







DATA NOTE

The genome sequence of the common yellow swallowtail, *Papilio machaon* (Linnaeus, 1758) [version 1; peer review: 1 approved, 1 approved with reservations]

Konrad Lohse ¹, Alex Hayward², Dominik R. Laetsch ¹, Roger Vila ³,
Wellcome Sanger Institute Tree of Life programme,
Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective,
Tree of Life Core Informatics collective, Tarunkishwor Yumnam ⁴,
Darwin Tree of Life Consortium

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

²College of Life and Environmental Sciences, Department of Biosciences, University of Exeter, Exeter, UK

³Institut de Biologia Evolutiva, CSIC - Universitat Pompeu Fabra, Barcelona, Spain

⁴Indian Institute of Science Education and Research, Thiruvananthapuram, India

V1 First published: 14 Oct 2022, 7:261
<https://doi.org/10.12688/wellcomeopenres.18119.1>
Latest published: 14 Oct 2022, 7:261
<https://doi.org/10.12688/wellcomeopenres.18119.1>

Abstract

We present a genome assembly from an individual female *Papilio machaon* (the common yellow swallowtail; Arthropoda; Insecta; Lepidoptera; Papilionidae). The genome sequence is 252 megabases in span. The majority of the assembly (99.97%) is scaffolded into 31 chromosomal pseudomolecules with the W and Z sex chromosomes assembled. The complete mitochondrial genome was also assembled and is 15.3 kilobases in length. Gene annotation of this assembly on Ensembl has identified 14,323 protein coding genes.

Keywords



Papilio machaon, common yellow swallowtail, genome sequence, chromosomal, Papilionidae





This article is included in the [Tree of Life gateway](#).

Open Peer Review

Approval Status  

	1	2
version 1		
14 Oct 2022	view	view

1. **Eliette Reboud** , Université de Montpellier, Montpellier, France
2. **Christopher W. Wheat** , Stockholm University, Stockholm, Sweden

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 – Original Draft Preparation, Writing – Review & Editing; **Hayward A:** Investigation, Resources, Writing – Review & Editing; **Laetsch DR:** Investigation, Resources, Writing – Review & Editing; **Vila R:** Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Yumnam T:** 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 and DRL are supported by an ERC grant (ModelGenomLand 757648). KL was also supported by a NERC fellowship (NE/L011522/1). AH is supported by a Biotechnology and Biological Sciences Research Council (BBSRC) David Phillips Fellowship (BB/N020146/1). RV was supported by the Spanish government through grant PID2019-107078GB-I00/MCIN/AEI/ 10.13039/501100011033.

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

Copyright: © 2022 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, Hayward A, Laetsch DR *et al.* **The genome sequence of the common yellow swallowtail, *Papilio machaon* (Linnaeus, 1758) [version 1; peer review: 1 approved, 1 approved with reservations]** Wellcome Open Research 2022, 7:261 <https://doi.org/10.12688/wellcomeopenres.18119.1>

First published: 14 Oct 2022, 7:261 <https://doi.org/10.12688/wellcomeopenres.18119.1>

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Papilionoidea; Papilionidae; Papilioninae; *Papilio*; *Papilio machaon* (Linnaeus, 1758) (NCBI:txid76193).

Background

Papilio machaon (Linnaeus, 1758), commonly known as the common yellow swallowtail or Old World swallowtail, is present in much of the Palearctic and parts of Asia and North America. It has 41 recognised subspecies, of which *P. m. britannicus* and *P. m. gorganus* are found in the UK. Larvae of *P. m. gorganus* are oligophagous on Apiaceae (mainly *Phoeniculum vulgare*, *Daucus carota*, *Angelica sylvestris* and *Heracleum* spp.) and Rutaceae (*Ruta* spp.) and adults are notably dispersive, reaching southern England occasionally. Meanwhile, *P. m. britannicus*, the largest butterfly of the UK, is an endemic subspecies with a confined range in the Norfolk Broads, resulting from sustained decline of its natural habitat. While *P. machaon* is listed as a species of Least Concern on the IUCN Red List of Europe (van Swaay *et al.*, 2010), *P. m. britannicus* is considered a species of conservation concern (Fox *et al.*, 2022) and is fully protected in the UK. It is a specialist on *Peucedanum palustre* (Apiaceae), and occurs only in open fens (Collins *et al.*, 2020). *P. m. britannicus* is often assumed to be most closely related to *P. m. gorganus*. However, this assumption warrants further investigation, and the availability of the common yellow swallowtail's genome sequence can enable researchers to resolve this question.

