scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Telomere-to-telomere-level genome assembly and annotation of the Zi goose *Anser cygnoides*

Ke Jiang[1,4], Hechuan Wang[1,4], Kexin Cong[1], Yunuo Liu[1], Jiaxin Yin[1], Xiaofang Ren[1], Zhifeng Chen[2], Kun Yang[2], Ying Zhang[1,3], Qiuju Wang[1,3] & Shengjun Liu[1,3] ✉

This study leveraged advanced genomic sequencing technologies, including Oxford Nanopore technology long reads, Pacific Biosciences HiFi sequencing reads, BGISEQ T7 short reads, and high-throughput chromatin conformation capture technology, to demonstrate the telomere-to-telomere genome assembly and annotation of the Zi goose, a distinctive breed in the cold regions of China. The assembled genome spans 1,305,488,053 bp with exceptional continuity (contig N50: 54,135,724 bp; scaffold N50: 81,055,700 bp). A total of 1,219,987,289 bp were anchored to 40 chromosomes (38 autosomes and a Z/W sex chromosome pair), with 13 chromosomes assembled gap-free, a BUSCO completeness score of 96.8%, and a protein completeness score of 91.7%. Genome annotation predicted a total of 18,359 protein-coding genes, among which 17,592 have been functionally characterized. In addition, 260,266,197 bp of sequences were identified as repetitive sequences, and 655,314 bp were identified as noncoding RNA sequences. This high-quality genome assembly provides valuable genetic resources and a theoretical basis for dissecting the important economic traits of the Zi goose and promoting the development of unique agriculture in cold regions.

Background & Summary

Geese are widely raised poultry with significant economic value, providing humans with nutrient-rich meat and eggs, high-quality hepatic fat, and soft down feathers for clothing production[1]. Furthermore, geese possess distinct physiological traits. For example, through overfeeding, their livers can expand to 5–10 times the weight of normal livers while remaining healthy[2]; this remarkable capacity for lipid accumulation also makes goose livers valuable models for studying hepatic steatosis in humans[3]. In terms of diet, despite being herbivorous poultry, geese avoid most broad-leaved plants and thus can be used for weed and pest control in various crops[4].

In recent years, considerable progress has been made in research on the goose genome. In 2015, Lu *et al.* performed genome sequencing and analysis on the Zhedong white goose, utilizing second-generation sequencing data and Soapdenovo software to generate the first goose genome assembly[5]. Chromosome-level genomes of other goose breeds and related species, including the wild Swan goose[6], Sichuan white goose[7], Tianfu goose[8], and bar-headed goose[9], have since been published. However, previously assembled goose genomes still contain numerous gaps, particularly in telomeric and centromeric regions enriched with repetitive sequences. These regions harbour genetic information crucial for chromosome structure and genetic regulation, yet their high repetitiveness poses significant challenges to accurate sequencing and assembly.

Whole-genome sequencing techniques allow for the comprehensive capture of genetic variation information at single-base resolution in target individuals of a species[10]. Currently, the widely used next-generation sequencing (NGS) technology offers high sequencing accuracy but has significant limitations in handling repetitive sequence regions of the genome because of its short read length (typically 100–150 bp). Third-generation sequencing technologies have largely addressed this drawback. For instance, the high-fidelity (HiFi) sequencing technique developed by Pacific Biosciences, which combines single-molecule real-time (SMRT) sequencing technology, can generate sequence data with both high accuracy and long read lengths[11], thereby significantly

[1]College of Animal Science and Veterinary Medicine, Heilongjiang Bayi Agricultural University, Daqing, 163319, P. R. China. [2]Branch of Animal Husbandry and Veterinary, Heilongjiang Academy of Agricultural Sciences, No. 2, Heyi Street, Qiqihar, 161005, P. R. China. [3]Heilongjiang Provincial Key Laboratory of Exploration and Innovative Utilization of White Goose Germplasm Resources in Cold Region, Daqing, 163319, P. R. China. [4]These authors contributed equally: Ke Jiang, Hechuan Wang. ✉e-mail: lsj4396@163.com

Platform	Tissue	Reads number	Total bases(G)	N50 length(bp)	GC (%)	Sequence coverage(X)
DNBSEQ-T7	blood	206,601,419	61.98	9,704	42.79	47.49
HiFi	blood	2,092,108	37.72	17,984	41.60	28.90
ONT	blood	546,957	47.18	100,001	43.73	36.15
Hi-C	liver	402,086,941	120.63	150	42.84	92.44
RNA-seq	kidney	75,411,696	11.31	150	51.07	8.67
	muscle	77,575,154	11.64	150	52.56	8.92
	ovary	119,436,558	17.92	150	49.74	13.73
	lung	63,681,134	9.55	150	50.34	7.32
	liver	65,081,278	9.76	150	49.09	7.48
	brain	90,572,256	13.59	150	49.16	10.41
	spleen	64,400,820	9.66	150	49.87	7.40
	heart	72,435,558	10.87	150	49.12	8.33
ISO-seq	Mixed	51,553,127	130.60	3,029	46.23	100.08

Table 1. Statistics of sequencing data results.

improving the completeness of the genome assembly. Moreover, the Oxford Nanopore Technologies (ONT) platform, allows the sequencing of ultralong fragments up to hundreds of thousands of nucleotides, exhibiting unique advantages in resolving complex genome structures[12,13]. Additionally, high-throughput chromatin conformation capture (Hi-C) technology can aid in the construction of high-quality chromosome-level genome assemblies, further enhancing the continuity and biological significance of the genome[14–16]. With the advancement of these long-read sequencing technologies, gap-free telomere-to-telomere (T2T) genome assemblies have been constructed for multiple species, including humans[17], cattle and sheep[18], fish[19], chickens[20],and pigeons[21]. In geese, relevant studies have also been reported. In 2024, Zhao *et al.* assembled the first T2T-level genome of geese using the Taihu goose, a characteristic breed from southern China, as the sample, with a total length of 1,197,991,206 bp and a scaffold N50 of 81,007,908 bp[22]. These studies provide invaluable insights into the structural characteristics of complex regions such as telomeres and centromeres.

The Zi goose (*Anser cygnoides*), a species of the genus *Anser* in the family *Anatidae*, is a high-quality local breed originating from Heilongjiang Province in northern China. Despite its small body size, Zi geese exhibit strong cold tolerance[23] and can adapt to the cold climatic conditions in Heilongjiang during the winter, which significantly reduces the breeding risks associated with low temperatures in local winter farming practices. In addition, compared with other goose breeds, the Zi goose has excellent egg-laying performance[24]; thus, it holds important economic value in the development of the local poultry industry. Given the current lack of research on the Zi goose genome, in this study, we performed *de novo* genome assembly using a combination of sequencing data: short-read data from second-generation DNBSEQ, long-read data from third-generation Hi-Fi and ONT technologies, and Hi-C chromatin conformation capture data. Concurrently, we conducted comprehensive genome annotation by integrating RNA-Seq transcriptome data, Iso-seq full-length transcriptome data, and protein databases, ultimately constructing a T2T-level genome for the Zi goose. This high-quality genome assembly will significantly increase our understanding of the genomic structural characteristics of goose breeds in northern China and provide critical theoretical support for genomic research on these breed.

Methods

Sample collection. An adult female Zi goose, sourced from the Heilongjiang Academy of Agricultural Sciences, Animal Husbandry Research Institute, China, served as the research sample. Before slaughter, 5 mL anticoagulant collection tubes (BD Vacutainer® EDTA) were used to collect blood from the wing vein, from which genomic DNA was extracted for second- and third-generation whole-genome sequencing. Eight tissues, including heart, liver, brain, spleen, lung, kidney, ovary, and muscle were harvested, cut into small pieces, placed in 2 mL cryovials, flash-frozen in liquid nitrogen, and then temporarily stored at −80 °C for transcriptome sequencing. For enhanced genome annotation accuracy, equal-volume mixtures of various tissues were used for third-generation full-length transcriptome sequencing. This protocol was approved by the Science and Technology Ethics Committee of Heilongjiang Bayi Agricultural University (Daqing, China; approval No. DWKJXY2024061).

