Check for updates

DATA NOTE

The genome sequence of Cynthia's Fritillary, *Euphydryas cynthia* (Dennis & Schiffermüller), 1775 (Lepidoptera: Nymphalidae)

[version 1; peer review: 3 approved, 1 approved with reservations]

Andreas Sanchez[1], Yannick Chittaro[1], Camille Cornet[2], Kay Lucek[2], Charlotte J. Wright[3], Joana I. Meier[3], Mark L. Blaxter[3], Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team, Wellcome Sanger Institute Scientific Operations: Sequencing Operations, Wellcome Sanger Institute Tree of Life Core Informatics team, Tree of Life Core Informatics collective, Project Psyche Community

[1]Info fauna, Neuchâtel, Switzerland
[2]University of Neuchâtel, Neuchâtel, Switzerland
[3]Tree of Life, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 15 Aug 2025, **10**:447
https://doi.org/10.12688/wellcomeopenres.24696.1
Latest published: 15 Aug 2025, **10**:447
https://doi.org/10.12688/wellcomeopenres.24696.1

Abstract

We present a genome assembly from a female specimen of *Euphydryas cynthia* (Cynthia's Fritillary; Arthropoda; Insecta; Lepidoptera; Nymphalidae). The assembly contains two haplotypes with total lengths of 605.62 megabases and 552.24 megabases. Most of haplotype 1 (99.7%) is scaffolded into 32 chromosomal pseudomolecules, including the W and Z sex chromosomes. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 15.32 kilobases.

Keywords

Euphydryas cynthia, Cynthia's Fritillary, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ? ✓

	1	2	3	4
version 1 15 Aug 2025	✓ view	✓ view	? view	✓ view

1. **Yang Mei**, Zhejiang University, Zhejiang, China
Jilin Agricultural University (Ringgold ID: 85112), Changchun, China
2. **Kristina Gagalova**, Curtin University, Perth, Australia
3. **Fedor Čiampor**, Slovak Academy of Sciences, Bratislava, Slovakia
4. **Somjit Homchan**, Naresuan University, Tha Pho, Thailand

Any reports and responses or comments on the

article can be found at the end of the article.

Corresponding author: Charlotte J. Wright (cw22@sanger.ac.uk)

Author roles: Sanchez A: Investigation, Resources; **Chittaro Y**: Writing – Original Draft Preparation; **Cornet C**: Writing – Review & Editing; **Lucek K**: Resources, Supervision; **Wright CJ**: Funding Acquisition, Project Administration, Supervision; **Meier JI**: Funding Acquisition, Project Administration, Supervision; **Blaxter ML**: Funding Acquisition, Project Administration, Supervision;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540). CC and KL were supported by the Swiss National Science Foundation Grant PCEFP3_202869.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2025 Sanchez A *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: Sanchez A, Chittaro Y, Cornet C *et al.* **The genome sequence of Cynthia's Fritillary, *Euphydryas cynthia* (Dennis & Schiffermüller), 1775 (Lepidoptera: Nymphalidae) [version 1; peer review: 3 approved, 1 approved with reservations]** Wellcome Open Research 2025, **10**:447 https://doi.org/10.12688/wellcomeopenres.24696.1

First published: 15 Aug 2025, **10**:447 https://doi.org/10.12688/wellcomeopenres.24696.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Amphiesmenoptera; Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Papilionoidea; Nymphalidae; Nymphalinae; Melitaeini; Euphydryina; *Euphydryas*; *Euphydryas cynthia* (Dennis & Schiffermüller), 1775 (NCBI:txid104506)

Background

The distribution of Cynthia's Fritillary – *Euphydryas cynthia* (Dennis & Schiffermüller, 1775) – is restricted to the Alps (France, Switzerland, Italy, Austria and Germany) and a few high peaks in southwestern Bulgaria (Rila and Pirin mountains) (Higgins, 1950; Kudrna *et al.*, 2011; Tshikolovets, 2011).

This species displays marked sexual dimorphism (Figure 1). The male may be distinguished by the large amount of white on the basal parts of the upperside and cannot be confused with any other species in its range. The female has an orange-brown chequered pattern and resembles other *Euphydryas* species. It can be distinguished by the presence of a row of black dots, not encircled in yellow, in the orange marginal area of the underside of the hindwings.

The species colonises open alpine grasslands, sometimes dotted with a few dwarf shrubs, and sunny rocky slopes in the alpine zone, mainly above 1 800 m (LSPN, 1987), although sightings are sometimes reported at lower altitudes. The males fly very fast and low over the vegetation.

The female lays her eggs in small clusters under leaves of *Plantago alpina* and *Plantago lanceolata*, but other host plants in the families Orobanchaceae, Plantaginaceae and Violaceae are also used (Clarke, 2024). The young caterpillars live socially in a silken web nest before becoming solitary. Larval development is generally biennial and requires two hibernations (sometimes more). Pupation takes place on or under a rock, sometimes in vegetation.

