DATA NOTE

The genome sequence of the Orange Footman, *Eilema sororcula* (Hufnagel, 1766) [version 1; peer review: 1 approved]

Douglas Boyes[1+], Owen T. Lewis[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 Jun 2023, **8**:282
https://doi.org/10.12688/wellcomeopenres.19626.1
Latest published: 28 Jun 2023, **8**:282
https://doi.org/10.12688/wellcomeopenres.19626.1

Abstract

We present a genome assembly from an individual male *Eilema sororcula* (the Orange Footman; Arthropoda; Insecta; Lepidoptera; Erebidae). The genome sequence is 729.4 megabases in span. Most of the assembly is scaffolded into 30 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled and is 15.46 kilobases in length. Gene annotation of this assembly on Ensembl identified 21,093 protein coding genes.

Keywords

Eilema sororcula, Orange Footman, genome sequence, chromosomal, Lepidoptera

Open Peer Review

Approval Status ✓

	1
version 1 28 Jun 2023	✓ view

1. **Michael Hiller**, Senckenberg Nature Research Society, Frankfurt Am Main, Germany

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Tree of Life gateway.

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

Author roles: Boyes D: Investigation, Resources; **Lewis OT**: 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) and the Darwin Tree of Life Discretionary Award (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, Lewis OT, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Orange Footman, *Eilema sororcula* (Hufnagel, 1766) [version 1; peer review: 1 approved]** Wellcome Open Research 2023, **8**:282 https://doi.org/10.12688/wellcomeopenres.19626.1

First published: 28 Jun 2023, **8**:282 https://doi.org/10.12688/wellcomeopenres.19626.1

Species taxonomy

Eukaryota; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Amphiesmenoptera; Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Noctuoidea; Erebidae; Arctiinae; Lithosiini; *Eilema*; *Eilema sororcula* (Hufnagel, 1766) (NCBI:txid987424).

Background

The Orange Footman *Eilema sororcula* had a local distribution in the south and east of the UK until the late 20th century, but like many related lichen-feeding Footman species, it has recently spread northwards and increased spectacularly in both its distribution and abundance (Randle *et al.*, 2019). At the time of writing there has been only a handful of records from Ireland (Moths Ireland, 2023), but it has been recorded across much of Europe and east across Eurasia to Korea and China (GBIF Secretariat, 2023).

The preferred habitats for *E. sororcula* are mature woodlands where its larvae feed on algae and lichens growing on trees (Henwood *et al.*, 2020). In Britain and Ireland, the adult moth is mostly observed during May and June, with the peak period of observations occurring a few weeks earlier in the year than in the 1970s (Randle *et al.*, 2019).

Here we present a chromosomally complete genome sequence for *E. sororcula* based on one male specimen from Wytham Woods, Oxfordshire, UK. A genome sequence for *E. sororcula* will facilitate studies into molecular adaptations to lichen-feeding and contribute to a growing data set of resources for understanding lepidopteran biology.

Genome sequence report

The genome was sequenced from one male *Eilema sororcula* (Figure 1) collected from Wytham Woods, Oxfordshire, UK (51.77, –1.34). A total of 43-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 93-fold coverage in 10X Genomics read clouds was generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 81 missing joins or mis-joins and removed 7 haplotypic duplications, reducing the assembly length by 0.15% and the scaffold number by 60.4%, and increasing the scaffold N50 by 5.25%.

Figure 1. Photograph of the *Eilema sororcula* (ilEilSoro1) specimen used for genome sequencing.

The final assembly has a total length of 729.4 Mb in 40 sequence scaffolds with a scaffold N50 of 25.3 Mb (Table 1). Most (99.94%) of the assembly sequence was assigned to 30 chromosomal-level scaffolds, representing 29 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 mitochondrial genome was also assembled and can be found as a contig within the multifasta file of the genome submission.

The estimated Quality Value (QV) of the final assembly is 61.3 with *k*-mer completeness of 100%, and the assembly has a BUSCO v5.3.2 completeness of 98.5% (single = 97.8%, duplicated = 0.7%), using the lepidoptera_odb10 reference set (n = 5,286).

Metadata for specimens, spectral estimates, sequencing runs, contaminants and pre-curation assembly statistics can be found at https://links.tol.sanger.ac.uk/species/987424.