To ensure sufficient gDNA quality and quantity, a NanoPhotometer® spectrophotometer (IMPLEN, CA, USA) was used to assess DNA purity, and a Qubit® DNA Assay Kit and a Qubit® 2.0 fluorometer (Life Technologies, CA, USA) were used to measure DNA concentration. Acceptable DNA samples had an $OD_{260}/_{280}$ ratio of 1.8–2.0 and a concentration >100 ng/μL, with integrity verified via 1% agarose gel electrophoresis. Those DNA samples were stored at −80 °C. Total RNA was extracted from tissues using the TRNzol method, and its concentration, purity, and integrity were determined with an Agilent 5400 Bioanalyzer. Eligible RNA samples were stored at −80 °C to preserve integrity for subsequent analyses.

Library construction and sequencing. For short-read sequencing, 1.5 μg of genomic DNA was first fragmented using Covaris-mediated ultrasonication, followed by end repair, A-tailing, and adapter ligation. After PCR amplification, paired-end sequencing (150 bp) was performed on the DNBSEQ-T7 platform (MGI Tech Co, Ltd, Shenzhen, China), generating 206,601,419 reads with a GC content of 42.79% (Table 1).

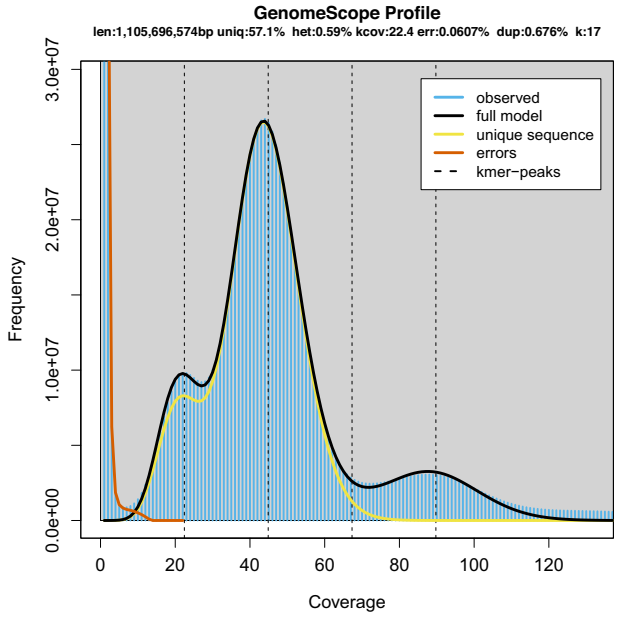

Fig. 1 GenomeScope genome size estimates for the Zi goose.

For HiFi sequencing, high-quality genomic DNA was sheared to an average size of 18 kb using Covaris g-TUBE tubes. Libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, CA, USA), followed by sequencing on the PacBio RS II platform. This process yielded 2,092,108 HiFi reads, with an N50 length of 17,984 bp and a GC content of 41.60% (Table 1).

For ONT sequencing, library preparation was performed using the Nanopore SQK-LSK109 kit (Oxford Nanopore Technologies, CA, USA). Afterwards, the library was loaded into an R9.4 nanopore flow cell and sequenced on the PromethION platform. A total of 546,957 reads were obtained, with an N50 length of 100,001 bp and a GC content of 43.73% (Table 1).

For Hi-C sequencing, liver tissue was homogenized in liquid nitrogen and cross-linked with 4% formaldehyde under vacuum at room temperature for 30 min. The nuclei were resuspended in NEB buffer, and the chromatin was solubilized with SDS. DNA was digested with MboI, followed by biotin labelling of sticky ends and removal of unligated biotin tags. Fragments were sonicated to approximately 350 bp, end-repaired, and used for library construction. Paired-end 150 bp sequencing on the Illumina NovaSeq 6000 platform (Illumina, CA, USA) generated 402,086,941 reads with a GC content of 42.84% (Table 1).

For transcriptome sequencing, total RNA isolated from heart, spleen, liver, brain, lung, kidney, ovary and pectoral muscle tissues. mRNA was enriched with Oligo (dT) magnetic beads and reverse transcribed into cDNA, and 150 bp end-to-end sequencing was performed on the DNBSEQ-T7 platform (MGI Tech Co, Ltd, Shenzhen, China). Among all the tissues, the ovaries produced the most reads (119,436,558), with a GC content of 49.74% (Table 1).

For full-length transcript acquisition, ISO-seq was performed using the PacBio Sequel system (Pacific Biosciences, CA, USA). Qualified RNA was used to construct cDNA libraries with the SMARTer PCR cDNA Synthesis Kit, followed by sequencing on the PacBio Sequel IIe platform. This generated 51,553,127 reads with an N50 length of 3,029 bp and a GC content of 46.23% (Table 1).

Genome survey. Following quality control of the DNBSEQ sequencing data using fastp v0.23.2 software[25] with default parameters, clean reads were subjected to K-mer analysis. First, Jellyfish v2.2.7 software[26] was used to calculate 17-mer frequencies and generate a K-mer frequency distribution table. Specifically, genome characteristics were estimated based on 17-mer profiles using a three-step Jellyfish workflow: (1) k-mer counting (jellyfish count -t 2 -m 17 -C -o kmercount.jf < input_reads.fasta >), (2) histogram generation (jellyfish histo kmercount.jf > 17merFreq.histo), and (3) statistical summary (jellyfish stats kmercount.jf > jelly.log). The k-mer size was set to 17, and the -C option was used to canonicalize k-mers from both strands. Subsequently, GenomeScope v2.0 software[27] was used to model and analyse the K-mer frequency data. The results revealed a Zi goose genome size of 1,105,696,574 bp, with a heterozygosity rate of 0.59% and a unique sequence proportion of 57.1% (Fig. 1).

Contig and scaffold assembly. For contig assembly, first, the raw HiFi sequencing data were processed to break the adapter regions and filter out the adapter sequences. The generated subreads were subjected to filtration with a minimum length threshold set at 50 bp. High-quality HiFi reads were subsequently generated using ccs software (parameters: min-passes = 3, min-rq = 0.99). Finally, *de novo* genome assembly was performed using Hifiasm v0.24.0 software[28] with default parameters. Hifiasm is a haplotype-aware assembler that constructs a phased assembly graph, enabling heterozygous alleles to be represented as alternative paths during assembly and thereby facilitating the resolution of heterozygous regions. The primary assembly was used as the main reference

Term	HiFi Contigs (bp)	Hi-C Scaffolds (bp)
N50	42,590,201	81,047,141
N90	7,047,446	6,485,889
Max length	165,737,842	222,224,803
Total length	1,303,912,393	1,303,921,693

Table 2. Statistics of genome assembly results.

Chromosome ID	Length (bp)	Chromosome ID	Length (bp)
Chr01	222,224,303	Chr19	13,864,324
Chr02	165,737,842	Chr20	12,539,600
Chr03	124,048,642	Chr21	11,987,524
ChrZ	85,693,147	Chr22	9,074,446
Chr04	81,046,941	Chr23	7,884,770
Chr05	66,200,929	Chr24	7,296,921
Chr06	42,090,201	Chr25	7,231,402
Chr07	39,171,359	Chr26	6,738,867
Chr08	33,468,513	Chr27	6,485,789
Chr09	27,719,449	Chr28	6,140,342
ChrW	25,512,936	Chr29	5,466,589
Chr10	23,936,533	Chr30	5,458,412
Chr11	22,340,373	Chr31	4,915,013
Chr12	22,164,948	Chr32	4,385,353
Chr13	21,511,482	Chr33	4,025,193
Chr14	19,955,523	Chr34	3,786,409
Chr15	18,185,937	Chr35	3,252,047
Chr16	17,815,345	Chr36	2,898,584
Chr17	17,764,318	Chr37	2,668,542
Chr18	15,482,897	Chr38	1,806,244

Table 3. Statistics of chromosome lengths after clustering results.

for downstream analyses, and alternate contigs were retained as supplementary haplotype information. A contig set with a total length of 1,303,912,393 bp was obtained, with a contig N50 of 42,590,201 bp (Table 2).

