Check for updates

DATA NOTE

The genome sequence of the Provençal Short-tailed Blue, *Cupido alcetas* (Hoffmannsegg, 1804) (Lepidoptera: Lycaenidae)

[version 1; peer review: 3 approved]

Kay Lucek[1], Zdeněk F. Fric[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]University of Neuchâtel, Neuchâtel, Switzerland
[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**:63
https://doi.org/10.12688/wellcomeopenres.25791.1
Latest published: 23 Jan 2026, **11**:63
https://doi.org/10.12688/wellcomeopenres.25791.1

Abstract

We present a genome assembly from an individual female *Cupido alcetas* (Provençal Short-tailed Blue; Arthropoda; Insecta; Lepidoptera; Lycaenidae). The assembly contains two haplotypes with total lengths of 493.02 megabases and 410.99 megabases. Most of haplotype 1 (99.32%) is scaffolded into 26 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 15.31 kilobases. This work is part of Project Psyche, a collaborative programme generating genomes for European butterflies and moths.

Keywords

Cupido alcetas; Provençal Short-tailed Blue; genome sequence; chromosomal; Lepidoptera

Open Peer Review

Approval Status ✓ ✓ ✓

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

1. **Melody Clark**, Natural Environment Research Council, Cambridge, UK
2. **Olli-Pekka Smolander**, University of Helsinki, Helsinki, Finland
3. **Oliver Rupp**, Justus-Liebig-University, Giessengiessen, Germany

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Tree of Life gateway.

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

Author roles: Lucek K: 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). KL was funded by Swiss National Science Foundation (SNSF) Grant ID 202869.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Lucek K *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Lucek K, Fric ZF, Wright CJ *et al.* **The genome sequence of the Provençal Short-tailed Blue, *Cupido alcetas* (Hoffmannsegg, 1804) (Lepidoptera: Lycaenidae) [version 1; peer review: 3 approved]** Wellcome Open Research 2026, **11**:63 https://doi.org/10.12688/wellcomeopenres.25791.1

First published: 23 Jan 2026, **11**:63 https://doi.org/10.12688/wellcomeopenres.25791.1

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 alcetas* (Hoffmannsegg, 1804) (NCBI:txid876058)

Background

The Provençal Short-tailed Blue – *Cupido alcetas* (Hoffmannsegg, 1804) – is a small butterfly with short tails on the hindwings, very similar to *C. argiades* and *C. decoloratus*. It has a single small orange spot near the tail on the underside of the hindwing, distinguishing it from the two similar species. The wing undersides are otherwise whitish with black spots. The wing uppersides are blue (males) or brownish (females). (SBN, 1987)

Cupido alcetas is known from open forests and forest margins, including substantial parts of the Ponto-Mediterranean region (SBN, 1987). It is rather common in south-western and southern Europe, it is very rare in central Europe (Austria, Czechia) and only occasionally found in northern Europe (Finland, Russia) (Beneš *et al.*, 2002). However, many records, especially from the eastern part of the range, have a low reliability, likely representing the expanding species *Cupido decoloratus*. The larval host plants are from the family Fabaceae: *Coronilla* and *Vicia*. The association with ants is weak, with two ant genera recorded: *Formica* and *Tapinoma* (Fiedler, 2021). The butterfly usually has two generations per year (Beneš *et al.*, 2002; SBN, 1987). The species is threatened in several European countries, near extinction in Spain, endangered in Romania and vulnerable in Bosnia and Herzegovina, Czechia, Slovakia, and Ukraine (Maes *et al.*, 2019). In Europe, the species forms two mitochondrial clusters, western and eastern, however, the related species *C. decoloratus* shares the western mitochondrial haplotypes (Sucháčková Bartoňová *et al.*, 2024).

We present a chromosome-level genome sequence for *C. alcetas*, sequenced as part of Project Psyche. The sequence data were derived from a female specimen (Figure 1) collected from Gadmen, Innertkirchen, Switzerland.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult female *Cupido alcetas* (specimen ID SAN28000069, ToLID ilCupAlce1; Figure 1), collected from Gadmen, Innertkirchen, Switzerland (latitude 46.7361, longitude 8.3572) on 2023-06-14. The specimen was collected and identified by Kay Lucek (University of Neuchâtel).

