42]. In total, 99.62% of the genes could be annotated to the databases (Table 6), demonstrating a high level of functional characterization of the genome.

Data Records

All sequencing data generated in this study have been deposited in public repositories. Raw sequencing data are available in the NCBI Sequence Read Archive (SRA) under accession numbers SRR34204860[43] (Hi-C), SRR32572587[44] (Illumina short reads), and SRR34216249[45] (PacBio HiFi long reads). The final genome assembly is available at NCBI under accession number GCA_051473905.1[46]. Genome annotation data, including repetitive element annotation, gene structure predictions, and functional annotations, are deposited in Figshare[47].

Data Overview

Comparative genomics analysis. To provide a concise summary of the *T. roborowskii* genome, we compared its annotated gene content with representative species, including lizards, mammals, aquatic species, birds, and other geckos. Gene structures were processed using gffread (v0.99)[48] and scripts to extract representative transcripts, and gene family clustering was performed with OrthoFinder (v2.5.5)[49]. Single-copy orthologs were extracted and aligned using MAFFT (v7.409)[50], followed by codon alignment with PAL2NAL (v14)[51] and refinement with Gblocks (v0.19b)[52]. Phylogenetic reconstruction was carried out with IQ-TREE (v2.2.0)[53] and divergence times calibrated using the TimeTree database, with further estimation by PAML (v4.9i)[54], supporting both annotation completeness and phylogenetic consistency (Fig. 2A). To assess gene model accuracy, a branch-site model analysis using Codeml identified 696 candidate genes with elevated dN/dS ratios, many associated with neuronal processes and development of the olfactory bulb and lobe, highlighting functional diversity (Fig. 2B,C). Gene family expansion and contraction analysis using CAFE (v4.2.1)[55] revealed 74 expanded families comprising 1,607 genes, demonstrating the dataset's capacity to capture gene family dynamics (Fig. 2D).

