DATA NOTE

The genome sequence of the Chalk Yellow-face Bee, *Hylaeus dilatatus* (Kirby, 1802) (Hymenoptera: Colletidae)

[version 1; peer review: 1 approved, 2 approved with reservations]

Clare Boyes[1], Natural History Museum Genome Acquisition Lab,
Darwin Tree of Life Barcoding Collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]Independent researcher, Welshpool, Wales, UK

v1 **First published:** 06 Jan 2026, **11**:18
https://doi.org/10.12688/wellcomeopenres.25424.1
Latest published: 06 Jan 2026, **11**:18
https://doi.org/10.12688/wellcomeopenres.25424.1

Abstract

We present a genome assembly from an individual female *Hylaeus dilatatus* (Chalk Yellow-face Bee; Arthropoda; Insecta; Hymenoptera; Colletidae). The assembly contains two haplotypes with total lengths of 307.38 megabases and 314.32 megabases. Most of haplotype 1 (82.9%) is scaffolded into 12 chromosomal pseudomolecules. Most of haplotype 2 (73.52%) is scaffolded into 12 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 17.78 kilobases. This assembly was generated as part of the Darwin Tree of Life project, which produces reference genomes for eukaryotic species found in Britain and Ireland.

Keywords

Hylaeus dilatatus; Chalk Yellow-face Bee; genome sequence; chromosomal; Hymenoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ? ?

	1	2	3
version 1 06 Jan 2026	✓ view	? view	? view

1. **Fatih Dikmen**, Istanbul University, Istanbul, Turkey
2. **Rick Masonbrink**, Iowa State University, Ames, USA
3. **Panagiotis Ioannidis**, Foundation for Research & Technology - Hellas, Crete, Greece

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: Boyes C: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540) and the Darwin Tree of Life Discretionary Award [218328, https://doi.org/10.35802/218328].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Boyes C, Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding Collective *et al.* **The genome sequence of the Chalk Yellow-face Bee, *Hylaeus dilatatus* (Kirby, 1802) (Hymenoptera: Colletidae) [version 1; peer review: 1 approved, 2 approved with reservations]** Wellcome Open Research 2026, **11**:18 https://doi.org/10.12688/wellcomeopenres.25424.1

First published: 06 Jan 2026, **11**:18 https://doi.org/10.12688/wellcomeopenres.25424.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Hymenoptera; Apocrita; Aculeata; Apoidea; Anthophila; Colletidae; Hylaeinae; *Hylaeus*; *Lambdopsis*; *Hylaeus dilatatus* (Kirby, 1802) (NCBI:txid1542591)

Background

Hylaeus dilatatus is a solitary bee in the family Colletidae which occurs from southern Scandanavia to Portugal, and east to Turkey and parts of Russia (GBIF Secretariat, 2025). In the UK, it is restricted to the south-east of England where it is local on calcareous grassland, chalk quarries, heaths and fenland.

This small bee (forewing length 3.5–4.5mm) is black with yellow markings on the face. The males have expanded black and yellow scapes (Falk & Lewington, 2015). The bee can be found between June and September. Unusually, the females do not carry the pollen back to the nest on a scopa, but carry both pollen and nectar back to their nest in their crop. The nest is lined with a waterproof substance which the female secretes from the Dufour's and salivary glands (Benton & Owens, 2023). The female nests mainly in dead stems of species such as *Rubus* and *Rosa*. It is polylectic, taking a wide range of pollen (Else & Edwards, 2018).

The assembly was produced using the Tree of Life pipeline using a specimen collected from Norfolk, UK (Figure 1). This assembly was generated as part of the Darwin Tree of Life Project.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult *Hylaeus dilatatus* (specimen ID NHMUK015059301, ToLID iyHylDila1; Figure 1), collected from Thompson Common, Norfolk, UK (latitude 52.53, longitude 0.85) on 2022-07-06. The specimen was collected and identified by Clare Boyes. Sample metadata were collected in line with the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

Figure 1. Photograph of the *Hylaeus dilatatus* (iyHylDila1) specimen used for genome sequencing.