Papilio machaon exhibits pupal colour plasticity (PCP) where diapausing pupae have a brown-white and non-diapausing pupae often have green-yellow colouration. It has been observed that the release of hormonal factor(s) from the head-thorax region results in brown-white pupae (Yamanaka *et al.*, 2013). *Ebony* is involved in the melanin biosynthesis and is expressed in the pupal stage (Li *et al.*, 2015), and hence, can potentially play a role in PCP. The availability of a high quality annotated genome sequence of *P. machaon* will enable researchers to investigate the regulatory mechanisms of pupal colouration in detail.

Genome sequence report

The genome was sequenced from a single female *P. machaon* collected near Gheorgheni, Romania (Figure 1). A total of 99-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 148-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected six missing/misjoins and removed one haplotypic duplications, reducing the assembly size by 0.76% and the scaffold number by 2.94%, and the scaffold N50 by 1.02%.

The final assembly has a total length of 252 Mb in 33 sequence scaffolds with a scaffold N50 of 8.8 Mb (Table 1). The majority, 99.97%, of the assembly sequence was assigned to 31 chromosomal-level scaffolds, representing 29 autosomes



Figure 1. Fore and hind wings of the *Papilio machaon* specimen from which the genome was sequenced. Dorsal (left) and ventral (right) surface view of wings from specimen RO_PM_1011 (ilPapMach1) from Gheorgheni, Romania, used to generate Pacific Biosciences, 10X genomics and Hi-C data.

Table 1. Genome data for *Papilio machaon*, ilPapMach1.1.

Project accession data	
Assembly identifier	ilPapMach1.1
Species	<i>Papilio machaon</i>
Specimen	ilPapMach1 (genome assembly, Hi-C)
NCBI taxonomy ID	76193
BioProject	PRJEB46295
BioSample ID	SAMEA7523121
Isolate information	Female. Thorax tissue (genome assembly); remaining whole organism tissue (Hi-C)
Raw data accessions	
PacificBiosciences SEQUEL II	ERR6594501
10X Genomics Illumina	ERR6363327-ERR6363330
Hi-C Illumina	ERR6363331
Genome assembly	
Assembly accession	GCA_912999745.1
Accession of alternate haplotype	GCA_912999765.2
Span (Mb)	252
Number of contigs	37
Contig N50 length (Mb)	8.6
Number of scaffolds	33
Scaffold N50 length (Mb)	8.8
Longest scaffold (Mb)	10.7
BUSCO* genome score	C:99.3%[S:98.3%,D:1.0%], F:0.1%,M:0.6%,n:5,286

*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/ilPapMach1.1/dataset/CAJWVC01/busco>.

