DATA NOTE

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

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

¹Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK

²University of Liverpool, Liverpool, UK

V1 First published: 25 Oct 2021, 6:287
<https://doi.org/10.12688/wellcomeopenres.17278.1>
Latest published: 25 Oct 2021, 6:287
<https://doi.org/10.12688/wellcomeopenres.17278.1>

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.

Open Peer Review

Approval Status

	1	2
version 1 25 Oct 2021	 view	 view

1. **Xuankun Li** , University of Florida, Gainesville, USA
2. **Xueyan Li** , Chinese Academy of Sciences, Yunnan, China

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: Lohse K: Investigation, Resources, Writing – Review & Editing; Taylor-Cox E: 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). KL is supported by a NERC fellowship (NE/L011522/1) and an ERC grant (ModelGenom Land 757648).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Lohse K *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 K, Taylor-Cox E, Darwin Tree of Life Barcoding collective *et al.* **The genome sequence of the speckled wood butterfly, *Pararge aegeria* (Linnaeus, 1758) [version 1; peer review: 1 approved, 1 approved with reservations]** Wellcome Open Research 2021, 6:287 <https://doi.org/10.12688/wellcomeopenres.17278.1>

First published: 25 Oct 2021, 6:287 <https://doi.org/10.12688/wellcomeopenres.17278.1>

Species taxonomy

Eukaryota; 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, *oblita* (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 Yellowcraig, 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 TRIzol, 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 EpiTect 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

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 (ilParAegt1) 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 (ilParAegt1) 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 (ilParAegt2) from Saltoun Forest, Scotland, UK, used to generate Hi-C data. (D) Ventral surface view of wings from specimen SC_PA_1194 (ilParAegt2) from Saltoun Forest, Scotland, UK, used to generate Hi-C data.

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.

Table 1. Genome data for *Pararge aegeria*, ilParAegt1.1.

Project accession data	
Assembly identifier	ilParAegt1
Species	<i>Pararge aegeria</i>
Specimen	ilParAegt1 (genome assembly); ilParAegt2 (Hi-C), ilParAegt3 (RNA-Seq)
NCBI taxonomy ID	NCBI:txid116150
BioProject	PRJEB42139
BioSample ID	SAMEA7532732
Isolate information	Female, whole organism (ilParAegt1, ilParAegt3); male, whole organism (ilParAegt2)
Raw data accessions	
PacificBiosciences SEQUEL II	ERR6594492, ERR6594493
10X Genomics Illumina	ERR6002715-ERR6002718
Hi-C Illumina	ERR6002719
Illumina polyA RNA-Seq	ERR6363260
Genome assembly	
Assembly accession	GCA_905163445.1
Accession of alternate haplotype	GCA_905163545.1
Span (Mb)	517
Number of contigs	81
Contig N50 length (Mb)	19
Number of scaffolds	70
Scaffold N50 length (Mb)	19
Longest scaffold (Mb)	21
BUSCO* genome score	C:98.8%[S:98.4%,D:0.4%],F:0.2%,M:1.0%,n:1658
Gene annotation	
Number of protein coding genes	12,888
Average coding sequence length (bp)	1,626
Average number of exons per transcript	9
Average exon size (bp)	362
Average intron size (bp)	2,461

*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/ilParAegt1.1/dataset/CAJHZN01.1/busco>.

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,

Scaffold statistics

- Log10 scaffold count (total 71)
- Scaffold length (total 520M)
- Longest scaffold (27M)
- N50 length (19M)
- N90 length (13M)

BUSCO

lepidoptera_odb10 (5286)

- Complete (98.6%)
- Fragmented (0.4%)
- Duplicated (0.4%)
- Missing (1.1%)

Figure 2. Genome assembly of *Pararge aegeria*, ilParAegt1.1: 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/ilParAegt1.1/dataset/CAJHZN01.1/snail>.

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>.

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 4. Genome assembly of *Pararge aegeria*, ilParAegt1.1: 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/ilParAegt1.1/dataset/CAJHZN01.1/cumulative>.

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.

Figure 5: Genome assembly of *Pararge aegeria*, ilParAegt1.1: Hi-C contact map. Hi-C contact map of the ilParAegt1.1 assembly, visualised in HiGlass. Chromosomes are shown in size order from left to right and top to bottom.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Pararge aegeria*, ilParAegt1.1.

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	36
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	36
LR990905.1	14	18.51	35.7
LR990906.1	15	18.47	35.7

INSDC accession	Chromosome	Size (Mb)	GC%
LR990907.1	16	18.41	36.1
LR990908.1	17	17.72	36.2
LR990909.1	18	17.53	35.9
LR990910.1	19	17.52	36.5
LR990911.1	20	17.49	35.9
LR990912.1	21	17.15	36.2
LR990913.1	22	15.15	36.2
LR990914.1	23	13.18	36.3
LR990915.1	24	13.05	37
LR990916.1	25	11.60	37.1
LR990917.1	26	9.87	37.4
LR990918.1	27	8.98	37.4
LR990919.1	W	2.08	36.1
LR990891.1	Z	27.17	35.6
LR990920.1	MT	0.02	19.5
-	Unplaced	6.73	37.8

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	Kerpedjiev 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 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 Tree of Life Core Informatics collective are listed here: <https://doi.org/10.5281/zenodo.5013542>.

