



DATA NOTE

The genome sequence of the Gorse Wanderer, *Brachmia*

blandella (Fabricius, 1798) (Lepidoptera: Gelechiidae)

[version 1; peer review: awaiting peer review]

Douglas Boyes¹, Clare Boyes²,
University of Oxford and Wytham Woods 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

¹UK Centre for Ecology & Hydrology, Wallingford, England, UK

²Independent researcher, Welshpool, Wales, UK

v1 First published: 06 Oct 2025, 10:551
<https://doi.org/10.12688/wellcomeopenres.24971.1>
Latest published: 06 Oct 2025, 10:551
<https://doi.org/10.12688/wellcomeopenres.24971.1>

Open Peer Review

Approval Status Awaiting Peer Review

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

Abstract

We present a genome assembly from an individual female *Brachmia blandella* (Gorse Wanderer, Gorse Crest; Arthropoda; Insecta; Lepidoptera; Gelechiidae). The genome sequence has a total length of 498.99 megabases. Most of the assembly (96.45%) is scaffolded into 31 chromosomal pseudomolecules, including the W and Z sex chromosomes. The mitochondrial genome has also been assembled, with a length of 15.62 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

Brachmia blandella; Gorse Wanderer; Gorse Crest; genome sequence; chromosomal; Lepidoptera



This article is included in the [Tree of Life](#) gateway.

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).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2025 Boyes D *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Boyes D, Boyes C, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Gorse Wanderer, *Brachmia blandella* (Fabricius, 1798) (Lepidoptera: Gelechiidae) [version 1; peer review: awaiting peer review]** Wellcome Open Research 2025, 10:551 <https://doi.org/10.12688/wellcomeopenres.24971.1>

First published: 06 Oct 2025, 10:551 <https://doi.org/10.12688/wellcomeopenres.24971.1>

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; Gelechioidea; Gelechiidae; Dichomeridinae; *Brachmia*; *Brachmia blandella* (Fabricius, 1798) (NCBI:txid1280453)

Background

Brachmia blandella, The Gorse Wanderer, is a micro-moth in the family Gelechiidae. It can be found in a range of habitats including grasslands, gardens, and woodlands. It is common in most of England and is local in Wales (Sterling *et al.*, 2023). It occurs throughout Europe (GBIF Secretariat, 2025).

The moth is very small (forewing length 5–6.5 mm) with a yellowish-brown forewing with some black marks (Sterling *et al.*, 2023). The larvae feed in a slender silken tube, mainly on withered gorse flowers (*Ulex europaeus*), but sometimes on *Cirsium palustre* seedheads, and very occasionally in galls on *Abies grandis* (Emmet & Langmaid, 2002). The larvae pupate in a flimsy cocoon, which incorporates frass and gorse fragments. The moth flies between late June and early August, and readily comes to light (Langmaid *et al.*, 2018).

We present a chromosome-level genome sequence for *Brachmia blandella*, the first high-quality genome for the genus *Brachmia* and one of 24 genomes available for the family Gelechiidae as of September 2025 (data obtained via NCBI datasets, O'Leary *et al.*, 2024). The assembly was produced using the Tree of Life pipeline from a specimen collected in Wytham Woods, Oxfordshire, United Kingdom (Figure 1).

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult female *Brachmia blandella* (specimen ID Ox000649, ToLID ilBraBlan1; Figure 1), collected from Wytham Woods, Oxfordshire, United Kingdom (latitude 51.772, longitude -1.338) on 2020-07-20. The specimen was collected and identified by Douglas Boyes (University of Oxford). A second specimen was used for Hi-C sequencing (specimen ID NHMUK014584841,

ToLID ilBraBlan2). It was collected from the Natural History Museum Wildlife Garden, London, UK (latitude 51.5, longitude -0.18) on 2022-06-23. The specimen was collected and identified by David Lees. 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 ilBraBlan1 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 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.

HMW DNA was extracted in the WSI Scientific Operations core 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



Figure 1. Photograph of the *Brachmia blandella* (ilBraBlan1) specimen used for genome sequencing.

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 μ L.

