DATA NOTE

The genome sequence of the Alpine Blue, *Agriades orbitulus* (de Prunner, 1798) (Lepidoptera: Lycaenidae)

[version 1; peer review: 4 approved]

Paula Escuer[1], Kay Lucek[1], Charlotte J. Wright[2], Joana I. Meier[2], Mark L. Blaxter[2],
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]Tree of Life, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 25 Jul 2025, **10**:362
https://doi.org/10.12688/wellcomeopenres.24607.1
Latest published: 25 Jul 2025, **10**:362
https://doi.org/10.12688/wellcomeopenres.24607.1

Abstract

We present a genome assembly from a male specimen of *Agriades orbitulus* (Alpine Blue; Arthropoda; Insecta; Lepidoptera; Lycaenidae). The assembly contains two haplotypes with total lengths of 520.38 megabases and 521.74 megabases. Most of haplotype 1 (99.85%) is scaffolded into 23 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 15.4 kilobases.

Keywords

Agriades orbitulus, Alpine Blue, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ✓ ✓

	1	2	3	4
version 1 25 Jul 2025	✓ view	✓ view	✓ view	✓ view

1. **Han Xu**, Beijing Forestry University, Beijing, China
2. **Melissa Toups**, University of Louisiana at Lafayette, Lafayette, USA
   Jaymes Awbrey, University of Louisiana at Lafayette Department of Biology Ringgold ID 171253, Lafayette, USA
3. **Nitin Ravikanthachari**, University of Montana, Missoula, USA
4. **Samridhi Chaturvedi**, Tulane University, New Orleans, USA
   Garima Setia, Tulane University, New Orleans, USA

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: Escuer P: Writing – Original Draft Preparation; **Lucek K**: Investigation, Resources, Supervision; **Wright CJ**: Funding Acquisition, Project Administration, Supervision; **Meier JI**: Funding Acquisition, Project Administration, Supervision; **Blaxter ML**: Funding Acquisition, Project Administration, Supervision;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540). The work was also funded by a Swiss National Science Foundation (SNSF) Grant ID 202869 to KL and to PE.
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: Escuer P, Lucek K, Wright CJ *et al.* **The genome sequence of the Alpine Blue, *Agriades orbitulus* (de Prunner, 1798) (Lepidoptera: Lycaenidae) [version 1; peer review: 4 approved]** Wellcome Open Research 2025, **10**:362 https://doi.org/10.12688/wellcomeopenres.24607.1

First published: 25 Jul 2025, **10**:362 https://doi.org/10.12688/wellcomeopenres.24607.1

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; *Agriades*; *Agriades orbitulus* (de Prunner, 1798) (NCBI:txid689475)

Background

The Alpine Argus or *Agriades orbitulus* (de Prunner, 1789) was first described in Piemonte (Italy) and belongs to the Polyommatinae subfamily. It has been identified in the Alps, Norway, Sweden and several locations in the Urals, the mountains of South Siberia, the Lower Amur region, northern Sikhote-Alin, northeast China and Mongolia (Stradomsky *et al.*, 2019).

The upper surface of the male's wings presents a steely blue colouration and the female has brown wings with a turquoise shade at the base of the wings. On the underside, both males and females have beige wings with small black dots on the forewings and large white spots on the hindwings. The larvae feed on several species of milkvetch (*Astragalus*), two *Oxytropis* species (*O. campestris* and *O. montana*) and the alpine sainfoin (*Hedysarum hedysaroides*; Clarke (2022)).

The genus *Agriades* originated 4.2 Mya ago and includes as subgenera the former genus *Albulina* (*orbitulus*; Talavera *et al.*, 2013). Additionally, several subspecies have been described among this group (Stradomsky *et al.*, 2019). The species has been assessed for the IUCN Red List of Threatened Species in 2009 and is listed as Least Concern (van Swaay *et al.*, 2010). This high-quality reference assembly provides the opportunity to shed light on the phylogeny of *Agriades* and species determination. This genome will facilitate comparative and population genomics studies within the genus and among the *Polyommatus* butterflies.