The initial identification was verified by an additional DNA barcoding process according to the framework developed by Twyford *et al.* (2024). A small sample was dissected from the specimen and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI) (see the protocol). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding are available on protocols.io.

Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The iyHylDila1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the whole organism was homogenised by powermashing using a PowerMasher II tissue disruptor.

HMW DNA was extracted in the WSI Scientific Operations core using the Manual MagAttract protocol. We used centrifuge-mediated fragmentation to produce DNA fragments in the 8–10 kb range, following the Covaris g-TUBE protocol for ultra-low input (ULI). Sheared DNA was purified by manual SPRI (solid-phase reversible immobilisation). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 10.32 ng/μL and a yield of 474.72 ng, with a fragment size of 9.7 kb.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Prior to library preparation, the DNA was fragmented to ~10 kb. Ultra-low-input (ULI) libraries were prepared using the PacBio SMRTbell® Express Template Prep Kit 2.0 and gDNA Sample Amplification Kit. Samples were normalised to 20 ng DNA. Single-strand overhang removal, DNA damage repair, and end-repair/A-tailing were performed according to the manufacturer's instructions, followed by adapter ligation. A 0.85× pre-PCR clean-up was carried out with Promega ProNex beads.

The DNA was evenly divided into two aliquots for dual PCR (reactions A and B), both following the manufacturer's protocol. A 0.85× post-PCR clean-up was performed with ProNex beads. DNA concentration was measured using a Qubit Fluorometer v4.0 (Thermo Fisher Scientific) with the Qubit HS Assay Kit, and fragment size was assessed on an Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit. PCR

reactions A and B were then pooled, ensuring a total mass of ≥500 ng in 47.4 μl.

The pooled sample underwent another round of DNA damage repair, end-repair/A-tailing, and hairpin adapter ligation. A 1× clean-up was performed with ProNex beads, followed by DNA quantification using the Qubit and fragment size analysis using the Agilent Femto Pulse. Size selection was performed on the Sage Sciences PippinHT system, with target fragment size determined by Femto Pulse analysis (typically 4–9 kb). Size-selected libraries were cleaned with 1.0× ProNex beads and normalised to 2 nM before sequencing.

The sample was sequenced on a Revio instrument (Pacific Biosciences). The prepared library was normalised to 2 nM, and 15 μL was used for making complexes. Primers were annealed and polymerases bound to generate circularised complexes, following the manufacturer's instructions. Complexes were purified using 1.2X SMRTbell beads, then diluted to the Revio loading concentration (200–300 pM) and spiked with a Revio sequencing internal control. The sample was sequenced on a Revio 25M SMRT cell. The SMRT Link software (Pacific Biosciences), a web-based workflow manager, was used to configure and monitor the run and to carry out primary and secondary data analysis.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the iyHylDila1 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

Biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size selected to 400–600 bp using SPRISelect beads. DNA was enriched with Arima-HiC v2 kit Enrichment beads. End repair, A-tailing, and adapter ligation were carried out with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs), following a modified protocol where library preparation occurs while DNA remains bound to the Enrichment beads. Library amplification was performed using KAPA HiFi HotStart mix and a custom Unique Dual Index (UDI) barcode set (Integrated DNA Technologies). Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, libraries were amplified with 10–16 PCR cycles. Post-PCR clean-up was performed with SPRISelect beads. Libraries were quantified using the AccuClear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech).

Prior to sequencing, libraries were normalised to 10 ng/μL. Normalised libraries were quantified again to create equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq 6000.

Genome assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts ($k = 31$) was generated from the filtered reads using FastK. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were assembled using Hifiasm in Hi-C phasing mode (Cheng *et al.*, 2021; Cheng *et al.*, 2022), producing two haplotypes. Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). Contigs were further scaffolded with Hi-C data in YaHS (Zhou *et al.*, 2023), using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023).