The imagos fly in one generation, from mid-June to the end of August, depending on altitude and annual weather conditions (date of snowmelt). Their numbers vary greatly from year to year and are regulated in particular by the specialised parasitic wasps *Cotesia cynthiae*, *Ichneumon cynthiae* and *I. obliteratus* (Lafranchis *et al.*, 2015). As with other *Euphydryas* species, the caterpillars, on the other hand, are unpalatable to birds (Bowers, 1981) since they sequester iridoid glycosides (Franke *et al.*, 1987) from their host plants as a chemical defence.

The butterfly is still widespread and common in alpine regions and is not considered globally threatened (Van Swaay *et al.*, 2010). However, global warming is certainly a latent threat to this species, particularly in the case of the isolated populations in Bulgaria. These have been linked to a distinct subspecies, *leonhardi* (Fruhstorfer, 1917), which is also genetically distinct at the COI barcode sequence (Dapporto *et al.*, 2022) and merits further study. The genome presented here will contribute to the conservation of alpine butterflies, and complements the work of De Lesse (1970) who had already studied the chromosomal formula of *E. cynthia* (n = 31). The sequence data were derived from a female specimen (Figure 1) collected from Oberems, Switzerland.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult female *Euphydryas cynthia* (specimen ID SAN28000107, ToLID ilEupCynt1; Figure 1), collected from Oberems, Switzerland (latitude 46.221, longitude 7.6689; elevation 2 658 m) on 22/07/2023. The specimen was collected and identified by Andreas Sanchez (Infofauna).

Figure 1. *Euphydryas cynthia* **a**) male; **b**) female; **c**) Voucher photograph of the specimen used for genome sequencing (ilEupCynt1).

Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The ilEupCynt1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the thorax was homogenised by powermashing using a PowerMasher II tissue disruptor.

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol. DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor®3 for LI PacBio protocol. Sheared DNA was purified by automated SPRI (solid-phase reversible immobilisation). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 70.24 ng/μL and a yield of 3 301.28 ng, with a fragment size of 14.7 kb. The 260/280 spectrophotometric ratio was 1.88, and the 260/230 ratio was 1.68.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Samples with an average fragment size greater than 8 kb and total mass exceeding 400 ng were eligible for the low-input SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences, California, USA), depending on genome size and required sequencing depth. Libraries were prepared using the SMRTbell Prep Kit 3.0 according to the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

The sample was sequenced on a Revio instrument (Pacific Biosciences). The prepared library was normalised to 2 nM, and 15 μL was used for making complexes. Primers were annealed and polymerases bound to generate circularised complexes, following the manufacturer's instructions. Complexes were purified using 1.2X SMRTbell beads, then diluted to the Revio loading concentration (200–300 pM) and spiked with a Revio sequencing internal control. The sample was sequenced on a Revio 25M SMRT cell. The SMRT Link software (Pacific Biosciences), a web-based workflow manager, was used to configure and monitor the run and to carry out primary and secondary data analysis.

Specimen details, sequencing platforms, and data yields are summarised in Table 1.

Table 1. Specimen and sequencing data for BioProject.

Platform	PacBio HiFi	Hi-C
ToLID	ilEupCynt1	ilEupCynt1
Specimen ID	SAN28000107	SAN28000107
BioSample (source individual)	SAMEA115109946	SAMEA115109946
BioSample (tissue)	SAMEA115109991	SAMEA115109989
Tissue	thorax	head
Sequencing platform and model	Revio	Illumina NovaSeq X
Run accessions	ERR13615513	ERR13621444
Read count total	2.01 million	1 055.79 million
Base count total	17.27 Gb	159.42 Gb

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen head tissue of the ilEupCynt1 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

Biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size selected to 400–600 bp using SPRISelect beads. DNA was enriched with Arima-HiC v2 kit Enrichment beads. End repair, A-tailing, and adapter ligation were carried out with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs), following a modified protocol where library preparation occurs while DNA remains bound to the Enrichment beads. Library amplification was performed using KAPA HiFi HotStart mix and a custom Unique Dual Index (UDI) barcode set (Integrated DNA Technologies). Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, libraries were amplified with 10–16 PCR cycles. Post-PCR clean-up was performed with SPRISelect beads. Libraries were quantified using the AccuClear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech).

Prior to sequencing, libraries were normalised to 10 ng/μL. Normalised libraries were quantified again and equimolar and/or

weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq X.

Specimen details, sequencing platforms, and data yields are summarised in Table 1.

Genome assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts (k = 31) was generated from the filtered reads using FastK. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were assembled using Hifiasm in Hi-C phasing mode (Cheng *et al.*, 2021; Cheng *et al.*, 2022), producing two haplotypes. Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). Contigs were further scaffolded with Hi-C data in YaHS (Zhou *et al.*, 2023), using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included 8 breaks and 60 joins. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

Chromosomal painting was performed using lep_busco_painter using Merian elements, which represent the 32 ancestral linkage groups in Lepidoptera (Wright *et al.*, 2024). Painting was based on gene locations from the lepidoptera_odb10 BUSCO analysis and chromosome lengths from the genome index produced using SAMtools faidx (Danecek *et al.*, 2021). Each complete BUSCO (including both single-copy and duplicated BUSCOs) was assigned to a Merian element using a reference database, and coloured positions were plotted along chromosomes drawn to scale.