We present a chromosome-level, haplotype-resolved genome sequence of the Alpine Blue, *Agriades orbitulus*, sequenced as part of Project Psyche. The sequence data were derived from a male specimen (Figure 1) collected from Gadmen, Innertkirchen, Switzerland.

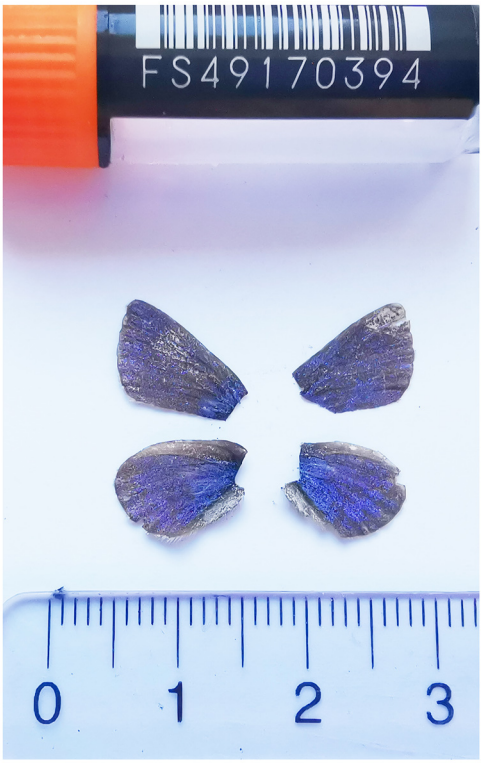

Figure 1. Voucher photograph of the *Agriades orbitulus* (ilAgrOrbi1) specimen used for genome sequencing.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Agriades orbitulus* (specimen ID SAN28000098, ToLID ilAgrOrbi1; Figure 1), collected from Gadmen, Innertkirchen, Switzerland (latitude 46.713, longitude 8.4169; elevation 2 091 m) on 2023-07-14. 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 ilAgrOrbi1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the thorax was homogenised by powermashing using a PowerMasher II tissue disruptor. HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol. DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor®3 for LI PacBio protocol. Sheared DNA was purified by automated SPRI (solid-phase reversible immobilisation). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 34.4 ng/μL and a yield of 1 616.80 ng, with a fragment size of 14.5 kb. The 260/280 spectrophotometric ratio was 1.94, and the 260/230 ratio was 3.48.

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), following 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 head tissue of the ilAgrOrbi1 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.

Table 1. Specimen and sequencing data for BioProject PRJEB78666.

Platform	PacBio HiFi	Hi-C
ToLID	ilAgrOrbi1	ilAgrOrbi1
Specimen ID	SAN28000098	SAN28000098
BioSample (source individual)	SAMEA115109871	SAMEA115109871
BioSample (tissue)	SAMEA115109897	SAMEA115109899
Tissue	thorax	head
Sequencing platform and model	Revio	Illumina NovaSeq X
Run accessions	ERR13485704	ERR13474169
Read count total	1.36 million	1 084.71 million
Base count total	13.74 Gb	163.79 Gb

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 Accu-Clear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech).

Prior to sequencing, libraries were normalised to 10 ng/µL. Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq X.

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

Genome assembly

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

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

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

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021).

Manual corrections included 4 breaks and 14 joins. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

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

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

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

Genome sequence report

Sequence data

The genome of a specimen of *Agriades orbitulus* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 13.74 Gb (gigabases) from 1.36 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 518.95 Mb, with a heterozygosity of 0.53% and repeat content of 43.38%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 26× coverage. Hi-C sequencing produced 163.79 Gb from 1 084.71 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 520.38 Mb in 51 scaffolds, with 183 gaps, and a scaffold N50 of 23.16 Mb (Table 2).

Most of the assembly sequence (99.85%) was assigned to 23 chromosomal-level scaffolds, representing 22 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 2; 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 3).