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 using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was in the range 40–135 pM. The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, and to perform primary and secondary analysis of the data upon completion.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the ilBraBlan2 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 Diagenode 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/ μ L. 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 (Cheng *et al.*, 2021) with the --primary option. The Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019), and the contigs were scaffolded in YaHS (Zhou *et al.*, 2023) with 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 141 breaks and 189 joins. The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation>. PretextViewSnapshot 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 the primary and alternate 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 *Brachmia blandella* specimen generated 26.80 Gb (gigabases) from 2.83 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 406.50 Mb, with a heterozygosity of 1.82% and repeat content of 33.53% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 42x coverage. Hi-C sequencing produced 118.87 Gb from 787.23 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

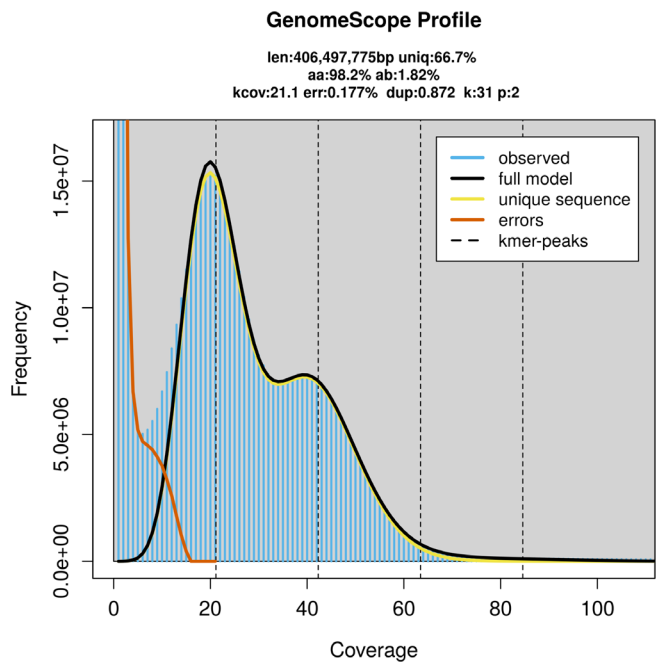


Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB71264.

Platform	PacBio HiFi	Hi-C
ToLID	ilBraBlan1	ilBraBlan2
Specimen ID	Ox000649	NHMUK014584841
BioSample (source individual)	SAMEA7701511	SAMEA115574999
BioSample (tissue)	SAMEA7701693	SAMEA115575154
Tissue	whole organism	whole organism
Instrument	Sequel IIe	Illumina NovaSeq X
Run accessions	ERR12370409	ERR13621414
Read count total	2.83 million	787.23 million
Base count total	26.80 Gb	118.87 Gb

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The final assembly has a total length of 498.99 Mb in 500 scaffolds, with 964 gaps, and a scaffold N50 of 16.58 Mb (Table 2).

Most of the assembly sequence (96.45%) was assigned to 31 chromosomal-level scaffolds, representing 29 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.

The combined primary and alternate assemblies achieve an estimated QV of 57.5. The *k*-mer completeness is 75.83% for the primary assembly, 69.04% for the alternate haplotype, and 98.48% for the combined assemblies (Figure 4).

BUSCO v.5.7.1 analysis using the lepidoptera_odb10 reference set (*n* = 5 286) identified 96.9% of the expected gene set (single = 94.8%, duplicated = 2.1%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for the primary assembly. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage.

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 primary assembly, is 5.C.Q56.

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
- 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.

Table 2. Genome assembly statistics.

Assembly name	ilBraBlan1.1
Assembly accession	GCA_965199015.1
Alternate haplotype accession	GCA_965198705.1
Assembly level	chromosome
Span (Mb)	498.99
Number of chromosomes	31
Number of contigs	1 464
Contig N50	0.76 Mb
Number of scaffolds	500
Scaffold N50	16.58 Mb
Sex chromosomes	W and Z
Organelles	Mitochondrion: 15.62 kb

