scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the short-faced mole (*Scaptochirus moschatus*)

Lei Chen[1]✉, Zenghao Gao[1,2], Chao Xue[1,2], Yue Zhao[1], Di Xu[1], Xiaohan Ma[1] & Yifan Zhang[1]

The short-faced mole (*Scaptochirus moschatus*) belonging to the family Talpidae in the order Eulipotyphla is a good model for studying adaptive evolution of mammals because of its morphological and ecological characteristics. However, the lack of genome of short-faced mole has hindered previous studies. In this study, we assembled the genome of the short-faced mole based on Illumina, PacBio HiFi and Hi-C sequencing, and acquired the genome of the short-faced mole with the size of 2.17 Gb. 99.6% of the assembled genome were identified as complete BUSCOs, including 90.7% as complete single-copy BUSCOs and 8.9% as complete duplicated BUSCOs. The assembled genome was anchored to 24 chromosomes with an anchor rate of 94.33%, of which the 24th chromosome (Chr 24) probably contained the X and Y chromosomes. A total of 21,139 coding genes were predicted, and 8.58 exons per gene were predicted.

Background & Summary

The short-faced mole (*Scaptochirus moschatus*) belonging to the genus *Scaptochirus* of the family Talpidae is endemic to China, mainly distributed in Hebei, Shandong, Inner Mongolia, Jilin, Shanxi, Gansu, etc[1]. It is similar to the Ussuri Mole (*Mogera robusta*), but it has a smaller body size. The short-faced moles are stout, with a short and sharp mouth. Their small eyes hidden in the fur. Their claws are flat, strong and sharp. Their whole body are covered with brown and metallic fine fur. The base of the fur is dark gray, and the top of the fur is dyed brown. The short-faced moles live underground for all the life. Their hearing and smell are very sensitive. They crawl fast in the ground and rarely climb out. They like to live in sandy areas with dry and loose soil and deep soil layer. The extensive ecotype and morphological features make the short-faced mole an interesting model for studying adaptive evolution[2].

Previous studies on moles have focused on morphology[2,3], taxonomy[4], karyology[5], phylogeny[6] and gut microbiology[1]. The follow-up evolutionary biological and molecular ecological studies on the short-faced mole is constrained by the limitation of the reference genome. In this study, multiple sequencing technologies, including short-read sequencing (Illumina), long-read PacBio high-fidelity sequencing (PacBio), and proximity ligation-based chromatin capture sequencing (Hi-C) were integrated to construct a high-quality genome assembly of the short-faced mole. The Illumina data was used to survey the genome yielded a genome size of 2.17 Gb. A total of 2.25 Gb genome contigs was assembled from the HiFi data, with a contig N50 of 67.66 Mb (Table 1). The genome was annotated to the chromosome level by assisted assembly using Hi-C data combined with HiFi data. The total length of genome scaffolds is 2.25 Gb, with a N50 of 110.51 Mb (Table 1). The assembled genome was successfully anchored to 24 chromosomes with the anchor rate of 94.33%. 0.88 Gb of repetitive sequences were identified, representing for about 39.41% of the genome. 21,139 genes were predicted of which 94.6% were functionally annotated, with an average gene length of 30.01Kb. 8.58 exons per gene were predicted. The average lengths of the exons and introns are 0.17Kb and 3.76Kb, respectively. The annotated gene set were evaluated by using the single-copy direct orthologous gene library, and they yield 94.1% complete BUSCOs.

[1]College of Life Sciences, Qufu Normal University, Qufu, 273165, Shandong Province, P.R. China. [2]These authors contributed equally: Zenghao Gao, Chao Xue. ✉e-mail: leisurechen@163.com

Sample ID	Contig length	Scaffold length	Contig number	Scaffold number
Total	2,252,979,606	2,252,983,806	297	255
Max	153,511,026	156,985,200	—	—
N50	67,659,624	110,510,357	13	9
N90	9,209,045	46,613,800	37	22

Table 1. Statistics of the assembled genome of *S. moschatus*.