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

Assembly quality metrics

For haplotype 1, the estimated QV is 65.1, and for haplotype 2, 63.8. When the two haplotypes are combined, the assembly achieves an estimated QV of 64.4. The *k*-mer completeness is 85.97% for haplotype 1, 86.04% for haplotype 2, and 99.29% for the combined haplotypes (Figure 4). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) (Kriventseva *et al.*, 2019) identified 97.0% of the expected gene set (single = 96.6%, duplicated = 0.4%) for haplotype 1.

Table 2. Genome assembly statistics.

Assembly name	ilAgrOrbi1.hap1.1	ilAgrOrbi1.hap2.1
Assembly accession	GCA_964258695.1	GCA_964258715.1
Assembly level	chromosome	scaffold
Span (Mb)	520.38	521.74
Number of chromosomes	23	N/A
Number of contigs	234	286
Contig N50	4.02 Mb	3.61 Mb
Number of scaffolds	51	91
Scaffold N50	23.16 Mb	23.17 Mb
Longest scaffold length (Mb)	34.25	N/A
Sex chromosomes	Z	N/A
Organelles	Mitochondrial genome: 15.4 kb	N/A

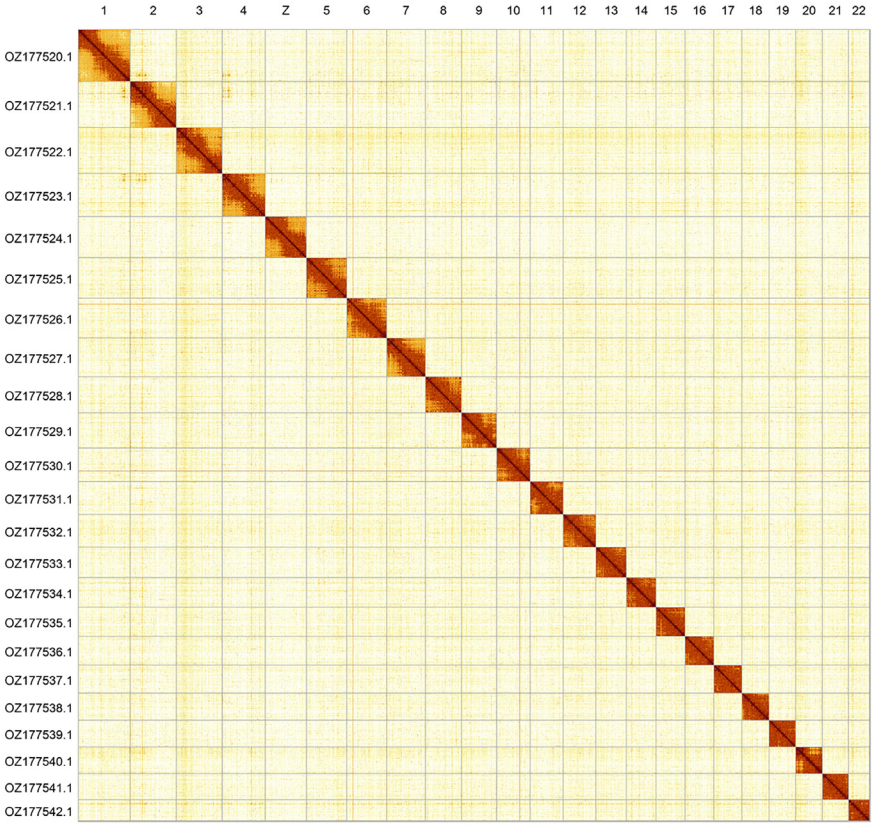

