DATA NOTE

The genome sequence of the Osiris Blue, *Cupido osiris* (Meigen, 1830) (Lepidoptera: Lycaenidae)

[version 1; peer review: 2 approved]

Andreas Sanchez[1], Yannick Chittaro[1], 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 Programme, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 23 Jan 2026, **11**:71
https://doi.org/10.12688/wellcomeopenres.25821.1
Latest published: 23 Jan 2026, **11**:71
https://doi.org/10.12688/wellcomeopenres.25821.1

Abstract

We present a genome assembly from a male *Cupido osiris* (Osiris Blue; Arthropoda; Insecta; Lepidoptera; Lycaenidae). The assembly contains two haplotypes with total lengths of 481.43 megabases and 479.83 megabases. Most of haplotype 1 (98.72%) is scaffolded into 24 chromosomal pseudomolecules, including the Z sex chromosome. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 15.36 kilobases. This work is part of Project Psyche, a collaborative programme generating genomes for European butterflies and moths.

Keywords

Cupido osiris; Osiris Blue; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 23 Jan 2026	✓ view	✓ view

1. **Sophie Tandonnet**, Universitat de Barcelona, Barcelona, Spain
María Morán Saiz, Universitat de Barcelona Departament de Genetica Microbiologia i Estadistica, Barcelona, Spain

2. **Sivasankaran Kuppusamy**, Loyola College, chennai, India

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; **Lucek K**: Supervision, Writing – Review & Editing; **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). KL received funding from 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.

How to cite this article: Sanchez A, Chittaro Y, Lucek K *et al.* **The genome sequence of the Osiris Blue, *Cupido osiris* (Meigen, 1830) (Lepidoptera: Lycaenidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2026, **11**:71 https://doi.org/10.12688/wellcomeopenres.25821.1

First published: 23 Jan 2026, **11**:71 https://doi.org/10.12688/wellcomeopenres.25821.1

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 osiris* (Meigen, 1830) (NCBI:txid282382)

Background

The Osiris blue – *Cupido osiris* (Meigen, 1829) – is widespread in Europe and Central Asia (Tolman & Lewington, 1997), ranging from the Iberian Peninsula to eastern Mongolia, including the southern part of France, Italy, the west and south of the Balkan Peninsula, Turkey, Kyrgyzstan and southern Russia. However, its distribution is highly fragmented, with large occurrence gaps, particularly in central Europe and the central Balkan Peninsula (Kudrna *et al.*, 2015; Tshikolovets, 2011).

The species shows strong sexual dimorphism, with males having a dark violet-blue upper surface, while females are brown. The latter is very similar to *Cupido minimus* (Füssli, 1775), *C. lorquinii* (Herrich-Schäffer, 1850) and *Cyaniris semiargus* (Rottemburg, 1775), and identification is not always easy. The underside of the wings is light grey with a large suffusion of blue at the base. It is adorned with small black eye spots.

The species is generally univoltine and adults are mainly seen from May to July, depending on altitude. A partial second generation is reported in the south of its range and at low altitudes. The species is found from the plains to altitudes above 2 000 m in the European Alps (Ligue Suisse pour la Protection de la Nature, 1987). It colonises sunny mesophilic calcareous grasslands (Van Swaay, 2002) and rocky roadsides, where sainfoins (*Onobrychis* spp.), its main host plants, grow in abundance. The occasional use of bladder senna (*Colutea arborescens*) is also reported (Clarke, 2024; Lafranchis *et al.*, 2015). Other suggested host plants (Tshikolovets, 2011; Verity, 1943), including common kidney vetch (*Anthyllis vulneraria*) and pea vine (*Lathyrus* spp.), do not seem to be confirmed.

