DATA NOTE

The genome sequence of the Purple-shot Copper, *Lycaena alciphron* (von Rottemburg, 1775) (Lepidoptera: Lycaenidae)

[version 1; peer review: 2 approved]

Paula Escuer[1], Alena Suchackova Bartonova[2], Kay Lucek[1], 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 Czech Academy of Sciences, Ceske Budejovice, South Bohemia, Czech Republic
[3]Tree of Life, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 28 Jul 2025, **10**:375
https://doi.org/10.12688/wellcomeopenres.24617.1
Latest published: 28 Jul 2025, **10**:375
https://doi.org/10.12688/wellcomeopenres.24617.1

Abstract

We present a genome assembly from a female specimen of *Lycaena alciphron* (Purple-shot Copper; Arthropoda; Insecta; Lepidoptera; Lycaenidae). The assembly contains two haplotypes with total lengths of 497.93 megabases and 461.30 megabases. Most of haplotype 1 (99.62%) is scaffolded into 25 chromosomal pseudomolecules, including the W and Z sex chromosomes. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 15.32 kilobases.

Keywords

Lycaena alciphron, Purple-shot Copper, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 28 Jul 2025	✓ view	✓ view

1. **Jason Charamis**, Foundation for Research and Technology - Hellas, Irákleion, Greece

2. **Sara E Mousa**, Assiut University, Assiut, Egypt
Gehad N Aboulnasr, Assiut University, Assiut, Egypt

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: Investigation, Resources; **Suchackova Bartonova A**: Writing – Original Draft Preparation, Writing – Review & Editing; **Lucek K**: Resources, Supervision; **Wright CJ**: Funding Acquisition, Project Administration, Supervision; **Meier JI**: Funding Acquisition, Project Administration, Supervision; **Blaxter ML**: Funding Acquisition, Project Administration, Supervision;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540). KL and PE were supported by the Swiss National Science Foundation Grant PCEFP3_202869. AL was supported by a grant from VEDA FELLOWSHIPS (CZ.02.01.01/00/22_010/0008117).
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, Suchackova Bartonova A, Lucek K *et al.* **The genome sequence of the Purple-shot Copper, *Lycaena alciphron* (von Rottemburg, 1775) (Lepidoptera: Lycaenidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, **10**:375 https://doi.org/10.12688/wellcomeopenres.24617.1

First published: 28 Jul 2025, **10**:375 https://doi.org/10.12688/wellcomeopenres.24617.1

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; Lycaeninae; *Lycaena*; *Lycaena alciphron* (von Rottemburg, 1775) (NCBI:txid282377)

Background

Lycaena alciphron (Rottemburg, 1775), the Purple-shot Copper, is a Western Palaearctic species, occurring from Morocco (Atlas Mountains) and Iberia in the south-west, to Central Siberia in the east, and Baltic countries in the north. It is present in all Mediterranean Peninsulas, reaching Northern Iran. It is absent from the Great Britain and Fennoscandia. The forewing length is 14–21 mm (Middleton-Welling *et al.*, 2020). Several subspecies were described throughout the range (Tolman & Lewington, 2009), though a formal biogeographic study has not been carried out. Males of the nominate form (Central and Eastern Europe to Siberia) are characterised by a coppery upper side of the wings with an iridescent violet sheen, while females are brown. *Lycaena a. gordius* (Sulzer, 1776) occurs in the southern and sestern Alps, Italy, Pyrenees, and Iberia; it is larger and both sexes have coppery wings, with reduced violet suffusion in males. Another form, *L. a. melibaeus* (Staudinger, 1878) occurs in the southern Balkans and Caucasus. Mitochondrial DNA barcoding studies show that the glacial separation by the Alps is a major barrier, and the Atlas population is divergent (Dapporto *et al.*, 2022; Sucháčková Bartoňová *et al.*, 2024).

