

Whole-genome sequences of the dwarf honey bee subgenus *Micrapis*: *Apis andreniformis* and *Apis florea*

Atma Ivancevic ^{1,†} Madison Sankovitz ^{2,†} Holly Allen,¹ Olivia Joyner,¹ Edward B. Chuong ^{1,*‡} Samuel D. Ramsey ^{3,4,*,‡}

¹BioFrontiers Institute and Department of Molecular, Cellular & Developmental Biology, University of Colorado Boulder, Boulder, CO 80309, United States

²BioFrontiers Institute, University of Colorado Boulder, Boulder, CO 80309, United States

³BioFrontiers Institute and Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, CO 80309, United States

⁴Ramsey Research Foundation, Temple Hills, MD 20748, United States

*Corresponding author: Samuel D. Ramsey, 334 UCB, Boulder, CO 80309, United States. Email: s.ramsey@colorado.edu; Edward B. Chuong, 596 UCB, Boulder, CO 80309, United States. Email: edward.chuong@colorado.edu

[†]Co-first authors

[‡]Co-last authors

The *Micrapis* subgenus, which includes the black dwarf honey bee (*Apis andreniformis*) and the red dwarf honey bee (*Apis florea*), remains underrepresented in genomic studies despite its ecological significance. Here, we present high-quality de novo genome assemblies for both species, generated using a hybrid sequencing approach combining Oxford Nanopore Technologies long reads with Illumina short reads. The final assemblies are highly contiguous, with contig N50 values of 5.0 Mb (*A. andreniformis*) and 4.3 Mb (*A. florea*), representing a major improvement over the previously published *A. florea* genome. Genome completeness assessments indicate high quality, with BUSCO scores exceeding 98.5% using the Hymenoptera database and *k*-mer analyses supporting base-level accuracy. Repeat annotation revealed a relatively low repetitive sequence content (~6%), consistent with other *Apis* species. Using RNA sequencing data, we annotated 12,189 genes for *A. andreniformis* and 12,207 genes for *A. florea*, with ~98% completeness in predicted proteomes. These genome assemblies provide a valuable resource for comparative and functional genomic studies, with the potential to offer new insights into the genetic basis of dwarf honey bee adaptations.

Keywords: dwarf honey bees; genome sequencing; genome assembly; long reads; *Micrapis*; *Apis florea*; *Apis andreniformis*

Introduction

Honey bees (*Apis* spp.) are among the most ecologically and economically significant insects the world over (Hristov et al. 2020; Khalifa et al. 2021; Papa et al. 2022). Renowned for their role in pollination, honey production, and as models for studying social behavior, these eusocial insects are indispensable to agricultural systems and natural ecosystems (Gill 1990; Zayed and Robinson 2012; Hoover and Ovinge 2018). The taxonomic standing of several species and subspecies in the genus *Apis* has been disputed for several decades. There are currently 8 recognized honey bee species that exhibit varying sizes, nest structures, and ecological niches; however, some evidence suggests the recognition of 11 species (Engel 1999; Michener 2007; Crane 2009; Hepburn and Radloff 2011). As part of the so-called “Honey Bee-nome” project, we aim to sequence all recognized species and subspecies within the genus *Apis*, with the goal of better understanding this complex system and identifying potential cryptic species. While the western honey bee *Apis mellifera* is the most extensively studied and globally distributed species, other honey bee species, including those in the subgenera *Apis*, *Megapis*, and *Micrapis*, exhibit remarkable diversity in their behaviors and adaptations, many of which remain underexplored.

The subgenus *Micrapis* consists of the dwarf honey bees (Fig. 1), black *A. andreniformis* and red *A. florea*, which live throughout south and southeast Asia (Oldroyd and Wongsiri 2009). *Apis florea* have been shown to contribute to the pollination of a diverse range of flora, playing an essential role in maintaining the biodiversity and ecosystem stability of its native regions (Abrol 2010; Ali et al. 2017; Shwetha et al. 2020). Colonies of *A. florea* are known to build their nests oriented in a northeast-southwest direction, potentially to maximize sunlight exposure (Ramyarani and Nagaraja 2024). Recently, this species has expanded its geographic range to include Malta and Sudan, with incursions also reported in Australia (El-Niweiri et al. 2019; Uzunov et al. 2024). These bees are the only species known to host the parasitic mite *Euvarroa sinhai*, and given the capacity of *E. sinhai* to parasitize *A. mellifera*, there are concerns about how the spread of *A. florea* may impact *A. mellifera* (Kapil and Aggarwal 1989; Mossadegh 1990).

Substantially less is known of *A. andreniformis*, likely because it has a smaller distribution than its sister species. Additionally, *A. andreniformis* was initially considered a subspecies of *A. florea*, suggesting that some previous work discussing *A. florea* may be describing the biology and behavioral characteristics of *A. andreniformis* (Hepburn and Radloff 2011). Their size and life

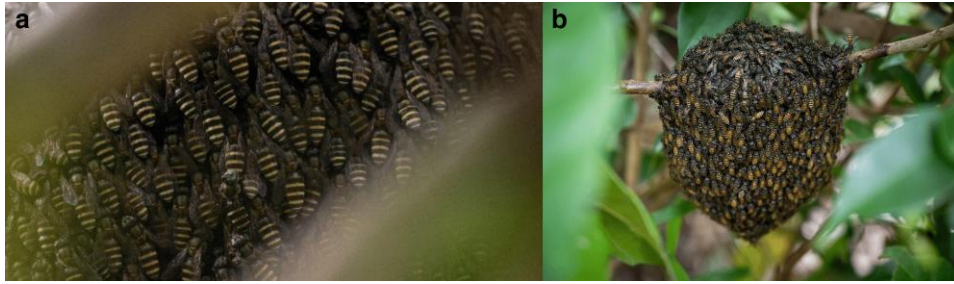


Fig. 1. The dwarf honey bees: a) *A. andreniformis* and b) *A. florea*. Photos by Sirachai Shin Arunrugstichai.

history are similar, suggesting that they may have a comparable ecological impact (Otis 1996). *Apis andreniformis* is host to *Euvarroa wongsirii*, a poorly studied parasite (Lekprayoon and Tangkanasing 1991). However, despite their ecological importance, both species of dwarf honey bee remain underexplored compared to the widely studied western honey bee.

Dwarf honey bees are morphologically and behaviorally distinct from western honey bees. Dwarf honey bees only live in specific regions in Asia, whereas western honey bees have a cosmopolitan distribution (Otis 1996). Workers of the dwarf honey bees are significantly smaller, about half the size of western honey bee workers (Oldroyd 2021). They also exhibit smaller colony sizes, a simplicity that extends to their nesting behavior; both species construct single, exposed combs in open-air settings, typically hanging from thin branches of low shrubs (Rinderer et al. 1996; Oldroyd 2021). This nest structure contrasts with the enclosed nests of western honey bees (Seeley and Morse 1976).

