Check for updates

DATA NOTE

REVISED The genome sequence of the Toadflax Pug, *Eupithecia linariata* ([Denis & Schiffermüller], 1775) (Lepidoptera: Geometridae)

[version 2; peer review: 2 approved]

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

v2 **First published:** 13 Jun 2026, **11**:361
https://doi.org/10.12688/wellcomeopenres.26703.1
Latest published: 21 Oct 2026, **11**:361
https://doi.org/10.12688/wellcomeopenres.26703.2

Abstract

We present a genome assembly from an individual male *Eupithecia linariata* (Toadflax Pug; Arthropoda; Insecta; Lepidoptera; Geometridae). The assembly contains two haplotypes with total lengths of 364.44 megabases and 364.02 megabases. Most of haplotype 1 (99.62%) is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 15.44 kilobases. Gene annotation of this assembly on Ensembl identified 12 684 protein-coding genes. This assembly was generated as part of the Darwin Tree of Life project, which produces genomes for eukaryotic species found in Britain and Ireland.

Keywords

Eupithecia linariata, Toadflax Pug, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 21 Oct 2026	✓ view	
version 1 13 Jun 2026	? view	✓ view

1. **Stuart J.E. Baird**, Academy of Sciences of the Czech Republic, Národní, Czech Republic

2. **Ricardo Rodriguez-de-la-vega**, Université Paris-Saclay, Gif-sur-Yvette, France

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)

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.

Copyright: © 2026 Natural History Museum Genome Acquisition Lab *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: Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding Collective, Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team *et al.* **The genome sequence of the Toadflax Pug, *Eupithecia linariata* ([Denis & Schiffermüller], 1775) (Lepidoptera: Geometridae) [version 2; peer review: 2 approved]** Wellcome Open Research 2026, **11**:361 https://doi.org/10.12688/wellcomeopenres.26703.2

First published: 13 Jun 2026, **11**:361 https://doi.org/10.12688/wellcomeopenres.26703.1

REVISED **Amendments from Version 1**

In Version 2 of this article, the Background has been revised in response to peer review comments to clarify the species' historical occurrence in Ireland and its food-plant associations.

Any further responses from the reviewers can be found at the end of the article

Species taxonomy

Eukaryota; Metazoa; Arthropoda; Insecta; Lepidoptera; Geometridae; *Eupithecia*; *Eupithecia linariata* (Denis & Schiffermüller, 1775) (NCBI:txid873508).

Background

Eupithecia linariata ([Denis & Schiffermüller], 1775) (Figure 1), the Toadflax Pug, is a geometrid moth that is widely recorded in the Palaearctic (GBIF Secretariat, 2026). In Great Britain, it is resident and common, being well distributed through much of England and Wales, more local in Scotland, rare in the Channel Islands, and recorded only historically from the Isle of Man. A historical Irish record is documented by a specimen collected at Favour Royal, Co. Tyrone, in 1903 (Bond *et al.*, 2006).

The larvae feed on Common Toadflax (*Linaria vulgaris*), first in the flowers and later in the seed capsules. The species also breeds on snapdragons (*Antirrhinum*) in some gardens (Waring *et al.*, 2017). The species is associated with rough chalk and limestone grassland, roadside verges, broken ground, derelict land and some urban sites. Adults fly mainly from late May to early September (Waring *et al.*, 2017).

We present a chromosome-level genome sequence for *Eupithecia linariata*, the Toadflax Pug. The assembly was produced using the Tree of Life pipeline using a specimen collected from Lucas Wood, High Wycombe, England, UK.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Eupithecia linariata* (specimen ID NHMUK014584890, ToLID ilEupLina1), collected from Lucas Wood, High Wycombe, England, UK (latitude 51.63, longitude −0.74) on 2022-07-07. The specimen was collected and identified by David Lees.

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.



Figure 1. Photograph of *Eupithecia linariata* by Ben Sale.

Nucleic acid extraction

Detailed protocols for nucleic acid extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The ilEupLina1 specimen was weighed and triaged to determine the appropriate extraction protocol. For HMW DNA extraction, 1 mg of tissue was used. Tissue from the whole organism was homogenised by powermashing using a PowerMasher II tissue disruptor. High molecular weight (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 18.2 ng/μL and a yield of 855.40 ng, with a fragment size of 14.1 kb. The 260/280 spectrophotometric ratio was 1.77, and the 260/230 ratio was 1.43. The Genomic Quality Number (GQN) was 7.

