scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of *Megachile lagopoda* (Linnaeus, 1761) (Hymenoptera: Megachilidae)

Dan Zhang[1,2], Jianfeng Jin[3], Zeqing Niu[2] ✉, Michael C. Orr[4], Feng Zhang[3], Rafael R. Ferrari[2,5], Qingtao Wu[2], Qingsong Zhou[2], Wa Da[6], Arong Luo[2,7] ✉ & Chaodong Zhu[2,8,9]

Megachile **is one of the largest bee genera, including nearly 1,500 species, but very few chromosome-level assemblies exist for this group or the family Megachilidae. Here, we report the chromosome-level genome assembly of *Megachile lagopoda* collected from Xizang, China. Using PacBio CLR long reads and Hi-C data, we assembled a genome of 256.83 Mb with 96.08% of the assembly located on 16 chromosomes. Our assembly contains 266 scaffolds, with a scaffold N50 length of 15.6 Mb, and BUSCO completeness of 99.20%. We masked 27.10% (69.61 Mb) of the assembly as repetitive elements, identified 459 non-coding RNAs, and predicted 11,157 protein-coding genes. This high-quality genome of *M. lagopoda* represents an important step forward for our knowledge of megachilid genomics and bee evolution overall.**

Background & Summary

Bees are the most important group of angiosperm-pollinating insects, comprising over 20,000 described species worldwide[1]. Bees exhibit a range of lifestyles and social behaviors, ranging from solitary to highly eusocial honey bees[2,3]. Their diversity of forms is deepened by suites of varied life history strategies, including what flowers they visit and how they construct their nests. The varied ways in which bees live make them exemplary foci for studies in ecology and evolution, but relatively little is presently known of bee genomics outside of highly social groups[4].

The subfamily Megachilinae is the largest within the family Megachilidae, containing excess of 4,000 described species[1,2]. This subfamily exhibits extraordinarily variable life histories among cavity-nesting bees, including mason bees using mud, resin bees using plant products, leaf-cutter bees using leaves, wool-carder bees using plant trichomes, and many different types of brood parasites[2,5–7]. Within this subfamily, *Megachile* is the most species-rich genus, with more than 1,400 known species[1]; some species are already managed for crop pollination, such as *Megachile rotundata*, making studies on them useful from both evolutionary and economic perspectives[2,8–11].

Despite being important and interesting bees, high-quality chromosome-level genome resources remain deficient for this group. To date, only four chromosome-level genomes of *Megachile* have been published (accessed on March 2024 from the NCBI database). Many assemblies are needed to perform substantial comparative genomic analyses, and the current lack of high-quality chromosome-level genome resources limits our comprehension of the behavior evolution and ecology of this group and bees in general.

[1]Characteristic Laboratory of Forensic Science in Universities of Shandong Province, Shandong University of Political Science and Law, Jinan, P. R. China. [2]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, P. R. China. [3]Department of Entomology, College of Plant Protection, Nanjing Agricultural University, Nanjing, P. R. China. [4]Entomologie, Staatliches Museum für Naturkunde Stuttgart, Stuttgart, Germany. [5]Environmental Science Training Center, Federal University of Southern Bahia, Porto Seguro, Brazil. [6]Tibet Plateau Institute of Biology, Tibet, P. R. China. [7]International College, University of Chinese Academy of Sciences, Beijing, P. R. China. [8]College of Biological Sciences, University of Chinese Academy of Sciences, Beijing, P. R. China. [9]State Key Laboratory of Integrated Pest Management, Institute of Zoology, Chinese Academy of Sciences, Beijing, P. R. China. ✉e-mail: niuzq@ioz.ac.cn; luoar@ioz.ac.cn

Libraries	Insert sizes (bp)	Clean data (Gb)	Sequencing coverage (X)
Illumina	350	26.92	104.77
PacBio	30 Kb	32.21	125.36
Hi-C	350	31.43	122.33
RNA	350	6.50	—

Table 1. Statistics of the sequencing data used for genome assembly.

