



DATA NOTE

The genome sequence of the Eastern Rock Grayling, *Hipparchia syriaca* (Staudinger, 1871) (Lepidoptera: Nymphalidae)

[version 1; peer review: 2 approved]

Roger Vila ¹, Alena Sucháčková Bartoňová^{2,3}, Charlotte J. Wright ⁴,
Joana I. Meier⁴, Mark L. Blaxter ⁴,
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

¹Institut de Biologia Evolutiva, CSIC - Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

²Biology Centre of the Academy of Sciences of the Czech Republic, České Budějovice, Czech Republic

³Senckenberg German Entomological Institute, Müncheberg, Germany

⁴Tree of Life Programme, Wellcome Sanger Institute, Hinxton, England, UK

v1 First published: 09 Feb 2026, 11:98
<https://doi.org/10.12688/wellcomeopenres.25847.1>
Latest published: 09 Feb 2026, 11:98
<https://doi.org/10.12688/wellcomeopenres.25847.1>

Abstract

We present a genome assembly from a female specimen of *Hipparchia syriaca* (Eastern Rock Grayling; Arthropoda; Insecta; Lepidoptera; Nymphalidae). The assembly contains two haplotypes with total lengths of 393.56 megabases and 365.79 megabases. Most of haplotype 1 (99.87%) is scaffolded into 30 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 15.47 kilobases. This work is part of Project Psyche, a collaborative programme generating genomes for European butterflies and moths.

Keywords

Hipparchia syriaca; Eastern Rock Grayling; genome sequence; chromosomal; Lepidoptera

Open Peer Review

Approval Status

	1	2
version 1		
09 Feb 2026	view	view

1. **Elena Pazhenkova** , University of Ljubljana, Ljubljana, Slovenia
2. **Marco Gerdol** , University of Trieste, Trieste, Italy

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: **Vila R:** Investigation, Resources, Supervision, Writing – Review & Editing; **Sucháčková Bartoňová A:** 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). ASB was supported by the programme Veda Fellowship (22_010/0008117-01). RV was funded by grant PID2022-139689NB-I00 (MICIU/ AEI/ 10.13039/501100011033 and ERDF, EU) and grant 2021-SGR-00420 (Departament de Recerca i Universitats, Generalitat de Catalunya). *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

Copyright: © 2026 Vila R *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: Vila R, Sucháčková Bartoňová A, Wright CJ *et al.* **The genome sequence of the Eastern Rock Grayling, *Hipparchia syriaca* (Staudinger, 1871) (Lepidoptera: Nymphalidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2026, 11:98 <https://doi.org/10.12688/wellcomeopenres.25847.1>

First published: 09 Feb 2026, 11:98 <https://doi.org/10.12688/wellcomeopenres.25847.1>

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; Nymphalidae; Satyrinae; Satyrini; Satyrina; *Hipparchia*; *Hipparchia syriaca* (Staudinger, 1871) (NCBI:txid876061)

Background

Hipparchia syriaca (Staudinger, 1871), the Eastern Rock Grayling, is an Eastern Mediterranean species of warm dry forests. It is a large, dark grey nymphalid butterfly with a yellow band stretching across both forewing and hindwing. Its forewing length is 31–33 mm (Middleton-Welling *et al.*, 2020). It closely resembles *H. fagi*, and both species might occur in sympatry (Dincă *et al.*, 2009). Its distribution stretches across the Balkans, Corfu, Kefalonia, the Aegean Islands, Turkey, Transcaucasia, Lebanon, and Syria to the Zagros Mountains in Iran, and it reaches its northern limits in the Velebit mountain range in Croatia.

The species has a single annual generation between May and October, with summer aestivation, when the adults retreat to forests and rest on tree trunks or rocks (Tolman & Lewington, 2008). In autumn, the butterflies mate and females lay eggs on dwarf Mediterranean shrubs or grasses. The larvae depend on autumn rains and resprouting of grasses, which they use as food plants (Wagner, 2025). Its habitat includes various hot, dry, bushy sites, often on rocky slopes, and open pine and oak woodlands.