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 25 M SMRT cell. The SMRT Link software (Pacific Biosciences), a web-based workflow manager, was used to configure and monitor the run and to carry out primary and secondary data analysis.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the ilEupLina1 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

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

Prior to sequencing, libraries were normalised to 10 ng/μL. Normalised libraries were quantified again to create equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq 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, 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 MerquryFK (Rhie *et al.*, 2020).

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

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included seven breaks and 15 joins. This reduced the scaffold count by 9.5%. The curation process is described at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The MerquryFK 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 database ($k = 31$) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

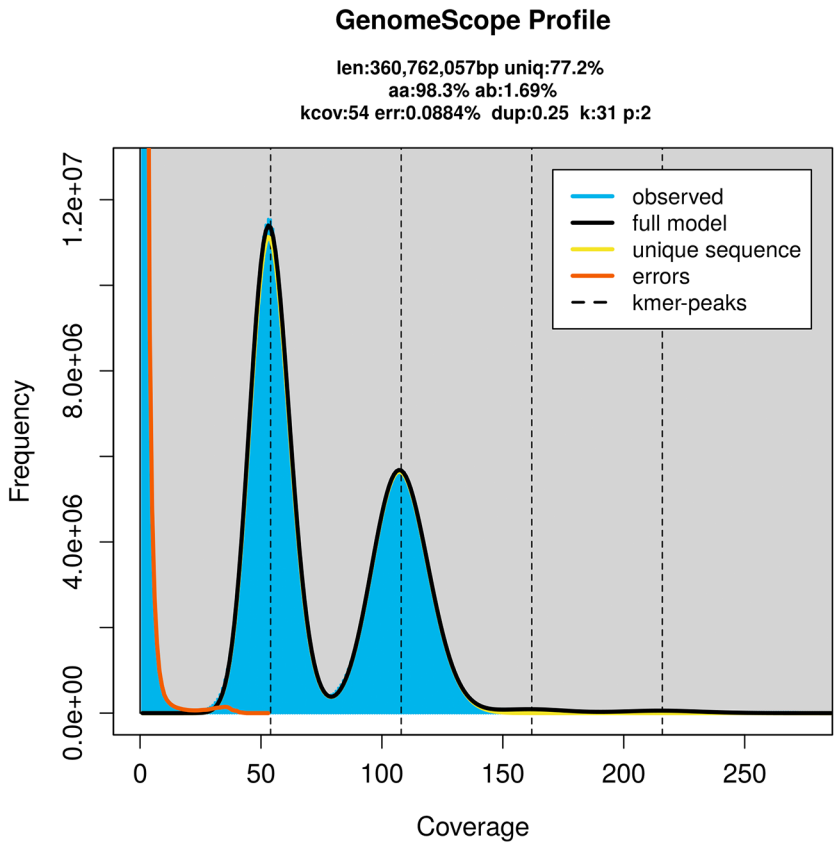


Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

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 *Eupithecia linariata* specimen generated 40.15 Gb (gigabases) from 3.70 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 360.76 Mb, with a heterozygosity of 1.69% and repeat content of 22.79% (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 110.14 Gb from 364.70 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Table 1. Specimen and sequencing data for *Eupithecia linariata* (BioProject PRJEB82155).

Platform	PacBio HiFi	Hi-C
ToLID	ilEupLina1	ilEupLina1
Specimen ID	NHMUK014584890	NHMUK014584890
BioSample (source individual)	SAMEA115574401	SAMEA115574401
BioSample (tissue)	SAMEA115599415	SAMEA115599415
Tissue	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13946455	ERR13947474
Read count total	3.70 million reads	364.70 million read pairs
Base count total	40.15 Gb	110.14 Gb

Table 2. Genome assembly statistics for *Eupithecia linariata.*

Genome assembly	Haplotype 1	Haplotype 2
Assembly name	ilEupLina1.hap1.1	ilEupLina1.hap2.1
Assembly accession	GCA_964340575.1	GCA_964340535.1
Assembly level	chromosome	scaffold
Span (Mb)	364.44	364.02
Number of chromosomes	31	scaffold-level
Number of contigs	132	108
Contig N50	6.44 Mb	6.71 Mb
Number of scaffolds	66	53
Scaffold N50	12.99 Mb	13.02 Mb
Longest scaffold length (Mb)	18.68	18.6
Sex chromosomes	Z	-
Organelles	Mitochondrial genome: 15.44 kb	-

