DATA NOTE

The genome sequence of the scotch argus butterfly, *Erebia aethiops* (Esper, 1777) [version 1; peer review: 2 approved]

Oskar Lohse¹, Konrad Lohse²,
Wellcome Sanger Institute Tree of Life programme,
Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective,
Tree of Life Core Informatics collective, Hannah Augustijnen³, Kay Lucek³,
Darwin Tree of Life Consortium

¹Independent collector, Edinburgh, UK
²Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
³Department of Environmental Sciences, University of Basel, Basel, Switzerland

**Abstract**

We present a genome assembly from an individual female *Erebia aethiops* (the scotch argus; Arthropoda; Insecta; Lepidoptera; Nymphalidae). The genome sequence is 473 megabases in span. The complete assembly is scaffolded into 20 chromosomal pseudomolecules, with the W and Z sex chromosomes assembled. The complete mitochondrial genome was also assembled and is 15.2 kilobases in length.

**Keywords**

Erebia aethiops, scotch argus, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.
Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: Lohse O: Investigation, Resources; Lohse K: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; Augustijnen H: Writing – Original Draft Preparation; Lucek K: 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) and the Darwin Tree of Life Discretionary Award (218328). KL is supported by an ERC grant (ModelGenom Land 757648) and a NERC fellowship (NE/L011522/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Lohse O 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: Lohse O, Lohse K, Wellcome Sanger Institute Tree of Life programme et al. The genome sequence of the scotch argus butterfly, Erebia aethiops (Esper, 1777) [version 1; peer review: 2 approved] Wellcome Open Research 2022, 7:217 https://doi.org/10.12688/wellcomeopenres.17927.1

First published: 19 Aug 2022, 7:217 https://doi.org/10.12688/wellcomeopenres.17927.1
Species taxonomy
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Papilionoidea; Nymphalidae; Satyrinae; Erebiini; Erebia; Erebia aethiops (Esper, 1777) (NCBI:txid447833).

Background
The Scotch argus, Erebia aethiops (Esper, 1777) has a wide distribution in the Palaearctic from Scotland to western Siberia and the Altai Mountains (Wendt et al., 2021). Unlike most other Erebia species, E. aethiops occurs in the lowland and montane zone. The species was first described from Scotland as the subspecies E. aethiops caledonia (Newland, 2012), though this taxonomy now refers only to populations in the west and southwest of Scotland (Newland, 2012; Thomson, 1980). Populations in the north and southeast of Scotland belong to the nominate subspecies E. aethiops aethiops (Thomas & Lewington, 2016). While the two subspecies differ in their larval foodplant preference and wing morphology, with caledonia having narrower forewings and a narrower orange band, their taxonomic status remains disputed (Kirkland, 1995).

In general, E. aethiops prefers meadows near forested areas and open woodlands (Loertscher, 1991); (Slamova et al., 2011; Wendt et al., 2021) with evidence for sex-specific preference in meso- and microhabitat use (Slamova et al., 2011; Slamova et al., 2013). E. aethiops is univoltine, with hibernating larvae and an adult flight period from mid-July to mid-August. Larvae feed on a wide range of grasses, including Bromus erectus, Brachypodium pinnatum and, in the UK, Molinia caerulea and Sesleria caerulea (Slamova et al., 2013; Thomas & Lewington, 2016). The species may be vulnerable to anthropogenic habitat fragmentation (Slamova et al., 2013; Wendt et al., 2021). Although UK populations have seen declines and northward range shifts over the last decades (Franco et al., 2006) and E. aethiops is now listed as Vulnerable on the UK Red List (Fox et al., 2022), it is listed as a species of Least Concern on the IUCN Red List of Europe (van Swaay et al., 2010). The karyotype of E. aethiops was first described as consisting of 21 chromosomes based on a single individual from Croatia (Lorković, 1941). Although we do not know whether this chromosome count included a W, it is inconsistent with the 20 chromosomal scaffolds of this assembly (Table 2).

Genome sequence report
The genome was sequenced from a single female E. aethiops (Figure 1) collected from Carrifran Wildwood, Scotland (latitude 55.4001, longitude -3.3352). A total of 35-fold coverage in Pacific Biosciences single-molecule circular consensus (HiFi) long reads and 61-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 30 missing/misjoins and removed two haplotypic duplications, reducing the assembly size by 0.04% and the scaffold number by 23.94%, and increasing the scaffold N50 by 8.42%.

