



OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the large carpenter bee *Xylocopa dejeanii* Lepeletier, 1841 (Hymenoptera: Apidae)

Dan Zhang^{1,2}, Jianfeng Jin^{3,4}, Zeqing Niu¹✉, Michael C. Orr¹, Feng Zhang⁴, Rafael R. Ferrari^{1,5}, Qingtao Wu¹, Qingsong Zhou¹, Wa Da⁶, Chaodong Zhu^{1,7,8} & Arong Luo^{1,7}✉

Xylocopinae, a diverse bee subfamily comprising over 1,000 bee species, and also a major model system for studying the pollination and evolution of sociality. The lack of chromosome-level genome assembly resources for the Xylocopinae limits our research of their biology and evolution. Here, we provided the first pseudo-chromosomes genome assembly of the *Xylocopa dejeanii* combined PacBio CLR long reads, Illumina sequences, and Hi-C data. The final genome is 194.44 Mb located in 16 chromosomes. Our assembly includes 141 scaffolds, with a scaffold N50 length of 13.15 Mb. BUSCO analysis revealed 99.00% completeness. Genome annotation identified 28.27 Mb of repetitive elements, 10,970 protein-coding genes, and 432 ncRNAs. This high-quality *X. dejeanii* assembly advances our understanding of Xylocopinae genomics and provides new insights into bee evolution.

Background & Summary

Bees are the most ecologically important group of pollinators, comprising over 20,000 described species worldwide¹, but much work remains to be done regarding their genomics². Among bees, the family Apidae, particularly the subfamily Xylocopinae, contains major pollinators of commercial crops, such as cardamom, passionfruit³, clover and more^{4,5}. The carpenter bees of the genus *Xylocopa* Latreille, the largest genus within Xylocopinae, encompass nearly 400 recognized species found on all continents but Antarctica^{4,6}.

Xylocopinae overall exhibits an exceptionally high variety of social behavior comparable to the diversity observed in Halictidae, making them an excellent model system for studying the evolution of sociality⁷. This subfamily spans the entire spectrum of social behaviors, from solitary bees such as those in the genus *Manuelia*⁸ to eusocial groups found in the tribe Allodapini^{9,10}. Some xylocopine present cooperatively breeding, while others display a unique strategy characterized by prolonged maternal-offspring associations and communal overwintering of males and females within family units (namely in *Xylocopa*).

Despite their ecological and evolutionary significance, high-quality chromosome-level genomic resources are insufficient for Xylocopinae. To date, only eight assemblies exist for the >1,000 species of Xylocopinae, with only one for the genus *Xylocopa* (*Xylocopa violacea* [GCA_963969225.2]) (accessed on April 2025 from the NCBI database). Currently, the absence of high-quality chromosome-level genome resources significantly limits our understanding of the evolutionary history and social behavior of Xylocopinae, and also for bees in general.

In this study, we present the first high-quality pseudo-chromosomes genome of *Xylocopa dejeanii* Lepeletier, 1841, a species distributed across the southern Palaearctic and northern Indo-Malayan

¹State Key Laboratory of Animal Biodiversity Conservation and Integrated Pest Management, Institute of Zoology, Chinese Academy of Sciences, Beijing, P.R. China. ²Characteristic Laboratory of Forensic Science in Universities of Shandong Province, Shandong University of Political Science and Law, Jinan, P.R. China. ³College of Life Sciences, Xinyang Normal University, Xinyang, P.R. China. ⁴Department of Entomology, College of Plant Protection, Nanjing Agricultural University, Nanjing, P.R. China. ⁵Environmental Science Training Center, Federal University of Southern Bahia, Porto Seguro, Brazil. ⁶Tibet Plateau Institute of Biology, Xizang, P.R. China. ⁷International College, University of Chinese Academy of Sciences, Beijing, P.R. China. ⁸College of Biological Sciences, University of 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	16.05	82.56
PacBio	30 Kb	31.25	160.70
Hi-C	350	34.37	174.85
RNA	350	7.31	—