The Merqury.FK tool (Rhie *et al.*, 2020), run in a Singularity container (Kurtzer *et al.*, 2017), was used to evaluate *k*-mer completeness and assembly quality for both haplotypes using the *k*-mer databases (k = 31) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

The genome was analysed using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake BlobToolKit pipeline (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. Simultaneously, it queries the GoaT database (Challis *et al.*, 2023) to identify relevant BUSCO lineages and runs BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all outputs into a blobdir for visualisation. The BlobToolKit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), with package management via Conda and Bioconda (Grüning *et al.*, 2018), and containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

The genome of a specimen of *Euphydryas cynthia* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 17.27 Gb (gigabases) from 2.01 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 574.89 Mb, with a heterozygosity of 1.09% and repeat content of 32.65%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 29× coverage. Hi-C sequencing produced 159.42 Gb from 1 055.79 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The genome was assembled into two haplotypes using Hi-C phasing. Haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The final assembly has a total length of 605.62 Mb in 85 scaffolds, with 154 gaps, and a scaffold N50 of 20.63 Mb (Table 2).

Most of the assembly sequence (99.7%) was assigned to 32 chromosomal-level scaffolds, representing 30 autosomes and the W and Z sex chromosomes. Chromosomes W and Z were identified by HiC signal. Contigs on the W chromosome have uncertain order and orientation. The chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 2; Table 3). Chromosome painting with Merian

Table 2. Genome assembly statistics.

Assembly name	ilEupCynt1.hap1.1	ilEupCynt1.hap2.1
Assembly accession	GCA_964274225.1	GCA_964274195.1
Assembly level	chromosome	scaffold
Span (Mb)	605.62	552.24
Number of chromosomes	32	N/A
Number of contigs	239	153
Contig N50	7.0 Mb	7.32 Mb
Number of scaffolds	85	69
Scaffold N50	20.63 Mb	20.38 Mb
Longest scaffold length (Mb)	26.33	N/A
Sex chromosomes	W and Z	N/A
Organelles	Mitochondrial genome: 15.32 kb	N/A

