DATA NOTE

The genome sequence of the Bee moth, *Aphomia sociella* (Linnaeus, 1758) (Lepidoptera: Pyralidae)

[version 1; peer review: 1 approved]

Douglas Boyes[1], Clare Boyes[2],
University of Oxford and Wytham Woods 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]UK Centre for Ecology & Hydrology, Wallingford, England, UK
[2]Independent researcher, Welshpool, Wales, UK

v1 **First published:** 13 Oct 2025, **10**:571
https://doi.org/10.12688/wellcomeopenres.24979.1
Latest published: 13 Oct 2025, **10**:571
https://doi.org/10.12688/wellcomeopenres.24979.1

Abstract

We present a genome assembly from an individual female *Aphomia sociella* (Bee moth; Arthropoda; Insecta; Lepidoptera; Pyralidae). The assembly contains two haplotypes with total lengths of 681.66 megabases and 601.52 megabases. Most of haplotype 1 (99.63%) is scaffolded into 30 chromosomal pseudomolecules, including the W and Z sex chromosomes. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 15.38 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

Aphomia sociella; Bee moth; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓

	1
version 1 13 Oct 2025	✓ view

1. **Bin Zhang**, Qingdao Agricultural University, Qingdao, China

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 D: Investigation, Resources; **Boyes C**: Writing – Original Draft Preparation;

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).

How to cite this article: Boyes D, Boyes C, University of Oxford and Wytham Woods Acquisition Lab *et al.* **The genome sequence of the Bee moth, *Aphomia sociella* (Linnaeus, 1758) (Lepidoptera: Pyralidae) [version 1; peer review: 1 approved]** Wellcome Open Research 2025, **10**:571 https://doi.org/10.12688/wellcomeopenres.24979.1

First published: 13 Oct 2025, **10**:571 https://doi.org/10.12688/wellcomeopenres.24979.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Amphiesmenoptera; Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Pyraloidea; Pyralidae; Galleriinae; *Aphomia*; *Aphomia sociella* (Linnaeus, 1758) (NCBI:txid688398)

Background

Aphomia sociella, the Bee moth, is a micromoth in the family Pyralidae, which is common throughout Britain and occurs throughout Europe, parts of Asia, North America and northern Australia (GBIF Secretariat, 2023). The moth lives in nests of bumblebees and social wasps, preferring those which are above ground (Goater, 1986). The moth can reduce the reproductive success of bumblebees; however, bumblebees which collect a wide range of pollens had a lower risk of damage to the colony (Schweiger *et al.*, 2022).

The moth is sexually dimorphic: the male is smaller (forewing length 12 mm) than the female, and is plainer. It has brown wings with two darker crosslines. The female (forewing length up to 17 mm) has browner forewings which are tinged with green and the crosslines are darker (Sterling *et al.*, 2023). Unusually for a Pyralid moth, it does not fold its antennae back along its body when at rest (Waring *et al.*, 2017). The larvae live in the nests of their hosts and initially feed on debris but will also eat the developing insects. They overwinter in a tough, silken cocoon. These can be found outside the hymenopteran nest in suitable dry places; and sometimes in bird boxes (Boyes, 2018). The moth is unusual in that as well as using pheromones to attract females, the males also sing ultrasonically using a special structure on the wing base which is only present in the males. The songs vary significantly between individuals; and it has been demonstrated that they initiate the female sexual response (Kindl *et al.*, 2011).

The genome of *Aphomia sociella* was sequenced as part of the Darwin Tree of Life Project, a collaborative effort to sequence of the named eukaryotic species in the Atlantic Archipelago of Britain and Ireland. Here we present a chromosomally complete genome sequence *Aphomia sociella* based on one specimen from Wytham Woods, Oxfordshire, UK. We present a chromosome-level genome sequence for *Aphomia sociella*, the Pyralid moth. The assembly was produced using the Tree of Life pipeline from a specimen collected in Wytham Woods, Oxfordshire, United Kingdom (Figure 1). This assembly was generated as part of the Darwin Tree of Life Project, which aims to generate high-quality reference genomes for all named eukaryotic species in Britain and Ireland to support research, conservation, and the sustainable use of biodiversity (Blaxter *et al.*, 2022). Another scaffold-level assembly for this species is also available (GCA_052324605.1; submitted by Norwegian University of Life Science) (data obtained via NCBI datasets, O'Leary *et al.*, 2024).

