DATA NOTE

The genome sequence of the Lime-speck Pug, *Eupithecia centaureata* (Denis & Schiffermuller, 1775) [version 1; peer review: awaiting peer review]

Douglas Boyes[1], Stephanie Fagan[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]Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 23 Mar 2023, **8**:132
https://doi.org/10.12688/wellcomeopenres.19249.1

Latest published: 23 Mar 2023, **8**:132
https://doi.org/10.12688/wellcomeopenres.19249.1

Abstract

We present a genome assembly from an individual male *Eupithecia centaureata* (the Lime-speck Pug; Arthropoda; Insect; Lepidoptera; Geometridae). The genome sequence is 465.6 megabases in span. Most of the assembly is scaffolded into 32 chromosomal pseudomolecules, including the assembled Z sex chromosome. The mitochondrial genome has also been assembled and is 15.3 kilobases in length. Gene annotation of this assembly on Ensembl identified 18,717 protein coding genes.

Keywords

Eupithecia centaureata, Lime-speck 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; **Fagan S**: Writing – Original Draft Preparation;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (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, Fagan S, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Lime-speck Pug, *Eupithecia centaureata* (Denis & Schiffermüller, 1775) [version 1; peer review: awaiting peer review]** Wellcome Open Research 2023, **8**:132 https://doi.org/10.12688/wellcomeopenres.19249.1

First published: 23 Mar 2023, **8**:132 https://doi.org/10.12688/wellcomeopenres.19249.1

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Geometroidea; Geometridae; Larentiinae; *Eupithecia*; *Eupithecia centaureata* (Denis & Schiffermüller, 1775) (NCBI:txid934844).

Background

The Lime-speck Pug, *Eupithecia centaureata* (Denis & Schiffermüller, 1775), is a moth belonging to the family Geometridae. The species has a distribution throughout the entire Palearctic region. In Britain, the Lime-speck Pug is considered common and widespread, chiefly around coastal areas such as Cornwall, but also as far north as Scotland and the Outer Hebrides (Waring *et al.*, 2017). Its status is considered of least concern (Fox *et al.*, 2019).

The *E. centaureata* species is relatively small, with a wingspan of 20–25 mm, and it has a distinctive appearance. The wings are primarily white in colour, with patches of black on the leading edge of the forewings (Waring *et al.*, 2017). It is believed that its visual resemblance to bird excrement could offer an advantage against predation, though the species is often attacked by parasitoid insects (Riley & Prior, 2003).

Adults are on the wing from April to September and have two overlapping generations, particularly in the southern populations (Riley & Prior, 2003). The species can be found in various open habitats, including urban gardens and hedgerows, often seen at rest on vertical surfaces, such as sheds (Waring *et al.*, 2017). Larvae feed on the flowers of low-growing, herbaceous plants during the summer and autumn, before overwintering as pupae in loose earth (Skinner & Wilson, 2009).

The genome of the Lime-speck Pug, *Eupithecia centaureata*, 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.

Genome sequence report

The genome was sequenced from one male *Eupithecia centaureata* (Figure 1) collected from Wytham Woods, Oxfordshire, UK (latitude 51.78, longitude –1.32). A total of 22-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 79-fold coverage in 10X Genomics read clouds was generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 137 missing joins or mis-joins and removed 66 haplotypic duplications, reducing the assembly length by 4.33% and the scaffold number by 23.33%, and increasing the scaffold N50 by 1.24%.

The final assembly has a total length of 465.6 Mb in 161 sequence scaffolds with a scaffold N50 of 15.0 Mb (Table 1). Most (96.5%) of the assembly sequence was assigned to 32 chromosomal-level scaffolds, representing 31 autosomes, and the Z sex chromosome. Chromosome-scale scaffolds 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 55.7, with *k*-mer completeness of 99.99%, and the assembly has a BUSCO v5.3.2 (Manni *et al.*, 2021) completeness of 97.9% (single 97.2%, duplicated 0.7%), using the lepidoptera_odb10 reference set ($n = 5{,}286$).

