DATA NOTE

The genome sequence of the Brindled Pug, *Eupithecia abbreviata* (Stephens, 1831) [version 1; peer review: awaiting peer review]

Douglas Boyes[1+], Peter W.H. Holland[2],
University of Oxford and Wytham Woods Genome Acquisition Lab,
Darwin Tree of Life Barcoding collective,
Wellcome Sanger Institute Tree of Life programme,
Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]UK Centre for Ecology & Hydrology, Wallingford, England, UK
[2]University of Oxford, Oxford, England, UK

[+] Deceased author

v1 **First published:** 28 Mar 2023, **8**:140
https://doi.org/10.12688/wellcomeopenres.19244.1
Latest published: 28 Mar 2023, **8**:140
https://doi.org/10.12688/wellcomeopenres.19244.1

Abstract

We present a genome assembly from an individual female *Eupithecia abbreviata* (the Brindled Pug; Arthropoda; Insecta; Lepidoptera; Geometridae). The genome sequence is 383.4 megabases in span. Most of the assembly is scaffolded into 32 chromosomal pseudomolecules, including the assembled Z and W sex chromosomes. The mitochondrial genome has also been assembled and is 15.3 kilobases in length. Gene annotation of this assembly on Ensembl identified 16,676 protein coding genes.

Keywords

Eupithecia abbreviata, Brindled Pug, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

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.

Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: Boyes D: Investigation, Resources; **Holland PWH**: Writing – Original Draft Preparation, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (206194, https://doi.org/10.35802/206194) and the Darwin Tree of Life Discretionary Award (218328, https://doi.org/10.35802/218328).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Boyes D, Holland PWH, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Brindled Pug, *Eupithecia abbreviata* (Stephens, 1831) [version 1; peer review: awaiting peer review]** Wellcome Open Research 2023, **8**:140 https://doi.org/10.12688/wellcomeopenres.19244.1

First published: 28 Mar 2023, **8**:140 https://doi.org/10.12688/wellcomeopenres.19244.1

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Geometroidea; Geometridae; Larentiinae; *Eupithecia*; *Eupithecia abbreviata* (Stephens, 1831) (NCBI:txid934840).

Background

The 'Pugs' are a group of small and delicate moths in the family Geometridae, most species of which are classified in the genus *Eupithecia*. Pugs are generally recognisable by the resting position of the adults: with wings closely pressed against the tree bark or other surface and with the leading edge of the left and right forewings usually forming a straight line at ninety degrees to the main axis of the body. The Brindled Pug *Eupithecia abbreviata* is typically the first Pug on the wing in spring in Britain and Ireland, seen as early as March and continuing to be recorded until late May (South, 1961; Waring *et al.*, 2017). The adjective 'brindled' in the common name refers to the angled and jagged streaks of black and ochreous brown that cross the grey, pointed forewings; 'brindled' has been described as "a peculiar colouration formerly seen on cows and some varieties of dog" (Newman, 1869). Although the ground colour of the wings is typically grey, there is considerable variation from very pale to black (Riley & Prior, 2003; South, 1961).

The Brindled Pug *E. abbreviata* is found across Europe, and can be common in woodlands and gardens in England, Wales, southern Scotland and the southern half of Northern Ireland (GBIF Secretariat, 2022; Thompson & Nelson, 2003). The larvae feed on pedunculate oak (*Quercus robur*), sessile oak (*Q. petraea*) and hawthorn (*Crataegus monogyna*). The maroon-striped larvae feed in June and July before pupating in cocoon in loose soil where the pupa overwinters (Stokoe, 1948).

A genome sequence for *E. abbreviata* will facilitate studies into wing colour polymorphisms and uncovering the molecular adaptations to feeding on oak.

Genome sequence report

The genome was sequenced from one female *E. abbreviata* specimen (Figure 1) collected from Wytham Woods, Oxfordshire, UK (latitude 51.772, longitude –1.34). A total of 48-fold coverage in Pacific Biosciences single-molecule HiFi long reads was generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected four missing or mis-joins and removed three haplotypic duplications, reducing the assembly length by 0.37% and the scaffold number by 9.21%.