It is considered to be of Least Concern in the European Red List of butterflies (Maes *et al.*, 2019; van Swaay *et al.*, 2025), though large fires in the Mediterranean and human structures such as wind power plants may destroy its habitats (Wagner, 2025).

We present a chromosome-level genome sequence for *Hipparchia syriaca*, sequenced as part of Project Psyche. The sequence data were derived from a female specimen (Figure 1) collected from Măcin, Romania.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult female *Hipparchia syriaca* (specimen ID SAN28000316, ToLID ilHipSyri1; Figure 1), collected from Macin, Romania (latitude 45.2449, longitude 28.1897) on 2021-09-08. The specimen was collected and identified by Roger Vila (Institut de Biologia Evolutiva).

Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The ilHipSyri1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from



Figure 1. Voucher photograph of the *Hipparchia syriaca* (ilHipSyri1) specimen used for genome sequencing.

the whole organism 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 26.2 ng/μL and a yield of 1 231.40 ng, with a fragment size of 13.3 kb. The 260/280 spectrophotometric ratio was 1.84, and the 260/230 ratio was 2.16.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), according to the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment.

Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

The sample was sequenced on a Revio instrument (Pacific Biosciences). The prepared library was normalised to 2 nM, and 15 µL was used for making complexes. Primers were annealed and polymerases bound to generate circularised complexes, following the manufacturer's instructions. Complexes were purified using 1.2X SMRTbell beads, then diluted to the Revio loading concentration (200–300 pM) and spiked with a Revio sequencing internal control. The sample was sequenced on a Revio 25M SMRT cell. The SMRT Link software (Pacific Biosciences), a web-based workflow manager, was used to configure and monitor the run and to carry out primary and secondary data analysis.

Specimen details, sequencing platforms, and data yields are summarised in [Table 1](#).

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the whole organism of the ilHipSyr1 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS

Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

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

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

Specimen details, sequencing platforms, and data yields are summarised in [Table 1](#).

Genome assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts (*k* = 31) was generated from the filtered reads using [FastK](#). GenomeScope2 ([Ranallo-Benavidez et al., 2020](#))

Table 1. Specimen and sequencing data for BioProject PRJEB82280.

Platform	PacBio HiFi	Hi-C
ToLID	ilHipSyr1	ilHipSyr1
Specimen ID	SAN28000316	SAN28000316
BioSample (source individual)	SAMEA115768873	SAMEA115768873
BioSample (tissue)	SAMEA115768889	SAMEA115768889
Tissue	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13946501	ERR13947529
Read count total	1.12 million	754.99 million
Base count total	11.37 Gb	114.00 Gb

was used to analyse the k -mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

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

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

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included ten breaks, 62 joins, and removal of one haplotypic duplication. This reduced the scaffold count by 39.0% and increased the total assembly length by 1.0%. The curation process is described at <https://gitlab.com/wtsi-grit/rapid-curation>. PretextViewSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

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

The genome was analysed using the BlobToolKit pipeline, a Nextflow (Di Tommaso *et al.*, 2017) implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified from the NCBI Taxonomy (Schoch *et al.*, 2020). For the three domain-level lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all outputs into a blobdir

for visualisation. The BlobToolKit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), with containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

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

Genome sequence report

Sequence data

PacBio sequencing of the *Hipparchia syriaca* specimen generated 11.37 Gb (gigabases) from 1.12 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 382.10 Mb, with a heterozygosity of 1.67% and repeat content of 28.14% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 29× coverage. Hi-C sequencing produced 114.00 Gb from 754.99 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 393.56 Mb in 49 scaffolds, with 148 gaps, and a scaffold N50 of 14.34 Mb (Table 2).

Most of the assembly sequence (99.87%) was assigned to 30 chromosomal-level scaffolds, representing 28 autosomes and the W and Z 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). The sex chromosomes were identified by read coverage. During curation, we observed that scaffolds on chromosome W have uncertain order and orientation.

