DATA NOTE

The genome sequence of the Arran brown, *Erebia ligea* (Linnaeus, 1758) [version 1; peer review: 2 approved]

Konrad Lohse\(^1\), Alex Hayward\(^2\), Dominik R. Laetsch\(^1\), Roger Vila\(^3\), Kay Lucek\(^4\), 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

\(^1\)Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
\(^2\)College of Life and Environmental Sciences, Department of Biosciences, University of Exeter, Exeter, UK
\(^3\)Institut de Biologia Evolutiva, CSIC - Universitat Pompeu Fabra, Barcelona, Spain
\(^4\)Department of Environmental Sciences, University of Basel, Basel, Switzerland

First published: 13 Oct 2022, 7:259
https://doi.org/10.12688/wellcomeopenres.18115.1

Abstract
We present a genome assembly from an individual male *Erebia ligea* (Arran brown; Arthropoda; Insecta; Lepidoptera; Nymphalidae). The genome sequence is 506 megabases in span. The majority (99.92%) of the assembly is scaffolded into 29 chromosomal pseudomolecules, with the Z sex chromosome assembled. The complete mitochondrial genome was also assembled and is 15.2 kilobases in length.

Keywords
Erebia ligea, Arran brown, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

| Approval Status | 1 | 2 |
|-----------------|---|---|
| version 1       | ✓ | ✓ |
| 13 Oct 2022     | view | view |

1. Markus Moest, University of Innsbruck, Innsbruck, Austria

2. Manuela Lopez Villavicencio, Museum National d'Histoire Naturelle Reseau des bibliotheques du museum, Paris, France

Any reports and responses or comments on the article can be found at the end of the article.
**Species taxonomy**
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Papilionoidea; Nymphalidae; Satyrinae; Satyrini; Erebiina; Erebia; Erebia ligea (Linnaeus, 1758) (NCBI: txid111903).

**Background**
The Arran brown, *Erebia ligea*, is one of the most widespread species of the genus *Erebia*, occurring from the Russian Kamchatka Peninsula and Japan in eastern Asia (Dubatolov et al., 1998) to central and northern Europe (Kudrna et al., 2015). Although the species takes its common name from the Isle of Arran in Scotland, where it was first recorded in 1803, the current and historic presence of this butterfly in the British Isles remains disputed (Salmon, 1995). The intraspecific phenotypic diversity present throughout the distribution of *E. ligea* has triggered the description of several subspecies (Dubatolov et al., 1998; Warren, 1937; Zakharova & Tatarinov, 2016), however, a formal biogeographic assessment remains lacking.

*E. ligea* is characterised as a woodland species associated with clearings and meadows, and occurs at relatively low altitudes compared to most other *Erebia* butterflies (Kleckova et al., 2014). Recorded host plants include a variety of grasses (Poaceae) and sedges (*Carex*, Cyperaceae). It is univoltine and in some northern localities it is recorded only every second year (Tolman & Lewington, 2008). Although *E. ligea* is considered a species of Least Concern according to the IUCN Red List (Europe) (van Swaay et al., 2010), the species can be locally endangered (Fichefet et al., 2008).

While the first karyotypic analysis suggested that male *Erebia ligea* from Finland have 29 chromosomes (Federley, 1938), Japanese individuals from Hokkaido were found to have only 28 chromosomes (Saitoh & Abe, 1997). These values are close to the most common and putatively ancestral chromosomal state for Lepidoptera (n=31; Robinson, 1971), although *Erebia* is one of the most karyologically diverse known genera of butterflies (Robinson, 1971; de Vos et al., 2020).

**Genome sequence report**
The genome was sequenced from a single male *E. ligea* (Figure 1) collected from Borzont, Joseni, Harghita, Romania (latitude 46.664, longitude 25.317). A total of 34-fold coverage of Pacific Biosciences single-molecule circular consensus (HiFi) long reads and 63-fold coverage of 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 47 missing/misjoins and removed 10 haplotypic duplications, reducing the assembly length by 3.59% and the scaffold number by 39.39%, and increased the scaffold N50 by 4.29%.

The final assembly has a total length of 506 Mb in 40 sequence scaffolds, with a scaffold N50 of 19.1 Mb (Table 1). The majority, 99.92%, of assembly sequence was assigned to 40 chromosomal-level scaffolds, representing 28 autosomes

---

*BUSCO scores based on the lepidoptera_odb10 BUSCO set using v5.2.2. 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/ilEreLige1.2/dataset/CAKAVA02/busco.*
(numbered by sequence length), and the Z sex chromosome (Figure 2–Figure 5; Table 2). The assembly has a BUSCO v5.2.2 (Manni et al., 2021) completeness of 97.9% (single 97.4%, duplicated 0.5%) 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
A single male *E. ligea* specimen (iEreLige1, genome assembly, HiC) was collected from Borzont, Joseni, Harghita, Romania (latitude 46.664, longitude 25.317) using a handnet by Konrad Lohse, Dominik Laetsch (both University of Edinburgh) and Alex Hayward (University of Exeter). The sample was identified by Roger Vila (Institut de Biologia Evolutiva, Barcelona) and snap-frozen from live in a dry shipper.