The final assembly has a total length of 383.4 Mb in 69 sequence scaffolds with a scaffold N50 of 13.1 Mb (Table 1). Most (99.98%) of the assembly sequence was assigned to 32 chromosomal-level scaffolds, representing 30 autosomes, and the Z and W sex chromosomes. Chromosome-scale scaffolds

Figure 1. Photograph of the *Eupithecia abbreviata* (ilEupAbbr2) specimen used for genome sequencing.

confirmed by the Hi-C data are named in order of size (Figure 2–Figure 5; Table 2). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The estimated Quality Value (QV) of the final assembly is 67.5 with *k*-mer completeness of 100%, and the assembly has a BUSCO v5.3.2 (Manni *et al.*, 2021) completeness of 97.9% (single 97.4%, duplicated 0.6%) using the lepidoptera_odb10 reference set (n = 5,286).

Genome annotation report

The *Eupithecia abbreviata* genome assembly GCA_943735965.1 (ilEupAbbr2.1) was annotated using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Eupithecia_abbreviata_GCA_943735965.1/Info/Index/). The resulting annotation includes 16,876 transcribed mRNAs from 16,676 protein-coding genes.

Methods

Sample acquisition and nucleic acid extraction

Two *Eupithecia abbreviata* specimens (ilEupAbbr2 and ilEupAbbr3) were collected from Wytham Woods, Oxfordshire, UK (biological vice-county: Berkshire) (latitude 51.77, longitude –1.34) on 31 March 2021. The specimens were taken from woodland habitat by Douglas Boyes (University of Oxford) using a light trap. The specimens were identified by the collector and snap-frozen on dry ice.

DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute (WSI). The ilEupAbbr2 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. Whole organism tissue was disrupted using a Nippi Powermasher fitted with a BioMasher pestle. High molecular weight (HMW) DNA was extracted using the Qiagen MagAttract HMW DNA extraction kit. HMW DNA

Table 1. **Genome data for *Eupithecia abbreviata*, ilEupAbbr2.1.**

Project accession data		
Assembly identifier	ilEupAbbr2.1	
Species	*Eupithecia abbreviata*	
Specimen	ilEupAbbr2	
NCBI taxonomy ID	934840	
BioProject	PRJEB52575	
BioSample ID	SAMEA10107031	
Isolate information	ilEupAbbr2, female (PacBio sequencing and Hi-C scaffolding) ilEupAbbr3 (RNA sequencing)	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	67.5	*≥ 50*
k-mer completeness	100%	*≥ 95%*
BUSCO**	C:97.9%[S:97.4%,D:0.6%], F:0.6%,M:1.4%,n:5,286	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	99.98%	*≥ 95%*
Sex chromosomes	Z and W chromosomes	*localised homologous pairs*
Organelles	Mitochondrial genome assembled	*complete single alleles*
Raw data accessions		
PacificBiosciences SEQUEL II	ERR9793190	
Hi-C Illumina	ERR9682490	
PolyA RNA-Seq Illumina	ERR10123697	
Genome assembly		
Assembly accession	GCA_943735965.1	
Accession of alternate haplotype	GCA_943735975.1	
Span (Mb)	383.4	
Number of contigs	74	
Contig N50 length (Mb)	13.1	
Number of scaffolds	69	
Scaffold N50 length (Mb)	13.1	
Longest scaffold (Mb)	19.1	
Genome annotation		
Number of protein-coding genes	16,676	
Number of gene transcripts	16,876	

* Assembly metric benchmarks are adapted from column VGP-2020 of "Table 1: Proposed standards and metrics for defining genome assembly quality" from (Rhie *et al.*, 2021).

** BUSCO scores based on the lepidoptera_odb10 BUSCO set using v5.3.2. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/ilEupAbbr2.1/dataset/CALSES01/busco.