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

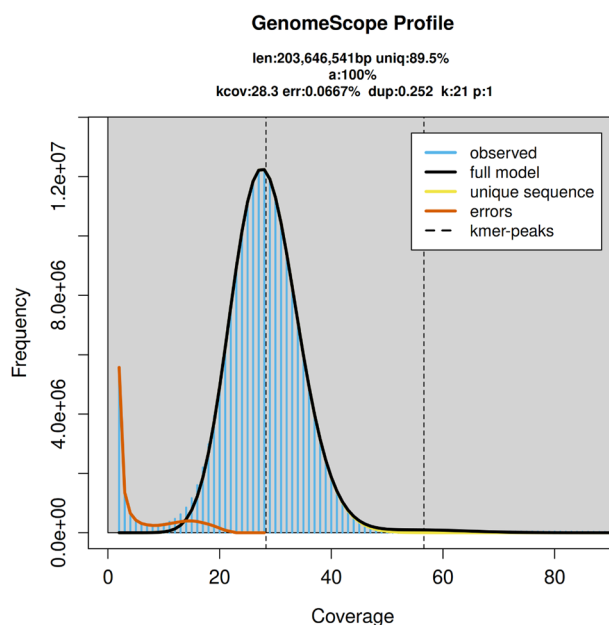


Fig. 1 GenomeScope genome size estimates for *Xylocopa dejeanii*.

regions. Using Illumina sequencing, PacBio long reads, and Hi-C data, we assembled a genome of 194.44 Mb located across 16 chromosomes, comprising 141 scaffolds, with an N50 length of 13.15 Mb and, GC content of 43.50%. Our annotation predicted 10,970 protein-coding genes, 432 ncRNA genes, and 28.27 Mb repeat elements. As a high-quality pseudo-chromosomes assembly for the genus *Xylocopa*, this genome resource provides a critical foundation for comparative genomic studies on bee social behavior and pollination ecology within Xylocopinae.

Methods

Sampling and sequencing. All samples of *X. dejeanii* were collected in Motuo County, Linzhi, Xizang, China (E95.2889, N29.3238, 771 m) on July 20, 2020 by D.Z. and Q.W. All samples were immediately flash-frozen in liquid nitrogen, and then stored at -80°C before subsequent DNA extraction in the lab. The bees were identified by Z.N., and all voucher specimens were deposited in the Institute of Zoology, Chinese Academy of Science. Prior to DNA extraction, we removed the metasomas of all individuals to minimize potential contamination from gut microbes. We used one male sample for PacBio, Illumina whole genome, Hi-C and Illumina transcriptome sequencing.

DNA extraction and sequencing were conducted by Berry Genomics (Beijing, China). PacBio, Illumina whole genome, Illumina transcriptome, and Hi-C sequencing were each performed using a single male individual. Genomic DNA for PacBio and Illumina whole-genome sequencing was done with a Qiagen Blood and Cell Culture DNA Mini Kit. The PacBio Sequel II platform was used for PacBio sequencing with 30 kb insert size sequencing libraries. For Illumina sequencing, a 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. Total RNA was extracted using TRIzol, and RNA libraries were prepared using a TruSeq RNA v2 kit. Hi-C sequencing was conducted on the BGI MGISEQ-2000 platform with 150 bp paired-end reads. In total, 88.98 Gb clean reads were generated, comprising 16.05 Gb ($\sim 82.56\times$) of Illumina whole-genome reads, 31.25 Gb ($\sim 160.70\times$) of PacBio reads, 34.37 Gb ($\sim 174.85\times$) of Hi-C reads, and 7.31 Gb RNA sequencing (Table 1).