Herein, we used PacBio long reads, Hi-C, and Illumina sequencing to achieve the first high-quality chromosome-level reference genome of *Megachile lagopoda* (Linnaeus, 1761). Our assembly resulted in a genome size of 256.83 Mb located across 16 chromosomes, with scaffold N50 lengths of 15.6 Mb. We annotated 11,157 protein-coding genes (PCGs), 459 non-coding RNAs (ncRNAs), and 69.61 Mb repeat elements. This high-quality chromosome-level genome of *M. lagopoda* provides significant new data resources for Megachilinae, while facilitating further comparative studies on the nesting and pollination biology of bees.

Methods

Sampling and sequencing. Specimens of *Megachile lagopoda* were collected from Zhaburang Village, Zanda County, Ali, Xizang, China (31.4676 N, 79.6709E, 3620 m) on June 23, 2020 by D.Z. and Q.W. All samples were deposited in liquid nitrogen first, and then stored at −80°C before DNA extraction. All species were identified by Z.N., and voucher specimens are deposited at the Institute of Zoology, Chinese Academy of Science (2020QZKK05010605-04446 to 2020QZKK05010605-04460). Identifications were later confirmed via the mitochondrial cytochrome oxidase subunit (COI) gene of the sample[12], this sequence has been deposited in GenBank under accession number PP652334, and then compared to a reliable GenBank resource (sequence ID: HM401107.1). The BLAST match score was higher than 98.00%, suggesting that they are highly likely to be conspecific according to currently accepted species concepts for this group. To reduce potential contamination from microbes, the metasoma of all samples was removed before sending them to the sequencing company (Berry Genomics, Beijing, China). One male sample each was used for PacBio, Illumina whole genome (polishing), Illumina transcriptome, and Hi-C sequencing, respectively.

Qiagen Blood and Cell Culture DNA Mini Kits were used to extract the genomic DNA for PacBio and Illumina whole genome sequencing. PacBio sequencing was conducted on the PacBio Sequel II platform, and sequencing libraries with 30 kb insert size were constructed. Truseq Nano DAN HT sample preparation Kit (Illumina USA) was used to generate sequencing libraries of 150 bp paired-end reads with an insert of 350 bp on the Novaseq6000 platform. RNA sequencing was extracted with TRIzol, and a TruSeq RNA v2 kit was used to generate the RNA library. Hi-C sequencing was conducted using the BGI MGISEQ-2000 platform with 150 bp paired-end reads. Finally, we generated 97.06 Gb clean reads in total, containing 26.96 Gb (~104.77X) Illumina reads, 6.5 Gb transcriptome data, 32.21 Gb (~125.36X) PacBio reads, and 31.43 Gb (~122.33X) Hi-C reads (Table 1).

Genome assembly. BBTools v38.82[13] was used to filter out duplicate and low-quality reads, specifically with the 'clumpify.sh' and 'bbduk.sh' scripts. K-mer analysis and k-distribution were performed via 'khist.sh', and then Genomescope v2.0[14] was used to estimate the heterozygosity, size, and repetitive elements of the *M. lagopoda* assembly. The genome survey suggested that *M. lagopoda* had a genome size of 294.51 Mb with low heterozygosity (0.71%) and repetitiveness (21.10%) based on the 21-kmer frequency distribution of short reads (Fig. 1).

Flye v2.9[15] was used to assemble PacBio long reads using minimum overlap between reads of 3,000 bp and one round of self-polishing ("-m 3,000 -i 1"). The primary assembly was polished with two rounds of short reads using NextPolish v1.3.0[16]. Minimap2 v2.23[17] was used to align the PacBio assembly and Illumina sequences. Hi-C reads were aligned and quality control was conducted on the assembly via Juicer v1.6.2[18]. 3D-DNA[19] was used to anchor the primary contigs into chromosomes. Juicebox v1.11.08[17] was then used to correct possible assembly errors through manual inspection and refinement. MMseqs. 2 v11[20] was used to detect possible contaminants with the "-min-seq-id" parameter set to 0.8 when compared with the UniVec and NCBI nucleotide databases. Sequences over 90% alignments were removed. BUSCO v5.4.4[21] was performed to assess genome completeness using the insecta_odb10 database (n = 1,367) as the reference database.

