scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-scale Genome Assembly of the Critically Endangered Blue-crowned Laughingthrush (*Pterorhinus courtoisi*, Leiothrichidae)

Yuxuan Ouyang[1,2] ✉, Lin Yang[2], Binbin Cheng[3], Chang Xiao[1] & Weiwei Zhang[1,2] ✉

The Blue-crowned Laughingthrush (*Pterorhinus courtoisi*) is a critically endangered species and listed as National First-class Protected Wildlife in China, with a small population size and highly restricted geographic distribution in Jiangxi Province. However, the genetic mechanisms underlying its endangered status remain unclear. In this study, we constructed a chromosome-level reference genome by integrating Illumina short-read, PacBio long-read, and Hi-C chromatin interaction data. The final assembled genome spans 1.255 Gb, with 1.158 Gb (92.32%) of the sequences anchored onto 39 pseudochromosomes. A total of 16,807 protein-coding genes were predicted, among which 15,574 genes (92.7%) were functionally annotated. This high-quality genome assembly provides a valuable genomic resource for future genetic studies and conservation efforts for the Blue-crowned Laughingthrush.

Background & Summary

The Blue-crowned Laughingthrush (*Pterorhinus courtoisi*) belongs to the family Leiothrichidae, order *Passeriformes*[1]. With an extremely limited distribution and a critically small population size, it was designated as National First-class Protected Wildlife in China in 2021[2], and was listed as Critically Endangered (CR) on the IUCN Red List of Threatened Species[3]. The nominate subspecies, *P. c. courtoisi*, is currently confined to the Wuyuan and Dexing regions in Jiangxi Province. The Simao subspecies, *P. c. simaoensis*, which once inhabited southern Yunnan, was officially declared extinct in the wild in 2017[4]. The genetic basis of its endangered status remains poorly understood.

In this study, we assembled a chromosome-level genome of *Pterorhinus courtoisi* using PacBio long-read sequencing and Hi-C chromatin conformation capture technology. A total of 101 Gb of raw sequencing data were generated, representing an estimated genome coverage of 78.85× (Table 1). The initial genome assembly had a scaffold N50 of 41.65 Mb, with a total length of 1,264,719,544 bp (Table 2). After Hi-C-assisted scaffolding, 1,158,634,107 bp of sequences were successfully anchored onto chromosomes, resulting in a final genome size of 1,255,045,218 bp and a chromosomal anchoring rate of 92.32% (Fig. 1 and Table 3).

Genome completeness was assessed using both BUSCO and CEGMA. BUSCO analysis showed that 96.6% of the expected single-copy orthologs were present in the genome assembly, including 10,388 (95.8%) complete single-copy BUSCOs, 86 (0.8%) duplicated BUSCOs, 65 (0.6%) fragmented, and 303 (2.8%) missing BUSCOs (Table 4). CEGMA analysis revealed that 228 out of 248 core eukaryotic genes (CEGs) were successfully assembled, with a completeness of 91.94% (Table 5).

Repeat annotation identified 322,662,110 bp of repetitive elements, with the majority classified as long terminal repeat (LTR) retrotransposons (233,056,295 bp, accounting for 18.57% of the genome) (Tables 7, 8). A

[1]National Conservation and Research Center for the Blue-crowned Laughingthrush, NanChang, China. [2]Jiangxi Provincial Key Laboratory of Conservation Biology, Nanchang, China. [3]Shaanxi Yellow River Wetland Provincial Nature Reserve Management Office, Weinan, Shaanxi, 714000, China. ✉e-mail: afaer.ouyang@qq.com; zhangweiwei_nefu@163.com

libraries	Insert size	Total data (G)	Read length (bp)	Sequence coverage (X)
Illumina reads	350 bp	57	150	44.5
PacBio reads	—	44	—	34.35
Total	—	101	—	78.85

Table 1. Summary of sequencing data for the *Pterorhinus courtoisi* genome.

Sample ID	Contig length	Scaffold length	Contig number	Scaffold number
Total	1,255,030,518	1,255,045,218	803	656
Max	105,723,693	120,087,461	—	—
Number >= 2000	—	—	803	656
N50	27,850,013	70,596,790	13	7
N60	19,321,900	37,358,479	18	10
N70	9,969,948	22,125,962	28	14
N80	5,492,398	15,355,625	45	21

Table 2. Summary statistics of the preliminary genome assembly of *Pterorhinus courtoisi*.

total of 16,807 protein-coding genes were predicted using a combination of de novo, homology-based, and transcriptome-assisted approaches, of which 15,574 (92.7%) were successfully functionally annotated (Table 11). Additionally, 5,360 non-coding RNAs (ncRNAs), including tRNAs predicted using tRNAscan-SE, were identified (Table 12).

We successfully assembled the first chromosome-level reference genome of the Blue-crowned Laughingthrush, which provides a fundamental resource for elucidating the genetic mechanisms underlying its critically endangered status and contributes to the conservation of its genetic diversity. Moreover, it lays a solid foundation for future studies on the biological and evolutionary history of this critically endangered species.

Methods

Sample collection and sequencing. A freshly deceased individual from Wuyuan, Jiangxi province (unsuccessfully resuscitated after injury) was obtained, and samples including blood and tissues were collected, including the tongue, pectoral muscle, liver, heart, eyes, and lungs. These samples were initially preserved at −20 °C, then transferred to a −80 °C ultra-low-temperature freezer within 24 hours for long-term storage and subsequent genomic DNA extraction and sequencing.