Genome annotation report

The *Eilema sororcula* genome assembly (GCA_914829495.1) was annotated using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Eilema_sororculum_GCA_914829495.1/Info/Index). The resulting annotation includes 21,274 transcribed mRNAs from 21,093 protein-coding genes.

Methods

Sample acquisition and nucleic acid extraction

A male *Eilema sororcula* (specimen ID Ox000399, individual ilEilSoro1) was collected in Wytham Woods, Oxfordshire (biological vice-county Berkshire), UK (latitude 51.77, longitude –1.34) on 2020-05-22, using a light trap. The specimen was collected and identified by Douglas Boyes (University of Oxford) and preserved on dry ice.

DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute (WSI). The ilEilSoro1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. Tissue from the whole organism 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 sheared into an

Table 1. Genome data for *Eilema sororcula*, ilEilSoro1.1.

Project accession data		
Assembly identifier	ilEilSoro1.1	
Species	*Eilema sororcula*	
Specimen	ilEilSoro1	
NCBI taxonomy ID	987424	
BioProject	PRJEB46305	
BioSample ID	SAMEA7631555	
Isolate information	ilEilSoro1, male: whole organism (DNA sequencing and Hi-C scaffolding)	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	61.3	*≥ 50*
k-mer completeness	100%	*≥ 95%*
BUSCO**	C:98.5%[S:97.8%,D:0.7%],F:0.3%,M:1.2%,n:5,286	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	99.94%	*≥ 95%*
Sex chromosomes	Z chromosome	*localised homologous pairs*
Organelles	Mitochondrial genome assembled	*complete single alleles*
Raw data accessions		
PacificBiosciences SEQUEL II	ERR6807993, ERR6939231	
10X Genomics Illumina	ERR6688449–ERR6688452	
Hi-C Illumina	ERR6688453	
Genome assembly		
Assembly accession	GCA_914829495.1	
Accession of alternate haplotype	GCA_914829255.1	
Span (Mb)	729.4	
Number of contigs	157	
Contig N50 length (Mb)	11.8	
Number of scaffolds	40	
Scaffold N50 length (Mb)	25.3	
Longest scaffold (Mb)	51.8	
Genome annotation		
Number of protein-coding genes	21,093	
Number of gene transcripts	21,274	

* 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/ilEilSoro1.1/dataset/CAJZCU01.1/busco.