The mitochondrial genome was also assembled (length 15.47 kb, OZ218331.1). 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 66.3, and for haplotype 2, 65.8. When the two haplotypes are combined, the assembly achieves an estimated QV of 66.1. The k -mer completeness is 72.67% for haplotype 1, 69.41% for haplotype 2, and

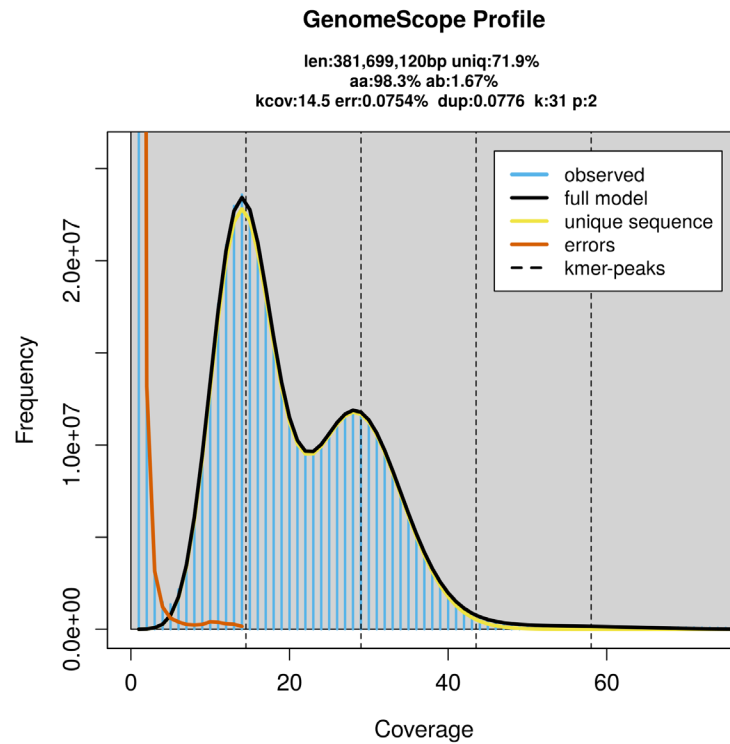


Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 2. Genome assembly statistics.

Assembly name	ilHipSyri1.hap1.1	ilHipSyri1.hap2.1
Assembly accession	GCA_965112445.1	GCA_965112535.1
Assembly level	chromosome	scaffold
Span (Mb)	393.56	365.79
Number of chromosomes	30	scaffold-level
Number of contigs	197	125
Contig N50	5.11 Mb	5.88 Mb
Number of scaffolds	49	49
Scaffold N50	14.34 Mb	14.09 Mb
Longest scaffold length (Mb)	17.15	-
Sex chromosomes	W and Z	-
Organelles	Mitochondrion: 15.47 kb	-

99.29% for the combined haplotypes (Figure 5). BUSCO analysis using the lepidoptera_odb10 reference set ($n = 5\,286$) identified 98.5% of the expected gene set (single = 98.1%, duplicated = 0.4%) in haplotype 1. For haplotype 2, BUSCO analysis identified 94.2% of the expected gene set

(single = 94.0%, duplicated = 0.3%). 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.