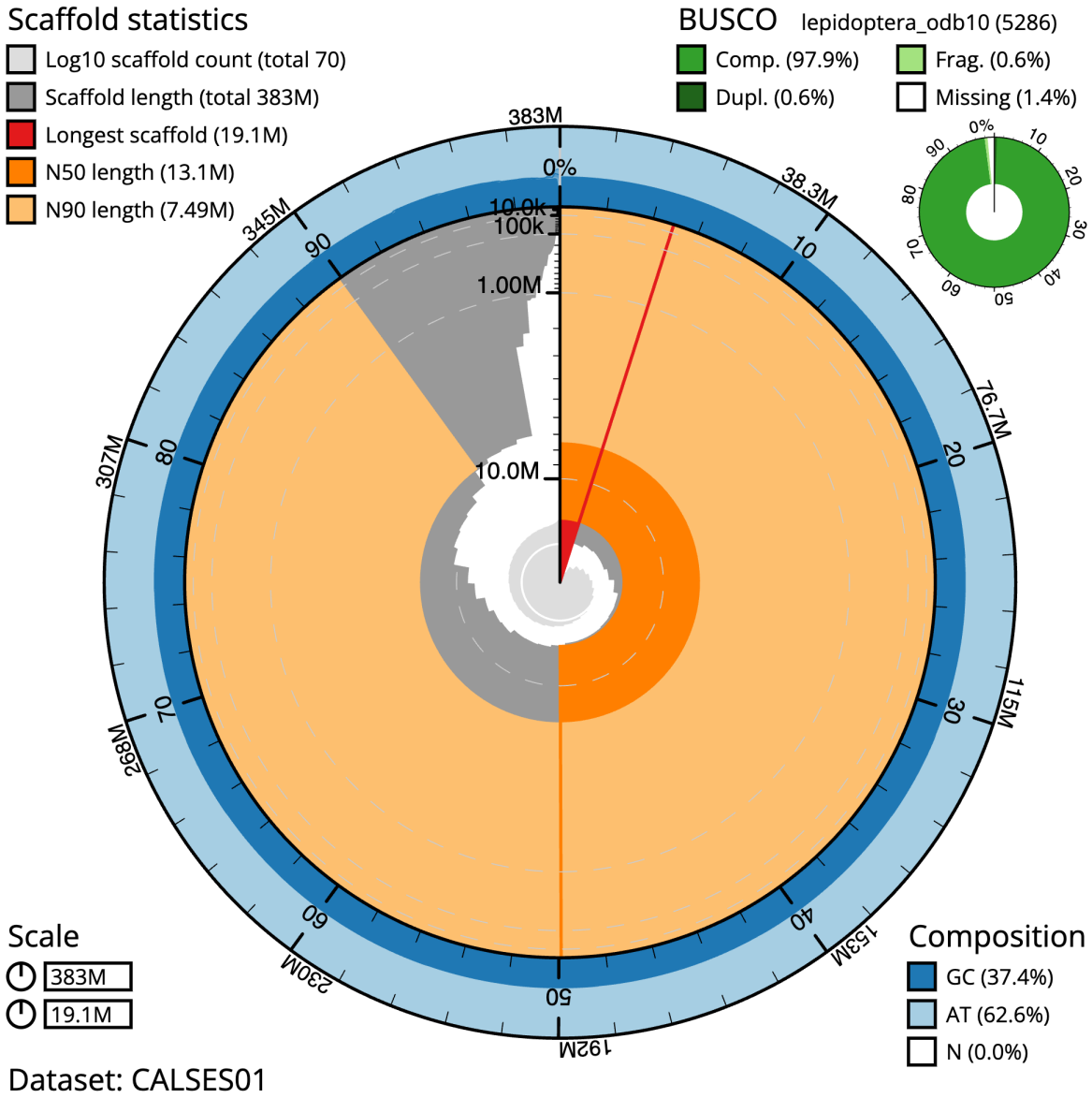

Figure 2. Genome assembly of *Eupithecia abbreviata*, ilEupAbbr2.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 383,462,667 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (19,063,336 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (13,098,345 and 7,494,420 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the lepidoptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilEupAbbr2.1/dataset/CALSES01/snail.

was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 30. Sheared DNA was purified by solid-phase reversible immobilisation using AMPure PB beads with a 1.8X ratio of beads to sample to remove the shorter fragments and concentrate the DNA sample. The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

RNA was extracted from whole organism tissue of ilEupAbbr3 in the Tree of Life Laboratory at the WSI using TRIzol, according to the manufacturer's instructions. RNA was then