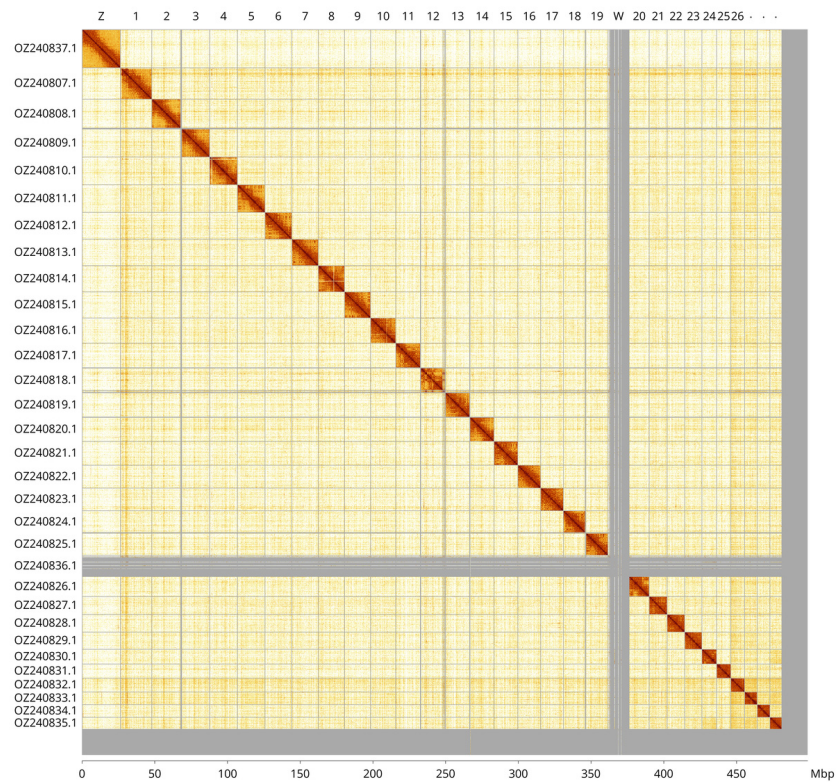


Figure 3. Hi-C contact map of the *Brachmia blandella* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Brachmia blandella* ilBraBlan1.

INSDC accession	Molecule	Length (Mb)	GC%
OZ240807.1	1	21.61	36.50
OZ240808.1	2	20.55	36.50
OZ240809.1	3	19.27	36.50
OZ240810.1	4	19.11	36.50
OZ240811.1	5	18.94	36.50
OZ240812.1	6	18.39	36.50
OZ240813.1	7	18.30	36
OZ240814.1	8	17.97	36.50
OZ240815.1	9	17.91	36
OZ240816.1	10	17.38	36.50
OZ240817.1	11	17.10	36
OZ240818.1	12	16.99	36.50
OZ240819.1	13	16.89	36
OZ240820.1	14	16.58	36.50

INSDC accession	Molecule	Length (Mb)	GC%
OZ240821.1	15	16.29	36
OZ240822.1	16	15.73	36.50
OZ240823.1	17	15.62	36
OZ240824.1	18	15.45	36.50
OZ240825.1	19	15.27	37
OZ240826.1	20	13.79	37
OZ240827.1	21	12.28	36.50
OZ240828.1	22	12.06	36.50
OZ240829.1	23	11.93	37
OZ240830.1	24	10.15	36.50
OZ240831.1	25	9.62	36.50
OZ240832.1	26	9.50	38
OZ240833.1	27	8.78	38.50
OZ240834.1	28	8.74	37
OZ240835.1	29	8.18	37.50
OZ240836.1	W	14.52	37
OZ240837.1	Z	26.39	36