Figure 1. Photograph of the *Aphomia sociella* (ilAphSoci2) specimen used for genome sequencing.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult female *Aphomia sociella* (specimen ID Ox000453, ToLID ilAphSoci2; Figure 1), collected from Wytham Woods, Oxfordshire, UK (latitude 51.764, longitude -1.337) on 2020-06-13. The specimen was collected and identified by Douglas Boyes (University of Oxford). For the Darwin Tree of Life sampling and metadata approach, refer to Lawniczak *et al.* (2022).

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 ilAphSoci2 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the abdomen was homogenised by powermashing using a PowerMasher II tissue disruptor. HMW DNA was extracted using the Automated MagAttract v2 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 automated 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.

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 head and thorax tissue of the ilAphSoci2 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 and 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), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

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 6 breaks and 155 joins. The curation process is documented 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 *Aphomia sociella* specimen generated 45.96 Gb (gigabases) from 4.84 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 626.91 Mb, with a heterozygosity of 1.49% and repeat content of 26.84% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 48× coverage. Hi-C sequencing produced 105.76 Gb from 700.37 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

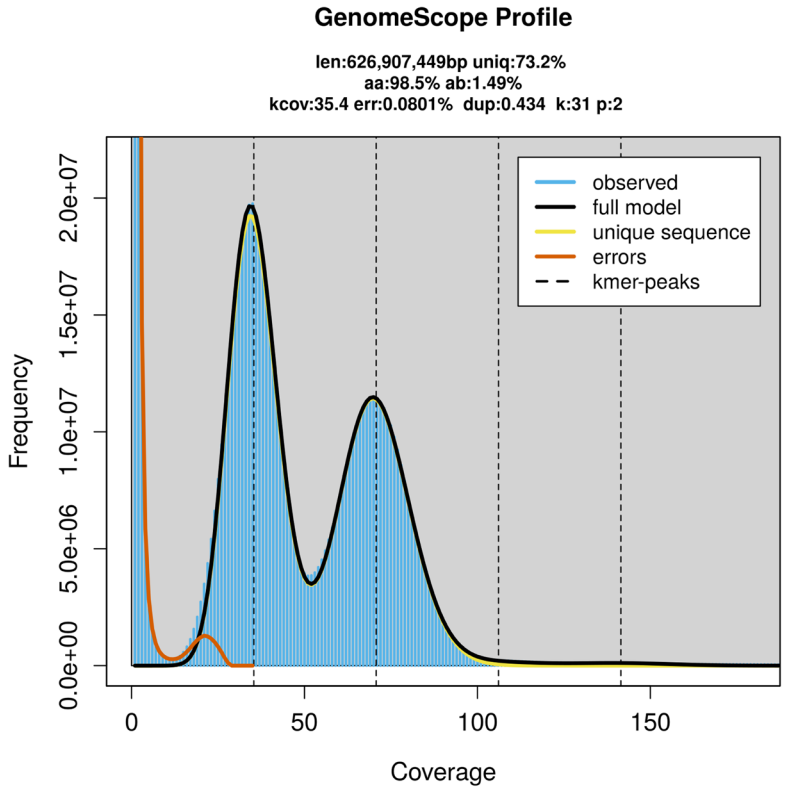

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

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 681.66 Mb in 87 scaffolds, with 218 gaps, and a scaffold N50 of 24.49 Mb (Table 2).

Most of the assembly sequence (99.63%) was assigned to 30 chromosomal-level scaffolds, representing 28 autosomes and the W and Z sex chromosomes. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Scaffolds on Chromosome W from ~3.68–39.21 Mb and Chromosome 24 from ~14.04–14.83 Mb have uncertain order and orientation.