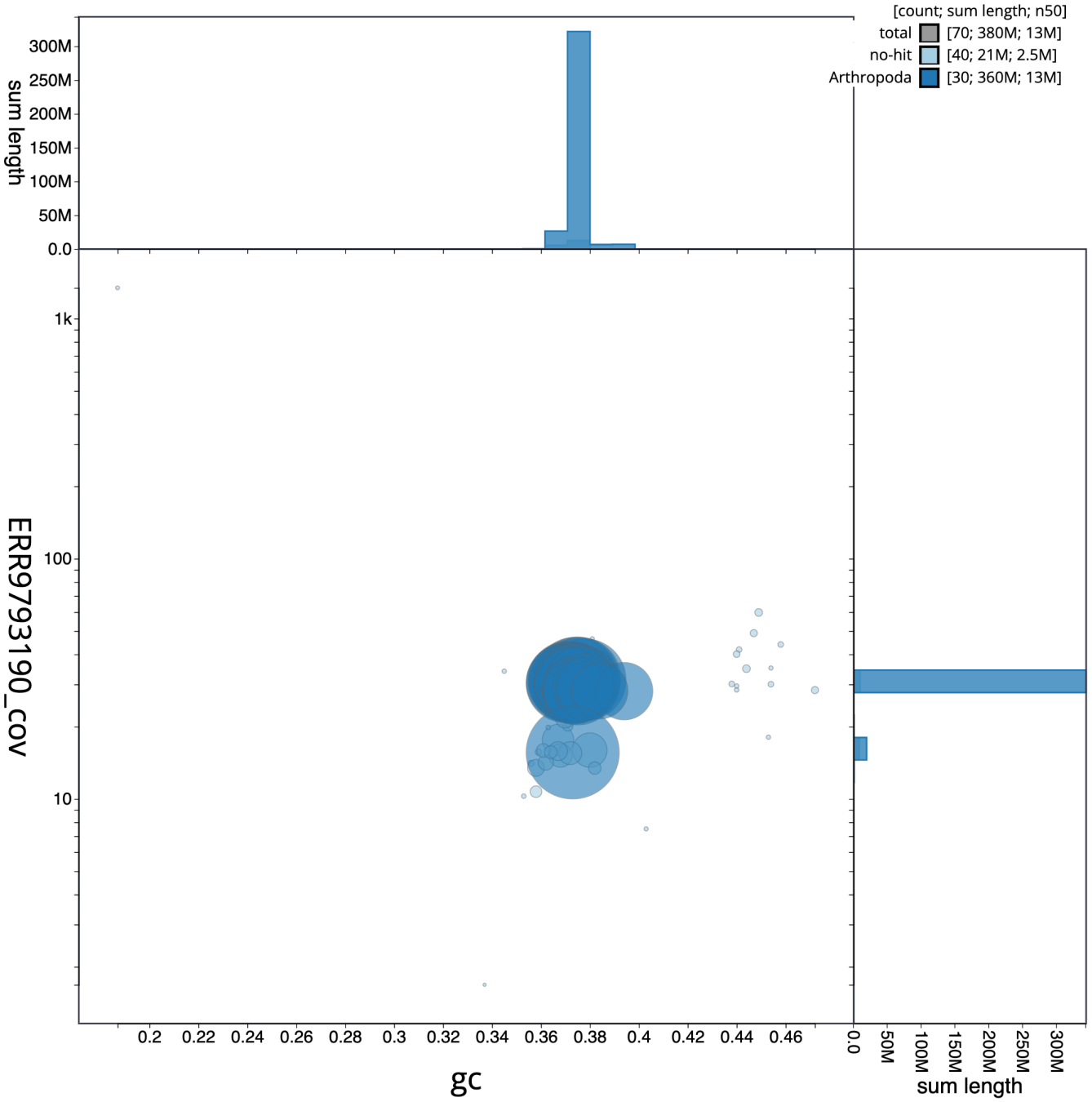

Figure 3. Genome assembly of *Eupithecia abbreviata*, ilEupAbbr2.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilEupAbbr2.1/dataset/CALSES01/blob.

eluted in 50 μl RNAse-free water and its concentration assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit RNA Broad-Range (BR) Assay kit. Analysis of the integrity of the RNA was done using Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

Sequencing

Pacific Biosciences HiFi circular consensus DNA sequencing libraries were constructed according to the manufacturers' instructions. Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit. DNA and RNA