Genomic DNA from blood samples was extracted using the Ezup Column Blood Genomic DNA Extraction Kit (Sangon Biotech, Shanghai), strictly following the manufacturer's protocol. DNA from tissue samples was extracted using the classical phenol–chloroform method. All downstream procedures, including library preparation, sequencing, genome assembly, and annotation were conducted by Novogene Co., Ltd. (Beijing, China). Upon sample arrival, DNA purity and integrity were assessed using Nanodrop, Qubit, and agarose gel electrophoresis. Only samples with OD260/280 values between 1.8 and 2.0, OD260/230 $\geq$ 1.5, and clear major bands were retained for library construction; those failing quality control were excluded.

HiFi sequencing was performed on the PacBio Sequel IIe platform. Raw subreads were filtered according to the official criterion (minimum read length = 50 bp), and high-quality HiFi reads (average quality > Q20) were generated using the ccs software (https://github.com/PacificBiosciences/ccs)[5]. For Hi-C sequencing, cells were crosslinked using formaldehyde to stabilize DNA–protein interactions. The chromatin was then subjected to restriction enzyme digestion, end repair, biotin labeling, proximity ligation, proteinase digestion, and crosslink reversal. The resulting DNA was randomly fragmented to ~350 bp, and libraries were constructed for paired-end sequencing on the Illumina HiSeq platform.

A total of 57 Gb of short-read (second-generation) sequencing data were produced, corresponding to a coverage of 44.5×, along with 44 Gb of long-read (third-generation) data, representing a coverage of 34.35× (Table 1).

Genome assembly. Prior to formal genome assembly, we estimated the genome size of *Pterorhinus courtoisi* using 17-mer frequency analysis with Jellyfish[6] (v2.2.7), setting k = 17. The initial genome size was estimated to be approximately 1,307.13 Mb, and after further correction, the final estimated size was determined to be 1,281.7 Mb. The heterozygosity rate was assessed at 0.68%, and repetitive elements were estimated to account for 44.56% of the genome.

An initial genome assembly was performed using SOAPdenovo2[7] with k = 41. This yielded a contig N50 of 2,364 bp and a total contig length of 1,243,813,211 bp. After scaffold construction, the scaffold N50 reached 4,165 bp, with a total scaffold length of 1,264,719,544 bp.

Subsequently, a high-quality genome assembly was generated using PacBio HiFi long reads and the Hifiasm assembler[8]. The resulting sequences exhibited normal base composition (A, T, G, C) with a GC content of 43.61% and no undetermined bases (N content = 0.00%), meeting the standard quality control threshold (<10%). The final assembly achieved a contig N50 of 27.85 Mb and a scaffold N50 of 70.59 Mb as well (Table 2).