The species favours colder habitats, such as piedmont and montane flowery meadows, pastures, heathlands, sandy soils, and shrublands (Beneš *et al.*, 2002). The presence of bare soil and sparse acidophilic vegetation are important for oviposition (Brau *et al.*, 2013; Ebert & Rennwald, 1993). Host plants are sorrels, such as *Rumex acetosella* L. and *R. acetosa* L. The larvae overwinter at the base of the host plant stem and continue their development in spring. The species forms a single adult generation in June and July (Tolman & Lewington, 2009).

According to the IUCN European Red List of butterflies, *L. alciphron* is considered of least concern in Europe, and near threatened in the European Union (Maes *et al.*, 2019; Van Swaay *et al.*, 2010). The species is threatened by the disappearance of small-scale pasture (Beneš *et al.*, 2002).

We present a chromosome-level, haplotype-resolved genome sequence of the Purple-shot Copper, *Lycaena alciphron*, sequenced as part of Project Psyche. The sequence data were derived from a female specimen (Figure 1) collected from Zwischbergen, Switzerland.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult female *Lycaena alciphron* (specimen ID SAN28000091, ToLID ilLycAlci1; Figure 1), collected from Zwischbergen, Switzerland (latitude 46.1694, longitude 8.1179; elevation 1 356 m) on 11/07/2023. The specimen was collected and identified by Paula Escuer (University of Neuchâtel).

Figure 1. Voucher photograph of the *Lycaena alciphron* (ilLycAlci1) specimen used for genome sequencing.

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 ilLycAlci1 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 39.43 ng/μL and a yield of 1 853.21 ng, with a fragment size of 15.6 kb. The 260/280 spectrophotometric ratio was 1.95, and the 260/230 ratio was 2.81.

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 ilLycAlci1 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

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

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

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

Table 1. Specimen and sequencing data for BioProject PRJEB80397.

Platform	PacBio HiFi	Hi-C
ToLID	ilLycAlci1	ilLycAlci1
Specimen ID	SAN28000091	SAN28000091
BioSample (source individual)	SAMEA115109879	SAMEA115109879
BioSample (tissue)	SAMEA115109921	SAMEA115109922
Tissue	thorax	head
Sequencing platform and model	Revio	Illumina NovaSeq X
Run accessions	ERR13725936	ERR13731895
Read count total	1.61 million	903.19 million
Base count total	19.02 Gb	136.38 Gb

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 3 breaks and 49 joins. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation.

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 *Lycaena alciphron* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 19.02 Gb (gigabases) from 1.61 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 477.03 Mb, with a heterozygosity of 1.10% and repeat content of 28.85%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 39× coverage. Hi-C sequencing produced 136.38 Gb from 903.19 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 497.93 Mb in 70 scaffolds, with 121 gaps, and a scaffold N50 of 20.47 Mb (Table 2).

Most of the assembly sequence (99.62%) was assigned to 25 chromosomal-level scaffolds, representing 23 autosomes and the W and Z sex chromosomes. Chromosomes Z and W were assigned based on Hi-C signal. The order and orientation of the scaffolds making up chromosome W are uncertain. 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 62.2, and for haplotype 2, 62.2. When the two haplotypes are combined, the assembly achieves an estimated QV of 62.2. The *k*-mer completeness is 79.65% for haplotype 1, 75.43% for haplotype 2, and 99.62% for the combined haplotypes (Figure 4). BUSCO analysis using the lepidoptera_odb10 reference set ($n = 5\,286$) (Kriventseva *et al.*, 2019) identified 98.4% of the expected gene set

Table 2. Genome assembly statistics.