Methods

Sample collection and ethics statement. A male short-faced mole was sampled from Liaocheng city, Shandong province, China, in May 13th, 2024. It was euthanized with ether. Samples of heart, liver, spleen, lungs, kidneys, muscles and blood were collected and subsequently stored at Qufu Normal University and kept at −80 °C before DNA and RNA extraction. All experiments were conducted in accordance with the Guidelines for the Care and Use of Laboratory Animals in China, and approved by the Biomedical Ethics Committee of Qufu Normal University with the registration number 2024118.

Genome sequencing. Blood samples were taken from the short-faced mole for Illumina and PacBio sequencing. Muscle tissues were used for Hi-C sequencing. Heart, liver, lung, spleen, kidney and muscle tissues were taken from the same mole for transcriptome sequencing. All sequencing analyses were performed by the Novogene Co., Ltd (Beijing, China). The QIAGEN AllPrep DNA/RNA Mini Kit (Qiagen, Germany) was used to extract Genomic DNA and total RNA. The quality of the DNA and RNA was checked by the NanoDrop 2000c spectrophotometer, Qubit 3.0 (Invitgen, USA) and the 2100 bioanalyzer (Agilent, USA).

Illumina sequencing library was generated using the Illumina PE Cluster Kit (Illumina, USA) according to the manufacturer's instructions. DNA libraries were sequenced on Illumina Novaseq platform and 150 bp paired-end reads were generated. A total of 103.14 Gb short reads were derived from Illumina sequencing. Muscle samples were used for extracting genome DNA for Hi-C sequencing after cross-linked by 4% formaldehyde solution and marked with biotin-14-dCTP. The obtained genome DNA was randomly interrupted into fragments by Covaris crusher and the Hi-C libraries were constructed according the standard protocol described previously[7]. Total RNA extracted from multi-tissues was used to construct cDNA libraries by TruSeq RNA Sample Prep Kit v2 (Illumina, USA). Both the libraries for Hi-C and transcriptome sequencing were sequenced on Illumina Novaseq platform by using a paired-end strategy. A total of 202.85 Gb data were derived from Hi-C sequencing, and each sample obtained more than 6 G data from transcriptome sequencing. Fastp v0.20 was used to clean the raw data with default parameters[8].

High quality DNA extracted from blood samples (primary band >30 kb) that passed the assay were selected to be randomly sheared into fragments (15–18 kb). Large fragments of DNA were enriched and purified using magnetic beads. Damage repair and end repair were performed on the fragmented DNA. Stem-loop sequencing junctions were attached to the ends of the DNA fragments, and exonucleases were used to remove fragments that failed to connect. The constructed library was sequenced by PacBio Revio/Sequel II/IIe platform followed the standard protocol (Pacific Biosciences, CA, USA), and it produced a total of 98.9 Gb HiFi long reads.

Genome survey and assembly. To estimate the genome size, the clean Illumina sequencing reads were used for *k* -mer analysis. Jellyfish v2.2.6[9] was used to calculate the optimal *k*-mer and GenomeScope v2.012[10] was used to estimate the genome size for corresponding *k*-mers. Referring to the genome size of affiliate species in family Talpidae[11], we selected *k*-mer = 17 for genome survey. The corresponding genome size of the short-faced mole is 2.17 Gb (Table S1, Fig. S1). The heterozygosity rate (obtained by calculating the proportion of heterozygous sites) in the surveyed genome is 0.34%.

The long reads obtained from PacBio sequencing were broken from junctions, and subreads were obtained after filtering out the junction sequences. The mean length and N50 length of the subreads were 18,852 bp and 18,761 bp, respectively. The CCS (https://github.com/PacificBiosciences/css) was used to generate high-precision HiFi reads, with the parameter of min-rq = 0.99. The HiFi reads obtained was used for genome assembly by using Hifiasm 0.19.9 software to give contigs[12]. A total of 2.25 Gb genome contigs were produced with an N50 of 67.66 Mb (Table 1). The GC content of the assembled contig genome was 42.87%. The integrity of the assembled genome was BUSCO v5.4.5[13] evaluated based on the Single-Copy Orthologs library (metazoa_odb10) using the software such as Metaeuk and HMMER. From 954 total BUSCO groups searched, 99.6% of complete BUSCOs (including 90.7% of complete and single-copy BUSCOs, 8.9% of complete duplicated BUSCOs), 0.2% of fragmented BUSCOs, and 0.2% of missing BUSCOs were identified, which further emphasized the precision and completeness of gene prediction.