Comparative genomics across honey bee species is critical for understanding the genetic basis of their distinctive traits, including morphology and ecological adaptations. Individual genomes of the western honey bee, *Apis mellifera*, have led to the identification of new subspecies, a better understanding of evolutionary lineages, and the discovery of a significant number of genes involved in adaptations and colony-level quantitative traits (Dogantzis and Zayed 2019). Genomics can enhance our understanding of invasive honey bee parasites and mitigate their impacts by selecting bees with resilient traits, elucidating the effects of parasites on bee health, and identifying biomarkers for rapid diagnostics (Grozinger and Zayed 2020). The dwarf honey bees have relatively few parasites that are otherwise invasive and highly damaging to western honey bee colonies worldwide. They are not hosts to *Varroa* mites, and *Tropilaelaps* mites have only been observed in *A. florea* colonies in India several decades ago (Abrol and Kakroo 1997). Therefore, dwarf honey bee genomic resources are timely and may be valuable in our collective efforts to control invasive parasites. While a high-quality nuclear genome assembly is available for the western honey bee (Wallberg et al. 2019), the genomes of the dwarf honey bees remain poorly characterized. The nuclear genome of *A. florea* was previously sequenced using legacy 454 sequencing technology (Fouks et al. 2021). This assembly has relatively low coverage (20.5×) and is highly fragmented, with 18,378 contigs and a contig N50 of only 24.9 kb, making it suboptimal for genomic analysis. The nuclear genome of *A. andreniformis* has not been sequenced. To address this gap, we combined Oxford Nanopore Technologies (ONT) long-read sequencing with Illumina short-read sequencing to sequence the nuclear genomes and transcriptomes of both species, generating high-quality genome assemblies and transcript-level gene annotations for both dwarf honey bees. These assemblies provide a valuable resource for comparative genomic analyses, offering

new insights into the genetic underpinnings of traits unique to dwarf honey bees.

Materials and methods

Sample collection

Genome sequencing was performed using specimens of *A. andreniformis* and *A. florea* collected from managed colonies in Singapore. Specifically, 1 female worker *A. andreniformis* pupa (Aa1SG1, diploid) was collected from a colony in Frankel, and 1 female worker *A. florea* pupa (Af1SG1, diploid) was collected from a colony in Admiralty.

To complement genome sequencing, additional samples from the same colonies were collected for RNA sequencing. These included

- Two female worker *A. andreniformis* pupae (diploid, samples Aa1SG2 and Aa1SG3)
- One female worker *A. florea* pupa (diploid, Af1SG2) and 1 male *A. florea* pupa (haploid, Af1SG3)

Pupae were extracted from comb cells with forceps after removing the wax cell capping with forceps. After collection, all specimens were flash-frozen in liquid nitrogen and stored at -80°C until extraction. They were then shipped in a cryo-shipper from Singapore to Colorado.

DNA extraction and sequencing

Genomic DNA was extracted from 1 female *A. andreniformis* pupa (Aa1SG1) and 1 female *A. florea* pupa (Af1SG1) using the Quick-DNA Tissue/Insect Kit (Zymo Research), according to the manufacturer's instructions. Both individuals were sequenced using both ONT long-read and Illumina short-read sequencing. For ONT long-read sequencing, a genomic library was prepared with the Native Barcoding Kit 24 V14 kit (SQK-NBD114.24) and sequenced on an R10.4.1 PromethION flow cell (FLO-PRO114M). For Illumina sequencing, a genomic library was prepared using the Ovation Ultralow System V2 kit (Tecan) and sequenced on an Illumina NovaSeq 6000 (University of Colorado Genomics Core) at the University of Colorado Anschutz Medical Campus. Paired-end 2×150 bp reads were generated.

RNA extraction and sequencing

RNA was extracted from 2 female *A. andreniformis* pupae, 1 female *A. florea* pupa, and 1 male *A. florea* pupa using a TRIzol/column-based method. In brief, whole pupae (30 to 110 mg) were homogenized in 1 ml TRIzol (ThermoFisher) and then centrifuged at $12,000 \times g$ for 5 min at 4°C to remove cell debris. The supernatant was incubated for 5 min at room temperature, then mixed with 200 μL of chloroform, and incubated for an additional 3 min. It

was subsequently centrifuged at $12,000 \times g$ for 5 min at 4 °C. The upper, aqueous phase was combined with a 1:1 ratio of 100% ethanol and then purified according to the Direct-zol RNA Miniprep Kit (Zymo Research).

All individuals were sequenced using both ONT long-read and Illumina short-read sequencing. For ONT sequencing, an RNA library was prepared with the PCR cDNA Barcoding Kit (SQK-PCB111.24) and sequenced on an R9.4.1 PromethION flow cell (FLO-PRO002). For Illumina sequencing, an RNA library was prepared using the KAPA mRNA HyperPrep kit (KK8581) and sequenced on an Illumina NovaSeq 6000 (University of Colorado Genomics Core) at the University of Colorado, Anschutz. Paired-end 2×150 bp reads were generated.

Genome assembly and polishing

High-quality genome assemblies for *A. andreniformis* and *A. florea* were generated using a combination of ONT long reads and Illumina short reads. The assembly process was conducted in several steps, including base-calling, demultiplexing, de novo assembly, initial polishing, and additional polishing using the short reads.

ONT raw signals in POD5 format were base called using Dorado v0.7.2 (ONT Public License v1.0) using the super high accuracy model (SUP) to ensure high-quality base calls and `-kit-name SQK-NBD114-24` to enable barcode classification in-line with base calling. Reads were demultiplexed using Dorado's demux function and flags `-no-classify` and `-emit-fastq` to produce FASTQ files for each barcode. Read statistics for demultiplexed reads were obtained using NanoStat v1.6.0 (De Coster et al. 2018).

De novo assemblies of the draft genomes for *A. andreniformis* and *A. florea* were conducted using Flye v2.9.5 (Kolmogorov et al. 2019) with options `-nano-hq` to indicate high-quality reads and the genome size estimate set to 230 Mb (`-genome-size 230 m`) for both bees. The mean coverage of each genome was calculated using Flye v2.9.5 (Kolmogorov et al. 2019). Draft assemblies were polished using Medaka v2.0.0 (ONT Public License v1.0), providing the raw ONT reads for each bee and the model `r1041_e82_400bps_sup_v5.0.0` as input. Further polishing was conducted using Pilon v1.24 (Walker et al. 2014), which leveraged Illumina short reads to correct small indels and base call mismatches.

Illumina short reads were first trimmed for adapter sequences, low-quality bases, and contaminants using the `bbduk.sh` script from BBMap v38.05 (<https://sourceforge.net/projects/bbmap>) and the following parameters: `-Xmx32 g ktrim = r k = 31 mink = 11 hdist = 1 tpe tbo qtrim = r trimq = 10`. Trimmed reads were aligned to their corresponding long-read draft assemblies using BWA-MEM v0.7.5 (Li and Durbin 2009), with alignments filtered to exclude low-quality and unmapped reads (`-q 10 -F 4`) and sorted with SAMtools v1.16.1 (Li et al. 2009; Danecek et al. 2021). The resulting BAM files were used in Pilon v1.24, as described above, to polish assemblies at the per-base level, correcting small base pair errors and indels. All software was executed using default settings unless otherwise specified.

After polishing, VSEARCH v2.14.1 (Rognes et al. 2016) was used with the `-sortbylength` function to sort all contigs by length, and the parameters `-maxseqlength 1000000000` and `-minseqlength 3000` were used to remove contigs shorter than 3 kb for each genome. These small contigs were removed because they often represent assembly artifacts, contamination, or unresolved repetitive sequences that provide little biological information (Murigneux et al. 2020; Rhie et al. 2021). Regardless, this filtering step removed only 0.16% of total sequence length for *A. andreniformis* and 0.27% for *A. florea*, confirming minimal impact on genomic

content. Before submission to GenBank, the final assemblies underwent contamination screening using NCBI's Foreign Contamination Screen Tool Suite v0.5.4 (Astashyn et al. 2024). This process consisted of 2 sequential steps: (i) screening for vector and adapter contamination using FCS-adaptor and (ii) identifying and removing foreign contaminants such as bacterial and viral sequences using FCS-GX, with *Apis*-specific taxonomic filtering (NCBI taxonomy ID 7464 for *A. andreniformis* and taxonomy ID 7463 for *A. florea*). The cleaned assemblies were then submitted to GenBank.

Quality assessment

The quality of each genome was assessed based on 3 criteria: assembly contiguity, genome completeness, and *k*-mer comparison to short reads.