After the contig assembly was completed, scaffold construction was performed. Following the filtering of low-quality sequences using fastp v0.23.2 software with default parameters, the Hi-C data were first processed with HiCUP v0.9.2 software[29] (parameters:–longest 800–shortest 0), and only fragments within the range of 0–800 bp were retained. ALLHiC software (parameters:–minREs50–maxlinkdensity 3–NonInformativeRabio 2) was subsequently used to cluster the contigs into homologous chromosome groups based on the Hi-C interaction intensity, optimizing the order and orientation of the contigs[30]. Finally, Juicebox v1.11.08 software with default parameters was employed to determine the final positions and orientations of the scaffolds in the chromosome set according to the interaction signals[31]. The results revealed that the total length of the scaffolds was 1,303,921,693 bp, with a scaffold N50 of 81,047,141 bp (Table 2). A total of 1,219,987,289 bp of scaffolds were anchored to 40 chromosomes, resulting in an anchoring rate of 93.56%. The longest and shortest chromosomes measured 222,224,303 bp and 1,806,224 bp in length, respectively (Table 3). The 40 chromosomes were easily distinguishable in the heatmap (Fig. 2), with strong interaction signals concentrated around the diagonal, indicating high-quality genome assembly. In addition, sequence alignment with the goose reference genome (GCA_040182565.1) retrieved from NCBI revealed chromosomal correspondence between the two genomes, enabling chromosome renaming based on the matched sequences of chromosomes 1–38, as well as the Z and W chromosomes. The alignment results revealed that the Z and W chromosome mapped to chromosomes 4 and 10 of the Zi goose, respectively (Table 3; Fig. 2).

Gap filling. During the scaffold construction process, some contigs were connected with "N" bases as gaps. To fill these gaps in the Hi-C-based chromosome-level genome, we combined the ultralong reads generated by Nanopore platform sequencing and used TGS-GapCloser v1.2.1 software[32] (parameters: "-tgstype ont-reads output.fasta -scaff gap.fasta -min_nread 5 -output"). Correction was performed using Racon v1.4.20 software[33] with default parameters. After gap filling, only 33 unfilled gaps remained in the genome, which were distributed across 27 chromosomes, with 13 chromosomes achieving complete gap-free assembly (Table 4). The assembly contained 305 contigs and 272 scaffolds, with a total scaffold length of 1,305,488,089 bp, a contig N50 of 54,135,724 Mb, and a scaffold N50 of 81,055,700 bp (Table 5). Compared with those in previous studies, such as the goose reference genome (GCA_040182565.1), our assembled Zi goose genome was larger and had longer contig N50 and scaffold

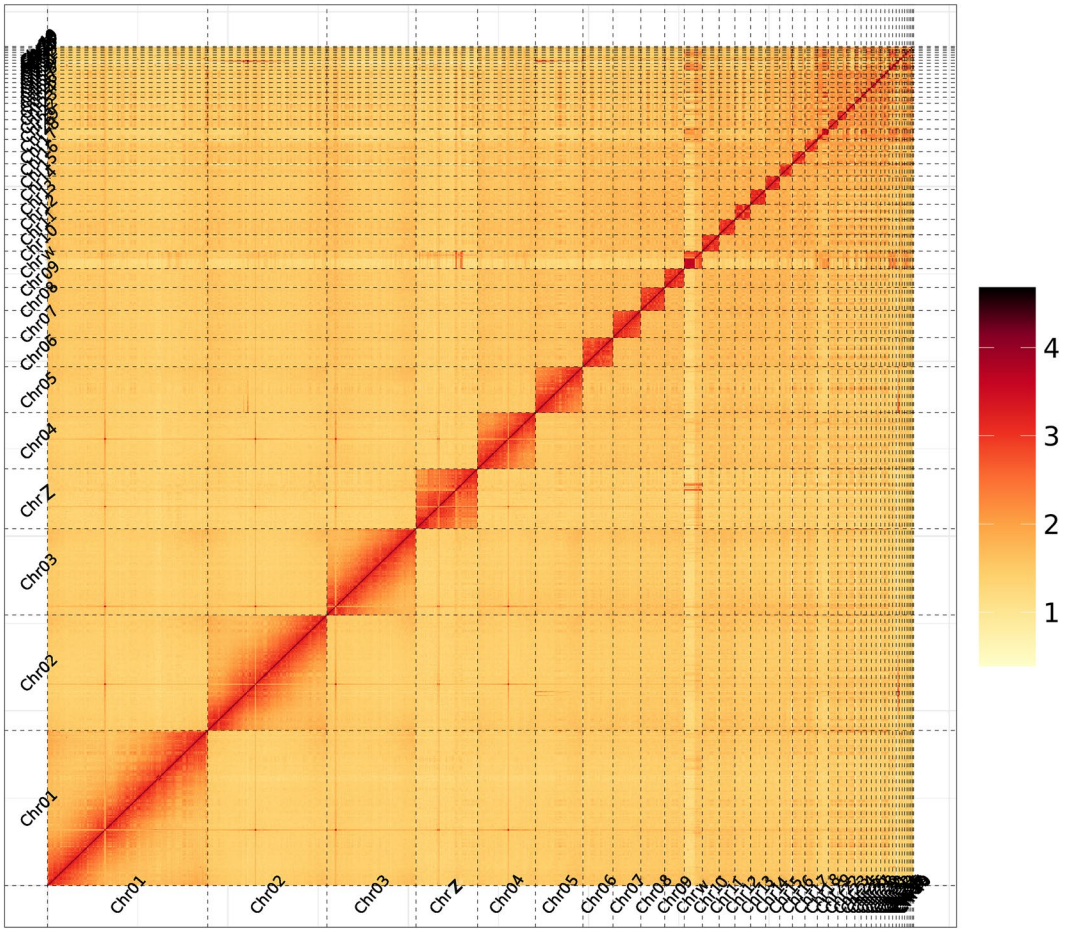

Fig. 2 Genome-wide Hi-C interaction heatmap for the Zi goose.

Chromosome ID	Gap Number	Chromosome ID	Gap Number
Chr01	2	Chr19	1
Chr02	0	Chr20	1
Chr03	1	Chr21	1
ChrZ	2	Chr22	0
Chr04	0	Chr23	1
Chr05	0	Chr24	2
Chr06	0	Chr25	1
Chr07	1	Chr26	0
Chr08	1	Chr27	1
Chr09	1	Chr28	1
ChrW	0	Chr29	0
Chr10	0	Chr30	1
Chr11	0	Chr31	1
Chr12	1	Chr32	0
Chr13	1	Chr33	2
Chr14	1	Chr34	0
Chr15	1	Chr35	1
Chr16	1	Chr36	2
Chr17	1	Chr37	0
Chr18	2	Chr38	1

Table 4. Statistics of genome gaps results.

Type	Contig Length	Contig Number	Scaffold Length	Scaffold Number
Total	1,305,484,789	305	1,305,488,053	272
Max	165,737,842	—	222,461,471	—
Number ≥ 2000	—	305	—	272
N50	54,135,724	7	81,055,700	5
N60	30,593,592	10	42,100,974	7
N70	19,447,608	15	23,933,904	12
N80	9,554,331	24	17,815,445	18
N90	3,753,254	46	6,485,889	29
Total Gap	—	—	3300	33

Table 5. Statistics of genome assembly after gap filling results.