After polishing, contaminant removal, and redundancy checks, the final *M. lagopoda* genome had a length of 256.83 Mb with 266 scaffolds, an N50 length of 15.60 Mb, and a GC content of 36.70% (Tables 2, 3). A total of 16 pseudo-chromosomes occupying 96.08% of the genome were assembled for *M. lagopoda* (Fig. 2), including 1,355 single-copy orthologs (99.10%), 2 duplicated BUSCOs (0.10%), and 10 missing BUSCOs (0.80%) (Table 2). Very few duplicated and missing BUSCOs, high mapping ratios of raw sequencing data, and superior assembly statistics all indicate that our assemblies are remarkably contiguous and comprehensive.

Genome annotation. RepeatModeler v2.0.2[22] was applied to generate a *de novo* repeat database with the LTR discovery pipeline (-LTRstruct), via specific repeat structures, and was then combined with the Dfam 3.5[23] and RepBase-20181026 databases[24] subsequently to establish a custom library. RepeatMasker v4.1.2[25] was used to identify repetitive elements against our custom repeat library. tRNAscan-SE v2.0.8[26] and Infernal v1.1.3[27] were performed to identify ncRNAs. TBtools[28] was used to produce the visual diagram of *M. lagopoda* genomic characteristics (LTR, LINE, SINE, DNA, GENE, GC, and Chr) (Fig. 3). We masked 27.10% (69.61 Mb) repetitive regions of the *M. lagopoda* genome. Specifically, 0.04% of repeat sequences were short interspersed elements (SINEs),

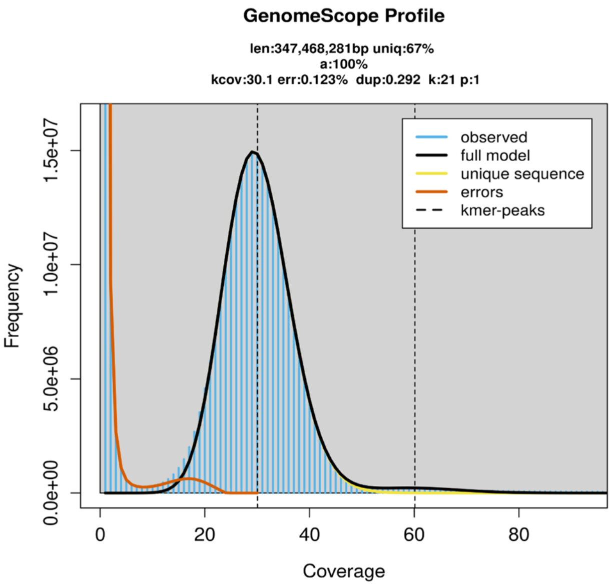

Fig. 1 GenomeScope genome size estimates for *Megachile lagopoda*.

Assembly	Total length (Mb)	Number scaffolds/ contigs (chromosomes)	Scaffold/contig N50 length (Mb)	GC (%)	BUSCO (n = 1,013) (%)			
					C	D	F	M
Flye	263.10	3,443/3,457	13.63/12.64	36.59	99.1	0.1	0.0	0.9
NextPolish	263.06	3,443/3,456	13.63/12.64	36.59	99.2	0.1	0.0	0.8
3D-DNA	256.83	694/2,240 (16)	15.60/8.91	36.70	99.2	0.1	0.0	0.8
Final	256.83	266/581 (16)	15.60/8.91	36.70	99.2	0.1	0.0	0.8

Table 2. Genome assembly statistics for *Megachile lagopoda*.