Assembly contiguity was evaluated using QUAST v5.2.0 (Mikheenko et al. 2018) to generate metrics such as N50, L50, total assembly length, and the number of contigs.

Genome completeness was assessed using BUSCO v5.7.1 (Manni et al. 2021) with OrthoDB v10 (Kriventseva et al. 2019) to evaluate the presence or absence of universal single-copy orthologous genes. Detected BUSCO genes were classified as single-copy, duplicated, or fragmented. The following lineage-specific OrthoDB databases were used:

- "hymenoptera_odb10" (creation date: 2024-01-08; 40 genomes; 5991 BUSCOs)
- "insecta_odb10" (creation date: 2024-01-08; 75 genomes; 1367 BUSCOs)
- "arthropoda_odb10" (creation date: 2024-01-08; 90 genomes; 1013 BUSCOs)
- "metazoa_odb10" (creation date: 2024-01-08; 65 genomes; 954 BUSCOs)

To assess potential diploid assembly duplications, we also tested `purge_dups v1.2.6` (Guan et al. 2020) for duplicate removal, which identified minimal duplications (0.13% of the genome for *A. florea* and 0.26% for *A. andreniformis*).

Completeness statistics were compared to 2 existing *Apis* genome assemblies: *A. mellifera* Amel_HAV3.1 (NCBI RefSeq assembly accession GCF_003254395.2) (Wallberg et al. 2019) and *A. florea* Aflo_1.1 (NCBI RefSeq assembly accession GCF_000184785.3) (Fouks et al. 2021).

K-mer analysis was performed using Meryl v1.4.1 (Rhie et al. 2020) and Merqury v1.3 (Rhie et al. 2020) to compare *k*-mers between the short-read data and the polished genome assemblies. Meryl was used to generate *k*-mer databases from the Illumina paired-end short reads. Merqury was used to compare these *k*-mer databases to the genome assemblies, providing estimates of *k*-mer completeness (i.e. the percentage of short-read *k*-mers present in the assembly) and assembly quality values. While the short reads in this study were only used for genome polishing, this approach nonetheless highlighted the concordance between the Illumina data and the final assemblies. All software was executed using default settings.

Whole-genome alignments

To validate assembly contiguity and chromosomal organization, we performed whole-genome alignments using D-GENIES (Cabanettes and Klopp 2018). Pairwise alignments were conducted between our dwarf honey bee assemblies and the *A. mellifera* reference genome (Amel_HAV3.1), as well as between the 2 dwarf species and against the previously published *A. florea*

Table 1. Sequencing statistics of raw reads for *A. andreniformis* and *A. florea*.

Species (sample)	Sequencing type	Number of raw reads	Number of bases	Estimated coverage	SRA accession
<i>A. andreniformis</i> (Aa1SG1)	ONT long read	11,909,664	54.0 Gb	235×	SRR32174174
	Illumina short read	54,425,614	16.4 Gb	71×	SRR32174718
<i>A. florea</i> (Af1SG1)	ONT long read	10,129,152	42.6 Gb	176×	SRR32174175
	Illumina short read	50,980,281	15.4 Gb	67×	SRR32174721

This table shows the number of raw reads (or read pairs for Illumina) and the total number of gigabases generated. For Illumina reads, estimated coverage is based on the number of sequence bases and an assumed genome size of ~230 Mb. For ONT long reads, estimated coverage is based on the mean coverage reported by Flye during the initial assembly step. All reads have been uploaded to NCBI SRA and can be downloaded using the listed accession numbers.

Table 2. Assembly contiguity statistics for the *A. andreniformis* and *A. florea* genomes compared to previously published *Apis* genomes.

Assembly	Sequencing technology	Total ungapped length	GC content	No. of contigs	Contig N50	Contig L50
Aa1SG1	ONT + Illumina for polishing	216.8 Mb	33.7%	333 contigs	5.0 Mb	14
Af1SG1	ONT + Illumina for polishing	221.6 Mb	33.7%	601 contigs	4.3 Mb	16
Aflo_1.1	454 shotgun sequencing	211.3 Mb	33.5%	6,983 scaffolds & 18,378 contigs	Scaffold N50: 2.9 Mb, contig N50: 24.9 kb	Scaffold L50: 22, contig L50: 2,398
Amel_HAv3.1	PacBio; 10× Chromium; Phase Genomics HiC; Bionano	223.9 Mb	32.5%	16 chr, 176 scaffolds, and 227 contigs	Scaffold N50: 13.6 Mb, contig N50: 5.4 Mb	Scaffold L50: 7, contig L50: 13

Assembly statistics for our genomes (Aa1SG1 for *A. andreniformis* and Af1SG1 for *A. florea*) were generated using QUAST on the polished assemblies. For the published assemblies (Aflo_1.1 and Amel_HAv3.1), assembly statistics were obtained from their publicly available NCBI genome pages.

assembly (Aflo_1.1). D-GENIES was configured to use MashMap v2.0 (Jain et al. 2018) as the alignment algorithm, which is optimized for highly similar sequences. Dot plots were generated to visualize collinear relationships and detect large-scale genomic rearrangements between species.

Repeat annotation

RepeatMasker v4.1.7 (Smit et al. 2013–2015, www.repeatmasker.org) was used to softmask repeats in the polished assemblies and generate repeat statistics for each genome. RepeatMasker was configured with Tandem Repeats Finder v4.0.9 (Benson 1999), RMBlast v2.14.0 (www.repeatmasker.org/rmbblast), and all 9 library partitions from Dfam Database v3.8 (Storer et al. 2021). RepeatMasker was executed with default parameters, except for -pa 8, -species Eukaryota, -noisy, and -xsmall. The soft-masked assemblies were used for repeat content analysis, while the unmasked assemblies were used for gene annotation to allow the detection of potential transposable element-derived sequences.

Transcriptome assembly and annotation

The transcriptomes of *A. andreniformis* and *A. florea* were characterized using long-read ONT cDNA sequencing data and short-read Illumina RNA sequencing data.

Long-read cDNA sequencing data were base-called using Dorado v0.7.2 (ONT Public License v1.0) with the super high accuracy (SUP) model and the -no-trim option to retain untrimmed reads, which are necessary for cDNA-specific downstream analyses. Demultiplexing was performed using Dorado's demux function with -no-trim enabled. Read quality and length distributions were assessed using NanoStat v1.6.0 (De Coster et al. 2018).

To refine the ONT reads, Pychopper v2.5.0 (ONT Public License v2.0) was used to identify and trim full-length cDNA reads, with -k PCB111 specified for the barcode kit. High-confidence full-length reads were merged with rescued reads to generate the final processed cDNA dataset. ONT reads from the male *A. florea* sample (Af1SG3) were excluded from further analyses due to poor

sequencing quality, an N50 of only 196 bp, and high barcode retention. The remaining 3 samples—2 *A. andreniformis* females and 1 *A. florea* female—demonstrated sufficient quality and coverage for downstream analyses, with N50 values exceeding 1,100 bp and total yields ranging from 31 to 40 Gb. Processed cDNA reads were aligned to the genome using Minimap2 v2.22 with the -x splice option.

Illumina RNA-seq reads were preprocessed using the bbdutk.sh script from BBMap v38.05 (<https://sourceforge.net/projects/bbmap>) and the following parameters: -Xmx32 g ktrim = r k = 31 mink = 11 hdist = 1 tpe tbo qtrim = r trimq = 10, to remove adapter sequences and low-quality bases. A Hisat2 genome index was generated for each species using Hisat2 v2.1.0 (Kim et al. 2019), and preprocessed reads were mapped to the genome assembly with default sensitivity and the -dta flag to optimize alignments for transcript assembly. BAM files were sorted and indexed using SAMtools v1.16.1 (Li et al. 2009; Danecek et al. 2021), and the 2 Illumina RNA-seq BAM files were merged for each species.

