DATA NOTE

The genome sequence of the Black Hairstreak, *Satyrium pruni* (Linnaeus, 1758) (Lepidoptera: Lycaenidae)

[version 1; peer review: 2 approved]

Jaakko Pohjoismaki[1], Marko Mutanen[2], Charlotte J. Wright[3], Joana I. Meier[3], Mark L. Blaxter[3],
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Project Psyche Community

[1]University of Eastern Finland, Joensuu, North Karelia, Finland
[2]University of Oulu, Oulu, Northern Ostrobothnia, Finland
[3]Tree of Life, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 28 Jul 2025, **10**:377
https://doi.org/10.12688/wellcomeopenres.24619.1
Latest published: 28 Jul 2025, **10**:377
https://doi.org/10.12688/wellcomeopenres.24619.1

Abstract

We present a genome assembly from a male specimen of *Satyrium pruni* (Black Hairstreak; Arthropoda; Insecta; Lepidoptera; Lycaenidae). The assembly contains two haplotypes with total lengths of 869.86 megabases and 870.75 megabases. Most of haplotype 1 (99.02%) 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 15.46 kilobases.

Keywords

Satyrium pruni, Black Hairstreak, 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. **Frederik Hendrickx**, Royal Belgian institute of natural Sciences, Brussels, Belgium
2. **Christopher J. Fields**, University of Illinois Urbana-Champaign, Urbana, 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: Pohjoismaki J: Investigation, Resources, Writing – Original Draft Preparation; **Mutanen M**: 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 funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Pohjoismaki J, Mutanen M, Wright CJ *et al.* **The genome sequence of the Black Hairstreak, *Satyrium pruni* (Linnaeus, 1758) (Lepidoptera: Lycaenidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, **10**:377 https://doi.org/10.12688/wellcomeopenres.24619.1

First published: 28 Jul 2025, **10**:377 https://doi.org/10.12688/wellcomeopenres.24619.1

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; Theclinae; Eumaeini; *Satyrium*; *Satyrium pruni* (Linnaeus, 1758) (NCBI:txid876075)

Background

The Black Hairstreak (*Satyrium pruni*) is a small butterfly in the family Lycaenidae, which have dark brown wings with subtle orange markings on the underside and a distinctive black "hairstreak" line running across the hindwings. Having a broad distribution range across Europe, as well as parts of Asia, this species inhabits woodland edges and scrubby areas, where its larval host plants (*Prunus* spp.) are found (Silvonen *et al.*, 2014).

Satyrium pruni is univoltine, with adults emerging from late June to mid-July. The species is rather inconspicuous and challenging to observe, flying relatively short distances and often sitting on leaves on higher vegetation. Females lay their eggs on blackthorn twigs, which overwinter and hatch in spring. Larvae feed on emerging *Prunus* leaves before pupating in early summer.

Although *S. pruni* is widely distributed, it is very localised in Europe and the species is sensitive to habitat degradation, making it endangered in several European countries. For example, in the UK it is considered to be endangered (EN) (Fox *et al.*, 2022), confined to a few isolated sites in the East Midlands, and is listed as a conservation priority (Butterfly Conservation, 2025). In Finland, the specimen for this work was collected, *S. pruni* is a relatively common but rarely abundant species (Silvonen *et al.*, 2014).

The reference genome assembly presented here for *Satyrium pruni* is an addition to the growing collection of butterfly genomes, offering valuable resource for future studies and genetic monitoring of populations. Furthermore, it provides a reference for comparative studies across Lycaenid butterflies, aiding in the investigation of evolutionary relationships, host plant co-adaptation, and population genetics. The sequence data were derived from a male specimen (Figure 1) collected from Noljakka, Joensuu, North Karelia, Finland.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult male *Satyrium pruni* (specimen ID SAN28000010, ToLID ilSatPrun1; Figure 1), collected from Noljakka, Joensuu, North Karelia, Finland (latitude 62.6226, longitude 29.7069) on 07/07/2023. The specimen was collected and identified by Jaakko Pohjoismaki (University of Eastern Finland).

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 ilSatPrun1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the whole organism was homogenised by powermashing using a PowerMasher II tissue disruptor.

Figure 1. Photograph of the *Satyrium pruni* (ilSatPrun1) specimen used for genome sequencing.

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 51.74 ng/μL and a yield of 2 431.78 ng, with a fragment size of 15.1 kb. The 260/280 spectrophotometric ratio was 1.88, and the 260/230 ratio was 2.53.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Samples with an average fragment size greater than 8 kb and total mass exceeding 400 ng were eligible for the low-input SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences, California, USA), depending on genome size and required sequencing depth. Libraries were prepared using the SMRTbell Prep Kit 3.0 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 whole organism tissue of the ilSatPrun1 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 PRJEB79118.

Platform	PacBio HiFi	Hi-C
ToLID	ilSatPrun1	ilSatPrun1
Specimen ID	SAN28000010	SAN28000010
BioSample (source individual)	SAMEA114539589	SAMEA114539589
BioSample (tissue)	SAMEA114539623	SAMEA114539623
Tissue	whole organism	whole organism
Sequencing platform and model	Revio	Illumina NovaSeq X
Run accessions	ERR13510362	ERR13621472
Read count total	2.40 million	938.48 million
Base count total	22.81 Gb	141.71 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 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.