The eggs are laid individually on the inflorescences of the host plants. The caterpillars feed on the flowers and fruits. Being facultative myrmecophiles, they are often accompanied by ants of the genera *Formica*, *Camponotus*, *Lasius* and *Plagiolepis* (Lafranchis *et al.*, 2015). Caterpillars spend the late summer, autumn and winter on the ground or on their host plant and pupate only in spring, usually among dead leaves on the ground. Adults can easily be seen nectaring on flowering sainfoin.

Although the species is not globally threatened (LC according to Van Swaay *et al.*, 2010) and may even be locally common, especially in the mountains, it has become very rare in its northern distribution range and is endangered in several countries, including Switzerland (Wermeille *et al.*, 2014), Austria (Höttinger & Pennerstorfer, 2005) or is already extinct as in Germany (Reinhardt & Bolz, 2011). The occupied habitats are often small and locally threatened by urbanisation and intensification of agricultural practices, including overfertilisation, overgrazing and excessive mowing. In the south, the abandonment of grasslands threatens the availability of its host plant and, consequently, the butterfly. Climate change represents an additional threat (Obregón *et al.*, 2016).

The reference genome for *Cupido osiris* provides an invaluable resource for studying the population structure of this species and for shedding light on the evolution of colour polymorphism in this group (Hinojosa *et al.*, 2020).

Methods

Sample acquisition

The specimen used for genome sequencing was an adult male *Cupido osiris* (specimen ID SAN28000112, ToLID ilCupOsir1; Figure 1), collected from Ormont-Dessus, Switzerland (latitude 46.3553, longitude 7.204; elevation 1 559 m) on 2023-06-28. The specimen was collected and identified by Andreas Sanchez (Info fauna).

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 ilCupOsir1 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

Figure 1. Voucher photograph of the *Cupido osiris* (ilCupOsir1) specimen used for genome sequencing.

FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 32.71 ng/μL and a yield of 1 537.37 ng, with a fragment size of 13.9 kb.

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 abdomen; head of the ilCupOsir1 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.

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 PRJEB82258.

Platform	PacBio HiFi	Hi-C
ToLID	ilCupOsir1	ilCupOsir1
Specimen ID	SAN28000112	SAN28000112
BioSample (source individual)	SAMEA115109942	SAMEA115109942
BioSample (tissue)	SAMEA115109978	SAMEA115109979; SAMEA115109977
Tissue	thorax	abdomen; head
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13946490	ERR13947518; ERR13947517
Read count total	2.82 million	1 259.01 million
Base count total	27.78 Gb	190.11 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 four breaks and ten joins. This reduced the scaffold count by 9.4%. 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 painted Merian elements, which represent the 32 ancestral linkage groups in Lepidoptera (Wright *et al.*, 2024), along chromosomes using lep_busco_painter. Painting used BUSCO gene locations from the lepidoptera_odb10 set (Kriventseva *et al.*, 2019) and chromosome lengths from NCBI Datasets. Each complete BUSCO (including both single-copy and duplicated BUSCOs) was assigned to a Merian element using a reference database. Positions were coloured and plotted along chromosomes drawn to scale.

Genome sequence report

Sequence data

PacBio sequencing of the *Cupido osiris* specimen generated 27.78 Gb (gigabases) from 2.82 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 473.78 Mb, with a heterozygosity of 1.27% and repeat content of 39.23% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 57× coverage. Hi-C sequencing produced 190.11 Gb from 1 259.01 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 481.43 Mb in 260 scaffolds, with 98 gaps, and a scaffold N50 of 20.3 Mb (Table 2).

Most of the assembly sequence (98.72%) was assigned to 24 chromosomal-level scaffolds, representing 23 autosomes and the Z sex chromosome. 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). Z chromosome identified based BUSCO gene painting with ancestral Merian elements (Wright *et al.*, 2024).

