DATA NOTE

The genome sequence of the Orange Beauty, *Commophila aeneana* (Hübner, 1799–1800) (Lepidoptera: Tortricidae)

[version 1; peer review: 3 approved]

David C. Lees[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]Natural History Museum, London, England, UK

v1 **First published:** 08 Sep 2025, **10**:489
https://doi.org/10.12688/wellcomeopenres.24865.1
Latest published: 08 Sep 2025, **10**:489
https://doi.org/10.12688/wellcomeopenres.24865.1

Abstract

We present a genome assembly from a female *Commophila aeneana* (Orange Beauty; Arthropoda; Insecta; Lepidoptera; Tortricidae). The assembly consists of two haplotypes with total lengths of 429.47 megabases and 322.66 megabases. Most of haplotype 1 (99.82%) is scaffolded into 25 chromosomal pseudomolecules, including the W and Z sex chromosomes., and most of haplotype 2 (99.71%) is scaffolded into 23 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 16.5 kilobases.

Keywords

Commophila aeneana; Orange Beauty; Orange Conch; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 1 08 Sep 2025	✓ view	✓ view	✓ view

1. **Nicola J Nadeau**, The University of Sheffield, Sheffield, UK
2. **Jonathan Badger**, National Cancer Institute, Misrata, Libya
National Cancer Institute, Bethesda, USA
3. **Xin Liu**, Beijing Genomics Institute, Beijing, 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: Lees DC: 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: Lees DC, Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding Collective *et al.* **The genome sequence of the Orange Beauty, *Commophila aeneana* (Hübner, 1799–1800) (Lepidoptera: Tortricidae) [version 1; peer review: 3 approved]** Wellcome Open Research 2025, **10**:489 https://doi.org/10.12688/wellcomeopenres.24865.1

First published: 08 Sep 2025, **10**:489 https://doi.org/10.12688/wellcomeopenres.24865.1

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; Apoditrysia; Tortricoidea; Tortricidae; Tortricinae; Cochylini; *Commophila*; *Commophila aeneana* (Hübner, 1799–1800) (NCBI:txid3127293)

Background

Commophila aeneana (Hübner, 1800), also known as the Orange Beauty (Sterling *et al.*, 2023) or Orange Conch, is a moth in the family Tortricidae with a 6–8 mm forewing length (Sterling *et al.*, 2023) or around 13–17 mm wingspan (Bradley *et al.*, 1973). The moth is one of the more distinctive and spectacular of tortricids, its forewing a bright orange yellow in background colour (sometimes fading to whitish) with scattered white scales, the broad median and pre-terminal fasciae being blackish with lines of raised metallic silvery-bluish scales and a similarly adorned basal black fascia tufted thoracically; the hindwing is a dark blackish-brown (Bradley *et al.*, 1973; Sterling *et al.*, 2023). The legs and head are also conspicuously orange. The moth appears to be fairly aposematic (warningly coloured) in its diurnal, buzzing flight.

The adult moth flies from early May (even late April) to late June (Sterling *et al.*, 2023) and can be disturbed by day, otherwise flying spontaneously in the mid-afternoon to early evening sunshine. The larva feeds principally on the roots of Hoary Ragwort, *Jacobaea erucifolia* (L.) G.Gaertn., B.Mey. & Scherb. (1801) (Sterling *et al.*, 2023: 217) with other '*Senecio*' species reported including Common Ragwort (*Jacobaea vulgaris* Gaertn.: Bradley *et al.*, 1973: 70), whilst Meyrick (1928: 498) reported it from Fen Ragwort *Jacobaea paludosa* (L.) G.Gaertn., B.Mey. & Scherb., which was rediscovered in 1972 with a single population near Ely, Cambridgeshire (Plant Atlas, 2020), and is also a known foodplant in Europe (Lepiforum, 2025).

The larva, with brownish head and pale-yellow prothoracic plate and whitish-yellowish abdomen (see Lepiforum, 2025), feeds from September in the rootstock of the foodplant (Bradley *et al.*, 1973). When full-fed in the late autumn, it forms a hibernaculum in the stem some 8 cm from ground level in which it pupates the following spring (Bradley *et al.*, 1973).