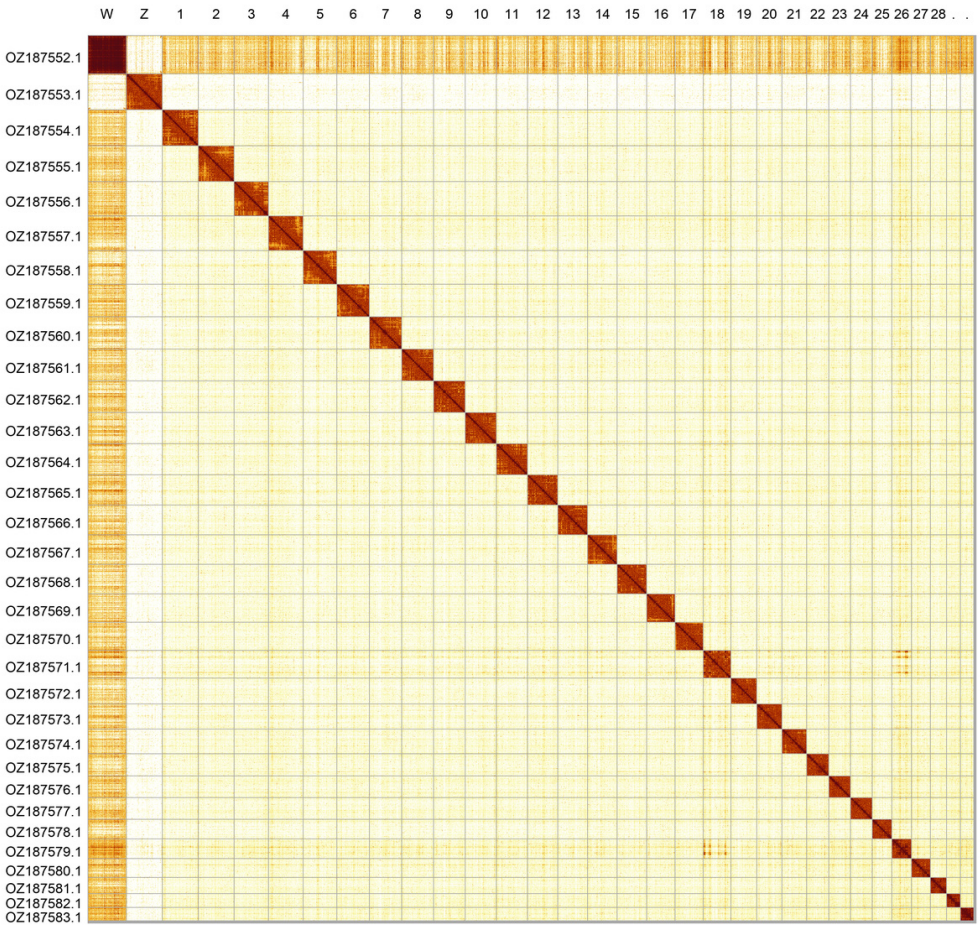

Figure 2. Hi-C contact map of the *Euphydryas cynthia* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Euphydryas cynthia* ilEupCynt1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ187554.1	1	24.47	35	M8
OZ187555.1	2	24.40	35	M2
OZ187556.1	3	23.56	35	M17;M20
OZ187557.1	4	23.55	35	M9
OZ187558.1	5	22.97	35	M1
OZ187559.1	6	22.12	35	M12
OZ187560.1	7	22.11	35	M3
OZ187561.1	8	21.69	34.50	M7
OZ187562.1	9	21.43	34.50	M18
OZ187563.1	10	21.38	35	M5
OZ187564.1	11	21.12	35	M6
OZ187565.1	12	20.63	35	M22
OZ187566.1	13	20.33	34.50	M16
OZ187567.1	14	20.17	34.50	M4
OZ187568.1	15	20.05	34.50	M21
OZ187569.1	16	19.42	35	M10
OZ187570.1	17	19.10	34.50	M15
OZ187571.1	18	18.90	35	M23
OZ187572.1	19	17.68	35	M11
OZ187573.1	20	17.17	35	M13
OZ187574.1	21	16.78	35	M14
OZ187575.1	22	15.07	34.50	M24
OZ187576.1	23	14.80	35.50	M19
OZ187577.1	24	14.76	34.50	M26
OZ187578.1	25	13.45	34.50	M28
OZ187579.1	26	13.32	37	M30
OZ187580.1	27	12.84	35	M27
OZ187581.1	28	11.02	35.50	M25
OZ187582.1	29	9.57	36	M29
OZ187583.1	30	9.02	37	M31
OZ187552.1	W	26.33	41	N/A
OZ187553.1	Z	24.56	34	MZ

elements illustrates the distribution of orthologues along chromosomes and highlights patterns of chromosomal evolution relative to Lepidopteran ancestral linkage groups (Figure 3).

The mitochondrial genome was also assembled. This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

For haplotype 1, the estimated QV is 60.0, and for haplotype 2, 60.0. When the two haplotypes are combined, the assembly achieves an estimated QV of 60.0. The *k*-mer completeness is 78.35% for haplotype 1, 74.92% for haplotype 2, and 99.51% for the combined haplotypes (Figure 4). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) (Kriventseva *et al.*, 2019) identified 98.8% of the expected gene set (single = 98.3%, duplicated = 0.5%) for haplotype 1. The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.Q60**, meeting the recommended reference standard.

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been 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, 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 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.

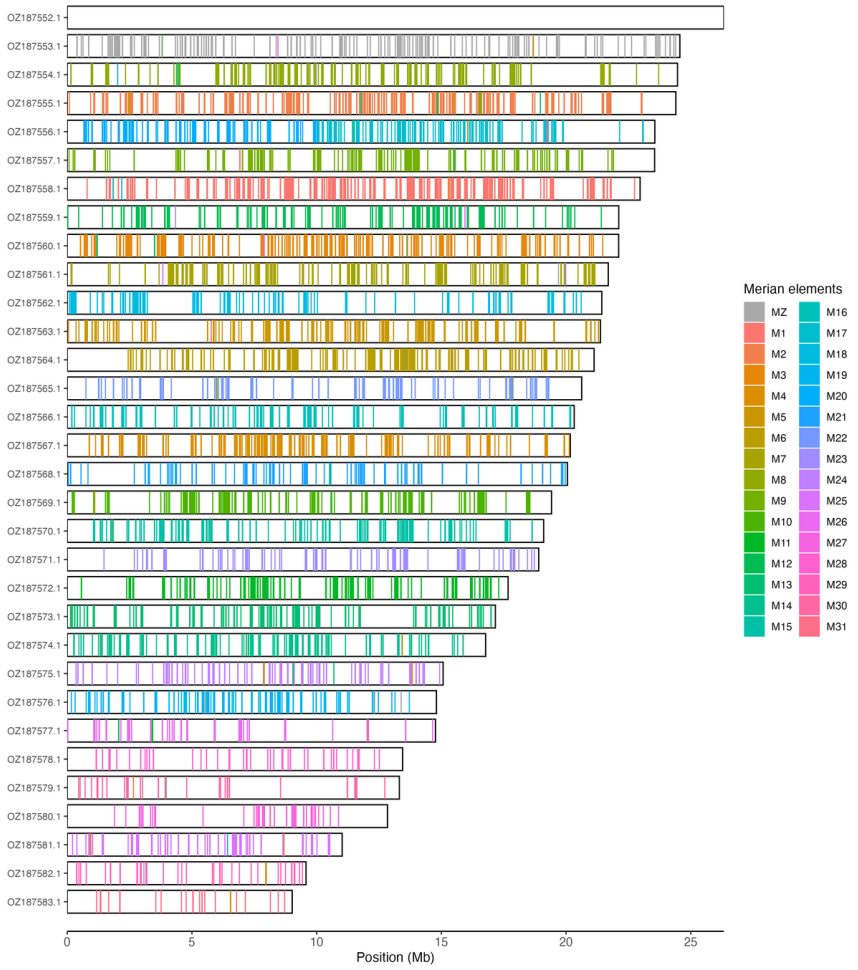

Figure 3. Merian elements painted across chromosomes in the ilEupCynt1.hap1.1 assembly of *Euphydryas cynthia*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