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

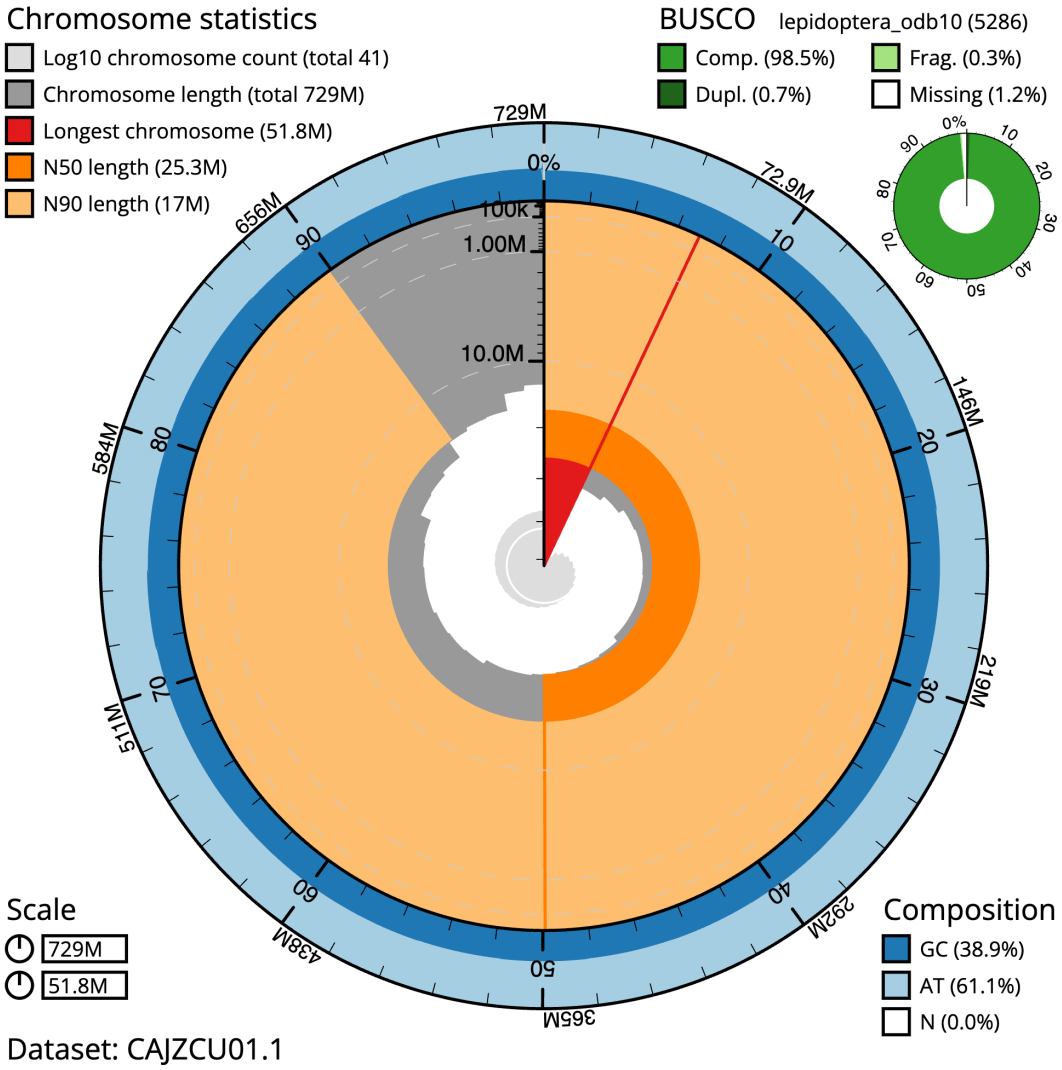

Figure 2. Genome assembly of *Eilema sororcula*, ilEilSoro1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 729,417,332 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (51,764,148 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (25,298,203 and 16,955,058 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/ilEilSoro1.1/dataset/CAJZCU01.1/snail.

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. DNA sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi) and Illumina NovaSeq 6000 (10X) instruments. Hi-C data were also generated from tissue of ilEilSoro1 using the Arima2 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). One round of polishing was performed by aligning 10X Genomics read data to the