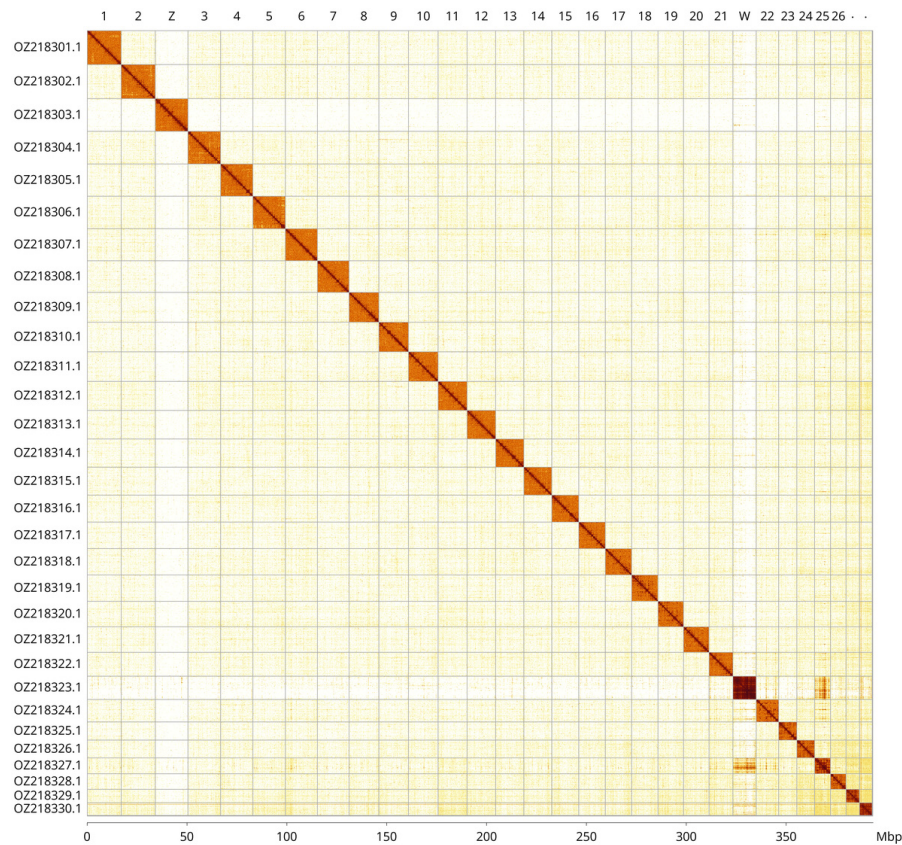


Figure 3. Hi-C contact map of the *Hipparchia syriaca* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Hipparchia syriaca* ilHipSyri1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ218301.1	1	17.15	36.50	M2
OZ218302.1	2	17.09	37	M1
OZ218304.1	3	16.30	37	M3
OZ218305.1	4	16.20	37	M8
OZ218306.1	5	16.18	37	M17;M20
OZ218307.1	6	16.07	37	M19;M26
OZ218308.1	7	15.84	36.50	M9
OZ218309.1	8	14.92	37	M5
OZ218310.1	9	14.89	36.50	M18
OZ218311.1	10	14.76	37	M12
OZ218312.1	11	14.51	37	M14;M29
OZ218313.1	12	14.34	37	M7
OZ218314.1	13	14.11	37	M16
OZ218315.1	14	13.98	37	M4

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ218316.1	15	13.51	37	M6
OZ218317.1	16	13.30	37	M21
OZ218318.1	17	13.17	37	M22
OZ218319.1	18	13.15	37.50	M15
OZ218320.1	19	12.86	37.50	M11
OZ218321.1	20	12.71	37	M10
OZ218322.1	21	12.03	37	M13
OZ218324.1	22	11.32	37.50	M23
OZ218325.1	23	9.10	37.50	M28
OZ218326.1	24	8.84	37.50	M24
OZ218327.1	25	8	39.50	M30
OZ218328.1	26	7.77	37.50	M27
OZ218329.1	27	6.67	37.50	M25
OZ218330.1	28	6.37	38	M31
OZ218323.1	W	11.57	37.50	-
OZ218303.1	Z	16.35	36.50	MZ