The mitochondrial genome was also assembled. This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

For haplotype 1, the estimated QV is 61.0, and for haplotype 2, 61.2. When the two haplotypes are combined, the assembly achieves an estimated QV of 61.1. The *k*-mer completeness is 76.23% for haplotype 1, 71.05% for haplotype 2, and 99.76% for the combined haplotypes (Figure 4).

BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 99.4% of the expected gene set (single = 98.9%, duplicated = 0.4%) for haplotype 1. The snail plot in Figure 5 summarises the scaffold length distribution

Table 1. Specimen and sequencing data for BioProject PRJEB84099.

Platform	PacBio HiFi	Hi-C
ToLID	ilAphSoci2	ilAphSoci2
Specimen ID	Ox000453	Ox000453
BioSample (source individual)	SAMEA7520671	SAMEA7520671
BioSample (tissue)	SAMEA7520757	SAMEA7520756
Tissue	abdomen	head and thorax
Instrument	Revio	Illumina NovaSeq 6000
Run accessions	ERR14121433; ERR14121434; ERR15170317	ERR14125360
Read count total	4.84 million	700.37 million
Base count total	45.96 Gb	105.76 Gb

Table 2. Genome assembly statistics.

Assembly name	ilAphSoci2.hap1.1	ilAphSoci2.hap2.1
Assembly accession	GCA_965363045.1	GCA_965363025.1
Assembly level	chromosome	scaffold
Span (Mb)	681.66	601.52
Number of chromosomes	30	N/A
Number of contigs	305	212
Contig N50	8.28 Mb	8.83 Mb
Number of scaffolds	87	128
Scaffold N50	24.49 Mb	24.0 Mb
Longest scaffold length (Mb)	42.46	N/A
Sex chromosomes	W and Z	N/A
Organelles	Mitochondrion: 15.38 kb	N/A

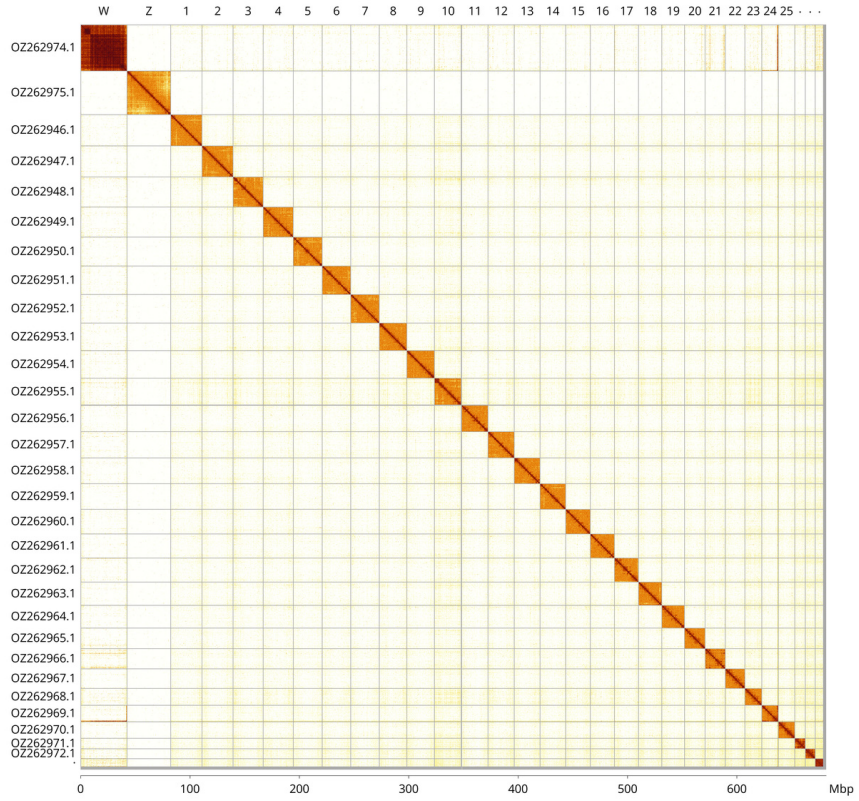

Figure 3. Hi-C contact map of the *Aphomia sociella* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Aphomia sociella* ilAphSoci2.