Characteristics	*M. lagopoda*
Genome Size (Mb)	256.83
Number of scaffolds	266
Number of chromosomes	16
Scaffold N50 length (Mb)	15.60
GC (%)	36.70
BUSCO completeness (%)	99.30
Protein–conding genes Number	11,157
Mean gene length (bp)	6,764.60
BUSCO completeness (%)	99.10
Repetitive elements Size (Mb)	69.61 (27.10%)
DNA transposons (Mb)	33.82 (13.11%)
SINEs (Kb)	120.52 (0.04%)
LINEs (Mb)	1.58 (0.63%)
LTRs (Mb)	5.06 (1.96%)
Unclassified (Mb)	20.81(8.10%)

Table 3. Genome and annotation statistics for two chromosome-level assemblies of *Megachile lagopoda*.

0.63% long interspersed elements (LINEs), 1.96% long terminal repeats (LTRs), 13.11% DNA transposons, and 36.21% unclassified (Table 3).

MAKER v3.01.03[29] was performed to predict PCGs and we ran EvidenceModeler using *ab initio*, transcript-based, and homology-based methods. BRAKER v2.1.6[30] was utilized to obtain *ab initio* gene predictions, employing GeneMark-ES/ET/EP 4.68_lic[31] and Augustus v3.4.0[32] and automatically trained them based on RNA sequence alignments and reference proteins mined from the OrthoDB v11 database[33]. For the transcript-based predictions, StringTie v2.1.634[34] was used to assemble transcript sequences after aligning the transcriptome data to the final genome assembly using Hisat2 v2.2.0[35]. Subsequently, protein homology and intron position conservation in GeMoMa v1.9[36] were used to predict genes with the