N50, values, indicating excellent continuity of the Zi goose genome.A circos map of the genome was generated using the Circos tool, thereby revealing the basic structure of the Zi goose genome(Fig. 3).

Telomere and centromere identification. Telomere identification on chromosomes of the Zi goose genome was performed using the telomeric repeat sequence TTAGGG, which is specific to most animal species, with tidk v0.2.0 software[34] (parameter: tidk search–string TTAGGG–output < prefix > –dir < output_dir > –window 100000 < genome.fa >). TTAGGG is a 6-bp telomeric repeat motif; therefore, TTAGGG signals were quantified in 100-kb windows at both chromosome termini, and terminal windows showing TTAGGG counts ≥ 100 with tandem repeat arrays were defined as candidate telomeric sequences to distinguish them from sporadic internal matches or non-telomeric terminal sequences.. Centromeric candidate regions were predicted using the CentroMiner module of the quarTeT toolkit[35] with default settings (parameter: quarTeT CentroMiner -i < genome.fa >). The results demonstrated that the assembled Zi goose genome contains a total of 39 telomeres, distributed as follows: telomeres were detected at both ends of 9 chromosomes and at one end of 21 chromosomes. In addition, 20 centromeric regions were identified in the genome assembly (Fig. 4).

Annotation of repetitive sequences. Repetitive sequences in the genome were annotated using a combined approach of homology-based alignment and *de novo* prediction. For homology-based alignment, RepeatMasker v2.0.3 software[36] (parameters: -nolow -no_is -norna -pa 30) was used, and its component RepeatProteinMask was run under default parameters. Genomic sequences were aligned against known transposable elements (TEs) in the widely used Repbase database to identify and annotate repetitive regions[37]. For *de novo* annotation, RepeatModeler v2.0.2 software[38] (parameters: -engine ncbi -pa 30 -LTRStruct) was used to construct a custom database of repetitive elements. Repetitive sequences with a length > 100 bp and unknown base (N) content < 5% were selected to construct the original TE library. Finally, we integrated the Repbase database with the self-constructed TE library that had undergone redundancy removal using UClust, establishing a comprehensive reference set of repetitive sequences. This set was used as input for RepeatMasker to perform genome-wide annotation and analysis of repetitive elements. The results revealed that the total length of the identified repetitive sequences in the goose genome was 260.267,197 bp, accounting for 19.94% of the genome. Within the identified repetitive sequences, short interspersed nuclear elements (SINEs) accounted for 0.04% (506,139 bp) of the genome, long interspersed nuclear elements (LINEs) represented 6.25% (81,652,793 bp) of the genome, long terminal repeats (LTRs) constituted 10.24% (133,711,538 bp) of the genome, DNA transposons made up 1.10% (14,382,088 bp) of the genome, and unclassified transposable elements accounted for 0.34% (4,451,441 bp) of the genome (Table 6).

Prediction of protein-coding genes. Protein-coding genes in the Zi goose genome were systematically annotated through an integrated approach encompassing ab initio prediction, homology alignment, and RNA-Seq-assisted annotation. For homology-based prediction, homologous protein sequences from related species, including Taihu goose (GCF_040182565.1), Sichuan white goose (GCF_002166845.1), Zhedong white goose (GCF_000971095.1), Beijing duck (GCF_047663525.1), and Northern pintail (GCF_963932015.1), were retrieved from the NCBI database. The protein sequences were aligned to the Zi goose genome using TBLASTN v2.2.26 software[39] (parameter settings: -e 1e-10 -v 10000 -b 10000), and the screening threshold was an E value ≤ 1e−5. Precise splicing site alignment was subsequently performed using GeneWise v2.4.1 software[40] (parameter:-tfor -genesf -gff -sum) to predict the corresponding gene structure. For ab initio prediction, automated gene prediction was performed using Augustus v3.5 software[41] (parameters:–species = pasa1–uniqueGeneId = TRUE–noInFrameStop = TRUE–GFF3 = on–genemodel = complete–strand = both) and SNAP v2013-11-29 software[42] (parameters: -gff pasa1.hmm). For RNA-Seq-assisted annotation, transcriptome data were assembled using Trinity v2.8.5 software[43] (parameter:–normalize_reads–full_cleanup–min_glue 2–min_kme_cov 2). Filtered long-read Iso-seq and short-read RNA-seq data were aligned to the Zi goose genome using HISAT2 v2.2.1 software[44] under default parameters to identify exons and splice sites. The alignment results were subsequently processed using StringTie v2.2.1 software[45] with default parameters to predict the transcripts. To integrate the prediction results from the three methods, we extracted the maximum integer value of the average gene length from the GFF files generated by each method as a reference standard, which was input into EvidenceModeler v1.1.1 software[46] (parameter:–segmentSize 200000–overlapSize 20000–min_intron_length 20) for integrated analysis. Moreover, we incorporated terminal exon support information provided by Program