Chromosome level genome assembly. The contig genome obtained from the Hifiasm 0.19.9 was combined with Hi-C data for chromosome clustering, orientation and sorting using ALLHiC 0.9.8[7] (parameters: enz = Dpn II, CLUSTER = n) to obtain the near-chromosome level genome. Juicebox v1.11.08 was employed to undertake an examination, perform a manual curation of the identification according to the chromosome interaction intensity[14], and to build the final chromosome-level genome assembly. After the above assembly, we obtained a total of 2.25 Gb scaffolds, with the lengths of N50 of 110.51 Mb, respectively (Table 1). Reflecting the previous report about the 2 N = 48 karyotype of the short-faced mole[5], a total of 2.13 Gb of genome data were anchored to 24 chromosomes with an anchor rate of 94.33% (Fig. 1), of which the 24th chromosome (Chr 24) probably contained the X and Y chromosomes. The anterior segment (about 16 Mb) of Chr 24 may be the

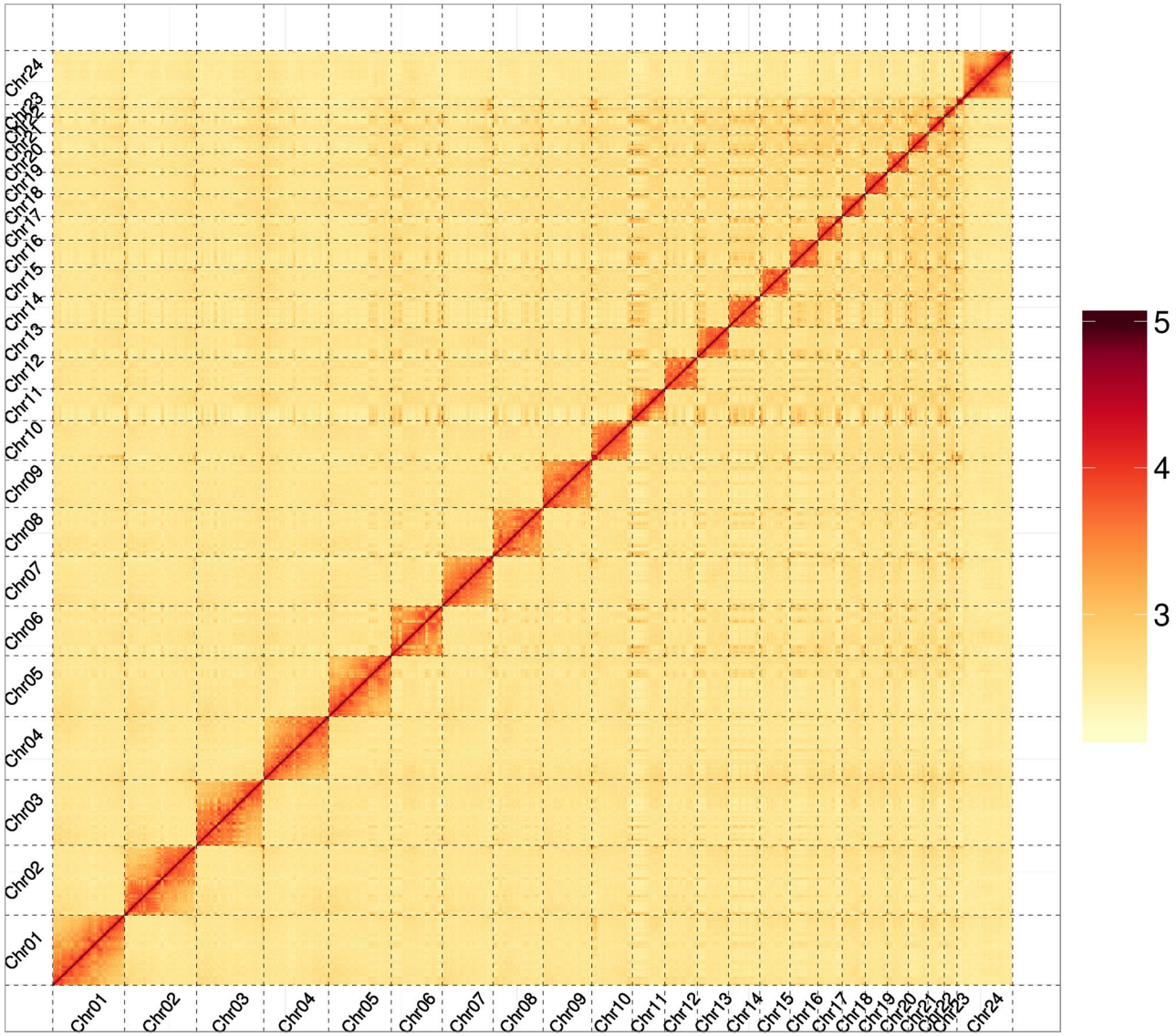

Fig. 1 Heat map of Hi-C linkage density of *Scaptochirus moschatus*. The x-axis and y-axis represent the genomic positions. Red dots indicate regions with a high density of paired reads, suggesting that they are more likely to be on the same chromosome.

Y chromosome and the posterior segment (about 103 Mb) may be the X chromosome (Fig. S2). However, since there is no other support for the annotation results of sex chromosomes of related species in the family Talpidae, this inference needs to be further verified. The length of the chromosomes ranged from 28.66 Mb to 156.99 Mb (Table S2 and Fig. 2).