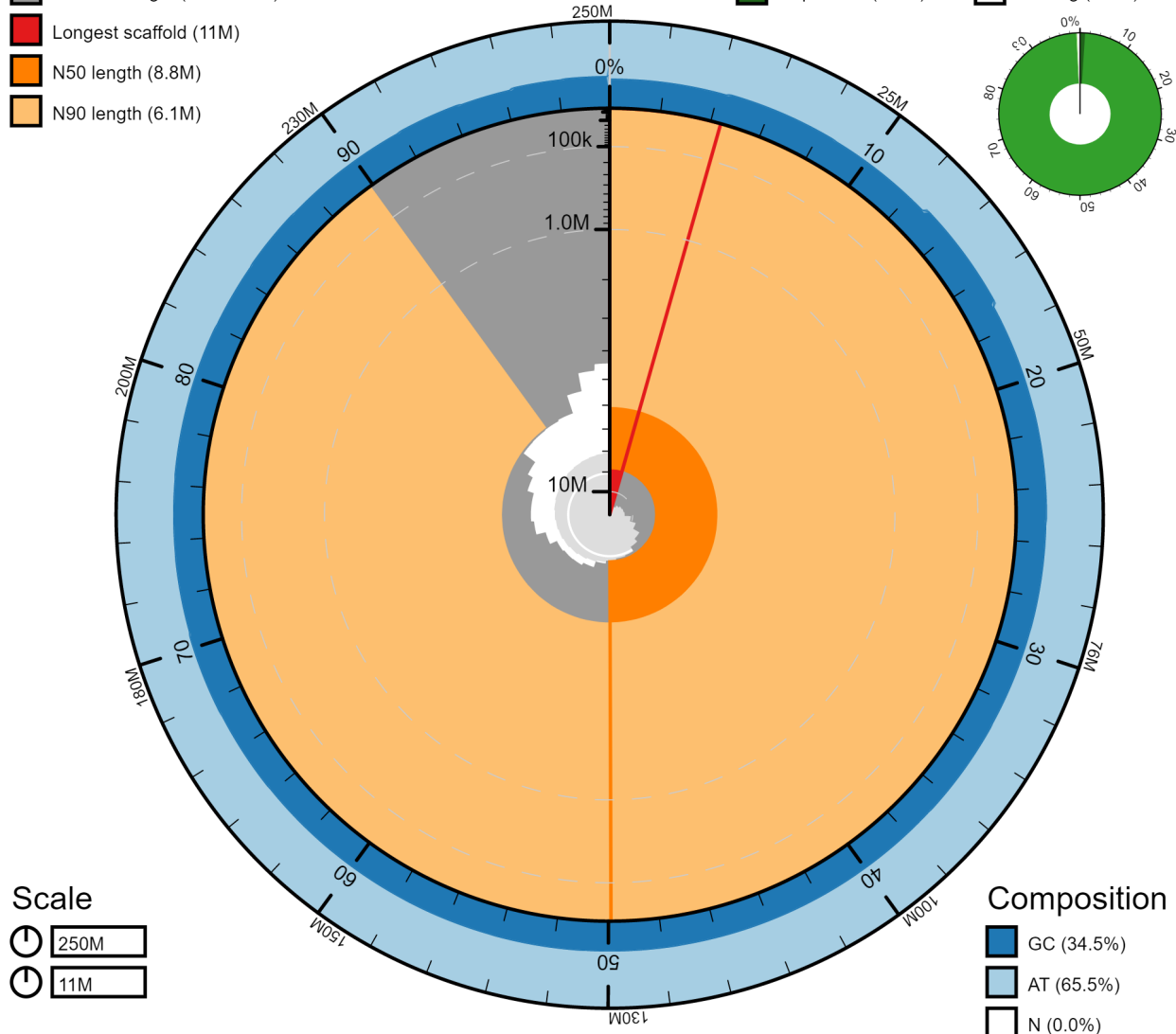
(numbered by sequence length) and the W and Z sex chromosomes (Figure 2–Figure 5; Table 2).

Scaffold statistics

- Log10 scaffold count (total 34)
- Scaffold length (total 250M)
- Longest scaffold (11M)
- N50 length (8.8M)
- N90 length (6.1M)

BUSCO lepidoptera_odb10 (5286)

- Complete (99.3%)
- Fragmented (0.1%)
- Duplicated (1.0%)
- Missing (0.6%)



Scale

- 250M
- 11M

Composition

- GC (34.5%)
- AT (65.5%)
- N (0.0%)

Dataset: CAJVWC01

Figure 2. Genome assembly of *Papilio machaon*, ilPapMach1.1: 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 252,129,366 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 (11,236,252 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (8,777,606 and 6,070,407 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/ilPapMach1.1/dataset/CAJVWC01/snail>.

The assembly has a BUSCO v5.3.2 (Manni *et al.*, 2021) completeness of 99.3% (single 98.3%, duplicated 1.0%) 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.

