scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Assembling a chromosome-level genome for the *Microtus fortis* using PacBio HiFi and Hi-C technologies

Du Zhang[1,2], Qi Hu[3], Tianqiong He[4,5], Junkang Zhou[4], Yixin Wen[4,5], Qian Liu[4,5], Jing Zhang[4,5], Wenlin Zhi[4,5], Lingxuan Ouyang[4,5], Suisui Gao[4,5], Ruotong Guan[4,5] & Zhijun Zhou[4,5] ✉

The reed vole (*Microtus fortis*) is an important rodent model for studying unique biological traits, such as its natural resistance to *Schistosoma japonicum*. To facilitate the genetic study of these phenotypes, we have produced the first high-quality, chromosome-level genome assembly for this species. The genome was assembled using PacBio HiFi long-read sequencing and scaffolded to the chromosome level with Hi-C data. The final 2.29 Gb assembly exhibits excellent continuity (contig N50 = 68.89 Mb; scaffold N50 = 91.23 Mb), with 97.7% of the sequence anchored into 26 pseudomolecules, consistent with the species' karyotype. Genome completeness was estimated at 96.3% via BUSCO analysis (glires_odb10). The annotation includes 23,678 protein-coding genes, with 97.5% assigned a putative function. This publicly available, high-quality genomic resource will be invaluable for future research, providing the necessary foundation to explore the genetic mechanisms behind the unique adaptations of *M. fortis*, including its innate immunity, digestive physiology, and disease models. The assembly will also serve as a key reference for comparative genomics, enriching our understanding of rodent evolution.

Background & Summary

The reed vole (*Microtus fortis*) is a small rodent with significant value as a model organism across multiple biological disciplines. Its unique physiological traits make it a compelling subject for genetic and biomedical research. These include a specialized digestive system adapted to a high-fiber diet, which offers insights into herbivore metabolism and gut microbiota interactions[1]. Furthermore, it serves as a rare natural model for spontaneous ovarian cancer, providing a clinically relevant system for studying tumorigenesis without artificial induction[2,3]. It is also utilized in behavioral studies to explore social dynamics and other complex behaviors[4]. Most notably, *M. fortis* possesses a remarkable innate resistance to parasites like *Schistosoma japonicum*[5–7], making it an invaluable non-permissive host model for dissecting the genetic underpinnings of anti-parasite immunity. However, the full exploration of the genetic basis for these characteristics has been significantly hampered by the lack of a high-quality genomic reference. Previous genomic resources for *M. fortis* were limited to transcriptomic data or highly fragmented draft assemblies[3,8–10]. Such resources are insufficient for studying large-scale genomic architecture, as they cannot be used to analyze synteny, identify large structural variants, or accurately resolve the structure and copy number of complex and tandemly arrayed gene families, such as those related to immunity. These limitations have prevented a deep investigation into the evolutionary adaptations and the molecular mechanisms underlying the vole's unique phenotypes.

High-quality, chromosome-level reference genomes are foundational for modern genomics, enabling comprehensive analyses of genome evolution, function, and regulation[11]. The advent of third-generation long-read sequencing technologies has revolutionized *de novo* genome assembly. Specifically, PacBio High-Fidelity (HiFi) sequencing, which generates long reads (>10 kb) with very high accuracy (>99.9%), is particularly effective

[1]Department of Medical Genetics, The Second Xiangya Hospital of Central South University, Changsha, 410011, China. [2]Hunan Province Clinical Medical Research Center for Genetic Birth Defects and Rare Diseases, Department of Medical Genetics, The Second Xiangya Hospital of Central South University, Changsha, 410011, China. [3]E-gene Biotechnology Co., Ltd., Shenzhen, 518038, China. [4]Department of Laboratory Animal Science, Xiangya School of Medicine College, Central South University, Changsha, 410013, China. [5]Hunan Key Laboratory of Animal Models for Human Diseases, Central South University, Changsha, 410013, China. ✉e-mail: zhouzhijun@csu.edu.cn

Source	Platform	Library Size	Clean Data	Read Length	Coverage
Genome-short reads	BGI DNBSEQ-T7	300 bp	195.46 Gb	150 bp	94×
Genome-long reads	PacBio Sequel	15 Kb	94.26 Gb	16,058 bp	45×
Hi-C	BGI DNBSEQ-T7	300 bp	233.49 Gb	150 bp	112×

Table 1. Statistics of the DNA sequencing data for the *M. fortis* genome assembly.

for resolving complex repeat regions and heterozygous sequences, thus generating highly contiguous initial assemblies with superior completeness[12]. To elevate such an assembly to a chromosomal scale, these contiguous sequences can be combined with data from chromatin conformation capture techniques like Hi-C. Hi-C data provides empirical, long-range information about the three-dimensional proximity of DNA segments within the nucleus. This orthogonal dataset is ideal for accurately ordering and orienting the assembled contigs into chromosome-length pseudomolecules, which is essential for validating karyotype and enabling studies of genome-wide structural organization[13].

To address the critical resource gap for *M. fortis*, we employed this powerful hybrid assembly strategy, integrating deep-coverage PacBio HiFi long-read data with extensive Hi-C-based scaffolding. This approach has allowed us to produce the first high-quality, chromosome-level reference genome for the species, overcoming the specific limitations of previous draft versions. This Data Descriptor provides a detailed account of the sample collection, sequencing protocols, assembly pipeline, and annotation methods used, presenting a robust and valuable genomic resource that will facilitate advanced, previously intractable research into the unique biology of *M. fortis*.