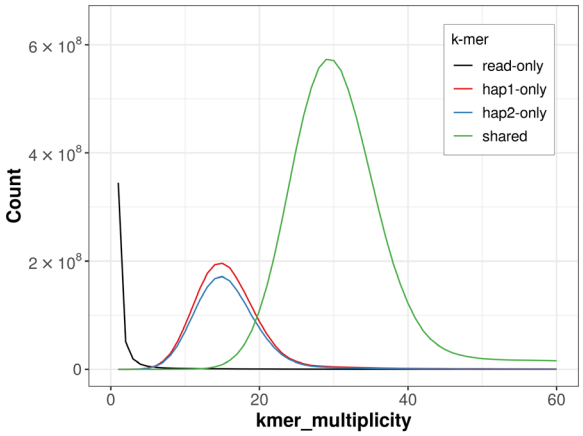

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve (the homozygous peak) corresponds to *k*-mers shared by both haplotypes and the red and blue curves (the heterozygous peaks) show *k*-mers found only in one of the haplotypes.

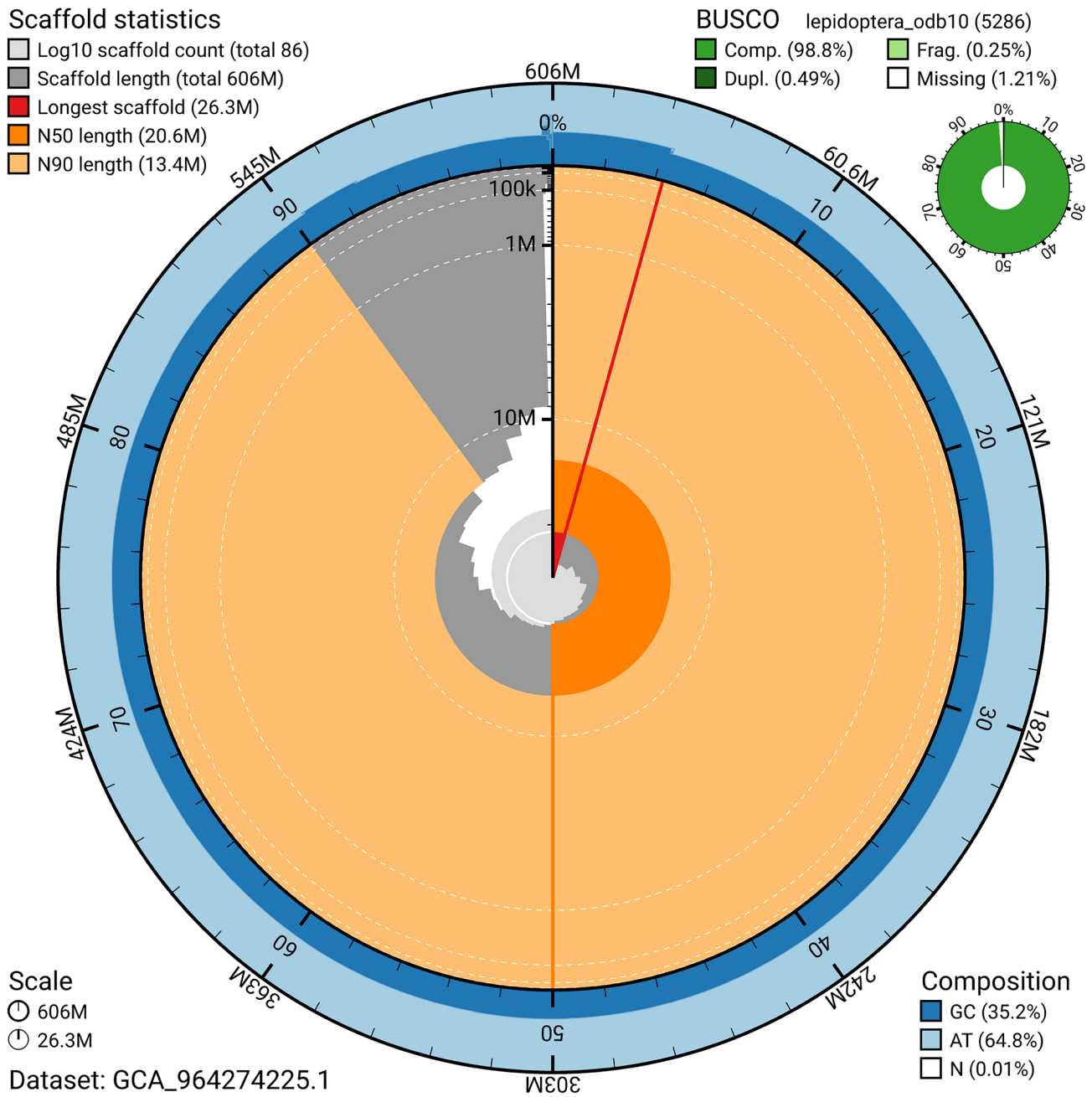

Figure 5. Assembly metrics for ilEupCynt1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