The species is very local, occurring mostly in the south-east of the United Kingdom with 81 records on NBN Gateway. It is characterised as Nationally Scarce B, occurring on rough ground, shingle, and other open habitats, predominantly on clay soils (neutral to calcareous). It is also localised in the north-west of Europe (GBIF Secretariat, 2025), where it occurs in Germany, Switzerland, Austria, Belgium, Netherlands, Luxembourg, France and Spain, but also in Romania (Lepiforum, 2025).

The DNA barcode from the mitogenome assembly (OZ205516.1) represents the COI-5P cluster Barcode Index Number (BIN) BOLD:AFT5420, which is sequenced from only UK specimens to date (NHMUK013697066, NHMUK013697090). This provides support for this as an evolutionarily isolated cochyline lineage that is over 9.17% *p*-distant from its nearest neighbours, including *Aethes smeathmanniana* (Fabricius, 1781) (BIN, BOLD:AAB1945) and *Cochylidia* Obraztsov, 1956 species. *C. aeneana* appears not have been included in previous molecular studies and the genome will be helpful in establishing the sister taxon of *Commophila* Razowski, 1999 within Cochylini and for other work on this scarce moth. It needs to be investigated for possible sequestration of pyrrolizidine alkaloids.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult female *Commophila aeneana* (specimen ID NHMUK013697066, ToLID ilComAene1; Figure 1), collected from Bradenham, England, United Kingdom (latitude 51.68, longitude –0.81) on 2022-04-07. The specimen was collected and identified by David Lees (Natural History Museum London).

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. Sample metadata was collected in line with the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

Figure 1. Photographs of the *Commophila aeneana* (ilComAene1) specimen used for genome sequencing.

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 ilComAene1 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 Automated MagAttract v2 protocol. DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor®3 for LI PacBio protocol. 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. For this sample, the final post-shearing DNA had a Qubit concentration of 21.76 ng/μL and a yield of 1 022.72 ng, with a fragment size of 14.1 kb. The 260/280 spectrophotometric ratio was 1.73, and the 260/230 ratio was 1.17.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), following the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

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 of the ilComAene1 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 X.

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 17 breaks and 167 joins. The Z chromosome was identified based on BUSCO gene painting with ancestral Merian elements (Wright *et al.*, 2024). The W chromosome was identified based on read coverage analysis and its single-copy status within a merged diploid map. The exact order and orientation of the contigs on chromosome W (980–47,680 Kbp) and on chromosome19 (4 500–5 600 Kbp) are unknown. 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 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 package management via Conda and Bioconda (Grüning *et al.*, 2018), and containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

Pacific Biosciences single-molecule HiFi sequencing of *Commophila aeneana* generated 41.36 Gb (gigabases) from 3.96 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 373.07 Mb, with a heterozygosity of 0.41% and repeat content of 23.52% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 108× coverage. Hi-C sequencing produced 134.83 Gb from 892.92 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 429.47 Mb in 40 scaffolds, with 178 gaps, and a scaffold N50 of 15.63 Mb (Table 2).

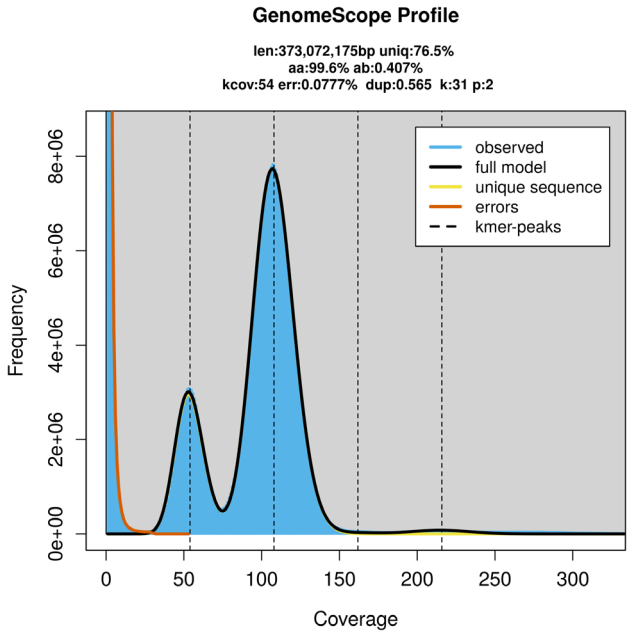

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. **Specimen and sequencing data for BioProject PRJEB82139.**