Methods

Sample collection and preparation. A healthy adult male *Microtus fortis* was sourced from the Dongting Lakes region and subsequently maintained at the Xiangya Medical College, Central South University (Changsha, China). To minimize allelic variation and simplify the assembly process, tissues from a single individual were used for genome sequencing. Skeletal muscle tissue was collected for whole-genome shotgun sequencing, Hi-C library construction, and RNA sequencing. An additional liver tissue was also collected for whole-genome shotgun sequencing. Immediately following dissection, all samples were immersed in liquid nitrogen for rapid freezing and subsequently stored at −80 °C to maintain integrity. All procedures were conducted in strict accordance with institutional guidelines and were approved by the Laboratory Animal Welfare and Ethics Committee of Central South University (Changsha, China; approval no. CSU-2022-0654).

Genome sequencing. Total genomic DNA was isolated from skeletal muscle tissue using a TIANamp Genomic DNA Kit (Tiangen Biotech, Beijing, China). For genome survey purposes, a short-read library with an insert size of approximately 300 bp was constructed using DNBSEQ technology. The genomic DNA was fragmented, and selected fragments underwent end-repair, A-tailing, and ligation of sequencing adapters. The adapter-ligated products were then heat-denatured and circularized by splint oligo hybridization. Unligated linear DNA molecules were digested, and the resulting single-stranded circular DNA molecules were amplified via rolling circle amplification (RCA) to create DNA Nanoballs (DNBs). The DNBs were loaded onto sequencing flow cells and sequenced on the DNBSEQ-T7 platform (MGI, Shenzhen, China), generating a total of 195.46 Gb of 150 bp paired-end data (Tables 1).

For long-read sequencing, a SMRTbell library with an insert size of approximately 15 kb was prepared using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA). High-molecular-weight DNA was sheared to the target fragment size. The fragmented DNA was then subjected to a DNA damage repair step, followed by an end-repair and A-tailing process. Hairpin SMRTbell adapters were ligated to both ends of the DNA fragments, creating a closed, single-stranded circular topology. Exonuclease treatment was performed to remove any remaining unligated linear DNA fragments. The resulting SMRTbell templates were size-selected to obtain the desired 15 kb library. A sequencing primer was annealed to the SMRTbell templates, which were then bound with DNA polymerase. This library was sequenced on a PacBio Sequel II platform (Pacific Biosciences, Menlo Park, CA, USA) using the circular consensus sequencing (CCS) mode, which generates highly accurate HiFi reads by sequencing the same molecule multiple times. This process yielded a total of 94.26 Gb of HiFi long reads (Table 1).

Hi-C sequencing. For chromosome-level scaffolding, an *in situ* Hi-C library was constructed following established protocols[14]. Approximately 1 g of skeletal muscle tissue was finely minced and treated with 2% formaldehyde for 30 minutes to crosslink proteins and DNA, thereby preserving the native three-dimensional chromatin architecture. The crosslinked cells were lysed, and the intact nuclei were purified. The chromatin was then digested overnight using the 4-cutter restriction enzyme DpnII, which creates sticky ends. The digested fragment ends were filled in with nucleotides, including a biotinylated dATP, to mark the original termini of the interacting fragments. Next, proximity ligation was performed in a dilute solution, which favors the ligation of fragments that were spatially adjacent within the nucleus. After ligation, the crosslinks were reversed by proteinase K treatment and overnight incubation at 65 °C. The DNA was purified, and non-ligated biotinylated ends were removed. The purified DNA, now enriched for chimeric ligation products, was sheared to a size range of 300–600 bp. The biotin-containing fragments, representing the Hi-C junctions, were captured and enriched using streptavidin-coated magnetic beads. Finally, the enriched fragments were processed into a sequencing library

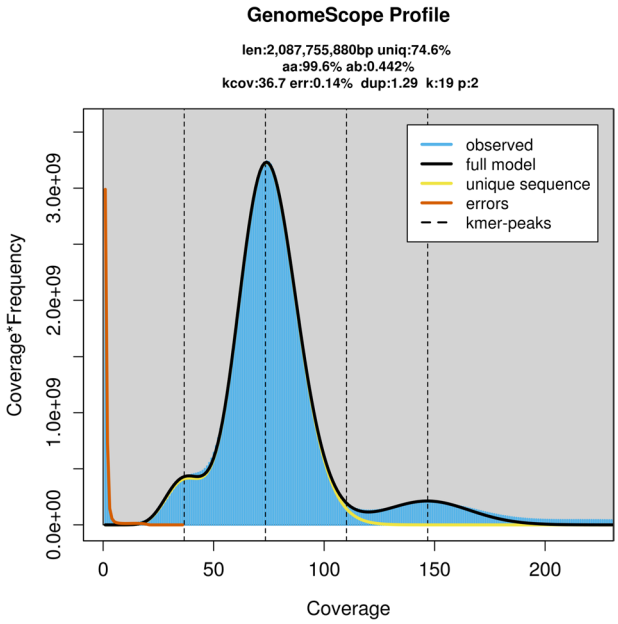

Fig. 1 The 19-mer count distribution for the genome size estimation. The genome size, heterozygous rate, and repeat content of *M. fortis* were estimated to be 2.09 Gb, 0.44% and 25.37% respectively.