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 ilCupAlce1 sample was weighed and triaged

Figure 1. Voucher photograph of the *Cupido alcetas* (ilCupAlce1) specimen used for genome sequencing.

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 22.68 ng/μL and a yield of 1 065.96 ng, with a fragment size of 14.6 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 head; abdomen of the ilCupAlce1 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) 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.

Table 1. Specimen and sequencing data for BioProject PRJEB82256.

Platform	PacBio HiFi	Hi-C
ToLID	ilCupAlce1	ilCupAlce1
Specimen ID	SAN28000069	SAN28000069
BioSample (source individual)	SAMEA115109747	SAMEA115109747
BioSample (tissue)	SAMEA115109794	SAMEA115109795; SAMEA115109793
Tissue	thorax	head; abdomen
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13946489	ERR13947515; ERR13947516
Read count total	1.74 million	1 433.91 million
Base count total	17.92 Gb	216.52 Gb

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 11 breaks and 50 joins. This reduced the scaffold count by 19.4%, increased scaffold N50 by 0.9%, and increased the total assembly length by 4.3%. 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 alcetas* specimen generated 17.92 Gb (gigabases) from 1.74 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 447.93 Mb, with a heterozygosity of 0.73% and repeat content of 38.61% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 39× coverage. Hi-C sequencing produced 216.52 Gb from 1 433.91 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

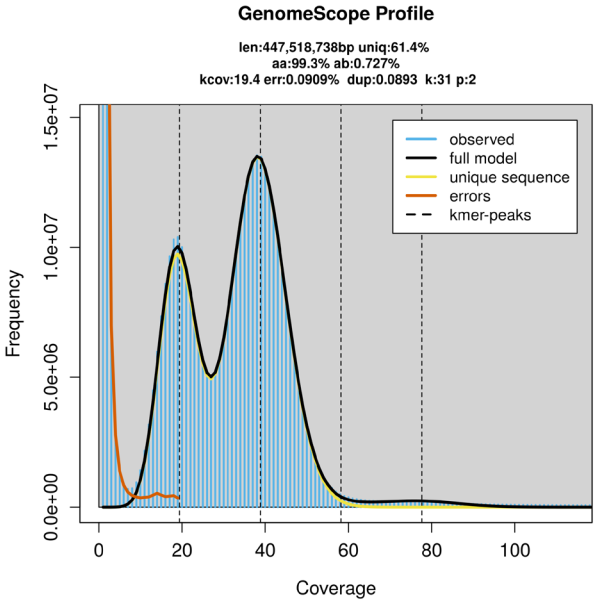

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

haplotype 2 was assembled to scaffold level. The final assembly has a total length of 493.02 Mb in 115 scaffolds, with 155 gaps, and a scaffold N50 of 19.06 Mb (Table 2).

Most of the assembly sequence (99.32%) was assigned to 26 chromosomal-level scaffolds, representing 23 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). Chromosome Z_1 was identified based on the presence of Merian element MZ.

The mitochondrial genome was also assembled (length 15.31 kb, OZ210016.1). 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 60.5, and for haplotype 2, 61.6. When the two haplotypes are combined, the assembly achieves an estimated QV of 61.0. The *k*-mer completeness is 86.16% for haplotype 1, 77.41% for haplotype 2, and 99.35% for the combined haplotypes (Figure 5). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 97.3% of the expected gene set (single = 93.2%, duplicated = 4.1%) in haplotype 1. For haplotype 2, BUSCO analysis identified 89.4% of the expected gene set (single = 89.0%, duplicated = 0.4%). The snail plot in 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.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.

Table 2. Genome assembly statistics.