Genome assembly. BBTools v38.29¹¹ was performed for quality control: duplicate reads were removed by clumpify.sh, and low-quality reads were filtered with bbduk.sh. K-mer and k-distribution analyses were performed with khist.sh, and Genomescope v2.0¹² was then used to estimate the size, repeat content, and heterozygosity of the *X. dejeanii* assembly, using a maximum k-mer coverage of 1,000. The genome survey estimated the genome size of *X. dejeanii* to be 203.65 Mb, based on the 21 k-mer frequency distribution of short reads (Fig. 1).

Assembly	Total length (Mb)	Number scaffolds/contigs (chromosomes)	Scaffold/contig N50 length (Mb)	GC (%)	BUSCO (n = 1,367) (%)			
					C	D	F	M
Flye	196.59	1,414/1,424	1.61/1.55	43.48	99	0.5	0.2	0.8
NextPolish	196.58	1,414/1,424	1.61/1.55	43.48	99	0.5	0.1	0.9
3D-DNA	195.36	303/661	13.15/1.43	43.50	99	0.5	0.1	0.9
Final	194.46	141/499	13.15/1.46	43.50	99	0.5	0.1	0.9

Table 2. Genome assembly statistics for *Xylocopa dejeanii*. Complete BUSCOs (C); Complete and single-copy BUSCOs (S); Complete and duplicated BUSCOs (D); Fragmented BUSCOs (F); Missing BUSCOs (M).

Characteristics	<i>X. dejeanii</i>
Genome assembly	
Size (Mb)	194.46
Number of scaffolds	141
Number of chromosomes	16
Scaffold N50 length (Mb)	13.15
GC (%)	43.50
BUSCO completeness (%)	99.0
Protein-coding genes	
Gene number	10,970
Mean gene length (bp)	6,715.5
BUSCO completeness (%)	99.0
Repetitive elements	
Size (Mb)	28.27 (14.54%)
DNA transposons (Mb)	5.85 (3.01%)
SINEs (kb)	14.13 (0.00%)
LINEs (Mb)	1.13 (0.58%)
LTRs (Mb)	5.61 (2.89%)
Unclassified (Mb)	7.84 (4.03%)
Number of ncRNA	432
rRNA	89
miRNA	67
snRNA	48

Table 3. Genome assembly and annotation statistics for the chromosome-level assemblies of *Xylocopa dejeanii*.

PacBio long reads were assembled via Flye v2.9¹³ using a minimum overlap between reads of 3000 bp and one round of self-polishing (“-m 3000 -i 1”). NextPolish v1.3.0¹⁴ was performed to polish the primary assembly with two rounds of short Illumina reads. Minimap2 v2.23¹⁵ was conducted to align the Illumina reads and PacBio assembly. Juicer v1.6.2¹⁶ was performed to Hi-C reads for alignment and quality control. Using 3D-DNA¹⁷ with default parameter, the primary contigs were anchored into chromosomal assembly. Juicebox v1.11.08¹⁶ was employed to correct potential errors. MMseqs2¹⁸ was used to identify potential contaminants by comparing sequences against the UniVec and BCBI nucleotide databases, with the “-min-seq-id” parameter, and sequences exhibiting more than 90% alignment were subsequently removed. BUSCO v5.4.4¹⁹ was run to assess genome completeness with the insecta_odb10 (n = 1,367) database as the reference.

Raw PacBio data were assembled with Flye, resulting in a 196.59 Mb assembly with a scaffold N50 length of 1.61 Mb in length (Table 2). After contaminant removal, polishing, and redundancy check, the final *X. dejeanii* assembly achieved a length of 194.44 Mb, including 141 scaffolds and 499 contigs, and the N50 length of 13.15 Mb, and GC content of 43.50% (Tables 2, 3), similar to others published bee genome sizes^{20–22}. Among them, 191.64 Mb (98.54%) of the genome were anchored into 16 pseudo-chromosomes (Fig. 2). The BUSCO completeness of the final genome assembly is 99%, including 1,347 single-copy orthologs (98.50%), seven duplicate BUSCOs (0.5%), two fragmented BUSCOs (0.1%), and 11 missing BUSCOs (0.90%) (Table 2). The mapping rates of PacBio, Illumina, and RNA reads to the final assembly, and the mapping rates were 99.52%, 97.94%, and 91.80%, respectively. The low proportion of duplicated and missing BUSCOs, combined with high mapping rates of raw sequencing data and outstanding assembly metrics, underscores the remarkable contiguity and completeness of our genome assembly.

