



DATA NOTE

The genome sequence of the Maple Pug moth, *Eupithecia inturbata* (Hübner, 1817)

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

Finley Hutchinson¹, Liam M. Crowley^{id}²,
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

¹University of Exeter, Penryn, England, UK

²University of Oxford, Oxford, England, UK

V1 First published: 18 Mar 2025, 10:141
<https://doi.org/10.12688/wellcomeopenres.23796.1>

Latest published: 18 Mar 2025, 10:141
<https://doi.org/10.12688/wellcomeopenres.23796.1>

Abstract

We present a genome assembly from a female *Eupithecia inturbata* (Maple Pug; Arthropoda; Insecta; Lepidoptera; Geometridae). The genome sequence has a total length of 427.76 megabases. Most of the assembly (99.94%) is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled, with a length of 15.33 kilobases. Gene annotation of this assembly on Ensembl identified 12,386 protein-coding genes.

Keywords

Eupithecia inturbata, Maple Pug moth, genome sequence, chromosomal, Lepidoptera



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

Open Peer Review

Approval Status ✓ ? ✓ ?

	1	2	3	4
version 1	✓	?	✓	?
18 Mar 2025	view	view	view	view

1. **Sivasankaran Kuppasamy** ^{id}, Loyola College, Chennai, India
2. **Annabel Whibley** ^{id}, The University of Auckland, Auckland, New Zealand
3. **Panagiotis Ioannidis** ^{id}, Foundation for Research & Technology - Hellas, Crete, Greece
4. **Maríndia Deprá**, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

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: Hutchinson F: Investigation, Resources; Crowley LM: Investigation;

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>].

Copyright: © 2025 Hutchinson F *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: Hutchinson F, Crowley LM, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Maple Pug moth, *Eupithecia inturbata* (Hübner, 1817) [version 1; peer review: 2 approved, 2 approved with reservations]** Wellcome Open Research 2025, **10**:141 <https://doi.org/10.12688/wellcomeopenres.23796.1>

First published: 18 Mar 2025, **10**:141 <https://doi.org/10.12688/wellcomeopenres.23796.1>

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; Geometroidea; Geometridae; Larentiinae; *Eupithecia*; *Eupithecia inturbata* (Hübner) (NCBI:txid934853)

Background

The genome of the Maple Pug, *Eupithecia inturbata*, was sequenced as part of the Darwin Tree of Life Project, a collaborative effort to sequence all named eukaryotic species in the Atlantic Archipelago of Britain and Ireland. Here we present a chromosome-level genome sequence for *Eupithecia inturbata*, based on a female specimen from Wytham Woods, Oxfordshire, United Kingdom (Figure 1).

Genome sequence report

Sequencing data

The genome of a specimen of *Eupithecia inturbata* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 22.05 Gb from 3.09 million reads. GenomeScope analysis of the PacBio HiFi data estimated the haploid genome size at 422.93 Mb, with a heterozygosity of 0.47% and repeat content of 23.00%. These values provide an initial assessment of genome complexity and the challenges anticipated during assembly. Based on this estimated genome size, the sequencing data provided approximately 50.0x coverage of the genome. Chromosome conformation Hi-C sequencing produced 110.68 Gb from 732.96 million reads.

Table 1 summarises the specimen and sequencing information, including the BioProject, study name, BioSample numbers, and sequencing data for each technology.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The assembly was improved by manual curation, which corrected 14 misjoins or missing joins and removed

7 haplotypic duplications. These interventions decreased the scaffold count by 18.18%. The final assembly has a total length of 427.76 Mb in 35 scaffolds, with 53 gaps, and a scaffold N50 of 14.97 Mb (Table 2).

The snail plot in Figure 2 provides a summary of the assembly statistics, indicating the distribution of scaffold lengths and other assembly metrics. Figure 3 shows the distribution of scaffolds by GC proportion and coverage. Figure 4 presents a cumulative assembly plot, with separate curves representing different scaffold subsets assigned to various phyla, illustrating the completeness of the assembly.

Most of the assembly sequence (99.95%) was assigned to 31 chromosomal-level scaffolds, representing 30 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3). The specimen is heterogametic, therefore female. During curation, chromosome Z was assigned based on read coverage statistics, but no W chromosome could be identified.

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 in GenBank.

Assembly quality metrics

The estimated Quality Value (QV) and *k*-mer completeness metrics, along with BUSCO completeness scores, were calculated for each haplotype and the combined assembly. The QV reflects the base-level accuracy of the assembly, while *k*-mer completeness indicates the proportion of expected *k*-mers identified in the assembly. BUSCO scores provide a measure of completeness based on benchmarking universal single-copy orthologues.