Assembly name	ilCupAlce1.hap1.1	ilCupAlce1.hap2.1
Assembly accession	GCA_964656375.1	GCA_964656505.1
Assembly level	chromosome	scaffold
Span (Mb)	493.02	410.99
Number of chromosomes	26	scaffold-level
Number of contigs	270	191
Contig N50	5.91 Mb	5.67 Mb
Number of scaffolds	115	106
Scaffold N50	19.06 Mb	18.64 Mb
Longest scaffold length (Mb)	37.2	-
Sex chromosomes	W, Z_1, and Z_2	-
Organelles	Mitochondrion: 15.31 kb	-

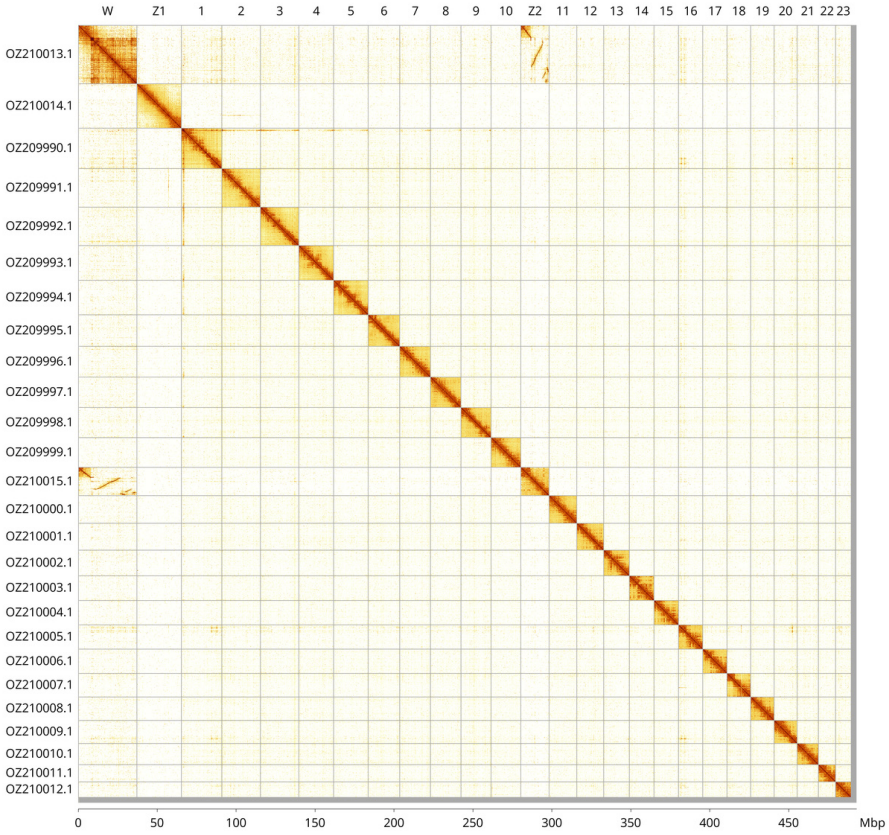

Figure 3. Hi-C contact map of the *Cupido alcetas* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Cupido alcetas* ilCupAlce1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ209990.1	1	25.54	36.50	M18;M30
OZ209991.1	2	24.54	37	M11;M23
OZ209992.1	3	24.36	36.50	M25;M5
OZ209993.1	4	21.94	36.50	M14;M26
OZ209994.1	5	21.91	36.50	M12;M29
OZ209995.1	6	20.02	37	M17;M20
OZ209996.1	7	19.42	37	M2
OZ209997.1	8	19.33	36.50	M8
OZ209998.1	9	19.06	37	M3
OZ209999.1	10	18.90	36.50	M9
OZ210000.1	11	17.56	37	M7
OZ210001.1	12	17.04	37	M4
OZ210002.1	13	16.25	36.50	M16
OZ210003.1	14	15.53	37	M22
OZ210004.1	15	15.50	37	M21
OZ210005.1	16	15.45	37	M13
OZ210006.1	17	15.30	37	M15
OZ210007.1	18	14.97	37	M28;M31
OZ210008.1	19	14.86	37	M10
OZ210009.1	20	14.63	37	M1;M19
OZ210010.1	21	13.46	37	M1
OZ210011.1	22	10.88	37	M24
OZ210012.1	23	9.88	37.50	M27
OZ210013.1	W	37.20	38	M6
OZ210014.1	Z1	28.19	37	MZ
OZ210015.1	Z2	17.93	37.50	M6