Genome annotation. RepeatModeler v2.0.2²³ was employed to generate a *de novo* repeat database using the LTR discovery pipeline (-LTRstruct), which identifies specific repeat structures. This output database was subsequently merged with the Dfam 3.5²⁴ and RepBase-20181026 databases²⁵ to establish a custom library. RepeatMasker v4.1.2²⁶ was performed to identify repetitive elements against our customized repeat library. The

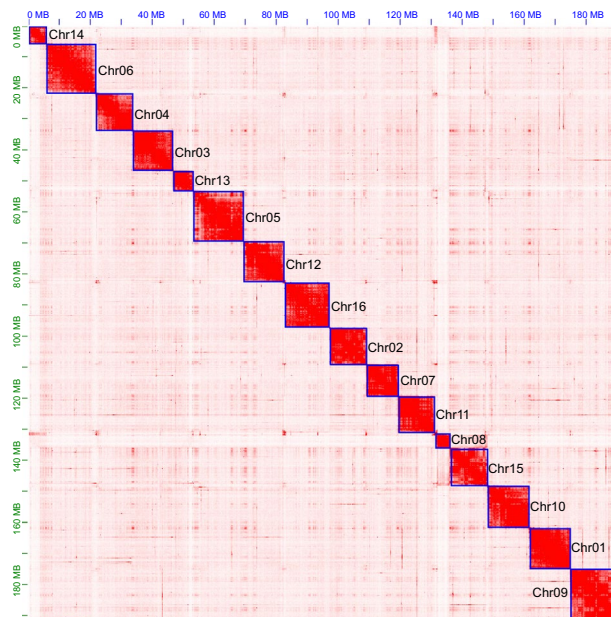


Fig. 2 Genome-wide chromosomal heatmap of *Xylocopa dejeanii*.

visual diagram of *X. dejeanii* genomic characteristics (LINE, LTR, DNA, SINE, GENE, GC, and Chr) were generated by TBtools²⁷ (Fig. 3). We identified 14.54% (28.27 Mb) repetitive regions of the *X. dejeanii* genome, including 0.58% (1.13 Mb) long interspersed elements (LINEs), 2.89% (5.61 Mb) long terminal repeats (LTRs), 3.01% (5.85 Mb) DNA transposons, and 4.03% (7.84 Mb) unclassified (Table 3).

MAKER v3.01.03²⁸ was utilized to predict protein-coding genes, which integrated evidence from *ab initio* predictions, transcript alignments, and homology-based approaches. BRAKER v2.1.6²⁹ was performed to obtain *ab initio* gene prediction, which was applied to trained GeneMark-ES/ET/EP 4.68_{lic}³⁰ and Augustus v3.4.0³¹, and integrating evidence from transcriptome and protein homology to accurately model sequence Properties. Hisat2 v2.2.0 was conducted to the transcriptome evidence in BAM alignments, and the reference proteins sourced from the OrthoDB v11³² database. For transcript prediction, the RNA-seq data were assembled with StringTie v2.1.6³³, and genome assembly was used as reference³⁴. Protein homology and intron position conservation in GeMoMa v1.9³⁵ were used for predicting genes with the parameters “GeMoMa.c = 0.4 GeMoMa.p = 10”. Protein sequences from six insect species: *Anthidium xuezhongii* (GCA_022405125.1), *Apis mellifera* (GCF_003254395.2), *Bombus vancouverensis* (GCF_011952275.1), *Rhopalosiphum maidis* (GCF_003676215.2), *Tribolium castaneum* (GCF_000002335.3), *Solenopsis invicta* (GCF_016802725.1) were employed to increase search sensitivity. The outputs from GeMoMa and BRAKER were then integrated and provided as the *ab initio* input for MAKER.