Item	Contigs	Scaffolds
Total length (bp)	2,286,496,162	2,286,498,262
Total number	107	88
GC content (%)	42.00	42.48
N50 (bp)	68,894,778	91,234,584
N90 (bp)	25,142,162	63,503,454
Max length (bp)	116,338,118	132,712,573

Table 2. Statistics of the final genome assembly of *M. fortis*.

through end-repair, A-tailing, and adapter ligation. The final Hi-C library was sequenced on the DNBSEQ-T7 platform (MGI, Shenzhen, China), producing approximately 233.49 Gb of 150 bp paired-end data (Table 1).

RNA-seq sequencing. Total RNA was extracted from skeletal muscle tissue using a Tiangen RNA extraction kit (Tiangen Biotech, Beijing, China). For transcriptome-assisted annotation, a DNBSEQ RNA-seq library was prepared. First, messenger RNA (mRNA) was enriched from the total RNA using oligo(dT)-coated magnetic beads to specifically capture polyadenylated transcripts. The enriched mRNA was then randomly fragmented. Using these fragments as templates, first-strand cDNA was synthesized with random hexamer primers and reverse transcriptase. Subsequently, second-strand cDNA was synthesized using DNA polymerase I and RNase H. The resulting double-stranded cDNA fragments underwent end-repair, A-tailing, and ligation of sequencing adapters. The adapter-ligated fragments were then amplified by PCR to create the final library, which was sequenced on the DNBSEQ platform (MGI, Shenzhen, China), generating approximately 12 Gb of new transcriptomic data from skeletal muscle. To further enhance the comprehensiveness of the gene annotation, this newly generated data was combined with previously published RNA-seq datasets from liver (NCBI BioProject: PRJNA395088) and ovary (NCBI BioProject: PRJNA687349) tissues[3,10].

Genome survey. To gain preliminary insights into the genomic landscape of *M. fortis* prior to assembly, a k-mer-based analysis was performed on the high-coverage short-read data. The frequency of all 19-base-pair subsequences (19-mers) was counted from approximately 195.46 Gb of quality-filtered reads using JELLYFISH v2.2.10[15], generating a k-mer frequency distribution histogram. The resulting distribution was then analyzed with GenomeScope v2.0[16], which models the genome's properties by fitting the k-mer profile to statistical expectations. The model identified a main peak corresponding to homozygous, single-copy k-mers and a smaller, secondary peak at half the depth, representing heterozygous k-mers (Fig. 1). Based on the position and relative size of these peaks, a haploid genome size of approximately 2.09 Gb, a genomic heterozygosity rate of 0.44%, and a repeat content of 25.37% were estimated. These pre-assembly metrics were crucial for guiding the assembly strategy and for later validation of the final assembly size and complexity.

Chromosome	Length(bp)	Contig number
chr1	132,712,573	2
chrX	128,540,061	8
chr2	116,338,118	1
chr3	110,622,173	2
chr4	107,954,820	3
chr5	102,180,330	3
chr6	101,054,257	2
chr7	97,308,709	1
chr8	95,173,695	1
chr9	94,808,429	2
chr10	91,234,584	1
chr11	89,400,603	1
chr12	87,512,534	2
chr13	82,670,725	1
chr14	77,762,427	1
chr15	74,370,358	1
chr16	74,325,793	1
chr17	71,851,052	1
chr18	71,417,391	1
chr19	71,008,403	2
chr20	69,131,911	2
chr21	66,553,439	1
chr22	63,503,454	1
chr23	61,786,958	2
chr24	55,423,275	1
chr25	39,920,996	3
chrUn	51,931,194	62

Table 3. Chromosome information of the *M. fortis* genome assembly.

Contig assembly. The initial contig-level assembly was generated from the PacBio HiFi long reads using Hifiasm v0.16.1-r375[17], which is specifically designed for long, accurate reads and constructs a phased assembly graph to resolve heterozygous regions. This initial step produced a primary assembly containing both primary contigs and alternative haplotigs. To create a single, non-redundant reference for scaffolding, this diploid assembly was processed with Purge_dups v1.2.5[18]. This tool identifies and removes redundant sequences corresponding to alternative alleles by analyzing read depth coverage across the assembled contigs. After this purging step, a clean, haploid representation of the genome was obtained. This final contig-level assembly spanned 2.29 Gb, comprising 107 contigs with a high contig N50 length of 68.89 Mb (Table 2), indicating excellent contiguity prior to scaffolding.