The final assembly has a total length of 473 Mb in 54 sequence scaffolds with a scaffold N50 of 25.9 Mb (Table 1). The complete assembly sequence was assigned to 20 chromosomal-level scaffolds, representing 18 autosomes (numbered by sequence length), and the W and Z sex chromosomes (Figure 2–Figure 5; Table 2). The assembly has a BUSCO v5.1.2 (Manni et al., 2021) completeness of 98.5% (single
Table 1. Genome data for *Erebia aethiops*, ilEreAeth2.1.

| Project accession data |   |
|------------------------|--|
| Assembly identifier     | ilEreAeth2.1 |
| Species                | *Erebia aethiops* |
| Specimen               | ilEreAeth2 (genome assembly); ilEreAeth1 (additional HiFi, 10X reads); ilEreAeth3 (Hi-C) |
| NCBI taxonomy ID        | 447833 |
| BioProject             | PRJEB47324 |
| BioSample ID           | SAMEA7523289 |
| Isolate information    | Female, whole organisms (ilEreAeth2, ilEreAeth1); male, whole organism (ilEreAeth3) |

| Raw data accessions   |   |
|-----------------------|--|
| PacificBiosciences SEQUEL II | ERR6808048 (ilEreAeth2); ERR6636094-ERR6636096, ERR6808047 (ilEreAeth1) |
| 10X Genomics Illumina | ERR6688769-ERR6688772 (ilEreAeth2); ERR6688764-ERR6688767 (ilEreAeth1) |
| Hi-C Illumina        | ERR6688768 (ilEreAeth3) |

| Genome assembly   |   |
|-------------------|---|
| Assembly accession | GCA_923060345.1 |
| Accession of alternate haplotype | GCA_923062935.1 |
| Span (Mb)          | 473 |
| Number of contigs  | 80 |
| Contig N50 length (Mb) | 21.4 |
| Number of scaffolds| 54 |
| Scaffold N50 length (Mb) | 25.9 |
| Longest scaffold (Mb) | 33.25 |
| BUSCO* genome score | C:98.5%[S:97.8%,D:0.7%],F:0.4%,M:1.1%,n:5286 |

*BUSCO scores based on the lepidoptera_odb10 BUSCO set using v5.1.2. 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/ilEreAeth2.1/dataset/CAKLPR01/busc](https://blobtoolkit.genomehubs.org/view/ilEreAeth2.1/dataset/CAKLPR01/busc).

97.8% duplicated (0.7%) using the lepidoptera_odb10 reference set (n=5,286). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

**Methods**

**Sample acquisition and nucleic acid extraction**

Two female *E. aethiops* specimens (ilEreAeth2, genome assembly; ilEreAeth1, additional HiFi and 10X reads) and a male (ilEreAeth3, Hi-C) were collected from Carrifran Wildwood, Scotland (latitude 55.4001, longitude -3.3352) using a net by Oskar and Konrad Lohse, who also identified the samples. Specimens were snap-frozen at -80°C.

DNA was extracted in the Tree of Life Laboratory at the Wellcome Sanger Institute. Whole organism tissue of ilEreAeth2 and ilEreAeth1 was cryogenically disrupted to a fine powder using a Covaris cryoPREP Automated Dry Pulveriser, receiving multiple impacts. Fragment size analysis of 0.01–0.5 ng of DNA was then performed using an Agilent FemtoPulse. High molecular weight (HMW) DNA was extracted using the Qiagen MagAttract HMW DNA extraction kit. Low molecular weight DNA was removed from a 200-ng aliquot of extracted DNA using 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 sheared into an average fragment size between 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
Figure 2. Genome assembly of *Erebia aethiops*, ilEreAeth2.1: 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 473,469,105 bp assembly. The distribution of chromosome lengths is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly (37,954,409 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (25,856,419 and 17,052,335 bp), respectively. The pale grey spiral shows the cumulative chromosome 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/ilEreAeth2.1/dataset/CAKLPR01/snail.

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 Erebia aethiops, iEreAeth2.1: 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/iEreAeth2.1/dataset/CAKLPR01/blob.

Sequencing
Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud DNA sequencing libraries were constructed for iEreAeth2 and iEreAeth1 according to the manufacturers’ instructions. DNA sequencing was performed by the Scientific Operations core at the WSI on Pacific
Figure 4. Genome assembly of *Erebia aethiops*, ilEreAeth2.1: 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/ilEreAeth2.1/dataset/CAKLPR01/cumulative.

Biosciences SEQUEL II, Illumina HiSeq X (ilEreAeth1, 10X) and Illumina NovaSeq 6000 (ilEreAeth2, 10X) instruments. Hi-C data were also generated from remaining whole organism tissue of ilEreAeth3 using the Arima v2 Hi-C kit and sequenced on an Illumina NovaSeq 6000 instrument.