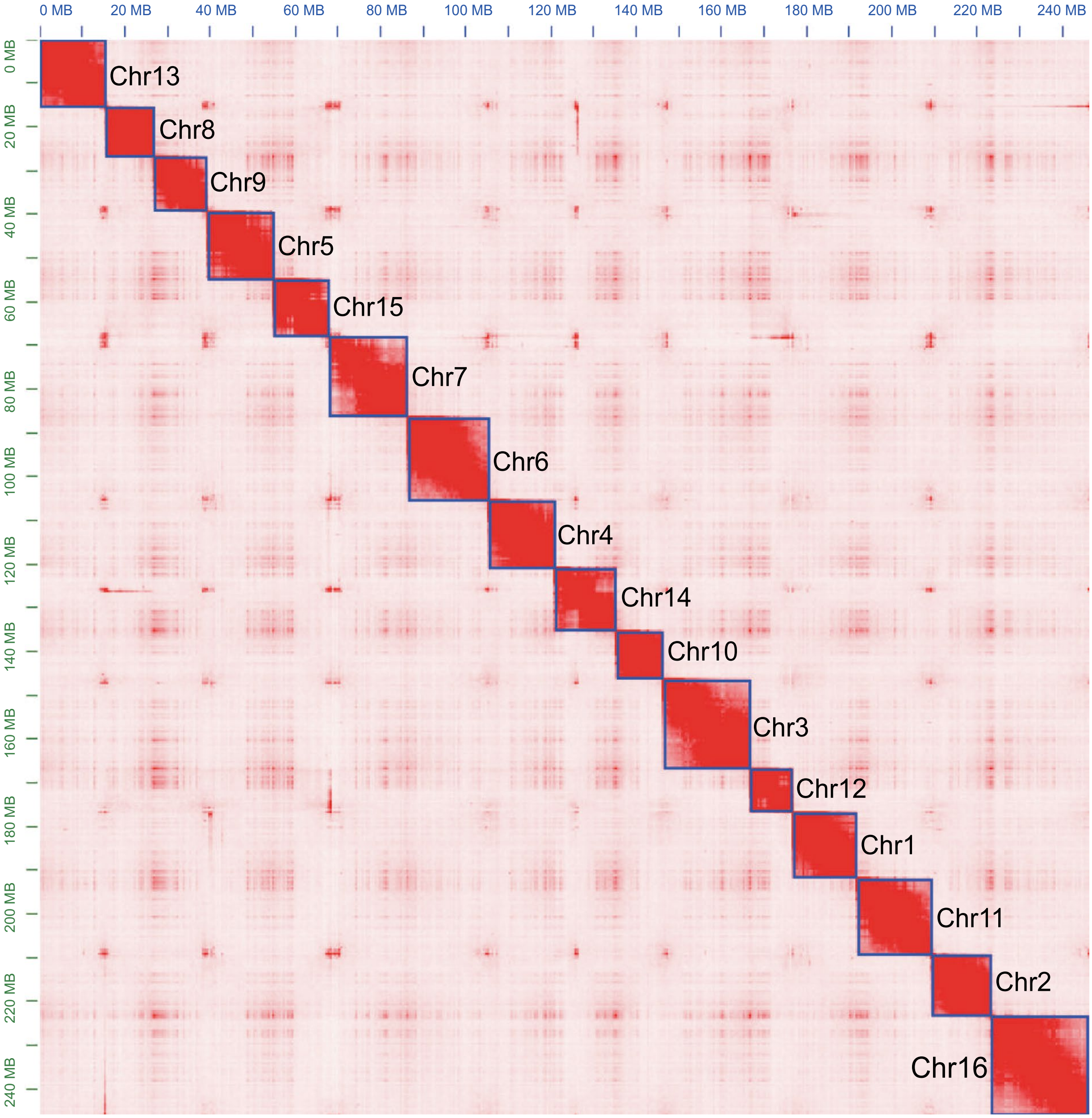

Fig. 2 Genome-wide chromosomal heatmap of *Megachile lagopoda*, the blue boxes show super scaffolds.

parameters "GeMoMa.c = 0.4 GeMoMa.p = 10", and the protein sequences of six species (*Anthidium xuezhongi* (GCA_022405125.1), *Apis mellifera* (GCF_003254395.2), *Bombus vancouverensis* (GCF_011952275.1), *Rhopalosiphum maidis* (GCF_003676215.2), *Tribolium castaneum* (GCF_000002335.3), and *Solenopsis invicta* (GCF_016802725.1) were used to increase search sensitivity. The results obtained from BRAKER and GeMoMa were combined and utilized as the ab initio input for MAKER.

We predicted 11,157 PCGs in the *M. lagopoda* genome, with an average length of 6,764.60 bp. The average number of exons, introns, and CDS of each gene were 7.10, 6.10, and 6.80, respectively, and their corresponding mean lengths were 350.20, 787, and 256 bp, respectively. The BUSCO completeness of predicted proteins exceeded 99.10% for all three types. Furthermore, functional annotation showed that 10,518 and 9,133 genes matched with the UniProtKB and InterProScan databases, respectively.

Two methods were used for gene function annotation: (1) Diamond v2.0.8[37] was performed to search the SwissProt and TrEMBL databases under 'very sensitive' mode (e-value of 1e-5). InterProScan 5.41–78.0[38] searched four public databases to predict protein domains: Pfam[39], SMART[40], Superfamily[41], and CDD[42]; (2) Gene Ontology (GO) and pathway (KEGG, Reactome) were annotated using InterProScan and eggnog-mapper v2.1.5[43], respectively.

ncRNAs were scanned with Infernal v1.1.4[27]. tRNAscan-SE v2.0.9[26] was applied to refine tRNAs, with low-confidence tRNAs filtered out using the built-in script 'Euk High Confidence Filter'. A total of 459 ncRNAs were identified for *M. lagopoda*, including 95 rRNA, 66 miRNA, and 193 tRNA genes. The snRNAs were