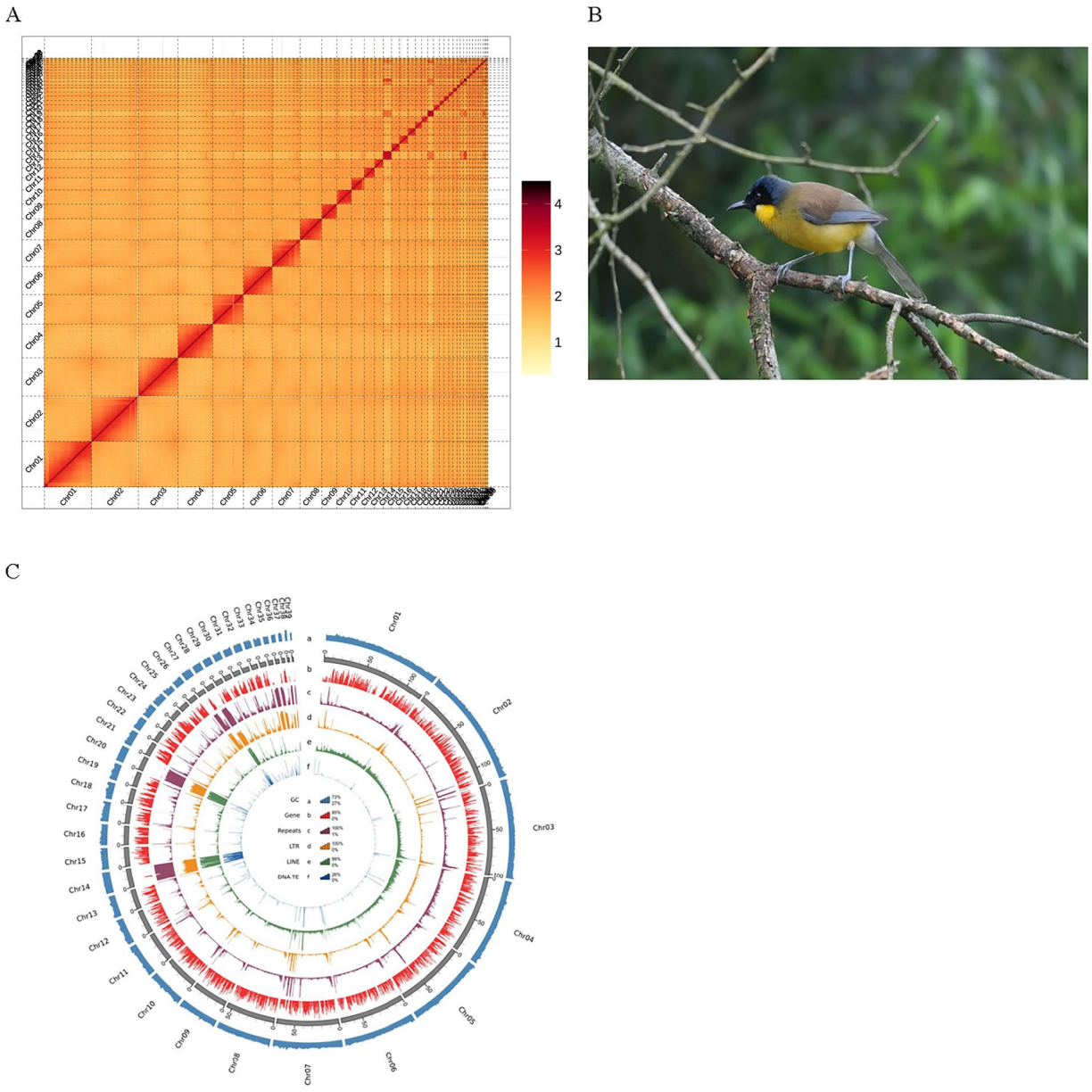

Fig. 1 Genomic features of *Pterorhinus courtoisi*. (**A**) Hi-C contact map showing chromosome-level scaffolding of the *Pterorhinus courtoisi* genome assembly. (**B**) The adult *Pterorhinus courtoisi*. (**C**) Circos plot illustrating genome-wide features: (**a**) GC content, (**b**) gene density, (**c**) repeat sequence density, (**d**) LTR element density, (**e**) LINE element density, (**f**) DNA transposable element (DNA-TE) density.

Chromosome-level assembly was then performed using Hi-C sequencing data. The AllHiC pipeline[9] was employed to anchor the contigs/scaffolds to a chromosomal scale, followed by manual refinement based on chromatin interaction signals using Juicebox[10,11]. This Hi-C-assisted assembly process[12] successfully anchored the genome into 39 pseudochromosomes, while 617 scaffolds remained unanchored. The total length of anchored sequences was 1,158,634,107 bp, accounting for 92.32% of the assembled genome (1,255,045,218 bp) (Fig. 1 and Table 3).

Genome assembly evaluation. The completeness of the *P. courtoisi* genome assembly was independently assessed using two standard strategies: BUSCO (Benchmarking Universal Single-Copy Orthologs) and CEGMA (Core Eukaryotic Genes Mapping Approach). BUSCO analysis against the eukaryota_odb10 database identified 96.6% complete single-copy orthologs[13] (Table 4). Meanwhile, CEGMA successfully detected 228 out of 248 core eukaryotic genes (CEGs), resulting in a completeness score of 91.94%[14] (Table 5). Both evaluations indicate that the assembled genome exhibits a high level of completeness.

To validate these results at the read level, we aligned reads from short-insert libraries to the assembled genome using BWA (http://bio-bwa.sourceforge.net/)[15,16]. Alignment metrics, including mapping rate, genome

Chromosome	Size(bp)	Chromosome	Size(bp)	Chromosome	Size(bp)
Chr01	120,087,461	Chr14	22,125,962	Chr27	9,478,776
Chr02	118,984,491	Chr15	21,613,396	Chr28	8,997,670
Chr03	101,455,316	Chr16	21,224,285	Chr29	8,435,579
Chr04	89,063,632	Chr17	20,091,129	Chr30	8,299,882
Chr05	77,504,021	Chr18	16,944,484	Chr31	8,242,621
Chr06	74,380,839	Chr19	16,810,415	Chr32	8,075,677
Chr07	70,596,790	Chr20	15,834,843	Chr33	7,136,133
Chr08	55,970,000	Chr21	15,355,625	Chr34	6,963,600
Chr09	39,853,124	Chr22	12,721,623	Chr35	5,313,226
Chr10	37,358,479	Chr23	12,696,370	Chr36	4,340,288
Chr11	33,702,399	Chr24	12,125,966	Chr37	3,547,891
Chr12	26,864,808	Chr25	10,466,502	Chr38	2,516,773
Chr13	22,349,836	Chr26	9,487,818	Chr39	1,616,377

Table 3. Length statistics of individual chromosomes in the *Pterorhinus courtoisi* genome assembly.

BUSCO notation assessment results	
Complete BUSCOs	96.6%
Complete and single-copy BUSCOs	95.8%
Complete Duplicated BUSCOs	0.8%
Fragmented BUSCOs	0.6%
Missing BUSCOs	2.8%
Total BUSCO groups searched	10844

Table 4. BUSCO assessment results of the *Pterorhinus courtoisi* genome assembly.

Complete		Complete + Partial	
# Prots	%Completeness	# Prots	%Completeness
224	90.32	228	91.94

Table 5. CEGMA assessment results of the *Pterorhinus courtoisi* genome assembly.

		% of percentage
Reads	Mapping rate (%)	99.15
Genome	Average sequencing depth	37.45
	Coverage (%)	99.48
	Coverage at least 4X (%)	99.14
	Coverage at least 10X (%)	98.12
	Coverage at least 20X (%)	93.27

Table 6. Summary of read mapping and genome coverage for the *Pterorhinus courtoisi* genome.

coverage, and depth distribution, were computed to evaluate assembly completeness and sequencing uniformity. The mapping results showed that 99.15% of reads could be uniquely or confidently mapped to the reference genome, with a total genome coverage of 99.48% (Table 6), indicating strong concordance between raw reads and the assembled genome.

To further assess GC bias and potential contamination, the assembled genome was divided into non-overlapping 10 kb windows to compute the two-dimensional distribution of GC content and sequencing depth. The results showed that GC content was predominantly centered around 43.61%, and no anomalous clustering or outlier subsets were observed in the scatter plot, suggesting that the sample was free from bacterial, viral, or other exogenous sequence contamination (Figs. 2, 3).

Finally, the base-level accuracy of the genome assembly was evaluated using the reference-free tool Merqury[17], based on k-mer spectra. The QV value estimated using the merqury-mash module was 46.3827, corresponding to a base error rate of approximately 2.3×10^{-5}, or a base accuracy of 99.9977%. This exceeds the "platinum-grade" genome assembly standard ($QV \geq 40$), indicating that the *P.courtoisi* genome assembled in this study is highly accurate and suitable for downstream applications in comparative genomics, population genetics, and functional annotation.In addition, k-mer–based completeness was estimated to be 85.73%, and the k-mer spectra plots further supported the high quality of the assembly(Fig. 4).