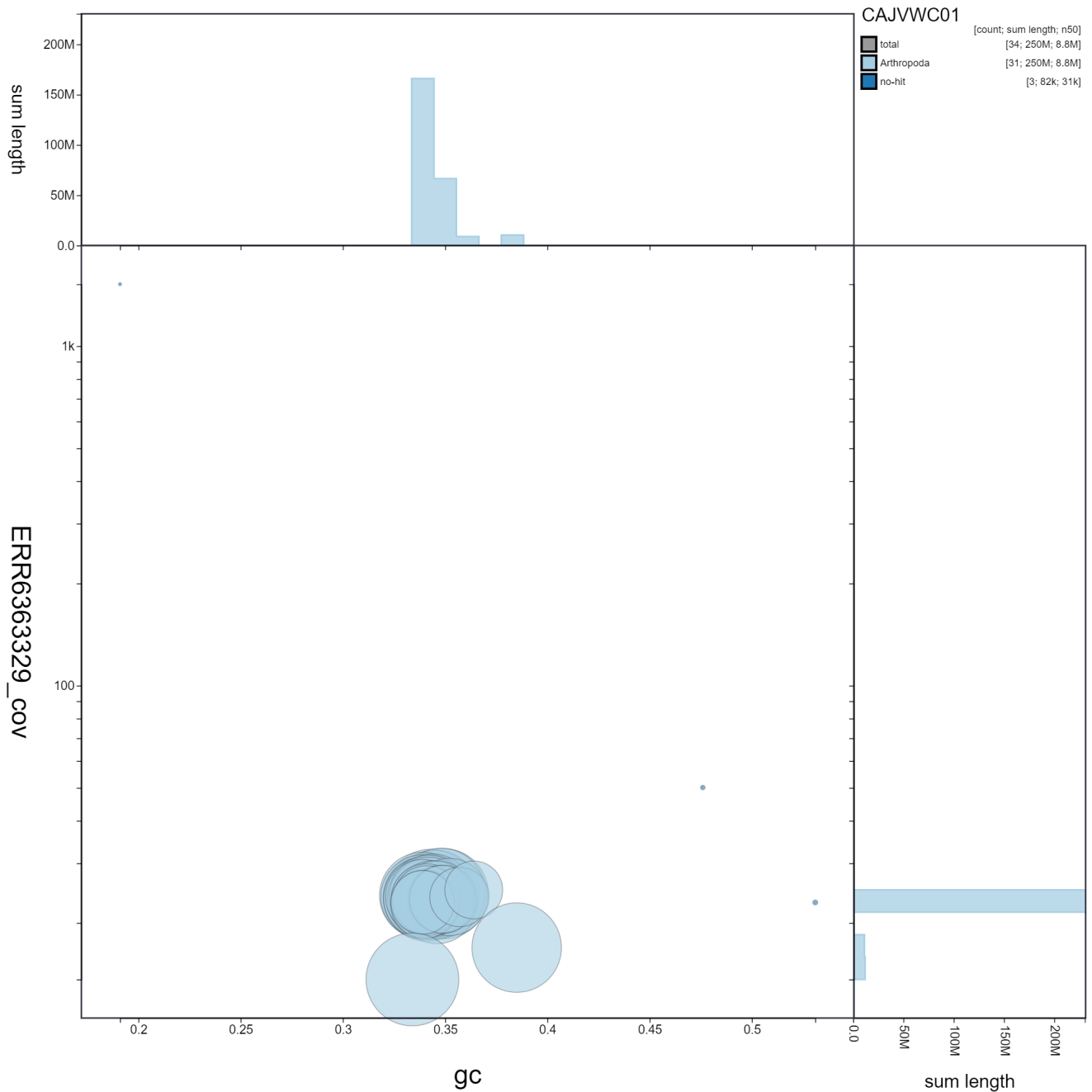


Figure 3. Genome assembly of *Papilio machaon*, ilPapMach1.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at <https://blobtoolkit.genomehubs.org/view/ilPapMach1.1/dataset/CAJVWC01/blob>.

Genome annotation report

The ilPapMach1.1 genome has been annotated using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Papilio_machaon_GCA_912999745.1/). The resulting annotation includes 36,706 transcribed mRNAs from 14,323 protein-coding and 6,538 non-coding genes.

Methods

Sample acquisition and nucleic acid extraction

A single female *P. machaon* specimen (ilPapMach1) was collected using a net near Gheorgheni, Romania (latitude 46.653, longitude 25.37) by Konrad Lohse, Dominik R Laetsch (both University of Edinburgh) and Alex Hayward