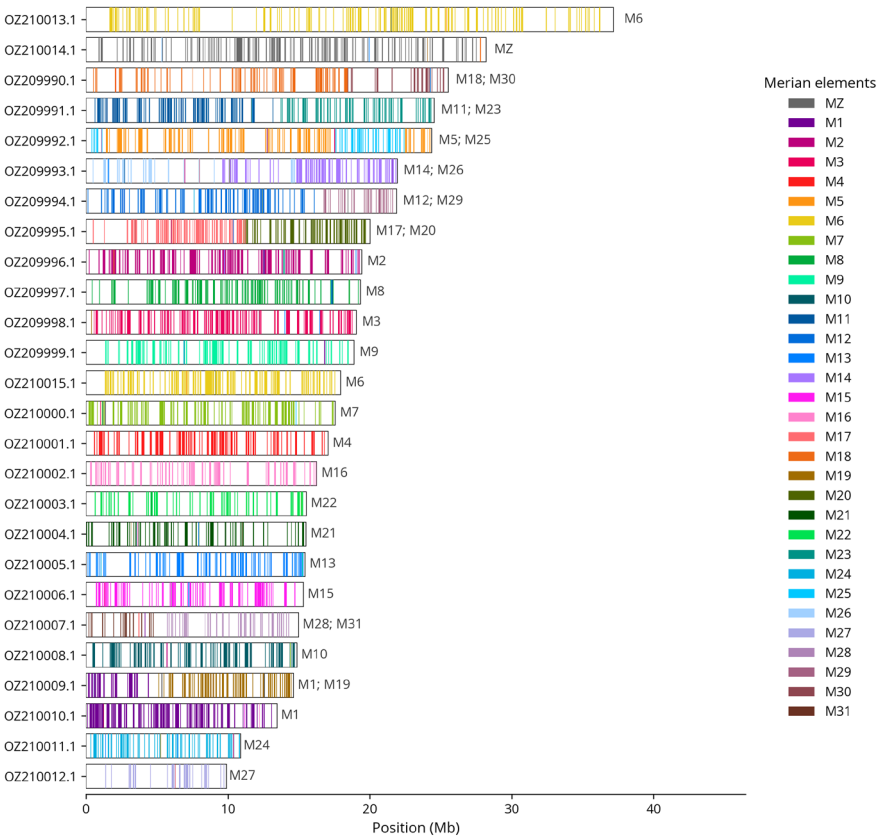

Figure 4. Merian elements painted across chromosomes in the ilCupAlce1.hap1.1 assembly of *Cupido alcetas*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