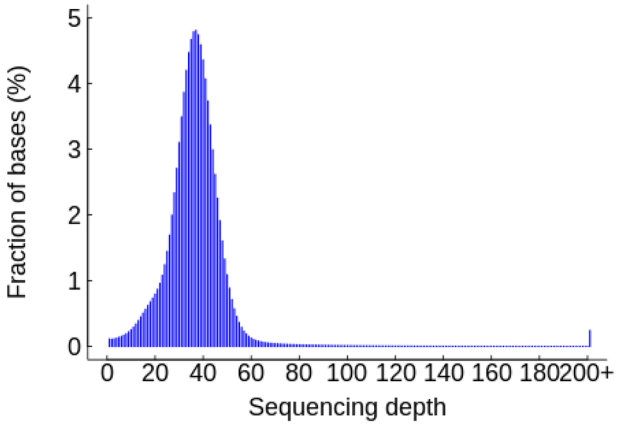

Fig. 2 Distribution of sequencing depth across the *Pterorhinus courtoisi* genome.

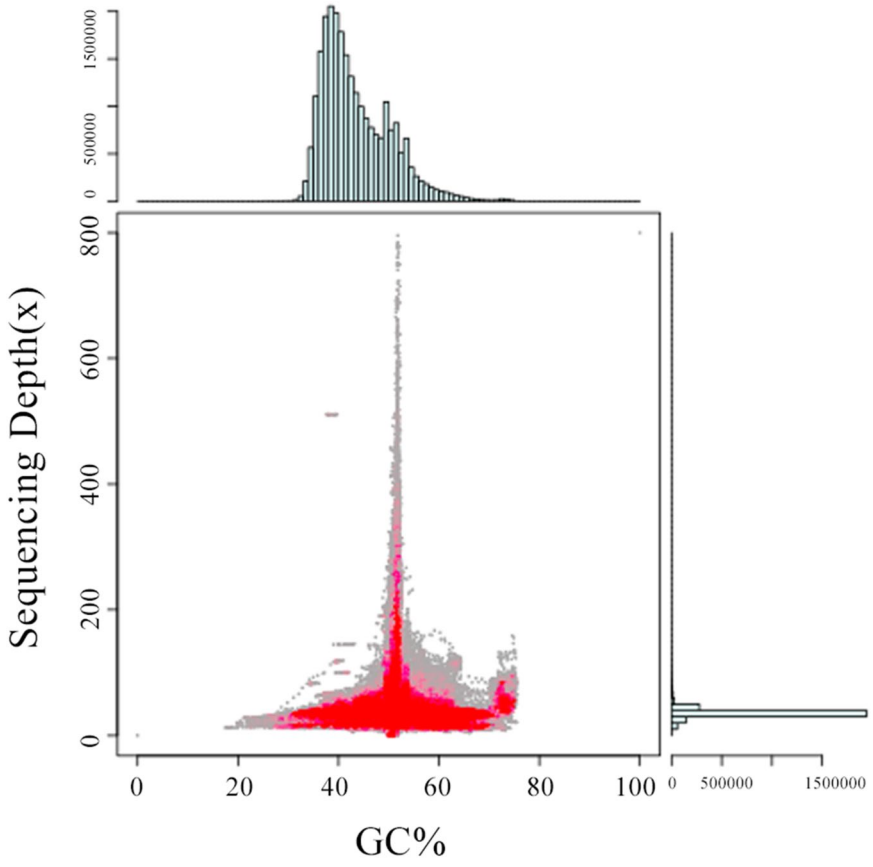

Fig. 3 Distribution of GC content and sequencing depth across the *Pterorhinus courtoisi* genome.

Repetitive element annotation. To systematically identify repetitive sequences across the *P. courtoisi* genome, we employed a combined strategy integrating both homology-based searches and *de novo* predictions.

For tandem repeats, we used Tandem Repeats Finder (TRF)[18] (http://tandem.bu.edu/trf/trf.html) under default parameters to scan and extract tandem repeat sequences through ab initio prediction.

For interspersed repetitive elements, we conducted both homology-based annotation and *de novo* identification:

The homology-based prediction was performed using the Repbase database[19] (http://www.girinst.org/repbase). Repeats were annotated at both DNA and protein levels using RepeatMasker[20,21] (http://www.repeatmasker.org/) and its embedded script RepeatProteinMask, both under default settings.

For *de novo* identification, we first employed LTR_FINDER[22] (http://tlife.fudan.edu.cn/ltr_finder/) with default settings to identify structural features of long terminal repeat (LTR) retrotransposons. Then, RepeatScout[23] (http://www.repeatmasker.org/) was used to construct a library of novel repeat families, followed

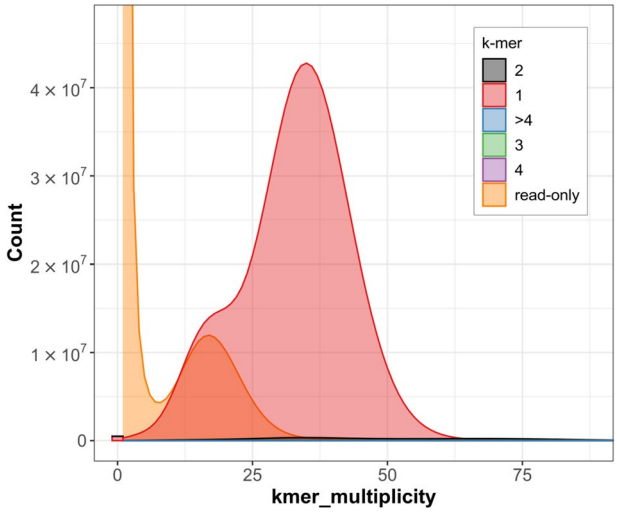

Fig. 4 Merqury k-mer spectra of the *Pterorhinus courtoisi* genome assembly.

