DATA NOTE

The genome sequence of the Short-tailed Blue, *Cupido argiades* (Pallas, 1771) (Lepidoptera: Lycaenidae)

[version 1; peer review: 3 approved]

Roger Vila[1], Zdeněk F. Fric[2], Charlotte J. Wright[3], Joana I. Meier[3], Mark L. Blaxter[3]

[1]Institut de Biologia Evolutiva (CSIC-UPF), Barcelona, Catalonia, Spain
[2]Biology Centre CAS, Institute of Entomology, Ceske Budejovice, Czech Republic
[3]Tree of Life Programme, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 23 Jan 2026, **11**:65
https://doi.org/10.12688/wellcomeopenres.25796.1
Latest published: 23 Jan 2026, **11**:65
https://doi.org/10.12688/wellcomeopenres.25796.1

Abstract

We present a genome assembly from a female specimen of *Cupido argiades* (Short-tailed Blue; Arthropoda; Insecta; Lepidoptera; Lycaenidae). The assembly contains two haplotypes with total lengths of 445.87 megabases and 384.46 megabases. Most of haplotype 1 (99.85%) is scaffolded into 25 chromosomal pseudomolecules, including the W, Z_1, and Z_2 sex chromosomes. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 15.41 kilobases. This work is part of Project Psyche, a collaborative programme generating genomes for European butterflies and moths.

Keywords

Cupido argiades; Short-tailed Blue; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 1 23 Jan 2026	✓ view	✓ view	✓ view

1. **Gurinder Kaur Walia**, Punjabi University, Patiala, India
2. **Zsófia Fekete**, University of Oulu, Oulu, Finland
   Niina Kiljunen, University of Oulu, Oulu, Finland
3. **Océane Seudre**, Queen Mary University of London, London, UK

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: Vila R: Investigation, Resources, Supervision, Writing – Review & Editing; **Fric ZF**: Writing – Original Draft Preparation; **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). RV was funded by Grant PID2022-139689NB-I00 (MICIU/ AEI/ 10.13039/501100011033 and ERDF, EU) and grant 2021-SGR-00420 (Departament de Recerca i Universitats, Generalitat de Catalunya).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Vila R, Fric ZF, Wright CJ *et al.* **The genome sequence of the Short-tailed Blue, *Cupido argiades* (Pallas, 1771) (Lepidoptera: Lycaenidae) [version 1; peer review: 3 approved]** Wellcome Open Research 2026, **11**:65 https://doi.org/10.12688/wellcomeopenres.25796.1

First published: 23 Jan 2026, **11**:65 https://doi.org/10.12688/wellcomeopenres.25796.1

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; Lycaenidae; Polyommatinae; *Cupido*; *Cupido argiades* (Pallas, 1771) (NCBI:txid596724)

Background

The Short-tailed Blue or Tailed Cupid (*Cupido argiades*) is a relatively small butterfly with short tails on the hindwings, with at least two prominent orange lunules close to the tails on the wing undersides. The wing undersides are otherwise whitish with black spots. The uppersides of the wings are blue (males) or brownish (females). The species lives in variable grassland habitats, but mostly close to shrublands or forest margins. The species is vagrant, distributed throughout most of the Palaearctic region, extending to the east in India, China and south-east Asia (Wang & Fan, 2002).

The larval host plants are various genera of the family Fabaceae such as *Lotus*, *Medicago*, *Trifolium*, or *Ulex*. *C. argiades* has only a weak association with ants (Fiedler, 2021). The butterfly is multivoltine and usually has two or three generations per year (Beneš *et al.*, 2002). While it is not threatened and even extending its range in south-eastern and central Europe, the species declined in western Europe in the second half of the 20th century with extinction in Belgium (Maes *et al.*, 2019). The species does not exhibit any mitochondrial variability across Europe (Sucháčková Bartonová *et al.*, 2024).

We present a chromosome-level genome sequence for *C. argiades*, sequenced as part of Project Psyche. The sequence data were derived from a female specimen (Figure 1) collected from Gréixer, Guardiola de Berguedà, Barcelona, Catalonia, Spain.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult female *Cupido argiades* (specimen ID SAN28000284, ToLID ilCupArgi2; Figure 1), collected from Gréixer, Guardiola de Berguedà, Barcelona, Catalonia, Spain (latitude 42.2884, longitude 1.841) on 2018-08-19. Another specimen (male) was used for RNA sequencing (specimen ID SAN28000283, ToLID ilCupArgi1). The specimens were collected and identified by Roger Vila (Institut de Biologia Evolutiva).