Assembly name	ilLycAlci1.hap1.1	ilLycAlci1.hap2.1
Assembly accession	GCA_964273875.1	GCA_964273735.1
Assembly level	chromosome	scaffold
Span (Mb)	497.93	461.30
Number of chromosomes	25	N/A
Number of contigs	191	152
Contig N50	7.91 Mb	8.44 Mb
Number of scaffolds	70	80
Scaffold N50	20.47 Mb	20.47 Mb
Longest scaffold length (Mb)	30.36	N/A
Sex chromosomes	W and Z	N/A
Organelles	Mitochondrial genome: 15.32 kb	N/A

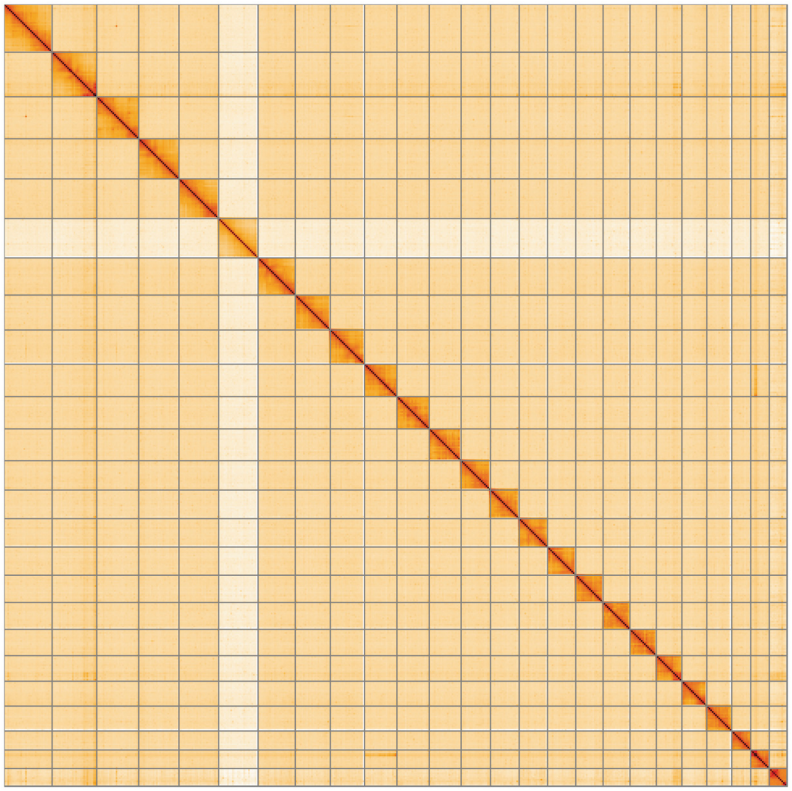

Figure 2. Hi-C contact map of the *Lycaena alciphron* genome assembly. Assembled chromosomes are shown in order of size.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Lycaena alciphron* ilLycAlci1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ187286.1	1	30.36	35.50	M1;M19
OZ187287.1	2	28.31	36	M18;M30
OZ187288.1	3	26.62	35.50	M11;M23
OZ187289.1	4	25.46	35.50	M12;M29
OZ187290.1	5	25.19	35.50	M25;M5
OZ187292.1	6	23.58	36	M14;M26
OZ187293.1	7	22.19	36	M2
OZ187294.1	8	21.69	36	M17;M20
OZ187295.1	9	20.55	36	M3
OZ187296.1	10	20.47	36	M8
OZ187297.1	11	20.21	35.50	M9
OZ187298.1	12	18.61	36	M7
OZ187299.1	13	18.17	36	M16
OZ187300.1	14	18.01	36	M6
OZ187301.1	15	17.88	36.50	M4
OZ187302.1	16	17.30	36	M21
OZ187303.1	17	17.05	36.50	M22
OZ187304.1	18	16.72	36.50	M15
OZ187305.1	19	16.19	37	M13
OZ187306.1	20	15.79	36.50	M28;M31
OZ187307.1	21	15.79	37	M10
OZ187308.1	22	12.06	37	M24
OZ187309.1	23	11.70	37.50	M27
OZ187310.1	W	11.19	37.50	N/A
OZ187291.1	Z	24.96	35	MZ
OZ187311.1	MT	0.02	18.50	N/A