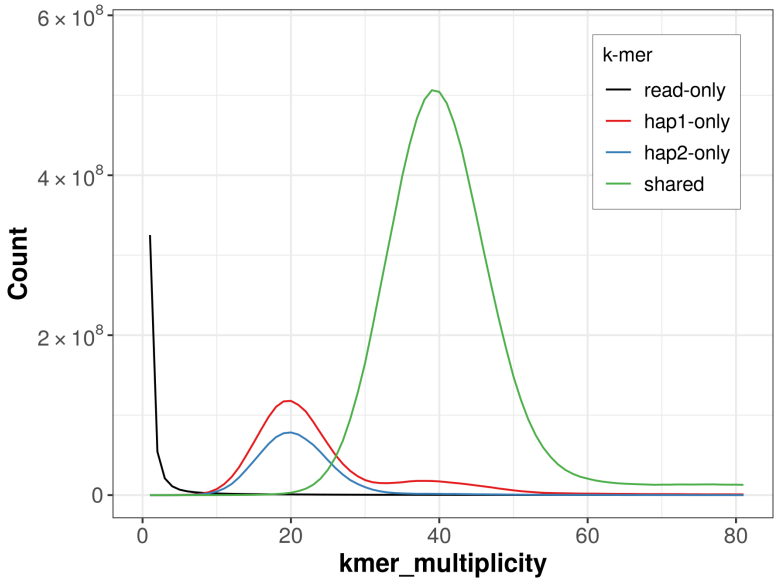

Figure 5. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

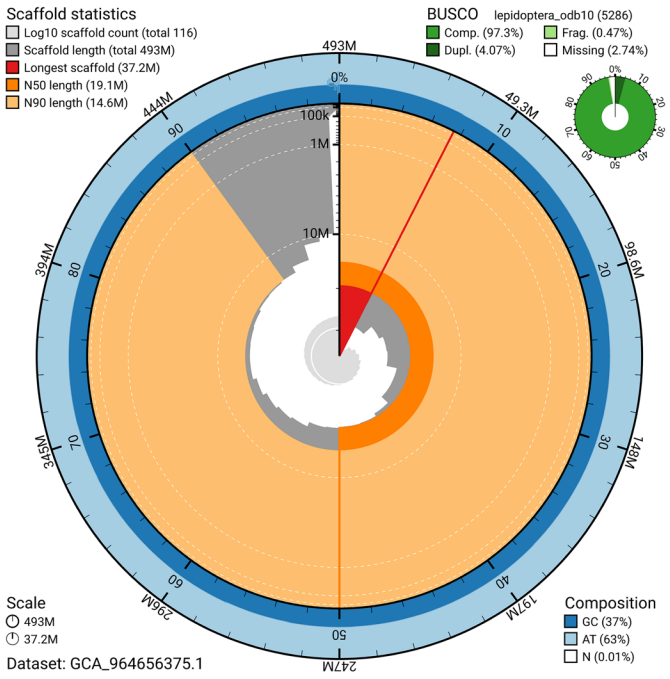

Figure 6. Assembly metrics for ilCupAlce1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

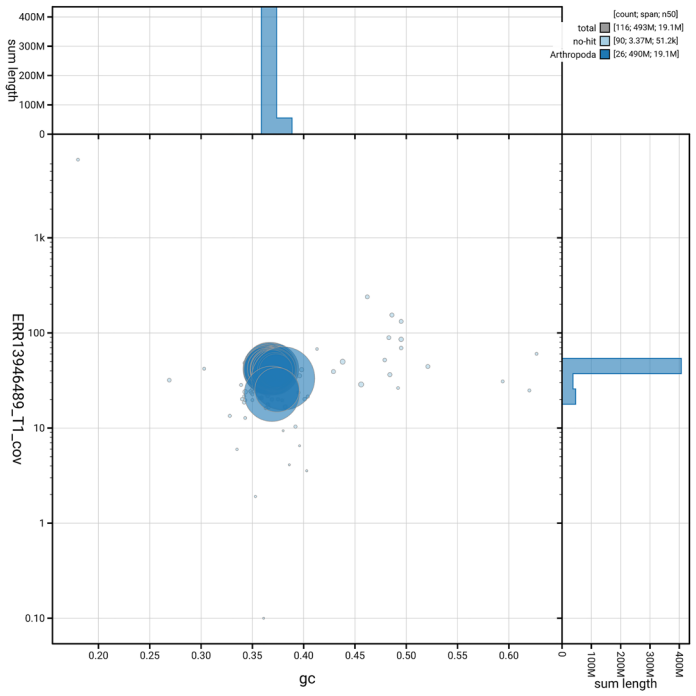

Figure 7. BlobToolKit GC-coverage plot for ilCupAlce1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 alcetas* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q60	6.C.Q40
Contig N50 length	5.91 Mb	≥ 1 Mb
Scaffold N50 length	19.06 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 60.5; haplotype 2: 61.6; combined: 61.0	≥ 40
k-mer completeness	Haplotype 1: 86.16%; Haplotype 2: 77.41%; combined: 99.35%	≥ 95%
BUSCO	C:97.3% [S:93.2%; D:4.1%]; F:0.5%; M:2.3%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.32%	≥ 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

Data availability

European Nucleotide Archive: Cupido alcetas (Provençal short-tailed blue). Accession number PRJEB82256. The genome sequence is released openly for reuse. The *Cupido alcetas* 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

Software	Version	Source
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
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**(1): 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

SBN: **Tagfalter Un Ihre Lebensräume. Arten, Gefährdung, Schutz. Band 1**. Basel: Fotorotar AG, 1987.
Reference Source

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**(8): 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

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 27 February 2026

https://doi.org/10.21956/wellcomeopenres.28392.r146608

© **2026 Rupp O.** 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.