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 9 breaks and 15 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 *Satyrium pruni* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 22.81 Gb (gigabases) from 2.40 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 859.49 Mb, with a heterozygosity of 0.57% and repeat content of 36.07%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 26× coverage. Hi-C sequencing produced 141.71 Gb from 938.48 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 869.86 Mb in 131 scaffolds, with 333 gaps, and a scaffold N50 of 37.75 Mb (Table 2).

Most of the assembly sequence (99.02%) was assigned to 23 chromosomal-level scaffolds, representing 22 autosomes and the Z sex chromosome. Chromosome Z was assigned by synteny to the genome of *Phengaris arion* (GCA_963565745.1). 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 59.6, and for haplotype 2, 59.7. When the two haplotypes are combined, the assembly achieves an estimated QV of 59.6. The *k*-mer completeness is 87.70% for haplotype 1, 87.71% for haplotype 2, and 99.16% for the combined haplotypes (Figure 4). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) (Kriventseva *et al.*, 2019) identified 97.9% of the expected

Table 2. Genome assembly statistics.

Assembly name	ilSatPrun1.hap1.1	ilSatPrun1.hap2.1
Assembly accession	GCA_964274205.1	GCA_964274185.1
Assembly level	chromosome	scaffold
Span (Mb)	869.86	870.75
Number of chromosomes	23	N/A
Number of contigs	464	453
Contig N50	4.1 Mb	4.22 Mb
Number of scaffolds	131	107
Scaffold N50	37.75 Mb	37.9 Mb
Longest scaffold length (Mb)	63.64	N/A
Sex chromosomes	Z	N/A
Organelles	Mitochondrial genome: 15.46 kb	N/A

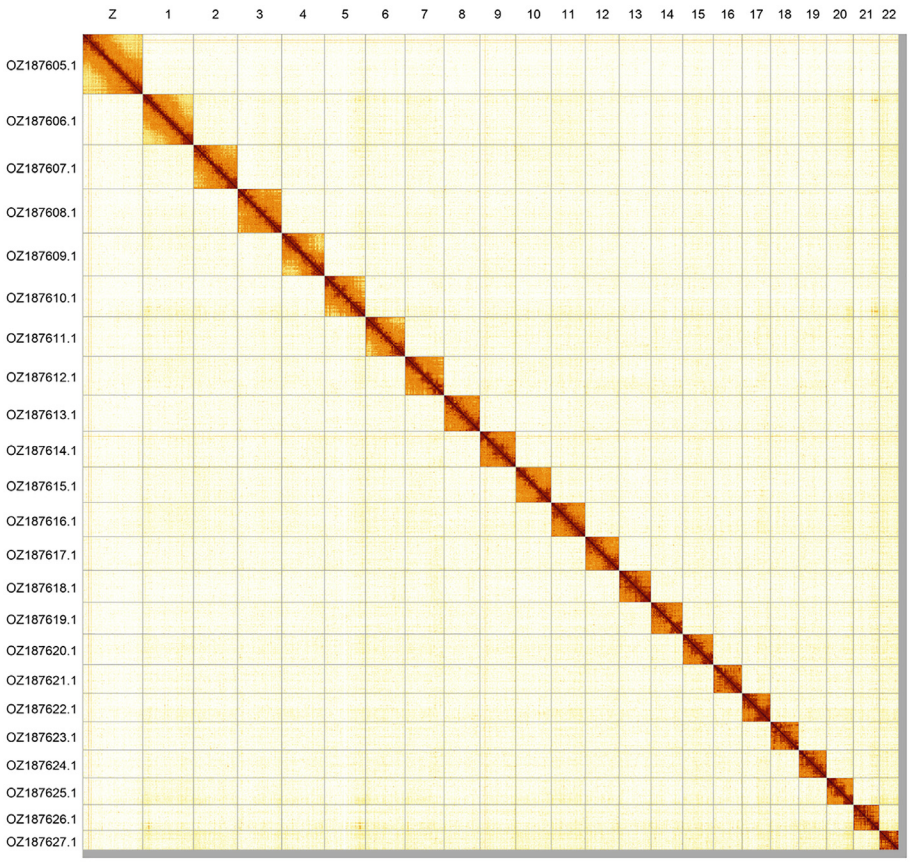

Figure 2. Hi-C contact map of the *Satyrium pruni* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Satyrium pruni* ilSatPrun1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ187606.1	1	53.58	38	M1;M19
OZ187607.1	2	46.56	38	M11;M23
OZ187608.1	3	46.46	38	M12;M29
OZ187609.1	4	45.29	38	M25;M5
OZ187610.1	5	43.14	38	M18;M30
OZ187611.1	6	41.66	38.50	M14;M26
OZ187612.1	7	41.21	38.50	M17;M20
OZ187613.1	8	37.87	38	M9
OZ187614.1	9	37.75	39	M8
OZ187615.1	10	37.66	38	M2
OZ187616.1	11	35.79	38.50	M3
OZ187617.1	12	35.69	38.50	M7
OZ187618.1	13	33.73	38	M16
OZ187619.1	14	33.37	38.50	M22
OZ187620.1	15	32.28	38.50	M6
OZ187621.1	16	30.35	38.50	M21
OZ187622.1	17	30.06	39	M4
OZ187623.1	18	29.73	39	M15
OZ187624.1	19	29.46	39	M10
OZ187625.1	20	28.33	39	M28;M31
OZ187626.1	21	27.25	39	M13
OZ187627.1	22	20.52	39	M27
OZ187605.1	Z	63.64	37.50	M24;MZ
OZ187628.1	MT	0.02	17	N/A