Type	Repeat Size(bp)	% of genome
Trf	132,617,866	10.57
Repeatmasker	295,992,812	23.58
Proteinmask	45,629,362	3.64
Total	322,662,110	25.71

Table 7. Summary of repetitive elements in the *Pterorhinus courtoisi* genome.

	Denovo + Repbase		TE Proteins		Combined TEs	
	Length(bp)	% in Genome	Length(bp)	% in Genome	Length(bp)	% in Genome
DNA	5,480,820	0.44	55,987	0	5,527,273	0.44
LINE	56,575,014	4.51	27,125,820	2.16	62,349,000	4.97
SINE	202,849	0.02	0	0	202,849	0.02
LTR	232,319,330	18.51	18,450,797	1.47	233,056,295	18.57
Unknown	20,195,183	1.61	0	0	20,195,183	1.61
Total	295,992,812	23.58	45,629,362	3.64	298,320,839	23.77

Table 8. Classification of repetitive elements in the *Pterorhinus courtoisi* genome.

Gene set		Number	Average transcript length(bp)	Average CDS length(bp)	Average exons per gene	Average exon length(bp)
De novo	Augustus	17,016	18,420.43	1,452.99	8.33	174.43
	GlimmerHMM	90,200	7,427.72	572.95	3.38	169.67
	SNAP	57,378	26,363.68	683.97	5.4	126.65
	Geneid	25,447	29,651.90	1,352.38	7.17	188.6
	Genscan	33,345	24,943.65	1,397.38	8.26	169.22
Homolog	Tgut	14,518	22,847.68	1,741.54	9.99	174.3
	Gcou	13,302	17,278.10	1,499.77	8.82	170.09
	Gsan	13,441	18,234.42	1,595.38	9.24	172.58
RNAseq	PASA	25,687	15,432.85	1,088.07	6.45	168.65
	Transcripts	73,679	17,766.22	1,898.31	5.89	322.24
EVM		18,573	21,262.69	1,519.15	9.03	168.27
Pasa-update		18,414	22,874.00	1,547.55	9.17	168.78
Final set		16,807	24,482.22	1,633.43	9.78	167.04

Table 9. Basic statistics of gene structure prediction for the *Pterorhinus courtoisi* genome.

Fig. 5 Evidence support statistics for the *Pterorhinus courtoisi* gene set.

Species	Number	Average transcript length(bp)	Average CDS length(bp)	Average exons per gene	Average exon length(bp)	Average intron length(bp)
*P. courtoisi** (this study)	16,807	24,482.22	1,633.43	9.78	167.04	2,602.80
Garrulax sannio	12,897	19,227.05	1,657.88	9.93	166.9	1,966.72
Taeniopygia guttata	16,348	27,680.62	1,819.67	10.73	169.65	2,659.01
P. courtoisi	13,534	17,512.99	1,540.14	9.31	165.47	1,922.65

Table 10. Basic statistics of gene structures in closely related species of *Pterorhinus courtoisi*.

by refinement and integration using RepeatModeler[24] (http://www.repeatmasker.org/RepeatModeler.html), also with default parameters. Candidate transposable element (TE) sequences were filtered to retain only those with a length $\geq$100 bp and an N-content $<$5%.

The final repeat annotation was obtained by merging the Repbase-based results with the de novo TE library, and removing redundancy using uclust[25] (100% identity and 100% coverage). This custom non-redundant library was then re-submitted to RepeatMasker to complete the DNA-level repeat identification.

Using this pipeline, we identified that 25.71% of the *P. courtoisi* genome consists of repetitive sequences (Tables 7, 8).

Gene annotation. Structural annotation of the genome was performed by integrating three strategies: *de novo* prediction, homology-based annotation, and transcriptome-assisted evidence, aiming to construct gene models with both completeness and accuracy.

For *de novo* gene prediction, five algorithms-Augustus[26] (v3.2.3), Geneid[27] (v1.4), Genescan[28] (v1.0), GlimmerHMM[29] (v3.04), and SNAP[30] (2013-11-29)—were executed in parallel with default parameters to generate candidate gene structures.

To incorporate transcriptional evidence, Trinity[31] (v2.1.1) was used to assemble RNA-seq reads de novo. Subsequently, RNA-seq reads from multiple tissues were aligned to the genome using Hisat[32] (v2.0.4) and TopHat[33] (v2.0.11) with default settings to accurately capture exon boundaries and splicing sites. The resulting alignments were used by StringTie[34] (v1.3.3) and Cufflinks[35] (v2.2.1) to reconstruct transcript models, providing transcript-level support for gene structure annotation.

For homology-based prediction, homologous protein sequences were downloaded from public databases such as Ensembl and NCBI. A two-step alignment strategy was applied: protein sequences were first mapped to the genome using TBLASTN[36] (E-value $\leq$ 1e-5), and candidate regions were then refined using GeneWise[37] (v2.4.1) to accurately determine exon-intron boundaries and recover full-length gene models. To improve annotation precision, protein sequences from *P. courtoisi*[38], *Garrulax sannio*[38] (NGDC BioProject accession: PRJCA005228), and *Taeniopygia guttata*(Ensembl assembly: taeGut1, Release 86)[39] were aligned to the target genome using BLAST and GeneWise to identify homologous gene loci. In addition, EST/cDNA sequences from closely related species were aligned to the genome using BLAT[40], further enriching gene structure information.