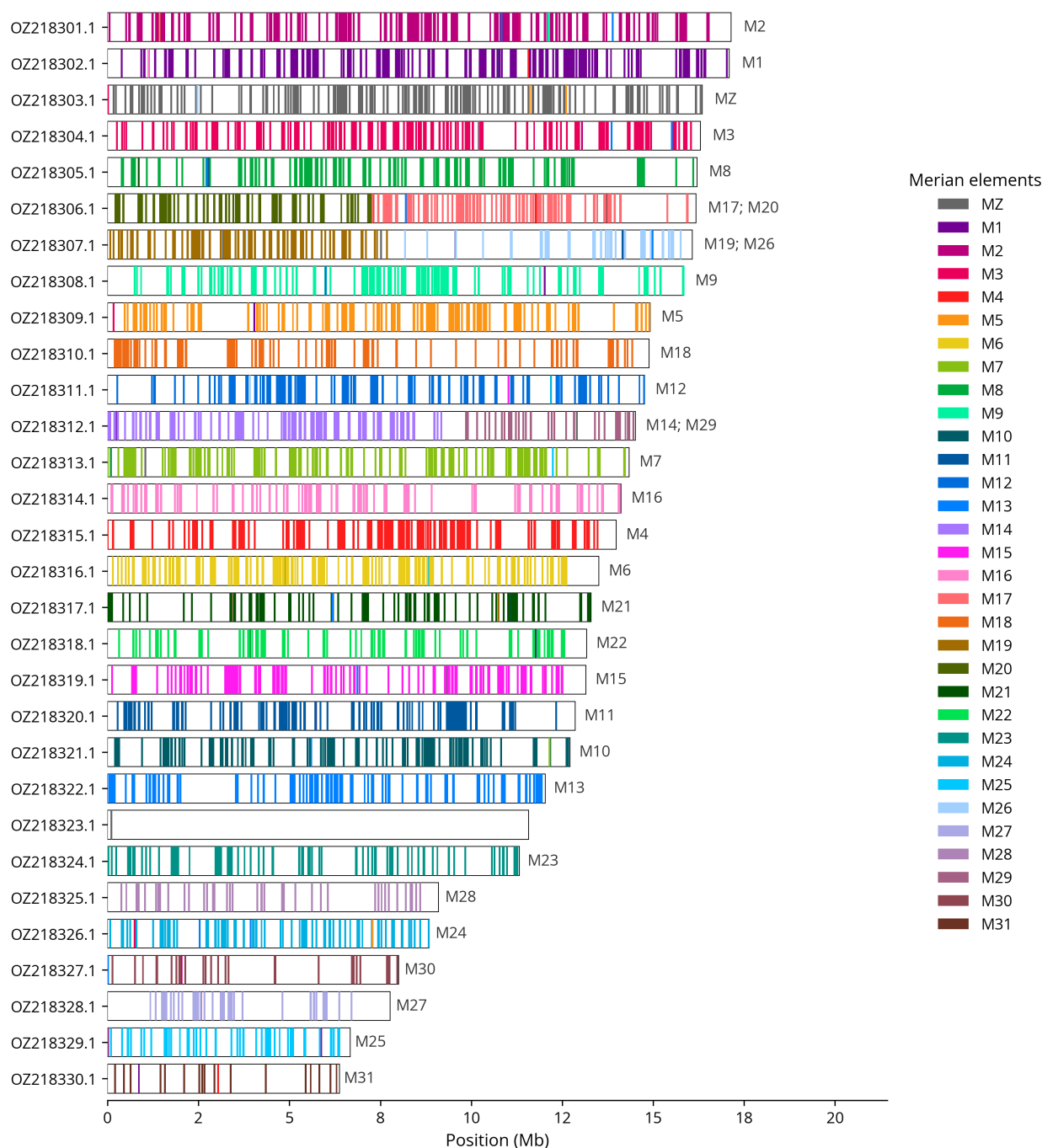


Figure 4. Merian elements painted across chromosomes in the ilHipSyri1.hap1.1 assembly of *Hipparchia syriaca*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

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.Q66, 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