The primary haplotype has a QV of 62.7, and the combined primary and alternate assemblies achieve an estimated QV of 62.2. The *k*-mer completeness for the primary haplotype is 90.00%, and for the alternate haplotype it is 83.95%. The combined primary and alternate assemblies achieve a *k*-mer completeness of 99.23%. BUSCO analysis using the lepidoptera_odb10 reference set ($n = 5,286$) gave a completeness score of 98.1% (single = 97.6%, duplicated = 0.6%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The assembly achieves the EBP reference standard of 6.C.Q62.

Genome annotation report

The *Eupithecia inturbata* genome assembly (GCA_963662085.1) was annotated at the European Bioinformatics Institute (EBI) on Ensembl Rapid Release. The resulting annotation includes 23,139 transcribed mRNAs from 12,386 protein-coding and 1,799 non-coding genes (Table 2; https://rapid.ensembl.org/Eupithecia_inturbata_GCA_963662085.1/Info/Index). The average transcript length is 15,332.34. There are 1.63 coding transcripts per gene and 7.30 exons per transcript.



Figure 1. Photograph of the *Eupithecia inturbata* (ilEupIntu1) specimen used for genome sequencing.

Table 1. Specimen and sequencing data for *Eupithecia inturbata*.

Project information			
Study title	Eupithecia inturbata (maple pug)		
Umbrella BioProject	PRJEB67611		
Species	<i>Eupithecia inturbata</i>		
BioSpecimen	SAMEA113425663		
NCBI taxonomy ID	934853		
Specimen information			
Technology	ToLID	BioSample accession	Organism part
PacBio long read sequencing	ilEupIntu1	SAMEA113426987	whole organism
Hi-C sequencing	ilEupIntu1	SAMEA113426987	whole organism
Sequencing information			
Platform	Run accession	Read count	Base count (Gb)
Hi-C Illumina NovaSeq 6000	ERR12144000	7.33e+08	110.68
PacBio Revio	ERR12205265	3.09e+06	22.05

Table 2. Genome assembly data for *Eupithecia inturbata*.

Genome assembly		
Assembly name	ilEupIntu1.1	
Assembly accession	GCA_963662085.1	
Alternate haplotype accession	GCA_963662135.1	
Assembly level for primary assembly	chromosome	
Span (Mb)	427.76	
Number of contigs	88	
Number of scaffolds	35	
Longest scaffold (Mb)	20.78	
Assembly metric	Measure	Benchmark
Contig N50 length	9.91 Mb	≥ 1 Mb
Scaffold N50 length	14.97 Mb	= chromosome N50
Consensus quality (QV)	Primary: 62.7; alternate: 61.8; combined 62.2	≥ 40
k-mer completeness	Primary: 90.00%; alternate: 83.95%; combined: 99.23%	≥ 95%
BUSCO*	C:98.1%[S:97.6%,D:0.6%], F:0.3%,M:1.5%,n:5,286	S > 90%; D < 5%
Percentage of assembly mapped to chromosomes	99.95%	≥ 90%
Sex chromosomes	Z	localised homologous pairs
Organelles	Mitochondrial genome: 15.33 kb	complete single alleles
Genome annotation of assembly GCA_963662085.1 at Ensembl		
Number of protein-coding genes	12,386	
Number of non-coding genes	1,799	
Number of gene transcripts	23,139	

* BUSCO scores based on the lepidoptera_odb10 BUSCO set using version 5.5.0. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison.