Figure 2. Hi-C contact map of the *Agriades orbitulus* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Agriades orbitulus* ilAgrOrbi1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ177520.1	1	34.25	37	M1;M19
OZ177521.1	2	30.35	37	M18;M30
OZ177522.1	3	30.10	36.50	M25;M5
OZ177523.1	4	28.19	37	M13;M27
OZ177525.1	5	26.40	36.50	M11;M23
OZ177526.1	6	26.27	36.50	M12;M29
OZ177527.1	7	25.40	36.50	M14;M26
OZ177528.1	8	23.63	36.50	M8
OZ177529.1	9	23.16	36.50	M17;M20
OZ177530.1	10	21.97	36.50	M2
OZ177531.1	11	21.62	36	M9
OZ177532.1	12	21.60	36.50	M3
OZ177533.1	13	20.01	36.50	M7
OZ177534.1	14	19.38	36.50	M6
OZ177535.1	15	19.08	36.50	M4
OZ177536.1	16	18.92	36	M16
OZ177537.1	17	18.44	36.50	M21
OZ177538.1	18	17.74	36.50	M15
OZ177539.1	19	17.52	36.50	M22
OZ177540.1	20	17.51	37	M28;M31
OZ177541.1	21	17.18	36.50	M10
OZ177542.1	22	13.77	36.50	M24
OZ177524.1	Z	27.12	36	MZ
OZ177543.1	MT	0.02	18	N/A

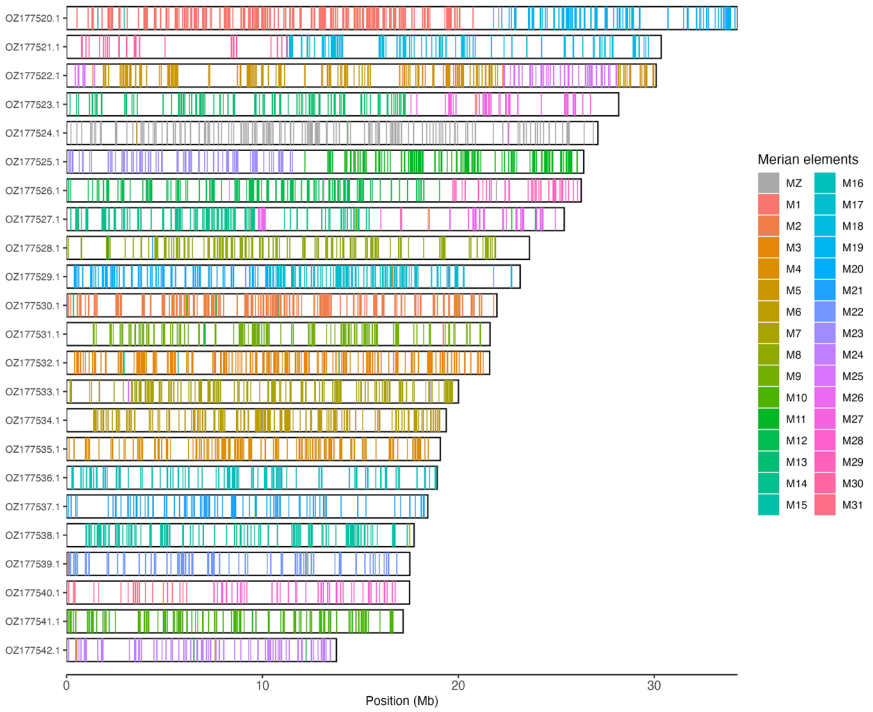

Figure 3. Merian elements painted across chromosomes in the ilAgrOrbi1.hap1.1 assembly of *Agriades orbitulus*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