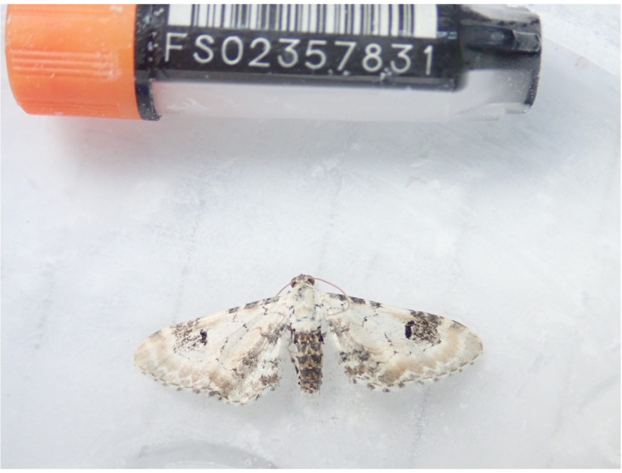

Figure 1. Photograph of the *Eupithecia centaureata* (ilEupCent1) specimen used for genome sequencing.

Genome annotation report

The *Eupithecia centaureata* genome assembly GCA_944547425.1 (ilEupCent1.1) was annotated using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Eupithecia_centaureata_GCA_944547425.1/Info/Index/). The resulting annotation includes 18,964 transcribed mRNAs from 18,717 protein-coding genes.

Methods

Sample acquisition and nucleic acid extraction

Two *Eupithecia centaureata* specimens (ilEupCent1 and ilEupCent2) were collected in Wytham Woods, Oxfordshire (biological vice-county Berkshire), UK (latitude 51.78, longitude –1.32) on 24 August 2019. 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 preserved on dry ice.

DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute (WSI). The ilEupCent1 sample was weighed and dissected on dry ice. 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. Low molecular weight DNA was removed from a 20 ng aliquot of extracted DNA using the 0.8X AMpure XP purification kit prior to 10X Chromium sequencing; a minimum of 50 ng DNA was submitted for 10X sequencing. HMW DNA was

Table 1. Genome data for *Eupithecia centaureata*, ilEupCent1.1.

Project accession data		
Assembly identifier	ilEupCent1.1	
Species	*Eupithecia centaureata*	
Specimen	ilEupCent1	
NCBI taxonomy ID	934844	
BioProject	PRJEB52801	
BioSample ID	SAMEA7520186	
Isolate information	ilEupCent1, male, whole organism (DNA sequencing) ilEupCent2, whole organism (Hi-C scaffolding)	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	55.7	*≥ 50*
k-mer completeness	99.99%	*≥ 95%*
BUSCO**	C:97.9%[S:97.2%,D:0.7%], F:0.5%,M:1.6%,n:5,286	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	96.5%	*≥ 95%*
Sex chromosomes	Z chromosome	*localised homologous pairs*
Organelles	Mitochondrial genome assembled.	*complete single alleles*
Raw data accessions		
PacificBiosciences SEQUEL II	ERR9763979	
10X Genomics Illumina	ERR9730869–ERR9730872	
Hi-C Illumina	ERR9730873	
Genome assembly		
Assembly accession	GCA_944547425.1	
Accession of alternate haplotype	GCA_944548335.1	
Span (Mb)	465.6	
Number of contigs	547	
Contig N50 length (Mb)	1.6	
Number of scaffolds	161	
Scaffold N50 length (Mb)	15.0	
Longest scaffold (Mb)	21.9	
Genome annotation		
Number of protein-coding genes	18,717	
Number of gene transcripts	18,964	

* 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/ilEupCent1.1/dataset/CALYMT01/busco.