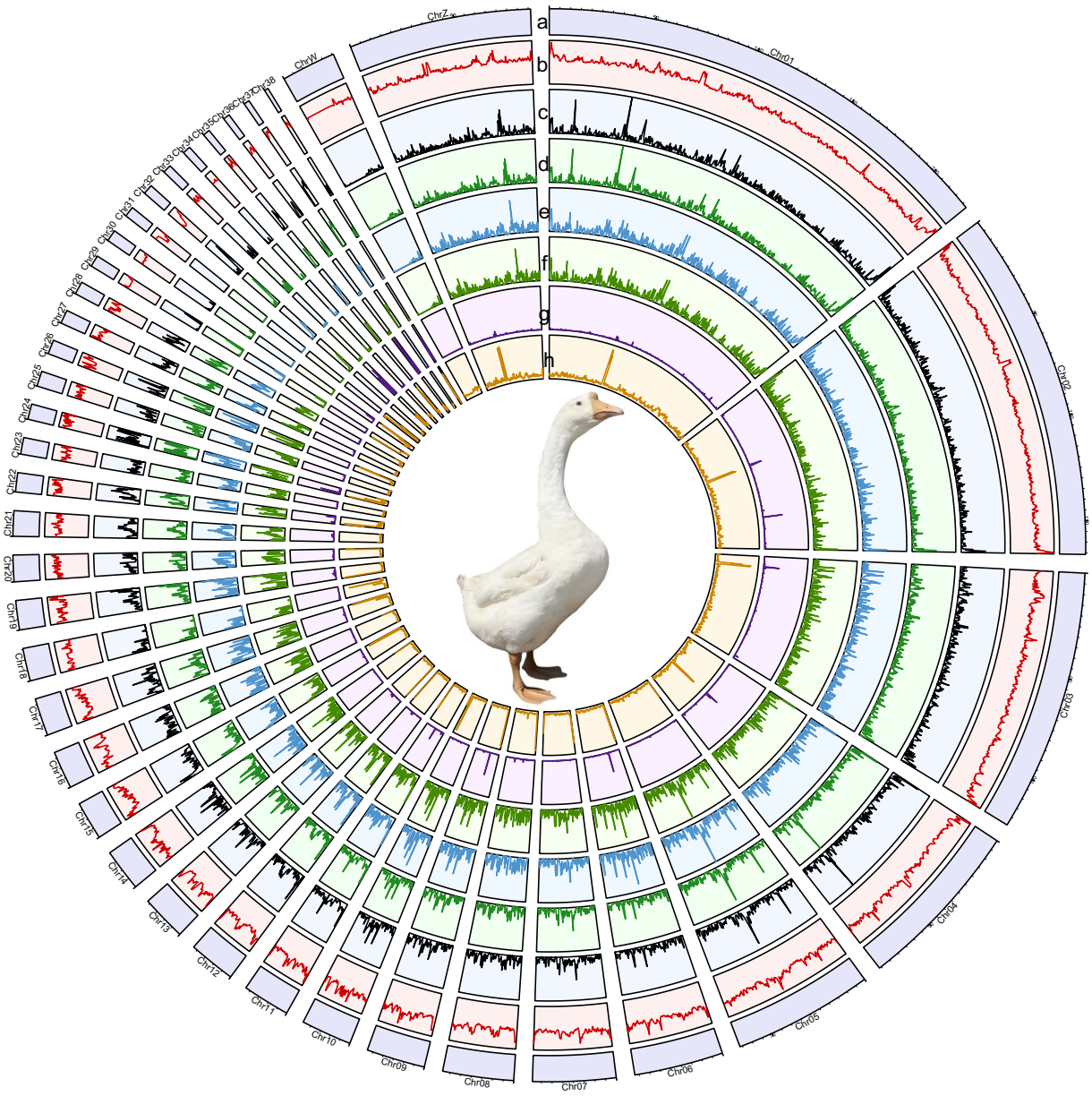

Fig. 3 Circos plot of chromosomes in the Zi goose genome. From outside to inside: (**a**) Chromosomes, (**b**) GC density, (**c**) Gene density, (**d**) mRNA density, (**e**) Exon density, (**f**) CDS density, (**g**) ncRNA density, (**h**) Repeat density, b-h are depicted at a resolution of 300 kb; The centre shows the body shape of the Zi goose used for assembly.

to Assemble Spliced Alignment (PASA) and included masked transposable elements in the prediction process, ultimately constructing a nonredundant high-quality reference gene set. Genome annotation predicted a total of 18,359 protein-coding genes (Table 7). Of these, 17,428 genes (94.93%) were supported by transcript-related evidence, including transcript alignments, rna_0.5 evidence, or expression evidence (RPKM > 1 in at least one sample). Specifically, 16,479 genes were supported by transcript alignments (rna_0.5). For all evidence types, support was defined as an overlap of more than 50% with the gene region (Fig. 5). These genes had a mean length of 26,080.24 bp, with coding sequences (CDSs) averaging 1,765.09 bp in length. Each gene contained, on average, 9.58 exons, with mean exon and intron lengths of 184.29 bp and 2,834.71 bp, respectively (Table 7). We subsequently compared the CDS region length, gene length, number of exons, exon length, and intron length of this species with those of five other closely related species. The results revealed that the distribution of these elements was similar among the closely related species (Fig. 6), which may indicate that this species shares a comparable distribution pattern in gene structure with species whose genomes have been published.

Functional annotation of protein-coding genes. To further annotate gene functions, first, the BLASTP v2.2.26 tool (parameter:-max_target_seqs. 1 -evalue 1e-5) was used to align protein sequences against the Swiss-Prot[47] database (E value ≤ 1e-5), with basic functional annotations assigned to genes based on the best

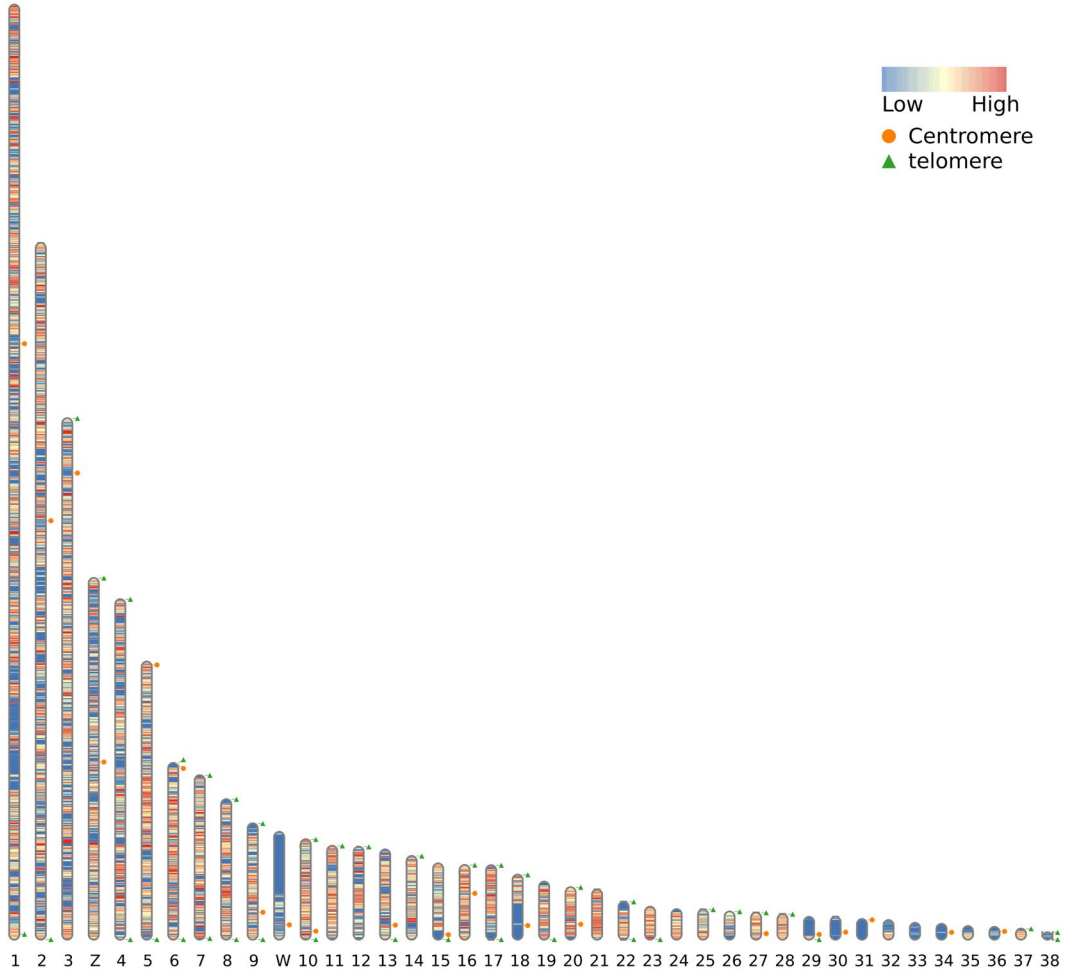

Fig. 4 Distribution of telomeres and centromeres on chromosomes. The numbers indicate the chromosome numbers. The green triangles represent telomeres, and the orange circles represent centromeres. The colour gradient from blue (low) to red (high) denotes the relative signal intensity.

Type	Genome Number	Length(bp)	Percentage(%)
SINE	4,119	506,139	0.04
LINE	162,526	81,652,793	6.25
LTR	546,211	133,711,538	10.24
DNA	67,733	14,382,088	1.10
Unknown	15,547	4,451,441	0.34
Total		260,266,197	19.94

Table 6. Statistics of repeat sequence classification results.