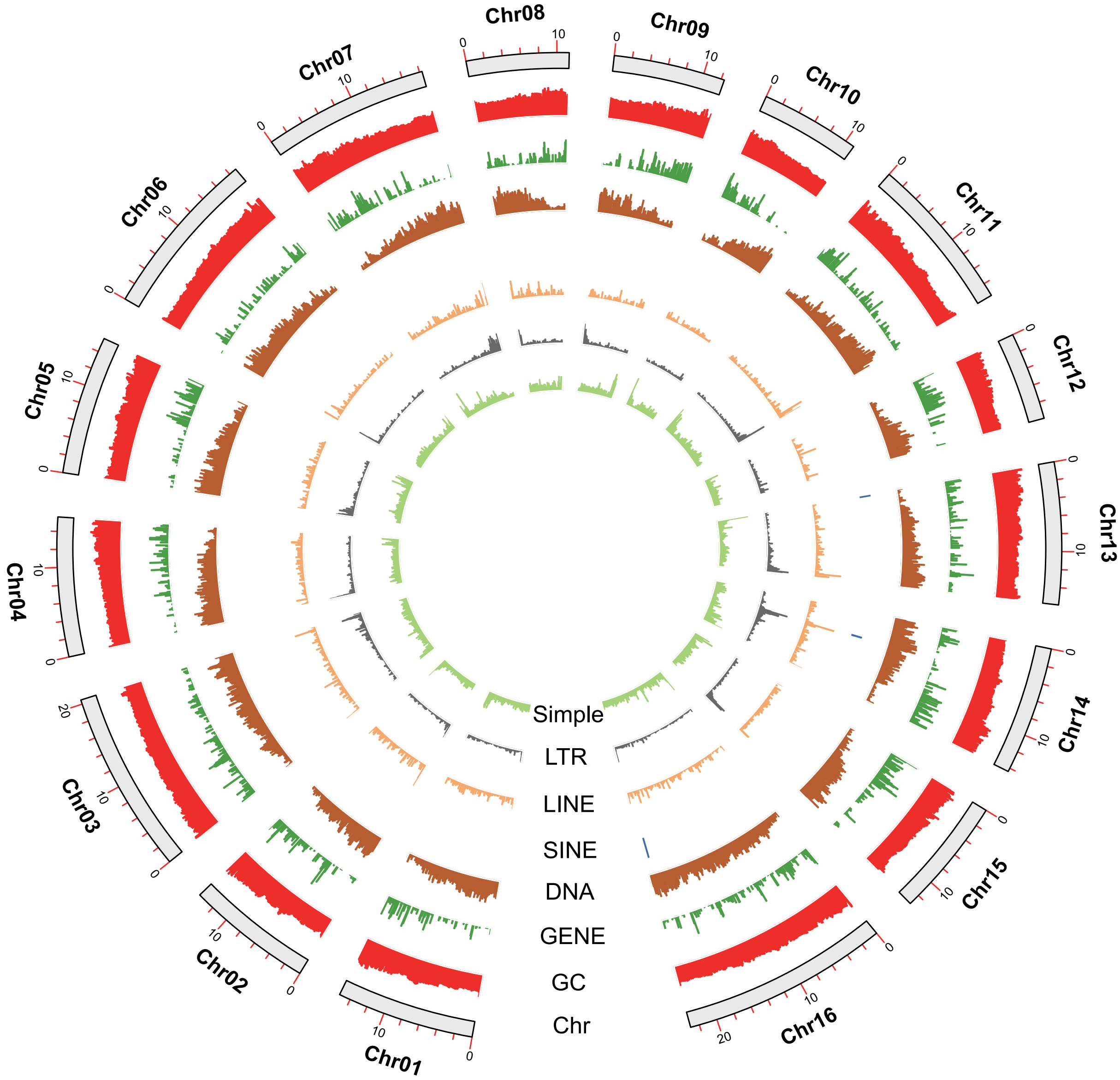

Fig. 3 Circos plot showing the genomic characters of *Megachile lagopoda* from outer to inner: chromosome length (Chr), GC content (GC), density of protein-coding genes (GENE), DNA transposons (DNA), SINE/LINE/LTR retrotransposons (SINE, LINE, LTR).

classified into 24 spliceosomal RNAs in eight groups (U1, U2, U3, U4, U5, U6, U11, and U12), two minor spliceosomal RNAs (U6atac and U4atac), eight C/D box snoRNAs, and one HACA-box snoRNA.