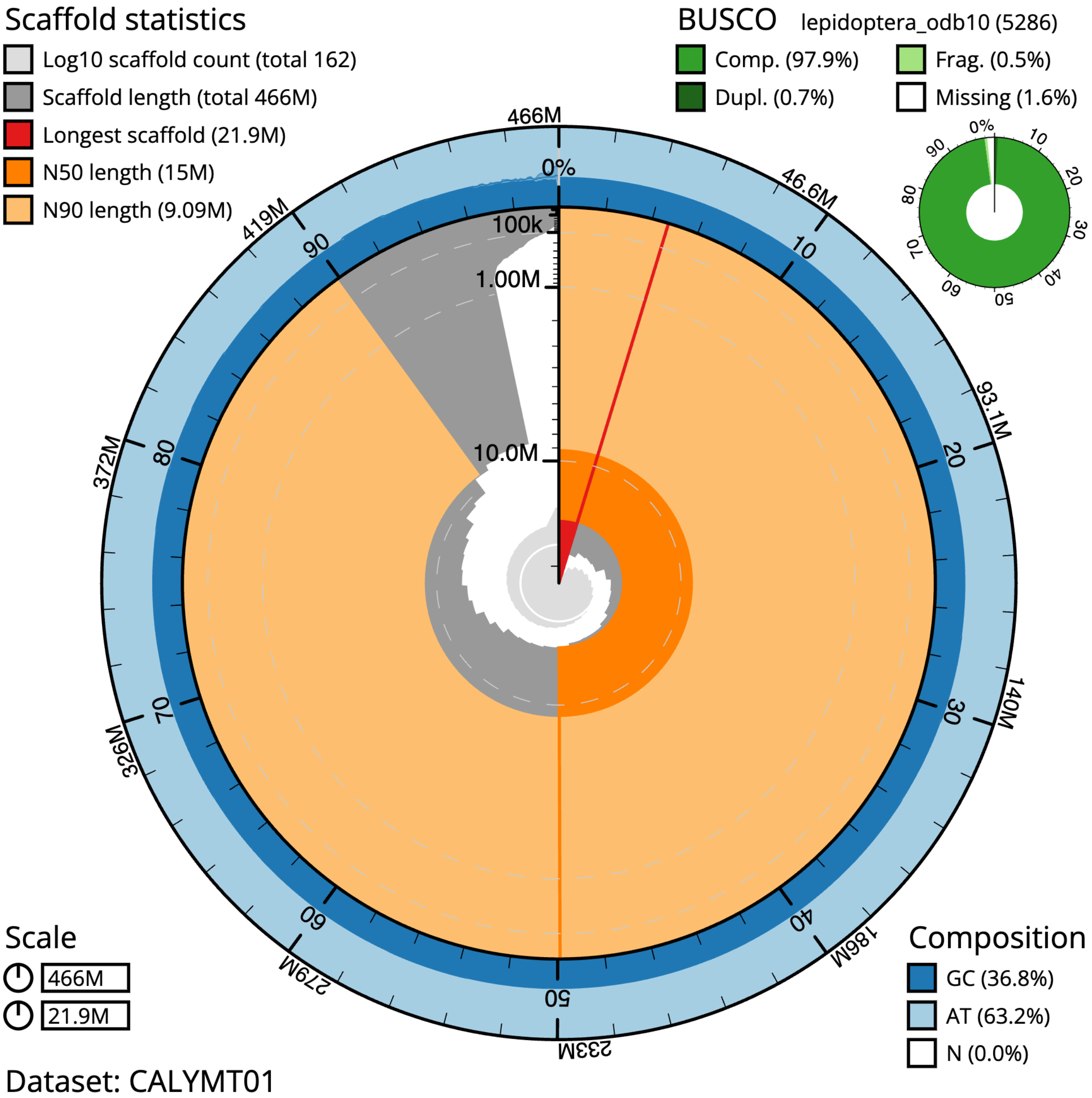

Figure 2. Genome assembly of *Eupithecia centaureata*, ilEupCent1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 465,593,408 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (21,918,189 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (15,011,971 and 9,086,237 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/ilEupCent1.1/dataset/CALYMT01/blob.

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.