Chromosome-level genome assembly. The highly contiguous contig-level assembly was scaffolded into chromosome-level pseudomolecules using the high-coverage Hi-C data. First, the raw Hi-C reads were aligned to the contig assembly. The Juicer pipeline (v1.6)[19] was then employed to process these alignments, filter out invalid pairs (e.g., self-ligations, random breaks), and generate a genome-wide contact map that quantifies the interaction frequency between all pairs of contigs. Subsequently, the 3D-DNA pipeline (v190716)[20] utilized this contact map to automatically cluster, order, and orient the contigs into large scaffolds corresponding to chromosomes. For quality control and refinement, the resulting scaffolds were subjected to a final polishing step via manual review in Juicebox (v1.11.08)[21]. The initial automated scaffolding was of high quality, requiring only minor corrective edits. This manual process involved resolving a small number of clear, unambiguous misjoins and connecting a few scaffolds where strong, contiguous Hi-C interaction signals provided clear evidence of their adjacency. All manual edits were conservative and strictly guided by the Hi-C contact patterns to ensure the final accuracy of the chromosome-level assembly. The final, polished chromosome-level assembly anchors 97.73% of the sequence into 26 pseudomolecules (Table 3 and Fig. 2), consistent with previous karyotype studies of *M. fortis* from the same geographical region[22]. The assembly exhibits exceptional long-range contiguity, with a scaffold N50 of 91.23 Mb. To assess its genic completeness, the assembly was evaluated with BUSCO v5.4.6[23] against the glires_odb10 lineage dataset, which identified 96.3% of the expected conserved genes as complete, confirming the high quality of the final genome assembly (Table 4).

Identification of sex chromosomes. Depth-of-coverage analysis of both short-read (DNBSEQ-T7) and long-read (PacBio HiFi) data revealed that chrX exhibited approximately half the average coverage compared to other chromosomes. This indicates a hemizygous state, confirming the sex chromosomes. This identification was further validated by genome-wide synteny analysis, which showed that chrX is highly homologous to the X

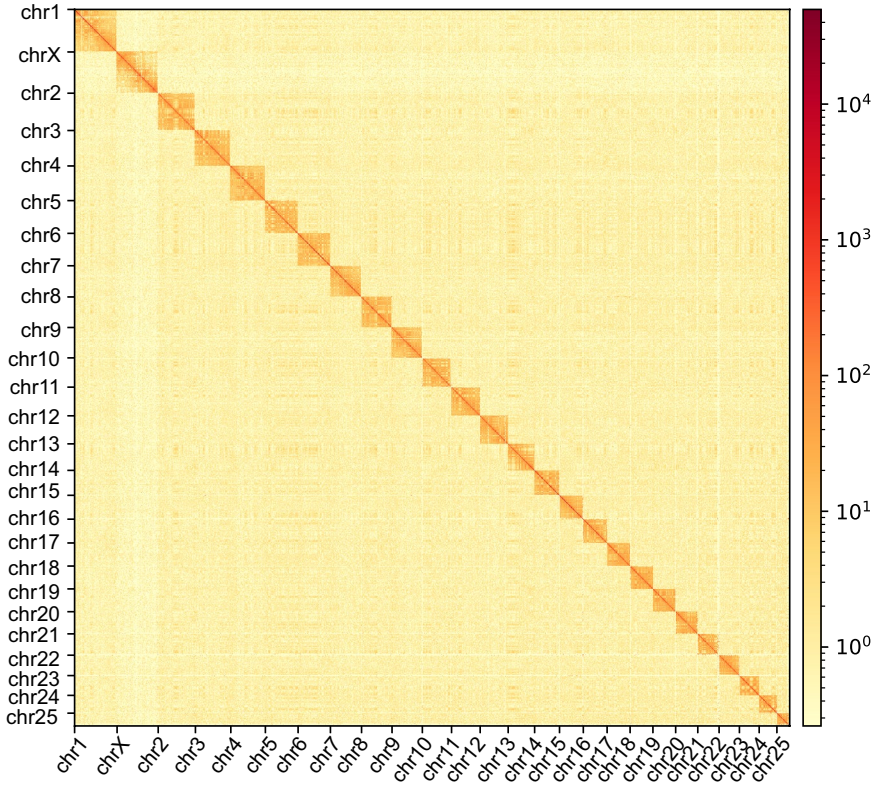

Fig. 2 Hi-C interaction heatmap of the *M. fortis* genome assembly. The heatmap illustrates the contact density between genomic regions. The 26 assembled chromosomes are arranged in order of size. The intense signal along the diagonal of each chromosome block indicates high interaction frequency within each chromosome, validating the high quality of the chromosome-level scaffolding. The color bar reflects the logarithm of contact density, from high (dark red) to low (light yellow).

Type	Proteins	Percentage (%)
Complete BUSCOs	13,290	96.3
Complete Single-Copy BUSCOs	12,931	93.7
Complete Duplicated BUSCOs	359	2.6
Fragmented BUSCOs	80	0.6
Missing BUSCOs	428	3.1
Total BUSCO groups searched	13,798	100

Table 4. BUSCO assessment of genome completeness for *M. fortis* (glires_odb10).

chromosomes of *Microtus ochrogaster* and *Mus musculus* (Supplementary Fig. 3). Due to its high repetitive content and relatively small size, the Y chromosome was not successfully scaffolded and remains distributed among the 62 unplaced scaffolds.

Gene prediction and functional annotation. To obtain a high-quality gene set, protein-coding genes (PCGs) were predicted on the repeat-masked genome using an integrated, evidence-based strategy that combined three distinct lines of evidence: homology-based prediction, *de novo* gene finding, and transcriptome-based evidence.

Homology-based prediction. Protein sequences from two closely related species (*Microtus ochrogaster* and *Microtus oregoni*) and from a previous *M. fortis* draft assembly were aligned to the new genome using TBLASTN. The gene structures corresponding to these alignments were then predicted using Exonerate v2.2.0[24].