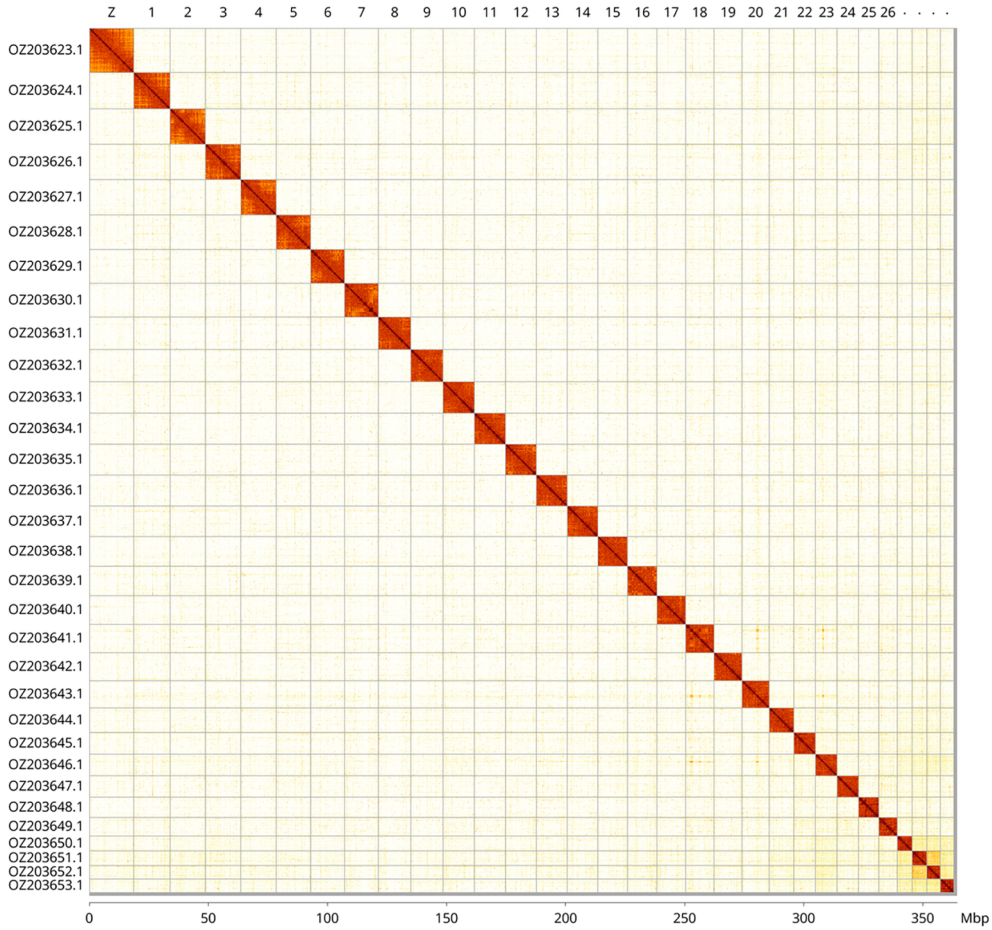


Figure 3. Hi-C contact map of the *Eupithecia linariata* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Eupithecia linariata* ilEupLina1.

INSDC accession	Molecule	Length (Mb)	GC%
OZ203624.1	1	15.27	37
OZ203625.1	2	14.90	37
OZ203626.1	3	14.87	37
OZ203627.1	4	14.79	37
OZ203628.1	5	14.51	36.50
OZ203629.1	6	14.23	36.50
OZ203630.1	7	14.13	36.50
OZ203631.1	8	13.69	36.50
OZ203632.1	9	13.51	36.50
OZ203633.1	10	13.16	36.50
OZ203634.1	11	13.07	36
OZ203635.1	12	12.99	36
OZ203636.1	13	12.95	36.50