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included 118 breaks and 429 joins. This reduced the scaffold count by 14.4% and increased the scaffold N50 by 30.2%. The curation process is described at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020) was run in a Singularity container (Kurtzer *et al.*, 2017) to evaluate *k*-mer completeness and assembly quality for both haplotypes using the *k*-mer databases ($k = 31$) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

The genome was analysed using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021)

to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified from the NCBI Taxonomy (Schoch *et al.*, 2020). For the three domain-level lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all outputs into a blobdir for visualisation. The BlobToolKit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), with containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

PacBio sequencing of the *Hylaeus dilatatus* specimen generated 60.26 Gb (gigabases) from 6.29 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 285.31 Mb, with a heterozygosity of 0.61% and repeat content of 24.94% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 185× coverage. Hi-C sequencing produced 106.02 Gb from 702.14 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The genome was assembled into two haplotypes using Hi-C phasing. Haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The final assembly has a total length of 307.38 Mb in 798 scaffolds, with 591 gaps, and a scaffold N50 of 19.1 Mb (Table 2).

Most of the haplotype 1 assembly sequence (82.9%) was assigned to 12 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Contigs on chromosome 2 (17.6–20.1 Mbp) and on chromosome 1 (13.8–17.2 Mbp) have uncertain order and orientation. During curation we noted that many sequences with centromeric repeats remain in the unscaffolded sequences.

The mitochondrial genome was also assembled (length 17.78 kb, OZ306364.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

For haplotype 1, the estimated QV is 60.9, and for haplotype 2, 59.2. When the two haplotypes are combined, the assembly achieves an estimated QV of 60.0. The *k*-mer completeness is 85.95% for haplotype 1, 86.19% for haplotype 2, and 98.44% for the combined haplotypes (Figure 4).

BUSCO analysis using the hymenoptera_odb10 reference set (n = 5 991) identified 98.9% of the expected gene set (single

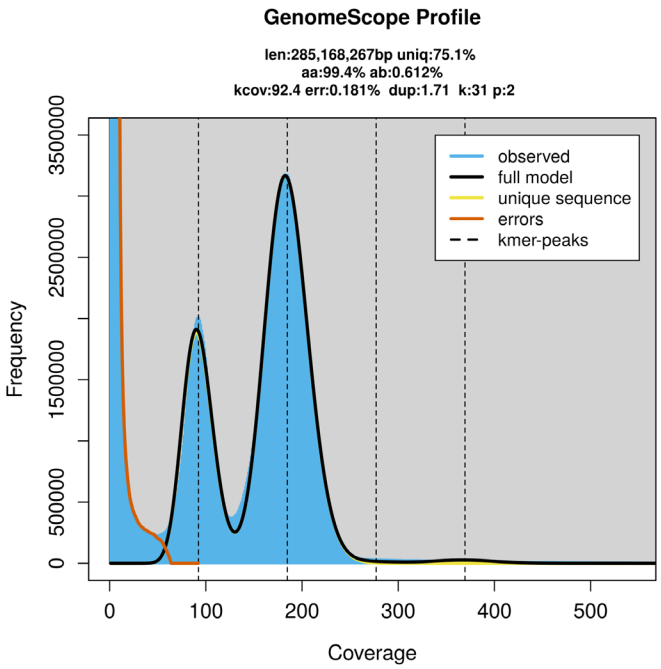

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. **Specimen and sequencing data for BioProject PRJEB81588.**

Platform	PacBio HiFi	Hi-C
ToLID	iyHylDila1	iyHylDila1
Specimen ID	NHMUK015059301	NHMUK015059301
BioSample (source individual)	SAMEA112963056	SAMEA112963056
BioSample (tissue)	SAMEA112963148	SAMEA112963148
Tissue	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq 6000
Run accessions	ERR13900436	ERR13907214
Read count total	6.29 million	702.14 million
Base count total	60.26 Gb	106.02 Gb

Table 2. **Genome assembly statistics.**

Assembly name	iyHylDila1.hap1.1	iyHylDila1.hap2.1
Assembly accession	GCA_966197075.1	GCA_966197085.1
Assembly level	chromosome	chromosome
Span (Mb)	307.38	314.32
Number of chromosomes	12	12
Number of contigs	1 389	1 822
Contig N50	0.67 Mb	0.64 Mb
Number of scaffolds	798	1 398
Scaffold N50	19.1 Mb	16.27 Mb
Longest scaffold length (Mb)	37.34	36.4
Organelles	Mitochondrion: 17.78 kb	-

= 97.3%, duplicated = 1.6%) in haplotype 1. For haplotype 2, BUSCO v.6.0.0 analysis identified 99.0% of the expected gene set (single = 97.0%, duplicated = 2.0%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric calculated for the haplotype 1 is **5.7.Q60**.