The mitochondrial genome was also assembled (length 15.36 kb, OZ218356.1). 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 59.3, and for haplotype 2, 59.1. When the two haplotypes are combined, the assembly achieves an estimated QV of 59.2. The *k*-mer completeness is 75.72% for haplotype 1, 75.66% for haplotype 2, and 99.53% for the combined haplotypes (Figure 5). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 97.4% of the expected gene set (single = 96.5%, duplicated = 0.9%) in haplotype 1. For haplotype 2, BUSCO analysis identified 97.6% of the expected gene set (single = 97.0%, duplicated = 0.5%).

The snail plot in Figure 6 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The

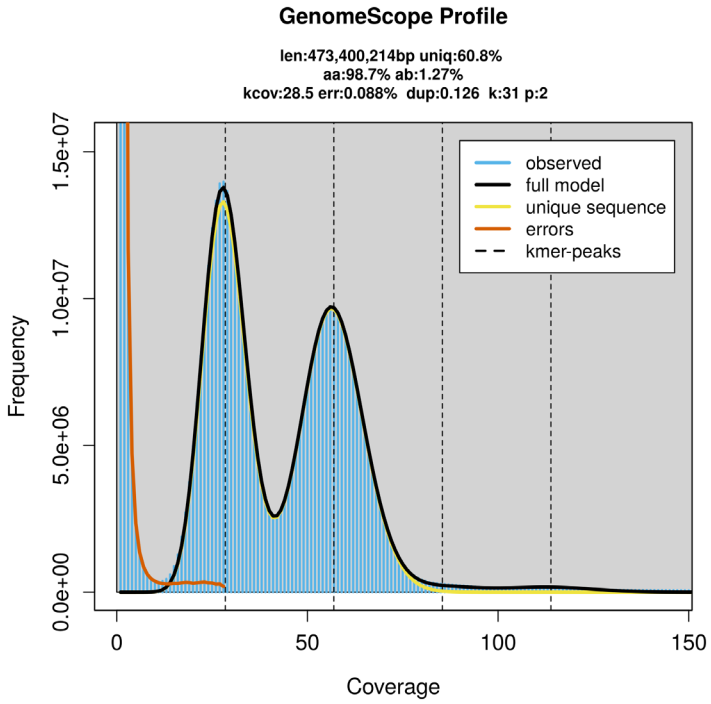

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 2. Genome assembly statistics.

Assembly name	ilCupOsir1.hap1.1	ilCupOsir1.hap2.1
Assembly accession	GCA_965112425.1	GCA_965112415.1
Assembly level	chromosome	scaffold
Span (Mb)	481.43	479.83
Number of chromosomes	24	scaffold-level
Number of contigs	358	250
Contig N50	7.22 Mb	7.02 Mb
Number of scaffolds	260	167
Scaffold N50	20.3 Mb	20.27 Mb
Longest scaffold length (Mb)	28.79	-
Sex chromosomes	Z	-
Organelles	Mitochondrion: 15.36 kb	-

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) and the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.Q59**, 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