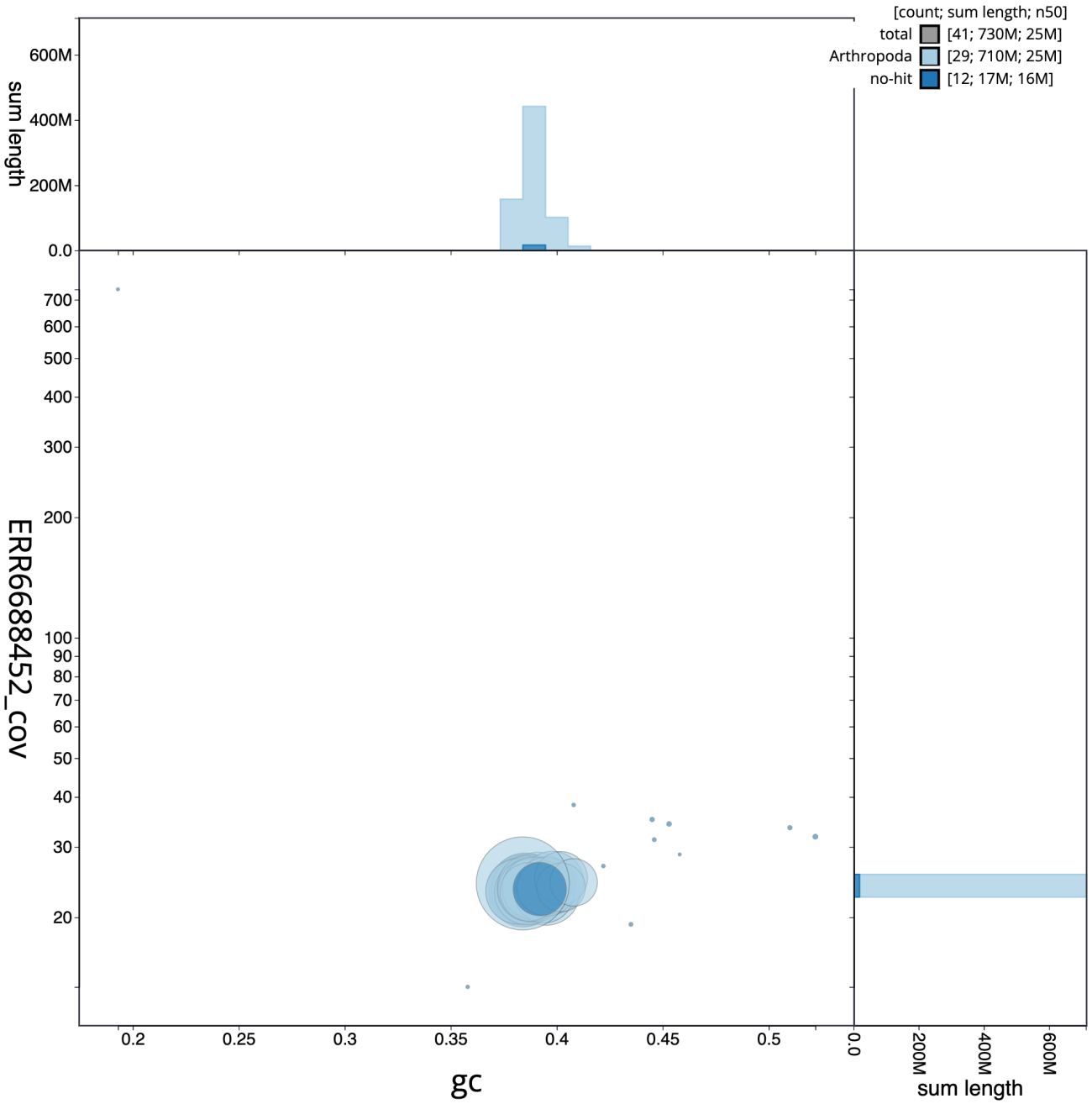

Figure 3. Genome assembly of *Eilema sororcula*, ilEilSoro1.1: BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilEilSoro1.1/dataset/CAJZCU01.1/blob.

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 SALSA2 (Ghurye *et al.*, 2019). The assembly was checked for contamination and corrected 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 runs MitoFinder (Allio *et al.*, 2020) or MITOS (Bernt *et al.*, 2013) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). To assess the assembly metrics,