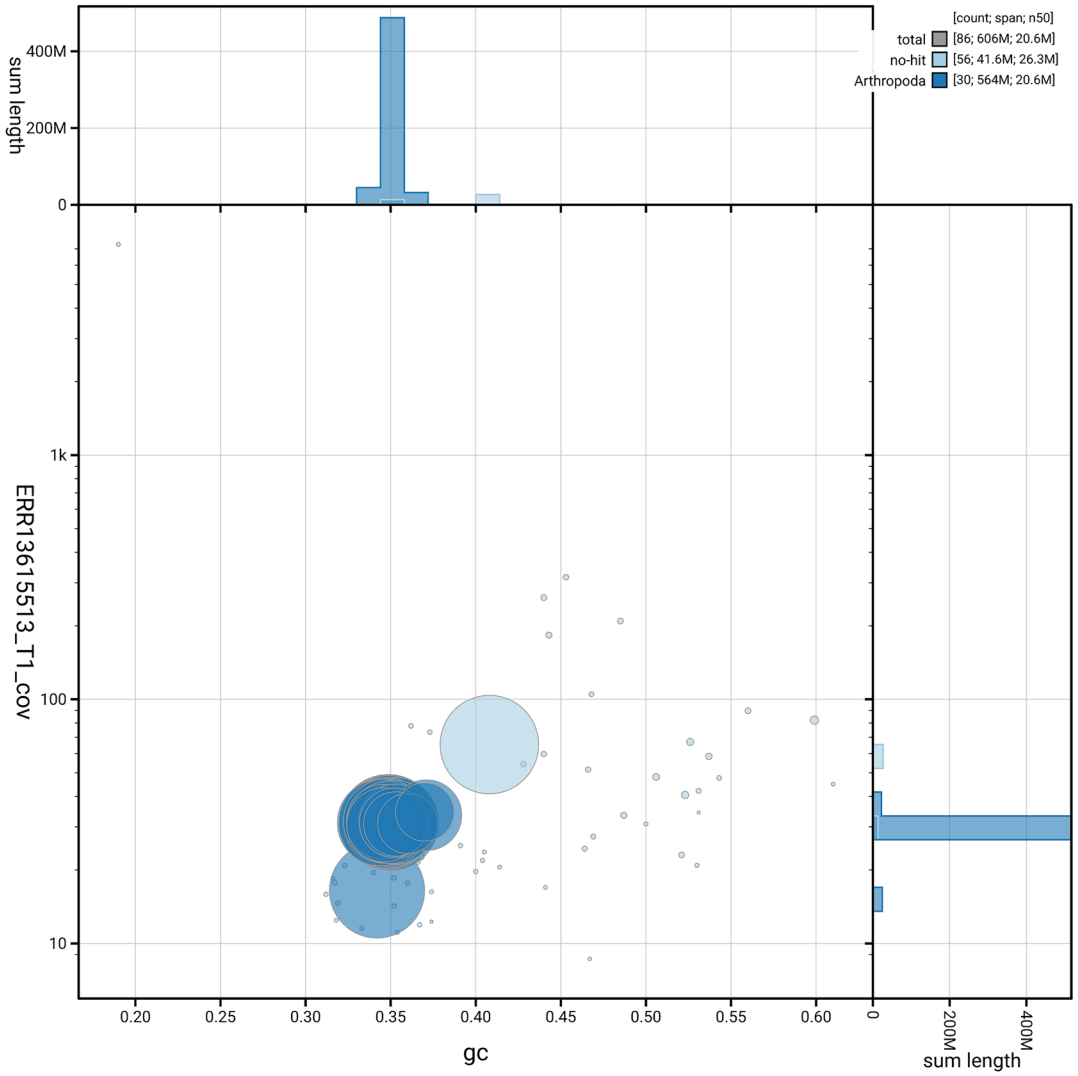

Figure 6. BlobToolKit GC-coverage plot for ilEupCynt1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the BlobToolKit viewer.

Table 4. Earth Biogenome Project summary metrics for the *Euphydryas cynthia* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.7.Q60	6.C.Q40
Contig N50 length	7 Mb	≥ 1 Mb
Scaffold N50 length	20.63 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 60.0; haplotype 2: 60.0; combined: 60.0	≥ 40
k-mer completeness	Haplotype 1: 78.35%; Haplotype 2: 74.92%; combined: 99.51%	≥ 95%
BUSCO	C:98.8%[S:98.3%,D:0.5%], F:0.2%,M:1.0%,n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.70%	≥ 90%

Data availability

European Nucleotide Archive: Euphydryas cynthia (Cynthia's fritillary). Accession number PRJEB79451. The genome sequence is released openly for reuse. The *Euphydryas cynthia* genome sequencing initiative is part of the Sanger Institute Tree of Life Programme (PRJEB43745) and Project Psyche (PRJEB71705). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through Ensembl at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

Pipelines used for genome assembly at the WSI Tree of Life are available at https://pipelines.tol.sanger.ac.uk/pipelines. Table 5 lists software versions used in this study.