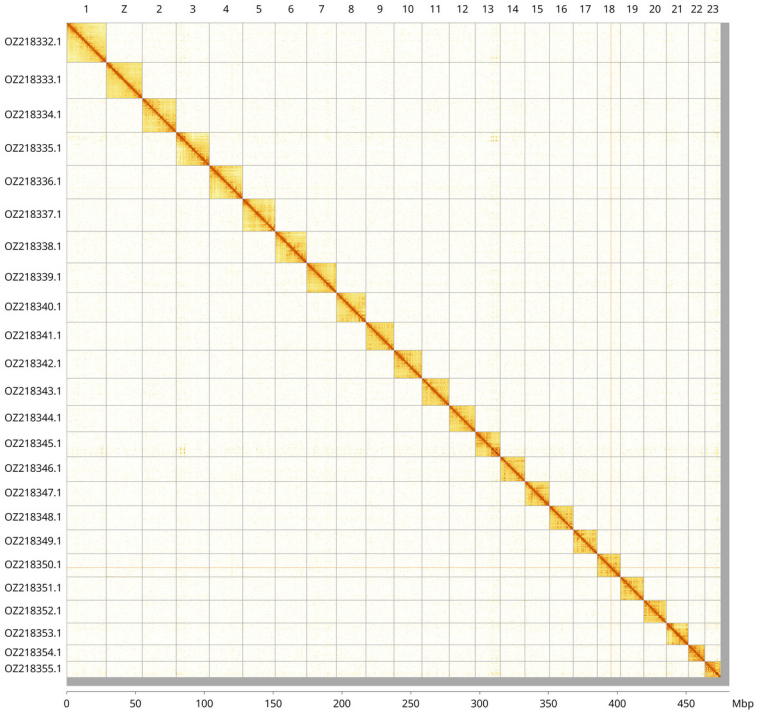

Figure 3. Hi-C contact map of the *Cupido osiris* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Cupido osiris* ilCupOsir1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ218332.1	1	28.79	37	M1;M19
OZ218334.1	2	24.55	37	M11;M23
OZ218335.1	3	24.22	37	M18;M30
OZ218336.1	4	24.05	37	M25;M5
OZ218337.1	5	23.69	37	M12;M29
OZ218338.1	6	22.83	37	M14;M26
OZ218339.1	7	21.61	36.50	M9
OZ218340.1	8	21.44	37	M17;M20
OZ218341.1	9	20.40	37	M8
OZ218342.1	10	20.30	37.50	M2
OZ218343.1	11	19.83	37.50	M3
OZ218344.1	12	18.96	37	M7
OZ218345.1	13	18.18	37	M13
OZ218346.1	14	17.89	37	M6
OZ218347.1	15	17.69	37	M16
OZ218348.1	16	17.48	37	M4

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ218349.1	17	17.37	37	M22
OZ218350.1	18	16.90	37.50	M10
OZ218351.1	19	16.80	37	M21
OZ218352.1	20	16.52	37.50	M15
OZ218353.1	21	15.92	37.50	M28;M31
OZ218354.1	22	12.04	37.50	M24
OZ218355.1	23	11.57	37.50	M27
OZ218333.1	Z	26.23	36.50	MZ