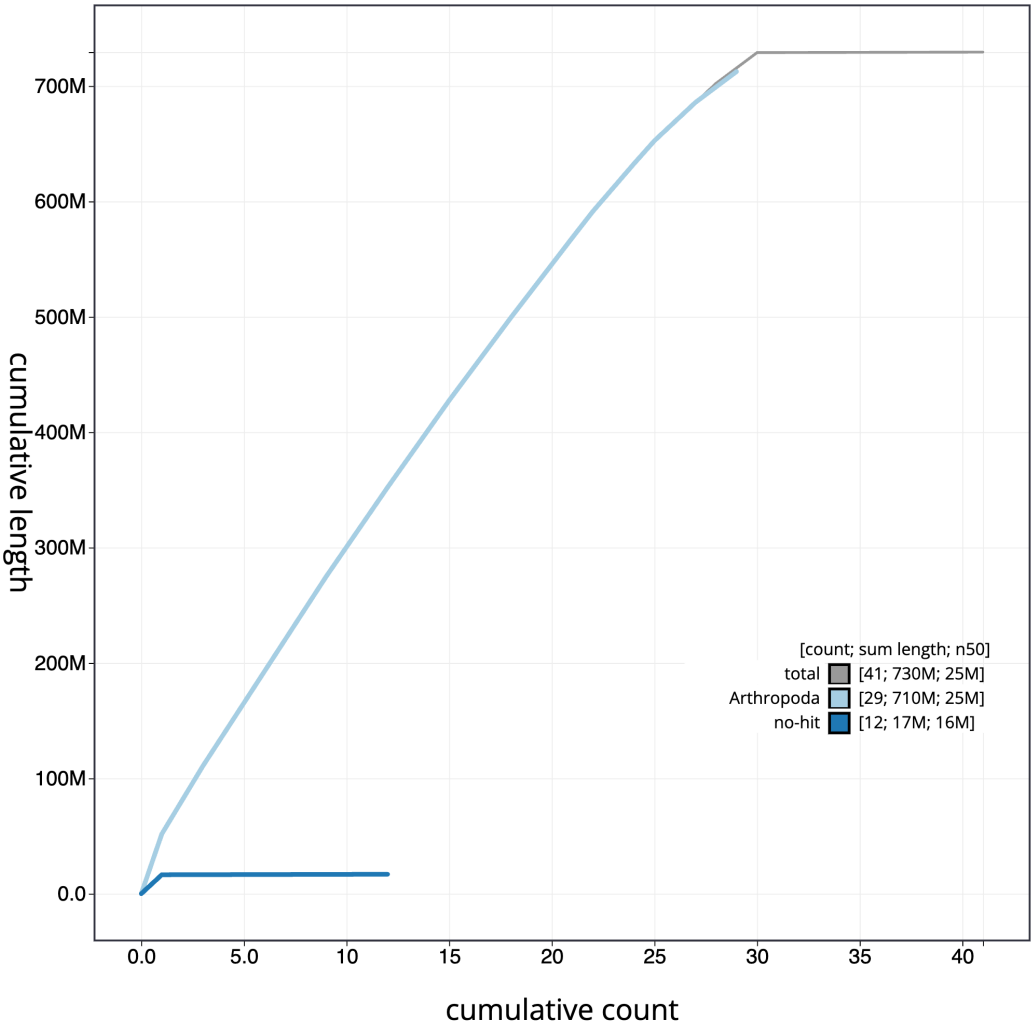

Figure 4. Genome assembly of *Eilema sororcula*, ilEilSoro1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilEilSoro1.1/dataset/CAJZCU01.1/cumulative.

the *k*-mer completeness and QV consensus quality values were calculated in Merqury (Rhie *et al.*, 2020). This work was done using Nextflow (Di Tommaso *et al.*, 2017) DSL2 pipelines "sanger-tol/readmapping" (Surana *et al.*, 2023a) and "sanger-tol/genomenote" (Surana *et al.*, 2023b). The genome was analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021; Simão *et al.*, 2015) were calculated.

Table 3 contains a list of relevant 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 *Eilema sororcula* assembly (GCA_914829495.1) in Ensembl Rapid Release.

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 here. 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.

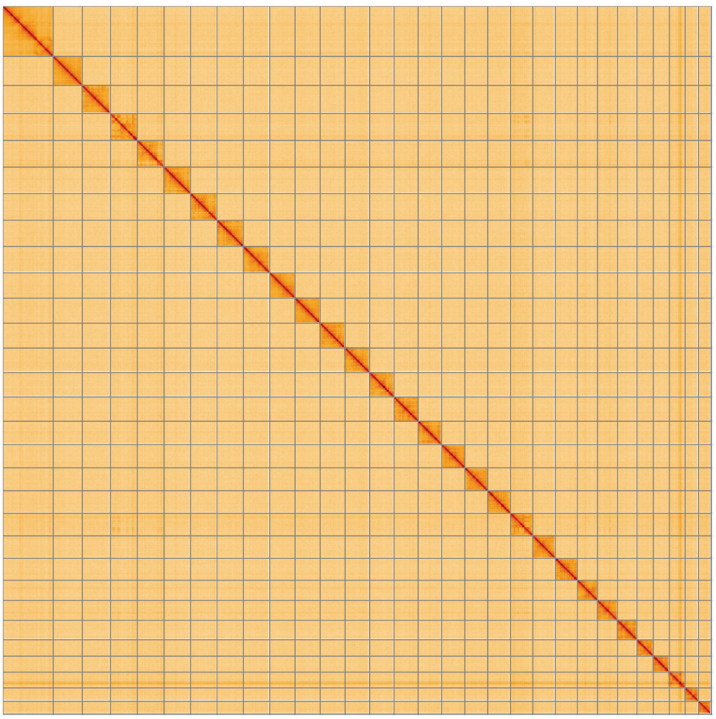

Figure 5. Genome assembly of *Eilema sororcula*, ilEilSoro1.1: Hi-C contact map of the ilEilSoro1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=IiePitoAQqCad3afW9wBMg.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Eilema sororcula*, ilEilSoro1.