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 ilCupArgi2 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the whole organism was homogenised by powermashing using a PowerMasher II tissue disruptor.

Figure 1. Voucher photographs of the *Cupido argiades* (ilCupArgi2) specimen used for genome sequencing. Top: dorsal view; bottom: ventral view.

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 32.93 ng/μL and a yield of 1 547.71 ng, with a fragment size of 17.2 kb.

RNA was extracted from whole organism tissue of ilCupArgi1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol. The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), 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.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the whole organism of the ilCupArgi2 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 to create 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.

RNA-seq library preparation and sequencing

Libraries were prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs), following the manufacturer's instructions. Poly(A) mRNA in the total RNA solution was isolated using oligo(dT) beads, converted to cDNA, and uniquely indexed; 14 PCR cycles were performed. Libraries were size-selected to produce fragment lengths of 100–300 bp. Libraries were quantified, normalised, pooled to a final concentration of 2.8 nM, and diluted to 150 pM for loading. Sequencing was carried out on the Illumina NovaSeq X to generate 150-bp paired-end reads.

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)

Table 1. Specimen and sequencing data for BioProject PRJEB80915.

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	ilCupArgi2	ilCupArgi2	ilCupArgi1
Specimen ID	SAN28000284	SAN28000284	SAN28000283
BioSample (source individual)	SAMEA115768736	SAMEA115768736	SAMEA115768734
BioSample (tissue)	SAMEA115768853	SAMEA115768853	SAMEA115768851
Tissue	whole organism	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq X	Illumina NovaSeq X
Run accessions	ERR13800475	ERR13802619	ERR15551481
Read count total	1.32 million	649.13 million	95.23 million
Base count total	14.83 Gb	98.02 Gb	14.38 Gb

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 nine breaks and 46 joins. This reduced the scaffold count by 36.6%, increased scaffold N50 by 1.8%, and increased the total assembly length by 3.9%. The curation process is described 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

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 database ($k = 31$) 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 (Di Tommaso *et al.*, 2017) implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified from the NCBI Taxonomy (Schoch *et al.*, 2020). For the three domain-level 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 containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

We used lep_busco_painter to paint Merian elements along chromosomes (Wright *et al.*, 2024). Merian elements represent the 32 ancestral linkage groups in Lepidoptera. The painting process utilised BUSCO gene locations from the lepidoptera_odb10 set (Kriventseva *et al.*, 2019) and chromosome lengths from NCBI Datasets. Each complete BUSCO gene (both single-copy and duplicated) was assigned to a Merian element based on a reference database, then plotted along chromosomes drawn to scale.

Genome sequence report

Sequence data

PacBio sequencing of the *Cupido argiades* specimen generated 14.83 Gb (gigabases) from 1.32 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 410.61 Mb, with a heterozygosity of 1.72% and repeat content of 36.25% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 35× coverage. Hi-C sequencing produced 98.02 Gb from 649.13 million reads, which were used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. 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 445.87 Mb in 44 scaffolds, with 101 gaps, and a scaffold N50 of 18.36 Mb (Table 2).

Most of the assembly sequence (99.85%) was assigned to 25 chromosomal-level scaffolds, representing 22 autosomes and the W, Z_1, and Z_2 sex chromosomes. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Chromosome painting with Merian elements illustrates the distribution of orthologues along chromosomes and highlights patterns of chromosomal evolution relative to Lepidopteran ancestral linkage groups (Figure 4). Chromosomes W, Z_1 and Z_2 were assigned based on the Hi-C signal.

The mitochondrial genome was also assembled (length 15.41 kb, OZ195136.1). 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 65.8, and for haplotype 2, 65.9. When the two haplotypes are combined, the assembly achieves an estimated QV of 65.8. The *k*-mer completeness is 73.28% for haplotype 1, 65.73% for haplotype 2, and 99.24% for the combined haplotypes (Figure 5).

BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 97.1% of the expected gene set (single = 94.2%, duplicated = 3.0%) in haplotype 1. For haplotype 2, BUSCO analysis identified 90.7% of the expected gene set (single = 90.2%, duplicated = 0.5%). The snail plot in

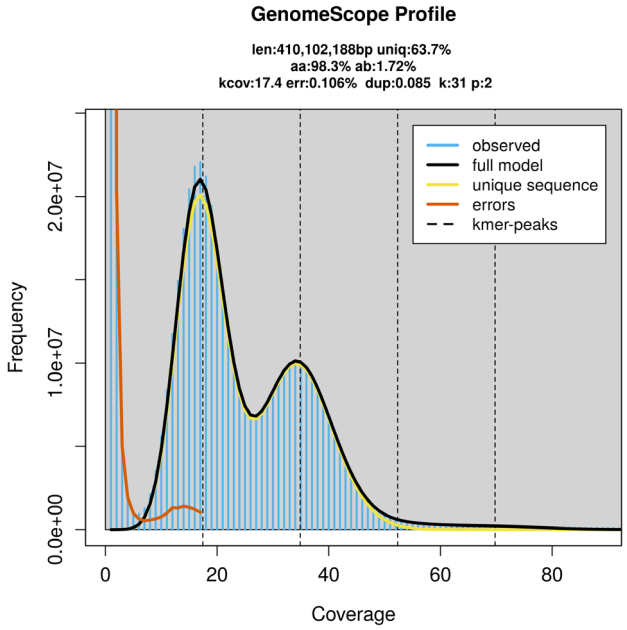

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 2. Genome assembly statistics.

Assembly name	ilCupArgi2.hap1.1	ilCupArgi2.hap2.1
Assembly accession	GCA_964277285.1	GCA_964277125.1
Assembly level	chromosome	scaffold
Span (Mb)	445.87	384.46
Number of chromosomes	25	scaffold-level
Number of contigs	145	112
Contig N50	8.09 Mb	7.63 Mb
Number of scaffolds	44	38
Scaffold N50	18.36 Mb	18.09 Mb
Longest scaffold length (Mb)	25.31	-
Sex chromosomes	W, Z_1, and Z_2	-
Organelles	Mitochondrion: 15.41 kb	-

Figure 6 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 7 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.Q65**, 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