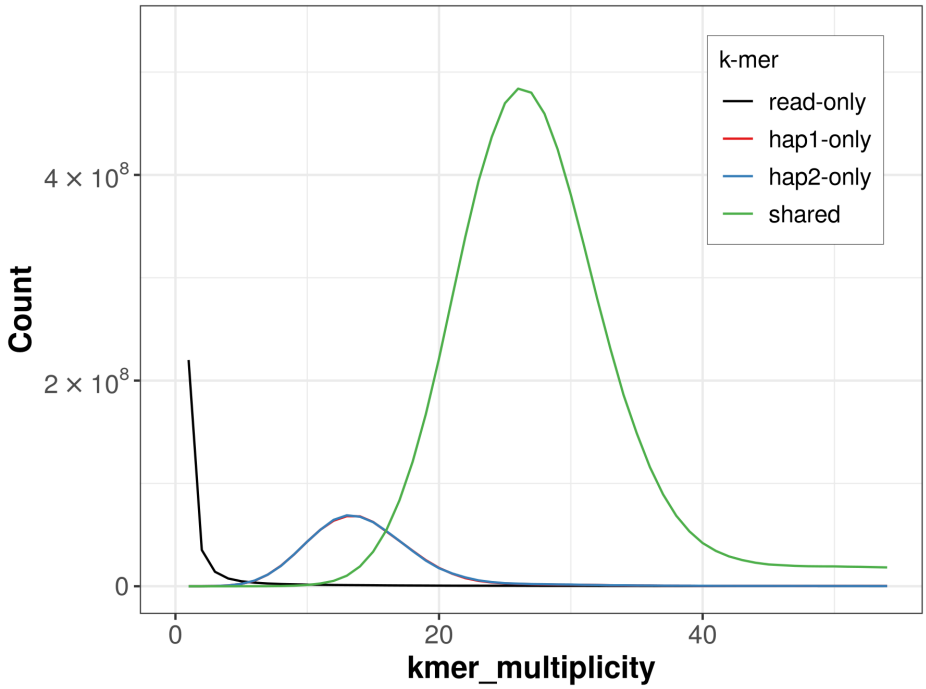

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

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