Platform	PacBio HiFi	Hi-C
ToLID	ilComAene1	ilComAene1
Specimen ID	NHMUK013697066	NHMUK013697066
BioSample (source individual)	SAMEA115574531	SAMEA115574531
BioSample (tissue)	SAMEA115599524	SAMEA115599524
Tissue	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13946447	ERR13947466
Read count total	3.96 million	892.92 million
Base count total	41.36 Gb	134.83 Gb

Table 2. **Genome assembly statistics.**

Assembly name	ilComAene1.hap1.1	ilComAene1.hap2.1
Assembly accession	GCA_964341535.1	GCA_964341585.1
Assembly level	chromosome	chromosome
Span (Mb)	429.47	322.66
Number of chromosomes	25	23
Number of contigs	218	126
Contig N50	7.62 Mb	7.65 Mb
Number of scaffolds	40	57
Scaffold N50	15.63 Mb	14.02 Mb
Longest scaffold length (Mb)	56.34	26.57
Sex chromosomes	W and Z	N/A
Organelles	Mitochondrion: 16.5 kb	N/A

Most of the assembly sequence (99.82%) was assigned to 25 chromosomal-level scaffolds, representing 23 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).

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.

Assembly quality metrics

For haplotype 1, the estimated QV is 62.8, and for haplotype 2, 62.2. When the two haplotypes are combined, the assembly achieves an estimated QV of 62.6. The *k*-mer completeness is 96.13% for haplotype 1, 82.07% for haplotype 2, and 99.06% for the combined haplotypes (Figure 4). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.4% of the expected gene set (single = 98.1%, duplicated = 0.3%) for haplotype 1. The snail plot in Figure 5