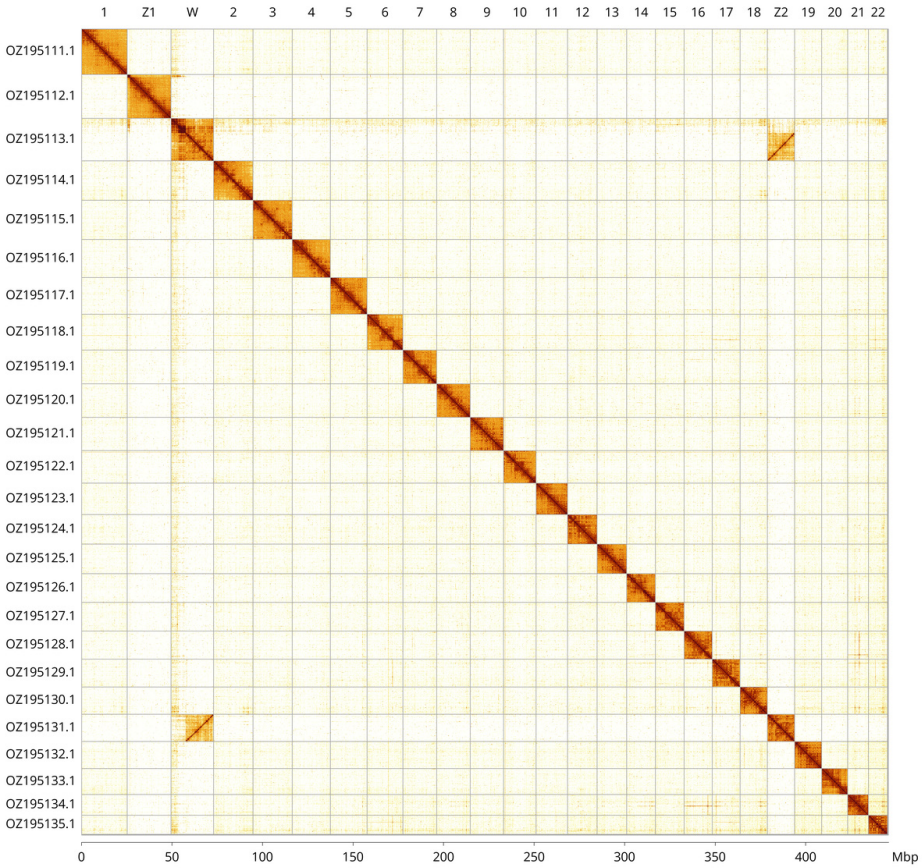

Figure 3. Hi-C contact map of the *Cupido argiades* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Cupido argiades* ilCupArgi2.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ195111.1	1	25.31	37	M1;M19
OZ195114.1	2	21.70	37	M18;M30
OZ195115.1	3	21.64	37	M11;M23
OZ195116.1	4	21.20	36.50	M25;M5
OZ195117.1	5	20.21	36.50	M12;M29
OZ195118.1	6	19.75	36.50	M14;M26
OZ195119.1	7	18.66	37	M17;M20
OZ195120.1	8	18.55	37	M2
OZ195121.1	9	18.36	36.50	M9
OZ195122.1	10	18.04	37	M8
OZ195123.1	11	17.34	37	M3
OZ195124.1	12	16.36	37	M4
OZ195125.1	13	16.31	37	M7
OZ195126.1	14	15.85	36.50	M16
OZ195127.1	15	15.83	37	M6
OZ195128.1	16	15.55	37	M21
OZ195129.1	17	15.39	37	M15
OZ195130.1	18	15.10	37	M13
OZ195132.1	19	14.92	37	M10
OZ195133.1	20	14.39	37.50	M28;M31
OZ195134.1	21	11.46	37	M24
OZ195135.1	22	10.51	37	M27
OZ195113.1	W	23.53	37.50	M22
OZ195112.1	Z_1	24.22	37.50	MZ
OZ195131.1	Z_2	15.04	37	M22

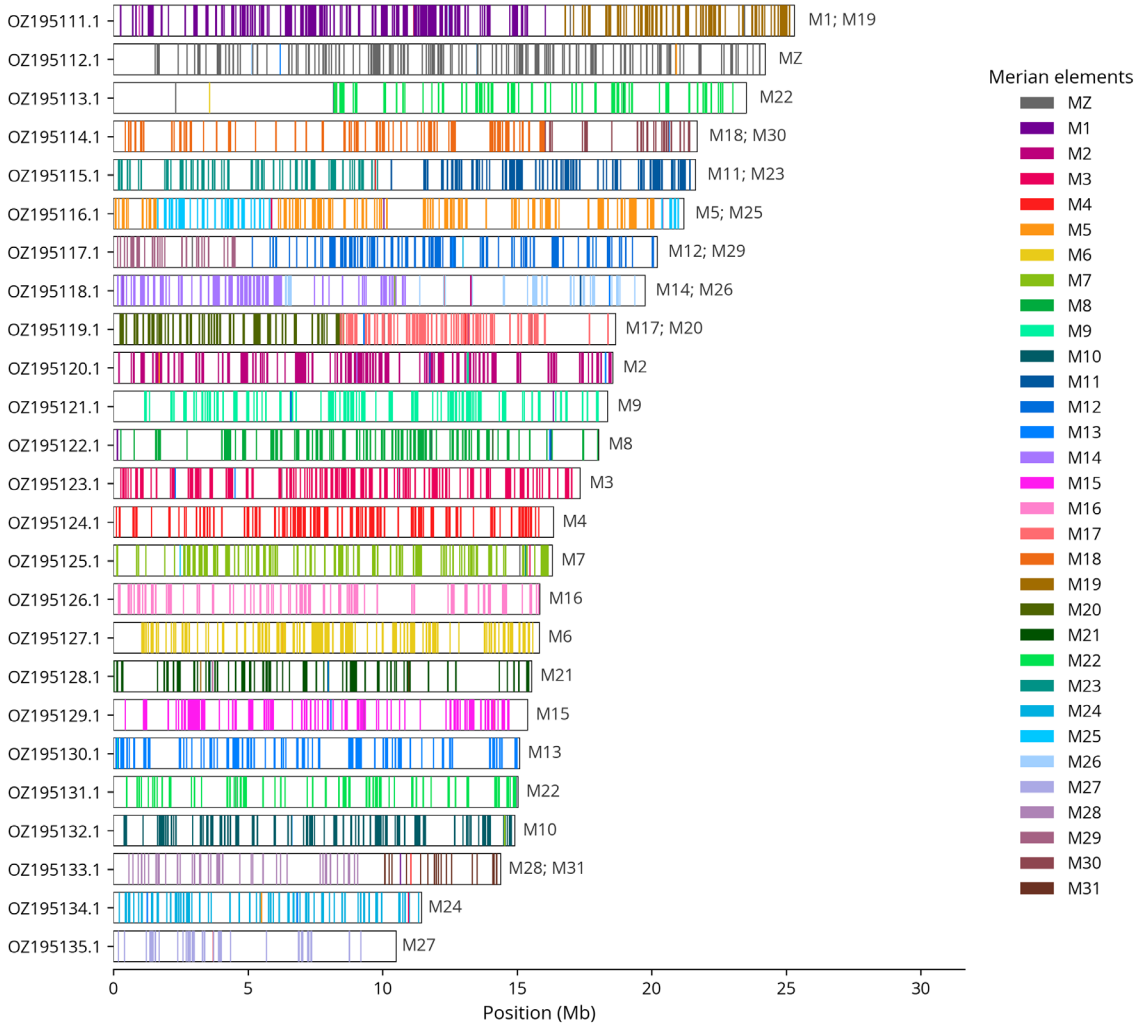

Figure 4. Merian elements painted across chromosomes in the ilCupArgi2.hap1.1 assembly of *Cupido argiades*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