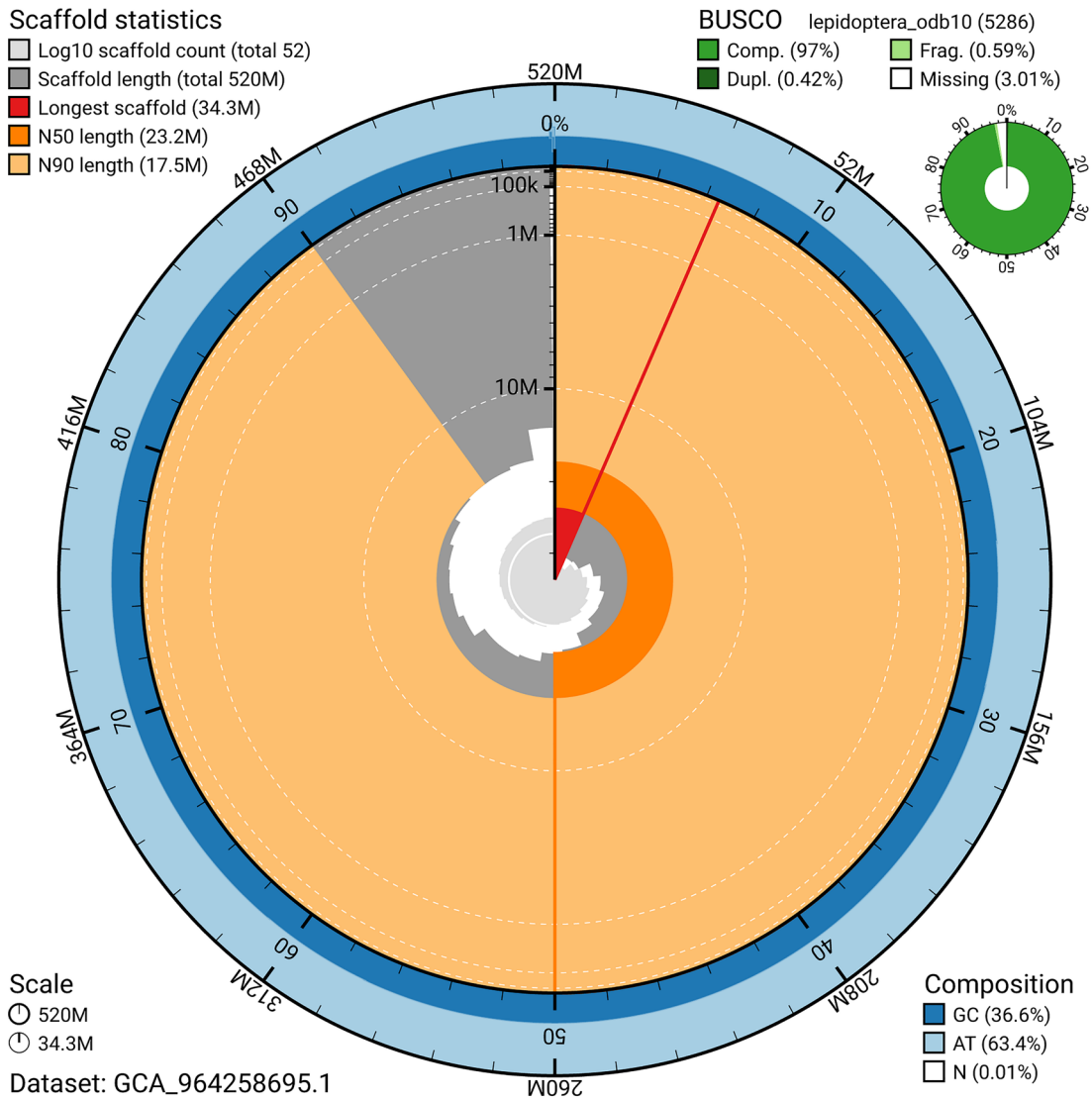

Figure 5. Assembly metrics for ilAgrOrbi1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