Table 3. *Continued*

INSDC accession	Molecule	Length (Mb)	GC%
OZ203637.1	14	12.87	36.50
OZ203638.1	15	12.42	36.50
OZ203639.1	16	12.37	37
OZ203640.1	17	12.01	36.50
OZ203641.1	18	11.92	36.50
OZ203642.1	19	11.79	36.50
OZ203643.1	20	11.43	36.50
OZ203644.1	21	10.30	36.50
OZ203645.1	22	9.11	36.50
OZ203646.1	23	9.11	37
OZ203647.1	24	8.98	36
OZ203648.1	25	8.47	36.50
OZ203649.1	26	7.79	36
OZ203650.1	27	6.26	36.50
OZ203651.1	28	6.10	36.50
OZ203652.1	29	5.72	37
OZ203653.1	30	5.66	38.50
OZ203623.1	Z	18.68	36.50

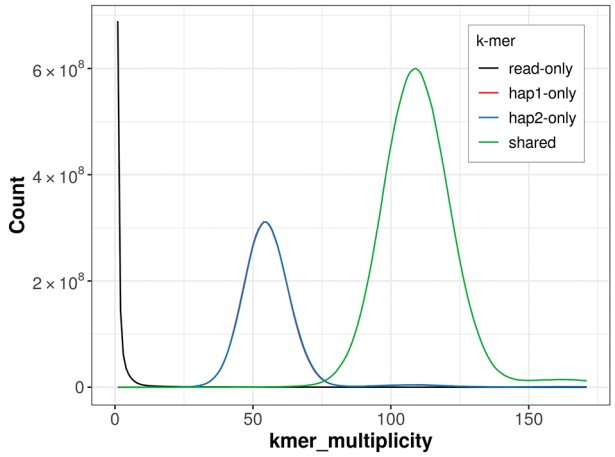


Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

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 364.44 Mb in 66 scaffolds, with 66 gaps, and a scaffold N50 of 12.99 Mb (Table 2).

Most of the haplotype 1 assembly sequence (99.62%) 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 3; Table 3). Chromosome Z was assigned by mapping of Merian elements (Wright *et al.*, 2024).

The mitochondrial genome was also assembled (length 15.44 kb, OZ203654.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

For the haplotype 1 assembly, the estimated QV is 69.5, and for haplotype 2, 69.4. The combined assembly achieves an estimated QV of 69.4. The *k*-mer completeness is 71.25% for haplotype 1, 71.24% for haplotype 2, and 99.82% for the combined haplotypes (Figure 4).

BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.1% of the expected gene set (single = 97.6%, duplicated = 0.5%) in the haplotype 1 assembly. For the haplotype 2, BUSCO v.5.5.0 analysis identified 98.0% of the expected gene set (single = 97.4%, duplicated = 0.6%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1 assembly.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards. The EBP metric, calculated for the haplotype 1 assembly, is **6.C.Q69**, meeting the recommended 6.C.Q40 reference standard.

Genome annotation report

The *Eupithecia linariata* genome assembly (GCA_964340575.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 24 677 transcribed mRNAs from 12 684 protein-coding and 2 342 non-