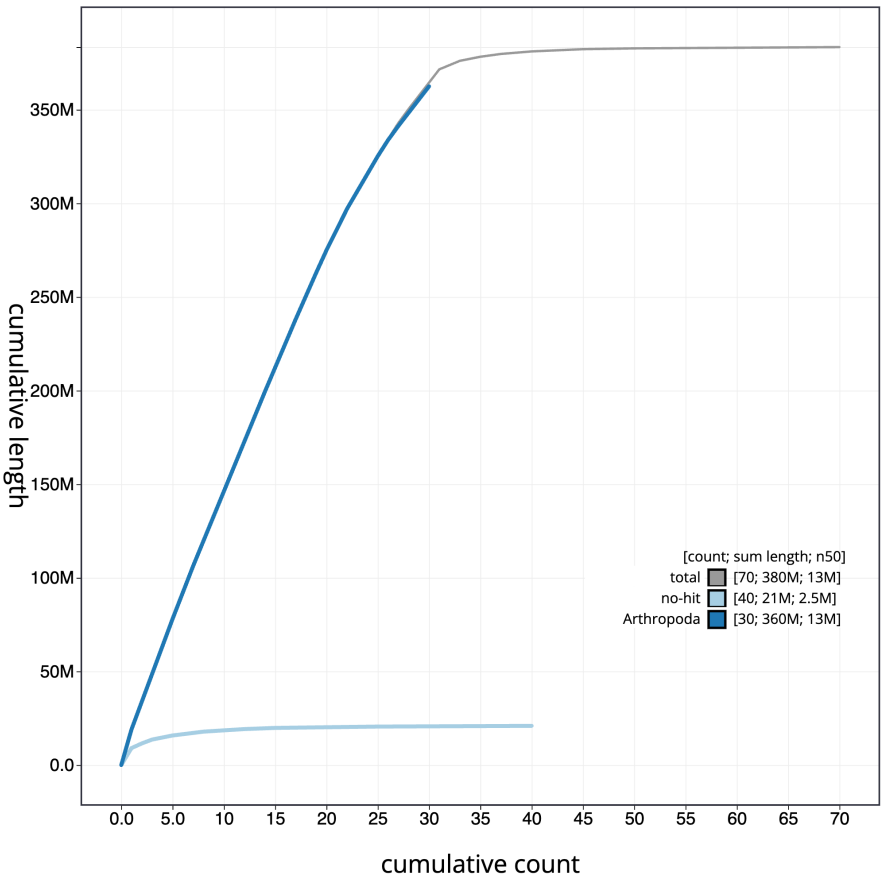

Figure 4. Genome assembly of *Eupithecia abbreviata*, ilEupAbbr2.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilEupAbbr2.1/dataset/CALSES01/cumulative.

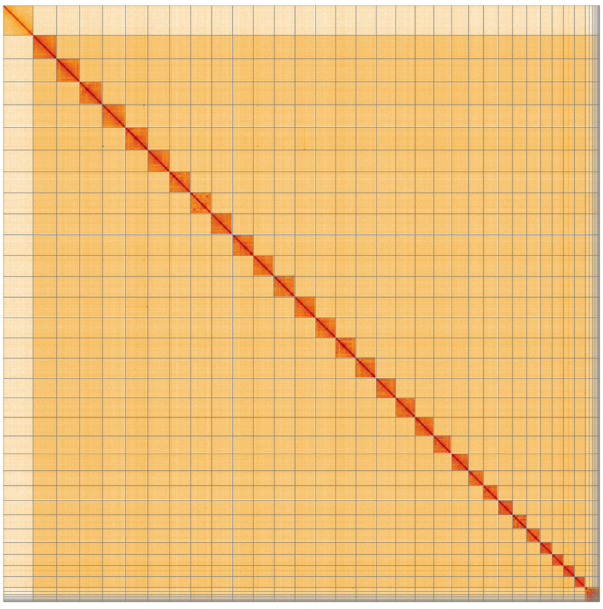

Figure 5. Genome assembly of *Eupithecia abbreviata*, ilEupAbbr2.1: Hi-C contact map. Hi-C contact map of the ilEupAbbr2.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=LCYsQChWQA-KzxTN1QsZbw.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Eupithecia abbreviata*, ilEupAbbr2.

INSDC accession	Chromosome	Size (Mb)	GC%
OX030974.1	1	15.02	37.4
OX030975.1	2	14.73	37.8
OX030976.1	3	14.62	37.5
OX030977.1	4	14.61	37.3
OX030978.1	5	14.33	37.6
OX030979.1	6	13.86	37.1
OX030980.1	7	13.45	37
OX030981.1	8	13.41	37.5
OX030982.1	9	13.36	37.2
OX030983.1	10	13.29	37.3
OX030984.1	11	13.29	37.1
OX030985.1	12	13.27	37
OX030986.1	13	13.1	37.3
OX030987.1	14	12.94	37.1
OX030988.1	15	12.88	37.4
OX030989.1	16	12.86	37.4
OX030990.1	17	12.48	37.3
OX030991.1	18	12.38	37.6
OX030992.1	19	11.99	37.4
OX030993.1	20	11.3	37.2
OX030995.1	21	10.78	37.6
OX030996.1	22	9.63	37.4
OX030997.1	23	9.33	37.4
OX030998.1	24	9.21	37.8
OX030999.1	25	9.07	37.5
OX031000.1	26	8.59	37.4
OX031001.1	27	7.22	37.8
OX031002.1	28	7.49	37.8
OX031003.1	29	7.14	39.4
OX031004.1	30	6.88	38.4
OX030994.1	W	2.49	38
OX030973.1	Z	19.06	37.3
OX031005.1	MT	0.02	18.9
-	unplaced	9.38	37.3