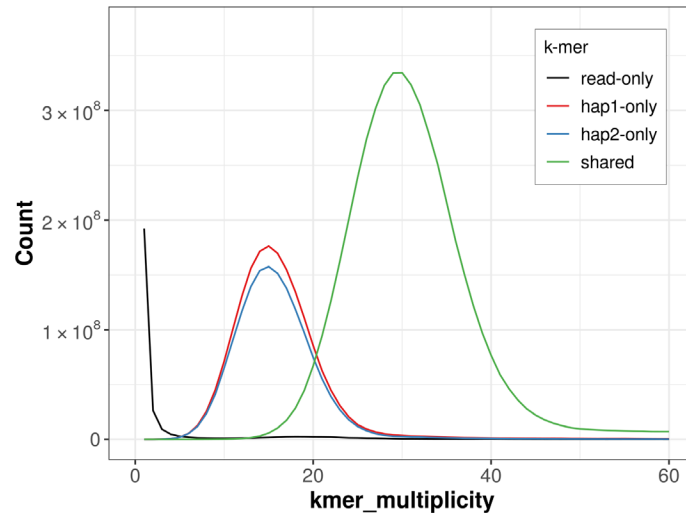


Figure 5. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

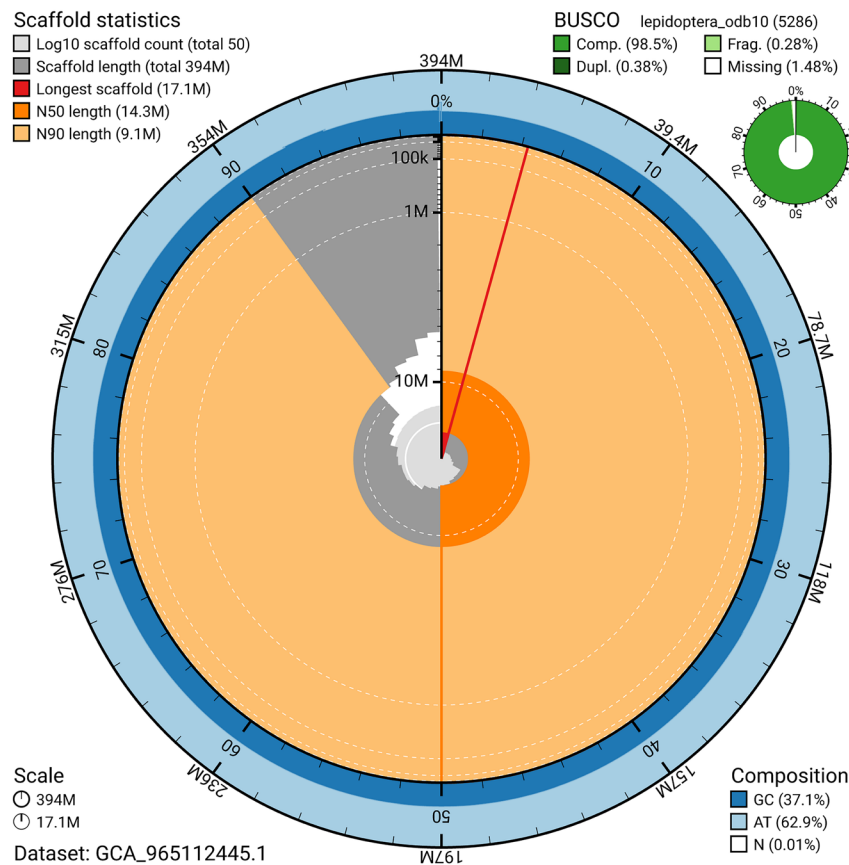


Figure 6. Assembly metrics for ilHipSyri1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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](#).