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the **'Darwin Tree of Life Project Sampling Code of Practice'**, which can be found in full on the Darwin Tree of Life website. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project. Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible. The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material

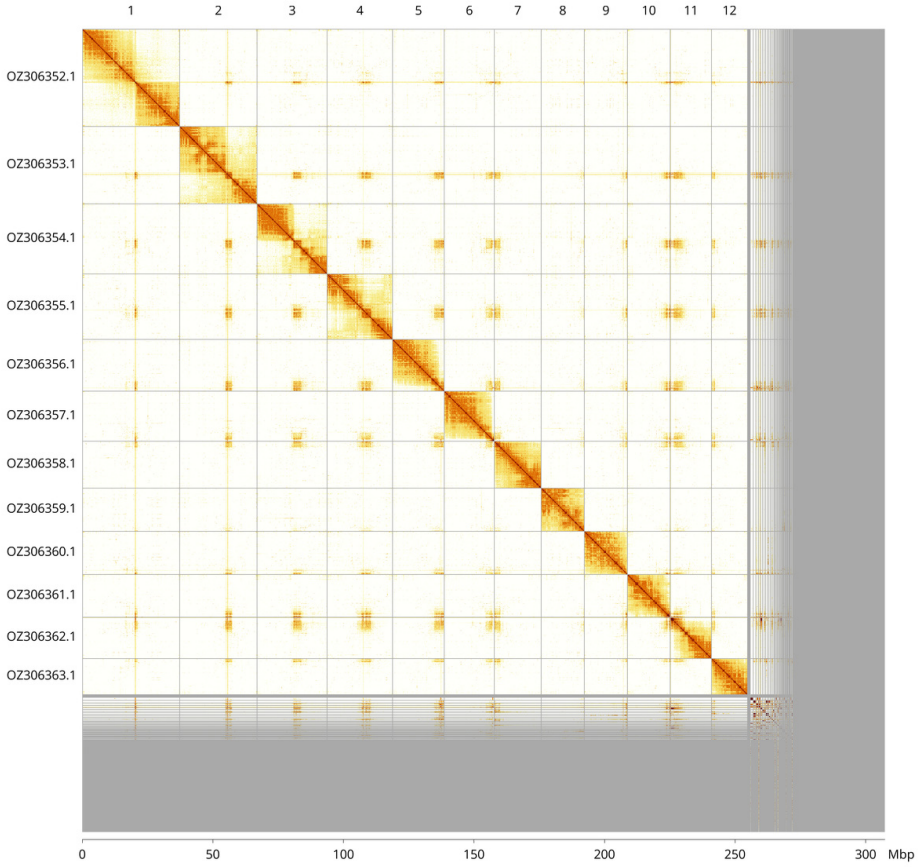

Figure 3. Hi-C contact map of the *Hylaeus dilatatus* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in both haplotypes of the genome assembly of *Hylaeus dilatatus*, iyHylDila1.

Haplotype 1				Haplotype 2			
INSDC accession	**Name**	**Length (Mb)**	**GC%**	**INSDC accession**	**Name**	**Length (Mb)**	**GC%**
OZ306352.1	1	37.34	41.50	OZ306370.1	1	36.40	41.50
OZ306353.1	2	29.63	42	OZ306371.1	2	27.45	41.50
OZ306354.1	3	26.82	42.50	OZ306372.1	3	24.29	42
OZ306355.1	4	25.08	42	OZ306373.1	4	21.51	41
OZ306356.1	5	19.81	42.50	OZ306374.1	5	16.27	42
OZ306357.1	6	19.10	41.50	OZ306375.1	6	17.29	41
OZ306358.1	7	18	40.50	OZ306376.1	7	15.67	40
OZ306359.1	8	16.54	41.50	OZ306377.1	8	16.53	41.50
OZ306360.1	9	16.49	42	OZ306378.1	9	15.61	42
OZ306361.1	10	16.34	41.50	OZ306379.1	10	15.06	41
OZ306362.1	11	15.88	41.50	OZ306380.1	11	11.64	40.50
OZ306363.1	12	13.80	41	OZ306381.1	12	13.38	40.50