De novo prediction. Gene models were predicted based on the intrinsic properties of the DNA sequence. First, high-quality proteins identified from the RNA-seq dataset were used to train gene prediction models for Augustus v3.3[25] and Genscan[26] using MAKER2 v2.31.10[27]. Then, these trained models were used to perform *ab initio* gene prediction across the genome.

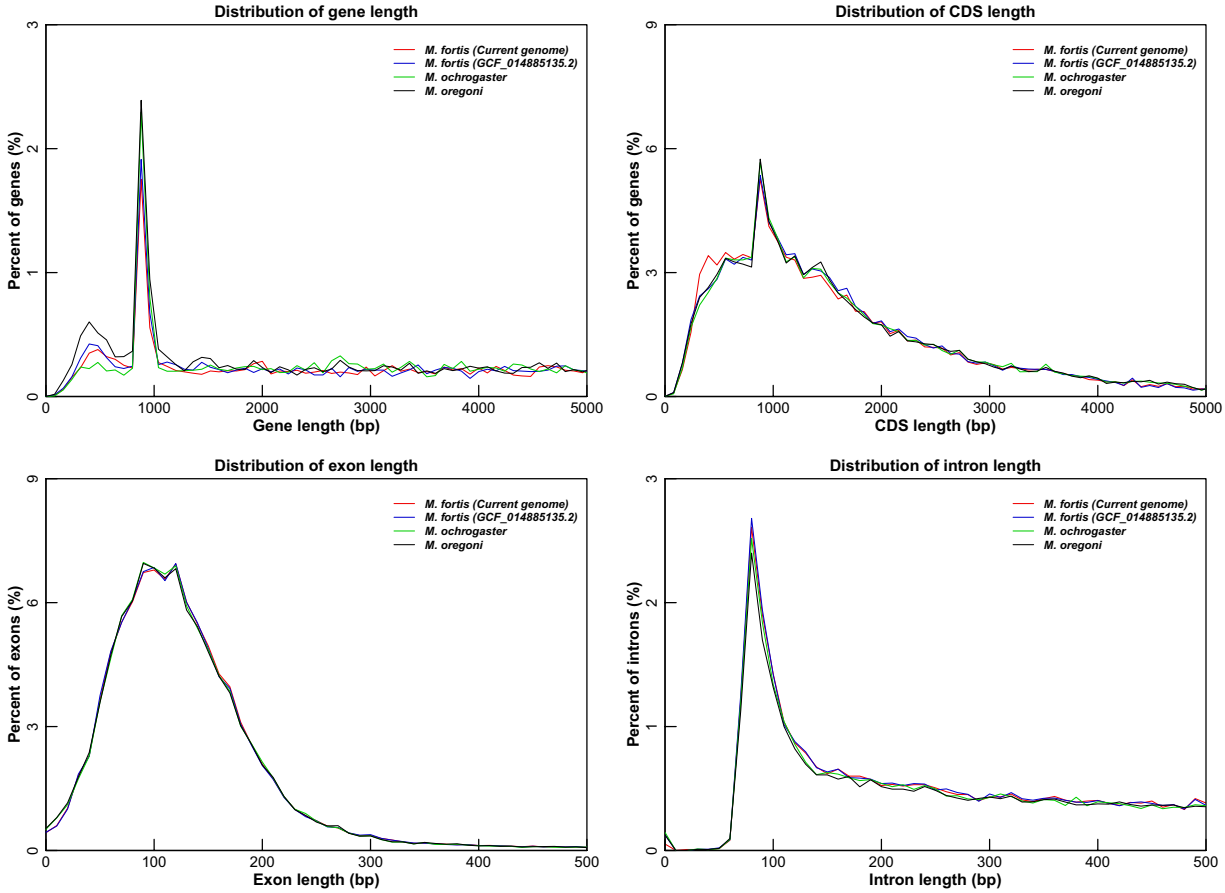

Fig. 3 Comparative analysis of gene element length distributions. The plots compare the length distributions of key gene structure elements between the current *M. fortis* genome (black), a previous *M. fortis* draft assembly (GCF_014885135.2, blue), and two other closely related species, *M. ochrogaster* (GCF_000317375.1, green) and *M. oregoni* (GCF_018167655.1, red). The four panels show distributions for: (top left) overall gene length, (top right) coding sequence (CDS) length, (bottom left) exon length, and (bottom right) intron length. The similar patterns observed across these species, particularly for CDS and exon lengths, support the high quality and accuracy of the gene prediction in the current assembly.

Transcriptome-based prediction. RNA-seq reads from multiple tissues were mapped to the genome using TopHat2 v2.1.1[28] and assembled into transcripts with StringTie v2.4.0[29]. Likely protein-coding regions within these transcripts were subsequently identified with TransDecoder v5.7.0.

Finally, the gene predictions from these three approaches were merged into a final, non-redundant consensus gene set using EVidenceModeler (EVM) v2.1.0[30], which weighs the different sources of evidence to generate the most reliable gene models. This comprehensive pipeline resulted in the identification of a total of 23,678 PCGs.