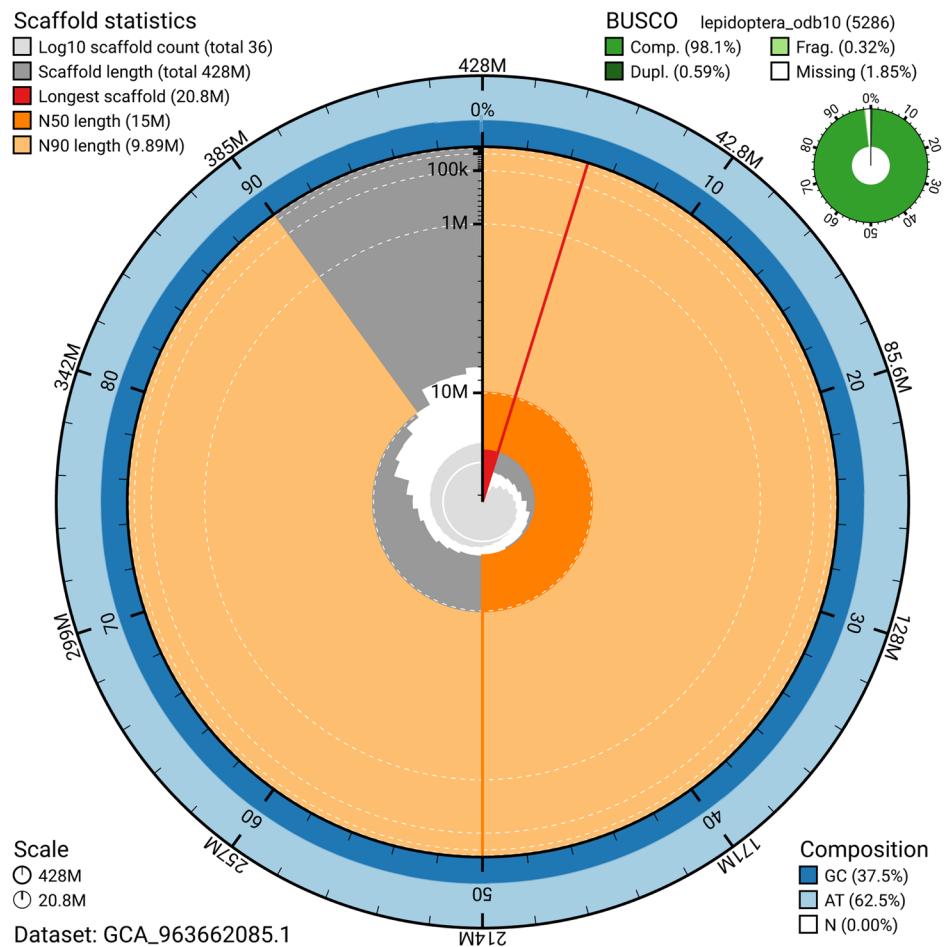


Figure 2. Genome assembly of *Eupithecia inturbata*, ilEupIntu1.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. 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 is available at https://blobtoolkit.genomehubs.org/view/GCA_963662085.1/dataset/GCA_963662085.1/snail.

Methods

Sample acquisition and DNA barcoding

An adult male *Eupithecia inturbata* (specimen ID Ox003039, ToLID ilEupIntu1) was collected from Wytham Woods, Oxfordshire, United Kingdom (latitude 51.77, longitude -1.34) on 2022-07-22 by light trap. The specimen was collected by Finley Hutchinson and Liam Crowley (University of Oxford), identified by Finley Hutchinson and preserved on dry ice.

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) (Pereira *et al.*, 2022). 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 have been deposited on protocols.io (Beasley *et al.*, 2023).

Metadata collection for samples adhered to the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of procedures: sample preparation and homogenisation, DNA extraction, fragmentation and