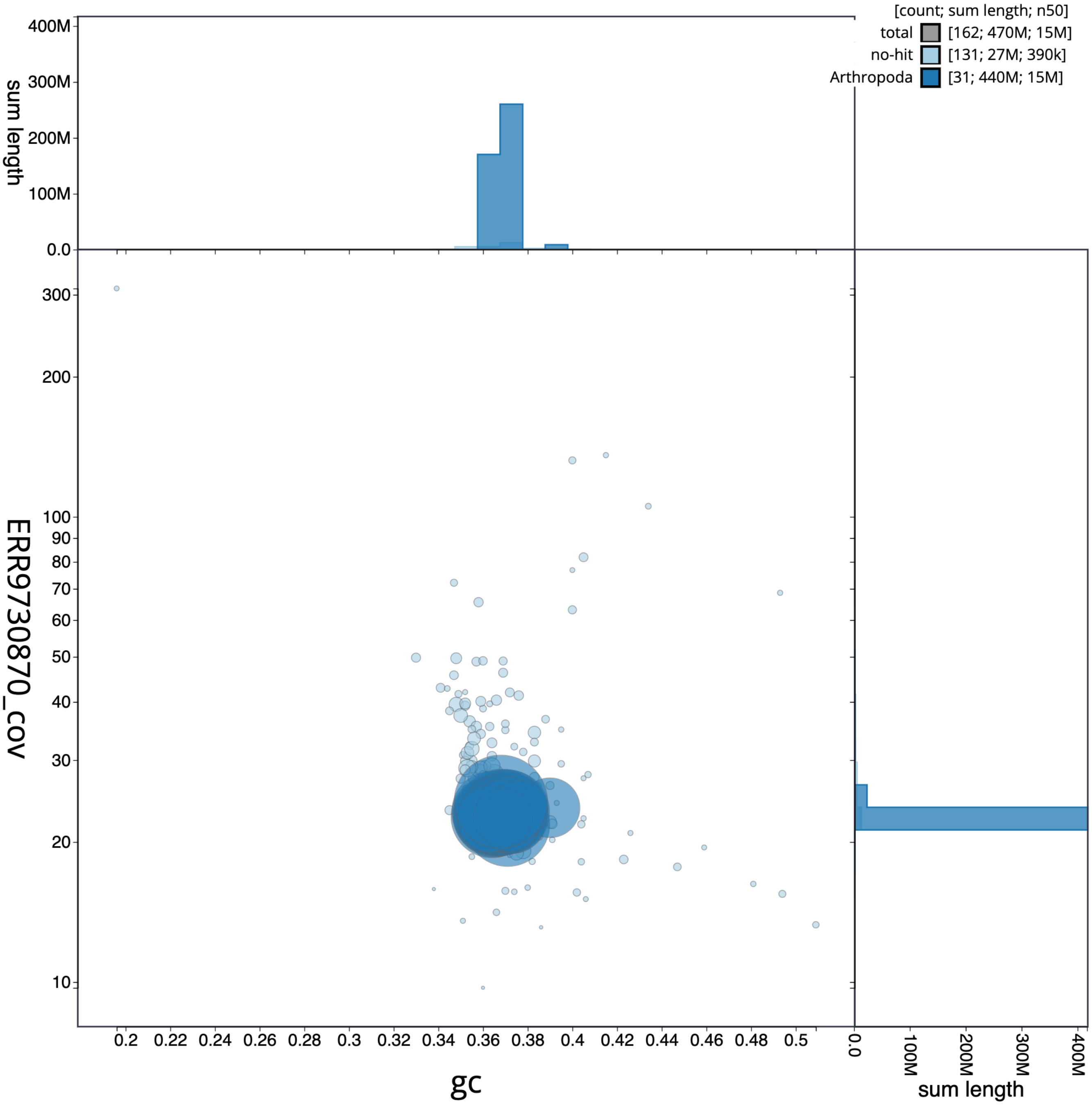

Figure 3. Genome assembly of *Eupithecia centaureata*, ilEupCent1.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilEupCent1.1/dataset/CALYMT01/blob.

Sequencing

Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud DNA sequencing libraries were constructed according to the manufacturers' instructions. DNA sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi) and HiSeq X Ten (10X) instruments. Hi-C data were also generated from tissue of ilEupCent2 using the Arima v1 kit and sequenced on the HiSeq X Ten 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). One round of polishing