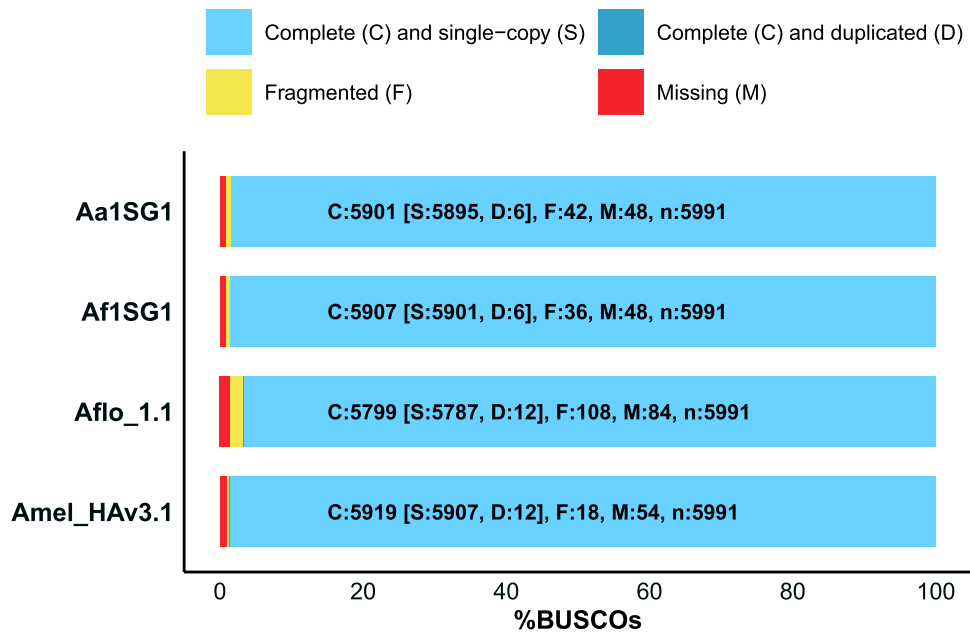
Genome annotation was performed using BRAKER3 (Gabriel et al. 2024), an automated gene prediction pipeline that integrates RNA-seq and protein homology evidence to generate gene models. Annotation was performed on the unmasked genome assemblies to enable the detection of transposable element-derived genes and isoforms. Gene prediction was performed using GeneMark-ETP (Brùna et al. 2024) and AUGUSTUS (Stanke et al. 2008), incorporating extrinsic evidence to refine gene structures. TSEBRA (Gabriel et al. 2021) was used to select the most well-supported transcript isoforms. StringTie2 (Kovaka et al. 2019) was used for transcriptome assembly. To improve homology-based predictions, we utilized the Arthropoda protein dataset from OrthoDB v12 (Tegenfeldt et al. 2025), which was downloaded from https://bioinf.uni-greifswald.de/bioinf/partitioned_odb12/Arthropoda.fa.gz. DIAMOND (Buchfink et al. 2015) was used to incorporate and align the Arthropoda protein sequences. Gene annotation outputs were further processed with GffRead (Pertea and Pertea 2020) for format conversion and filtering.

Table 3. Genome completeness statistics for the *A. andreniformis* and *A. florea* genomes compared to previously published *Apis* genomes.

Assembly	Complete [single, duplicated]	Fragmented	Missing
Hymenoptera			
Aa1SG1	98.5% [98.4%,0.1%]	0.7%	0.8%
Af1SG1	98.6% [98.5%,0.1%]	0.6%	0.8%
Aflo_1.1	96.8% [96.6%,0.2%]	1.8%	1.4%
Amel_HAv3.1	98.8% [98.6%,0.2%]	0.3%	0.9%
Insecta			
Aa1SG1	99.7% [99.7%,0.0%]	0.1%	0.2%
Af1SG1	99.6% [99.6%,0.0%]	0.1%	0.3%
Aflo_1.1	99.2% [99.1%, 0.1%]	0.4%	0.4%
Amel_HAv3.1	99.5% [99.4%, 0.1%]	0.1%	0.4%

BUSCO analysis was performed using the OrthoDB v10 datasets for Hymenoptera (5,991 genes) and Insecta (1,367 genes), with results shown below. Additional analyses using datasets for Arthropoda (1,013 genes) and Metazoa (954 genes) are summarized in [Supplementary Table 1](#) in [Supplementary File 1](#).

BUSCO Assessment Results

**Fig. 2.** Bar chart showing the completeness of *A. andreniformis* and *A. florea* genome assemblies compared to previously published *Apis* genomes, based on BUSCO analysis against the Hymenoptera database.

In addition to our own ONT cDNA and Illumina RNA-seq datasets, we incorporated publicly available *A. mellifera* whole-body RNA-seq data (BioProject PRJNA1263005; [Kwon et al. 2024](#)), which consisted of 3 pooled adult samples (10 individuals per pool, totaling 30 bees from SRA accession numbers SRR33569126, SRR33569127, and SRR33569128). As described above, reads were trimmed and aligned to each dwarf honey bee genome using Hisat2 v2.1.0 ([Kim et al. 2019](#)) and then merged into a single BAM file per species using SAMtools v1.16.1 ([Li et al. 2009](#); [Danecek et al. 2021](#)). These BAMs were included as extrinsic evidence in BRAKER3 alongside our original datasets.

For *A. andreniformis*, BRAKER3 was run using 2 ONT cDNA BAM files (1 per sample), a merged Illumina RNA-seq BAM file from our species-specific data, the merged *A. mellifera* Illumina RNA-seq BAM file, and protein sequences from Arthropoda as extrinsic evidence. For *A. florea*, BRAKER3 was run with 1 ONT cDNA BAM file (from sample Af1SG2), a single species-specific merged Illumina RNA-seq BAM file, a merged species-specific Illumina RNA-seq

BAM file, the merged *A. mellifera* BAM file, and Arthropoda protein sequences.

To avoid isoform inflation in BUSCO analyses of predicted proteomes, we generated a non-redundant protein set by retaining only the longest isoform per gene. BUSCO analyses were repeated on these non-redundant proteomes, which reduced inflated duplication estimates and yielded rates consistent with the genome-level BUSCO results.

To compare gene content across *Apis* species, we performed ortholog analysis using OrthoFinder v3.1.0 ([Emms and Kelly 2019](#)). Input protein FASTA files included BRAKER3-predicted proteomes for *A. andreniformis* and *A. florea*, as well as the published proteome for *A. mellifera* (i.e. GCF_003254395.2_Amel_HAv3.1_protein.faa.gz). OrthoFinder was run with MSA-based orthogroup inference (-M msa -og). We used the output gene counts to summarize (i) total orthogroups, (ii) orthogroups shared by all 3 species, (iii) pairwise-exclusive orthogroups, and (iv) species-specific orthogroups. All software was executed using default settings unless otherwise specified.