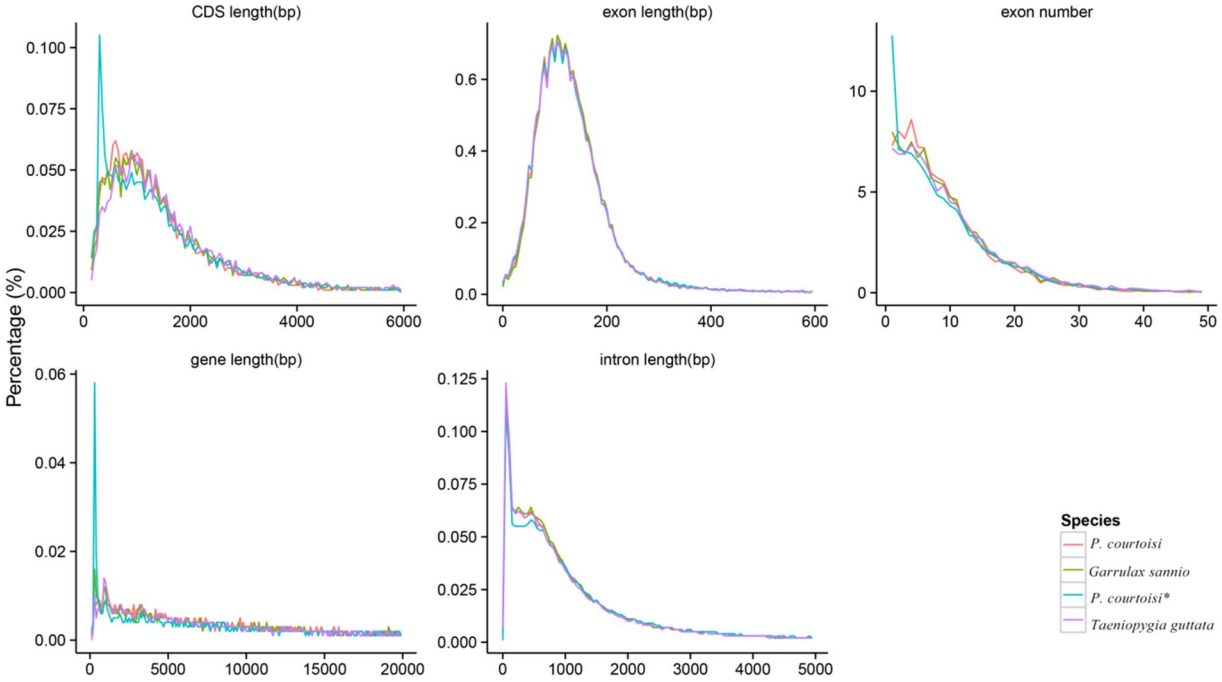

Fig. 6 Comparative analysis of genomic elements among closely related species of *Pterorhinus courtoisi*.

Functional annotation	Number	Percent(%)
Swissprot	14,610	86.9
Nr	15,068	89.7
KEGG	13,688	81.4
InterPro	15,399	91.6
GO	11,290	67.2
Pfam	13,503	80.3
Annotated	15,574	92.7
Unannotated	1,233	7.3
Total	16,807	—

Table 11. Functional annotation of genes in the *Pterorhinus courtoisi* genome.

All prediction results from de novo, homology, and transcriptome sources were integrated using EVidenceModeler[41] (EVM) to produce a non-redundant and high-confidence gene set. The resulting gene models were further refined using PASA[42], which incorporated assembled transcripts from Trinity to add 5′/3′ UTR annotations and alternative splicing information, thus finalizing the annotated gene set.

As a result, 16,807 protein-coding genes were annotated in the *P. courtoisi* genome. The basic statistics of these gene models are presented in Table 9, and Fig. 5 illustrates the proportion of gene models supported by different evidence types. Comparative statistics of gene structures among *P. courtoisi*, *G. sannio*, and *T. guttata* are shown in Table 10 and Fig. 6, respectively.

For functional annotation, protein sequences were first aligned to the Swiss-Prot[43] database using BLASTP[44] (E-value $\leq$ 1e-5), with the best-matching entry used as the primary functional assignment. Conserved domains and functional motifs were identified using InterProScan[45] (v5.31), which searched across multiple databases including ProDom, PRINTS, Pfam, SMART, PANTHER, and PROSITE. Gene Ontology (GO) terms were assigned based on InterPro entries. To further expand annotation coverage, DIAMOND[46] (v0.8.22) and BLAST searches (E-value $<$ 1e-5) were conducted against the NR database, and KEGG[47] pathway mapping was used to associate genes with biological pathways.

In total, 92.7% of the predicted genes were functionally annotated with high confidence. Detailed functional annotation statistics are presented in Table 11 and Fig. 7.

Non-coding RNA (ncRNA) annotation. The annotation of non-coding RNAs (ncRNAs) was conducted through separate workflows targeting tRNAs, rRNAs, and other small RNAs.

Firstly, genome-wide tRNA scanning was performed using tRNAscan-SE[48] (http://lowelab.ucsc.edu/tRNAscan-SE/) with default parameters. The software outputs included predicted secondary structures, anticodon types, and confidence scores.