DNA was extracted at the Scientific Operations Core, Wellcome Sanger Institute. The iEreLige1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. Whole organism tissue was disrupted by manual grinding with a disposable pestle. 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

---

**Figure 2.** Genome assembly of *Erebia ligea*, iEreLige1.2: 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 506,397,422 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 (22,722,498 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (19,149,538 and 12,368,103 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.genomeshubs.org/view/iEreLige1.2/dataset/CAKAVA02/snail](https://blobtoolkit.genomeshubs.org/view/iEreLige1.2/dataset/CAKAVA02/snail).
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 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.

Sequencing

Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud DNA sequencing libraries were constructed according to the manufacturers’ instructions. Sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi) and Illumina HiSeq X (10X) instruments. Hi-C data were also generated from remaining whole organism tissue of ilEreLige1 using the Arima v1 Hi-C kit and sequenced on an Illumina HiSeq X (10X) instrument.

Figure 3. Genome assembly of Erebia ligea, ilEreLige1.2: 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/ilEreLige1.2/dataset/CAKAVA02/blob.
Genome assembly

Assembly was carried out with Hiiasm (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

Figure 4. Genome assembly of Erebia ligea, iEreLige1.2: 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/iEreLige1.2/dataset/CAKAVA02/cumulative.
Table 2. Chromosomal pseudomolecules in the genome assembly of *Erebia ligea*, ilEreLige1.2.

| INSDC accession | Chromosome | Size (Mb) | GC%  |
|-----------------|------------|-----------|------|
| OU785219.1      | 1          | 22.72     | 37.2 |
| OU785220.1      | 2          | 22.11     | 37.3 |
| OU785221.1      | 3          | 22.01     | 37.4 |
| OU785223.1      | 4          | 21.42     | 37.3 |
| OU785224.1      | 5          | 20.97     | 37.3 |
| OU785225.1      | 6          | 20.77     | 37.1 |
| OU785226.1      | 7          | 20.52     | 37.1 |
| OU785227.1      | 8          | 20.11     | 37.4 |
| OU785228.1      | 9          | 20.06     | 37.3 |
| OU785229.1      | 10         | 19.31     | 37.4 |
| OU785230.1      | 11         | 19.22     | 37.3 |
| OU785231.1      | 12         | 19.15     | 37.2 |
| OU785232.1      | 13         | 18.89     | 37.3 |
| OU785233.1      | 14         | 18.5      | 37.4 |
| OU785234.1      | 15         | 18.36     | 37.3 |

**Figure 5. Genome assembly of Erebia ligea, ilEreLige1.2: Hi-C contact map.** Hi-C contact map of the ilEreLige1.2 assembly, visualised in HiGlass. Chromosomes are shown in size order from left to right and top to bottom. The interactive Hi-C map can be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=L3267sJyakmh-bayAgPg.
performed 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 3. Software tools used.

| Software tool   | Version | Source                      |
|-----------------|---------|-----------------------------|
| Hifiasm         | 0.12    | Guan D, McCarthy SA, Wood 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 |
| purge_dups      | 1.2.3   | Guan et al., 2020           |
| SALSA2          | 2.2     | Ghurye et al., 2019         |
| longranger align | 2.2.2   | https://support.10xgenomics.com/ genome-exome/software/pipelines/latest/advanced/other-pipelines |
| freebayes       | 1.3.1-17-gaa2ace8 | Garrison & Marth, 2012 |
| MitoHiFi        | 1       | Uliano-Silva et al., 2021   |
| HiGlass         | 1.11.6  | Kerpedjiev et al., 2018     |
| PretextView      | 0.1.x gEVAL — a Web-Based Browser for Evaluating Genome Assemblies. Bioinformatics. 2016; 32(16): 2508-10. PubMed Abstract | Publisher Full Text | Free Full Text |
| BlobToolKit     | 3.0.5   | Challis et al., 2020        |

#### References

Allio R, Somaker-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

Challis R, Richards E, Rajan J, et al.: BlobToolkit - interactive quality assessment of genome assemblies. G3 (Bethesda). 2020; 10(4): 1361-74. 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): 170-75. PubMed Abstract | Publisher Full Text | Free Full Text

Chow W, Brugger K, Caccamo M, et al.: gEVAL — a Web-Based Browser for Evaluating Genome Assemblies. Bioinformatics. 2016; 32(16): 2508-10. PubMed Abstract | Publisher Full Text | Free Full Text

de Vos JM, Augustijnen H, Bätscher L, et al.: Speciation through Chromosomal Fusion and Fission in Lepidoptera. Philos Trans R Soc Lond B Biol Sci. 2020; 375(1806): 20190539. PubMed Abstract | Publisher Full Text | Free Full Text

Dubatolov VV, Korshunov YP, Garbunov PV, et al.: A Review of the Erebia ligea-Complex (Lepidoptera, Satyrifidae) from Eastern Asia. Lepidoptera Science. 1998; 49(3): 177-93. PubMed Abstract | Publisher Full Text | Free Full Text

Federley H: Chromosomenzahlen Finnländischer Lepidopteren: 1. Rhopalocera. Hereditas. 1938; 24(4): 397-464.