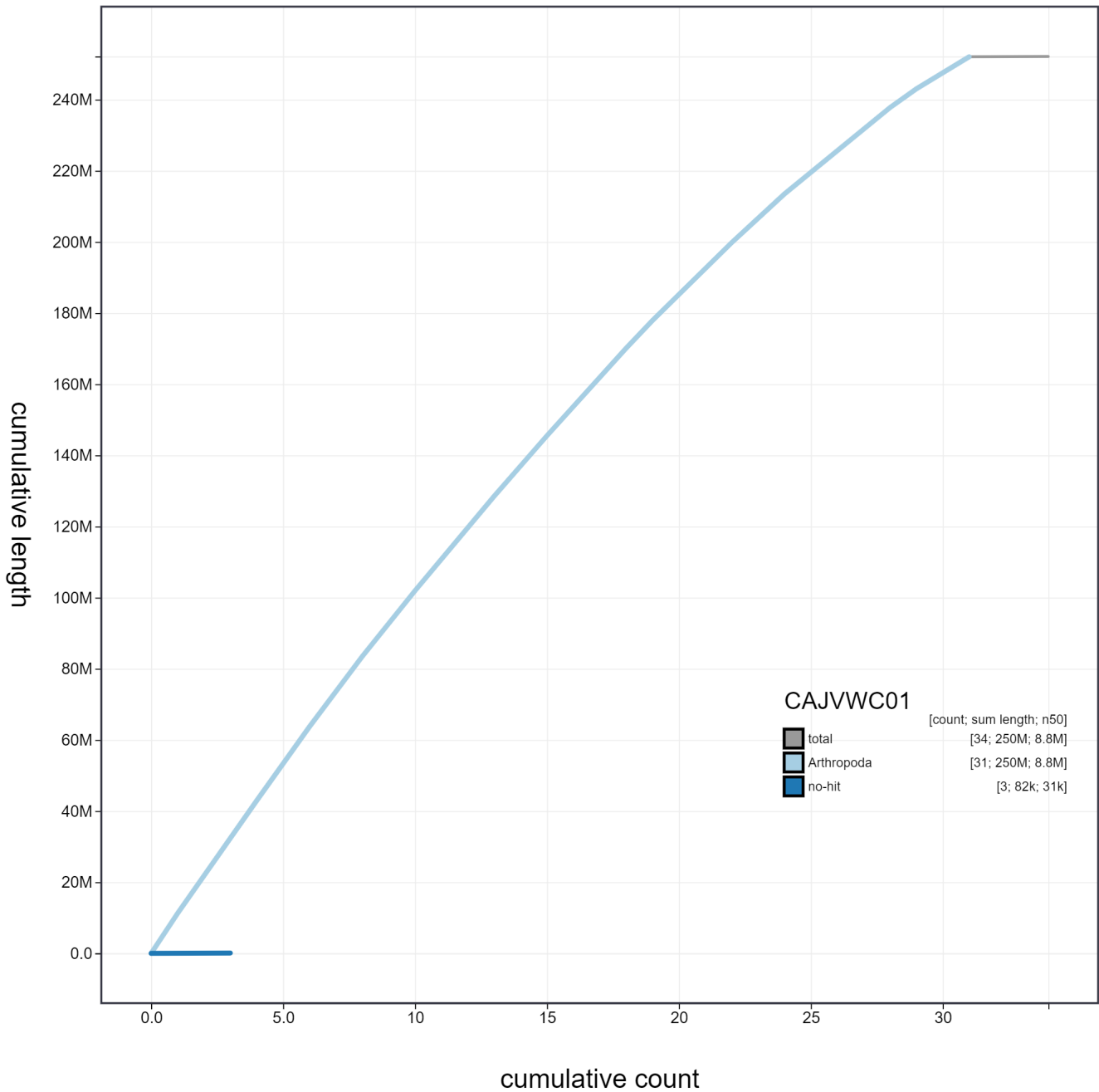


Figure 4. Genome assembly of *Papilio machaon*, iIPapMach1.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxruler. An interactive version of this figure is available at <https://blobtoolkit.genomehubs.org/view/iIPapMach1.1/dataset/CAJVWC01/cumulative>.

(University of Exeter). The specimen was identified by Roger Vila (Institut de Biologia Evolutiva, Barcelona) and snap-frozen at -80°C.

DNA was extracted at the Scientific Operations Core, Wellcome Sanger Institute. The iIPapMach1 sample was weighed and dissected on dry ice with tissue set aside for