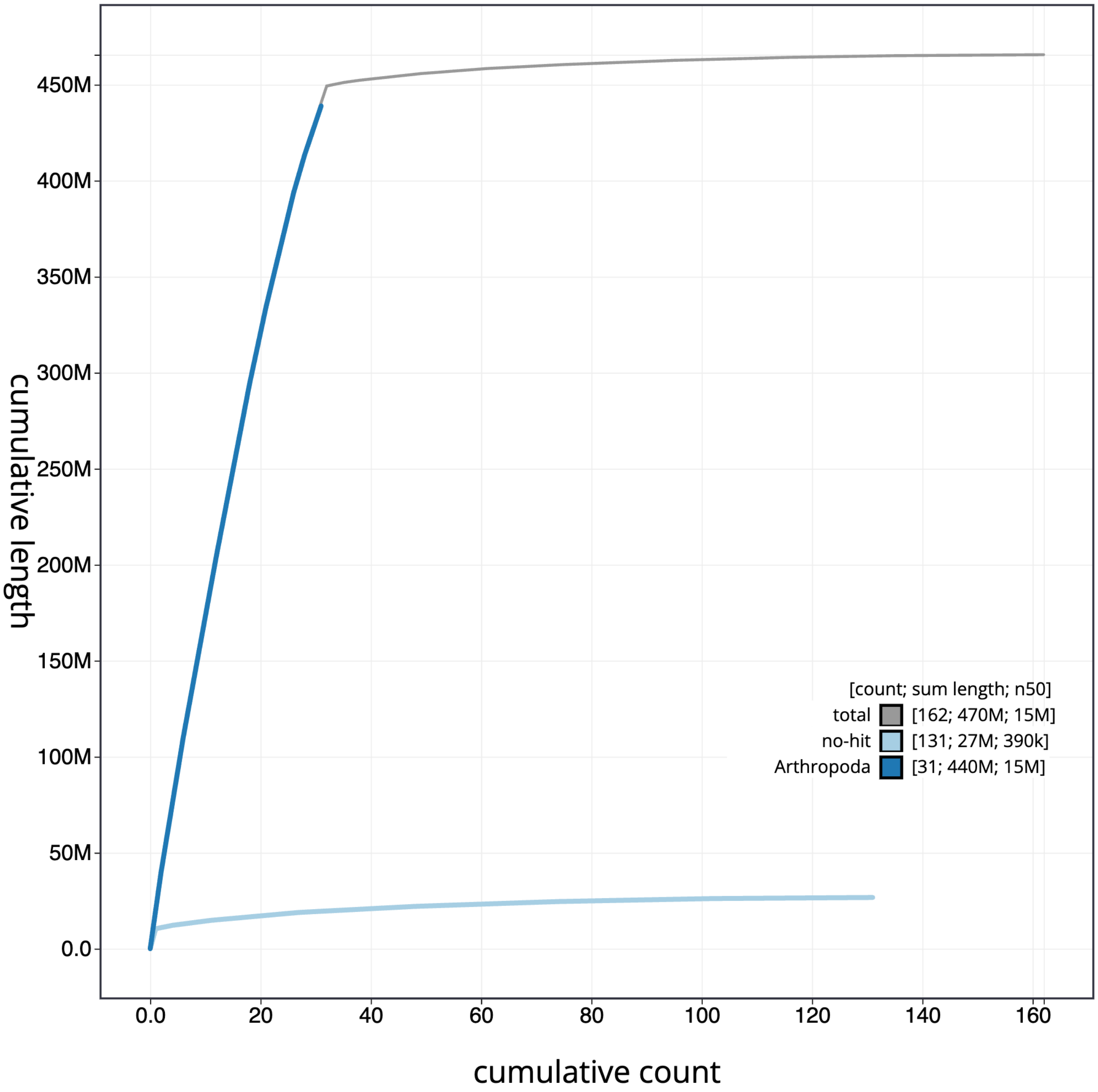

Figure 4. Genome assembly of *Eupithecia centaureata*, ilEupCent1.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilEupCent1.1/dataset/CALYMT01/cumulative.

was performed by aligning 10X Genomics read data to the assembly with Long Ranger ALIGN, calling variants with FreeBayes (Garrison & Marth, 2012). The assembly was then 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.