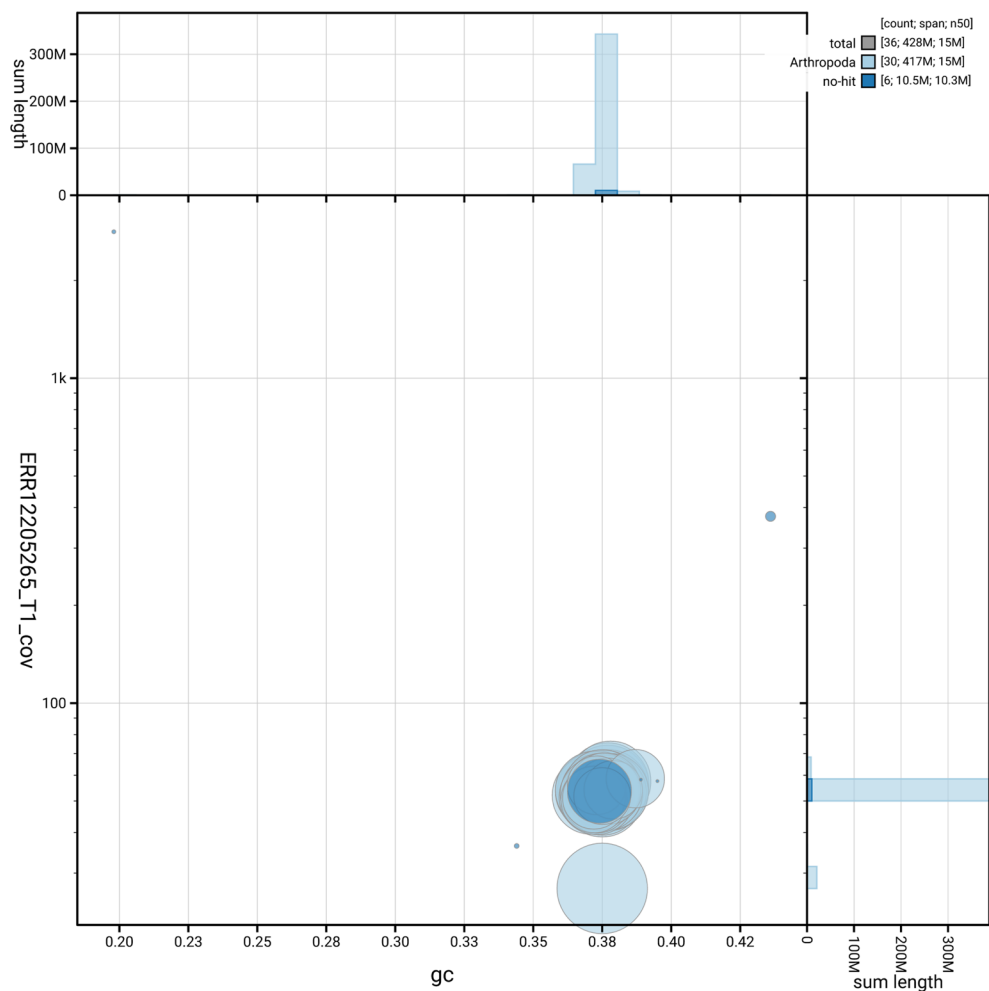


Figure 3. Genome assembly of *Eupithecia inturbata*, ilEupIntu1.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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 at https://blobtoolkit.genomehubs.org/view/GCA_963662085.1/blob.

purification. Detailed protocols are available on protocols.io (Denton *et al.*, 2023b). The ilEupIntu1 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the whole organism was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a).

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023). The DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Bates *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Strickland *et al.*, 2023). 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.

Hi-C sample preparation

Tissue from the whole organism of the ilEupIntu1 sample was processed for Hi-C sequencing at the WSI Scientific Operations core, using the Arima-HiC v2 kit. In brief, 20–50 mg of frozen tissue (stored at -80°C) was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde concentration. After crosslinking, the tissue was homogenised using the Diagenode Power Masher-II and BioMasher-II tubes and pestles. Following the Arima-HiC v2 kit manufacturer's instructions, crosslinked DNA was digested using a restriction enzyme master mix. The 5'-overhangs were filled in and labelled with biotinylated nucleotides and proximally ligated.

