DATA NOTE

The genome sequence of the speckled wood butterfly, *Pararge aegeria* (Linnaeus, 1758) [version 1; peer review: awaiting peer review]

Konrad Lohse\(^1\), Evelyn Taylor-Cox\(^2\), 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

\(^1\)Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
\(^2\)University of Liverpool, Liverpool, UK

**Abstract**

We present a genome assembly from an individual female *Pararge aegeria* (the speckled wood butterfly; Arthropoda; Insecta; Lepidoptera; Nymphalidae). The genome sequence is 517 megabases in span. The majority of the assembly (99.68%) is scaffolded into 29 chromosomal pseudomolecules, with the W and Z sex chromosome assembled. Gene annotation of this assembly on Ensembl has identified 12,288 protein coding genes.

**Keywords**

*Pararge aegeria*, speckled wood butterfly, genome sequence, chromosomal

This article is included in the Tree of Life gateway.
Species taxonomy
Eukarya; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Papilionoidea; Nymphalidae; Satyrinae; Satyrini; Parargina; Pararge; Pararge aegeria (Linnaeus, 1758) (NCBI:txid116150).

Introduction
The speckled wood butterfly, Pararge aegeria (Linnaeus, 1758), is a common woodland species that occurs across the Palearctic, from north Africa to Europe and western Asia, reflecting wide environmental tolerances. Pararge aegeria is a multivoltine species, producing a variable number of generations each year and can over-winter in two life stages; as a larva or pupa, using a variety of grasses as the larval host plant (Shreeve, 1986). Two main lineages are recognised: subspecies aegeria in north Africa (Linnaeus, 1758) and tircis in Eurasia (Godart, 1821; Livraghi et al., 2018). Two additional subspecies, oblixa (western Scotland) (Harrison, 1949) and insula (Isles of Scilly) (Howarth, 1971), have been described based on wing pattern variation. The distribution of P. aegeria in the UK has been dynamic in recent history: the range contracted from a widespread distribution, due to cooler climates in the 1950s, to south-west England and Wales, as well as a refugial population in western Scotland (Hill et al., 1999). Since the 1970s, a warming climate has enabled P. aegeria to expand its range and to recolonise northern England and the whole of Scotland. Morphological differences in wing size, shape and colour have been described across the UK distribution (Taylor-Cox et al., 2020) and life-history evolution and plasticity have been studied extensively in this species (Nylin et al., 1995). The standard haploid karyotype of P. aegeria consists of 27 autosomes and one sex chromosome (Federley, 2010), and the female is heterogametic (WZ).

Genome sequence report
The genome was sequenced from a single female P. aegeria (ilParAegt1) collected from West Saltoun Forest, Scotland (latitude 55.884404, longitude -2.857514). Hi-C data were generated from a male P. aegeria (ilParAegt2) collected from the same location (Figure 1). A total of 42-fold coverage in Pacific Biosciences single-molecule long reads and 66-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 12 missing/misjoins, reducing the scaffold number by 2.78%.

The final assembly has a total length of 517 Mb in 70 sequence scaffolds with a scaffold N50 of 19 Mb (Table 1). Of the assembly sequence, 99.68% was assigned to 29 chromosomal-level scaffolds, representing 27 autosomes (numbered by sequence length), and the W and Z sex chromosome (Figure 2–Figure 5; Table 2). The assembly has a BUSCO (Simão et al., 2015) v5.1.2 completeness of 98.6% using the lepidoptera_odb10 reference set. While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

Gene annotation
The Ensembl gene annotation system (Aken et al., 2016) was used to generate annotation for the Pararge aegeria assembly (GCA_905163445.1, see https://rapid.ensembl.org/Pararge_aegeria_GCA_905163445.1/; Table 1). The annotation was created primarily through alignment of transcriptomic data to the genome, with gap filling via protein to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 2019) and OrthoDB (Kriventseva et al., 2008). Prediction tools, CPC2 (Kang et al., 2017) and RNAsamba (Camargo et al., 2020), were used to aid determination of protein coding genes.

Methods
Sample acquisition and nucleic acid extraction
One female (ilParAegt1, genome assembly) and one male (ilParAegt2, Hi-C) P. aegeria specimen were collected from West Saltoun Forest, Scotland (latitude 55.884404, longitude -2.857514) by Konrad Lohse, University of Edinburgh. A further female specimen, ilParAegt3 (RNA-Seq), was collected from Yellow Craig, Scotland (latitude 56.062447, longitude 2.769836) by the same individual, who also identified both specimens. All samples were snap-frozen in liquid nitrogen.

DNA was extracted from the whole organism of ilParAegt1 at the Wellcome Sanger Institute (WSI) Scientific Operations core from the whole organism using the Qiagen MagAttract HMW DNA kit, according to the manufacturer’s instructions. RNA (also from the whole organism) was extracted from ilParAegt3 in the Tree of Life Laboratory at the WSI using TRlzol, according to the manufacturer’s instructions. RNA was then eluted in 50 µl RNase-free water and its concentration RNA assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit RNA Broad-Range (BR) Assay kit. Analysis of the integrity of the RNA was done using Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

Sequencing
Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud DNA sequencing libraries were constructed according to the manufacturers’ instructions. Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit. DNA and RNA sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences Sequel II (HiFi), Illumina HiSeq X (10X) and Illumina HiSeq 4000 (RNA-Seq) instruments. Hi-C data were generated from the whole organism of ilParAegt2 using the Qiagen Epitector Hi-C kit and sequenced on HiSeq X.