Data Records

All raw sequences and the genome assembly of *M. lagopoda* have been submitted to NCBI under the BioProject IDs: PRJNA977040. The data for the following are also accessioned: transcriptome (SRR24955928)[44], Hi-C (SRR24955929)[45], Illumina (SRR24955930)[46] and PacBio data (SRR24955931)[47]. The NCBI accession number of the assembled genome is GCA_036983795.1[48]. Results of annotation for repetitive and other gene prediction are available in the Figshare data upload[49].

Technical Validation

Completeness and accuracy were used to evaluate the genome quality of *M. lagopoda*. BUSCO was used to assess the completeness of the *M. lagopoda* genome with the insects_odb10 database (n = 1,367). The BUSCO completeness of the final genome assembly was 99.30%, containing 1,356 (99.20%) single-copy BUSCOs, two (0.10%) duplicated BUSCOs, and nine (0.70%) missing BUSCOs. We calculated mapping rates by aligning PacBio, Illumina, and RNA reads to the final assembly to verify accuracy. The mapping rates were 99.57%, 97.71%, and 97.59%, respectively. The Hi-C assembly was subjected to manual correction to ensure accuracy, and the heatmap indicated a highly organized pattern of interactions at the chromosomal level (Fig. 2). The above two approaches confirm the high quality and accuracy of the *M. lagopoda* chromosome-level assembly.

Code availability

The scripts used for genome assembly and annotation in this study were submitted to figshare[49]. All bioinformatics software was performed according to the manual, commands, and/or pipelines of the corresponding software packages.

Received: 13 May 2024; Accepted: 23 October 2024;
Published online: 29 October 2024

References

1. Ascher, J. S., Pickering, J. Discover life bee species guide and world checklist (Hymenoptera: Apoidea: *Anthophila*). Available from: http://www.discoverlife.org/mp/20q?guide=Apoidea_species (accessed March 20, 2024) (2024).
2. Michener, C. D. The Bees of the World. Baltimore, London: John Hopkins University Press. 953 p. (2007).
3. Danforth, B. N., Minckley, R. L., Neff, J. L. & Fawcett, F. *The solitary bees: biology, evolution, conservation*. (Princeton University Press, 2019).
4. Branstetter, M. G. *et al*. Genomes of the Hymenoptera. *Curr. Opin. Insect. Sci.* **25**, 65–75 (2018).
5. Michener, C. D. *The bees of the world*. **Vol. 1** (JHU press, 2000).
6. Pitts-Singer, T. L. & Bosch, J. J. Nest establishment, pollination efficiency, and reproductive success of *Megachile rotundata* (Hymenoptera: Megachilidae) in relation to resource availability in field enclosures. *Environ. Entomol.* **39**, 149–158 (2010).
7. Kemp, W. P. & Bosch, J. Development and emergence of the alfalfa pollinator *Megachile rotundata* (Hymenoptera: Megachilidae). *Ann. Entomol. Soc. Am.* **93**, 904–911 (2000).
8. Bosch, J. & Kemp, W. Development and emergence of the orchard pollinator *Osmia lignaria* (Hymenoptera: Megachilidae). *Environ. Entomol.* **29**, 8–13 (2000).
9. Bosch, J., Kemp, W. P. & Peterson, S. S. Management of *Osmia lignaria* (Hymenoptera: Megachilidae) populations for almond pollination: methods to advance bee emergence. *Environ. Entomol.* **29**, 874–883 (2000).
10. Pasteels, J. J. (1977, October). Une Revue Comparative de l'Éthologie des Anthidiinae Nidificateurs de l'Ancien Monde (Hymenoptera, Megachilidae). In *Annales de la Société entomologique de France (NS)* (Vol. 13, No. 4, pp. 651–667). Taylor & Francis.
11. Gess, S. K. & Gess, F. W. Notes on nesting and flower visiting of some anthidiine bees (Hymenoptera: Megachilidae: Megachilinae: Anthidiini) in southern Africa. (2007).
12. Hebert, P. D., Ratnasingham, S. & De Waard, J. R. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**(suppl_1), S96–S99 (2003).
13. Bushnell, B. BBtools. Available online: https://sourceforge.net/projects/bbmap/ (accessed on 1 October 2023) (2014)
14. Ranallo–Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference–free profiling of polyploid genomes. *Nat. Commu.* **11**, 1432 (2020).
15. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
16. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics.* **36**, 2253–2255 (2020).
17. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* **34**, 3094–3100 (2018).
18. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
19. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science.* **356**, 92–95 (2017).
20. Steinegger, M. & Soding, J. MMseqs. 2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
21. 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).
22. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
23. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).
24. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. Dna.* **6**, 11 (2015).
25. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. Available online: http://www.repeatmasker.org (accessed on 1 October 2022) (2013–2015).
26. Chan, P. P. & Lowe, T. M. TRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* **1962**, 1–14 (2019).
27. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.* **29**, 2933–2935 (2013).
28. Chen, C. *et al*. Tbtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant.* **13**, 1194–1202 (2020).
29. Holt, C. & Yandell, M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *Bmc Bioinformatics.* **12**, 491 (2011).
30. Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *Nar. Genom. Bioinform.* **3**, lqaa108 (2021).
31. Bruna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP: Eukaryotic gene prediction with self-training in the space of genes and proteins. *Nar Genom. Bioinform.* **2**, lqaa26 (2020).
32. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids. Res.* **32**, W309–W312 (2004).
33. Kriventseva, E. V. *et al*. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids. Res.* **47**, D807–D811 (2019).
34. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
35. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods.* **12**, 357–360 (2015).
36. Keilwagen, J., Hartung, F., Grau, J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA–seq data. *Gene prediction: Methods and protocols*. 161–177 (2019).
37. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods.* **12**, 59–60 (2015).
38. Finn, R. D. *et al*. InterPro in 2017-Beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
39. El-Gebali, S. *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
40. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, D493–D496 (2018).
41. Wilson, D. *et al*. SUPERFAMILY—Sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic. Acids. Res.* **37**, D380–D386 (2009).