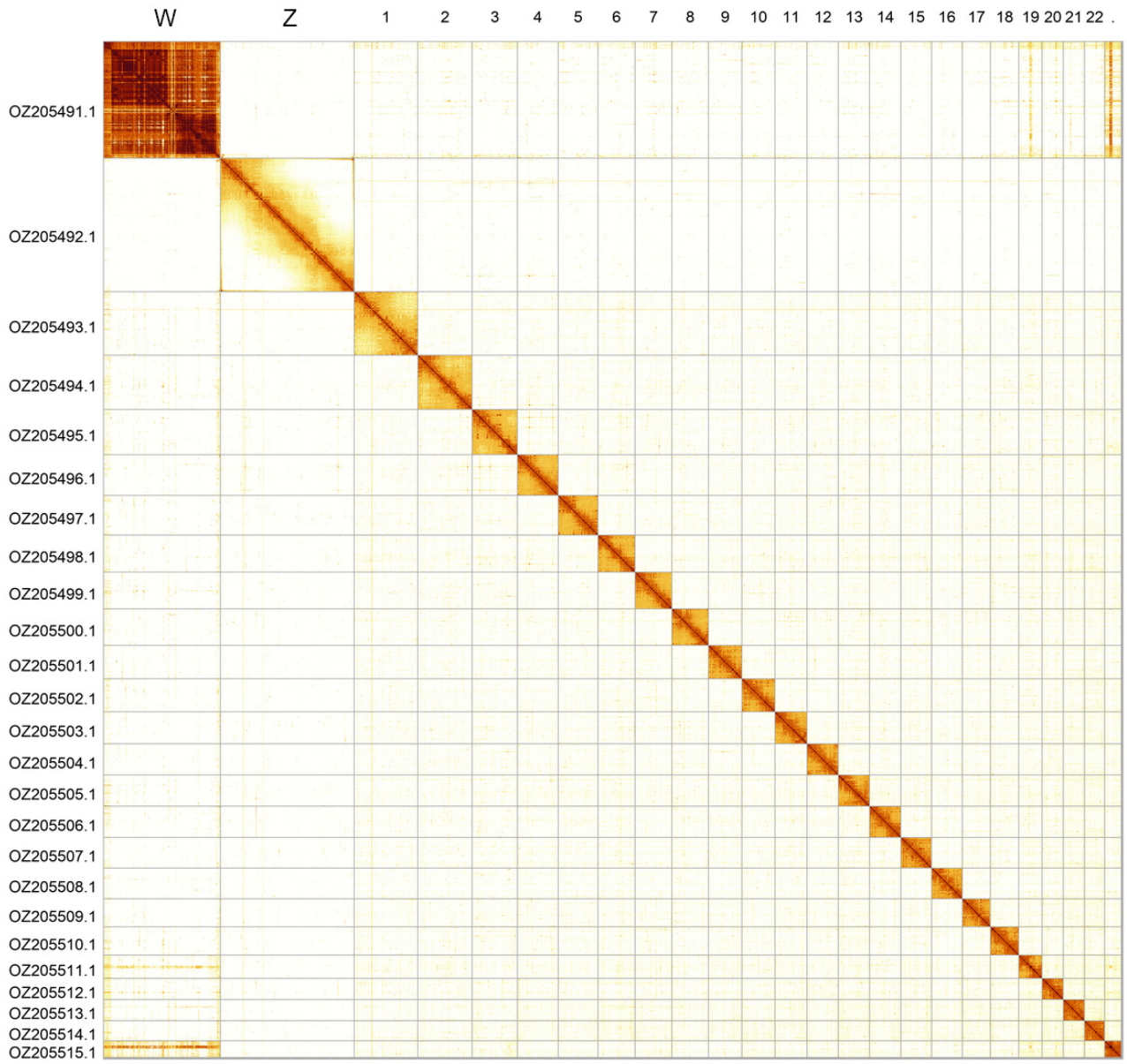

Figure 3. Hi-C contact map of the *Commophila aeneana* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

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) the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.Q62**, 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**

Table 3. Chromosomal pseudomolecules in both haplotypes of the genome assembly of *Calluna vulgaris*, ddCalVulg4.

Haplotype 1				Haplotype 2			
INSDC accession	**Name**	**Length (Mb)**	**GC%**	**INSDC accession**	**Name**	**Length (Mb)**	**GC%**
OZ205493.1	1	26.77	38	OZ205436.1	1	26.57	38
OZ205494.1	2	22.90	38	OZ205437.1	2	22.68	38
OZ205495.1	3	19.09	39	OZ205438.1	3	19.06	39
OZ205496.1	4	17.10	38.50	OZ205439.1	4	17.16	38.50
OZ205497.1	5	16.73	38	OZ205440.1	5	16.43	38
OZ205498.1	6	15.63	38.50	OZ205441.1	6	15.70	38.50
OZ205499.1	7	15.57	38	OZ205442.1	7	15.63	38
OZ205500.1	8	15.39	38	OZ205443.1	8	15.27	38
OZ205501.1	9	14.06	38	OZ205444.1	9	14.02	37.50
OZ205502.1	10	13.92	38	OZ205445.1	10	13.74	38
OZ205503.1	11	13.44	37.50	OZ205446.1	11	13.42	37.50
OZ205504.1	12	13.24	38.50	OZ205447.1	12	13.20	38
OZ205505.1	13	13.16	38	OZ205448.1	13	13.17	38
OZ205506.1	14	13.09	38	OZ205449.1	14	13.11	38.50
OZ205507.1	15	12.95	38	OZ205450.1	15	13.03	38
OZ205508.1	16	12.92	37.50	OZ205451.1	16	12.94	38
OZ205509.1	17	11.90	38	OZ205452.1	17	11.90	38.50
OZ205510.1	18	11.89	38.50	OZ205453.1	18	11.79	38.50
OZ205511.1	19	9.85	40	OZ205454.1	19	11.17	41
OZ205512.1	20	8.92	39	OZ205455.1	20	8.89	39
OZ205513.1	21	8.87	38.50	OZ205456.1	21	8.74	38.50
OZ205514.1	22	8.56	38	OZ205457.1	22	8.51	38
OZ205515.1	23	7.02	39	OZ205458.1	23	5.58	39.50
OZ205492.1	W	49.37	36.50				
OZ205491.1	Z	56.34	37.50				