Fichefet V, Barbier Y, Baugnée JY, et al.: Papillons de jour de Wallonie: (1985-2007). Reference Source

Garrison E, Marth G: Haplotype-Based Variant Detection from Short-Read Sequencing. 2012; arXiv:1207.3907. Publisher Full Text

Ghurye J, Rihe A, Walenz BP, et al.: Integrating Hi-C Links with Assembly

Data availability

European Nucleotide Archive: *Erebia ligea* (Arran brown). Accession number PRJEB42125; https://identifiers.org/ena.embl/PRJEB42125.

The genome sequence is released openly for reuse. The *E. ligea* 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 Wellcome Sanger Institute Tree of Life programme are listed here: https://doi.org/10.5281/zenodo.6866293.

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.6418363.
Open Peer Review

Current Peer Review Status: ✔️ ✔️

Version 1

Reviewer Report 19 June 2023

https://doi.org/10.21956/wellcomeopenres.20086.r60428

© 2023 Lopez Villavicencio M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

- ✔️ Manuela Lopez Villavicencio

Museum National d'Histoire Naturelle Reseau des bibliothèques du museum, Paris, Île-de-France, France

This article reports the genome assembly of a male individual of the very widespread butterfly species *Erebia ligea*. The assembly included the autosomes, the Z sexual chromosome and the mitochondrial genome.

The article is clearly written and pleasant to read. It shows convincing evidences for a high-quality assembly based on BUSCO scores. The methods for genome assembly, quality test and Hi-C scaffolding are relevant and up-to-date. There are sufficient details of methods and materials provided to allow replication by others.

It is important to highlight that the assembly sequence was assigned to 40 chromosomal-level scaffolds, representing 28 autosomes plus the sexual Z chromosome and this is relevant because previous karyotypic analysis have found different chromosomal numbers for this species.

As I am interested in butterfly genomes, I would have loved to have other interesting information as genome-wide level of heterozygosity (estimated with jellyfish and genomescope).

Overall, I think the release of this well-assembled and annotated genome is a very useful contribution and I recommend the indexing of this article.

**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
Lohse et al. present a high-quality reference genome for a common Eurasian butterfly species. The quality of the assembly is impressive and represents a very valuable resource for future studies on biogeography and population genomics of the species as well as speciation research in the genus Erebia.

I only have few minor comments/questions:

Background:

“relatively low altitudes”... I would suggest to provide an approximate altitudinal range or range of altitudinal zones that are inhabited by this species

“in some northern localities it is recorded only every second year“...is the entire population really only present every other year or should this just mean that the completion of development can take up to 2 years?

Genome Sequence Report:

"sequence was assigned to 40 chromosomal-level scaffolds, representing 28 autosomes (numbered by sequence length), and the Z sex chromosome"... this section and the fact that in the legend in Figure2 41 scaffolds are indicated confuses me. My interpretation of "chromosomal-level scaffolds" is 1 scaffold = 1 chromosome, but that does not seem to be the case here (40 scaffolds = 28 autosomes + 1 Z) - could well be that my interpretation is the problem here. The 40 scaffolds in the text and the 41 scaffolds in Figure2 are maybe due to including/excluding the mitochondrial scaffold but I think this is not explicitly stated.
Methods:

Some more details on the specific tissues that were used for extraction would be helpful - e.g. whether only thorax or muscle tissues were used, whether the abdomen and guts were excluded, etc. as this information is relevant in order to assess the risk of potential contamination. Moreover, the description of the DNA extraction procedure is a bit unclear. For example, was the DNA used for PacBio sequencing first size selected in the same way as for the 10X sequencing, how much DNA was used for PacBio and how was the tissue set aside for Hi-C processed?

"A single male \textit{E. ligea} specimen (ilEreLige1, genome assembly, HiC) was collected from Borzont, Joseni, Harghita, Romania (latitude 46.664, longitude 25.317) using a handnet by Konrad Lohse, Dominik Laetsch (both University of Edinburgh) and Alex Hayward (University of Exeter)."...obviously an extraordinarily strong specimen...

General comment to Methods in data notes (not a comment to the authors!):

I think it would be useful to provide a link to a general but more detailed description of the - I assume standardised - sequencing methods used in the DToL project, if that's available. This would provide interested readers with some more details and background and the method sections in the notes can remain short, concise and focussed on deviations from the general protocol.

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

\textbf{Competing Interests:} No competing interests were disclosed.

\textbf{Reviewer Expertise:} Evolutionary Ecology

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.