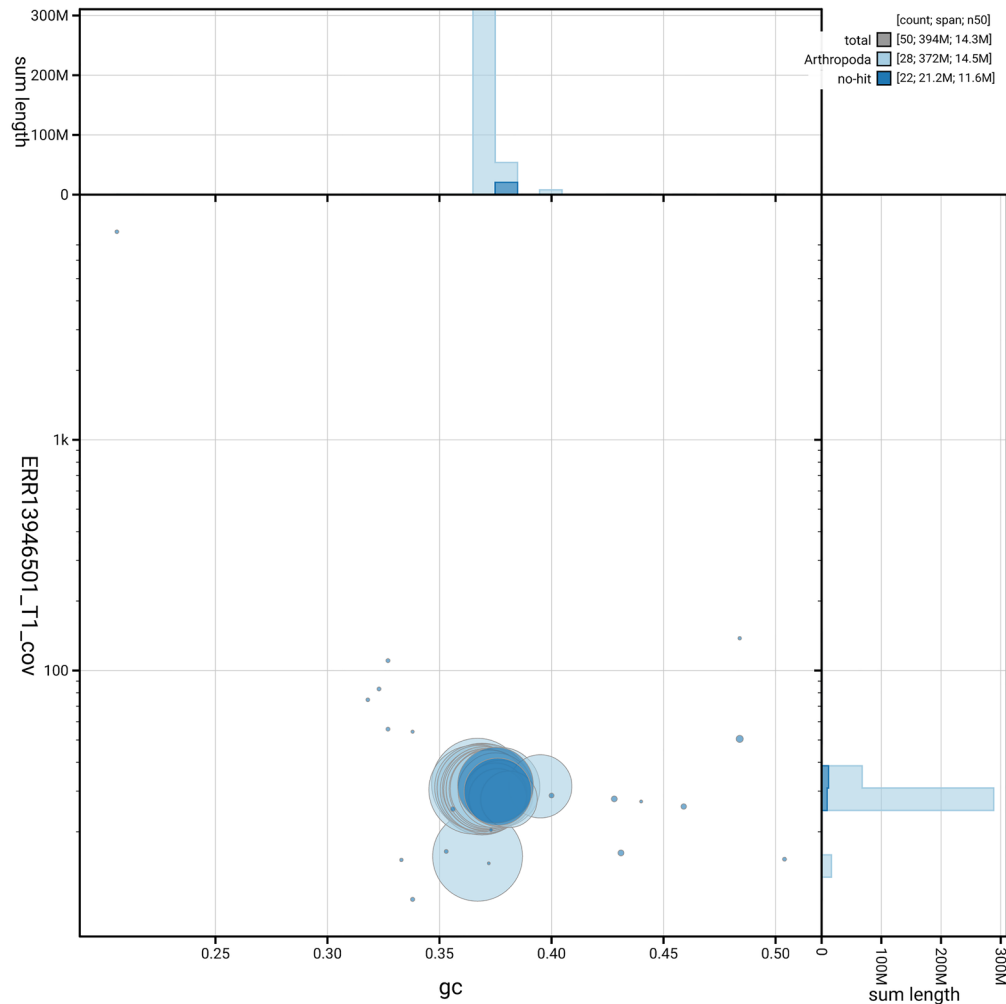


Figure 7. BlobToolKit GC-coverage plot for ilHipSyri1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Hipparchia syriaca* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q66	6.C.Q40
Contig N50 length	5.11 Mb	≥ 1 Mb
Scaffold N50 length	14.34 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 66.3; haplotype 2: 65.8; combined: 66.1	≥ 40
k-mer completeness	Haplotype 1: 72.67%; Haplotype 2: 69.41%; combined: 99.29%	≥ 95%
BUSCO	C:98.5% [S:98.1%; D:0.4%]; F:0.3%; M:1.2%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.87%	≥ 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

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: *Hipparchia syriaca*. Accession number [PRJEB82280](#). The genome sequence is released openly for reuse. The *Hipparchia syriaca* genome sequencing

initiative is part of the Sanger Institute Tree of Life Programme (PRJEB43745) and Project Psyche (PRJEB71705). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through [Ensembl](#) at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in [Table 1](#) and [Table 2](#).