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

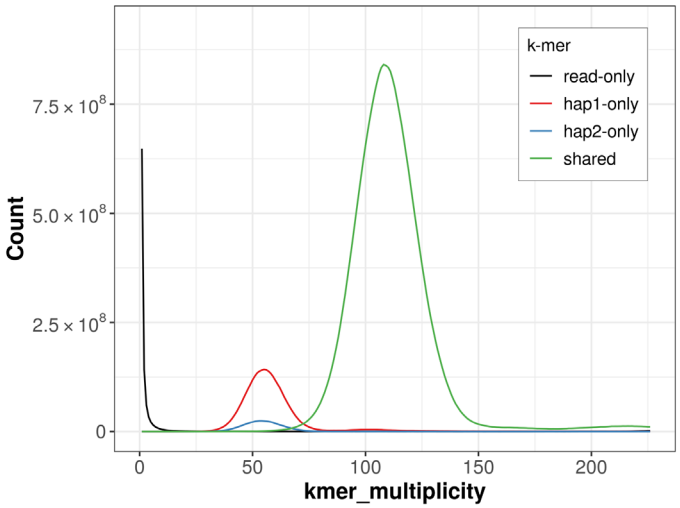

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

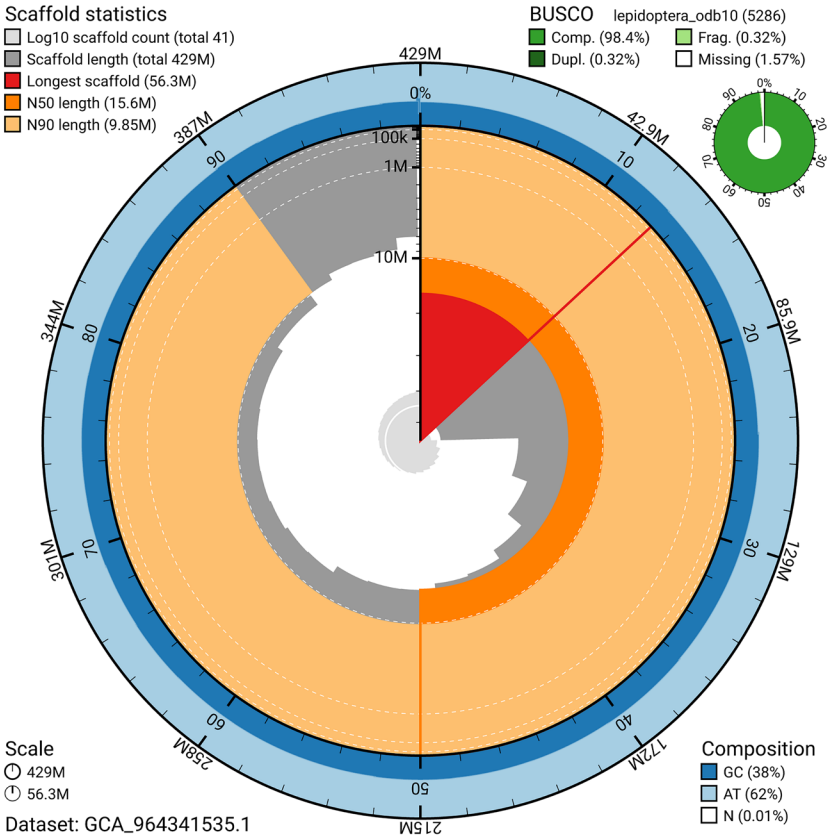

Figure 5. Assembly metrics for ilComAene1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost 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.