Genome Number	Average gene length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
18,359	26,080.24	1,765.09	9.58	184.29	2,834.71

Table 7. Statistics of structural prediction results.

match results. Concurrently, Diamond v0.8.22 software[48] (parameter:–more-sensitive -k 10 -e 1e-5 -f 6 qseqid-qlen qstart qend sseqid slen sstart sendpident ppos qcovhsp bitscore evalue–salltitles–threads 10) was employed to compare protein sequences with the NR database[49] (E value $\leq$ 1e-5), and functional information from homologous sequences was transferred to supplement gene function annotations. For the analysis of protein structural characteristics, the Pfam database[50] was retrieved using InterProScan v5.39 software[51] (parameter:-cpu 20 -format tsv -appl ProDom,SMART,ProSiteProfiles,PRINTS,Pfam,Panther -iprlookup-dp -goterms) to determine the conserved motifs and domains, and the corresponding GO[52] entries were obtained by association with InterPro

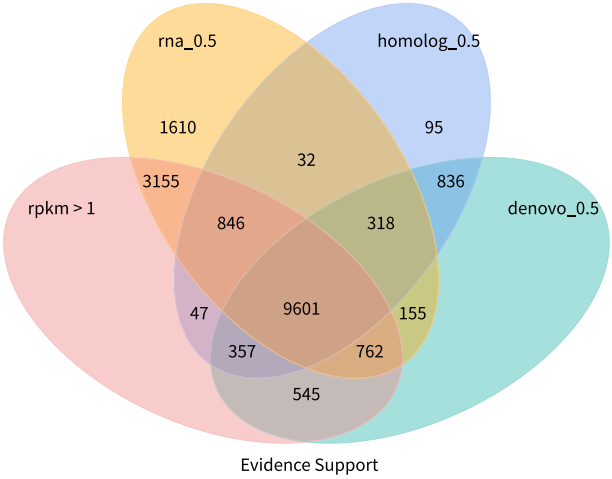

Fig. 5 Venn diagram of gene evidence supporting the Zi goose genome.Regions represent genes supported by *de novo* prediction (denovo_0.5), homologue-based prediction (homolog_0.5), RNA-seq (rna_0.5), active transcription (rpkm > 1), and CDS annotations. Overlap > 50% indicates evidence support; numbers indicate gene counts. Central overlap highlights high-confidence genes, and peripheral regions show unique supported genes.

entries[53]. Additionally, gene sets were mapped to the KEGG pathway database[54] to clarify their functional roles in metabolic and signalling pathways. Ultimately, 17,592 genes were successfully annotated, accounting for 95.82% of all the predicted genes (Table 8).

Annotation of noncoding RNAs. To identify noncoding RNAs (ncRNAs) in the Zi goose genome, we employed a combined strategy integrating database searches and model prediction. First, tRNA sequences were predicted using tRNAscan-SE v1.4[55] with default parameters.. BLASTN v2.2.26 (parameters: -e 1e-10 -v 10000 -b 10000) was subsequently used to align the genomic sequences against a database of known rRNA sequences to identify rRNA-encoding regions. For the identification of miRNA and snRNA sequences, Infernal v1.1.4[56] was used with default parameters to perform alignments against the Rfam database[57]. In total, 2,654 ncRNAs were annotated, including 326 miRNAs, 1,188 rRNAs, 832 tRNAs, and 308 snRNAs (Table 9).

Data Records

The short-read genomic sequencing data generated via DNBSEQ have been deposited in the NCBI Sequence Read Archive under accession number SRR33458851[58].

Long-read genomic sequencing data produced using PacBio HiFi technology are available in the NCBI Sequence Read Archive with the accession number SRR33458852[59].

Ultralong-read genomic sequencing data obtained through Nanopore sequencing have been stored in the NCBI Sequence Read Archive under accession number SRR33593404[60].

Hi-C sequencing data obtained via the Illumina platform have been deposited in the NCBI Sequence Read Archive under accession number SRR33454779[61].

Transcriptome sequencing data from heart, liver, brain, spleen, lung, kidney, muscle, and ovary tissues, along with full-length transcriptome sequencing data from mixed tissues, have been deposited in the NCBI Sequence Read Archive under the following accession numbers: SRR33514019[62], SRR33514020[63], SRR33514021[64], SRR33514022[65], SRR33514023[66], SRR33514024[67], SRR33514025[68], SRR33514026[69], and SRR33513674[70].

The genome assembly files have been submitted to NCBI GenBank under the accession number JBPZTX0000000000[71].

Genome assembly files and annotation files have also been uploaded to Figshare. (https://doi.org/10.6084/m9.figshare.29909237)[72].

Technical Validation

Genome evaluation. The assembly quality of the Zi goose genome was evaluated using BUSCOv5.4.7 software[73] by alignment with the aves_odb10 and vertebrata_odb10 databases of single-copy orthologous genes. For the aves_odb10 library (parameters: -i genome.fa -l aves_odb10 -m genome), the results revealed that the overall completeness of the genome was 96.8%. Specifically, 96.2% of the genes were complete and single-copy genes, 0.6% were complete and duplicated, 0.6% were fragmented, and 2.6% were missing. For the vertebrata_odb10 library (parameters: -i genome.fa -l vertebrata_odb10 -m genome), the results revealed that the overall completeness of the genome was 95.6%. Specifically, 94.7% of the genes were complete and single-copy genes, 0.9% were complete and duplicated, 2.1% were fragmented, and 2.3% were missing. In addition, the completeness of the genome protein sequences was evaluated using BUSCO v5.4.7. For the aves_odb10 library (parameters: -i proteins.fa -l aves_odb10 -m proteins), the results revealed that the overall completeness of the genome was 91.7%. Specifically, 90.9% of the genes were complete and single-copy genes, 0.8% were complete and duplicated, 2.9% were fragmented, and 5.4% were missing. For the vertebrata_odb10 library (parameters: -i proteins.fa -l vertebrata_odb10 -m proteins), the results revealed that the overall completeness of the genome was 90.5%.

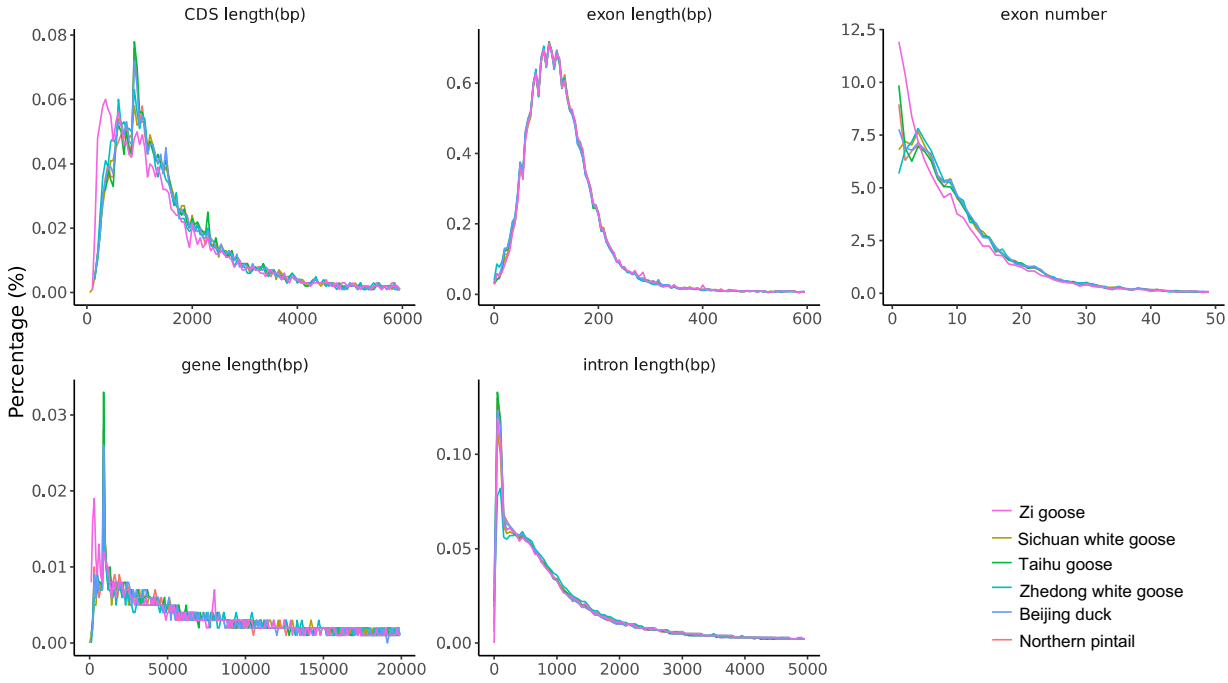

Fig. 6 Comparison of various components among the genomes of closely related species.