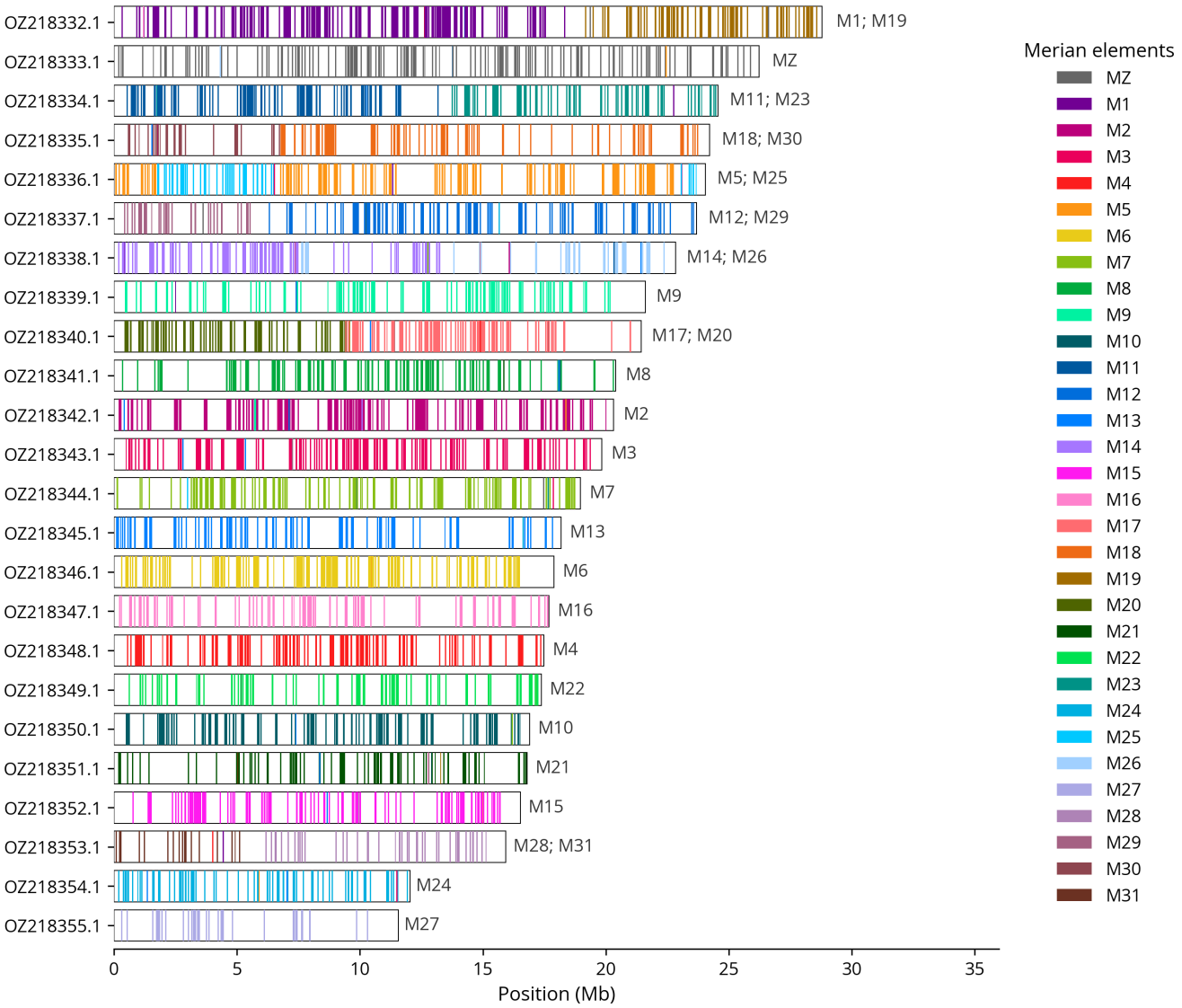

Figure 4. Merian elements painted across chromosomes in the ilCupOsir1.hap1.1 assembly of *Cupido osiris*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