INSDC accession	Molecule	Length (Mb)	GC%
OZ262946.1	1	28.71	35.50
OZ262947.1	2	28.24	36
OZ262948.1	3	27.57	36
OZ262949.1	4	27.50	35.50
OZ262950.1	5	26.49	35.50
OZ262951.1	6	26.13	35.50
OZ262952.1	7	26.03	35.50
OZ262953.1	8	25.28	35.50
OZ262954.1	9	25.18	36
OZ262955.1	10	24.86	38
OZ262956.1	11	24.19	35.50
OZ262957.1	12	23.96	35.50
OZ262958.1	13	23.52	36
OZ262959.1	14	23.43	36
OZ262960.1	15	22.63	36
OZ262961.1	16	22.03	36
OZ262962.1	17	22.01	36
OZ262963.1	18	21.25	36
OZ262964.1	19	20.86	36.50
OZ262965.1	20	18.74	36.50
OZ262966.1	21	18.49	36.50
OZ262967.1	22	17.86	36.50
OZ262968.1	23	15.52	37
OZ262969.1	24	15.04	37.50
OZ262970.1	25	15.03	37.50
OZ262971.1	26	9.71	38
OZ262972.1	27	9.03	38.50
OZ262973.1	28	7.47	39
OZ262974.1	W	42.46	38.50
OZ262975.1	Z	39.95	35

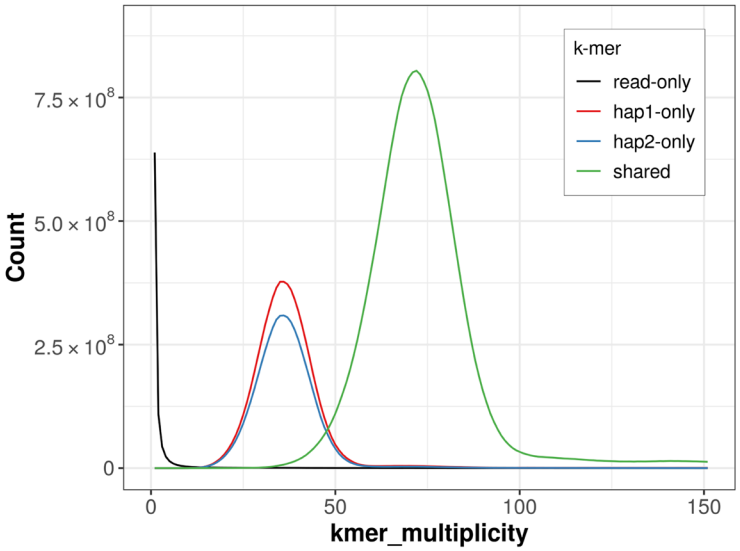

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers 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.