For functional annotation, the amino acid sequences of these 23,678 predicted genes were aligned against multiple public databases using BLASTP (E-value < 1e-5), including the NCBI non-redundant protein (Nr) and Uniprot (Swiss-Prot/TrEMBL) databases[31]. Further functional information was derived by searching for conserved protein domains and families using InterProScan[32], which scans against databases like Pfam. This process also assigned Gene Ontology (GO) terms and associated genes with biological pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and eukaryotic Orthologous Groups (KOG) databases via eggNOG-mapper[33]. This multi-faceted approach successfully assigned putative functions to 23,088 genes (97.5% of the total predicted set).

Repeat annotation. The identification and masking of repetitive elements is a critical step before gene annotation to prevent spurious predictions. We employed a comprehensive strategy combining both homology-based and *de novo* approaches. For homology-based prediction, which identifies known repeats, the genome was screened using RepeatMasker v.open-4.0.9[34] and RepeatProteinMask against the RepBase library[35], a curated database of known transposable elements. For *de novo* prediction, which discovers novel repeat families specific to the *M. fortis* genome, we first used RepeatModeler v1.0.4[36] to construct a species-specific repeat library. Tandem repeats, which are simple, consecutively repeated sequences, were identified separately using Tandem Repeats Finder v4.07[37]. The results from these different approaches were then integrated to create a final, comprehensive repeat annotation. This analysis revealed that repetitive sequences constitute 41.87% of the *M. fortis*

Type	Length (bp)	% in genome
DNA	64,099,798	2.80
LINE	346,658,019	15.16
SINE	206,978,069	9.05
LTR	392,309,727	17.16
LTR-Gypsy	16,635,454	0.73
LTR-Copia	3,659,796	0.16
Other	5,224	0.00
Unknown	9,084,763	0.40
Total	957,337,009	41.87

Table 5. Repetitive element annotations in the *M. fortis* genome.

Type		Copy	Average length(bp)	Total length(bp)	% of genome
miRNA		566	83	46841	0.002
tRNA		34634	70	2416927	0.1057
rRNA	rRNA	1398	174	243511	0.0106
	18S	10	1868	18680	0.0008
	28S	10	6539	65393	0.0029
	5.8S	737	114	84012	0.0037
	5S	641	118	75426	0.0033
snRNA	snRNA	1964	109	214218	0.0094
	CD-box	775	87	67789	0.003
	HACA-box	303	134	40605	0.0018
	splicing	834	118	98140	0.0043
	scaRNA	46	159	7321	0.0003

Table 6. Summary of non-coding RNA annotation in the *M. fortis* genome.

genome (Table 5). The most abundant categories of transposable elements were Long Terminal Repeats (LTRs) at 17.16%, followed by Long Interspersed Nuclear Elements (LINEs) at 15.16%, Short Interspersed Nuclear Elements (SINEs) at 9.05%, and DNA transposons at 2.8%.

Non-coding RNA annotation. A comprehensive annotation of non-coding RNAs (ncRNAs) was performed to identify key functional RNA molecules within the genome. Different classes of ncRNAs were identified using specialized bioinformatics tools. Transfer RNAs (tRNAs), which are essential for translating mRNA into protein, were predicted using tRNAscan-SE v2.0[38], which searches for the characteristic cloverleaf secondary structure of tRNA genes. Other major classes of ncRNAs, including ribosomal RNAs (rRNAs), microRNAs (miRNAs), and small nuclear RNAs (snRNAs), were identified by searching the genome against the Rfam database (v14.7)[39] using the Infernal v1.1.2 software[40]. Infernal employs covariance models that account for both sequence and secondary structure conservation, allowing for highly sensitive and specific identification of structured ncRNA families. The combined results from these analyses identified a total of 1,398 rRNA genes, 566 miRNA genes, and 1,964 snRNA genes (Table 6).

Mitochondrial genome assembly and annotation. The complete mitochondrial genome was assembled de novo using the MitoZ (v1.3) software[41], which utilized both the DNBSEQ short-read data and the PacBio HiFi long-read data. The resulting circular mitogenome was annotated using both MitoZ and the MitoS2 server[42], selecting the appropriate reference database based on the species. To ensure high accuracy, all protein-coding genes in the final annotation were manually checked and curated by comparing them to the mitogenome of a closely related species, *Microtus obscurus* (NC_087845.1). The final annotated mitogenome map was visualized using the OGDRAW software[43]. We successfully assembled and annotated the complete, circular mitochondrial genome using the MitoZ pipeline. The final mitogenome is 16,367 bp in length and contains the full set of 37 typical vertebrate mitochondrial genes (Supplementary Fig. 5).

Data Records

All sequencing data generated for this study have been deposited in the NCBI database under BioProject accession number PRJNA1271721. The raw sequencing reads are available in the Sequence Read Archive (SRA) under study accession SRP589748[44]. Specifically, DNBSEQ genomic sequencing data are deposited under accession SRR33821528[45], and PacBio HiFi genomic sequencing data are available under accessions SRR33821526[46] and SRR33821527[47]. Additionally, Hi-C interaction data and transcriptomic (RNA-seq) data have been deposited under accessions SRR33821525[48] and SRR33821524[49], respectively. The complete mitochondrial

genome assembly is available in GenBank in.fasta format under accession number PX549189.1[50]. The final chromosome-level genome assembly has been deposited in GenBank under accession JBQVRV000000000.1[51] and in the Genome Warehouse (GWH) at the National Genomics Data Center (NGDC) under accession GWHESEF00000000[52]. The GWH entry includes the genome assembly, genome annotation, coding sequences, and protein sequences available for download.