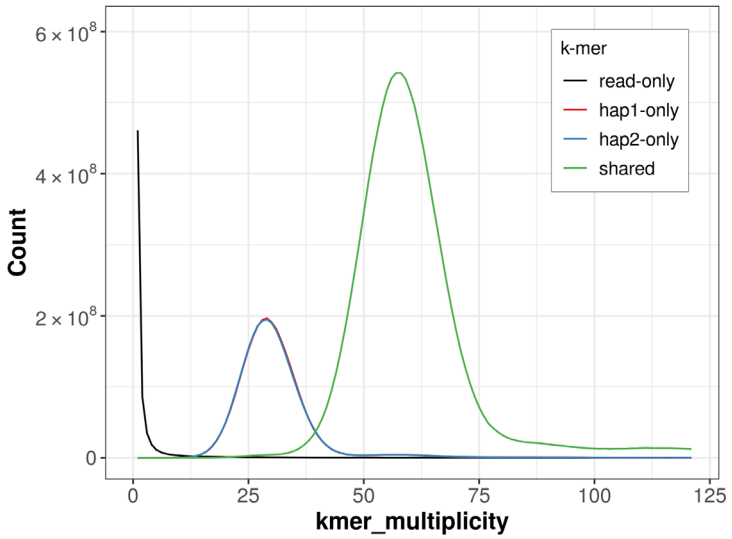

Figure 5. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

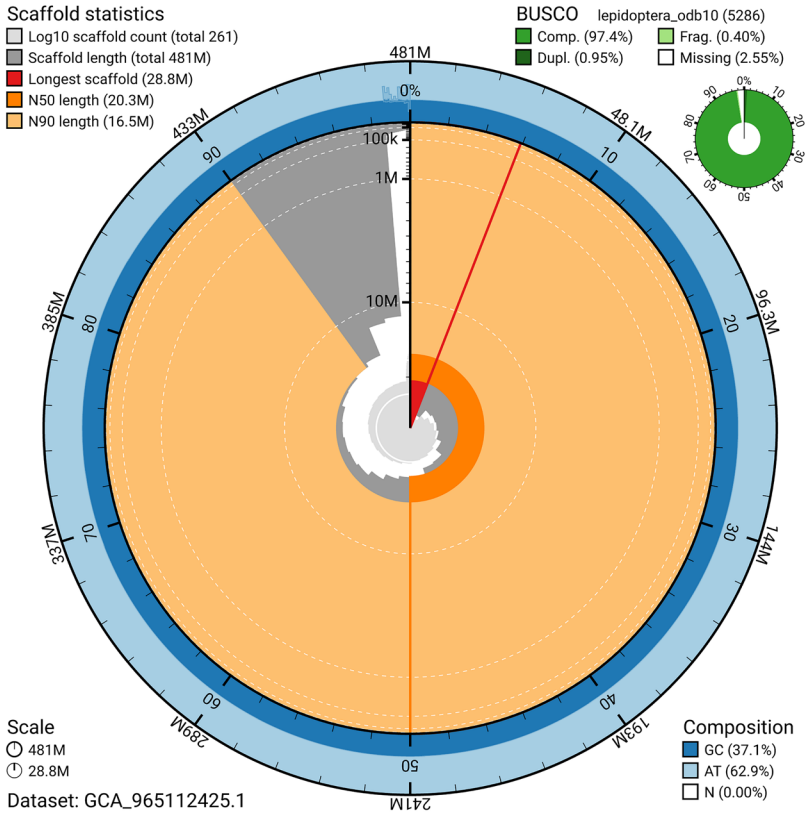

Figure 6. Assembly metrics for ilCupOsir1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