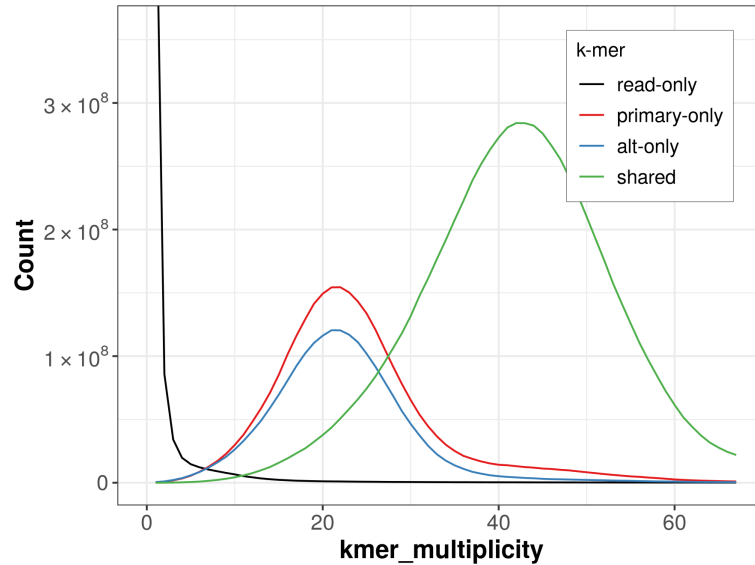


Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

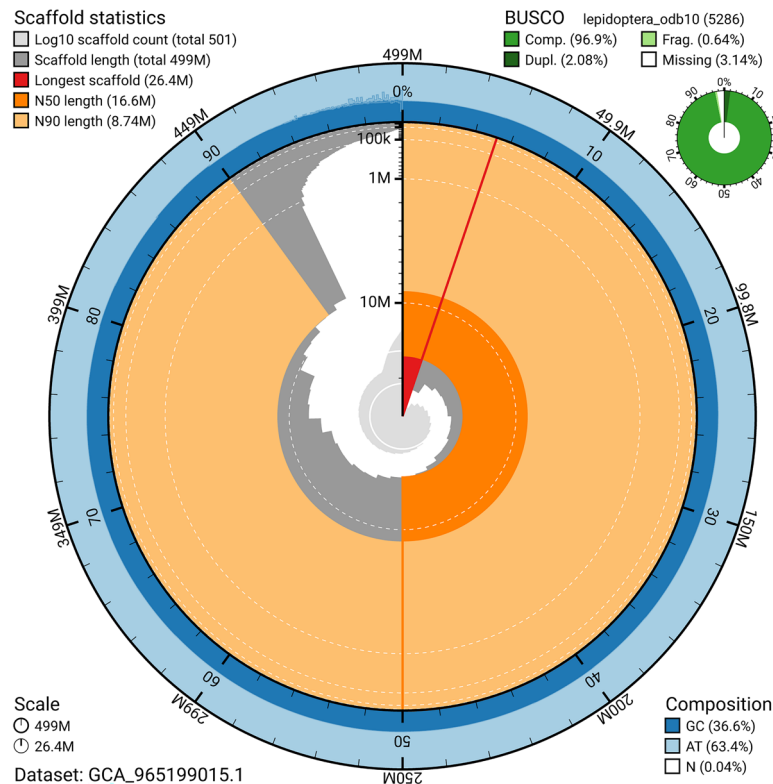


Figure 5. Assembly metrics for ilBraBlan1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue 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 lepidoptera_odb10 set is presented at the top right. An interactive version of this figure can be accessed on the [BlobToolKit viewer](#).