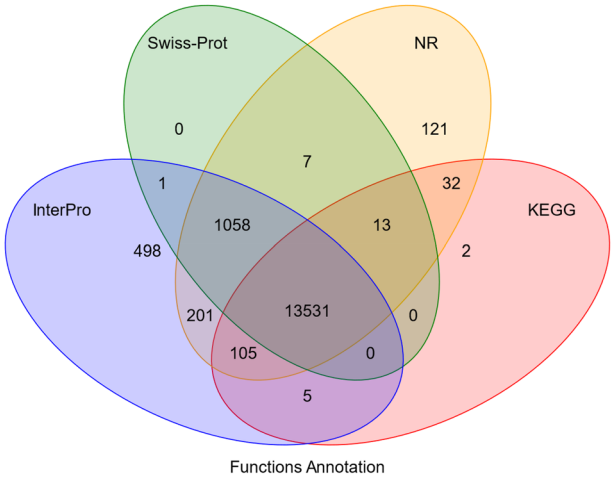

Fig. 7 Summary statistics of gene functional annotations in the *Pterorhinus courtoisi* genome.

Type		Copy number	Average length(bp)	Total length(bp)	% of genome
miRNA		322	90.46	29,129	0.002321
tRNA		345	74.77	25,795	0.002055
rRNA	rRNA	2,071	418.39	866,479	0.06904
	18S	321	875.85	281,147	0.022401
	28S	1,399	384.66	538,140	0.042878
	5.8S	162	154.95	25,102	0.002
	5S	189	116.88	22,090	0.00176
snRNA	snRNA	276	124.58	34,384	0.00274
	CD-box	131	94	12,314	0.000981
	HACA-box	86	149.55	12,861	0.001025
	splicing	42	147.93	6,213	0.000495
	scaRNA	16	183.44	2,935	0.000234

Table 12. Summary of non-coding RNA annotation in the *Pterorhinus courtoisi* genome.

Secondly, due to the high conservation of rRNAs, known rRNA sequences from *P.courtoisi* (Gcou), *G. sannio* (Gsan), and the zebra finch (*T. guttata*, Tgut) were selected as reference templates. These sequences were aligned to the target genome using BLAST to locate the 18S, 5.8S, 28S, and 5S rRNA gene loci.

Other ncRNAs, including miRNAs, snRNAs, C/D-box snoRNAs, and others, were identified using Infernal[49] (http://eddylab.org/infernal) under default parameters by searching against covariance models from the Rfam database[50].

All predicted ncRNA annotations were then cross-validated against experimentally verified ncRNA databases to obtain a high-confidence ncRNA annotation set for the Pterorhinus courtoisi genome (see Table 12).

Data Records

The sequencing reads generated in this study have been deposited in the Genome Sequence Archive (GSA)[51,52] under the BioProject accession number CRA030745, which includes the genome survey data (CRX2037144[53], CRX2037145[54], CRX2037146[55], CRX2037147[56]), Hi-C data (CRX2037148[57], CRX2037149[58]), and Iso-Seq/RNA-Seq data (CRX2037150[59], CRX2037151[60], CRX2037152[61], CRX2037153[62], CRX2037154[63], CRX2037155[64], CRX2037156[65]).The chromosome-level assembled genome has been deposited in the NCBI GenBank database (BioProject accession: PRJNA1406269; Assembly accession: GCA_054913815.1[66]).

Technical Validation

To generate a high-quality reference genome, rigorous quality control and data filtering pipelines were applied separately to second- and third-generation sequencing data.

For the Illumina paired-end reads, raw data were processed using Novogene's proprietary pipeline pk_qc.v2 for fine-grained filtering, supplemented by redup.v2 to remove redundant sequences. Specifically, any reads containing adapter contamination were discarded outright; reads with undetermined bases (N) exceeding 10% of their length, along with their paired mates, were removed. Additionally, paired reads were excluded if more than 20% of bases in either read had a quality score below 5. Only high-quality clean reads passing these filters were output in FASTQ format and used as input for all downstream analyses to minimize the impact of low-quality sequences.

For PacBio Sequel raw data, systematic preprocessing was also performed. Raw polymerase reads exhibit a characteristic "dumbbell" structure with adapters at both ends. These adapter sequences were trimmed at cleavage sites to yield subreads. Subreads shorter than 50 bp were filtered out. High-fidelity (HiFi) reads with quality scores above Q20 were generated using the PacBio ccs program (https://github.com/PacificBiosciences/ccs) with parameters min_passes = 3 and min_rq = 0.99. This HiFi dataset constituted the core input for genome assembly and validation.

To comprehensively assess the completeness and accuracy of the final assembly, five independent and complementary quality evaluation strategies were integrated:

K-mer spectrum analysis to estimate genome size, repeat content, and heterozygosity;

GC-depth distribution analysis to detect systematic biases in sequencing or assembly;

BUSCO and CEGMA analyses to evaluate assembly completeness based on single-copy orthologous gene sets;

Merqury, a reference-free method, to calculate consensus quality values (QV) and genome completeness by comparing k-mer consistency between reads and assembly.

These multidimensional metrics mutually corroborated each other, providing a robust foundation for subsequent functional annotation and comparative genomics.

Data availability

The sequencing reads generated in this study, including genome survey, Hi-C, and RNA-Seq/Iso-Seq data, have been deposited in the Genome Sequence Archive of the National Genomics Data Center under BioProject accession number CRA030745. The chromosome-level genome assembly is available in the NCBI GenBank database under BioProject accession PRJNA1406269 and Assembly accession GCA_054913815.1.

Code availability

No custom code was generated for the data curation or validation in this study. All data preprocessing was conducted using Novogene's proprietary and standardized bioinformatics pipelines (including the pk_qc.v2 and redup.v2 modules). The specific algorithms and implementation details of these modules are confidential and proprietary to Novogene.