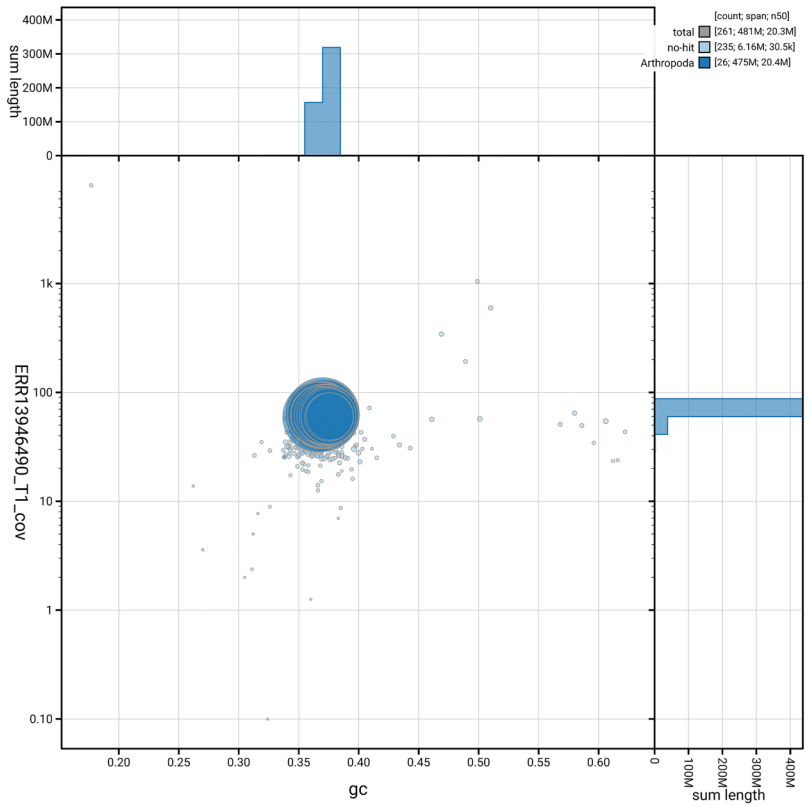

Figure 7. BlobToolKit GC-coverage plot for ilCupOsir1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 osiris* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q59	6.C.Q40
Contig N50 length	7.22 Mb	≥ 1 Mb
Scaffold N50 length	20.30 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 59.3; haplotype 2: 59.1; combined: 59.2	≥ 40
k-mer completeness	Haplotype 1: 75.72%; Haplotype 2: 75.66%; combined: 99.53%	≥ 95%
BUSCO	C:97.4% [S:96.5%; D:0.9%]; F:0.4%; M:2.2%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	98.72%	≥ 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

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.

Data availability

European Nucleotide Archive: Cupido osiris (Osiris blue). Accession number PRJEB82258. The genome sequence is released openly for reuse. The *Cupido osiris* 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

Software	Version	Source
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
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): 403–10.
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

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

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

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

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

Hinojosa JC, Koubínová D, Dincă V, *et al.*: **Rapid colour shift by reproductive character displacement in *Cupido* butterflies.** *Mol Ecol.* 2020; **29**(24): 4942–55.
PubMed Abstract | Publisher Full Text

Höttinger H, Pennerstorfer J: **Rote liste der Tagschmetterlinge Österreichs (Lepidoptera: Papilionoidea & Hesperioidea)**. In: K. P. Zulka (ed), *Rote Listen Gefährdeter Tiere Österreichs. Checklisten, Gefährdungsanalysen, Handlungsbedarf. Teil 1: Säugetiere, Vögel, Heuschrecken, Wasserkäfer, Netzflügler, Schnabelfliegen, Tagfalter, Vols. Band 14/1.* Wien: Böhlau, 2005; 313–54.

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, Pennerstorfer J, Lux K: **Distribution atlas of European butterflies and skippers**. Wissenschaftlicher Verlag PEKS, 2015.
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

Ligue Suisse pour la Protection de la Nature: **Les papillons de jour et leurs biotopes. Espèces, dangers qui les menacent, protection. Volume 1**. Fotorotar AG, 1987.
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

Obregón R, Fernández Haeger J, Jordano D: **Effects of climate change on three species of *Cupido* (Lepidoptera, Lycaenidae) with different biogeographic distribution in Andalusia, southern Spain.** *Anim Biodivers Conserv.* 2016; **39**(1): 115–28.
Publisher Full Text

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**

Reinhardt R, Bolz R: **Rote Liste und Gesamtartenliste der Tagfalter (Rhopalocera) (Lepidoptera: Papilionoidea et Hesperioidea) Deutschlands**. In M. Binot-Hafke, S. Balzer, N. Becker, *et al*. (eds), *Rote Listen Gefährdeter Tiere, Pflanzen Und Pilze Deutschlands, Band 3: Wirbellose Tiere (Teil 1).* Münster: Landwirtschaftsverlag, 2011; **70**: 167–94.

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**

Tolman T, Lewington R: **Butterflies of Britain and Europe**. Collins, 1997.
Reference Source

Tshikolovets VV: **Butterflies of Europe & the Mediterranean area**. 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.
Reference Source

Van Swaay CAM: **The importance of calcareous grasslands for butterflies in Europe.** *Biol Conserv.* 2002; **104**(3): 315–18.
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

Verity R: **Le farfalle diurne d'italia. 2. Divisione lycaenidae**. Firenze: Marzocco, 1943.