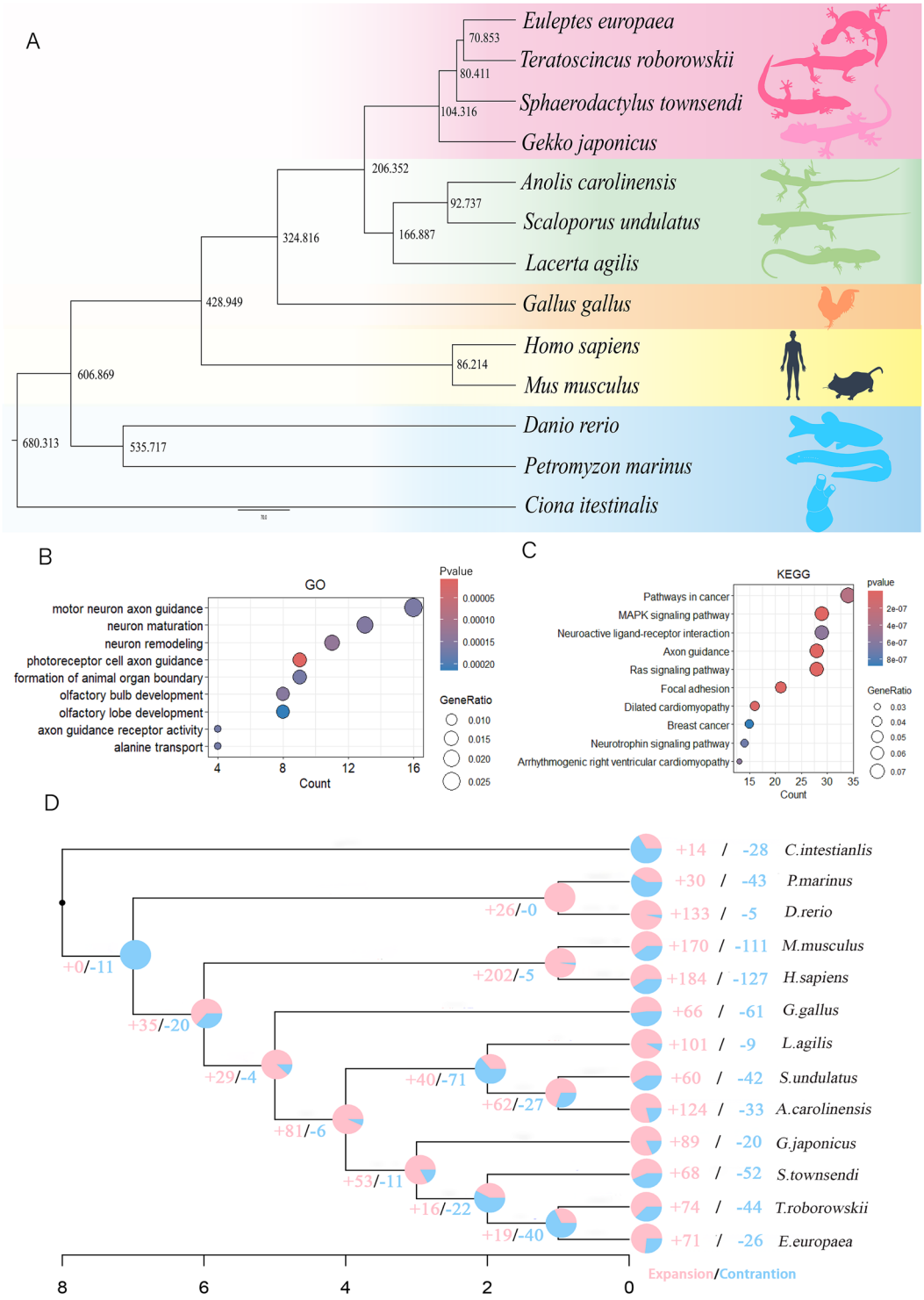

Fig. 2 Comparative genomic analysis of *T. roborowskii* and 11 reference species. (**A**) Phylogenetic tree and divergence times inferred from single-copy orthologs. (**B,C**) Annotation results for high dN/dS genes identified by PAML. (**D**) Expansion and contraction of gene families in *T. roborowskii*.

Technical Validation

Assessing genome size. The HiFi sequencing generated a data volume of 100.76 Gb, with a sequencing depth of approximately 52 X, and the average reads length of 16.5Kb. Using Hifiasm for genome assembly, the total length of the genomic contig sequence was 2.09 Gb, with Contig N50 of 117.38 Mb. After performing BLAST to remove contaminants, plasmids (plants-specific), mitochondria, and applying Hi-C to reduce redundancy. the final genomic contig sequence length was refined to 2.08 Gb, with the Contig N50 remaining at 117.38 Mb. Additionally, Hi-C sequencing provided about 180.38 Gb of data. The contig-level genome assembly

Evaluation Metric	Result	Description
CEGMA completeness	98.03%	Core eukaryotic gene completeness
BUSCO completeness	95.23%	Based on vertebrata dataset from OrthoDB (v10)[60]
Short-read mapping rate	99.11%	Second-generation (Illumina) reads mapped to assembly
Short-read genome coverage	99.98%	Genome coverage by second-generation reads
Short-read average depth	78×	Mean sequencing depth for second-generation reads
Long-read mapping rate	99.93%	Third-generation (PacBio) reads mapping
Long-read genome coverage	99.99%	Genome coverage by third-generation reads
Long-read average depth	49×	Mean sequencing depth for third-generation reads

Table 7. Summary of genome assembly quality evaluation.

was successfully anchored to chromosomes, with a total of 2,077.46 Mb of sequence anchored to 18 chromosomes, representing 99.92% of the genome. Of the sequences localized to chromosomes, 2,070.96 Mb (99.69%) had their order and orientation determined. These results demonstrate excellent continuity and integrity, making this assembly highly suitable for downstream genomic research.