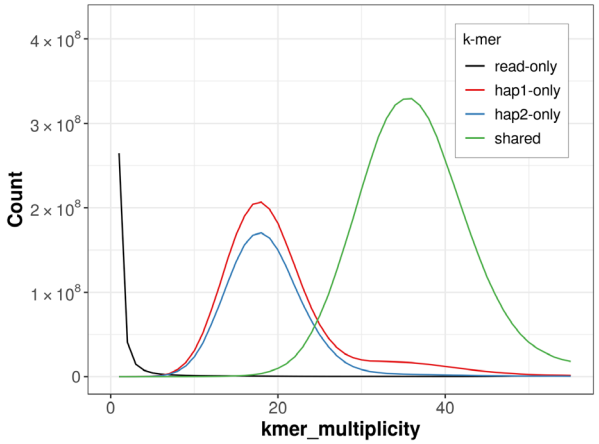

Figure 5. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers 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.

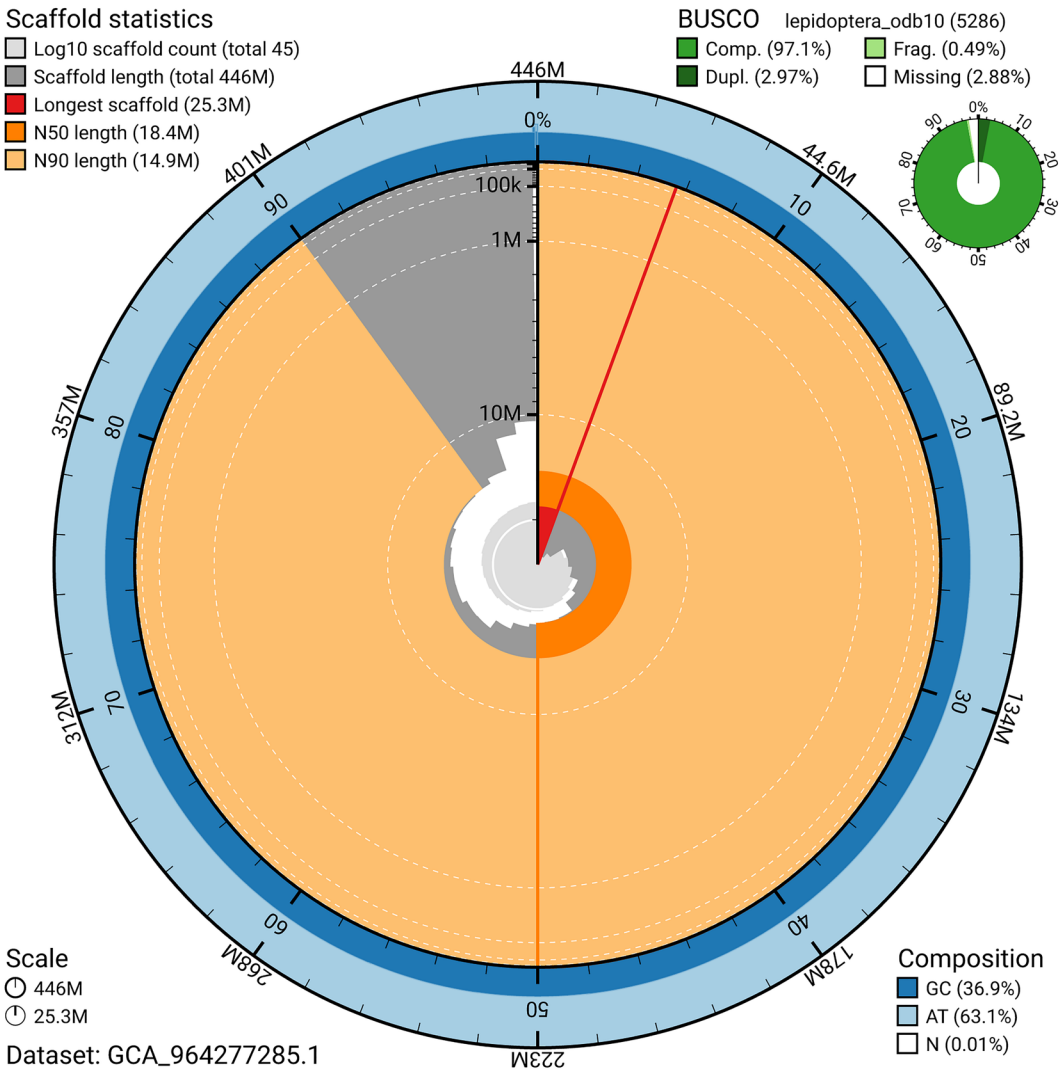

Figure 6. Assembly metrics for ilCupArgi2.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

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).

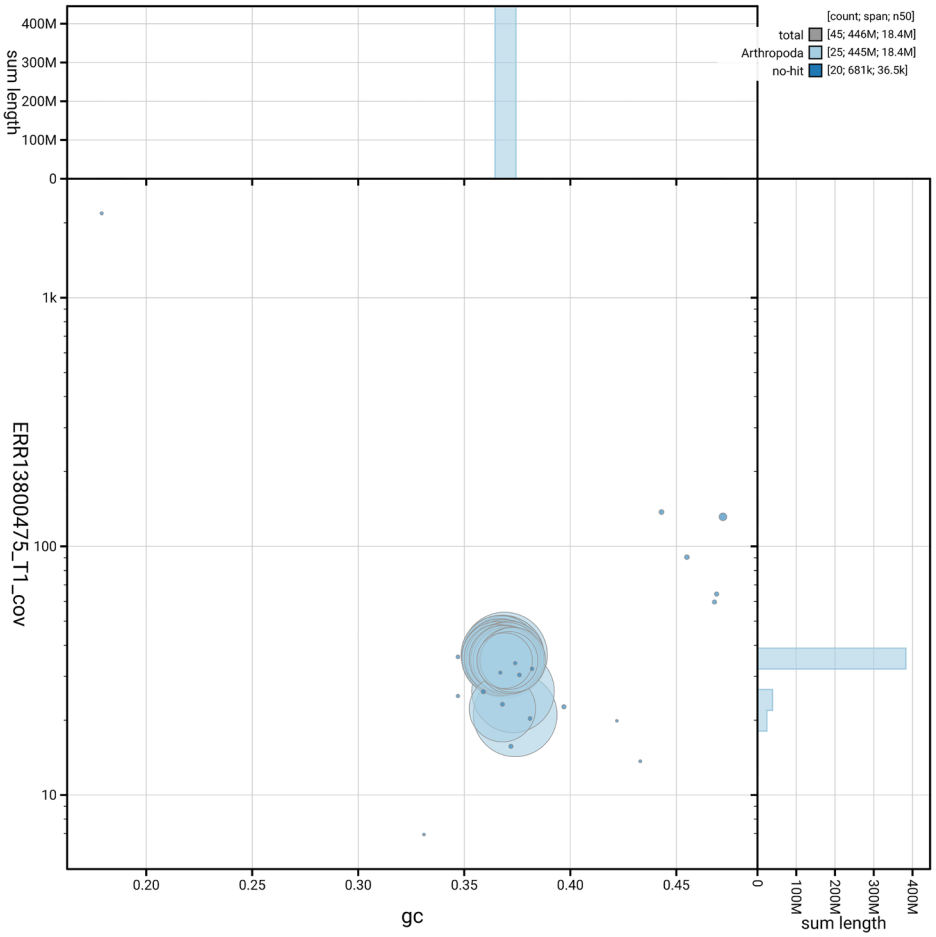

Figure 7. BlobToolKit GC-coverage plot for ilCupArgi2.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Cupido argiades* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q65	6.C.Q40
Contig N50 length	8.09 Mb	≥ 1 Mb
Scaffold N50 length	18.36 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 65.8; haplotype 2: 65.9; combined: 65.8	≥ 40
k-mer completeness	Haplotype 1: 73.28%; Haplotype 2: 65.73%; combined: 99.24%	≥ 95%
BUSCO	C:97.1% [S:94.2%; D:3.0%]; F:0.5%; M:2.4%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.85%	≥ 90%

Notes: EBP summary uses log10(Contig N50); chromosome-level (C) or log10(Scaffold N50); Q (Merqury QV). BUSCO: C=complete; S=single-copy; D=duplicated; F=fragmented; M=missing; n=orthologues

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.