Author information

Contributors are listed at the following links:

- Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team
- Wellcome Sanger Institute Scientific Operations – Sequencing Operations
- Wellcome Sanger Institute Tree of Life Core Informatics team
- Tree of Life Core Informatics collective
- Project Psyche Community.

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextSnapshot	N/A	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
samtools	1.19.2	https://github.com/samtools/samtools

Software	Version	Source
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.6.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

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

Altschul SF, Gish W, Miller W, *et al.*: **Basic local alignment search tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text

Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the Universal Protein Knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531.
PubMed Abstract | Publisher Full Text | Free Full Text

Bowers MD: **Unpalatability as a defense strategy of western checkerspot butterflies (*Euphydryas* Scudder, Nymphalidae).** *Evolution.* 1981; **35**(2): 367–75.
PubMed Abstract | Publisher Full Text

Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368.
PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24.
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

Cheng H, Jarvis ED, Fedrigo O, *et al.*: **Haplotype-resolved assembly of diploid genomes without parental data.** *Nat Biotechnol.* 2022; **40**(9): 1332–1335.
PubMed Abstract | Publisher Full Text | Free Full Text

Clarke HE: **A checklist of European butterfly larval foodplants.** *Ecol Evol.* 2024; **14**(1): e10834.
PubMed Abstract | Publisher Full Text | Free Full Text

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008.
PubMed Abstract | Publisher Full Text | Free Full Text

Dapporto L, Menchetti M, Vodă R, *et al.*: **The atlas of mitochondrial genetic diversity for Western Palaearctic butterflies.** *Global Ecology and Biogeography.* 2022; **31**(11): 2184–90.
Publisher Full Text

de Lesse H: **Formules chromosomiques de quelques Rhopalocères paléarctiques (Lep.).** *Bulletin de la Société entomologique de France.* 1970; **75**(7–8): 214–16.
Publisher Full Text

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
PubMed Abstract | Publisher Full Text

Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
PubMed Abstract | Publisher Full Text | Free Full Text

Franke A, Rimpler H, Schneider D: **Iridoid glycosides in the butterfly *Euphydryas cynthia* (Lepidoptera, Nymphalidae).** *Phytochemistry.* 1987; **26**(1): 103–6.
Publisher Full Text

Fruhstorfer H: **Neue rhopaloceren aus der sammlung leonhard.** *Archiv für Naturgeschichte.* 1917; **82A**(2): 1–28.

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
PubMed Abstract | Publisher Full Text | Free Full Text

Higgins LG: **A descriptive catalogue of the Palearctic *Euphydryas*.** *Transactions of the Royal Entomological Society, London.* 1950; **101**(12): 435–87.
Publisher Full Text

Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025.
Publisher Full Text

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 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

Kriventseva EV, Kuznetsov D, Tegenfeldt F, *et al.*: **OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs.** *Nucleic Acids Res.* 2019; **47**(D1): D807–D811.
PubMed Abstract | Publisher Full Text | Free Full Text

Kudrna O, Harpke A, Lux K, *et al.*: **The distribution atlas of butterflies in Europe**. Gesellschaft für Schmetterlingsschutz, 2011.
Reference Source

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
PubMed Abstract | Publisher Full Text | Free Full Text

Lafranchis T, Jutzeler D, Guillosson JY, *et al.*: **La vie des papillons. Écologie, biologie et comportement des rhopalocères de France**. Diatheo, 2015.
Reference Source

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text | Free Full Text

LSPN: **Les papillons de jour et leurs biotopes. espèces, dangers qui les menacent, protection.** Fotorotar AG, 1987; **1**.
Reference Source

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

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2.
Reference Source

Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat Commun.* 2020; **11**(1): 1432.
PubMed Abstract | Publisher Full Text | Free 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): 1665–1680.
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

Tshikolovets VV: **Butterflies of Europe & the mediterranean area**. Kiev: Tshikolovets Publications, 2011.
Reference Source

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.
PubMed Abstract | Publisher Full Text | Free Full Text

Van Swaay C, Cuttelod A, Collins S, *et al.*: **European red list of butterflies**. Luxembourg: Publications Office of the European Union, 2010.
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

Wright CJ, Stevens L, Mackintosh A, *et al.*: **Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera.** *Nat Ecol Evol.* 2024; **8**(4): 777–90.
PubMed Abstract | Publisher Full Text | Free Full Text

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔ ? ✔

Version 1

Reviewer Report 29 December 2025

https://doi.org/10.21956/wellcomeopenres.27212.r139794

© **2025 Homchan S.** 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.