Type	Genome Number	Percent(%)
Total	18,359	—
NR	16,961	92.39
Swissport	15,974	87.01
KEGG	15,193	82.76
InterPro	17,115	93.22
Pfam	14,461	78.77
GO	14,353	78.18
Annotated	17,592	95.82%
Unannotated	767	4.18

Table 8. Statistics of gene function annotation results.

Type	Copy number	Average length(bp)	Total length(bp)	Percent(%)
miRNA	326	87.423	28,500	0.002
tRNA	832	73.304	60,989	0.005
rRNA	1,188	443.505	526,884	0.040
snRNA	308	126.432	38,941	0.003

Table 9. Statistics of non-coding RNA annotations results.

Specifically, 89.4% of the genes were complete and single-copy genes, 1.1% were complete and duplicated, 4.6% were fragmented, and 4.9% were missing (Fig. 7). Short-read data were mapped to the genome sequence using BWA v0.7.17 software[74] with default parameters, with the read mapping rate and genome coverage rate reaching 99.75% and 99.61%, Additionally, HiFi long-read sequencing data were aligned to the assembled genome using minimap2 v2.30 (parameters: minimap2 -x map-hifi -t 50 -a genome.fa hifi.fq.gz), achieving a read mapping rate of 99.98% and a genome coverage rate of 99.50%, respectively. ONT long-read sequencing data were also aligned to the assembled genome using minimap2 v2.30 (parameters: minimap2 -x map-ont -t 50 -a genome.fa ont.fq.gz), resulting in a read mapping rate of 99.99% and a genome coverage rate of 99.26%. Collectively, these results demonstrate the high accuracy, consistency, and completeness of the genome assembly. SNP statistics of the genome were conducted using SAMtools v1.9 software[75] with default parameters, and the results revealed that the percentage of all SNPs in the genome was 0.2411%, and the percentage of homologous SNPs was 0.000418%, indicating extremely high single-base accuracy of the genome. Through k-mer analysis using Mercury v1.3 software[76] with default parameters, the genome quality value (QV) estimated based on short-read data was 38.729, corresponding to a base accuracy of approximately 99.99% and confirming the high accuracy of the genome.

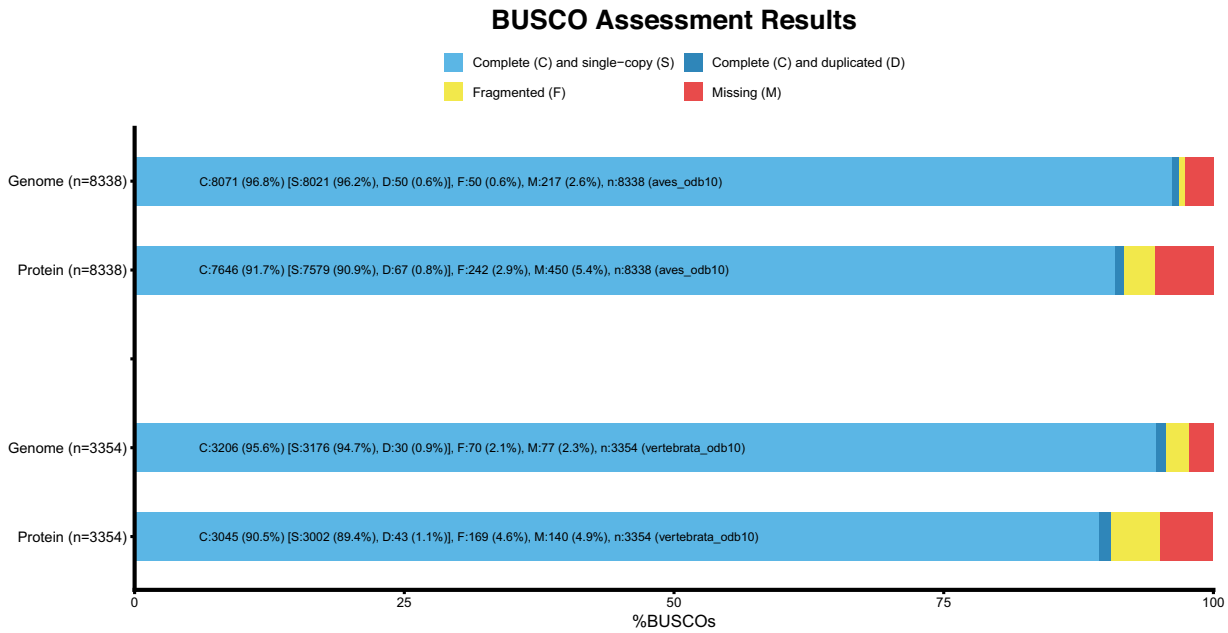

Fig. 7 BUSCO assessment results for Genome and Protein.

Data availability

All raw sequencing data and the genome assembly generated in this study are publicly available. Short-read, long-read, ultralong-read, and high-throughput chromosome conformation capture (Hi-C) genomic sequencing data are available in the NCBI Sequence Read Archive (SRA) under accessions SRR33458851, SRR33458852, SRR33593404, and SRR33454779. Transcriptome sequencing data from heart, liver, brain, spleen, lung, kidney, muscle, and ovary, together with full-length transcriptome sequencing data from a mixed-tissue sample, are available in the SRA under accessions SRR33514019, SRR33514020, SRR33514021, SRR33514022, SRR33514023, SRR33514024, SRR33514025, SRR33514026, and SRR33513674. The genome assembly has been deposited in NCBI GenBank under accession JBPZTX0000000000. The genome annotation files have been deposited in the Figshare database (https://doi.org/10.6084/m9.figshare.29909237).

Code availability

All the software utilized in this study is publicly available, and detailed parameter settings are explicitly described in the Methods section. In cases where specific parameters were not specified, the default values recommended by the software developers were adopted. No custom scripts or programming codes were developed or employed.

Received: 21 August 2025; Accepted: 26 March 2026;
Published online: 31 March 2026

References

1. Yang, H. *et al*. Microbial community and short-chain fatty acid profile in gastrointestinal tract of goose. *Poult Sci* **97**, 1420–1428, https://doi.org/10.3382/ps/pex438 (2018).
2. Hermier, D., Rousselot-Pailley, D., Peresson, R. & Sellier, N. Influence of orotic acid and estrogen on hepatic lipid storage and secretion in the goose susceptible to liver steatosis. *Biochim Biophys Acta* **1211**, 97–106, https://doi.org/10.1016/0005-2760(94)90143-0 (1994).
3. Qin, J. *et al*. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60, https://doi.org/10.1038/nature11450 (2012).
4. Weng, K. *et al*. Effects of marketable ages on meat quality through fiber characteristics in the goose. *Poult Sci* **100**, 728–737, https://doi.org/10.1016/j.psj.2020.11.053 (2021).
5. Lu, L. *et al*. The goose genome sequence leads to insights into the evolution of waterfowl and susceptibility to fatty liver. *Genome Biol* **16**, 89, https://doi.org/10.1186/s13059-015-0652-y (2015).
6. Zhang, Y. *et al*. *De novo* assembly of a wild swan goose (Anser cygnoides) genome. *Anim Genet* **53**, 878–880, https://doi.org/10.1111/age.13262 (2022).
7. Gao, G. *et al*. Genome and metagenome analyses reveal adaptive evolution of the host and interaction with the gut microbiota in the goose. *Sci Rep* **6**, 32961, https://doi.org/10.1038/srep32961 (2016).
8. Li, Y. *et al*. Pacific Biosciences assembly with Hi-C mapping generates an improved, chromosome-level goose genome. *Gigascience* **9**, https://doi.org/10.1093/gigascience/giaa114 (2020).
9. Zhang, Y. *et al*. Chromosome-level genome assembly of the bar-headed goose (Anser indicus). *Sci Data* **9**, 668, https://doi.org/10.1038/s41597-022-01801-9 (2022).
10. Ng, P. C. & Kirkness, E. F. Whole genome sequencing. *Methods Mol Biol* **628**, 215–226, https://doi.org/10.1007/978-1-60327-367-1_12 (2010).
11. Barthélémy, D. *et al*. Direct Comparative Analysis of a Pharmacogenomics Panel with PacBio Hifi(®) Long-Read and Illumina Short-Read Sequencing. *J Pers Med* **13**, https://doi.org/10.3390/jpm13121655 (2023).
12. Gao, Y. Chromosome-level assembly of the Clinopodium gracile genome. *Front Plant Sci* **15**, 1489102, https://doi.org/10.3389/fpls.2024.1489102 (2024).