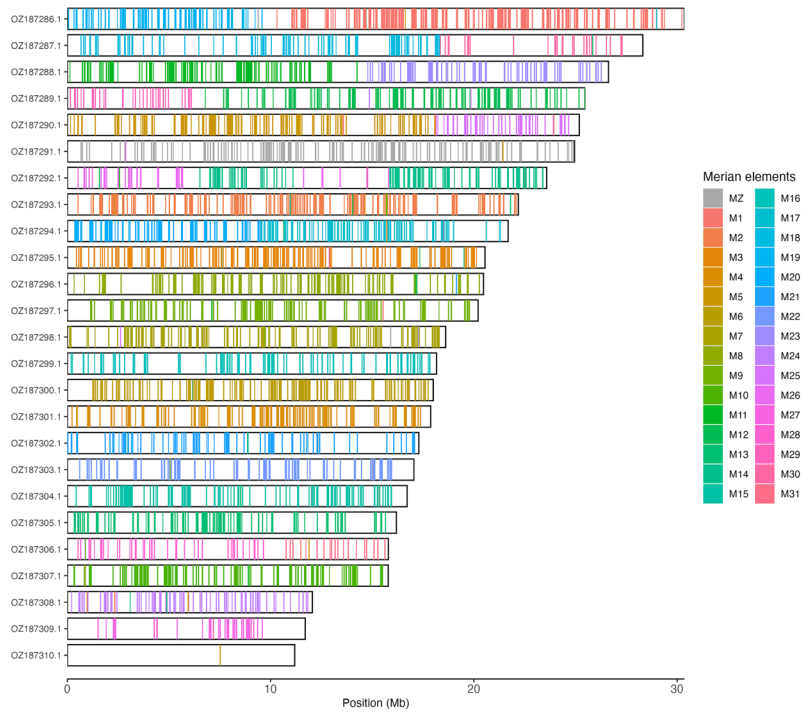

Figure 3. Merian elements painted across chromosomes in the ilLycAlci1.hap1.1 assembly of *Lycaena alciphron*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

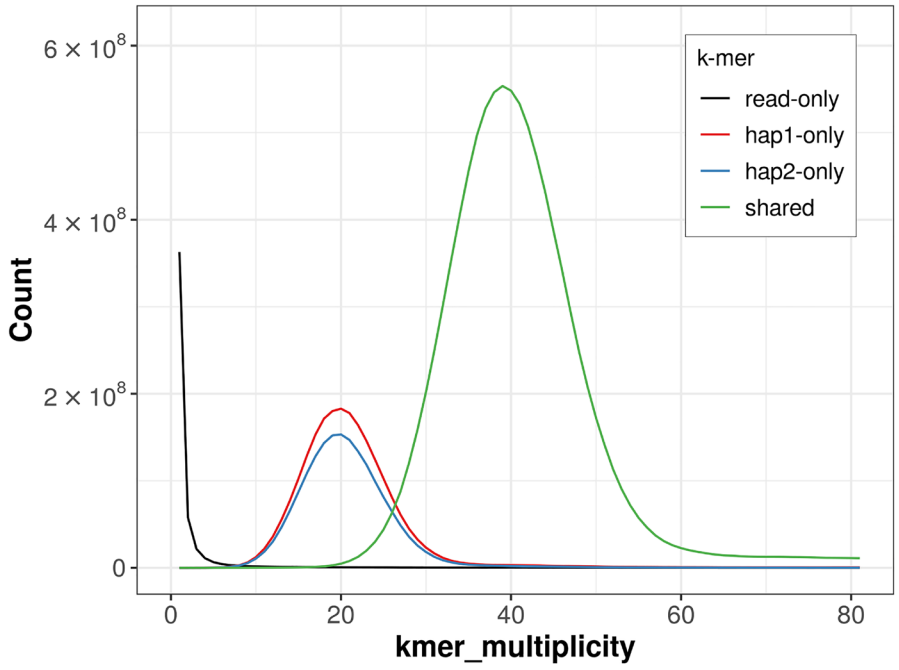

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers 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.

(single = 98.0%, duplicated = 0.4%) for haplotype 1. The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.Q62**, meeting the recommended reference standard.