INSDC accession	Name	Size (Mb)	GC%
OU618533.1	1	30.01	38.5
OU618534.1	2	28.88	38.6
OU618535.1	3	27.72	39.1
OU618536.1	4	27.46	39.5
OU618537.1	5	27.37	38.5
OU618538.1	6	27.23	38.4
OU618539.1	7	27.15	38.4
OU618540.1	8	27.12	38.5
OU618541.1	9	26.1	38.7
OU618542.1	10	25.84	38.5
OU618543.1	11	25.59	38.2
OU618544.1	12	25.3	38.7
OU618545.1	13	25.05	38.2
OU618546.1	14	24.98	38.6
OU618547.1	15	24.1	38.7
OU618548.1	16	23.88	38.8
OU618549.1	17	23.65	38.9
OU618550.1	18	23.21	38.7
OU618551.1	19	23.2	39.8
OU618552.1	20	23.01	38.7
OU618553.1	21	22.79	38.9
OU618554.1	22	20.73	39.4
OU618555.1	23	20.53	39.5
OU618556.1	24	19.97	38.7
OU618557.1	25	16.96	39.4
OU618558.1	26	16.45	39.2
OU618559.1	27	16.3	40.2
OU618560.1	28	13.71	40.2
OU618561.1	29	12.97	40.8
OU618532.1	Z	51.76	38.4
OU618562.1	MT	0.02	19.5
-	-	0.4	45.8

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	0.15.3	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
SALSA	2.2	https://github.com/salsa-rs/salsa
sanger-tol/genomenote	v1.0	https://github.com/sanger-tol/genomenote
sanger-tol/readmapping	1.1.0	https://github.com/sanger-tol/readmapping/tree/1.1.0

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: *Eilema sororcula* (orange footman). Accession number PRJEB46305; https://identifiers.org/ena.embl/PRJEB46305. (Wellcome Sanger Institute, 2021)

The genome sequence is released openly for reuse. The *Eilema sororcula* 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

Abdennur N, Mirny LA: **Cooler: Scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Bernt M, Donath A, Jühling F, *et al.*: **MITOS: Improved *de novo* metazoan mitochondrial genome annotation.** *Mol Phylogenet Evol.* 2013; **69**(2): 313–9.
PubMed Abstract | Publisher 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

Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.
PubMed Abstract | Publisher Full Text

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

GBIF Secretariat: ***Eilema sororcula*.** GBIF Backbone Taxonomy. 2023; [Accessed 3 June 2023].
Reference Source

Ghurye J, Rhie A, Walenz BP, *et al.*: **Integrating Hi-C links with assembly graphs for chromosome-scale assembly.** *PLoS Comput Biol.* 2019; **15**(8): e1007273.
PubMed 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): 2896–2898.
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

Henwood B, Sterling P, Lewington R: **Field Guide to the Caterpillars of Great Britain and Ireland.** London: Bloomsbury, 2020; 448.
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

Moths Ireland: **Eilema sorocula, MothsIreland: Mapping Ireland's Moths.** 2023; [Accessed 3 June 2023].
Reference Source

Randle Z, Evans-Hill LJ, Parsons MS, *et al.*: **Atlas of Britain & Ireland's Larger Moths.** Newbury: NatureBureau, 2019.
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): 1665–80.
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

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

Surana P, Muffato M, Qi G: **sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo*. 2023a.
Publisher Full Text

Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote v1.0.dev (v1.0.dev).** *Zenodo*. 2023b; [Accessed 17 April 2023].
Publisher Full Text

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

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.
Publisher Full Text

Wellcome Sanger Institute: **The genome sequence of the Orange Footman, *Eilema sororcula* (Hufnagel, 1766).** European Nucleotide Archive. [dataset], accession number PRJEB46305, 2021.

Open Peer Review

Current Peer Review Status: ✔

Version 1

Reviewer Report 30 June 2023

https://doi.org/10.21956/wellcomeopenres.21742.r61541

✔ **Michael Hiller**
LOEWE-Centre for Translational Biodiversity Genomics (TBG), Senckenberg Nature Research Society, Frankfurt Am Main, Germany

This data note reports a reference-quality genome assembly of *Eilema sororcula*, together with a gene annotation produced by Ensembl. The generated data and methods used are according to the latest standards. The resulting assembly has a very high quality.

I have 3 minor comments:

1. typo: collected from| Wytham Woods

2. "Most (99.94%) of the assembly sequence was assigned to 30 chromosomal-level scaffolds" - Although I understand that these sentences are generated by filling placeholders with the real values, I think the word 'most' does not capture that essentially the entire assembly (99.94%) is in chrom-level scaffolds. I would suggest to write "The vast majority of the assembly" or something similar, here and in the abstract.

3. "While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. " - I don't understand this point. The authors likely produced a primary assembly (and provide alternative contigs in the 'alt assembly'), although this is not precisely mentioned in the methods (HiFiasm has different assembly modes). If so, the primary is likely a mix of both haplotypes, and not 'one haplotype' as stated. This should please be clarified.

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: Genome assembly / comparative 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.