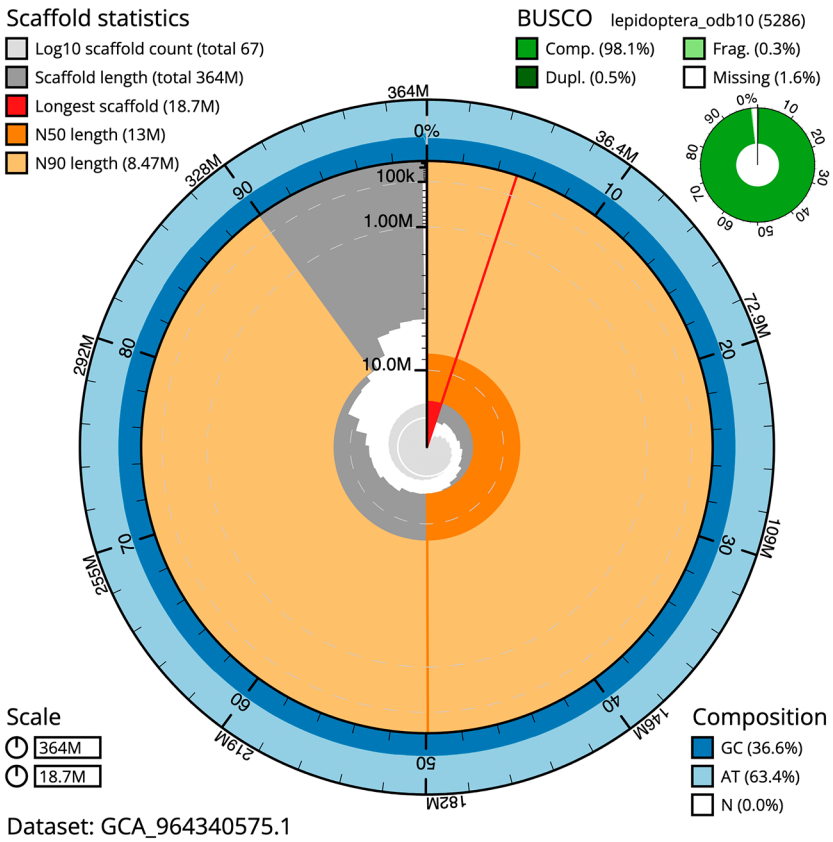


Figure 5. Assembly metrics for ilEupLina1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost 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.