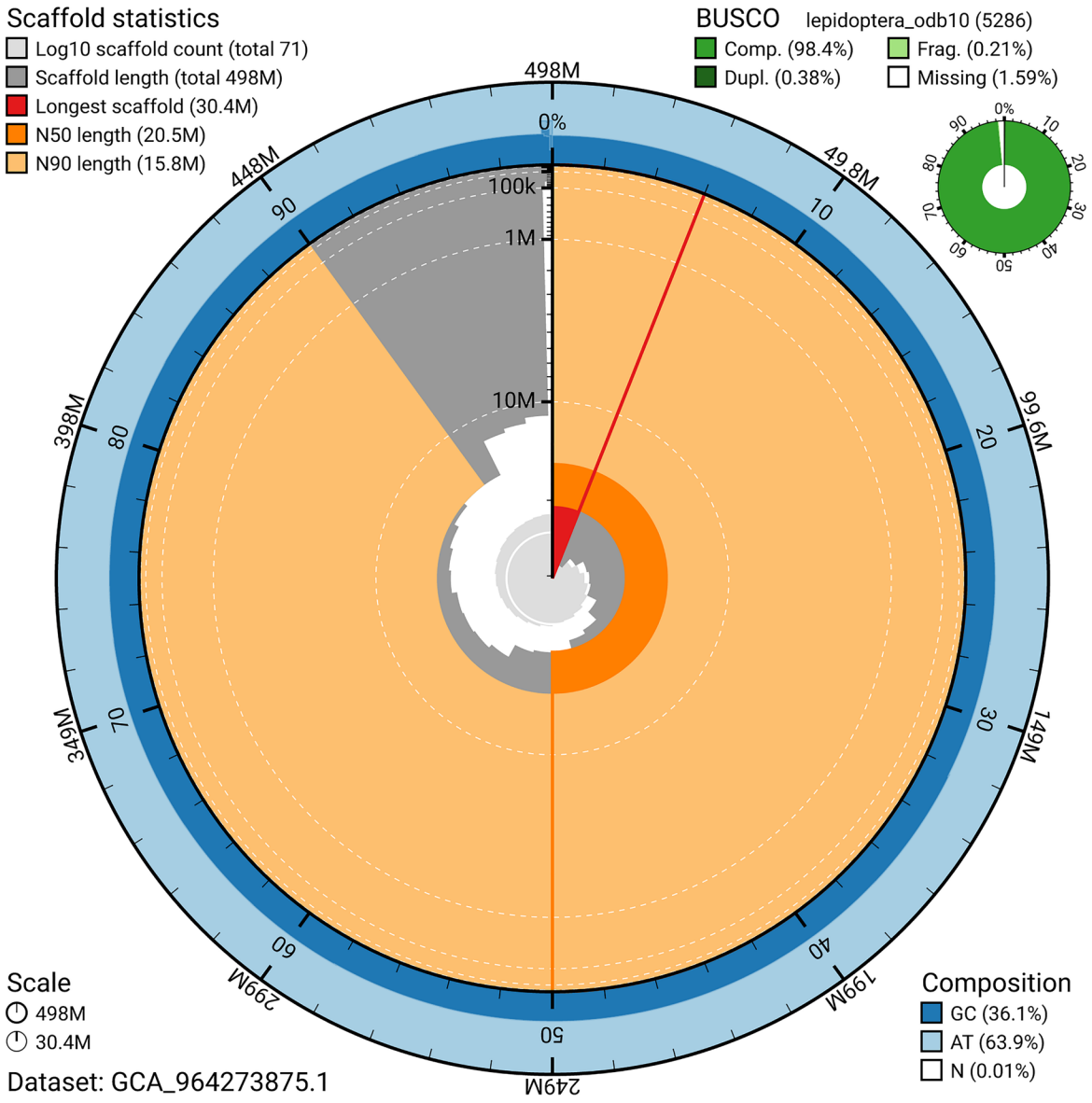

Figure 5. Assembly metrics for ilLycAlci1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