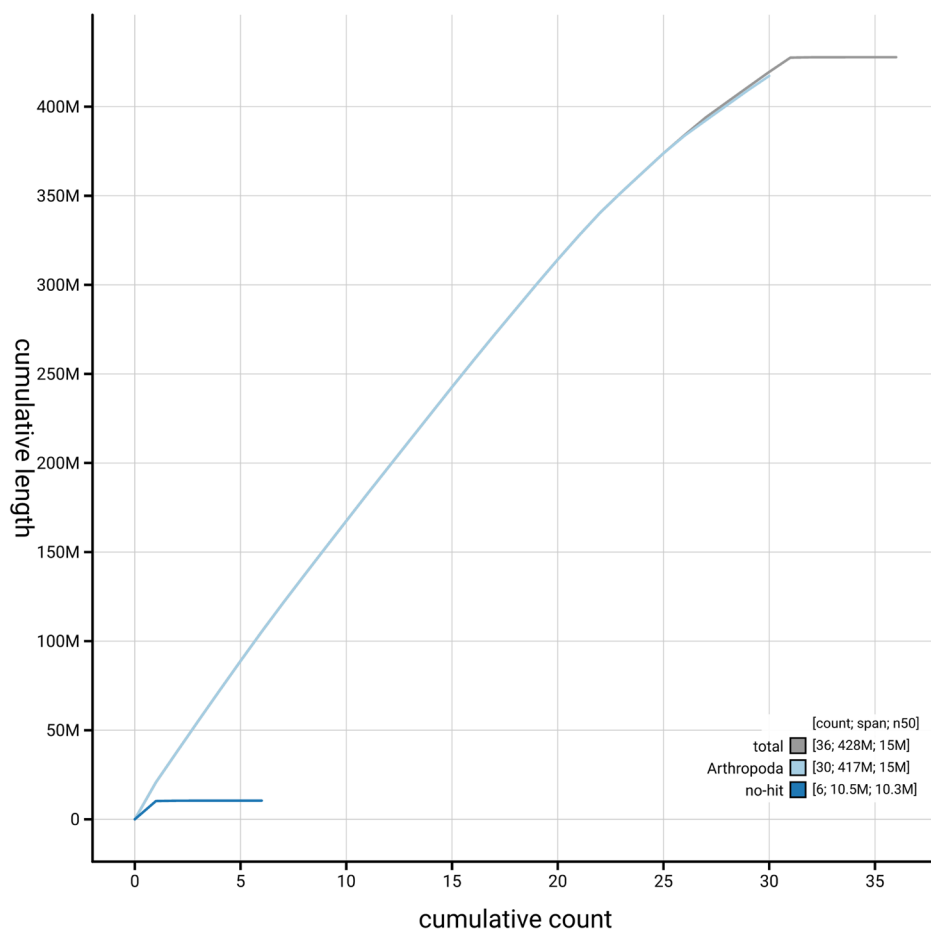


Figure 4. Genome assembly of *Eupithecia inturbata*, ilEupIntu1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963662085.1/dataset/GCA_963662085.1/cumulative.

An overnight incubation was carried out for enzymes to digest remaining proteins and for crosslinks to reverse. A clean up was performed with SPRIselect beads prior to library preparation. Additionally, the biotinylation percentage was estimated using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit and Arima-HiC v2 QC beads.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core.

PacBio HiFi

At a minimum, samples were required to have an average fragment size exceeding 8 kb and a total mass over 400 ng to proceed to the low input SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences, California, USA), depending on genome size and sequencing depth required. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA) as per the manufacturer's instructions. The kit includes the reagents required for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead cleanup, and nuclease treatment.

Following the manufacturer's instructions, size selection and clean up was carried out using diluted AMPure PB beads (Pacific Biosciences, California, USA). DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) with Qubit 1X dsDNA HS assay kit and the final library fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and gDNA 55kb BAC analysis kit.

Samples were 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, as well as perform primary and secondary analysis of the data upon completion.

Samples were sequenced on a Revio instrument (Pacific Biosciences, California, USA). Prepared libraries were normalised to 2 nM, and 15 µL was used for making complexes. Primers were annealed and polymerases were hybridised to create

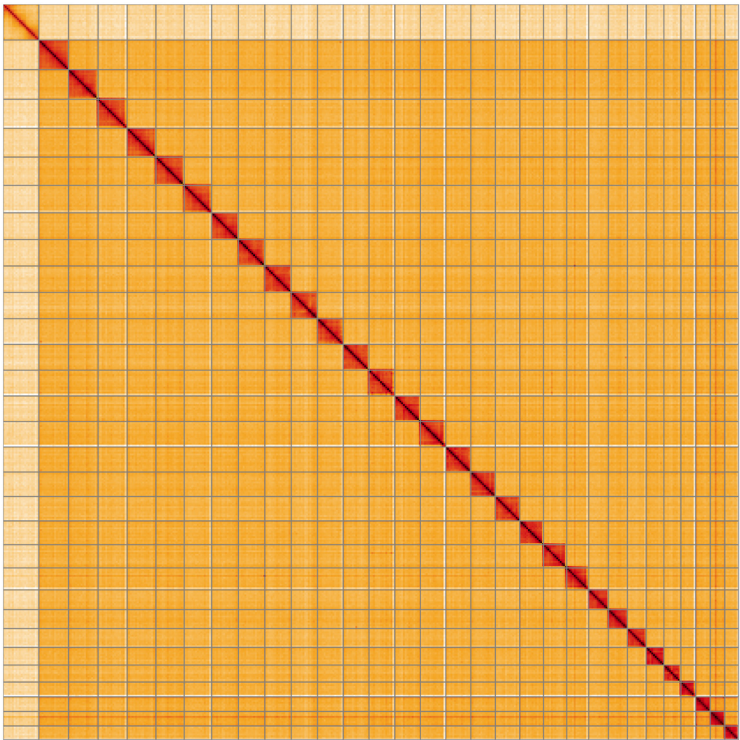


Figure 5. Genome assembly of *Eupithecia inturbata*: Hi-C contact map of the ilEupIntu1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at <https://genome-note-higlass.tol.sanger.ac.uk/l/?d=FFENAPbrRG-iKuz10tLzmA>.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Eupithecia inturbata*, ilEupIntu1.