Repeats and non-coding RNAs annotations. Following the completion of the genome assembly, the annotation was performed on three kinds of repetitive sequences, and on ncRNAs and PCGs. Repetitive elements were delineated by employing RepeatModeler v2.0.1[15] following its preset parameters, and a new repeat sequence library was assembled. A customized library was built in conjunction with Dfam 3.1[16] and Repbase databases[17]. Repeated elements in custom libraries were masked in homology prediction by using RepeatMasker v4.1.0[18]. The analysis yielded the detection of 0.88 GB of repetitive sequences within the short-faced mole genome, representing 39.41% of the total genome. The primary categories of TEs (tandem repeats) and IEs (interspersed repeats) identified are as follows: long interspersed nuclear elements (LINEs) at 26.45%, short interspersed nuclear elements (SINEs) at 0.44%, long terminal repeats (LTRs) at 8.19% and DNA transposon elements at 3.90% (Table 2).

The annotation of all non-coding RNAs (rRNAs, tRNAs, snRNAs, and miRNAs) was performed using Infernal v1.1.3[19] and tRNAscan-SE v2.0.7[20]. Non-coding RNA annotation is performed at the genome-wide level, and genome that are masked for repetitive sequences are used for subsequent gene structure annotation. In our annotation process, transcriptome data are used only for prediction of genes and their structures, and are not used for prediction of non-coding RNAs. ncRNA predictions are obtained from sequence comparisons of