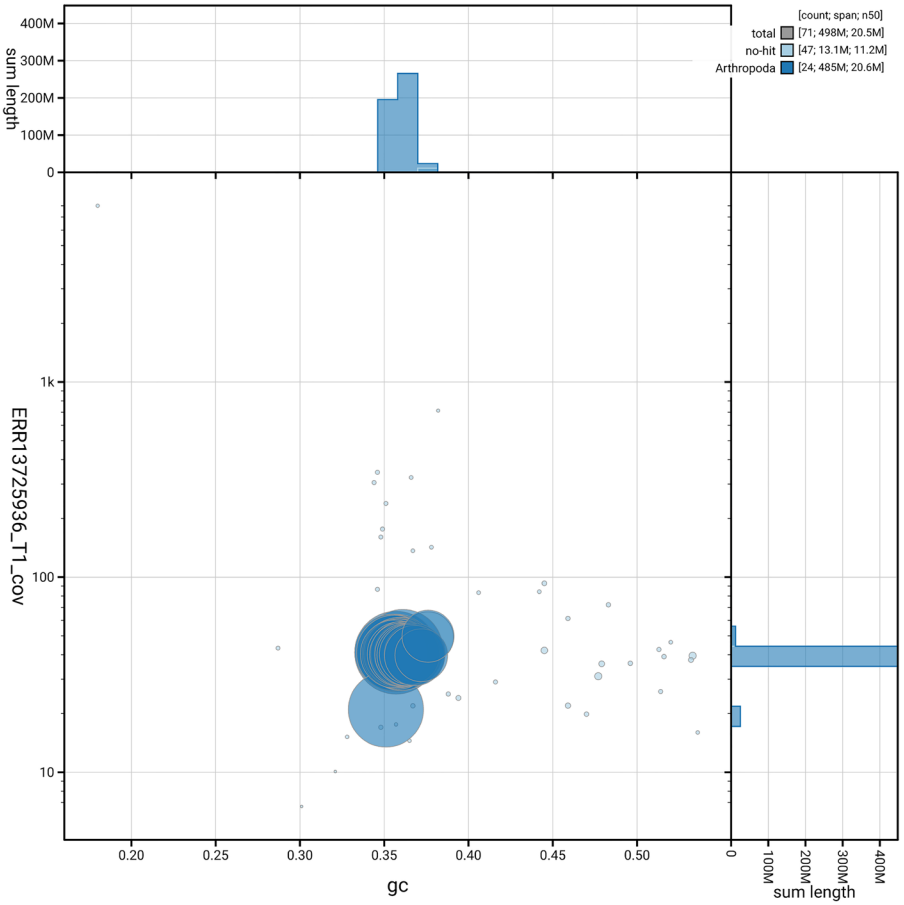

Figure 6. BlobToolKit GC-coverage plot for ilLycAlci1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Lycaena alciphron* assembly.

Measure (Benchmark)	Value
EBP summary (haplotype 1)	6.C.Q62
Contig N50 length (≥ 1 Mb)	7.91 Mb
Scaffold N50 length (= chromosome N50)	20.47 Mb
Consensus quality (QV) (≥ 40)	Haplotype 1: 62.2; haplotype 2: 62.2; combined: 62.2
k-mer completeness (≥ 95%)	Haplotype 1: 79.65%; Haplotype 2: 75.43%; combined: 99.62%
BUSCO* (S > 90%; D < 5%)	C:98.4%[S:98.0%,D:0.4%],F:0.2%,M:1.4%,n:5286
Percentage of assembly assigned to chromosomes (≥ 90%)	99.62%

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.

Data availability

European Nucleotide Archive: Lycaena alciphron (purple-shot copper). Accession number PRJEB80397. The genome sequence is released openly for reuse. The *Lycaena alciphron* genome sequencing initiative is part of the Sanger Institute Tree of Life Programme (PRJEB43745) and Project Psyche (PRJEB71705). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through Ensembl at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

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

Author information

Contributors are listed at the following links:

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

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	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	44086069ee7d4d3f6f3f0012569789ec138f42b84 aa44357826c0b6753eb28de	https://github.com/higlass/higlass
lep_busco_painter	1.0.0	https://github.com/charlottewright/lep_busco_painter

Software	Version	Source
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
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

Beneš J, Konvička M, Dvořák J, *et al.*: **Motýli České Republiky: Rozšíření a ochrana I, II**. Společnost pro ochranu motýlů, 2002.