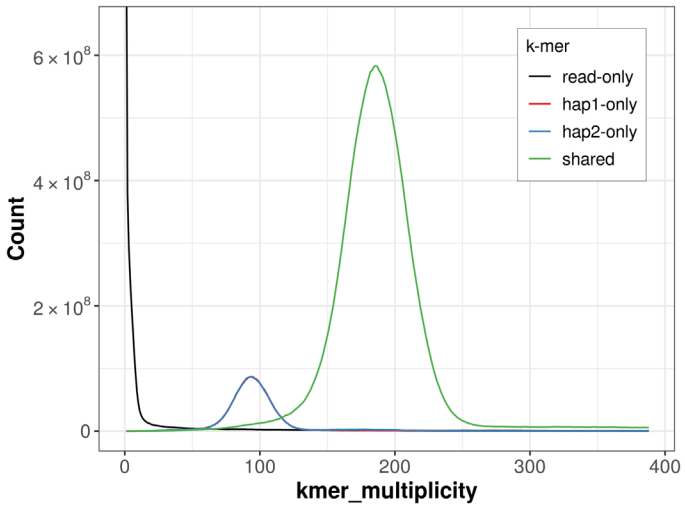

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve corresponds to *k*-mers shared by both haplotypes, and the red and blue curves show *k*-mers found only in one of the haplotypes.

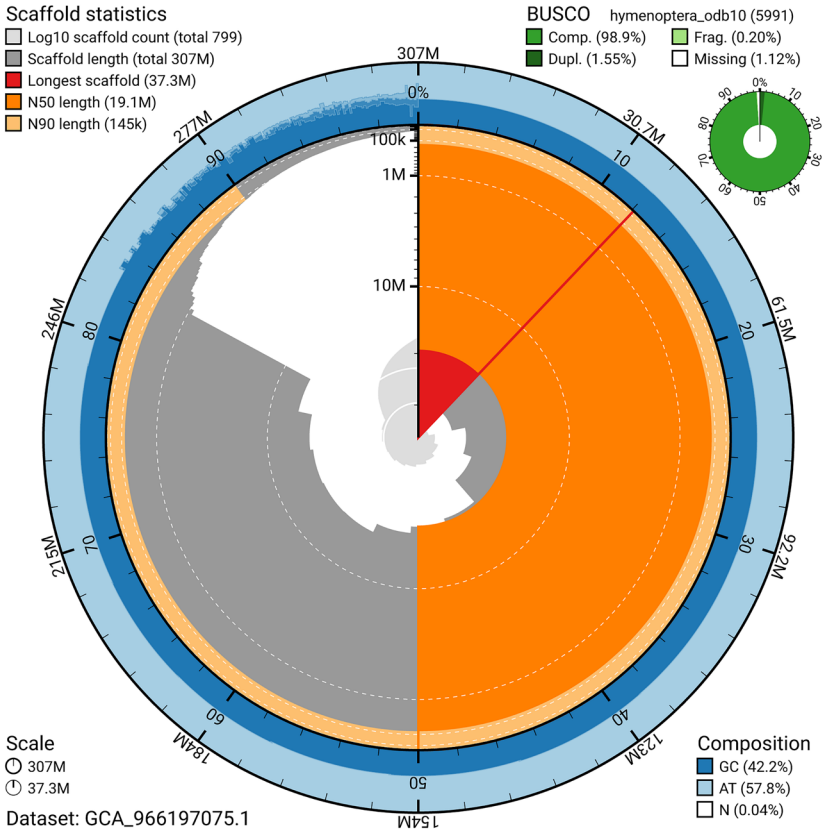

Figure 5. Assembly metrics for iyHylDila1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