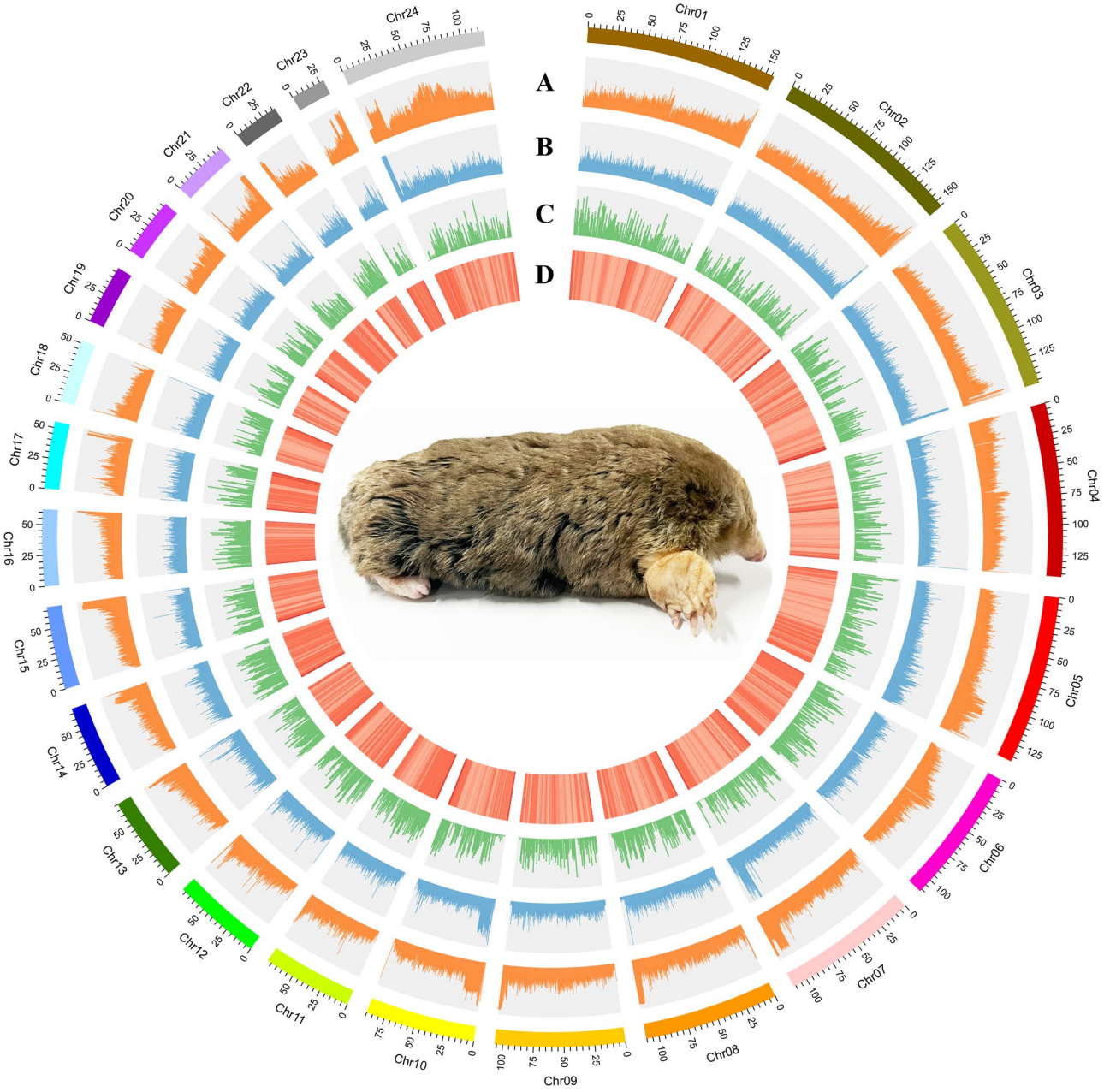

Fig. 2 Genome characteristics of *Scaptochirus moschatus*. From the outer ring to the inner ring are the distributions of RNA TEs, DNA TEs, gene density and GC content.

Type	Number	Length (bp)	Percentage (%)
SINE	79,325	9,896,501	0.44
LINE	1,495,956	595,998,526	26.45
LTR	674,358	184,486,499	8.19
DNA	614,716	87,876,629	3.9
Unknown	52,962	6,528,814	0.29
Total		887,811,039	39.41

Table 2. Statistics of repeat sequence classification results.

databases with the whole genome, and there are no predictions of overlap with protein-coding genes. A variety of non-coding RNAs were catalogued, resulting in the identification of 89,096 rRNAs, 5,958 transfer RNAs, 3,386 snRNAs, and 4,915 miRNAs (Table 3).

Type	Copy number	Average length(bp)	Total length(bp)	% of genome
miRNA	4,915	108.949	535,484	0.024
tRNA	5,958	80.355	478,753	0.021
rRNA	89,096	77.112	6,870,358	0.305
snRNA	3,386	126.211	427,351	0.019

Table 3. Annotations of non-coding RNAs in the *S. moschatus* genome.

Species	Number	Average gene length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
S. moschatus	21,139	30,008.97	1,490.70	8.58	173.79	3,763.59
R. norvegicus	22,313	32,787.21	1,532.42	8.85	173.15	3,981.46
C. cristata	17,608	36,553.98	1,686.02	9.88	170.73	3,928.53
T. occidentalis	21,483	35,128.25	1,636.61	9.30	175.93	4,033.98

Table 4. Basic statistical results on gene structure of affiliate species and a model species.