Apis mellifera (Amel_HAv3.1)

Apis andreniformis (Aa1SG1)

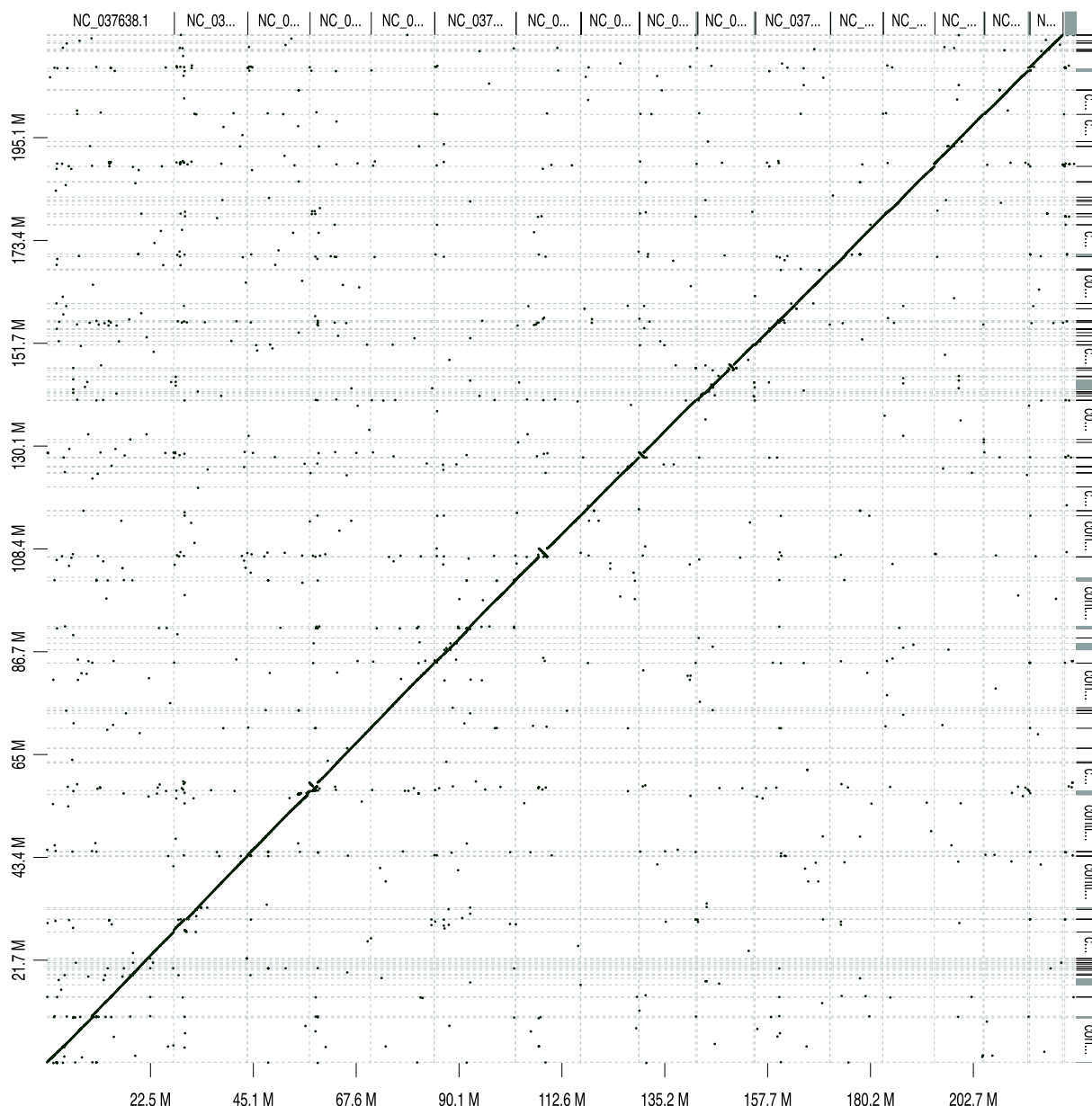


Fig. 3. D-GENIES dot plot showing collinear relationships between *A. andreniformis* (Aa1SG1) and *A. mellifera* (Amel_HAv3.1). The x axis represents *A. mellifera* chromosomes, and the y axis represents *A. andreniformis* contigs. Diagonal straight lines indicate collinearity between the genomes, with dark green lines indicating regions that share greater than 75% sequence identity.

Results and discussion

Genome sequence data summary

Sequencing statistics are summarized in Table 1. In brief, for *A. andreniformis*, long-read sequencing using ONT (Native Barcoding V14 Kit sequenced on R10.4.1 PromethION flow cells) yielded 54.0 Gb of raw reads (read length N50 of 6,657 bp; mean read quality of 15.9). Using an estimated genome size of ~230 Mb gave an initial coverage estimate of 235x. De novo assembly using Flye (Kolmogorov et al. 2019) yielded a reported mean coverage of 235x. Whole-genome sequencing using Illumina yielded 16.4 Gb (54,425,614 150 bp read pairs; mean quality score of 38.5), producing an estimated short read coverage of 71x. Pilon (Walker et al. 2014)

was used to polish the long-read genome assembly with these Illumina short reads. All raw sequencing reads were uploaded to the NCBI Sequence Read Archive (SRA), which revealed that the ONT reads contained <0.01% bacteria and other contaminants. In contrast, the Illumina reads contained many unidentified reads (32.43%, likely derived from adapter sequences) and contaminants (e.g. 6.84% bacteria). Therefore, the bduk.sh script from BBMap was used to filter and trim the short reads before aligning them to the long-read assembly.

For *A. florea*, we generated 42.6 Gb of long reads using ONT sequencing (read length N50 of 6,132 bp; mean read quality of 15.7). This amount provided a coverage estimate of 185x, based on the assumed genome size of ~230 Mb. De novo assembly using Flye