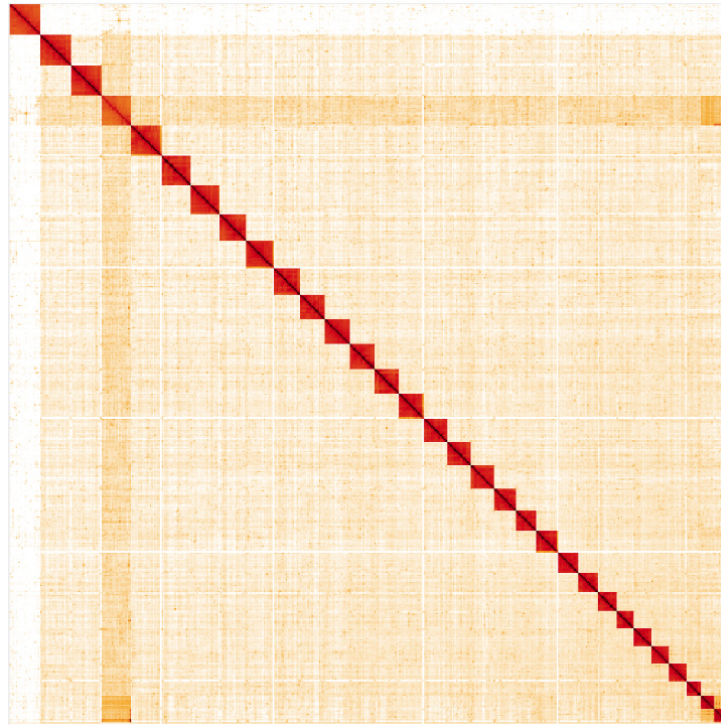


Figure 5. Genome assembly of *Papilio machaon*, ilPapMach1.1: Hi-C contact map. Hi-C contact map of the ilPapMach1.1 assembly, visualised in HiGlass. Chromosomes are arranged in size order from left to right and top to bottom. The interactive Hi-C map can be viewed at <https://genome-note-higlass.tol.sanger.ac.uk/?d=eYBrtPdPTqeG54yJChnrYw>.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Papilio machaon*, ilPapMach1.1.

INSDC accession	Chromosome	Size (Mb)	GC%
OU538757.1	1	10.69	34.9
OU538758.1	2	10.55	34.3
OU538760.1	3	10.41	34.5
OU538761.1	4	10.37	34.8
OU538762.1	5	10.03	34.1
OU538763.1	6	9.72	33.9
OU538764.1	7	9.35	34.3
OU538765.1	8	9.16	33.9
OU538766.1	9	8.91	34
OU538767.1	10	8.87	34
OU538768.1	11	8.78	34.2
OU538769.1	12	8.64	34.4
OU538770.1	13	8.36	34.2
OU538771.1	14	8.23	34.1
OU538772.1	15	8.22	34.2
OU538773.1	16	8.16	34.6

INSDC accession	Chromosome	Size (Mb)	GC%
OU538774.1	17	7.81	35.2
OU538775.1	18	7.38	34.2
OU538776.1	19	7.25	33.9
OU538777.1	20	7.25	34.4
OU538778.1	21	6.96	34.6
OU538779.1	22	6.71	34.1
OU538780.1	23	6.11	34.8
OU538781.1	24	6.09	34.1
OU538782.1	25	6.07	34
OU538783.1	26	6.02	34.9
OU538784.1	27	5.28	33.9
OU538785.1	28	4.61	35.7
OU538786.1	29	4.39	36.4
OU538759.1	W	10.45	38.5
OU538756.1	Z	11.24	33.4
OU538787.1	MT	0.02	19.4
-	Unplaced	0.07	50.7

Hi-C sequencing. Thorax 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 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 Chromium read cloud sequencing libraries were constructed according to the manufacturers' instructions. Sequencing was performed by the Scientific Operations Core at the Wellcome Sanger Institute on Pacific Biosciences SEQUEL II (HiFi) and Illumina HiSeq (10X) instruments. Hi-C data were generated in the Tree of Life laboratory from remaining whole organism tissue of *ilPapMach1* using the Arima v2 kit and sequenced on a 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 as described previously (Howe *et al.*, 2021). Manual curation was performed using 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.

Genome annotation

The Ensembl gene annotation system (Aken *et al.*, 2016) was used to generate annotation for the *Papilio machaon* assembly (GCA_912999745.1). Annotation was created primarily through alignment of transcriptomic data to the

Table 3. Software tools used.

Software tool	Version	Source
Hifiasm	0.15.1	Cheng <i>et al.</i> , 2021
purge_dups	1.2.3	Guan <i>et al.</i> , 2020
SALSA2	2.2	Ghurye <i>et al.</i> , 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	2.0	Uliano-Silva <i>et al.</i> , 2021
HiGlass	1.11.6	Kerpedjiev <i>et al.</i> , 2018
PretextView	0.2.x	https://github.com/wtsi-hpag/PretextView
BlobToolKit	3.2.6	Challis <i>et al.</i> , 2020

genome, with gap filling via protein-to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 2019).