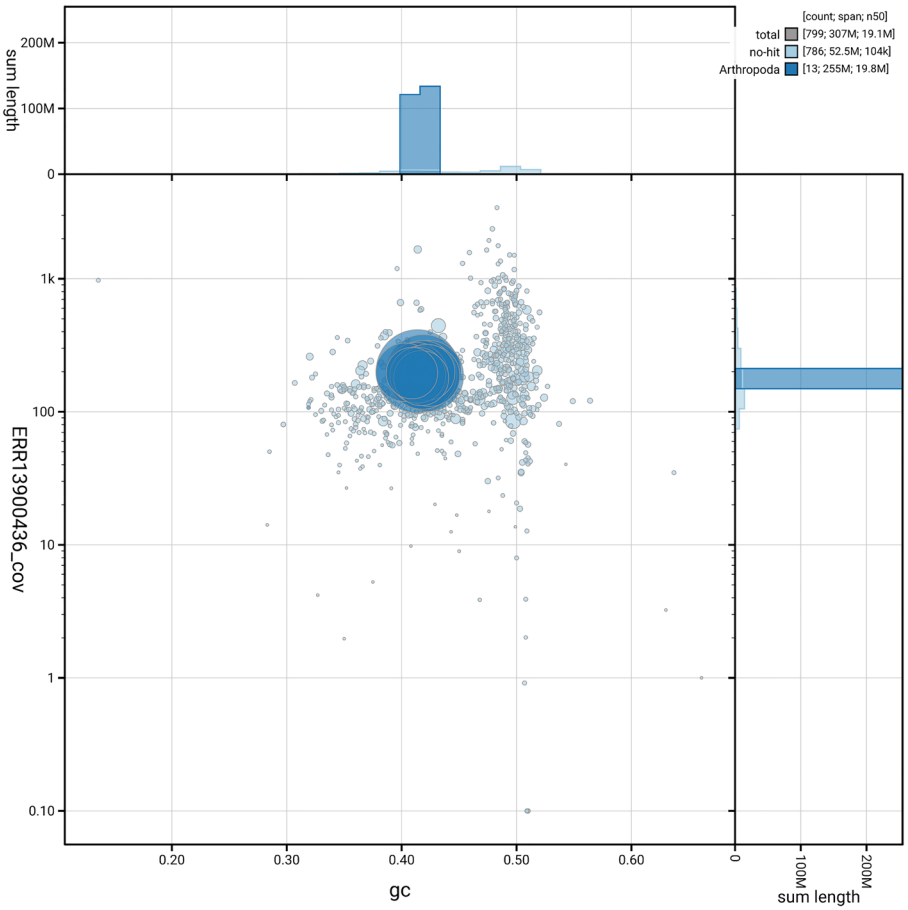

Figure 6. BlobToolKit GC-coverage plot for iyHylDila1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the BlobToolKit viewer.

Table 4. Earth Biogenome Project summary metrics for the *Hylaeus dilatatus* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	5.7.Q60	6.C.Q40
Contig N50 length	0.67 Mb	≥ 1 Mb
Scaffold N50 length	19.10 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 60.9; haplotype 2: 59.2; combined: 60.0	≥ 40
k-mer completeness	Haplotype 1: 85.95%; Haplotype 2: 86.19%; combined: 98.44%	≥ 95%
BUSCO	C:98.9% [S:97.3%; D:1.6%]; F:0.2%; M:0.9%; n:5 991	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	82.90%	≥ 90%

- Legality of collection, transfer and use (national and international)

Each transfer of samples is further undertaken according to a Research Collaboration Agreement or Material Transfer

Agreement entered into by the Darwin Tree of Life Partner, Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances, other Darwin Tree of Life collaborators.

Data availability

European Nucleotide Archive: Hylaeus dilatatus. Accession number PRJEB81588. The genome sequence is released openly for reuse. The *Hylaeus dilatatus* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665) and the Sanger Institute Tree of Life Programme (PRJEB43745). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

Production code used in genome assembly at the WSI Tree of Life is available at https://github.com/sanger-tol. Table 5 lists software versions used in this study.