Genome assembly
Assembly was carried out with Hifiasm (Cheng *et al.*, 2021); haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2020). One round of polishing was performed by aligning 10X Genomics read data to the
assembly with longranger align, calling variants with freebayes (Garrison & Marth, 2012). The assembly was then scaffolded with Hi-C data (Rao et al., 2014) using SALSA2 (Ghurye et al., 2019). The assembly was checked for contamination and corrected using the gEVAL system (Chow et al., 2016) as described previously (Howe et al., 2021). Manual curation (Howe et al., 2021) was performed using gEVAL, HiGlass (Kerpedjiev et al., 2018) and Pretext. The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva et al., 2021), which performs annotation using MitoFinder (Allio et al., 2020). The genome was analysed and BUSCO scores generated within the BlobToolKit environment (Challis et al., 2020). Table 3 contains a list of all software tool versions used, where appropriate.
Table 2. Chromosomal pseudomolecules in the genome assembly of *Erebia aethiops*, llEreAeth2.1.

| INSDC accession | Chromosome | Size (Mb) | GC%  |
|-----------------|------------|-----------|------|
| OV281080.1      | 1          | 33.25     | 37.1 |
| OV281081.1      | 2          | 32.76     | 37.1 |
| OV281082.1      | 3          | 32.72     | 37.2 |
| OV281083.1      | 4          | 30.44     | 37.1 |
| OV281084.1      | 5          | 30.01     | 37.5 |
| OV281085.1      | 6          | 26.26     | 37.4 |
| OV281086.1      | 7          | 25.86     | 37.4 |
| OV281087.1      | 8          | 23.96     | 37.2 |
| OV281088.1      | 9          | 20.72     | 37.3 |
| OV281089.1      | 10         | 20.45     | 37.1 |
| OV281090.1      | 11         | 20.15     | 37.3 |
| OV281091.1      | 12         | 19.45     | 37.3 |
| OV281092.1      | 13         | 19.3      | 37.3 |
| OV281093.1      | 14         | 18.42     | 37.2 |
| OV281094.1      | 15         | 17.95     | 37.3 |
| OV281095.1      | 16         | 17.05     | 37.3 |
| OV281096.1      | 17         | 15.92     | 37.4 |
| OV281097.1      | 18         | 15.76     | 37.7 |
| OV281098.1      | W          | 3.11      | 37.7 |
| OV281079.1      | Z          | 3.795     | 36.8 |
| OV281099.1      | MT         | 0.02      | 19.6 |
| -               | Unplaced   | 11.97     | 37.5 |

Table 3. Software tools used.

| Software tool   | Version          | Source                                      |
|-----------------|------------------|---------------------------------------------|
| Hifiasm         | 0.15.3-r339      | Cheng *et al.*, 2021                       |
| purge_dups      | 1.2.3            | Guan *et al.*, 2020                        |
| SALSA2          | 2.2              | Ghurrye *et al.*, 2019                     |
| longranger align| 2.2.2            | https://support.10xgenomics.com/...          |
| freebayes       | 1.3.1-17...      | Garrison & Marth, 2012                     |
| MitoHiFi        | 2.0              | Uliano-Silva *et al.*, 2021                |
| HiGlass         | 1.11.6           | Kerpedjiev *et al.*, 2018                  |
| PretextView     | 0.2.x            | https://github.com/wtsi-hpag/PretextView    |
| BlobToolKit     | 3.0.5            | Challis *et al.*, 2020                     |
Data availability
European Nucleotide Archive: Erebia aethiops (Scotch argus). Accession number PRJEB47324; https://identifiers.org/ena.embl/PRJEB47324.

The genome sequence is released openly for reuse. The *E. aethiops* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project. The genome will be annotated and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1.

Author information
Members of the Wellcome Sanger Institute Tree of Life programme are listed here: https://doi.org/10.5281/zenodo.6125027.

Members of Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.5746904.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.6125046.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.5638618.