sequencing were performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi) and Illumina NovaSeq 6000 (RNA-Seq) instruments. Hi-C data were also generated from tissue of ilEupAbbr2 using the Arima v2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021) and haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2020). The assembly was scaffolded with Hi-C data (Rao *et al.*, 2014) using YaHS (Zhou *et al.*, 2023). The assembly was checked for contamination as described previously (Howe *et al.*, 2021). Manual curation was performed using HiGlass (Kerpedjiev *et al.*, 2018) and Pretext (Harry, 2022). The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2022), which performed annotation using MitoFinder (Allio *et al.*, 2020). To evaluate the assembly, MerquryFK was used to estimate consensus quality (QV) scores and *k*-mer completeness (Rhie *et al.*, 2020). The genome was analysed and BUSCO scores (Manni *et al.*, 2021; Simão *et al.*, 2015) were generated within the BlobToolKit environment (Challis *et al.*, 2020). Table 3 contains a list of software tool versions and sources.

Genome annotation

The BRAKER2 pipeline (Brůna *et al.*, 2021) was used in the default protein mode to generate annotation for the *Eupithecia abbreviata* assembly (GCA_943735965.1) in Ensembl Rapid Release.

Ethics and compliance issues

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. 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. All efforts are undertaken to minimise the suffering of animals used for sequencing. 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 3. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	4.0.7	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.3.2	https://gitlab.com/ezlab/busco
Hifiasm	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	1.11.6	https://github.com/higlass/higlass
Merqury	MerquryFK	https://github.com/thegenemyers/MERQURY.FK
MitoHiFi	2	https://github.com/marcelauliano/MitoHiFi
PretextView	0.2	https://github.com/wtsi-hpag/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
YaHS	yahs-1.1.91eebc2	https://github.com/c-zhou/yahs

Data availability

European Nucleotide Archive: *Eupithecia abbreviata* (brindled pug). Accession number PRJEB52575; https://identifiers.org/ena.embl/PRJEB52575. (Wellcome Sanger Institute, 2022)

The genome sequence is released openly for reuse. The *Eupithecia abbreviata* 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.

Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.4789928.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.4893703.

Members of the Wellcome Sanger Institute Tree of Life programme are listed here: https://doi.org/10.5281/zenodo.4783585.

Members of Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.4790455.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.5013541.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783558.

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

Brůna T, Hoff KJ, Lomsadze A, *et al.*: **BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database.** *NAR Genom Bioinform.* 2021; **3**(1): lqaa108.
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

GBIF Secretariat: **Eupithecia abbreviata Stephens, 1831, GBIF Backbone Taxonomy.** 2022; (Accessed: 6 March 2023).
Publisher 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; (Accessed: 19 October 2022).
Reference Source

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* Oxford University Press. 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

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

Newman E: **An Illustrated Natural History of British Moths.** London: Hardwicke and Bogue. 1869.
Reference Source

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 Biology.* 2020; **21**(1): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

Riley A, Prior G: **British and Irish Pug Moths – a Guide to Their Identification and Biology.** Leiden, Netherlands: BRILL. 2003.
Reference Source

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–3212.
PubMed Abstract | Publisher Full Text

South R: **Moths of the British Isles.** *New edition*. London: Frederick Warne and Co. 1961.

Stokoe WJ: **The Caterpillars of British Moths.** Series 2. London and New York: Warne and Co. 1948.

Thompson RS, Nelson B: **Brindled Pug.** In: *The Butterflies and Moths of Northern Ireland*. Ulster Museum. 2003; (Accessed: 6 March 2023).
Reference Source

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio High Fidelity reads.** *bioRxiv.* 2022; [Preprint].
Publisher Full Text

Waring P, Townsend M, Lewington R: **Field Guide to the Moths of Great Britain and Ireland: Third Edition.** Bloomsbury Wildlife Guides. 2017.
Reference Source

Wellcome Sanger Institute: **The genome sequence of the Brindled Pug, *Eupithecia abbreviata* (Stephens, 1831).** European Nucleotide Archive. [dataset], accession number PRJEB52575. 2022.

Zhou C, McCarthy SA, Durbin R: **YaHS: yet another Hi-C scaffolding tool.** *Bioinformatics.* Edited by C. Alkan, 2023; **39**(1): btac808.
PubMed Abstract | Publisher Full Text | Free Full Text