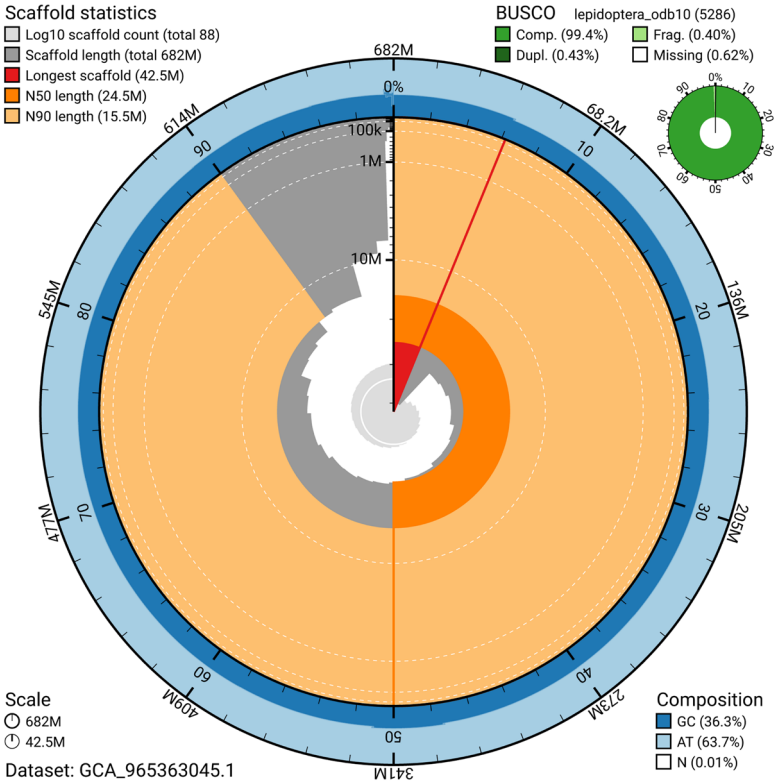

Figure 5. Assembly metrics for ilAphSoci2.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost 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.

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) the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.Q61**, meeting the recommended reference standard.

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

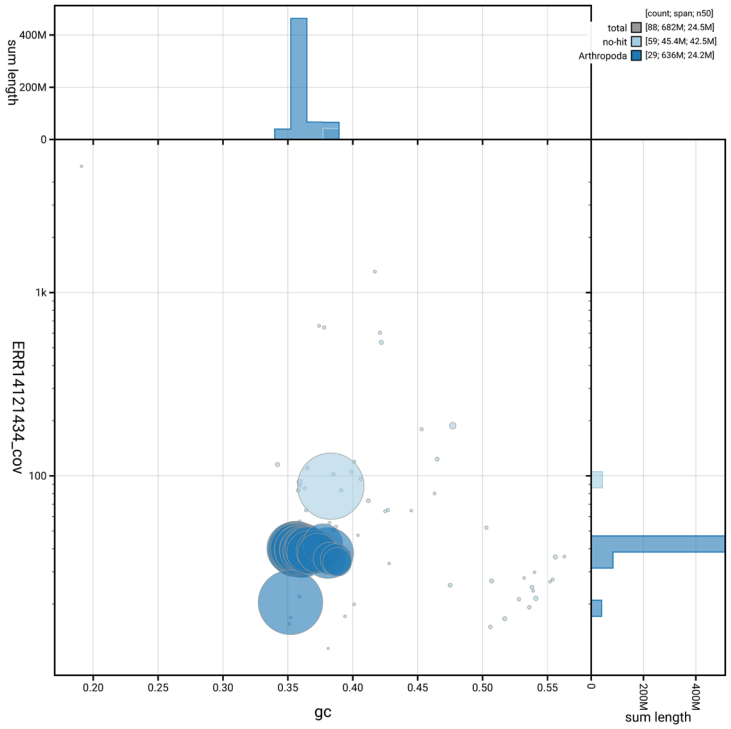

Figure 6. BlobToolKit GC-coverage plot for ilAphSoci2.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Aphomia sociella* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q61	6.C.Q40
Contig N50 length	8.28 Mb	≥ 1 Mb
Scaffold N50 length	24.49 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 61.0; haplotype 2: 61.2; combined: 61.1	≥ 40
k-mer completeness	Haplotype 1: 76.23%; Haplotype 2: 71.05%; combined: 99.76%	≥ 95%
BUSCO	C:99.4% [S:98.9%; D:0.4%]; F:0.4%; M:0.2%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.63%	≥ 90%

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
- 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: Aphomia sociella (bee moth). Accession number PRJEB84099. The genome sequence is released openly for reuse. The *Aphomia sociella* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665), the Sanger Institute Tree of Life Programme (PRJEB43745) and Project Psyche (PRJEB71705). 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 University of Oxford and Wytham Woods 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	5.8.3	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

Software	Version	Source
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	N/A	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
samtools	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.8.0	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	1.2.2	https://github.com/c-zhou/yahs