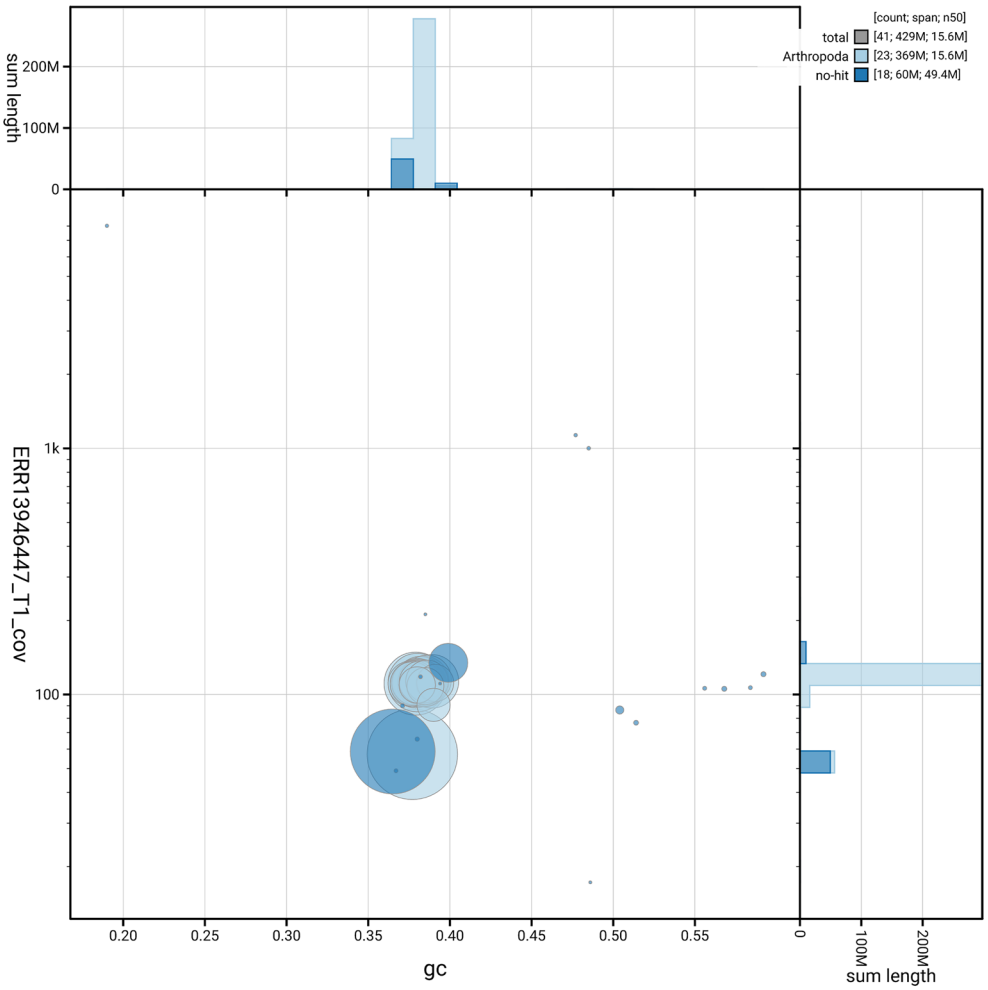

Figure 6. BlobToolKit GC-coverage plot for ilComAene1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Commophila aeneana* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q62	6.C.Q40
Contig N50 length	7.62 Mb	≥ 1 Mb
Scaffold N50 length	15.63 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 62.8; haplotype 2: 62.2; combined: 62.6	≥ 40
k-mer completeness	Haplotype 1: 96.13%; Haplotype 2: 82.07%; combined: 99.06%	≥ 95%
BUSCO	C:98.4% [S:98.1%; D:0.3%]; F:0.3%; M:1.2%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.82%	≥ 90%

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: Commophila aeneana. Accession number PRJEB82139. The genome sequence is released openly for reuse. The *Commophila aeneana* 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.

Pipelines used for genome assembly at the WSI Tree of Life are available at https://pipelines.tol.sanger.ac.uk/pipelines. Table 5 lists software versions used in this study.