Author information

Contributors are listed at the following links:

- Members of the Natural History Museum Genome Acquisition Lab
- Members of the Darwin Tree of Life Barcoding collective
- Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team
- Members of Wellcome Sanger Institute Scientific Operations – Sequencing Operations
- Members of the Wellcome Sanger Institute Tree of Life Core Informatics team
- Members of the Tree of Life Core Informatics collective
- Members of the Darwin Tree of Life Consortium

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.4.6	https://github.com/blobtoolkit/blobtoolkit
BUSCO	6.0.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.28-r1209	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	24.10.4	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView

Software	Version	Source
samtools	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.9.0	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text

Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the Universal Protein Knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531.
PubMed Abstract | Publisher Full Text | Free Full Text

Benton T, Owens N: **Solitary bees**. London: Harper Collins, 2023.
Reference Source

Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368.
PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374.
PubMed Abstract | Publisher Full Text | Free Full Text

Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175.
PubMed Abstract | Publisher Full Text | Free Full Text

Cheng H, Jarvis ED, Fedrigo O, *et al.*: **Haplotype-resolved assembly of diploid genomes without parental data.** *Nat Biotechnol.* 2022; **40**(9): 1332–1335.
PubMed Abstract | Publisher Full Text | Free Full Text

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial Arthropod fauna [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 123.
PubMed Abstract | Publisher Full Text | Free Full Text

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008.
PubMed Abstract | Publisher Full Text | Free Full Text

Else G, Edwards M: **Handbook of the bees of Great Britain and Ireland**. London: The Ray Society, 2018.

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
PubMed Abstract | Publisher Full Text

Falk SJ, Lewington R: **Field guide to the bees of Great Britain and Ireland**. London: Bloomsbury, 2015.
Reference Source

Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
PubMed Abstract | Publisher Full Text | Free Full Text

GBIF Secretariat: ***Hylaeus dilatatus* (Kirby, 1802)**. [WWW document]. 2025.
Reference Source

Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025.
Publisher Full Text

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): giaa153.
PubMed Abstract | Publisher Full Text | Free Full Text

Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HiGlass: web-based visual exploration and analysis of genome interaction maps.** *Genome Biol.* 2018; **19**(1): 125.
PubMed Abstract | Publisher Full Text | Free Full Text

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
PubMed Abstract | Publisher Full Text | Free Full Text

Lawniczak MKN, Davey RP, Rajan J, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project [version 1; peer review: 2 approved with reservations].** *Wellcome Open Res.* 2022; **7**: 187.
Publisher Full Text

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text | Free Full Text

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | Publisher Full Text | Free Full Text

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2.
Reference Source

Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat Commun.* 2020; **11**(1): 1432.
PubMed Abstract | Publisher Full Text | Free Full Text

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI Taxonomy: a comprehensive update on curation, resources and tools.** *Database (Oxford).* 2020; **2020**: baaa062.
PubMed Abstract | Publisher Full Text | Free Full Text

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for**

taxonomic verification in the Darwin Tree of Life project [version 1; peer review: 2 approved]. *Wellcome Open Res.* 2024; **9**: 339.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.
Publisher Full Text

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✔ ? ?

Version 1

Reviewer Report 06 February 2026

https://doi.org/10.21956/wellcomeopenres.28010.r144914

?

Panagiotis Ioannidis
Foundation for Research & Technology - Hellas, Crete, Greece

This paper reports the sequencing and assembly of the hymenopteran species Hylaeus dilatatus.

Firstly, the sequencing coverage (185x) is pretty good; in fact, higher than whatever I've seen so far. However, despite such a good coverage and a relatively small genome (285 Mbp), the final assembly contains 798 scaffolds, which are a lot! Especially given that the expected number of chromosomes is 12. Only 82.9% of the haplotype 1 is assigned to one of these 12 chromosome-level scaffolds. In contrast, this number in most other genome assemblies I have reviewed so far is >99%.

All other quality metrics provided by the authors are very good; QV score = 60, k-mer completeness = 98.44%, BUSCO = 98.9%.

As a summary I am expecting that the authors make a comment on the possible causes of such a fragmented assembly. The blob plot (Figure 6) is a very nice summary in this respect; apart from the large scaffolds near the center of the plot, there is a "clould" of small scaffolds at the right (i.e. of higher %GC) and also another, smaller "cloud" at the left (i.e. of lower %GC). Such a figure cannot be presented without a comment.