In total, 10,970 protein-coding genes in the genome *X. dejeanii* were predicted, with an average length of 6715.5 bp. The number of introns, exons, and CDS per gene was 6.3, 7.4, and 6.9, respectively, and their mean lengths were 732.2, 355.8, and 252.5 bp, respectively. The BUSCO completeness assessment for protein sequences reached 99% (n = 1,367), reflecting the high quality of these predictions.

Gene functional annotation was performed using two distinct methods: the SwissProt and TrEMBL databases were searched via Diamond v2.0.8³⁶ on “very sensitive” mode (e-value of 1e-5). InterProScan 5.41-78.0³⁷ was used to predict protein domains by searching four public databases: SMART³⁸, Pfam³⁹, CDD⁴⁰, and Superfamily⁴¹. InterProScan and eggnohmapper v2.1.5⁴² were conducted to annotate the Gene Ontology (GO) and pathway (KEGG, Reactome). Functional annotation attached 10,343 and 8,974 genes, matching the UniProtKB and InterProScan databases, respectively. Combining InterProScan and eggnohmapper annotation results, we identified 8,461 GO items and 3,959 KEGG items.

Two approaches were used for ncRNA annotation: Infernal v1.1.4⁴³ was performed to scan Rfam database to annotate rRNA, snRNA and miRNA; tRNAs were refined via tRNAscan-SE v2.0.9⁴⁴, using the built-in script ‘Euk High Confidence Filter’ to filter out the low-confidence tRNAs. Totally, 432 ncRNAs were captured, including 198 tRNAs, 89 rRNAs, and 67 miRNAs. snRNAs were classed into 34 spliceosomal RNAs in eight groups (U1, U2, U3, U4, U5, U6, U11, and U12), two minor spliceosomal RNAs (U4atac, U6atac), eight C/D box snoRNAs, and one HACA-box snoRNA.

Data Records

All raw data and the assembly of *X. dejeanii* have been submitted to NCBI under Project PRJNA977042, including transcriptome (SRR24955308)⁴⁵, Hi-C (SRR24955309)⁴⁶, Illumina (SRR24955310)⁴⁷, and PacBio (SRR24955311)⁴⁸ data. The NCBI accession number of the assembled genome is GCA_036983795.148⁴⁹. Annotation results of repetitive and other gene predictions are submitted in the Figshare database⁵⁰.