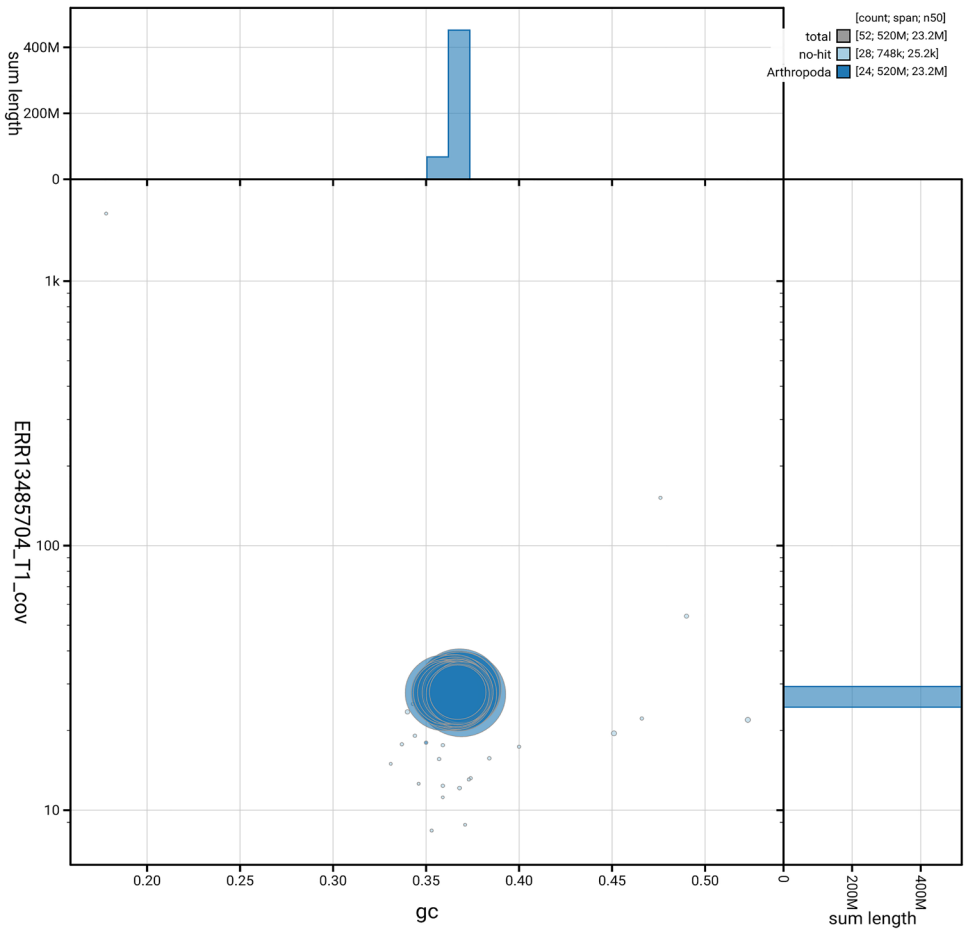

Figure 6. BlobToolKit GC-coverage plot for ilAgrOrbi1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Agriades orbitulus* assembly.

Measure	Value	(Benchmark)
EBP summary	6.C.Q65 (haplotype 1)	6.C.Q40
Contig N50 length	4.02 Mb	≥ 1 Mb
Scaffold N50 length	23.16 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 65.1; haplotype 2: 63.8; combined: 64.4	≥ 40
k-mer completeness	Haplotype 1: 85.97%; Haplotype 2: 86.04%; combined: 99.29%	≥ 95%
BUSCO	C:97.0%[S:96.6%,D:0.4%], F:0.6%,M:2.4%,n:5 286	S > 90%; D < 5%
Percentage of sequence assigned to chromosomes	99.85%	≥ 90%

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: Agriades orbitulus (alpine argus). Accession number PRJEB78666. The genome sequence is released openly for reuse. The *Agriades orbitulus* 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

- 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	666652151335353eef2fcd58880bcef5bc2928e1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84aa44357826c0b6753eb28de	https://github.com/higlass/higlass
MerquryFK	d00d98157618f4e8d1a9190026b19b471055b22e	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		https://github.com/sanger-tol/PretextSnapshot

Software	Version	Source
PretextView	0.2.5	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
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

Allio R, Schomaker-Bastos A, Romiguier J, *et al.*: **MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics.** *Mol Ecol Resour.* 2020; **20**(4): 892–905.
PubMed Abstract | Publisher Full Text | Free Full Text

Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text

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

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

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

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374.
PubMed Abstract | Publisher Full Text | Free Full Text

Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Clarke HE: **A provisional checklist of European butterfly larval foodplants.** *Nota Lepidopterologica.* 2022; **45**(3): 139–167.
Publisher 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

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

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

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

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

Stradomsky BV, Yakovlev RV, Fomina ES: **Molecular genetic confirmation of the subspecies status of blues *Agriades orbitulus pheretimus* (Staudinger, 1892) and *S. orbitulus jugnei* Churkin, 2004 (Lepidoptera: Lycaenidae: Polyommatinae).** *Russ Entomol J.* 2019; **28**(2): 186–188.
Publisher Full Text

Talavera G, Lukhtanov VA, Pierce NE, *et al.*: **Establishing criteria for higher level taxonomic classification using molecular data: the systematics of *Polyommatus* blue butterflies (Lepidoptera, Lycaenidae).** *Cladistics.* 2013; **29**(2): 166–192.
PubMed Abstract | **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**

Van Swaay C, Wynhoff I, Verovnik R, *et al.*: ***Plebejus orbitulus* (Europe assessment)**. 2010.

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 25 October 2025

https://doi.org/10.21956/wellcomeopenres.27117.r134196

✔ **Samridhi Chaturvedi**
Tulane University, New Orleans, Louisiana, USA
Garima Setia
Ecology and Evolutionary Biology, Tulane University, New Orleans, Louisiana, USA