✔ **Oliver Rupp**
Justus-Liebig-University, Giessengiessen, Germany

This article presents the diploid genome assembly, including the mitochondrial genome, of Cupido alcetas, a small butterfly. The sequencing methods, including PacBio HiFi and Hi-C sequencing, follow the state-of-the-art methods used for eukaryotic genome sequencing. The subsequent assembly pipeline included well established tools including hifiasm, for the initial assembly, YaHS for scaffolding, and MerquryFK and BUSCO for the quality assessment. The haplotype 1 was manually curated into a chromosome level assembly. All quality metrics point out a high quality assembly. The assembly pipelines are available and the data was submitted to the European Nucleotide Archive.

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: eukaryotic genome assembly

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 26 February 2026

https://doi.org/10.21956/wellcomeopenres.28392.r146614

© 2026 Smolander O. 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.

Olli-Pekka Smolander
University of Helsinki, Helsinki, Finland

The manuscript describes the construction of a high quality chromosome level genome assembly for the Provencal Short-tailed Blue butterfly. The genome has been assembled based on PacBio HiFi and Hi-C data from a single female specimen yielding a primary assembly spanning 493.02 megabases. The assembly is organized into 26 chromosomal pseudomolecules including a unique sex chromosome arrangement involving the W and two Z chromosomes.

The biological significance of the work is based on the conservation status of the species: it is threatened and near extinction in several European countries. The authors also highlight the taxonomic complexity and potential for misidentification with related species. However in the background-section text they mention that records from some regions have low reliability and likely represent a different expanding species without presenting evidence or citations for these claims. To enhance the scientific rationale the authors should clarify this. Is the morphological spot is sometimes insufficient for certain identification? This would allow more clearly justifying the need for these genomic data. Additionally, the mention of associations with the ant genera *Formica* and *Tapinoma* is a useful biological detail and the section would be even more cohesive if a brief sentence explained the significance of these interactions within the family.

The methodology includes current benchmarks for reference quality genomes. The use of PacBio HiFi for sequencing and the Hi-C data for scaffolding is a high-level standard for such work. The assembly process and quality assessment has been done rigorously using appropriate tools.

The manuscript provides sufficient experimental details such as the specific geolocation coordinates, collection date, laboratory procedures and software versions well documented. However, because this butterfly is very similar to other species the authors should mention the specific diagnostic criteria or tools used during sample acquisition to ensure the specimen was not an individual of the similar expanding species mentioned in the background. Including these details would strengthen the confidence in the taxonomic identification of the biological source material

The data are well organized and adhere to open science standards. All raw data and the final assembly are deposited in the European Nucleotide Archive. Detailed assembly statistics are clearly tabulated and visualisations are provided.

The article presents a technically robust genome assembly but would benefit from additional context and evidence in the text to be fully rigorous. Specifically the authors should provide citations or evidence for the claim of low reliability in existing records and the expansion of the

related species. The methods should also clarify the identification process for the specific specimen used. Finally, the results describe a sex chromosome arrangement involving the W and two Z chromosomes. Adding a brief note of context regarding the nature or origin of the Z duplication would help readers appreciate the evolutionary significance of this chromosomal structure. Addressing these would further improve the manuscript and increase its 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: Genomics, bioinformatics, ML/AI, computational biology, microbiome, gut-brain axis

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 16 February 2026

https://doi.org/10.21956/wellcomeopenres.28392.r146615

© 2026 Clark M. 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.

Melody Clark
Natural Environment Research Council, Cambridge, UK

This Wellcome Open Research article describes the haplotype-resolved genome of the short-tailed blue butterfly (female). The morphology distinguishing this specimen from closely related species is well described, as is the biology and biogeography. This species was sequenced under the European Project Psyche and is a priority for sequencing given its threatened and near extinction status in several European countries.

All methods are well described with full acknowledgement of freely available methodologies. The PacBio HiFi and Hi-C assembly comprises 26 haplotype resolved pseudochromosomes including the W and Z1, Z2 sex chromosomes. The assembly includes contamination removal, in addition to

some manual corrections.

The description of the final assembly is well described using standard Open Source tools with the addition of chromosome painting of Meridian elements to illustrate chromosome evolution relative to the Lepidopteran ancestor. As with all Wellcome Sanger Institute produced genomes, this one is of a very high standard.

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: transcriptomics and genomics of invertebrates and fish

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.