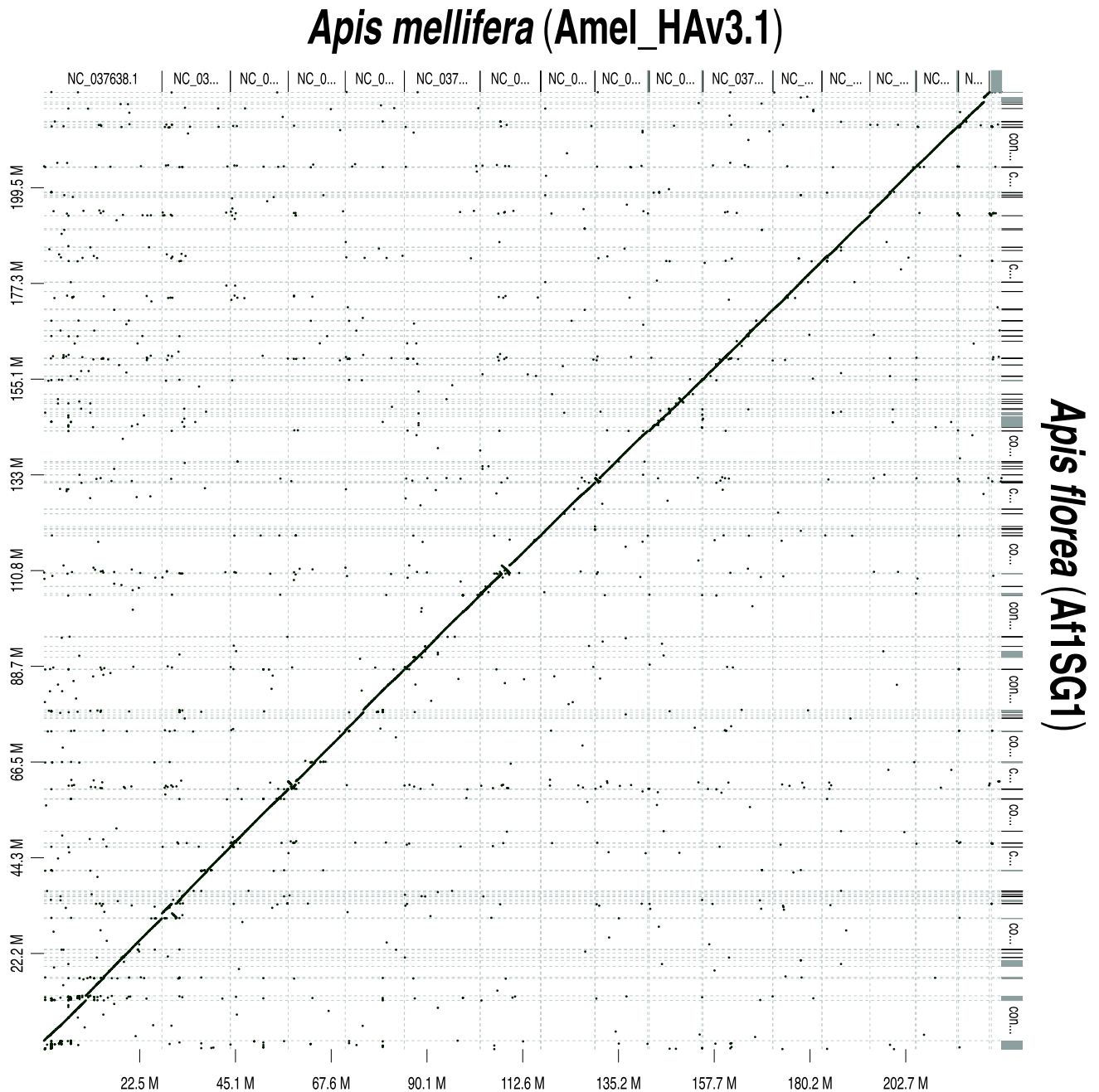


Fig. 4. D-GENIES dot plot showing collinear relationships between *A. florea* (Af1SG1) and *A. mellifera* (Amel_HAv3.1). The x axis represents *A. mellifera* chromosomes, and the y axis represents *A. florea* contigs. Diagonal straight lines indicate collinearity between the genomes, with dark green lines indicating regions that share greater than 75% sequence identity.

(Kolmogorov et al. 2019) resulted in a reported mean coverage of 176x, which is slightly less than our initial estimate. Whole-genome sequencing using Illumina yielded 15.4 GB (50,980,281 read pairs; mean quality score of 38.4), producing an estimated short-read coverage of 67x, which we used for base-level polishing of the long-read assembly. As observed with *A. andreniformis*, the ONT reads for *A. florea* showed <0.01 contaminants, whereas the Illumina reads contained 2.47% bacteria and required trimming prior to alignment. Final assemblies were additionally screened using NCBI's Foreign Contamination Screen (Astashyn et al. 2024) to remove any remaining vector or microbial contaminants before submission to GenBank.

Assembly quality and completeness

We generated high-quality genome assemblies for *A. andreniformis* and *A. florea* using ONT long-read sequencing, followed by Illumina short-read polishing. The final *A. andreniformis* assembly (Aa1SG1) consisted of 333 contigs totaling 216.8 Mb, a contig N50 of 5.0 Mb, and a GC content of 33.7% (Table 2). The *A. florea* assembly (Af1SG1) had a slightly larger total length of 221.6 Mb, with 601 contigs and a contig N50 of 4.3 Mb. Compared to the previously published *A. florea* assembly, Aflo_1.1 (Fouks et al. 2021), which was sequenced using 454 shotgun sequencing, our assemblies show significant improvements in contiguity and completeness. The Aflo_1.1 assembly comprises 6,983 scaffolds and 18,378

Table 4. Summary statistics of BRAKER3 gene annotation for both genome assemblies.

Feature	Aa1SG1	Af1SG1
Number of genes	12,189	12,207
Number of transcripts	16,479	16,511
Number of exons	115,245	115,828
Number of introns	98,766	99,317
Mean gene length (bp)	8501	8173
Mean exon length (bp)	252	252
Mean intron length (bp)	1579	1486

contigs, with a contig N50 of 24.9 kb, which is orders of magnitude lower than our assemblies (Table 2).

Genome completeness estimates using BUSCO confirm the high quality of our assemblies. When evaluated against the Hymenoptera dataset, *A. andreniformis* (Aa1SG1) and *A. florea* (Af1SG1) achieved completeness scores of 98.5 and 98.6%, respectively, with low levels of fragmentation (0.6 to 0.7%) or missing genes (0.8%; Table 3, Fig. 2). BUSCO analysis using the broader Insecta dataset showed even higher completeness estimates of 99.7% for both species (Table 3). K-mer analysis using Merqury further supported the completeness and base-level accuracy of our assemblies. Since Illumina short reads were used exclusively for polishing rather than assembly, the k-mer completeness estimates primarily reflect how well the short-read data align with the final genomes. K-mer completeness was estimated at ~94% for both species (see Supplementary Table 2 and Figs. 1 and 2 in Supplementary File 1), suggesting that some genomic regions, such as highly repetitive sequences, may be underrepresented in the short-read data. However, the low estimated per-base error rates (~0.0002; see Supplementary Table 2 in Supplementary File 1) indicate a high base-level accuracy, consistent with the strong BUSCO scores.

While our assemblies demonstrate strong overall quality, it is worth noting that our Merqury quality scores of approximately 37 (error rate of ~2 in 10,000 bases; see Supplementary Table 2 in Supplementary File 1) are lower than some recently published hymenopteran genomes. For example, Toga et al. (2024) reported a Merqury quality value of 68.08 for the primary contigs of their *Copidosoma floridanum* assembly (66.41 when primary and alternate contigs were combined), generated using PacBio HiFi reads. This difference likely reflects the inherent advantages of PacBio HiFi sequencing technology over ONT long reads for base-level accuracy, as well as potential differences in sequencing coverage, genome-specific complexity, or polishing strategies. Nonetheless, our assemblies represent substantial improvements over previous resources, with dramatically enhanced contiguity and gene completeness that enable comparative analyses.

D-GENIES' whole-genome alignment demonstrated strong collinear relationships between our assemblies and the *A. mellifera* reference genome (Figs. 3 and 4). 74.12% of the *A. andreniformis* assembly aligned with >75% sequence identity to *A. mellifera* (Fig. 3), while 80.10% of the *A. florea* assembly aligned with >75% identity to *A. mellifera* (Fig. 4). Both species showed clear diagonal patterns in dot plots, indicating preserved chromosomal collinearity and confirming that large contigs represent correctly assembled chromosome-scale segments rather than misassemblies.

The comparison between the 2 dwarf species revealed exceptionally high collinearity, with 98.42% of the *A. andreniformis* assembly aligning with greater than 75% sequence identity to *A. florea* (see Supplementary Fig. 3 in Supplementary File 1), reflecting their close evolutionary relationship within the *Micrapis*

Table 5. BUSCO completeness statistics for the predicted proteomes of *A. andreniformis* and *A. florea*.

Proteome	Complete [single, duplicated]	Fragmented	Missing
Hymenoptera			
Aa1SG1	98.1% [97.9%,0.2%]	0.5%	1.4%
Af1SG1	97.9% [97.7%,0.2%]	0.5%	1.6%
Insecta			
Aa1SG1	99.5% [99.1%,0.4%]	0.1%	0.4%
Af1SG1	99.3% [98.9%,0.4%]	0.1%	0.6%

BUSCO analysis was performed using the OrthoDB v10 datasets for Hymenoptera (5,991 genes) and Insecta (1,367 genes). Results indicate the percentage of complete, fragmented, and missing orthologs in each assembly, with complete genes further categorized as single-copy or duplicated.

subgenus. To further validate assembly quality, we also compared our *A. florea* assembly to the previously published Aflo_1.1 genome. We observed 95.18% alignment at >75% sequence identity (see Supplementary Fig. 4 in Supplementary File 1), demonstrating very high sequence conservation as expected between assemblies of the same species.

Finally, repeat analysis identified a diverse set of repetitive elements (see Supplementary Tables 3 and 4 in Supplementary File 1), with total repeat content estimated at 6.44% for *A. andreniformis* and 6.17% for *A. florea*. This finding is consistent with previous studies in *Apis dorsata* (6.89%; Oppenheim et al. 2020) and *A. mellifera* (7.78%; Wallberg et al. 2019). Altogether, our results indicate that these genome assemblies represent the highest-quality *Micrapis* dwarf honey bee genomes published to date.