The manuscript "The genome sequence of the Alpine Blue, *Agriades orbitulus* (de Prunner, 1798) (Lepidoptera: Lycaenidae)" presents a high-quality genome assembly with good scaffold N50 and BUSCO completeness. The authors have used appropriate tools for genome assembly and sequence data analysis. The manuscript is overall well written.
However, I propose the following minor corrections:
• More background information about the ecology, life cycle, and behavior of the species would be helpful.
• The introduction should explain more clearly the relevance of this species/ genome assembly.
• Please include the versions of all programs used for the assembly and analysis to ensure reproducibility.
• The k-mer completeness is below the expected benchmark of 95 percent. This should be addressed in the text whether the lower individual completeness is acceptable, expected, or not.
• The manuscript should clarify why haplotype 2 was not assembled to chromosome level.
• The section- “Sample acquisition and DNA barcoding” does not actually include any information about DNA barcoding. This part seems to have been accidentally left out and should be completed.
• It would also be helpful to include more information explaining the figures, either in the captions or in the main text.
This genome will be a valuable resource for future research in Lepidoptera. I recommend that the authors revise and resubmit the manuscript after making these minor changes.

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

Are the protocols appropriate and is the work technically sound?

Partly

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolutionary genomics, speciation, hybridization, high-throughput genome sequencing and assembly

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 25 October 2025

https://doi.org/10.21956/wellcomeopenres.27117.r134195

Nitin Ravikanthachari
University of Montana, Missoula, Montana, USA

The authors report a high-quality genome for the Lycaenid *Agriades orbitulus* using Pac-Bio and Hi-Fi sequencing. The study is a critical resource in increasing genome sequences for Lepidoptera. The study is well designed, and the methods are thorough. The quality metrics reported suggests that the genome is of sufficiently high quality. Overall, the manuscript is well written and ready for Indexing.

I have one minor comment to add. I believe that the manuscript will benefit from additional explanation of Merian elements. While the authors provide citation for the relevant paper, I believe a short summary (2-3 lines) about Merian elements and its use in genome assembly would make the paper better.

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: Population genetics, bioinformatics, evolutionary genetics, plant-insect interactions, insect-microbe interactions.

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 24 October 2025

https://doi.org/10.21956/wellcomeopenres.27117.r134194

Melissa Toups
University of Louisiana at Lafayette, Lafayette, Louisiana, USA
Jaymes Awbrey
University of Louisiana at Lafayette Department of Biology Ringgold ID 171253, Lafayette, Louisiana, USA

The authors present a chromosome-level genome assembly for the Alpine Blue, *Agriades orbitulus*. The methods are described in detail and are the current best practices for genome assembly. This is a useful genomics resource. I have three comments.

First, a *de novo* gene annotation would further enhance the utility of this assembly for researchers interested in comparative, population, and functional genomics.

Second, the Z is identified based on homology to other Lepidopteran assemblies. While this is very likely the Z chromosome, the caveat should be added that this has not been independently verified using coverage analyses or sex-linked polymorphism.

Third, Fig. 4 has extensive overlap for the lines for Hap1 and Hap2. Perhaps a dotted or dashed line would allow the reader to view both lines.

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: sex-chromosome evolution, comparative genomics, population 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.

Reviewer Report 30 September 2025

https://doi.org/10.21956/wellcomeopenres.27117.r132882

Han Xu
Beijing Forestry University, Beijing, Beijing, China

This study reports the chromosome-level, haplotype-resolved genome assembly of Agriades orbitulus, an important addition to the genomic resources of Lepidoptera. The study design is well-designed, the methods advanced, and the data quality high, with assembly metrics meeting current standards for high-quality genomes. The manuscript is clearly structured, and the data are transparent and open, meeting the indexing requirements for data-based papers. Acceptance for indexing is recommended after minor revisions.
The manuscript mentions the sample collection site and collaborators but does not clearly state whether relevant collection permits or ethical review approvals were obtained. It is recommended that relevant statements or license numbers be included in the "Methods". And, the title mentioned "DNA barcoding" but the process of molecular identification was not stated.
The legends of Figure 3 (Merian element distribution) and Figure 4 (k-mer distribution) can be further simplified, and it is recommended that the main conclusions of the figures be briefly stated in the text to enhance readability.

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

Are the protocols appropriate and is the work technically sound?
Partly

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: taxonomy, molecular phylogeny, genomes

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.