INSDC accession	Name	Length (Mb)	GC%
OY759177.1	1	17.26	37.5
OY759178.1	2	17.15	37.5
OY759179.1	3	16.94	37.5
OY759180.1	4	16.73	37.5
OY759181.1	5	16.45	38
OY759182.1	6	15.94	37.5
OY759183.1	7	15.52	37.5
OY759184.1	8	15.39	37.5
OY759185.1	9	15.28	37.5
OY759186.1	10	15.27	37
OY759187.1	11	15.0	37
OY759188.1	12	14.99	37
OY759189.1	13	14.97	37.5
OY759190.1	14	14.95	37.5
OY759191.1	15	14.73	37.5

INSDC accession	Name	Length (Mb)	GC%
OY759192.1	16	14.53	38
OY759193.1	17	14.23	37.5
OY759194.1	18	14.21	37.5
OY759195.1	19	13.88	37.5
OY759196.1	20	13.42	37.5
OY759197.1	21	12.82	37.5
OY759198.1	22	11.42	37.5
OY759199.1	23	11.01	37.5
OY759200.1	24	10.96	37
OY759201.1	25	10.25	37.5
OY759202.1	26	9.89	37
OY759203.1	27	8.57	37.5
OY759204.1	28	8.51	38
OY759205.1	29	8.46	38.5
OY759206.1	30	8.01	37.5
OY759176.1	Z	20.78	37.5
OY759207.1	MT	0.02	20

circularised complexes according to manufacturer's instructions. The complexes were purified with the 1.2X clean up with SMRTbell beads. The purified complexes were then diluted to the Revio loading concentration (in the range 200–300 pM), and spiked with a Revio sequencing internal control. Samples were sequenced on Revio 25M SMRT cells (Pacific Biosciences, California, USA). The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, as well as perform primary and secondary analysis of the data upon completion.

Hi-C

For Hi-C library preparation, DNA was fragmented using the Covaris E220 sonicator (Covaris) and size selected using SPRISelect beads to 400 to 600 bp. The DNA was then enriched using the Arima-HiC v2 kit Enrichment beads. Using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) for end repair, A-tailing, and adapter ligation. This uses a custom protocol which resembles the standard NEBNext Ultra II DNA Library Prep protocol but where library preparation occurs while DNA is bound to the Enrichment beads. For library amplification, 10 to 16 PCR cycles were required, determined by the sample biotinylation percentage. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on an Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

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 first assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). The Hi-C reads were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). The contigs were further scaffolded using the provided Hi-C data (Rao *et al.*, 2014) 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 (article in preparation). Flat files and maps used in curation were generated in TreeVal (Pointon *et al.*, 2023). Manual curation was primarily conducted using PretextView (Harry, 2022), with

additional insights provided by JBrowse2 (Diesh *et al.*, 2023) and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were corrected, and duplicate sequences were tagged and removed. Sex chromosomes were identified by read coverage statistics. The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation> (article in preparation).

Assembly quality assessment

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

A Hi-C contact map was produced for the final version of the assembly. The Hi-C reads were aligned using bwa-mem2 (Vasimuddin *et al.*, 2019) and the alignment files were combined using SAMtools (Danecek *et al.*, 2021). The Hi-C alignments were converted into a contact map using BEDTools (Quinlan & Hall, 2010) and the Cooler tool suite (Abdennur & Mirny, 2020). The contact map is visualised in HiGlass (Kerpedjiev *et al.*, 2018).

The blobtoolkit pipeline is a Nextflow (Di Tommaso *et al.*, 2017) port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads in SAMtools and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoaT database (Challis *et al.*, 2023) to identify all matching BUSCO lineages to run BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineages, the pipeline aligns the BUSCO genes to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) with DIAMOND blastp (Buchfink *et al.*, 2021). The genome is also divided into chunks according to the density of the BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database using DIAMOND blastx. Genome sequences without a hit are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The blobtools suite combines all these 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), relying on the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), as well as the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

Table 4 contains a list of relevant software tool versions and sources.

Genome annotation

The Ensembl Genebuild annotation system (Aken *et al.*, 2016) was used to generate annotation for the *Eupithecia inturbata* assembly (GCA_963662085.1) in Ensembl Rapid Release at

Table 4. Software tools: versions and sources.

Software tool	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	427104ea91c78c3b8b8b49f1a7d6bbeaa869ba1c	https://github.com/thegenemyers/FASTK
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.5-r587	https://github.com/chhyllp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84aa44357826c0b6753eb28de	https://github.com/higlass/higlass
MercuryFK	d00d98157618f4e8d1a9190026b19b471055b22e	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	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
NCBI Datasets	15.12.0	https://github.com/ncbi/datasets
Nextflow	23.04.1	https://github.com/nextflow-io/nextflow
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	-	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.5.1	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

the EBI. Annotation was created primarily through alignment of transcriptomic data to the genome, with gap filling via protein-to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 2019).

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 [here](#). 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.