Author information

- 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	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	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	2.24-r1122	https://github.com/lh3/minimap2

Software	Version	Source
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.10.0	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	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/ blobtoolkit	0.6.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.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

Bradley JD, Tremewan WG, Smith A: **British tortricoid moths. Cochylidae and tortricidae: tortricinae**. London: Ray Society, 1973.
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): 1332–35.
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: ***Commophila aeneana* (Hübner, 1799–1800)**. 2025.
Reference Source

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–76.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Lepiforum: ***Commophila aeneana* (Hübner, [1800])**. 2025.
Reference Source

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

Meyrick E: **A revised handbook of the British Lepidoptera**. London, 1928.
Reference Source

Plant Atlas: **Fen ragwort.** *Jacobaea paludosa.* (L.) G.Gaertn., B.Mey. & Scherb, 2020.
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

Sterling P, Parsons M, Lewington R: **Field guide to the micro-moths of great Britain and Ireland**. London: Bloomsbury Publishing, 2023.
Reference Source

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

Wright CJ, Stevens L, Mackintosh A, *et al.*: **Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera.** *Nat Ecol Evol.* 2024; **8**(4): 777–90.
PubMed Abstract | Publisher Full Text | Free 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 10 October 2025

https://doi.org/10.21956/wellcomeopenres.27389.r133409

✔ **Xin Liu**
Beijing Genomics Institute, Beijing, China

David C. Lees's article provides a clear description of the genome sequencing, assembly, and evaluation of the Orange Beauty, *Commophila aeneana*. The methodology is well-documented, and the tools and software used are consistent with those commonly applied in similar genomic projects. The resulting genome data are already accessible in public databases. As a data note, I believe this article is already of high quality and presents no major issues. However, I have two suggestions for the author to consider:

1. I noticed a significant size difference between the two haplotypes. Based on the data in Table 3, it appears that the second haplotype assembly may lack the ZW sex chromosomes. The current description of the assembly methods makes it difficult to understand why one haplotype includes both sex chromosomes. I recommend discussing the possible reasons for the length discrepancy between the two haplotypes and, in the sections on chromosome anchoring and final assembly, clarifying how the two sex chromosomes and haplotype data were ultimately integrated into the assembly of haplotype 1.

2. In the Data Availability section, the authors mention RNA-seq and gene annotation. I believe that for a genome dataset to be considered complete—especially for downstream applications—it should include gene annotation results. Even if the available RNA-seq data are not yet comprehensive, I suggest including a preliminary gene annotation in this version of the dataset, following the standard workflows used in similar genomic studies. Further updates to the annotation can be made as more RNA-seq data become available.

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: genomics, 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.

Reviewer Report 06 October 2025

https://doi.org/10.21956/wellcomeopenres.27389.r133404

Jonathan Badger
[1] National Cancer Institute, Misrata, Libya
[2] National Cancer Institute, Bethesda, MD, USA

This is a straightforward and well written genome sequencing manuscript with adequate description of methods.

It might be useful for a reader for a phylogenetic tree of related insects to be included to get a sense of where *C. aeneana* fits in, but I suppose this is not absolutely necessary.

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: I am a computational biologist with a focus on genomics, particularly microbial genomics, although I have also worked in insect genomics.

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.

Reviewer Report 17 September 2025

https://doi.org/10.21956/wellcomeopenres.27389.r132174

Nicola J Nadeau
The University of Sheffield, Sheffield, UK

There is a good level of background information provided on the species. The methods are clear. The genome is clearly of excellent quality, having 2 fully assembled haplotypes and both sex chromosomes.

There are some minor points that could be improved/clarified:

Hi-C library preparation: Were multiple libraries prepared? If so, how many? If a single library then this should be clarified and specific details of the methods for that library included (eg: the number of PCR cycles).

There are some inconsistencies about whether the second haplotype is assembled to chromosome or scaffold level. Most of the text and tables suggest it is assembled to chromosome level but “assembly statistics” text section says scaffold level.

The Fig 6 (blob plot) text is a bit small and hard to read

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: Evolutionary genetics and genomics. Lepidopteran ecology and evolution.

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.