Data availability

European Nucleotide Archive: Cupido argiades. Accession number PRJEB80915. The genome sequence is released openly for reuse. The *Cupido argiades* 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:

- Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team
- Members of Wellcome Sanger Institute Scientific Operations – Sequencing Operations
- Members of the Wellcome Sanger Institute Tree of Life Core Informatics team
- Members of the Tree of Life Core Informatics collective
- Members of the 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
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
lep_busco_painter	1.0.0	https://github.com/charlottewright/lep_busco_painter
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	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	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
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	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

Beneš J, Konvička M, Dvořák J, *et al.*: **Motýli České republiky: Rozšíření a Ochrana/Butterflies of the Czech Republic: distribution and conservation. I., II**. Prague: SOM, 2002.
Reference Source

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, 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

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

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

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

Fiedler K: **The ant associates of Lycaenidae butterfly caterpillars – revisited.** *Nota Lepidopterol.* 2021; **44**: 159–174.
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

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

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

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

Maes D, Verovnik R, Wiemers M, *et al.*: **Integrating national Red Lists for prioritising conservation actions for European butterflies.** *J Insect Conserv.* 2019; **23**(2): 301–330.
Publisher Full Text

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI Taxonomy: a comprehensive update on curation, resources and tools.** *Database (Oxford).* 2020; **2020**: baaa062.
PubMed Abstract | Publisher Full Text | Free Full Text

Sucháčková Bartoňová A, Škopek P, Konvička M, *et al.*: **Czech Republic butterfly barcoding reveals that distribution of genetic lineages depends on species traits.** *J Biogeogr.* 2024; **51**: 1575–1586.
Publisher Full Text

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python**

pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics.* 2023; **24**(1): 288.
PubMed Abstract | **Publisher Full Text** | **Free 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

Wang M, Fan X: **Butterflies Fauna Sinica: Lycaenidae**. Zhengzhou: Henan Science; Technology Publishing House, 2002.

Wright CJ, Stevens L, Mackintosh A, *et al.*: **Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera.** *Nat Ecol Evol.* 2024; **8**(4): 777–790.
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 16 March 2026

https://doi.org/10.21956/wellcomeopenres.28397.r148553

✔ **Océane Seudre**
Queen Mary University of London, London, England, UK

This Data Note presents a chromosome-level genome assembly for the Short-tailed Blue, *Cupido argiades,* generated as part of Project Psyche and the Sanger Tree of Life Programme. The manuscript is clearly organised and reports a strong genome resource, including a chromosome-level primary haplotype, a secondary scaffold-level haplotype, and a mitochondrial assembly, together with the associated raw data and accession information.
The rationale for generating the dataset is clear and appropriate. The species background is sufficient for a genome note of this type, and the manuscript places the resource well within wider comparative and biodiversity genomics efforts in Lepidoptera.
The protocols are appropriate and the work appears technically sound. The sequencing and assembly strategy, combining PacBio HiFi, Hi-C scaffolding/phasing, RNA-seq support, manual curation, and multiple quality assessment metrics, is in line with current best practice. The reported assembly statistics are strong and support the overall quality of the resource.
While I answered “Yes” for whether sufficient methodological detail is provided to allow full replication by others, computational reproducibility could be improved by including a few more practical details. In particular, it would help to report key non-default parameters or decision thresholds for the main assembly and QC steps, and to provide a slightly clearer account of the manual curation decisions, including the basis for breaks and joins. It would also be helpful to briefly clarify how the Hi-C signal was used to identify and assign the W, Z1, and Z2 sex chromosomes.
Overall, these are improvements to clarity and reproducibility rather than concerns about the validity of the dataset itself.

Points to be addressed

- Add key command settings or non-default parameters for the main computational steps, including assembly, scaffolding, decontamination, BUSCO/MerquryFK assessment, and mitochondrial assembly.
- Briefly clarify how the Hi-C signal was used to identify and assign the W, Z1, and Z2 sex chromosomes.
- Provide justification or criteria for the manual curation step described as “Manual

corrections included nine breaks and 46 joins," specifically what signals or evidence were used to support these corrections.
- Optionally, provide direct hyperlinks for the accession identifiers in the tables, so readers can more easily access the BioProject, BioSamples, sequencing runs, and assembly records.

These are minor suggestions intended to improve reproducibility and interpretability. I do not see major issues with the underlying dataset or its presentation.