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. Published Abstract | Publisher Full Text | Free Full Text |
| --- | --- | --- |
| Chaliss R, Richards E, Rajan J, et al.: BlobToolKit - Interactive Quality Assessment of Genome Assemblies. G3 (Bethesda) 2020; 10(4): 1361–74. Published Abstract | Publisher Full Text | Free Full Text |
| Cheng H, Concepcion GT, Feng X, et al.: Haplotyping-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm. Nat Methods. 2021; 18(2): 170–75. Published Abstract | Publisher Full Text | Free Full Text |
| Chow W, Bruggker K, Caccamo M, et al.: eVAl - a web-based browser for evaluating genome assemblies. Bioinformatics. 2016; 32(16): 2508–10. Published Abstract | Publisher Full Text | Free Full Text |
| Fox R, Dennis EB, Brown AF, et al.: A Revised Red List of British Butterflies. Insect Conservation and Diversity / Royal Entomological Society of London. May. 2022. Publisher Full Text |
| Franco AMA, Hill JK, Kitschke C, et al.: Impacts of Climate Warming and Habitat Loss on Extinctions at Species’ Low-Latitude Range Boundaries. Global Change Biology. 2006; 12(8): 1545–53. Publisher Full Text |
| Garrison E, Marth G: Haplotype-Based Variant Detection from Short-Read Sequencing. 2012; arXiv:1207.3907. Reference Source |
| Ghurye J, Rhee A, Walenz BP, et al.: Integrating Hi-C Links with Assembly Graphs for Chromosome-Scale Assembly. PLoS Comput Biol. 2019; 15(8): e1007273. Published 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): 2896–98. Published Abstract | Publisher Full Text | Free Full Text |
| Howe K, Chow W, Collins J, et al.: Significantly Improving the Quality of Genome Assemblies through Curation. Gigascience. 2021; 10(1): giaa153. Published Abstract | Publisher Full Text | Free Full Text |
| Kerpedjiev P, Abdennur N, Lekshas F, et al.: HiGlass: Web-Based Visual Exploration and Analysis of Genome Interaction Maps. Genome Biol. 2018; 19(1): 125. Published Abstract | Publisher Full Text | Free Full Text |

Kirkland KL: A Review of the Distribution, Ecology and Behaviour of the Scotch Argus. Brit Ecol Soc Bull. 1995; 26: 95–102.

Loertscher M: Population Biology of Two Satyrine Butterflies Erebia Meolans (de Prunner, 1798) and Erebia Aethiops (Esper, 1777) (Lepidoptera, Satyrinae). 1991; 2: 22–31.

Lorkovic Z: Die Chromosomenszahlen in Der Spermatogenese Der Tagfalter. Chromosoma. 1941; 2(1): 155–91. Publisher 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): 4647–54. Published Abstract | Publisher Full Text | Free Full Text |

Newland DE: Variation in Erebia Aethiops Caledonia Viricy, 1911 (Lepidoptera: Nymphalidae, Satyrinae). 2012; 63(4): 246–47.

Rao SS, 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. Published Abstract | Publisher Full Text | Free Full Text |

Slamova I, Klecka J, Konvicka M: Diurnal Behavior and Habitat Preferences of Erebia Aethiops, an Aberrant Lowland Species of a Mountain Butterfly Clade. Insect Behav. 2011; 24(3): 230–46. Publisher Full Text |

Slamova I, Klecka J, Konvicka M: Woodland and Grassland Mosaic from a Butterfly Perspective: Habitat Use by Erebia Aethiops (Lepidoptera: Satyrinae). Insect Conservation and Diversity / Royal Entomological Society of London. 2013; 6(3): 243–54. Publisher Full Text |

Thomas J, Lewington R: The Butterflies of Britain and Ireland. Bloomsbury, London. 2016.

Thomson GD: Butterflies of Scotland. Croom Helm, 1980. Reference Source |

Uliano-Silva M, Nunes JG, Krasheninnikova K, et al.: marcelauliano/MitoHiFi: mitohifi_v2.0. 2021. Publisher Full Text |

van Swaay C, Cuthelod A, Collins S, et al.: European Red List of Butterflies. 2010. Reference Source |

Wendt M, Husemann M, Kramp K, et al.: Reconstruction of Forest Dynamics in the Western Palaearctic Based on Phylogeographic Analysis of the Ringlet Butterfly Erebia Aethiops. Sci Rep. 2021; 11(1): 201. Published Abstract | Publisher Full Text | Free Full Text |
Jing Zhang
Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA

Authors present the high-quality chromosome-level genome assembly for *Erebia aethiops*. The assembly includes W chromosome which is rarely included in other Lepidoptera genomes. This high quality assembly will benefit future studies of *E. aethiops* and other related insects. The assembly pipeline is described in detail and also repeated by the Darwin Tree of Life (DToL) project for numerous other high quality assemblies too.

**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, Computational biology, 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.
Wei Zhang
State Key Laboratory of Protein and Plant Gene Research, Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, School of Life Sciences, Peking University, Beijing, China

Lohse et al. report a chromosome-level genome assembly of the scotch argus butterflies with both W and Z chromosomes assembled. This genome assembly has a considerable N50 value and good BUSCO statistics. The methods section is well written as well. I think the overall quality of this genome assembly is valid and will benefit relevant studies. Therefore, I recommend this manuscript be accepted 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: 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.