I'm not saying that such a genome shouldn't be indexed. I'm just saying that, at the very least, the authors should state that their assembly is fragmented; more fragmented than it should be, to be precise. I appreciate the effort on both the side of the authors, as well as on the side of the journal, but such findings shouldn't be left without comment. I appreciate that there a particular format, but I would expect a sentence or two about the findings.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: insect genomics, bionformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 03 February 2026

https://doi.org/10.21956/wellcomeopenres.28010.r144241

?

Rick Masonbrink
Iowa State University, Ames, Iowa, USA

This is a significant contribution to the diversity of Hymenoptera genomes available. They've assembled a high-quality phased genome at chromosomal resolution using PacBio and Hi-C data. Every quality metric used points to this genome being high quality, though there are some minor issues that need to be addressed.
Photograph 1 does not seem to be of sufficient quality and hides a large part of the external characteristics.
The TC buffer used in the Hi-C sample preparation needs to be written out.
Table 5 should be referenced in the methods.
The repeat content estimation by genomescope may be misleading as it also contains PCR/Sequencing errors. May be important to mention this caveat.
Under Assembly statistics the two phased genomes are described differently, with one being chromosomal and one being scaffolded. It appears to me that both have been assigned to pseudomolecules and are thus both chromosomal. The sentence after this does inform the reader of which genome the authors are referring to.
Figure 3 seems odd. The visible Hi-C signal indicates the genome size is around ~270Mb, with a large section shown in gray. Is this gray area comprised of exceedingly small contigs? Did the authors not perform small contig elimination? How did the authors decide that this species has 12 chromosomes? If there is no external evidence, like a chromosome squash, then it appears that some of these chromosomes are sharing the same sequences between centromere and telomere. This is unlikely and may indicate some scaffolding errors, unless this species has

telocentric/acrocentric chromosomes.
There is an incongruity with the BUSCO version and the database used. BUSCO v6 was released with the hymenoptera_odb12, but the text says BUSCO v6 was used with hymenoptera_odb10. So, was Busco v6 used on an older database?
What BLAST database was used for Blobtools? This seems like an extraordinary number of unannotated contigs.
Figure 4 does not appear to have a red line.
Altogether this submission warrants indexing after these minor details are investigated, as it is of high value to the scientific community.
Rick Masonbrink

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics; genome assembly, gene annotation, Hi-C, cytology, molecular biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 21 January 2026

https://doi.org/10.21956/wellcomeopenres.28010.r144909

Fatih Dikmen
Istanbul University, Istanbul, Turkey

The manuscript presents a phased, chromosome-level genome assembly for the yellow-faced bee, Hylaeus dilatatus. The study is a significant and timely contribution to the field of insect genomics, providing a high-quality genomic resource for an important pollinator genus with unique

biological characteristics.
Hylaeus bees are particularly interesting from an evolutionary standpoint as they lack external pollen-carrying structures (scopa), instead transporting pollen internally. A high-quality reference genome for a representative species like H. dilatatus provides an essential foundation for investigating the genetic underpinnings of this and other unique adaptive traits.
The technical quality of the genome assembly is good. The combination of PacBio long-read sequencing, providing deep coverage of approximately 185×, and Hi-C data for scaffolding has proven highly effective. The resulting quality metrics are fair. The assembly boasts a combined QV score of 60.0, which translates to an extremely low error rate, indicating a high-fidelity genomic sequence. This is further supported by the high k-mer completeness of 98.44%.
From a functional perspective, the genome is remarkably complete. The BUSCO analysis, identifying approximately 99% of the expected single-copy orthologs from the Hymenoptera database, confirms that the assembly contains a near-complete set of genes.
In conclusion, the authors have produced a gold-standard genomic resource for an ecologically and evolutionarily significant bee species. The methodology is sound, the results are robustly validated, and the resulting assembly is of the highest quality. This work will be good groundbase for future studies on pollinator genetics, the evolution of social behavior, and the conservation of this vital insect group. The manuscript is well-written, and the data presented clearly supports the authors' claims of a high-quality assembly. It is a valuable addition to the scientific literature.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genetics, Taxonomy, Hymenoptera,

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.