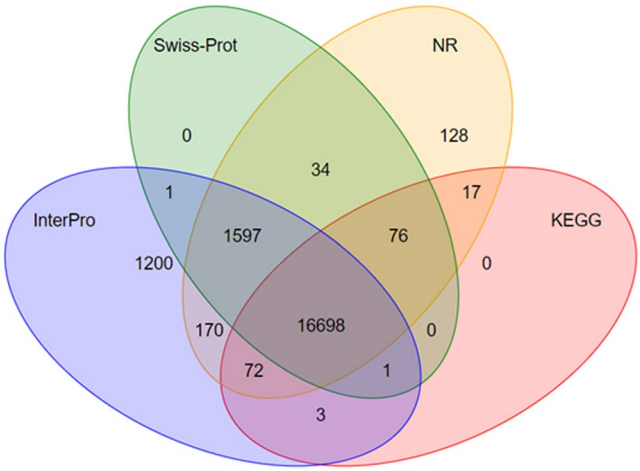

Fig. 3 Statistical results of gene function annotation of the *Scaptochirus moschatus* genome.

Gene structure and function predictions. The *De novo* annotation[21] was performed using Augustus v3.5 and Snap v2013.11.29. Homologous encoded protein prediction[22] was performed using Blastall v2.2.26, Solar v0.9.6 and GeneWise v2.4.1. The gene annotation results were integrated using EVM v1.1.1 with default parameters. The NCBI RefSeq database[23] was queried to obtain the protein sequences of two affiliate species (*Talpa occidentalis*, *Condylura cristata*) and a model species (*Rattus norvegicus*), and the gene structure was predicted by blat (http://genome.ucsc.edu/cgi-bin/hgBlat)[24]. The gene structure was annotated by using transcriptome data. Combining the above prediction results and the transcriptome alignment data, EVidenceModeler (http://evidencemodeler.sourceforge.net)[25] was used to integrate the gene sets predicted by various methods into a non-redundant, more complete gene set. Finally, PASA (http://pasa.sourceforge.net) was used to combine transcriptome assembly results to correct the EVM, and UTR and variable shear were added to obtain the final gene set. There were 21,139 protein coding genes in the genome of the short-faced mole, 94.6% of the genes were predicted to have a function, and the mean length of these genes was 30.01 Kb. The mean count of exons per gene stands at 8.58, while the typical exon spans 0.17 Kb, an average CDs extends to 1.49Kb and an average intron extends to 3.76 Kb (Table 4, Fig. S3).

The gene sets obtained from annotation of the gene structure were used for gene function prediction based on the NR (non-redundant protein records) (http://www.ncbi.nlm.nih.gov/protein), Swiss-Prot[26] (http://www.uniprot.org), Pfam (http://pfam.xfam.org), KEGG[27] (http://www.genome.jp/kegg), and InterPro (https://www.ebi.ac.uk/interpro) databases by using Interproscan v5.59-91.0[28], Blastp v2.2.26 and Diamond v0.8.22. 94.6% of the 21,139 annotated genes were predicted successfully, each identified as having at least one homologous gene, with the information corroborated across three public databases (Fig. 3).

Data Records

All sequencing data and genome assembly have been deposited in the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1127535), which include SRR31617600[29], SRR31478213[30], SRR32108084[31] in SRA and JBFCTN000000000[32] in GenBank. The genome annotations deposited in the Figshare database (https://doi.org/10.6084/m9.figshare.28430021)[33].

Technical Validation

BUSCO[13] was used to assess the integrity of the assembled genome yielded 99.6% of the genome were identified as complete BUSCOs (metazoa_odb10). BWA 0.7.8 software[34] was used to calculate the alignment of the fragment reads to the assembled genome to evaluate the completeness of the assembly and the uniformity of sequencing. It yielded a 99.68% alignment rate and a 99.98% genome coverage indicating good consistency between reads and assembled genome. The SNP calling of the alignment results performed using SAMtools 0.1.19[35] yielded a 0.001% heterozygous SNP ratio and a 0.000% homozygous SNP ratio, indicating that the assembly has a high single base accuracy (Table S3). The genomic sequence accuracy was evaluated using Merqury[36] yielded a quality value of 50.97 indicating that the sequence accuracy was more than 99.99%. The assembled genome was evaluated by multiple methods and showed good genomic consistency, completeness, and accuracy.