Technical Validation

DNA quantification and qualification. The quality of the starting biological material was rigorously assessed prior to library construction to establish a foundation of high-quality data. High-molecular-weight genomic DNA, essential for long-read sequencing, was extracted from muscle tissue. Its integrity was confirmed by 1% agarose gel electrophoresis, which revealed a distinct, high-molecular-weight band with no visible smearing, indicating that the DNA was not degraded and was suitable for generating long PacBio reads. The concentration of the DNA was precisely quantified using a Qubit Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA), which employs a dsDNA-specific dye for accuracy, ensuring optimal input for library preparation. Purity was assessed with a NanoDrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA); an A260/280 ratio of approximately 1.8 confirmed that the sample was largely free of protein contamination, which can inhibit downstream enzymatic reactions.

Quality control of raw sequencing data. The raw sequencing data from all platforms underwent a stringent quality control process to remove low-quality reads and artifacts. For the DNBSEQ short-read data, raw reads were processed to remove adapter sequences, reads with a high proportion of N bases (>5%), and low-quality reads where more than 50% of the bases had a phred quality score below 20. This ensured that only high-quality reads were used for the downstream genome survey. For the PacBio sequencing, the raw subread data was processed on-instrument to generate highly accurate HiFi reads (>99.9% accuracy) through circular consensus sequencing (CCS), which inherently filters out low-quality reads. For Hi-C data, raw reads were first processed to remove adapter sequences and low-quality reads. The resulting clean data was then aligned to the reference genome and filtered for valid interaction pairs. The filtering steps included removing unmapped reads, invalid pairs (such as self-circles and dangling ends), and PCR duplicates. Only the valid, unique read pairs were used for the subsequent scaffolding process.

RNA quality evaluation. For the transcriptome analysis, the quality and integrity of the total RNA extracted from skeletal tissues were paramount. RNA quality was evaluated on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The resulting electropherograms for all samples used in library preparation yielded an RNA Integrity Number (RIN) greater than 8.0. This high RIN value, derived from the ratio of the 28S to 18S ribosomal RNA subunits, confirms that the RNA was intact and had not undergone significant degradation. This is critical for ensuring that the RNA-seq data accurately represents the full-length transcript repertoire without 3' bias, thus providing reliable evidence for gene annotation.

Evaluation of the assembled genome. The final genome assembly demonstrates exceptional quality across evaluations of its continuity, completeness, and accuracy. The assembly exhibits exceptional continuity, with a contig N50 of 68.89 Mb and a scaffold N50 of 91.23 Mb. A total of 97.73% of the assembled sequence was successfully anchored into 26 chromosome-level pseudomolecules, which is consistent with the known karyotype of *M. fortis*. The completeness was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.4.6[23] against the glires_odb10 lineage dataset, yielding a complete score of 96.3% (C:96.3% [S:93.7%, D:2.6%], F:0.6%, M:3.1%, n:13798). The assembly accuracy was further validated using Merqury, which estimated a consensus quality (QV) score of 59.76 and a base error rate of 1.056×10^{-6}. The k-mer copy number and assembly spectra (Supplementary Figs. 1, 2) showed a clean distribution with minimal artificial duplications, confirming the high fidelity of the chromosome-level assembly. The accuracy of the assembly was also confirmed by a high read mapping rate of 99.98% with Inspector[53] and by the clean intra-chromosomal interaction patterns shown in the Hi-C contact map (Fig. 2). The quality of the gene annotation was also assessed by comparing the length distributions of gene elements to those of related species, which showed similar patterns and supported the accuracy of the gene prediction pipeline (Fig. 3). Finally, a contamination screen using BlobToolKit[54] confirmed the high purity of the assembly, showing that the vast majority of contigs belonged to a single, high-coverage cluster assigned to Chordata, with no evidence of significant contamination (Supplementary Fig. 4).

Data availability

The raw sequencing data (including Illumina, PacBio HiFi, Hi-C, and transcriptomic reads) generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) under study accession number SRP589748[44]. The complete mitochondrial genome is available in GenBank under accession number PX549189.1[50]. The final chromosome-level genome assembly and annotation have been deposited in GenBank under accession number JBQVRV000000000.1[51] and in the Genome Warehouse (GWH) at the National Genomics Data Center (NGDC) under accession number GWHESEF00000000[52]. All data are associated with BioProject accession number PRJNA1271721.

Code availability

No custom code was generated for this study. All analyses were performed using publicly available bioinformatics tools as described in the methods section.