Data availability

European Nucleotide Archive: *Papilio machaon* (common yellow swallowtail). Accession number PRJEB46295; <https://identifiers.org/ena.embl/PRJEB46295> (Wellcome Sanger Institute, 2022)

The genome sequence is released openly for reuse. The *P. machaon* 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>.

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](#)
- 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](#)
- 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): 170–175.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Collins NM, Barkham PJ, Blencowe M, *et al.*: **Ecology and Conservation of the British Swallowtail Butterfly, *Papilio Machaon Britannicus*: Old Questions, New Challenges and Potential Opportunities.** *Insect Conserv Divers*. 2020; **13**(1): 1–9.
[Publisher Full Text](#)
- Fox R, Dennis EB, Brown AF, *et al.*: **A Revised Red List of British Butterflies.** *Insect Conserv Divers*. 2022.
[Publisher Full Text](#)
- Garrison E, Marth G: **Haplotype-Based Variant Detection from Short-Read Sequencing.** 2012; arXiv:1207.3907.
[Publisher Full Text](#)
- 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): 2896–98.
[PubMed 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): g1aa153.
[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](#)
- Li X, Fan D, Zhang W, *et al.*: **Outbred Genome Sequencing and CRISPR/Cas9 Gene Editing in Butterflies.** *Nat Commun*. 2015; **6**: 8212.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Manni M, Berkeley MR, Seppy 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](#)
- 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.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Uliano-Silva M, Nunes JGF, Krasheninnikova 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–D515.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- van Swaay C, Cuttelod A, Collins S, *et al.*: **European Red List of Butterflies.** 2010.
[Publisher Full Text](#)
- Wellcome Sanger Institute: **The genome sequence of the common yellow swallowtail, *Papilio machaon* (Linnaeus, 1758).** European Nucleotide Archive [dataset], 2022.
<https://identifiers.org/ena.embl/PRJEB46295>
- Yamanaka A, Tsujimura Y, Oda Y, *et al.*: **Regulatory Mechanisms in Phenotypic Plasticity of Diapause and Nondiapause Pupal Colouration of the Swallowtail butterfly *Papilio Machaon*.** *Physiol Entomol*. 2013; **38**(2): 133–39.
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status: ? ✓

Version 1

Reviewer Report 28 October 2022

<https://doi.org/10.21956/wellcomeopenres.20090.r52879>

© 2022 Wheat C. 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.

✓ **Christopher W. Wheat** 

Department of Zoology, Faculty of Science, Stockholm University, Stockholm, Sweden

Lohse and co-authors report an exceptionally high-quality genome as a naked genome report, for a common species. The methodological details are clear. I note that much has been done to improve genome reports in this format, notably the detailed legend for the BlobToolKit Snailplot, and the full breakdown of the BUSCO results, showing the completeness of expected genes, and their haploid nature in the genome. Excellent.

Minor critique regarding data accessibility:

Unfortunately, this publication continues to provide a dead end when using the provided links in the Data Accessibility section. Once at the ENA Browser site, finding the actual data desired (e.g. genome fasta, annotation gff, annotation fasta) is simply not easy. I direct readers to an alternative route:

- Take the Assembly Identifier from Table 1, which is iIPapMach1.1 in this case, and simply google that. The top link takes you NIH, and their webpage for this genome. Click on the blue Download Assembly button, and poof, you have all the options of datasets you could desire. I have no idea why ENA doesn't do this.

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: Butterfly genome assembly and study.

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.

Reviewer Report 28 October 2022

<https://doi.org/10.21956/wellcomeopenres.20090.r52953>

© 2022 Reboud E. 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.



Eliette Reboud 

CNRS, Institut des Sciences de l'Evolution de Montpellier, Université de Montpellier, Montpellier, France

This is a high quality, carefully sequenced and assembled genome. It will surely be useful to the community and deserves to be published and approved in the Wellcome Open Research journal. The sequencing and assembling methods are of the highest standard currently available, and sufficiently well-presented and explained.

Given the quality of this assembly, I do not have many comments on the substance of this article, but I do have some reservations about its form and presentation. Overall, I am very confused about the background part of this study. As I read at the end of the article, this initiative to sequence the *Papilio machaon* genome is part of the Darwin Tree of Life (DToL) project. As Threlfall and Blaxter¹ (2022) explain, this project "aims to sequence the genomes of all eukaryotic species in the Atlantic Archipelago of Britain and Ireland". Unfortunately, the Darwin project and its objective are not presented in the article, and it is therefore very difficult for the reader to understand why the two British subspecies of *Papilio machaon* (*P. m. britannicus* and *P. m. gorganus*) are presented in the beginning of the article (background part), while the *P. machaon* specimen of this study was sampled in Romania. Since nothing is said about the subspecies of this Romanian specimen, there is no apparent link between this individual and the UK.