Data availability

European Nucleotide Archive: *Eupithecia inturbata* (maple pug). Accession number PRJEB67611; <https://identifiers.org/ena.embl/PRJEB67611>. The genome sequence is released openly for reuse. The *Eupithecia inturbata* genome sequencing initiative is part of the Darwin Tree of Life (DTOL) project. All raw sequence data and the assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

References

- Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Aken BL, Ayling S, Barrell D, et al.: **The Ensembl gene annotation system.** *Database (Oxford).* 2016; **2016**: baw093.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 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](#)
- Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor[®]3 for LI PacBio.** *protocols.io.* 2023.
[Publisher Full Text](#)
- Beasley J, Uhl R, Forrest LL, et al.: **DNA barcoding SOPs for the Darwin Tree of Life project.** *protocols.io.* 2023; [Accessed 25 June 2024].
[Publisher 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, Kumar S, Sotero-Caio C, et al.: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24.
[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](#)
- da Veiga Leprevost F, Grüning BA, Alves Aflitos S, et al.: **BioContainers: an open-source and community-driven framework for software standardization.** *Bioinformatics.* 2017; **33**(16): 2580–2582.
[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](#)
- Denton A, Oatley G, Cornwell C, et al.: **Sanger Tree of Life sample homogenisation: PowerMash.** *protocols.io.* 2023a.
[Publisher Full Text](#)
- Denton A, Yatsenko H, Jay J, et al.: **Sanger Tree of Life wet laboratory protocol collection V.1.** *protocols.io.* 2023b.
[Publisher Full Text](#)
- Di Tommaso P, Chatzou M, Floden EW, et al.: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Diesh C, Stevens GJ, Xie P, et al.: **JBrowse 2: a modular genome browser with views of synteny and structural variation.** *Genome Biol.* 2023; **24**(1): 74.
[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](#)

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods*. 2018; **15**(7): 475–476.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics*. 2020; **36**(9): 2896–2898.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Harry E: **PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps.** 2022.

[Reference Source](#)

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience*. 2021; **10**(1): gaa153.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life sample preparation: triage and dissection.** *protocols.io*. 2023.

[Publisher Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One*. 2017; **12**(5): e0177459.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[Publisher Full Text](#)

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics*. 2018; **34**(18): 3094–3100.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Manni M, Berkeley MR, 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, [Accessed 2 April 2024].

[Reference Source](#)

Oatley G, Denton A, Howard C: **Sanger Tree of Life HMW DNA extraction: Automated MagAttract v.2.** *protocols.io*. 2023.

[Publisher Full Text](#)

Pereira L, Sivell O, Sivess L, *et al.*: **DTOL taxon-specific standard operating**

procedure for the terrestrial and freshwater arthropods working group. 2022.

[Publisher Full Text](#)

Pointon DL, Eagles W, Sims Y, *et al.*: **sanger-tol/treeval v1.0.0 – Ancient Atlantis.** 2023.

[Publisher Full Text](#)

Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics*. 2010; **26**(6): 841–842.

[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.*: **Merquy: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol*. 2020; **21**(1): 245.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *protocols.io*. 2023.

[Publisher Full Text](#)

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer review: 2 approved].** *Wellcome Open Res*. 2024; **9**: 339.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Uliano-Silva M, Ferreira JGRN, 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](#)

UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res*. 2019; **47**(D1): D506–D515.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[Publisher Full Text](#)

Zhou C, McCarthy SA, Durbin R: **YaHS: yet another Hi-C scaffolding tool.** *Bioinformatics*. 2023; **39**(1): btac808.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:    

Version 1

Reviewer Report 25 April 2025

<https://doi.org/10.21956/wellcomeopenres.26246.r121415>

© 2025 Deprá M. This is an open access peer review report 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.



Maríndia Deprá

Federal University of Rio Grande do Sul, Porto Alegre, Brazil

The paper titled "The genome sequence of the Maple Pug moth, *Eupithecia inturbata* (Hübner, 1817)" is a Data Note presenting the nuclear and mitochondrial genome assemblies of *Eupithecia inturbata*. The nuclear genome spans 427.76 Mb and has been scaffolded into 31 chromosomal pseudomolecules, representing 30 autosomes and the Z sex chromosome. The mitochondrial genome is 15.33 Kb in length.

The methods employed follow standard practices for bioinformatics analyses, utilizing well-established pipelines such as Nextflow, BUSCO, and DIAMOND. The manuscript provides detailed descriptions of the software tools, databases, and parameters used, along with explanation of each step in the data processing workflow.

Although the manuscript does not include direct links to the datasets, the accession numbers are provided, ensuring data accessibility and enabling potential reuse.