13. Lu, S. *et al*. Chromosomal-level reference genome of Chinese peacock butterfly (Papilio bianor) based on third-generation DNA sequencing and Hi-C analysis. *Gigascience* **8**, https://doi.org/10.1093/gigascience/giz128 (2019).
14. Belaghzal, H., Dekker, J. & Gibcus, J. H. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* **123**, 56–65, https://doi.org/10.1016/j.ymeth.2017.04.004 (2017).
15. Gong, G. *et al*. A chromosome-level genome assembly of the darkbarbel catfish Pelteobagrus vachelli. *Sci Data* **10**, 598, https://doi.org/10.1038/s41597-023-02509-0 (2023).
16. Belton, J. M. *et al*. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276, https://doi.org/10.1016/j.ymeth.2012.05.001 (2012).
17. Yang, C. *et al*. The complete and fully-phased diploid genome of a male Han Chinese. *Cell Res* **33**, 745–761, https://doi.org/10.1038/s41422-023-00849-5 (2023).
18. Olagunju, T. A. *et al*. Telomere-to-telomere assemblies of cattle and sheep Y-chromosomes uncover divergent structure and gene content. *Nat Commun* **15**, 8277, https://doi.org/10.1038/s41467-024-52384-5 (2024).
19. Bian, C., Hu, C., He, Z., Li, Z. & Shi, Q. A complete telomere-to-telomere chromosome-level genome assembly of X-ray tetra (Pristella maxillaris). *Sci Data* **12**, 496, https://doi.org/10.1038/s41597-025-04824-0 (2025).
20. Huang, Z. *et al*. Evolutionary analysis of a complete chicken genome. *Proc Natl Acad Sci USA* **120**, e2216641120, https://doi.org/10.1073/pnas.2216641120 (2023).
21. Yang, W. *et al*. Near telomere-to-telomere assembly of the Tarim pigeon (Columba livia) genome. *Sci Data* **11**, 1455, https://doi.org/10.1038/s41597-024-04350-5 (2024).
22. Zhao, H. *et al*. Telomere-to-telomere genome assembly of the goose Anser cygnoides. *Sci Data* **11**, 741, https://doi.org/10.1038/s41597-024-03567-8 (2024).
23. Ni, H. *et al*. Multiomics analysis uncovers host-microbiota interactions regulate hybrid vigor traits in geese. *Poult Sci* **104**, 105289, https://doi.org/10.1016/j.psj.2025.105289 (2025).
24. Ni, H. *et al*. Potential regulator of meat quality in geese: C1QTNF1 implications on cell proliferation and muscle growth. *Poult Sci* **103**, 103927, https://doi.org/10.1016/j.psj.2024.103927 (2024).
25. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
26. Marçais, G. & 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).
27. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153 (2017).
28. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
29. Wingett, S. *et al*. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310, https://doi.org/10.12688/f1000research.7334.1 (2015).
30. 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).
31. Robinson, J. T. *et al*. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst* **6**, 256–258.e251, https://doi.org/10.1016/j.cels.2018.01.001 (2018).
32. Xu, M. *et al*. TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience* **9**, https://doi.org/10.1093/gigascience/giaa094 (2020).
33. Safar, H. A. *et al*. The impact of applying various **de novo** assembly and correction tools on the identification of genome characterization, drug resistance, and virulence factors of clinical isolates using ONT sequencing. *BMC Biotechnol* **23**, 26, https://doi.org/10.1186/s12896-023-00797-3 (2023).
34. Brown, M. R., Manuel Gonzalez de La Rosa, P. & Blaxter, M. tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. *Bioinformatics* **41**, https://doi.org/10.1093/bioinformatics/btaf049 (2025).
35. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res* **10**, uhad127, https://doi.org/10.1093/hr/uhad127 (2023).
36. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, 4.10.11–14.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
37. 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).
38. 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).
39. Gertz, E. M., Yu, Y. K., Agarwala, R., Schäffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol* **4**, 41, https://doi.org/10.1186/1741-7007-4-41 (2006).
40. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988–995, https://doi.org/10.1101/gr.1865504 (2004).
41. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–439, https://doi.org/10.1093/nar/gkl200 (2006).
42. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59, https://doi.org/10.1186/1471-2105-5-59 (2004).
43. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
44. 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).
45. 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).
46. 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).
47. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45–48, https://doi.org/10.1093/nar/28.1.45 (2000).
48. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368, https://doi.org/10.1038/s41592-021-01101-x (2021).
49. O'Leary, N. A. *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–745, https://doi.org/10.1093/nar/gkv1189 (2016).
50. Mistry, J. *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412–d419, https://doi.org/10.1093/nar/gkaa913 (2021).
51. Quevillon, E. *et al*. InterProScan: protein domains identifier. *Nucleic Acids Res* **33**, W116–120, https://doi.org/10.1093/nar/gki442 (2005).
52. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
53. Finn, R. D. *et al*. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res* **45**, D190–d199, https://doi.org/10.1093/nar/gkw1107 (2017).

54. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457–462, https://doi.org/10.1093/nar/gkv1070 (2016).
55. 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).
56. 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).
57. Kalvari, I. *et al*. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* **46**, D335–d342, https://doi.org/10.1093/nar/gkx1038 (2018).
58. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33458851 (2025).
59. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33458852 (2025).
60. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33593404 (2025).
61. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33454779 (2025).
62. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33514019 (2025).
63. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33514020 (2025).
64. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33514021 (2025).
65. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33514022 (2025).
66. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33514023 (2025).
67. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33514024 (2025).
68. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33514025 (2025).
69. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33514026 (2025).
70. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33513674 (2025).
71. Jiang, K. *et al*. (Anser cygnoides) Zi_goose_T2T_genome. *GenBank* https://identifiers.org/ncbi/insdc:JBPZTX0000000000 (2025).
72. Jiang, K., Wang, H., Ren, X., Yin, J. & Liu, S. Genome annotation files of Zi goose. *figshare* https://doi.org/10.6084/m9.figshare.29909237 (2025).
73. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227–245, https://doi.org/10.1007/978-1-4939-9173-0_14 (2019).
74. 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).
75. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
76. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).

Acknowledgements

This work was supported by the Bio-breeding Industry Innovation and Development Project of Heilongjiang Province (2024), the "Double First-Class" Characteristic Discipline Platform Project of Heilongjiang Province (HLJ2022TSXK), the Bio-breeding Industry Innovation and Development Project of Heilongjiang Province (No. 2023-401), the National Natural Science Foundation of China (32202664), and the Natural Science Foundation of Heilongjiang Province (LH2022C067), the Bio-breeding Industry Innovation and Development Project of Heilongjiang Province "Genome-Wide Association Study of Economically Traits in Zi Geese"(No. 2024-515).

Author contributions

K.J. and S.J.L. co-led the research project. K.J., Z.F.C.,K.Y. and H.C.W. designed the experiments and collected the samples. K.J., H.C.W., K.X.C., Y.N.L., J.X.Y. and X.F.R. conducted the data analysis. Y.Z., Q.J.W. and S.J.L. supervised the study and provided guidance. K.J. wrote the manuscript. All the authors read, revised, and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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