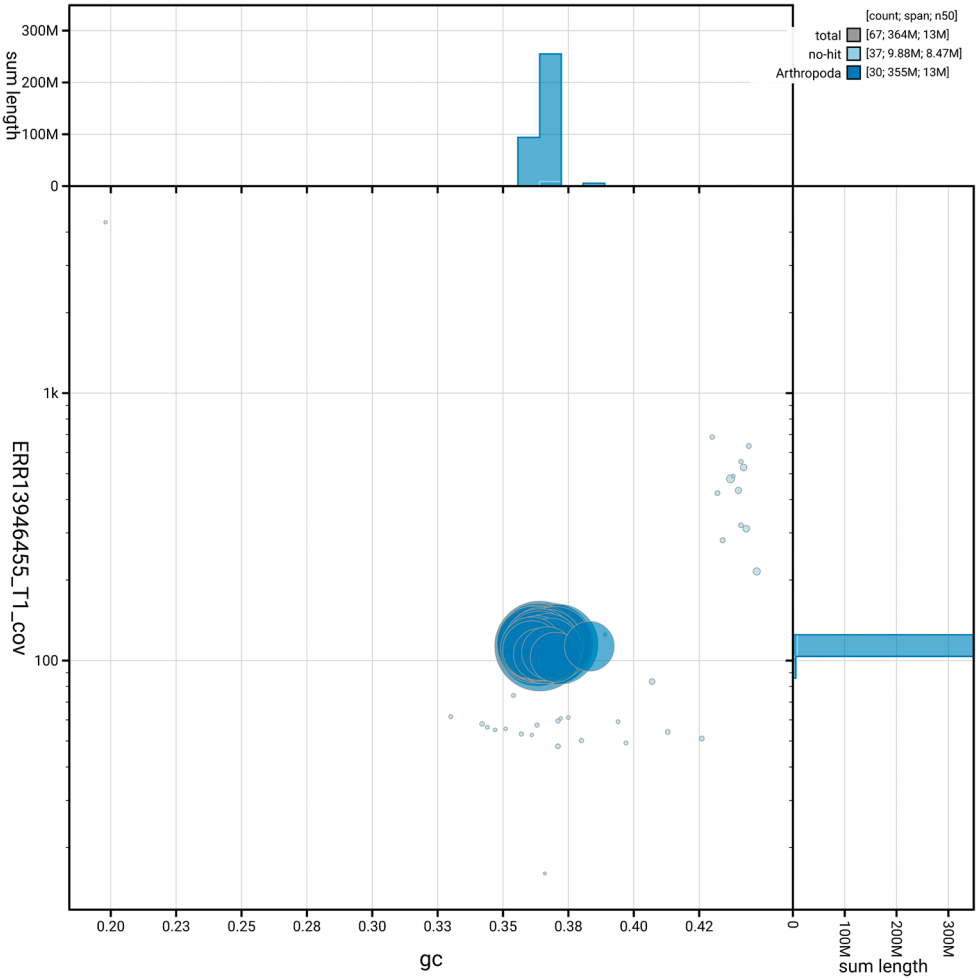


Figure 6. BlobToolKit blob plot for ilEupLina1.hap1.1. The plot shows base coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the 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 *Eupithecia linariata* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q69	6.C.Q40
Contig N50 length	6.44 Mb	≥ 1 Mb
Scaffold N50 length	12.99 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 69.5; haplotype 2: 69.4; combined: 69.4	≥ 40
k-mer completeness	Haplotype 1: 71.25%; Haplotype 2: 71.24%; combined: 99.82%	≥ 95%
BUSCO	Haplotype 1: C:98.1% [S:97.6%, D:0.5%], F:0.3%, M:1.6%, n:5 286; Haplotype 2: C:98.0% [S:97.4%, D:0.6%], F:0.4%, M:1.6%, n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.62%	≥ 90%

Notes: The EBP summary uses log10(Contig N50); chromosome-level (C) or log10(Scaffold N50); Q (Merqury QV). BUSCO: C = complete; S = single-copy; D = duplicated; F = fragmented; M = missing; n = orthologues.

coding genes. The average transcript length is 14 289.98 bp, with an average of 1.64 coding transcripts per gene and 7.41 exons per transcript. Further details of this annotation are available from the Ensembl annotation page.

Author information

Contributors are listed at the following links:

- 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

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.

Data availability

European Nucleotide Archive: Eupithecia linariata. Accession number PRJEB82155; https://identifiers.org/ena.embl/PRJEB82155. The genome sequence is released openly for reuse. The *Eupithecia linariata* 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. Raw data and assembly accession identifiers are reported in Tables 1 and 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.

Table 5. Software versions and sources used for *Eupithecia linariata*.

Software	Version	Source
BLAST	2.14.0	https://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

Table 5. *Continued*

Software	Version	Source
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
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
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	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	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
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

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

Bateman A, Martin M-J, 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

Bond KGM, Nash R, O'Connor JP, *et al.*: **An Annotated Checklist of the Irish Butterflies and Moths (Lepidoptera).** *Irish Biogeographical Society and National Museum of Ireland, Dublin.* 2006.

Buchfink B, Reuter K, Drost H-G: **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 Genes|Genomes|Genetics.* 2020; **10**(4): 1361–1374.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial Arthropod fauna.** *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.** *Nature Biotechnology.* 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: **GBIF backbone taxonomy: *Eupithecia linariata* (Denis & Schiffermüller), 1775.** 2026.
Reference Source

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

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

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

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

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.
Publisher Full Text

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

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

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

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

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

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life Project.** *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.** *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE; 2019; 314–324.
Publisher Full Text

Waring P, Townsend M, Lewington R: *Field Guide to the Moths of Great Britain and Ireland.* London: Bloomsbury Wildlife; 2017.

Wright CJ, Stevens L, Mackintosh A, *et al.*: **Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera.** *Nat Ecol Evol.* 2024; **8**(4): 777–790.
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 2

Reviewer Report 22 October 2026

https://doi.org/10.21956/wellcomeopenres.30954.r175250

© **2026 Baird 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.

✔ **Stuart J.E. Baird**
Academy of Sciences of the Czech Republic, Národní, Czech Republic

The authiors have addressed the minor issue I flagged.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Admixture 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.

Version 1

Reviewer Report 14 October 2026

https://doi.org/10.21956/wellcomeopenres.29404.r167408

© **2026 Rodriguez-de-la-vega R.** 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.

✔ **Ricardo Rodriguez-de-la-vega**
Université Paris-Saclay, Gif-sur-Yvette, France

Comments on NHM Genome Acquisition Lab, DToL Barcoding Collective, et al., "The genome sequence of the Toadflax Pug ..."

This collective authors' Data Note from the Darwin Tree of Life (DToL) project presents Earth

BioGenome Project reference level haplotype resolved assemblies of a male specimen of *Eupithecia linariata* (commonly known as the toadflax pug). Assemblies were obtained with hifiasm on HiFi long-reads scaffolded with Arima-HiC reads, resulting in two haplotypes, the primary of which was manually curated in PretextView and HiGlass using Hi-C data to generate a chromosome level assembly, leaving the other haplotype at the scaffold level, notably quality of both assemblies meet the Earth Biogenome Project reference.