Code availability

No custom code was used for this study. All data analyses were conducted using published bioinformatics software with default settings, unless otherwise specified.

Received: 15 October 2024; Accepted: 20 February 2025;
Published online: 03 March 2025

References

1. Chen, L. *et al*. Habitat environmental factors influence intestinal microbial diversity of the short-faced moles (*Scaptochirus moschata*). *AMB Express*. **11** (2021).
2. Platonov, V. V. Peculiarities of the thoracic-lumbar vertebral morphology in common mole, Talpa europaea (Lipotyphla, Talpidae) in connection with its burrowing activity. *Russian Journal of Theriology*. **1**, 111–115 (2002).
3. Endo, H. *et al*. Skull Morphology and Mitochondrial DNA Sequence Analysis in the Lesser Japanese Mole (*Mogera imaizumii*) from the Imperial Palace (Tokyo, Japan). *The Journal of Veterinary Medical Science*. **61**, 1087–1091 (1999).
4. He, K., Shinohara, A., Jiang, X.-L. & Campbell, K. L. Multilocus phylogeny of talpine moles (Talpini, Talpidae, Eulipotyphla) and its implications for systematics. *Molecular Phylogenetics and Evolution*. **70**, 513–521 (2014).
5. Kawada, S. I., Harada, M., Koyasu, K. & Oda, S. I. Karyological note on the short-faced mole, *Scaptochirus moschatus* (Insectivora, Talpidae). *Mammal Study* **27**, 91–94 (2002).
6. He, K. *et al*. Talpid mole phylogeny unites shrew moles and illuminates overlooked cryptic species diversity. *Molecular Biology and Evolution*. (2016).
7. Belton, J.-M. *et al*. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*. **58**, 268–276 (2012).
8. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
9. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. **27**, 764–770 (2011).
10. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. **33**, 2202–2204 (2017).
11. Real, F. M. *et al*. The mole genome reveals regulatory rearrangements associated with adaptive intersexuality. *Science* **370**(6513), 208–214 (2020).
12. Cheng, H. *et al*. Haplotype-resolved assembly of diploid genomes without parental data. *Nature Biotechnology*. **40**, 1332–1335 (2022).
13. 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. *Mol. Biol. Evol*. **38**, 4647–4654 (2021).
14. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems*. **3**, 95–98 (2016).
15. 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).
16. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic Acids Research*. **44**, D81–D89 (2016).
17. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*. **6**, 11 (2015).
18. Smit, A. F., Hubley, R. & Green, P. Repeat Masker Open-4.0. (2015).
19. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. **29**, 2933–2935 (2013).
20. Lowe Todd, M. & Eddy Sean, R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*. **25** (1997).
21. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics*. **21**, i351–i358 (2005).
22. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Research*. **14**, 988–995 (2004).
23. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*. **110**, 462–467 (2005).
24. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Research*. **12**, 656–664 (2002).
25. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*. **9** (2008).
26. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*. **28**, 45–48 (2000).
27. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. **28**, 27–30 (2000).
28. Zdobnov, E. M. & Apweiler, R. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics Applications Note*. **17**, 847–848 (2001).
29. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31617600 (2024).
30. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31478213 (2024).
31. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32108084 (2025).
32. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_042919465.1 (2024).
33. Chen, L. Genome annotations of *Scaptochirus moschatus*. *Figshare* https://doi.org/10.6084/m9.figshare.28430021 (2025).
34. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*. **25**, 1754–1760 (2009).
35. Li, H. *et al*. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).
36. Rhie, A. *et al*. Merqury: reference-free quality and phasing assessment for genome assemblies. *Genome Biol*. **21**, 245 (2020).

Acknowledgements
This work was supported by the National Natural Science Foundation of China (No. 31400473 and No. 32370446).

Author contributions
C.L. conceived and managed the project. C.L. and G.Z.H. analyzed the data and wrote the paper. X.C., Z.Y. and X.D. modify the manuscript. M.X.H. and Z.Y.F. collected samples. All authors have read and agreed to the published version of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04691-9.

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

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.