Quality assessment of genome assembly. The quality of the genome assembly was evaluated using the BUSCO (v5.2.2)[56] software, which assesses the completeness of gene prediction against a vertebrate database containing 3,354 conserved core genes. Our analysis showed that 98.27% of the BUSCO genes were present in the predicted genes. Specifically, 97.08% of the single-copy BUSCO genes were complete, and 1.19% were multi-copy genes. The analysis revealed that 0.03% of BUSCO genes were incomplete, with no missing or out-of-order BUSCOs. Furthermore, CEGMA[15] analysis of the genome assembly revealed a completeness score of 98.03%, confirming the high quality of the gene predictions and the robustness of the assembly (Table 7).

Data availability

The genomic data of *T. roborowskii* have been deposited in the NCBI database under BioProject accession number PRJNA1229158[57] and BioSample accession number SAMN47122675[58]. The chromosome-level genome assembly of *T. roborowskii* has been submitted to GenBank under the accession number JBPQTH000000000[59]. All datasets are publicly accessible. Raw reads have been deposited in the NCBI SRA (Hi-C: SRR34204860[43]; Illumina: SRR32572587[44]; PacBio HiFi: SRR34216249[45]). The final genome assembly is available under accession GCA_051473905.1[46]. Genome annotations, including repeats, gene models, and functional data, are deposited in Figshare[47].

Code availability

All data processing was carried out using bioinformatics software tools, selected in accordance with the methods detailed in the Methods section. Each software tool was executed following the instructions in the respective manuals. In instances where specific parameter settings were not explicitly provided in the documentation, default settings of the software were applied. No custom code was used in the curation or validation of the dataset described in this study.

Received: 10 April 2025; Accepted: 10 September 2025;
Published online: 23 October 2025

References

1. Shi, L., Zhou, Y. & Yuan, H. Reptile fauna and geographic zoning of Xinjiang Uygur Autonomous Region. *Sichuan Journal of Zoology* **21**, 152–157 (In Chinese) (2002).
2. Du, Y., Zhou, R. & Zhang, K. Characterization and source analysis of urban atmospheric particulate matter in an arid continental climate zone—Tulufan City as an example. *Environment and Development* **36**, 38–46+65 (In Chinese) (2024).
3. Liao, B. Y. *et al.* Degeneration of the olfactory system in a murid rodent that evolved diurnalism. *Mol. Biol. Evol.* **41**, 37 (2024).
4. Jacyniak, K., McDonald, R. P. & Vickaryous, M. K. Tail regeneration and other phenomena of wound healing and tissue restoration in lizards. *J. Exp. Biol.* **220**, 2858–2869 (2017).
5. Huang, R., Zhang, J., Lu, L., Huang, S. & Li, C. High-quality genome assembly and annotation of the crested gecko (*Correlophus ciliatus*). *G3 (Bethesda)* **15**, jkae265 (2025).
6. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
7. Pinto, B. J. *et al.* Chromosome-level genome assembly reveals dynamic sex chromosomes in Neotropical leaf-litter geckos (*Sphaerodactylidae: Sphaerodactylus*). *J. Hered.* **113**, 272–287 (2022).
8. Nurhidayat, L. *et al.* Tokay gecko tail regeneration involves temporally collinear expression of HOXC genes and early expression of satellite cell markers. *BMC Biol.* **23**, 6 (2025).
9. Ghatak, S., Muthukumaran, R. B. & Nachimuthu, S. K. A simple method of genomic DNA extraction from human samples for PCR-RFLP analysis. *J. Biomol. Tech.* **24**, 224–231 (2013).
10. Li, Y. & Xu, X.X. Research progress of high-throughput sequencing technology. *Chinese Medical Engineering* **27**, 32–37 (In Chinese) (2019).
11. Mulqueen, R. M. *et al.* High-content single-cell combinatorial indexing. *Nat. Biotechnol.* **39**, 1574–1580 (2021).
12. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
13. Wang, Y. *et al.* Sequencing and assembly of polyploid genomes. *Methods Mol. Biol.* **2545**, 429–458 (2023).
14. Yekefenhazi, D. *et al.* Chromosome-level genome assembly of *Nibea coibor* using PacBio HiFi reads and Hi-C technologies. *Sci. Data* **9**, 670 (2022).

15. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
16. Gong, D., Zhang, Y. & Li, X. Molecular and chromosomal evolution of scorpions. *Chin J Zool* **33**, 62–64 (2018).
17. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**, e49 (2012).
18. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457 (2020).
19. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12**, 1269–1276 (2002).
20. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics 21 Suppl* **1**, i351–i358 (2005).
21. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol* **176**, 1410–1422 (2018).
22. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
23. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265–W268 (2007).
24. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics Chapter* **4**, Unit 4.10 (2004).
25. Beier, S. *et al*. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585 (2017).
26. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
27. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* **32**, W309–W312 (2004).
28. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
29. Keilwagen, J. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res* **44**, e89 (2016).
30. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360 (2015).
31. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
32. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research* **43**, e78 (2015).
33. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644–652 (2011).
34. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666 (2003).
35. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7 (2008).
36. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **33**, D501–D504 (2005).
37. 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*. **8**, 47(D1), D309-D314 (2019).
38. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).
39. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457–D462 (2016).
40. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* **31**, 365–370 (2003).
41. Tatusov, R. L. *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
42. Finn, R. D. *et al*. Pfam: clans, web tools and services. *Nucleic Acids Research* **34**, D247–D251 (2006).
43. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34204860 (2025).
44. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32572587 (2025).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34216249 (2025).
46. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_051473905.1 (2025).
47. Ping, Y. Genome annotation. *Figshare* https://doi.org/10.6084/m9.figshare.28748555 (2025).
48. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Research* **9**, 304 (2020).
49. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).
50. Katoh, K., Asimenos, G. & Toh, H. Multiple alignment of DNA sequences with MAFFT. *Methods in Molecular Biology* **537**, 39–64 (2009).
51. Zhang, C., Wang, J., Long, M. & Fan, C. gKaKs: the pipeline for genome-level Ka/Ks calculation. *Bioinformatics* **29**, 645–646 (2013).
52. Kück, P. *et al*. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Frontiers in Zoology* **7**, 10 (2010).
53. Trifinopoulos, J., Nguyen, L. T., von, Haeseler, A. & Minh, B. Q. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research* **44**, W232–W235 (2016).
54. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586–1591 (2007).
55. Lu, Y. Y. *et al*. CAFE: aCcelerated Alignment-FrEe sequence analysis. *Nucleic Acids Research* **45**, W554–W559 (2017).
56. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227–245 (2019).
57. *NCBI BioProject* https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1229158 (2025).
58. *NCBI BioSample* https://www.ncbi.nlm.nih.gov/biosample/SAMN47122675 (2025).
59. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBPQTH000000000 (2025).
60. Kriventseva, E. V. *et al*. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research* **47**, D807–D811 (2019).

Acknowledgements

Thanks to Prof. Lei Shi for his support. This work was fnancially supported by Xinjiang Uygur Autonomous Region Tianchi Talent Introduction Program ("Tianchi Outstanding Young Doctoral Talent Award"), the Xinjiang Uygur Autonomous Region Universities' Basic Research Operating Expenses Project (No. XJEDU2022P050) and Xinjiang Key Laboratory for Ecological Adaptation and Evolution of Extreme Environment Biology, College of Life Sciences, Xinjiang Agricultural University (No. KFKT2402).

Author contributions

P.Y. was responsible for manuscript writing, experimental manipulation and data analysis, and drew all the graphs, C.W. was involved in the experimental process, H.X.C. and X.Y.L. were involved in data organization, and J.W. was responsible for manuscript revision and experimental design. All authors discussed the results and implications and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.W.

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.