Received: 11 June 2025; Accepted: 3 February 2026;
Published online: 14 February 2026

References

1. Wang, S. *et al*. The feeding preference and bite response between Microtus fortis and Broussonetia papyrifera. *Frontiers in Plant Science* **15**, 1361311 (2024).
2. Okada, K. & Kageyama, A. Assisted reproductive technologies in Microtus genus. *Reproductive Medicine and Biology* **18**, 121–127 (2019).
3. Hu, Q. *et al*. De novo assembly and transcriptome characterization: Novel insights into the mechanisms of primary ovarian cancer in Microtus fortis. *Molecular Medicine Reports* **25**, 64 (2022).
4. Ueoka, I., Pham, H. T. N., Matsumoto, K. & Yamaguchi, M. Autism spectrum disorder-related syndromes: modeling with Drosophila and rodents. *International Journal of Molecular Sciences* **20**, 4071 (2019).
5. Zhu, L., Qi, Z., Wen, Y. C., Min, J. Z. & Song, Q. K. The complete mitochondrial genome of Microtus fortis pelliceus (Arvicolinae, Rodentia) from China and its phylogenetic analysis. *Mitochondrial DNA Part B* **4**, 2039–2041 (2019).
6. He, T. *et al*. Metabolomic analysis of the intrinsic resistance mechanisms of Microtus fortis against Schistosoma japonicum infection. *Scientific Reports* **15**, 7147 (2025).
7. Shen, J. *et al*. Macrophage-mediated trogocytosis contributes to destroying human schistosomes in a non-susceptible rodent host, Microtus fortis. *Cell Discovery* **9**, 101 (2023).
8. Xiong, D. *et al*. Transcriptional profiling of Microtus fortis responses to S. japonicum: New sight into Mf-Hsp90 α resistance mechanism. *Parasite Immunology* **43**, e12842 (2021).
9. Hu, Y. *et al*. De novo assembly and transcriptome characterization: novel insights into the natural resistance mechanisms of Microtus fortis against Schistosoma japonicum. *BMC Genomics* **15**, 1–13 (2014).
10. Li, H. *et al*. Genome assembly and transcriptome analysis provide insights into the antischistosome mechanism of Microtus fortis. *Journal of Genetics and Genomics* **47**, 743–755 (2020).
11. Rhie, A. *et al*. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
12. Wenger, A. M. *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* **37**(10), 1155–1162 (2019).
13. Burton, J. N. *et al*. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology* **31**(12), 1119–1125 (2013).
14. Rao, S. S. P. *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
15. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
16. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**, 1432 (2020).
17. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175 (2021).
18. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
19. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3**, 95–98 (2016).
20. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
21. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems* **3**, 99–101 (2016).
22. Pan, Y. Q. *et al*. Analysis of chromosome number and chromosome bands of Microtus fortis from different regions in China. *Chinese Journal of Laboratory Animal Science* **12**(3), 147–150 (2002).
23. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* **38**, 4647–4654 (2021).
24. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 1–11 (2005).
25. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–W439 (2006).
26. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**(1), 78–94 (1997).
27. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 1–14 (2011).
28. Kim, D. *et al*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**, 1–13 (2013).
29. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**, 290–295 (2015).
30. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, 1–22 (2008).
31. UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47**, D506–D515 (2019).
32. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
33. Huerta-Cepas, J. *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution* **34**, 2115–2122 (2017).
34. Bergman, C. M. & Quesneville, H. Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics* **8**, 382–392 (2007).
35. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 1–6 (2015).
36. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457 (2020).
37. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
38. Lowe, T. M. & Chan, P. P. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research* **44**(W1), W54–W57 (2016).
39. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* **33**, D121–D124 (2005).
40. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
41. Meng, G., Li, Y., Yang, C. & Liu, S. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic acids research* **47**(11), e63 (2019).

42. Bernt, M. *et al.* MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, 69(2) (2013).
43. Stephan, G., Pascal, L. & Ralph, B. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Research* **47**(W1), W59–W64 (2019).
44. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP589748 (2025).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra/SRR33821528 (2025).
46. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra/SRR33821526 (2025).
47. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra/SRR33821527 (2025).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra/SRR33821525 (2025).
49. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra/SRR33821524 (2025).
50. *NCBI GenBank* https://identifiers.org/ncbi/insdc:PX549189.1 (2025).
51. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBQVRV000000000.1 (2025).
52. *CSDC Genome Warehouse* https://ngdc.cncb.ac.cn/gwh/Assembly/84475/show (2025).
53. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biology* **22**, 1–21 (2021).
54. Challis, R. *et al.* BlobToolKit-interactive quality assessment of genome assemblies[J]. *G3: Genes, Genomes, Genetics* **10**(4), 1361–1374 (2020).

Acknowledgements

This work was supported by the Changsha Major Special Project of Science and Technology (Grant No. kh2301027), the Natural Science Foundation of Hunan Province (Grant Nos. 2024JJ5422 and 2024JJ6494), and the Key Research and Development Program of Hunan Province (Grant No. 2024DK2001). The authors are grateful to Benagen Technology (Wuhan, China) for their valuable advice on bioinformatic analysis. We also sincerely thank the journal editors and the reviewers for their constructive suggestions and insightful comments, which significantly contributed to the improvement of this manuscript.

Author contributions

Z.J.Z. and Q.H. conceived and designed the study; T.Q.H. and J.K.Z. conducted the collection of the reed vole samples; J.K.Z., Y.X.W., Q.L., J.Z., W.L.Z., L.X.O., S.S.G., R.T.G. contributed to experimental design and data collection. D.Z. analyzed the data and wrote the draft manuscript; D.Z., T.Q.H., Q.H. and Z.J.Z. discussed the results and improved and revised the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare that they have no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-06813-3.

Correspondence and requests for materials should be addressed to Z.Z.