References

Allio R, Schomaker-Bastos A, Romiguier J, *et al.*: **MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics.** *Mol Ecol Resour.* 2020; **20**(4): 892–905.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Blaxter M, Mieszkowska N, Di Palma F, *et al.*: **Sequence locally, think globally: the Darwin Tree of Life project.** *Proc Natl Acad Sci U S A.* 2022; **119**(4): e2115642118.
PubMed Abstract | Publisher Full Text | Free Full Text

Boyes D: **Natural history of Lepidoptera associated with bird nests in Mid-Wales.** *Entomol Record J Var.* 2018; **130**(5): 249–259.
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

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

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: ***Aphomia sociella* (Linnaeus, 1758)**. 2023.
Reference Source

Goater B: **British pyralid moths.** Colchester: Harley Books, 1986.
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

Kindl J, Kalinová B, Červenka M, *et al.*: **Male moth songs tempt females to accept mating: the role of acoustic and pheromonal communication in the reproductive behaviour of *Aphomia sociella*.** *PLoS One.* 2011; **6**(10): e26476.
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

O'Leary NA, Cox E, Holmes JB, *et al.*: **Exploring and retrieving sequence and metadata for species across the Tree of Life with NCBI Datasets.** *Sci Data.* 2024; **11**(1): 732.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

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**

Schweiger SE, Beyer N, Hass AL, *et al.*: **Pollen and landscape diversity as well as wax moth depredation determine reproductive success of bumblebees in agricultural landscapes.** *Agric Ecosyst Environ.* 2022; **326**: 107788.
Publisher Full Text

Sterling P, Parsons M, Lewington R: **Field guide to the micro moths of Great Britain and Ireland.** Dorset: British Wildlife Publishing, 2023.

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

Waring P, Townsend M, Lewington R: **Field guide to the moths of Great Britain and Ireland.** London, UK: Bloomsbury, 2017.
Reference Source

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 25 October 2025

https://doi.org/10.21956/wellcomeopenres.27514.r136178

✔ **Bin Zhang**
Qingdao Agricultural University, Qingdao, Shandong, China

This data note presents a chromosome-level genome assembly of *Aphomia sociella* (Bee moth) as part of the Darwin Tree of Life Project, which fills a gap in genomic resources for the family Pyralidae (Lepidoptera). The assembly meets key standards of the Earth BioGenome Project (EBP) and provides comprehensive sequencing and assembly details. However, several critical issues related to sample representativeness, genomic detail, and comparative context need to be addressed to enhance the manuscript's scientific rigor and utility for future research.

1. The manuscript notes that segments of Chromosome W (3.68–39.21 Mb) and Chromosome 24 (14.04–14.83 Mb) have "uncertain order and orientation" but provides no explanation for these ambiguities (e.g., low Hi-C contact frequency, repetitive sequences) or plans for future validation (e.g., long-read sequencing of gap regions, FISH). This uncertainty may affect downstream analyses relying on chromosomal structure (e.g., synteny studies).

2. The manuscript mentions that the genome "will be annotated using available RNA-Seq data" but provides no preliminary annotation results (e.g., number of protein-coding genes, functional categories of BUSCO genes) or timelines for annotation release. Without even basic functional insights, the assembly's utility for studying *A. sociella*'s biology (e.g., genes related to host-nest interaction, ultrasonic communication) is severely limited.

3. While the mitochondrial genome length (15.38 kb) is reported, key features (e.g., gene order, GC content, presence of non-coding regions like control regions) are not described. Comparisons with mitochondrial genomes of congeneric (Aphomia) or confamilial (Pyralidae) species are absent, hindering its use in evolutionary and phylogenetic analyses.

4. A scaffold-level assembly of *A. sociella* (GCA_052324605.1, Norwegian University of Life Science) is referenced but not compared to the current chromosome-level assembly. Key differences (e.g., contig N50, BUSCO completeness, chromosome assignment rate) are unreported, making it difficult for readers to assess the novelty and improvement of the current work.

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: Entomology and Bioinformatics

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.