Overall, the manuscript is well structured and technically sound. However, it would benefit from the inclusion of additional information about the species, such as aspects of its biology and habitat.

One methodological point that could be clarified is the identification of the Z chromosome: the manuscript states that this was based on read coverage statistics, but it is unclear which metrics were used. Wouldn't homology-based identification be more appropriate? Furthermore, identification of the W chromosome would also be desirable.

Finally, I suggest revising the sections that refer to the sex of the sequenced specimen. In some parts of the manuscript (e.g., the abstract and page 3), it is described as female, while in others it is referred to as male.

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

No

Are the protocols appropriate and is the work technically sound?

Yes

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

Yes

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: genetics and 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, however I have significant reservations, as outlined above.

Reviewer Report 11 April 2025

<https://doi.org/10.21956/wellcomeopenres.26246.r121412>

© 2025 Ioannidis P. This is an open access peer review report 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.



Panagiotis Ioannidis 

Foundation for Research & Technology - Hellas, Crete, Greece

The present manuscript presents the genome assembly and initial annotation of a lepidopteran species.

I would like to see a more detailed overview of the biology of this insect in the Background section. Currently the authors only mention the site from where they collected the sample.

Other than this the manuscript is very well-written with many technical details explaining all the methodology. Moreover, the figure legends are among the best I've seen!

Like other genome reports in this journal, a gene set was generated with the Ensembl Rapid Release pipeline. My first comment on this is that it would be nice for the reader to have at least a few details on the methodology followed for gene prediction (e.g. was RNAseq used as evidence, which proteins were used as evidence, etc). I guess that a detailed report can be found at Ensembl's web site, and that's why I'm not asking for details; a few basic things would be enough. My second comment is one that I have mentioned in my other reviews for this journal, but I should mention it again here; it is almost obligatory (!) to have the BUSCO scores for the generated gene set and put them side-by-side with the BUSCO scores from the corresponding

genome. This way the reader can see how good is the gene prediction pipeline (normally, the gene set score should be higher than the genome score).

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: bioinformatics, insect genomics, insect transcriptomics

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 11 April 2025

<https://doi.org/10.21956/wellcomeopenres.26246.r121416>

© 2025 Whibley A. This is an open access peer review report 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.



Annabel Whibley 

The University of Auckland, Auckland, New Zealand

This Data Note reports the genome assembly and Ensembl gene annotation of the Maple Pug moth *Eupithecia inturbata*.

Please note that there is an error in the methods: the specimen is recorded as male in the "Sample acquisition" section. This should be corrected to avoid confusion.

The assembly is of high quality, with clear reporting of metadata and methods and use of appropriate bioinformatics tools. The accession links are functional. The report follows DToL templating with clear presentation of relevant information.

The apparent absence of the W chromosome from this female is interesting. The reduced dosage of the Z chromosome compared to autosomes is consistent with the female identification of the specimen. Atypical sex chromatin has been described in Geometrids, including in the family Larentiinae (e.g. Cells 2021, 10(9), 2230; <https://doi.org/10.3390/cells10092230>). It would be

interesting to explore this finding further (and to compare to closely related species).

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, however I have significant reservations, as outlined above.

Reviewer Report 05 April 2025

<https://doi.org/10.21956/wellcomeopenres.26246.r120962>

© 2025 Kuppusamy S. This is an open access peer review report 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.



Sivasankaran Kuppusamy 

Loyola College, Chennai, India

The authors have done chromosome-level genome assembly of the Maple Pug moth, *Eupithecia inturbata* (Hubner, 1817). The species' genome size was estimated at 427.76 Mb. They have extracted and sequenced DNA using the correct methodology. Furthermore, they assembled and annotated using the appropriate software. The comprehensive data will be necessary for the phylogenomic studies of the lepidopteran moths.

Comments on the manuscript:

- The first sentence of the Background can include the author's name of the species. The sentence may change like this "The genome of the Maple Pug, *Eupithecia inturbata* (Hubner, 1817)"....
- The authors consistently included the complete name of the genus, *Eupithecia inturbata*, throughout the article. If at all possible, start with the full genus name and then abbreviate

it to *E. inturbata*.

- Authors have mentioned the sequence was done based on a female specimen in the Abstract, Background and Genome sequence report. But in the Methods authors have mentioned "An adult male *E. inturbata* was collected from Wytham Woods". I think there is an error. Authors can correct it.

Above all, I confirm that the manuscript meets the necessary scientific standard and is suitable for indexing

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: Phylogeny of Lepidoptera

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.