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: Comparative genomics, including Lepidoptera genome assembly and annotation

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 09 March 2026

https://doi.org/10.21956/wellcomeopenres.28397.r145918

Zsófia Fekete
Ecology and Genetics Research Unit, University of Oulu, Oulu, Finland
Niina Kiljunen
Ecology and Genetics Research Unit, University of Oulu, Oulu, Finland

This article describes the process of producing a high-quality genome assembly for the Short-tailed Blue (*Cupido argiades*), which did not previously have a full genome assembly available.

The taxonomy and background information for the species *Cupido argiades* are presented in detail, providing a sufficient introduction to the species for the reader.

Methods used for Sample acquisition, Nucleic acid extraction and PacBio HiFi library preparation and sequencing, Sample preparation and crosslinking, Hi-C library preparation and sequencing, RNA-seq library preparation and sequencing, Genome assembly, Assembly curation, and Assembly curation assessment follow the standard methodology used for specimen processing in the Project Psyche.

I would suggest some clarification to the interpretation of the Fig. 3.; There seems to be a connection between these two: OZ195131.1 (Z2) and OZ195113.1 (W).

Fig.4. and the chromosome annotation is hard to understand if one is not familiar with Merian elements. I would clarify why OZ195113.1 is assigned as W and OZ195131.1 as Z2 while having the same Merian elements, and only OZ195112.1 (Z1) has MZ elements.

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: DNA-barcoding, DNA-based species delimitation, entomology, gall midges, dark taxa

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 23 February 2026

https://doi.org/10.21956/wellcomeopenres.28397.r146975

Gurinder Kaur Walia
Punjabi University, Patiala, Punjab, India

The manuscript presents a high-quality, chromosome-level genome assembly of *Cupido argiades*, a widely distributed Lycaenid butterfly. The study is part of Project Psyche and utilizes state-of-the-

art sequencing technologies (PacBio HiFi and Hi-C). The assembly achieves an exceptional consensus quality (QV 65.8) and high gene completeness (97.1% BUSCO).The data provided is of significant value to the lepidopterists' community, particularly for those studying chromosomal evolution and molecular phylogeny. The identification of a complex sex-chromosome system (W, Z_1, Z_2) and a reduced karyotype (n=25) are major highlights.

Recommendation: Accepted with Minor Modifications.

1. Haplotype Size Discrepancy:

The authors report a total length of 445.87 Mb for Haplotype 1 and 384.46 Mb for Haplotype 2.This results in a discrepancy of 61.41 Mb (approximately 14%) between the two haplotypes.

Required Correction: While Haplotype 1 was curated to the chromosome level and Haplotype 2 remained at the scaffold level, the authors should clarify if this gap is due to the collapse of repetitive regions in the uncurated haplotype or biological hemizygosity. This is essential for users of the alternate assembly.

2. Chromosomal & Merian Element Labeling (Table 3)

In Table 3, the sex chromosome Z_2 (Molecule OZ195131.1) is assigned Merian element M22.Simultaneously, the W chromosome (Molecule OZ195113.1) is also assigned Merian element M22.

Required Correction: This assignment is likely a typographical error or an oversimplification. Ancestral linkage group MZ is correctly assigned to Z_1, but Z_2 and W usually involve distinct autosomal fusions. The authors should re-verify Figure 4 and correct the labeling in Table 3 to reflect the actual ancestral linkage groups involved in the Z_1Z_2W split.

3. Geographic Context of Genetic Variability

The "Background" section states that the species exhibits no mitochondrial variability across Europe.However, the authors also note that the species distribution extends east into India, China, and South-East Asia.

Required Correction: The claim of "no variability" should be explicitly limited to "European populations". This ensures that researchers working on Asian subspecies (which may harbor significant genetic divergence) are not misled by a claim based solely on European sampling.

4. Mitochondrial-Nuclear Analysis (NUMTs)

The assembly provides both a complete 15.41 kb mitochondrial genome and a chromosomal nuclear genome.

Required Correction: Given that *C. argiades* is a common subject for DNA barcoding, a brief mention of whether significant NUMTs (Nuclear Mitochondrial DNAs) were identified would be of high value. This prevents future taxonomic errors where nuclear pseudogenes might be mistaken for the mitochondrial COI gene.

5. Technical Clarification in Table 4

Table 4 shows k-mer completeness for individual haplotypes as 73.28% and 65.73%, while the combined completeness is 99.24%.

Required Correction: A brief note should be added to explain that these lower single-haplotype values are a result of the high heterozygosity (1.72%), which partitions k-mers into haplotype-specific peaks rather than a single homozygous peak.

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: Cytogenetics and Molecular studies of Odonata and Lepidoptera

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.