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 was performed using gEVAL, HiGlass (Kerpedjiev et al., 2018) and Pretext.

Figure 1. Fore and hind wings of *Pararge aegeria* specimens from which the genome was sequenced. (A) Dorsal surface view of wings from specimen SC_PA_1193 (iiParAegt1) from Saltoun Forest, Scotland, UK, used to generate Pacific Biosciences and 10X genomics data. (B) Ventral surface view of wings from specimen SC_PA_1193 (iiParAegt1) from Saltoun Forest, Scotland, UK, used to generate Pacific Biosciences and 10X genomics data. (C) Dorsal surface view of wings from specimen SC_PA_1194 (iiParAegt2) from Saltoun Forest, Scotland, UK, used to generate Hi-C data. (D) Ventral surface view of wings from specimen SC_PA_1194 (iiParAegt2) from Saltoun Forest, Scotland, UK, used to generate HiC data.
The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva et al., 2021). 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.

**Ethical/compliance issues**

The materials that have contributed to this genome note were supplied by a Tree of Life collaborator. The Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves,
Figure 2. Genome assembly of Pararge aegeria, iParAegt1.1: 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 516,586,240 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 (27,171,456 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (19,375,358 and 13,052,829 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/iParAegt1.1/dataset/CAJHZN01.1/snail.
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.

Figure 3. Genome assembly of Pararge aegeria, ilParAegt1.1: 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/ilParAegt1.1/dataset/CAJHZN01.1/blob,
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 undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Tree of Life collaborator, Genome Research Limited (operating as the Wellcome Sanger Institute) and in some circumstances other Tree of Life collaborators.
Table 2. Chromosomal pseudomolecules in the genome assembly of *Pararge aegeria*, iIParAegt1.1.

| INSDC accession | Chromosome | Size (Mb) | GC% |
|-----------------|------------|-----------|-----|
| LR990895.1      | 1          | 21.30     | 35.6|
| LR990892.1      | 2          | 22.24     | 35.6|
| LR990893.1      | 3          | 21.66     | 35.8|
| LR990894.1      | 4          | 21.56     | 35.6|
| LR990896.1      | 5          | 20.98     | 35.7|
| LR990897.1      | 6          | 20.78     | 35.9|
| LR990899.1      | 7          | 20.31     | 35.7|
| LR990898.1      | 8          | 20.39     | 35.6|
| LR990900.1      | 9          | 19.84     | 35.8|
| LR990901.1      | 10         | 19.54     | 35.7|
| LR990902.1      | 11         | 19.42     | 35.7|
| LR990903.1      | 12         | 19.38     | 35.7|
| LR990904.1      | 13         | 18.56     | 35.7|
| LR990905.1      | 14         | 18.51     | 35.7|
| LR990906.1      | 15         | 18.47     | 35.7|
| LR990891.1      | Z          | 27.17     | 35.6|
| LR990899.1      | MT         | 0.02      | 19.5|
| LR990909.1      | W          | 2.08      | 36.1|
| LR990891.1      | Unplaced   | 6.73      | 37.8|

**Figure 5:** Genome assembly of *Pararge aegeria*, iIParAegt1.1: Hi-C contact map. Hi-C contact map of the iIParAegt1.1 assembly, visualised in HiGlass. Chromosomes are shown in size order from left to right and top to bottom.
### Table 3. Software tools used.

| Software tool      | Version | Source                                           |
|-------------------|---------|--------------------------------------------------|
| Hifiasm           | 0.12    | Cheng et al., 2021                               |
| 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                        |
| gEVAL             | N/A     | Chow et al., 2016                                |
| HiGlass           | 1.11.6  | Kerpedjiév et al., 2018                          |
| PretextView       | 0.1.x   | https://github.com/wtsi-hpag/PretextView         |
| BlobToolKit       | 2.6.2   | Challis et al., 2020                             |

### Data availability

European Nucleotide Archive: Pararge aegeria (speckled wood). Accession number PRJEB42139; https://identifiers.org/ena.embl/PRJEB42139.

The genome sequence is released openly for reuse. The *P. aegeria* 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.

### Acknowledgements

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.4893704.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.5013542.

Members of the Wellcome Sanger Institute Tree of Life programme collective are listed here: https://doi.org/10.5281/zenodo.5377053.

Members of Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.4790456.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783559.

### References

Aken BL, Ayling S, Barrell D, et al.: The Ensembl Gene Annotation System. Database (Oxford). 2016. 2016: baw093. PubMed Abstract | Publisher Full Text | Free Full Text

Camargo AP, Sourkov V, Pereira GAG, et al.: RNAsamba: Neural Network-Based Assessment of the Protein-Coding Potential of RNA Sequences. NAR Genom Bioinform. 2020; 2(1): lqz024. 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, Consecioon 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

Federley H: Chromosomenzahlen Finnlän-Discher Lepidopteren. Hereditas. 2010; 24(4): 397-464. Publisher Full Text

Garrison E, Marth G: Chromosomenzahlen Finnlän-Discher Lepidopter. arXiv: 1207.3907. 2012.

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

Godart JB: Histoire naturelle des lépidoptères ou papillons de France. Méquignon-Marvis fils. 1821. Publisher Full Text

Guan D, McCarthy SA, Wood j, et al.: Identifying and Removing Haplotypic