Brau M, Bolz R, Kolbeck H, *et al.*: **Tagfalter in Bayern**. Ulmer, 2013.
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, 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

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

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

Ebert G, Rennwald E: **Die Schmetterlinge Baden-Wüttembergs, Bd. 2, Tagfalter II**. Ulmer Eugen Verlag, 1993.

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

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): 301–30.
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

Middleton-Welling J, Dapporto L, García-Barros E, *et al.*: **A new comprehensive trait database of European and Maghreb butterflies, Papilionoidea.** *Sci Data.* 2020; **7**(1): 351.
PubMed Abstract | Publisher Full Text | Free 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

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

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): 1575–86.
Publisher Full Text

Tolman T, Lewington R: **Collins butterfly guide**. HarperCollins Publishers, 2009.
Reference Source

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

Van Swaay C, Cuttelod A, Collins S, *et al.*: **European Red List of Butterflies.** Luxembourg: Publications Office of the European Union, 2010.
Publisher Full Text

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.
Publisher Full Text

Wright CJ, Stevens L, Mackintosh A, *et al.*: **Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera.** *Nat Ecol Evol.* 2024; **8**(4): 777–790.
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 29 October 2025

https://doi.org/10.21956/wellcomeopenres.27127.r134611

✔ **Sara E Mousa**
Plant Protection Department, Faculty of Agriculture, Assiut University, Assiut, Egypt
Gehad N Aboulnasr
Zoology and Entomology Department, Faculty of Science, Assiut University, Assiut, Egypt

The most important aspects of this paper, which is a **Data Note** presenting a new genomic resource, are:
1. The Creation of a High-Quality Genomic Resource
The central importance lies in the successful generation of the **first genome assembly** for the **Purple-shot Copper butterfly, *Lycaena alciphron***.

- **Chromosome-Level Assembly:** The research produced a **chromosome-level, haplotype-resolved reference genome**. This means the assembled DNA sequences are not just fragments but are organized into the correct 25 chromosomes (23 autosomes, Z, and W sex chromosomes).
- **High Quality Standard:** The assembly meets the demanding **Earth BioGenome Project (EBP) reference standard** (6.C.Q62), with an exceptionally high quality value (**QV of 62.2**), ensuring its reliability for future studies.

2. Significance for Evolutionary Biology
The primary utility of this dataset is in advancing the study of insect evolution.

- **Chromosomal Evolution:** The genome provides a crucial foundation for investigating the **patterns of chromosomal evolution** and changes in linkage groups (Merian elements) within the Lepidoptera (butterfly and moth) order.
- **Species-Specific Resource:** As the first of its kind for *L. alciphron*, it provides tools for population genetics, conservation biology, and identifying genes related to key traits in this specific species.

3. Technical Soundness and Accessibility
The importance is reinforced by the methods used and the accessibility of the final data.

- **Technical Rigor:** The work is deemed **technically sound** with appropriate protocols, ensuring the reliability of the assembly.

- **Replicability:** Sufficient details on methods and materials are provided, making the work **fully replicable**.
- **Accessibility:** The final dataset is publicly available through the **European Nucleotide Archive (ENA)**, ensuring the genome is a lasting and accessible resource for the global scientific community.

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: Entomology, Molecular Genetics, Ecology, Biological Control, Plant Protection

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 06 October 2025

https://doi.org/10.21956/wellcomeopenres.27127.r134616

Jason Charamis
Foundation for Research and Technology - Hellas, Irákleion, Greece

This manuscript reports the first genome assembly of *Lycaena alciphron.* The work is technically sound and the methods used are clearly described. My only suggestion is to also add a gene annotation to this assembly, which -as described- in the Data Availability section will be performed through Ensembl at the European Bioinformatics Institute.

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: Arthropod Comparative Genomics

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