Sample acquisition, nucleic acid extraction, library preparation and crosslinking, sequencing and assembly followed the well-established DToL's standard operating procedures, dully referenced in the Data Note. I spotted no technical issues. As common for recent (since early 2026) DToL-derived Data Notes, supporting figures include:
- GenomeScope kmer frequency vs kmer coverage of HiFi reads,
- Hi-C contact map of the primary assembly,
- MerquryFK kmer multiplicity of reads and assemblies,
- BlobToolKit snail plot of the primary assembly and
- Coverage (minimap2 of reads on assembly) vs gc content length-scaled scatter plot with length distribution histograms.

Although the Z chromosome is said to have been identified by mapping Merian elements, how this was done is not disclosed. In the original reference (Wright et all., 2024 in the references' list) authors used the tool syngraph (Mackintosh et al., 2023) to define the Merian elements. It is not clear to me if authors of this Data Note have followed a similar approach or have been directly use the shiny app (https://charlottejwright.shinyapps.io/busco_painter/ input: busco full_table.tsv run with --datasets_version odb10) or python code associated with the original reference (https://github.com/charlottewright/lep_busco_painter). If the Data Note gets updated, it would be good that this procedure is more clearly explained.

I have run BUSCO 6.0.0 with -l lepidoptera --datasets_version odb10 for both assemblies. Worth noticing is that there is a perfect match between the assembled chromosomes in the manually curated primary assembly and the scaffolds in the second haplotype, both in terms of composition and with regard to location of Merian elements (i.e. the assemblies are collinear). It would be interesting to check how often this is the case among DToL or other EBGP project (e.g. Psyche also released through Wellcome Open Research).

For the sake of convenience, I am including links to the DToL quality check and Genomes on a Tree page for this specimen/species.
https://tolqc.cog.sanger.ac.uk/darwin/insects/Eupithecia_linariata
https://goat.genomehubs.org/record?recordId=873508&result=taxon&taxonomy=ncbi#Eupithecia%20linariata

References

1. Mackintosh A, de la Rosa P, Martin S, Lohse K, et al.: Inferring inter-chromosomal rearrangements and ancestral linkage groups from synteny. *bioRxiv*. 2023. Publisher Full Text

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

Yes

Are the protocols appropriate and is the work technically sound?

Yes

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

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Evolutionary 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 08 October 2026

https://doi.org/10.21956/wellcomeopenres.29404.r167409

© 2026 Baird 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.

?

Stuart J.E. Baird
Academy of Sciences of the Czech Republic, Národní, Czech Republic

The genome note follows the excellent DToL template.

I have two issues only:

(1)

Background
“In Great Britain and Ireland, it is resident and common, being well distributed through much of England and Wales, more local in Scotland, rare in the Channel Islands, and recorded historically from the Isle of Man.”

The species is also historically recorded in Ireland.

(2)

"The species is associated with rough chalk and limestone grassland, roadside verges, broken ground, derelict land and some urban sites. Adults fly mainly from late May to early September, and the larvae feed on Common Toadflax, first in the flowers and later in the seed capsules"

The species is usually described as associated first with two food plants, rather than the context in which they in turn are found: Common Toadflax (*Linaria vulgaris*) and cultivated Snapdragons (

Antirrhinum). I find it strange to frontload the substrate association, as the foodplant association explains not only the common name of this geometrid, but also the scarcity of the Toadflax Pug in Ireland: despite its name, Common Toadflax is extremely rare in Ireland (and likely introduced). Likewise, Toadflax is a rare site on the Isle of Man: it appears the main constraint of the species distribution is the food plant, not soils that food plants could in principle grow on.

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: Admixture 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.

Author Response 14 Oct 2026

Tree of Life Team Sanger

Thank you for your helpful comments. We have corrected the text of the Background.

The Tree of Life team

Competing Interests: No competing interests were disclosed.

Comments on this article

Version 1

Reader Comment 15 Jun 2026

jim mallet, Organismic and Evolutionary Biology, Harvard University Faculty of Arts and Sciences,

Cambridge, USA

I hope this serves as a review, as well as a comment. This is a very nice, clean example of Darwin Tree of Life genome sequencing of a British moth. It is also a beautiful species I know well from my youthful moth-trapping in Kent, UK. I am so happy we now know so much more about it, including Z chromosome, complete mitochondrial genome, etc. Keep up the good work Darwin Tree of Lifers, and all of you collaborators!

Competing Interests: none