✔ **Somjit Homchan**
Naresuan University, Tha Pho, Thailand

This manuscript is well-prepared, and the protocols employed in this study are robust, providing a comprehensive report of the whole genome sequence. I have a few suggestions to further strengthen the manuscript for Indexing.
1. Species taxonomy: Please check species, *Euphydryas cynthia* (Dennis & Schiffermüller, 1775), or *Euphydryas cynthia* (Denis & Schiffermüller, 1775). Additionally, I found that the species name is sometimes cited with only one authorship as *Euphydryas cynthia* (Schiffermüller, 1775). Therefore, which one is the right one
2. Regarding the two haplotypes reported in this genome sequencing, haplotype 1 and haplotype 2 show significant differences. Please clarify for the reader how these differences may have arisen.
3. The author reported the heterozygosity rate of 1.09%. Is this level typical for this butterfly species? is it low or high? It would be beneficial to include a brief discussion and explain about this value.

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: Molecular genetic markers, Population genetics, Cytogenetics

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 25 November 2025

https://doi.org/10.21956/wellcomeopenres.27212.r139790

© 2025 Čiampor F. 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.

?

Fedor Čiampor
Slovak Academy of Sciences, Bratislava, Slovakia

I am not an expert on Lepidoptera taxonomy, but according to gbif.org, *Euphydryas cynthia* (Denis & Schiffermüller), 1775 is a synonym of the species *Hypodryas cynthia* (Denis & Schiffermüller), 1775, but in the European butterfly dataset, which is also indexed on gbif.org, the species is listed in the genus Euphydryas. I cannot judge which is correct, but before the article is indexed, it definitely needs to be verified, as this is basic information about sequenced species. It seems to me that Higgins 1978 described a new genus *Hypodryas* and moved the species *E.cynthia* into it. Perhaps this is potentially incorrect and lepidopterists adhere to the original classification in the genus *Euphydryas*, but no one has formally corrected the taxonomy, so the correct name to use should be *Hypodryas cynthia*. However, maybe the synonymization of *Hypodryas* with *Euphydryas* has been indexed, but I have not found this information. It needs to be found and the currently valid name used. I have received a reaction to my request from the gbif: „GBIF will soon update the taxonomic backbone to the extended release of Catalogue of Life. According to this resource, *Euphydryas cynthia (Denis & Schiffermüller, 1775)* is a synonym of *Hypodryas cynthia ([Denis & Schiffermüller], 1775)*."
If invalid/incorrect names are used, this causes serious problems in the use of data. In addition, you list the authors as Dennis & Schiffermüller, but the correct is Denis & Schiffermüller.
Furthermore, in the taxonomy of the species, you refer to (NCBI:txid104506), where *Euphydryas cynthia* is listed with incorrect authorship: (Schiffermuller, 1775). I would recommend not refer to NCBI in taxonomy because it is not a very reliable database from a taxonomic point of view.
I found information about isolated occurrences of the species in Belarus, Montenegro, and Spain. It is good to follow current information about distribution
The genome of a related butterfly (Coenonympha arcania) from the same family was recently indexed. The authors should comment on whether it is OK that there is such a significant difference in the size of their genomes and the number of chromosomes. What could be the cause of this? This would be helpful for the future use of such generated data.

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?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: invertebrate taxonomy, phylogeny, DNA (meta)barcoding

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 25 November 2025

https://doi.org/10.21956/wellcomeopenres.27212.r139791

© 2025 Gagalova K. 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.

Kristina Gagalova
Curtin University, Perth, Australia

The work is well developed and presented. I only have a few comments

Major
* What do you think is the difference in size between the two haplotypes, 50Mbp? This may suggest biological and sequencing issues, could you elaborate?
are the following relevant to explain the discrepancy: repetitive element expansion, hemizygous W regions, assembly overphasing, or uncollapsed haplotigs
* 1.09% heterozygosity was assumed for the whole genome, is there any difference between the sex and somatic chromosomes? Clarify whether heterozygosity was recalculated after W chromosome filtering and whether any significant difference may be expected
* Sex chromosomes have a quite different features from somatic chromosomes, can you please separate these from the rest of the results and mention how these contigs were identified and curated, separately from the rest? That will be of value for others assembling similar genomes
* BlobToolKit was use but there is no description whether biological contaminants are found. Can you please elaborate on this? low-coverage contaminants usually exist, especially with whole-body DNA extractions from insects. How much was found of these and were they removed?

Minor
* Which FastK, GenomeScope settings were used
* Specific Hifiasm parameters
* Hi-C read processing and filtering thresholds
* Manual curation criteria (i.e. 8 breaks and 60 joins, why?)
* Versioning of Nextflow/conda profiles used in pipelines.

* Is the mitochondrion genome circular?

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 assemblies, genomics, plant and insects genomes, complex genomes, contamination screening, pangenomes, genome annotations, nextflow pipelines development, HPC workflows

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 05 September 2025

https://doi.org/10.21956/wellcomeopenres.27212.r130721

© 2025 Mei Y. 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.

Yang Mei
[1] Zhejiang University, Zhejiang, China
[2] Jilin Agricultural University (Ringgold ID: 85112), Changchun, Jilin, China

The authors provide a high-quality assembly of *Euphydryas cynthia*, especially including the W chromosome. The methods and results were described well and clearly. I think it could be indexed in the current version.
But I also wonder why there is 50 mb difference between the two haplotypes. It is because of the high heterozygosity?
Why is the heterozygosity high based on the one adult sample?

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: Insect 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.