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.: **RNAse: 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, 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](#)
- Federley H: **Chromosomenzahlen Finnland-Discher Lepidopteren. Hereditas.** 2010; **24**(4): 397–464.
[Publisher Full Text](#)
- Garrison E, Marth G: **Hahromosomenzahlen Finnland-Discher Lepidopt. 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**

Duplication in Primary Genome Assemblies. *Bioinformatics*. 2020; **36**(9): 2896–98.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Harrison JWH: **Rhopalocera in the Scottish Western Isles in 1948 with an Account of Two New Forms of Pararge Aegeria L.**(Lepidoptera, Satyridae). *Entomologist's Monthly Magazine*. 1949; **85**: 25–28.

Hill JK, Thomas CD, Huntley B: **Climate and Habitat Availability Determine 20th Century Changes in a Butterfly's Range Margin.** *Proceedings of the Royal Society of London. Series B: Biological Sciences*. 1999; **266**(1425): 1197–1206. [Publisher Full Text](#)

Howarth TG: **Descriptions of a New British Subspecies of Pararge Aegeria (L.)**(Lepidoptera, Satyridae) and an Aberration of Cupido Minimus (Fuessly)(Lepidoptera, Lycaenidae). *Entomologist's Gazette*. 1971. [Reference Source](#)

Howe K, Chow W, Collins J, *et al.*: **Significantly Improving the Quality of Genome Assemblies through Curation.** *GigaScience*. 2021; **10**(1): g1aa153. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Kang YJ, Yang DC, Kong L, *et al.*: **CPC2: A Fast and Accurate Coding Potential Calculator Based on Sequence Intrinsic Features.** *Nucleic Acids Res*. 2017; **45**(W1): W12–16. [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](#)

Kriventseva EV, Rahman N, Espinosa O, *et al.*: **OrthoDB: The Hierarchical Catalog of Eukaryotic Orthologs.** *Nucleic Acids Res*. 2008; **36**(Database issue): D271–75. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Livraghi L, Vodá R, Evans LC, *et al.*: **Historical and Current Patterns of Gene**

Flow in the butterfly Pararge Aegeria. *J Biogeogr*. 2018; **45**(7): 1628–39. [Publisher Full Text](#)

Nylin S, Wickman PO, Wiklund C: **Life-Cycle Regulation and Life History Plasticity in the Speckled Wood Butterfly: Are Reaction Norms Predictable?** *Biological Journal of the Linnean Society. Linnean Society of London*. 1995; **55**(2): 143–57. [Publisher 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](#)

Shreeve TG: **Egg-Laying by the Speckled Wood Butterfly (*Pararge Aegeria*): The Role of Female Behaviour, Host Plant Abundance and Temperature.** *Ecological Entomology*. 1986. [Publisher 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): 3210–12. [PubMed Abstract](#) | [Publisher Full Text](#)

Taylor-Cox ED, Macgregor CJ, Corthine A, *et al.*: **Wing Morphological Responses to Latitude and Colonisation in a Range Expanding Butterfly.** *PeerJ*. 2020; **8**: e10352. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Uliano-Silva M, Gabriel Ferreira Nunes J, Krashenninnikova K, *et al.*: **marcelauliano/MitoHiFi: mitohifi_v2.0.** 2021. [Publisher Full Text](#)

UniProt Consortium: **UniProt: A Worldwide Hub of Protein Knowledge.** *Nucleic Acids Res*. 2019; **47**(D1): D506–15. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 20 December 2021

<https://doi.org/10.21956/wellcomeopenres.19102.r47353>

© 2021 Li X. 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.

Xueyan Li

State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Yunnan, China

This paper "The genome sequence of the speckled wood butterfly, *Pararge aegeria* (Linnaeus, 1758)" assembled the chromosomal-level genome of *Pararge aegeria* which is an excellent research model on population genetics, evolutionary biology, and ecology. Due to the HiFi and 10X Genomics RNA-seq, the assembly quality is relatively high, and the annotation analyses are reliable. In general, this paper is well written with concise description and methods, and analyses are transparent, and the high-quality genome assembly will provide a good genetic resource for the subsequent analyses.

Several revisions are needed about the data:

1. First, I suggest there should be some genome size assessment, e.g., the *Kmer* analysis and FCM (flow cytometry). If there has been some public data known, it can be quoted here.
2. Second, please provide the mapping depth and synteny analysis to verify the identification of the sex chromosome.
3. Third, the average intron size (AIL) is relatively high, I suggest the authors can estimate the QV and completeness using Merqury software.

Minor revisions:

1. Introduction: "*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)", please remove the semi-colon after "life stages".
2. Introduction: "Since the 1970s, a warming climate has enabled *P. aegeria* to expand its range and to recolonise northern England...", please change "recolonise" to "recolonize".

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.

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, however I have significant reservations, as outlined above.

Reviewer Report 07 December 2021

<https://doi.org/10.21956/wellcomeopenres.19102.r47352>

© 2021 Li X. 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.

Xuankun Li

McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville, FL, USA

Authors sequenced, assembled, and annotated the chromosome-level genome of the speckled wood butterfly, *Pararge aegeria*. The paper is well-presented, with methods clearly explained, and results properly described. It is one of the series of products reflecting matured sequencing protocol and assembling pipeline.

Minor issues:

- **Species taxonomy:** 'Pararge' should be in italics.
- **Introduction:** 'two life stages;' change to 'two life stages:'.
- **Methods:** 'who also identified both specimens' change to 'who also identified all specimens'.

This paper provides an important resource for genome studies. It is a valuable contribution to entomology and molecular evolution. Therefore, I recommend accepting this paper.

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: Entomology, genome biology

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.