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

Author information

Contributors are listed at the following links:

- Members of the [Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team](#)
- Members of [Wellcome Sanger Institute Scientific Operations – Sequencing Operations](#)
- Members of the [Wellcome Sanger Institute Tree of Life Core Informatics team](#)
- Members of the [Tree of Life Core Informatics collective](#)
- Members of the [Project Psyche Community](#)

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast/
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Hifiasm	0.19.8-r603	https://github.com/chhy123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
lep_buscoPainter	1.0.0	https://github.com/charlottewright/lep_buscoPainter
MercuryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2

Software	Version	Source
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
PretextView	0.0.5	https://github.com/sanger-tol/PretextView
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](#)
- 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](#)
- Dincă V, Cuvelier S, Székely L, *et al.*: **New data on the Rhopalocera (Lepidoptera) of Dobrogea (South-Eastern Romania).** *Phegea.* 2009; **37**(1): 1–21.
[Reference Source](#)
- Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 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): g1aa153.
[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): D807–11.
[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, Seppely 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](#)
- 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](#)

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](#)

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI taxonomy: a comprehensive**

update on curation, resources and tools. *Database (Oxford).* 2020; **2020**: baaa062.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Tolman T, Lewington R: **Collins butterfly guide: the most complete field guide to the butterflies of Britain and Europe.** London: HarperCollins Publishers, 2008.

Uliano-Silva M, Ferreira JGRN, Krashennikova 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, Warren M, Ellis S, *et al.*: **European red list of butterflies: measuring the Pulse of European biodiversity.** Brussels, Belgium: European Commission, 2025.

[Reference Source](#)

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](#)

Wagner W: ***Hipparchia syriaca* (Staudinger, 1871).** 2025.

[Reference Source](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 28 March 2026

<https://doi.org/10.21956/wellcomeopenres.28448.r150670>

© 2026 Gerdol 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.



Marco Gerdol 

University of Trieste, Trieste, Italy

This manuscript provides a report about the genome assembly of *Hipparchia syriaca*,

From a methodological point of view, it follows a gold standard pipeline, so there is no major concern about the technical soundness of this work.

The assembly obtained by the authors is of high quality, it reaches a chromosome scale and is characterized by good completeness, even though the alternative haplotype was assembled at a scaffold level.

Overall, this work does not require revisions and can be accepted for indexing, in my opinion, in its present form.

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

Yes

Are the protocols appropriate and is the work technically sound?

Yes

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

Yes

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Comparative genomics

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 28 March 2026

<https://doi.org/10.21956/wellcomeopenres.28448.r150220>

© 2026 Pazhenkova E. 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.

**Elena Pazhenkova** 

University of Ljubljana, Ljubljana, Slovenia

The article presents a comprehensive genome assembly of the Eastern Rock Grayling, *Hipparchia syriaca* (Lepidoptera, Nymphalidae), generated as part of the Tree of Life and Project Psyche initiatives. The authors produced a high-quality assembly from a female specimen using PacBio HiFi long-read sequencing combined with Hi-C data for scaffolding and phasing. The resulting assembly comprises two haplotypes with total lengths of 393.56 megabases and 365.79 megabases. 99.87% of haplotype 1 is scaffolded into 30 chromosomal pseudomolecules, including both W and Z sex chromosomes, while haplotype 2 remains at scaffold level. The mitochondrial genome was also assembled, with a length of 15.47 kilobases.

GenomeScope analysis estimated a haploid genome size of 382.10 megabases, with heterozygosity of 1.67% and repeat content of 28.14%, providing a solid framework for assembly expectations. The sequencing effort yielded approximately 29× coverage of the nuclear genome, complemented by extensive Hi-C data that enabled chromosome-level scaffolding and careful manual curation.

Importantly, the study incorporates chromosome painting using Merian elements, representing the 32 ancestral linkage groups of Lepidoptera. The distribution of these elements across the assembled chromosomes provides insight into chromosomal organisation and evolutionary rearrangements in *H. syriaca*. This adds an evolutionary genomics perspective to the assembly, enabling comparisons with other lepidopteran genomes and contributing to a broader understanding of chromosome evolution within the group.

Overall, this chromosomal-level genome assembly represents a valuable genomic resource for *H. syriaca*. It provides an important foundation for future studies on population genomics, adaptation, and chromosomal evolution in Satyrinae butterflies, particularly given the species' broad distribution and potential sympatry with closely related taxa. The combination of robust sequencing depth, high assembly quality, and thorough curation makes this work a significant contribution to Lepidoptera genomics and biodiversity research.

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, bioinformatics, population 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.