42. Marchler-Bauer, A. *et al*. CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids. Res.* **45**, D200–D203 (2017).
43. Huerta-Cepas, J. *et al*. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
44. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24955928 (2024).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24955929 (2024).
46. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24955930 (2024).
47. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24955931 (2024).
48. *NCBI Assembly* https://identifiers.org/ncbi/insdc.gca:GCA_036983795.1 (2024).
49. Zhang, D. Genome annotation of *Megachile lagopoda* (Hymenoptera: Megachilidae) (repeats, ncRNAs, and protein–coding genes). *figshare Dataset*. https://doi.org/10.6084/m9.figshare.25138703 (2024).

Acknowledgements

This work was supported by the Second Tibetan Plateau Scientific Expedition and Research program (2019QZKK05010605); the Baseline Survey on Diversity of Pollinating Insects in Qomolangma National Nature Reserve (BIRIPDENNR001); the National Natural Science Foundation, China (31772487, 32070465, 32300382); the National Science Fund for Excellent Young Scholars (32122016) and Shandong Provincial Natural Science Foundation (ZR2023QH353). CDZ's Lab is funded by both the Key Program of the National Natural Science Foundation of China (32330013) and grant from the Key Laboratory of the Zoological Systematics and Evolution of the Chinese Academy of Sciences (grant number 2008DP173354). MCO was funded by the President's International Fellowship Initiative for Visiting Scientists (2024PVC0046).

Author contributions

A.L., Z.N. and C.Z., conceived and supervised this study; Z.D., Q.W. and W.D. collected species; Z.N. identified species; J.F. and Z.D. performed genomic analyses; D.Z., J.F., M.C.O., F.Z., R.F.F. and Q.Z. wrote the paper; A.L., C.Z., Z.N. reviewed and edited the paper. All authors have read and agreed to the current version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.N. or A.L.

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.