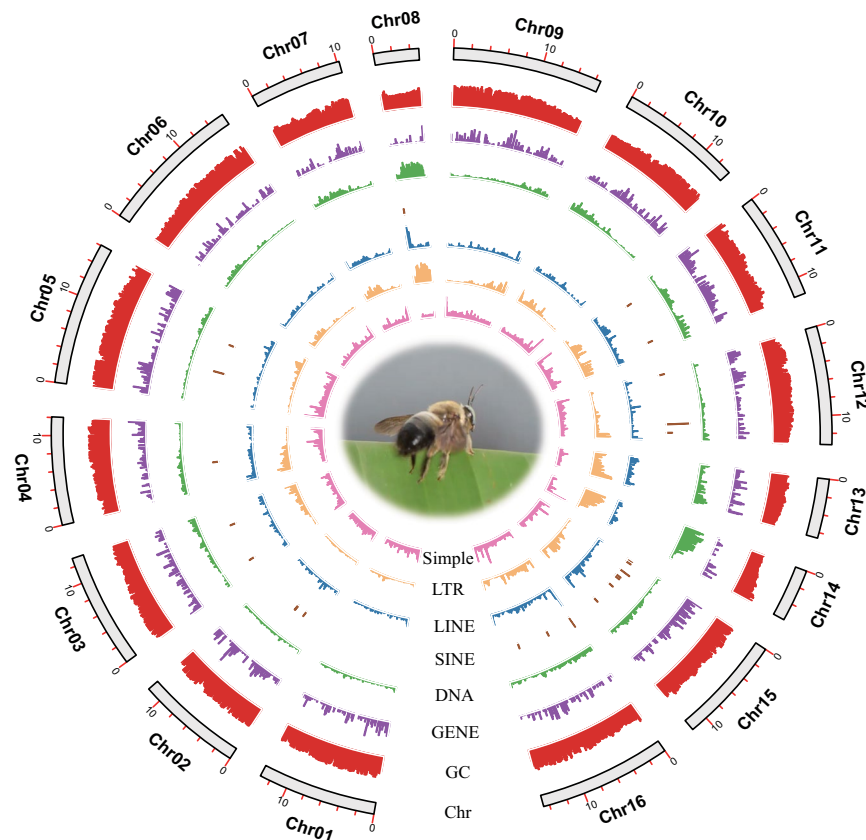


Fig. 3 Circos plot showing the genomic characters of *Xylocopa dejeanii* from outer to inner: chromosome length, GC content, density of protein-coding genes, DNA transposons, SINE/LINE/LTR retrotransposons, and simple repeats.

Technical Validation

The genome quality of *X. dejeanii* was evaluated for completeness and accuracy. Genome completeness was analyzed using BUSCO, via the insects_odb10 database ($n = 1,367$), revealing 99.00% completeness, including 1,347 (98.50%) single-copy orthologs, seven duplicate BUSCOs (0.10%), two fragmented BUSCOs (0.1%), and 11 missing BUSCOs (0.90%). To assess assembly accuracy, we calculated the mapping rates of PacBio, Illumina, and RNA reads to the final assembly, and the mapping rates were 99.52%, 97.94%, and 91.80%, respectively. Manual correction was used for Hi-C assembly to ensure accuracy, and the heatmap revealed a highly structured pattern of interactions at the chromosomal level (Fig. 2).

Code availability

No specific script was used in this work. All bioinformatics analysis was performed according to relevant manuals, commands, and/or pipelines of the corresponding software packages.

Received: 28 April 2025; Accepted: 17 July 2025;

Published online: 23 July 2025

References

- 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 (2025).
- Branstetter, M. G. *et al.* Genomes of the Hymenoptera. *Curr. Opin. Insect. Sci.* **25**, 65–75 (2018).
- Barrera, W. B. Jr, Trinidad, K. A. D. & Presas, J. A. Hand pollination and natural pollination by carpenter bees (*Xylocopa* spp.) in *Passiflora edulis* Sims. f. *flavicarpa* Deg. (yellow passion fruit). *J. Apic. Res.* **60**, 845–852 (2021).
- Michener, C. D. *The Bees of the World*. Baltimore, London: John Hopkins University Press. 953 p. (2007).
- Abrol, D. P. *Pollination biology: biodiversity conservation and agricultural production*. Vol. 792 (Springer, 2012).
- Danforth, B. N., Minckley, R. L., Neff, J. L. & Fawcett, F. *The solitary bees: biology, evolution, conservation*. (Princeton University Press, 2019).
- Sless, T. & Rehan, S. Phylogeny of the carpenter bees (Apidae: Xylocopinae) highlights repeated evolution of sociality. *Biol. Letters*. **19**, 20230252 (2023).
- Flores-Prado, L., Chiappa, E. & Niemeyer, H. M. Nesting biology, life cycle, and interactions between females of *Manuelia postica*, a solitary species of the Xylocopinae (Hymenoptera: Apidae). *New Zeal. J. Zool.* **35**, 93–102 (2008).
- Schwarz, M. P., Bull, N. J. & Cooper, S. J. Molecular phylogenetics of allodapine bees, with implications for the evolution of sociality and progressive rearing. *Syst. Biol.* **52**, 1–14 (2003).
- Schwarz, M. P., Tierney, S. M., Rehan, S. M., Chenoweth, L. B. & Cooper, S. J. The evolution of eusociality in allodapine bees: workers began by waiting. *Biol. letters*. **7**, 277–280 (2011).