Gene annotation

RNA was extracted from 2 pupae per species and sequenced using a combination of ONT long-read cDNA sequencing and Illumina short-read RNA-seq. For *A. andreniformis*, we generated 70,960,146 ONT long reads with read length N50s exceeding 1 kb (1,178 bp N50 for sample Aa1SG2; 1,145 bp N50 for sample Aa1SG3; see Supplementary Table 5 in Supplementary File 1), along with 25.6 million high-quality Illumina read pairs. Similarly, for *A. florea*, we obtained 34,084,378 Nanopore reads with read length N50s of 1,098 bp (sample Af1SG2) and 196 bp (sample Af1SG3, excluded from downstream analyses), as well as 21.1 million Illumina read pairs (see Supplementary Table 5 in Supplementary File 1).

Previous annotations for *A. florea* (Fouks et al. 2021) and *A. mellifera* (Wallberg et al. 2019) identified 12,573 and 12,398 genes, respectively. Our BRAKER3 annotation yielded similar numbers, with 12,189 putative genes for *A. andreniformis* and 12,207 genes for *A. florea* (Table 4). BUSCO analysis of the predicted proteomes identified ~98% of orthologs using the Hymenoptera database and >99% of orthologs using the broader Insecta database (Table 5), indicating that our annotation successfully captured nearly all expected genes for these species.

To assess genome quality through gene content conservation, we performed ortholog analysis of the predicted proteomes of *A. andreniformis*, *A. florea*, and *A. mellifera*. The analysis identified 11,774 total orthogroups, which are sets of genes descended from a single ancestral gene. Of these, 9,458 (80.3%) contained genes from all 3 species, suggesting that our assemblies capture the vast majority of conserved honey bee gene families. An additional 1,213 orthogroups were shared exclusively between the 2 dwarf honey bee species, potentially representing *Micrapis*-specific adaptations. Species-specific orthogroups were relatively few: 36 for *A. andreniformis*, 48 for *A. florea*, and 472 for *A. mellifera*. The high proportion

of conserved orthogroups, together with the low number of species-specific gene families in our assemblies relative to the *A. mellifera* reference genome, supports the completeness and quality of our annotations.

Conclusion

Here, we report new high-quality reference genome assemblies for *A. andreniformis*, sampled from Frankel, Singapore, and *A. florea*, sampled from Admiralty, Singapore. These genomes will provide a crucial resource for further biological, population, and evolutionary studies of dwarf honey bees. Our assemblies were generated using a hybrid sequencing approach that combines ONT long reads (with R10.4.1 flow cells) and Illumina short reads. While we sequenced diploid female individuals to enable us to call heterozygous sites, given our relatively modest contig N50 values (5.0 Mb for *A. andreniformis*; 4.3 Mb for *A. florea*), we did not conduct haplotype resolution for these initial assemblies. Ultra-long reads and/or Hi-C data would be necessary to achieve haplotype-resolved and chromosome-level assemblies. Future work sequencing additional individuals from other populations will be critical for developing a comprehensive catalog of genomic variation, including structural variants.

A key strength of our approach is its efficiency, as it required only a single pupa per species to generate genome assemblies with over 98.5% completeness. This approach contrasts with previous studies, such as Aflo_1.1 (Fouks et al. 2021) and Amel_HAV3.1 (Wallberg et al. 2019), which pooled multiple adult drones. While our assemblies achieve high contiguity and completeness comparable to these references at the contig level, the *A. mellifera* reference (Amel_HAV3.1) represents a superior, chromosome-level assembly, scaffolded using additional technologies such as Hi-C and optical mapping. Nonetheless, our results underscore the potential of using hybrid sequencing to study rare or endangered species, such as these dwarf honey bees, where access to adult specimens is limited or unavailable, and even pupae are considered extremely precious.

Data availability

All raw sequencing data used for genome and transcriptome sequencing are available on the NCBI SRA (www.ncbi.nlm.nih.gov/sra) under the following BioProjects:

A. andreniformis

- 1) Genome sequencing and assembly: NCBI BioProject PRJNA1217036
- 2) Transcriptome sequencing: NCBI BioProject PRJNA1217108

A. florea

- 1) Genome sequencing and assembly: NCBI BioProject PRJNA1217041
- 2) Transcriptome sequencing: NCBI BioProject PRJNA1217110

Genome assemblies have been deposited in GenBank under accession GCA_048593525.1 for *A. andreniformis* and GCA_048593485.1 for *A. florea*. Gene annotations generated using BRAKER3 are available on Zenodo (<https://zenodo.org/records/17451668>). Source code and workflows are available on GitHub (https://github.com/atmaivancevic/Micrapis_genome_project).

Supplemental material available at G3 online.

Acknowledgments

We thank the University of Colorado Genomics Shared Resource and BioFrontiers Computing Core for their valuable technical support and for providing access to the high-performance computing clusters that enabled this work.

Funding

M.S. and S.D.R. were funded by grant ID AP23PPQS&T00C157, titled “Study of Honey Bee Pest Diversity to Support Development of Emergency Response Plan,” from the United States Department of Agriculture Animal and Plant Health Inspection Service, as well as the 2022 National Geographic Wayfinder Award. A.I., H.A., O.J., and E.B.C. were funded by NIGMS grant ID 2R35GM128822.

Conflicts of interest. None declared.

Literature cited

- Abrol DP. 2010. Foraging behaviour of *Apis florea* F., an important pollinator of *Allium cepa* L. *Journal Apicultural Research*. 49: 318–325. <https://doi.org/10.3896/IBRA.1.49.4.04>.
- Abrol DP, Kakroo SK. 1997. Observations on concurrent parasitism by mites on four honeybee species in India. *Tropical Agriculture*. 74. <https://journals.sta.uwi.edu/ojs/index.php/ta/article/view/2497>.
- Ali M, Sajjad A, Saeed S. 2017. Yearlong association of *Apis dorsata* and *Apis florea* with flowering plants: planted forest vs. agricultural landscape. *Sociobiology*. 64:18–25. <https://doi.org/10.13102/sociobiology.v64i1.995>.
- Astashyn A et al. 2024. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol*. 25:60. <https://doi.org/10.1186/s13059-024-03198-7>.
- Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 27:573–580. <https://doi.org/10.1093/nar/27.2.573>.
- Brůna T, Lomsadze A, Borodovsky M. 2024. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *Genome Res*. 34:757–768. <https://doi.org/10.1101/gr.278373.123>.
- Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 12:59–60. <https://doi.org/10.1038/nmeth.3176>.
- Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*. 6:e4958. <https://doi.org/10.7717/peerj.4958>.
- Crane E. 2009. *Apis* species:(honey bees). In: *Encyclopedia of insects*. Academic Press. p. 31–32. <https://doi.org/10.1016/B978-0-12-374144-8.00009-6>.
- Danecek P et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience*. 10:giab008. <https://doi.org/10.1093/gigascience/giab008>.
- De Coster W, D’Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*. 34:2666–2669. <https://doi.org/10.1093/bioinformatics/bty149>.
- Dogantzis KA, Zayed A. 2019. Recent advances in population and quantitative genomics of honey bees. *Curr Opin Insect Sci*. 31: 93–98. <https://doi.org/10.1016/j.cois.2018.11.010>.
- El-Niweiri MA, Moritz RF, Lattorff HMG. 2019. The invasion of the dwarf honeybee, *Apis florea*, along the river Nile in Sudan. *Insects*. 10:405. <https://doi.org/10.3390/insects10110405>.
- Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 20:238. <https://doi.org/10.1186/s13059-019-1832-y>.