Received: 10 October 2025; Accepted: 23 February 2026;
Published online: 20 March 2026

References

1. Zheng, G. M. Zhongguo Niaolei Fenlei Yu Fenbu Minglu [A Checklist on the Classification and Distribution of the Birds of China] 4th edn (Science Press, 2023).
2. National Forestry and Grassland Administration & Ministry of Agriculture and Rural Affairs of China. Guojia zhongdian baohu yesheng dongwu minglu (revised 1 February 2021). *Yesheng Dongwu Xuebao* **42**, 605–640 (2021).
3. IUCN. The IUCN Red List of Threatened Species. https://www.iucnredlist.org/species/22732350/131890764 (2025).
4. Shi, J. Z. Wuyuan Languan Zaomei (Garrulax courtoisi) fanzhi shengtai ji zhongqun shengcunli fenxi [Breeding Ecology and Population Viability Analysis of the Blue-crowned Laughingthrush (Garrulax courtoisi) in Wuyuan]. Master's thesis, Northeast Forestry University (2017).
5. Wenger, A. M. *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162, https://doi.org/10.1038/s41587-019-0217-9 (2019).
6. Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
7. Li, R. *et al*. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272, https://doi.org/10.1101/gr.097261.109 (2010).
8. Cheng, H. *et al*. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
9. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845, https://doi.org/10.1038/s41477-019-0487-8 (2019).
10. Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I. & Mesirov, J. P. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
11. Dudchenko, O. *et al*. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds. *Cell Syst.* **6**, 104–115.e3, https://doi.org/10.1016/j.cels.2018.01.011 (2018).
12. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* **43**, 1059–1065, https://doi.org/10.1038/ng.947 (2011).
13. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.* **1**, e323, https://doi.org/10.1002/cpz1.323 (2021).
14. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067, https://doi.org/10.1093/bioinformatics/btm071 (2007).
15. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
16. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
17. Rhie, A. *et al*. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
18. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
19. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11, https://doi.org/10.1186/s13100-015-0041-9 (2015).
20. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0 http://www.repeatmasker.org (2013–2015).
21. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25**, 4.10.1–4.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
22. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268, https://doi.org/10.1093/nar/gkm286 (2007).

23. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358, https://doi.org/10.1093/bioinformatics/bti1018 (2005).
24. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA.* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
25. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461, https://doi.org/10.1093/bioinformatics/btq461 (2010).
26. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644, https://doi.org/10.1093/bioinformatics/btm613 (2008).
27. Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Curr. Protoc. Bioinformatics* **18**, 4.3.1–4.3.28, https://doi.org/10.1002/0471250953.bi0403s18 (2007).
28. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94, https://doi.org/10.1006/jmbi.1997.0951 (1997).
29. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open-source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879, https://doi.org/10.1093/bioinformatics/bth315 (2004).
30. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59, https://doi.org/10.1186/1471-2105-5-59 (2004).
31. Grabherr, M. G. *et al*. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
32. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360, https://doi.org/10.1038/nmeth.3317 (2015).
33. Kim, D. *et al*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36, https://doi.org/10.1186/gb-2013-14-4-r36 (2013).
34. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
35. Trapnell, C. *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515, https://doi.org/10.1038/nbt.1621 (2010).
36. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402, https://doi.org/10.1093/nar/25.17.3389 (1997).
37. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995, https://doi.org/10.1101/gr.1865504 (2004).
38. Chen, H. *et al*. Genomic signatures and evolutionary history of the endangered blue-crowned laughingthrush and other Garrulax species. *BMC Biol.* **20**, 188, https://doi.org/10.1186/s12915-022-01386-0 (2022).
39. Ensembl. Taeniopygia guttata (zebra finch) genome assembly taeGut1, release 86. EMBL-EBI & Wellcome Sanger Institute https://www.ensembl.org/Taeniopygia_guttata/Info/Index (2016).
40. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664, https://doi.org/10.1101/gr.229202 (2002).
41. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
42. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
43. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402, https://doi.org/10.1093/nar/25.17.3389 (1997).
44. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531, https://doi.org/10.1093/nar/gkac1052 (2023).
45. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
46. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60, https://doi.org/10.1038/nmeth.3176 (2015).
47. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361, https://doi.org/10.1093/nar/gkw1092 (2017).
48. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964, https://doi.org/10.1093/nar/25.5.955 (1997).
49. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
50. Kalvari, I. *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–D200, https://doi.org/10.1093/nar/gkaa1047 (2021).
51. The Genome Sequence Archive Family. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. *Genomics, Proteomics & Bioinformatics* **19**, 578–583, https://doi.org/10.1016/j.gpb.2021.08.001 (2021).
52. China National Center for Bioinformation. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2025. *Nucleic Acids Res.* **53**, D30–D44, https://doi.org/10.1093/nar/gkae978 (2025).
53. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA030745/CRX2037144 (2025).
54. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA030745/CRX2037145 (2025).
55. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA030745/CRX2037146 (2025).
56. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA030745/CRX2037147 (2025).
57. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA030745/CRX2037148 (2025).
58. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA030745/CRX2037149 (2025).
59. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA030745/CRX2037150 (2025).
60. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA030745/CRX2037151 (2025).
61. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA030745/CRX2037152 (2025).
62. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA030745/CRX2037153 (2025).
63. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA030745/CRX2037154 (2025).
64. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA030745/CRX2037155 (2025).
65. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA030745/CRX2037156 (2025).
66. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_054913815.1 (2026).

Acknowledgements

We thank Minling Li, Jie Liu, Jie Sun, Xinghe Gao, Dandan Wang, Jutao He, Hangmin Gu, and Zhiming Cao for their support and assistance during fieldwork, and we are grateful to Qiang Yang and Yongtao Xu for their guidance on data analysis. This work was supported by the National Natural Science Foundation of China (Grant No. 32360251).

Author contributions

Ouyang Yuxuan was responsible for drafting the manuscript, Zhang Weiwei and Yang Lin revised it and provided guidance, and Cheng Binbin and Xiao Chang handled sample collection and sequencing data organization.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.O. or W.Z.

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.