It would be understandable that *P. m. britannicus* was not sequenced (perhaps because of permits?) although this somewhat misses the point of the DToL. Indeed, the authors quite rightly mention that *P. m. britannicus* is endemic to the UK, endangered and protected and its genome would have been of great interest for all these reasons. I therefore assume that the subspecies sampled is *P. m. gorganus* (since it seems to me that it is both present in Romania and an occasional migrant in the UK), but if this is the case, this logical link should be presented much more clearly in the background section of this study. If this is not the case, please find or specify the subspecies of the sampled individual and clearly establish the link between the Romanian individual and the UK.

I also have minor comments on specific parts of the article:

- In the paragraph describing the UK subspecies, the author states that "*P. m. britannicus* is often assumed to be most closely related to *P. m. gorganus*". I would like to know where this statement comes from as I have not been able to find it in the several sources treating the phylogenetic relationships of *Papilio machaon* subspecies^{2,3}?
- In the same paragraph it is stated that *P. m. britannicus* is the largest butterfly in the UK. Yet Riley⁴ (2007) or Koutodistrou and Nudds⁵ (2020) state that *ssp. britannicus* is slightly smaller than *ssp. gorganus*. As the authors state that *ssp. gorganus* is found in the UK, I would recommend that they instead state at the beginning of the paragraph that *P. machaon* (without naming the subspecies) is the largest butterfly in the UK.
- At the end of the background section, the authors briefly mention the pupal colour plasticity (PCP) of *P. machaon*. I don't find this paragraph really relevant. If I understand it correctly, the authors have proposed a question to be answered with their genome, but this seems to me rather artificial and not useful. This genome is of very good quality, and I have no doubt that people could use it for several ongoing research questions.
- In the *Genome sequence report* section, it is stated "*Manual assembly correction [...] removed one haplotypic duplication*" Why is there a "s" to duplication if only one was removed? Is this one haplotypic duplication removal the result of the `purge_dups` analysis?
- In the Methods – Genome assembly section: "*The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva et al., 2021), which performs annotation using MitoFinder*". I understand that this mitochondrion was assembled and annotated but this annotation does not appear on the NCBI BioProject or the Darwin Tree of Life Data Portal, does it? MitoFinder couldn't find the mitochondrial sequence in the .fasta file released and I could not find its position in the .gff file of the project. Would not this mitochondrial sequence and annotation be useful for the community?

Conclusion of the review:

Overall, I find this paper to be very good material for Wellcome Open Research journal, but it would need to be reworked on several points before being fully approved, the most essential being the subspecies sampled and its link to the UK and/or the Darwin of life project. The mitochondrion should also be clearly made available and/or its position indicated in the genome files or a in a separated file.

References

1. Threlfall J, Blaxter M: Launching the Tree of Life Gateway. *Wellcome Open Research*. 2021; **6**. [Publisher Full Text](#)
2. DEMPSTER J, KING M, LAKHANI K: The status of the swallowtail butterfly in Britain. *Ecological Entomology*. 1976; **1** (2): 71-84 [Publisher Full Text](#)
3. Domagała PJ, Lis JA: One Species, Hundreds of Subspecies? New Insight into the Intraspecific Classification of the Old World Swallowtail (*Papilio machaon* Linnaeus, 1758) Based on Two Mitochondrial DNA Markers. *Insects*. 2022; **13** (8). [PubMed Abstract](#) | [Publisher Full Text](#)
4. Riley AM: British and Irish Butterflies-the Complete Identification, Field and Site. *Taunton: Brambleby Books*. 2007. 72
5. Koutouditsou LK, Nudds RL: No evidence of sexual dimorphism in the tails of the swallowtail butterflies *Papilio machaon gorganus* and *P. m. britannicus*. *Ecol Evol*. 2021; **11** (9): 4744-4749

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

Is the rationale for creating the dataset(s) clearly described?

No

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?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Papilionidae, phylogenomics, lepidopteran long-reads and short-reads assemblies

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.