Wermeille E, Chittaro Y, Gonseth Y: **Liste rouge Papillons diurnes et Zygènes. Espèces menacées en Suisse, état 2012**. Office fédéral de l'environnement, Berne; Centre Suisse de Cartographie de la Faune, Neuchâtel, 2014.
Reference Source

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 05 March 2026

https://doi.org/10.21956/wellcomeopenres.28422.r146650

✔ **Sivasankaran Kuppusamy**
Loyola College, chennai, Tamil Nadu, India

The study reports a high-quality chromosome-level genome sequence of *Cupido osiris* (Meigen, 1830). The assembly consists of two haplotypes with lengths of 481.43 Mb and 479.83 Mb, respectively. This research provides important molecular data for Lepidopteran studies and contributes valuable resources for evolutionary, comparative, and conservation genomics.
The use of PacBio HiFi sequencing combined with Hi-C scaffolding ensures high accuracy and strong structural integrity of the assembly. The bioinformatics pipeline appears standard, reproducible, and well-documented.
The application of advanced sequencing and assembly methods is clearly demonstrated by the successful separation into two haplotypes, which increases precision by distinguishing maternal and paternal chromosome sets.
Approximately 98.72% of haplotype 1 was scaffolded into 24 chromosomal pseudomolecules, indicating near chromosome-level completeness. Genome completeness assessment using BUSCO demonstrates high contiguity, strong completeness, and excellent base-level accuracy.
Comments on the Manuscript
A total of 207.89 Gb of sequencing data were generated from two different platforms. However, following assembly and annotation, the manuscript does not report the number of predicted protein-coding genes, non-coding genes, or gene transcripts. These important annotation statistics should be clearly presented in the main text and summarized in a table for clarity.
Overall Comments
The manuscript is well prepared and suitable for scientific indexing.

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: Phylogenetic analysis of superfamily Noctuoidea moths using mitogenome sequences

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 19 February 2026

https://doi.org/10.21956/wellcomeopenres.28422.r146031

Sophie Tandonnet
Universitat de Barcelona, Barcelona, Spain
María Morán Saiz
Universitat de Barcelona Departament de Genetica Microbiologia i Estadistica, Barcelona, Catalonia, Spain

In this genome note, the authors report the chromosome-scale assembly for the genome of Osiris Blue butter fly (*Cupido osiris*). The assembly originates from a single male (the homogametic sex) was assembled in two haplotypes. Haplotype 1 was curated to chromosome level and contains 24 pseudo-chromosomal molecules (including the Z chromosome). This chromosome number differs from the usual 31 chromosomes found in Lepidoptera. Looking at the Merian element pattern, some recent fusions are quite clear. This haplotype is of high quality (QV 59.3), completeness (97.4% complete BUSCOs) and highly continuous (98.72% of the genome in chromosome-scale scaffolds). The assembly is based on Pacbio HiFi long reads (thorax used) and scaffolded into pseudo-chromosomal molecules using Hi-C data (head and abdomen used). The same individual (ilCupOsir1) was used for the Pacbio HiFi sequencing and the Hi-C. The assembly doesn't have an annotation yet. The paper is straightforward and easy to follow.
All in all, this note reports a new, high-quality and high-continuity genome assembly of a species not yet sequenced.

Main Comments

- It is unfortunate that the homogametic sex was sequenced instead of the heterogametic sex. The W chromosome is therefore missing and no coverage analysis can be performed to identify the Z chromosome. This severely limits studies focusing on sex determination/differentiation, sex chromosome evolution, general comparative genomics

analyses and studies aiming to understand the sexual dimorphism in this species. Although Lepidoptera's macrosynteny is very conserved (we particularly like the inclusion of the Merian element painting), coverage analysis to uncover the Z chromosome should be preferred and used in conjunction with homology techniques to identify sex chromosomes. Just homology cannot identify sex chromosomes that have newly evolved (appearance of neo-W and neo-Z).

- No annotation or RNA-seq data is available for this species which also limits its usage as a resource. Nevertheless, this could be done at a later stage.
- The link to the data is correct and the data available for download.
- The background on *Cupido osiris* was nice to read.
- The picture of the specimen is lacking of the parts that were used for the sequencing.

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: Evolutionary Biology, sex determination, genetics, genomics

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.