- Engel MS. 1999. The taxonomy of recent and fossil honey bees (Hymenoptera: Apidae; Apis). *Journal of Hymenoptera Research*. 8: 165–196. <https://biostor.org/reference/28973>.
- Fouks B et al. 2021. The genomic basis of evolutionary differentiation among honey bees. *Genome Res*. 31:1203–1215. <https://doi.org/10.1101/gr.272310.120>.
- Gabriel L et al. 2024. BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res*. 34:769–777. <https://doi.org/10.1101/gr.278090.123>.
- Gabriel L, Hoff KJ, Bruna T, Borodovsky M, Stanke M. 2021. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics*. 22:566. <https://doi.org/10.1186/s12859-021-04482-0>.
- Gill RA. 1990. The value of honeybee pollination to society. In VI International Symposium on Pollination, Tilburg, Netherlands, ISHS Acta Horticulturae 288, p. 62–68. <https://doi.org/10.17660/ActaHortic.1991.288.4>
- Grozier CM, Zayed A. 2020. Improving bee health through genomics. *Nat Rev Genet*. 21:277–291. <https://doi.org/10.1038/s41576-020-0216-1>.
- Guan D et al. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 36:2896–2898. <https://doi.org/10.1093/bioinformatics/btaa025>.
- Hepburn HR, Radloff SE. 2011. Biogeography of the dwarf honeybees, *Apis andreniformis* and *Apis florea*. *Apidologie*. 42:293–300. <https://doi.org/10.1007/s13592-011-0024-x>.
- Hoover SE, Ovinge LP. 2018. Pollen collection, honey production, and pollination services: managing honey bees in an agricultural setting. *J Econ Entomol*. 111:1509–1516. <https://doi.org/10.1093/jee/toy125>.
- Hristov P, Neov B, Shumkova R, Palova N. 2020. Significance of Apoidea as main pollinators. Ecological and economic impact and implications for human nutrition. *Diversity (Basel)*. 12:280. <https://doi.org/10.3390/d12070280>.
- Jain C, Koren S, Dilthey A, Phillippy AM, Aluru S. 2018. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics*. 34:i748–i756. <https://doi.org/10.1093/bioinformatics/bty597>.
- Kapil RP, Aggarwal K. 1989. Observations on reproduction and seasonal population trends of *Euvarroa sinhai* (Mesostigmata: Varroidae) in India. *Progress in Acarology*. 2:277–281. <https://www.cabidigitallibrary.org/doi/full/10.5555/19930513573>.
- Khalifa SAM et al. 2021. Overview of bee pollination and its economic value for crop production. *Insects*. 12:688. <https://doi.org/10.3390/insects12080688>.
- Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 37:907–915. <https://doi.org/10.1038/s41587-019-0201-4>.
- Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. 37: 540–546. <https://doi.org/10.1038/s41587-019-0072-8>.
- Kovaka S et al. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol*. 20:278. <https://doi.org/10.1186/s13059-019-1910-1>.
- Kriventseva EV et al. 2019. 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. <https://doi.org/10.1093/nar/gky1053>.
- Kwon M, Zakhidovna RN, Abdiyazovovich KB, Jung C, Kil EJ. 2024. First metagenomic analysis of virome in Uzbekistan honey bee (*Apis mellifera*): investigating basic information on honey bee viruses. *J Invertebr Pathol*. 206:108171. <https://doi.org/10.1016/j.jip.2024.108171>.
- Lekprayoon C, Tangkanasing P. 1991. *Euvarroa wongsirii*, a new species of bee mite from Thailand. *Int J Acarol*. 17:255–258. <https://doi.org/10.1080/01647959108683915>.
- Li H et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*. 25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25:1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>.
- Manni M, Berkeley MR, Seppely M, Simão FA, Zdobnov EM. 2021. 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. <https://doi.org/10.1093/molbev/msab199>.
- Michener CD. 2007. The bees of the world. JHU press.
- Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*. 34:i142–i150. <https://doi.org/10.1093/bioinformatics/bty266>.
- Mossadegh MS. 1990. Development of *Euvarroa sinhai* (Acarina: Mesostigmata), a parasitic mite of *Apis florea*, on *A. mellifera* worker brood. *Exp Appl Acarol*. 9:73–78. <https://doi.org/10.1007/BF01198984>.
- Murigneux V et al. 2020. Comparison of long-read methods for sequencing and assembly of a plant genome. *GigaScience*. 9: g146. <https://doi.org/10.1093/gigascience/giaa146>.
- Oldroyd BP. 2021. Dwarf honey bees (*Apis (Micrapis)*). In: *Encyclopedia of social insects*. Springer International Publishing. p. 333–339. https://doi.org/10.1007/978-3-030-28102-1_38.
- Oldroyd BP, Wongsiri S. 2009. Asian Honey bees: biology, conservation, and human interactions. Harvard University Press.
- Oppenheim S et al. 2020. Whole genome sequencing and assembly of the Asian honey bee *Apis dorsata*. *Genome Biol Evol*. 12: 3677–3683. <https://doi.org/10.1093/gbe/evz277>.
- Otis GW. 1996. Distributions of recently recognized species of honey bees (Hymenoptera: Apidae; Apis) in Asia. *Journal of the Kansas Entomological Society*. 69:311–333.
- Papa G et al. 2022. The honey bee *Apis mellifera*: an insect at the interface between human and ecosystem health. *Biology (Basel)*. 11:233. <https://doi.org/10.3390/biology11020233>.
- Pertea G, Pertea M. 2020. GFF utilities: GffRead and GffCompare. *F1000Res*. 9:304. <https://doi.org/10.12688/f1000research.23297.2>.
- Ramyanani SK, Nagaraja N. 2024. Nesting behavior of the red dwarf honeybee, *Apis florea* Fabricius (Hymenoptera: Apidae). *Apidologie*. 55:27. <https://doi.org/10.1007/s13592-02401067-7>.
- Rhie A et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. 592:737–746. <https://doi.org/10.1038/s41586-021-03451-0>.
- Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 21:245. <https://doi.org/10.1186/s13059-020-02134-9>.
- Rinderer TE et al. 1996. Comparative nest architecture of the dwarf honey bees. *J Apic Res*. 35:19–26. <https://doi.org/10.1080/00218839.1996.11100909>.
- Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*. 4:e2584. <https://doi.org/10.7717/peerj.2584>.
- Seeley TD, Morse RA. 1976. The nest of the honey bee (*Apis mellifera* L.). *Insectes Soc*. 23:495–512. <https://doi.org/10.1007/BF02223477>.
- Shwetha BV, Bhat NS, Neethu T. 2020. Utilization of *Apis florea* in crop pollination. In: *The future role of dwarf honey bees in natural and agricultural systems*. CRC Press. p. 107–123.

- Smit AFA, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. <http://www.repeatmasker.org>.
- Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*. 24:637–644. <https://doi.org/10.1093/bioinformatics/btn013>.
- Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA*. 12:2. <https://doi.org/10.1186/s13100-020-00230-y>.
- Tegenfeldt F et al. 2025. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res*. 53:D516–D522. <https://doi.org/10.1093/nar/gkae987>.
- Toga K et al. 2024. Long-read genome assembly of the Japanese parasitic wasp *Copidosoma floridanum* (Hymenoptera: Encyrtidae). 14:jkae127. <https://doi.org/10.1093/g3journal/jkae127>.
- Uzunov A et al. 2024. *Apis florea* in Europe: first report of the dwarf honey bee in Malta. *J Apic Res*. 63:1122–1125. <https://doi.org/10.1080/00218839.2024.2386888>.
- Walker BJ et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 9:e112963. <https://doi.org/10.1371/journal.pone.0112963>.
- Wallberg A et al. 2019. A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. *BMC Genomics*. 20:275. <https://doi.org/10.1186/s12864-019-5642-0>.
- Zayed A, Robinson GE. 2012. Understanding the relationship between brain gene expression and social behavior: lessons from the honey bee. *Annu Rev Genet*. 46:591–615. <https://doi.org/10.1146/annurev-genet-110711-155517>.

Editor: B. Jones