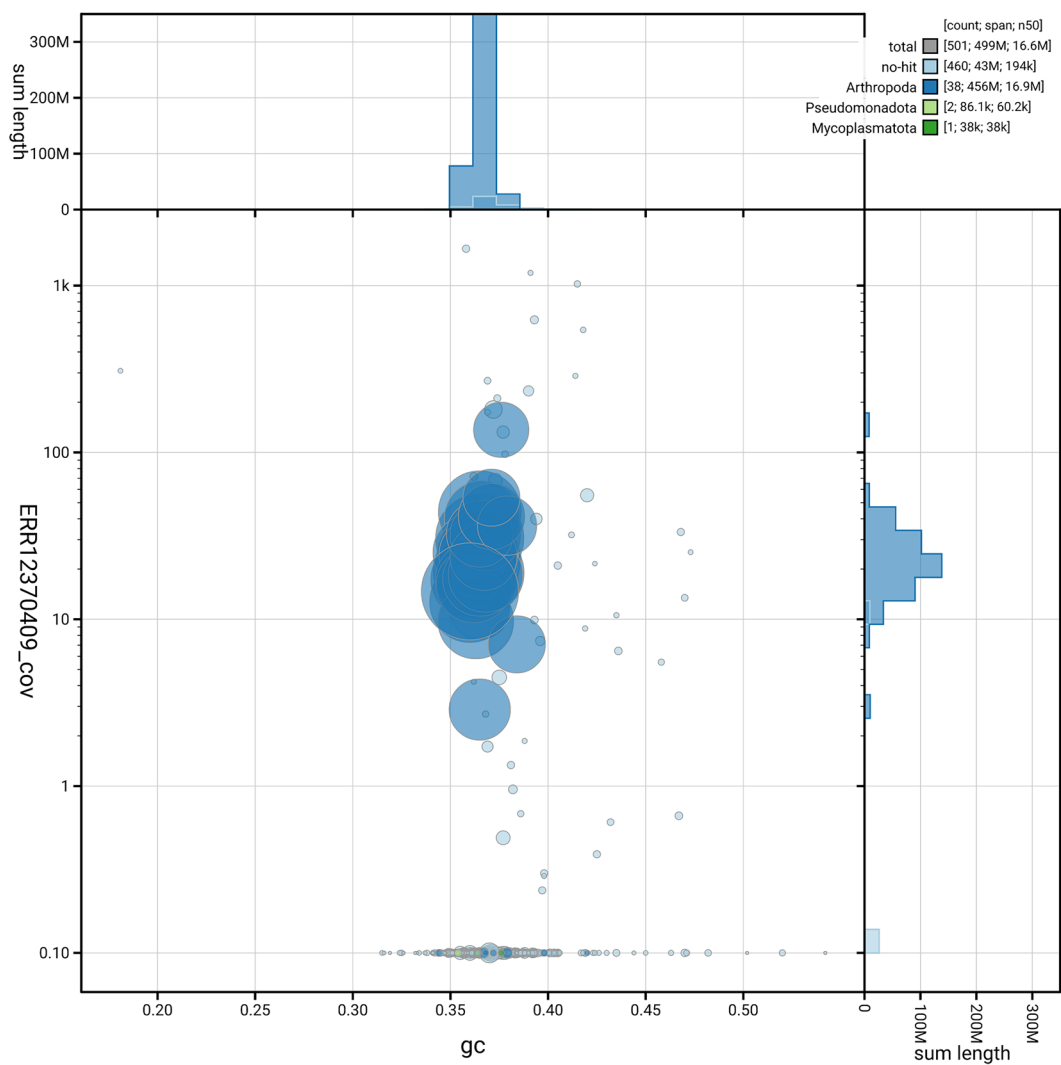


Figure 6. BlobToolKit GC-coverage plot for ilBraBlan1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display 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 *Brachmia blandella* assembly.

Measure	Value	Benchmark
EBP summary (primary)	5.C.Q56	6.C.Q40
Contig N50 length	0.76 Mb	≥ 1 Mb
Scaffold N50 length	16.58 Mb	= chromosome N50
Consensus quality (QV)	Primary: 56.8; alternate: 58.0; combined: 57.5	≥ 40
<i>k</i> -mer completeness	Primary: 75.83%; alternate: 69.04%; combined: 98.48%	≥ 95%
BUSCO	C:96.9% [S:94.8%; D:2.1%]; F:0.6%; M:2.5%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	96.45%	≥ 90%

Data availability

European Nucleotide Archive: *Brachmia blandella* (gorse crest). Accession number [PRJEB71264](#). The genome sequence is released openly for reuse. The *Brachmia blandella* 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	4.4.5	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.7.1	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/chhy123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MercuryFK	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	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

Software	Version	Source
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.7.1	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

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](#)

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](#)

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](#)

Emmet AM, Langmaid JR: **The moths and butterflies of Great Britain and Ireland.** Gelechiidae. Colchester: Harley Books, 2002. [Reference Source](#)

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: ***Brachmia blandella* (Fabricius, 1798) in GBIF Backbone Taxonomy.** 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): g1aa153. [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](#)

Langmaid JR, Palmer SM, Young M: **A field guide to the smaller moths of Great Britain and Ireland.** Reading: The British Entomological; Natural History Society, 2018. [Reference Source](#)

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, Seppely 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.*: **Merqurey: 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.** Dorset: British Wildlife 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, Krashenninnikova 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](#)