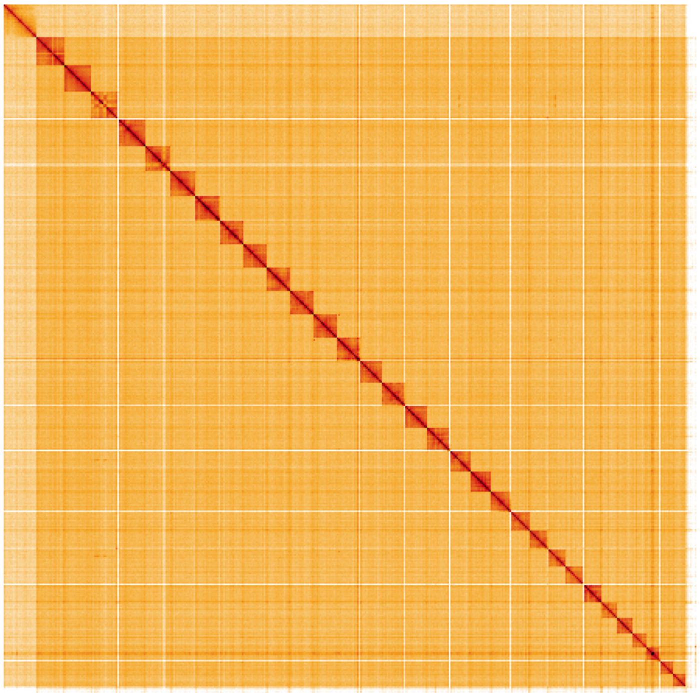

Figure 5. Genome assembly of *Eupithecia centaureata*, ilEupCent1.1: Hi-C contact map. Hi-C contact map of the ilEupCent1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=ZNwA48_HTUuX75UFZw-0oQ.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Eupithecia centaureata*, ilEupCent1.

INSDC accession	Chromosomes	Size (Mb)	GC%
OX155676.1	1	18.43	36.8
OX155677.1	2	17.86	37
OX155678.1	3	17.7	37.1
OX155679.1	4	17.57	37.1
OX155680.1	5	16.65	36.4
OX155681.1	6	16.18	36.5
OX155682.1	7	16.08	37.1
OX155683.1	8	15.93	36.4
OX155684.1	9	15.56	36.9
OX155685.1	10	15.33	36.8
OX155686.1	11	15.33	36.7
OX155687.1	12	15.28	36.5
OX155688.1	13	15.01	36.4
OX155689.1	14	14.95	36.6
OX155690.1	15	14.83	37.1
OX155691.1	16	14.72	36.8
OX155692.1	17	14.59	36.8
OX155693.1	18	14.01	36.7
OX155694.1	19	13.48	37
OX155695.1	20	12.75	37
OX155696.1	21	12.45	36.7
OX155697.1	22	12	36.7
OX155698.1	23	11.86	37.3
OX155699.1	24	11.75	36.4
OX155700.1	25	11.43	37.2
OX155701.1	26	10.66	36.3
OX155702.1	27	10.4	37
OX155703.1	28	9.09	37.2
OX155704.1	29	8.68	39
OX155705.1	30	8.44	37.6
OX155706.1	31	8.4	36.9
OX155675.1	Z	21.92	36.8
OX155707.1	MT	0.02	20

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
FreeBayes	1.3.1-17-gaa2ace8	https://github.com/freebayes/freebayes
Hifiasm	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	1.11.6	https://github.com/higlass/higlass
Long Ranger ALIGN	2.2.2	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines
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

Genome annotation

The BRAKER2 pipeline (Brůna *et al.*, 2021) was used in the default protein mode to generate annotation for the *Eupithecia centaureata* assembly (GCA_944547425.1) 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.

Data availability

European Nucleotide Archive: *Eupithecia centaureata* (limespeck pug). Accession number PRJEB52801; https://identifiers.org/ena.embl/PRJEB52801. (Wellcome Sanger Institute, 2022)

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

Fox R, Parsons MS, Harrower CA: **A review of the status of the macro-moths of Great Britain.** Dorset, UK: Butterfly Conservation. 2019.
Reference Source

Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing.** 2012.
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**

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): 1665–80.
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**

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): 3210–2.
PubMed Abstract | **Publisher Full Text**

Skinner B, Wilson D: **Colour Identification Guide to the Moths of the British Isles.** Leiden, Netherlands: BRILL. 2009.
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.* [Preprint], 2022.
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 Lime-speck Pug *Eupithecia centaureata* (Denis & Schiffermüller, 1775).** European Nucleotide Archive. [dataset], accession number PRJEB52801, 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**