11. Bushnell, B. BBtools. Available online: <https://sourceforge.net/projects/bbmap/> (2014).
12. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
13. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
14. Hu, J., Fan, J., Sun, Z. & Liu, S. J. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
15. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
16. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**, 95–98 (2016).
17. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
18. Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
19. 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).
20. Zhang, D. *et al.* Chromosome-level Genome Assembly of *Anthidium xuezhongii* Niu & Zhu, 2020 (Hymenoptera: Apoidea: Megachilidae: Anthidiini). *Genome Biol. Evol.* **14**, evac014 (2022).
21. Zhang, D. *et al.* Chromosome-level genome assembly of *Megachile lagopoda* (Linnaeus, 1761) (Hymenoptera: Megachilidae). *Sci. Data* **11**, 1171 (2024).
22. Nash, W. J. *et al.* The genome sequence of the violet carpenter bee, *Xylocopa violacea* (Linnaeus, 1785): a hymenopteran species undergoing range expansion. *Heredity* **133**, 381–387 (2024).
23. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *PNAS* **117**, 9451–9457 (2020).
24. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids. Res.* **44**, D81–D89 (2016).
25. Bao, W., Kojima, K. K. & Kohany, O. J. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. Dna.* **6**, 1–6 (2015).
26. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. Available online: <http://www.repeatmasker.org> (accessed on 1 (2013–2015)).
27. Chen, C. *et al.* TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant.* **13**, 1194–1202 (2020).
28. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *Bmc Bioinformatics* **12**, 1–14 (2011).
29. 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).
30. 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**, lqaa026 (2020).
31. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids. Res.* **32**, W309–W312 (2004).
32. 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).
33. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**, 1–13 (2019).
34. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Genome Biol* **12**, 357–360 (2015).
35. 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).
36. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods.* **12**, 59–60 (2015).
37. Finn, R. D. *et al.* InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids. Res.* **45**, D190–D199 (2017).
38. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids. Res.* **46**, D493–D496 (2018).
39. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids. Res.* **47**, D427–D432 (2019).
40. Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids. Res.* **45**, D200–D203 (2017).
41. Wilson, D. *et al.* SUPERFAMILY-sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic. Acids. Res.* **37**, D380–D386 (2009).
42. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
43. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
44. Chan, P. P. & Lowe, T. M. TRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* **1962**, 1–14 (2019).
45. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR24955308> (2025).
46. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR24955309> (2025).
47. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR24955310> (2025).
48. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR24955311> (2025).
49. NCBI Assembly https://identifiers.org/ncbi/insdc.gca:GCA_049004755.1 (2025).
50. Zhang, D. Genome annotation (repeats, ncRNAs, and protein-coding genes). *figshare. Dataset*. <https://doi.org/10.6084/m9.figshare.25138748.v1> (2025).

Acknowledgements

This work was supported by the Second Tibetan Plateau Scientific Expedition and Research program (2019QZKK05010605); the National Natural Science Foundation of China (32270465, 32470473, 32300382); the National Science Fund for Excellent Young Scholars (32122016) and Shandong Provincial Natural Science Foundation (ZR2023QH353). CDZ's Lab is funded by the Key Program of the National Natural Science Foundation of China (32330013), and State Key Laboratory of Animal Biodiversity Conservation and Integrated Pest Management (SKLA2501). WD was supported by the Central-guided local projects from the Xizang Science and Technology Department (XZ202301YD0007C).

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. wrote the paper; A.L., C.Z., M.C.O., 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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025