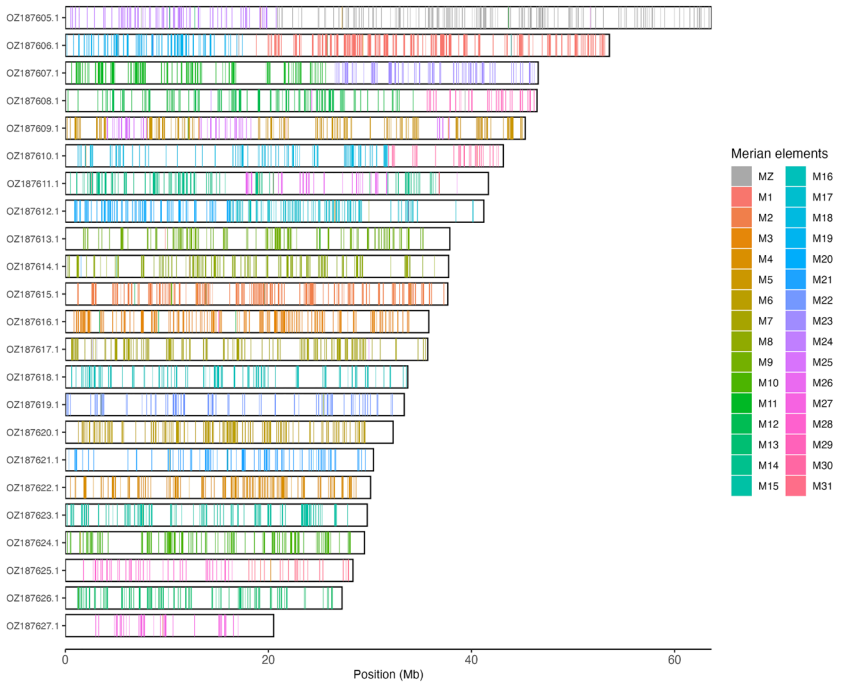

Figure 3. Merian elements painted across chromosomes in the ilSatPrun1.hap1.1 assembly of *Satyrium pruni*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

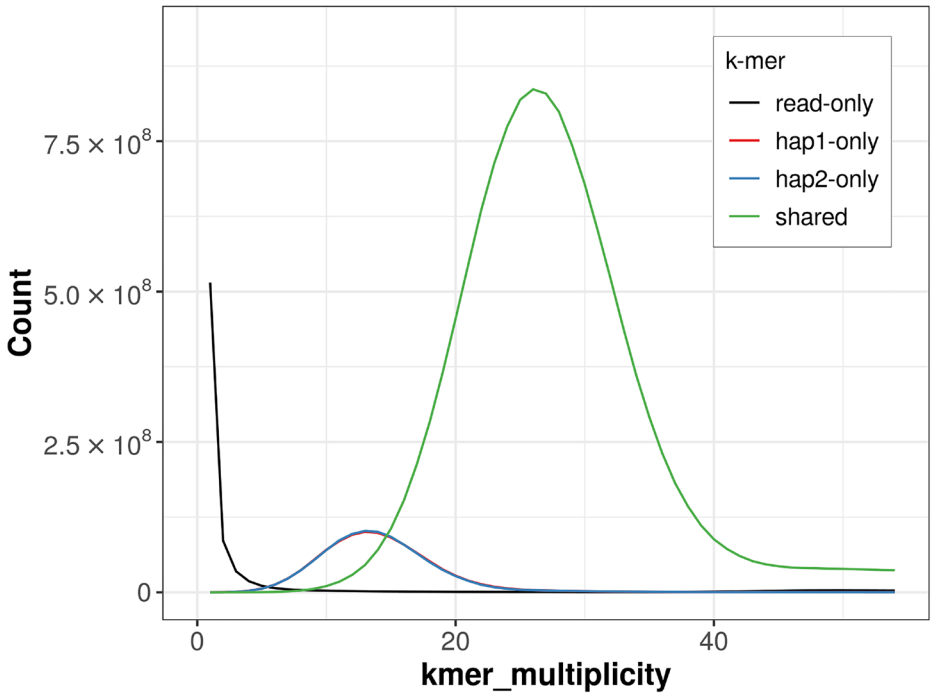

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

gene set (single = 97.2%, duplicated = 0.7%) 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.Q59**, meeting the recommended reference standard.

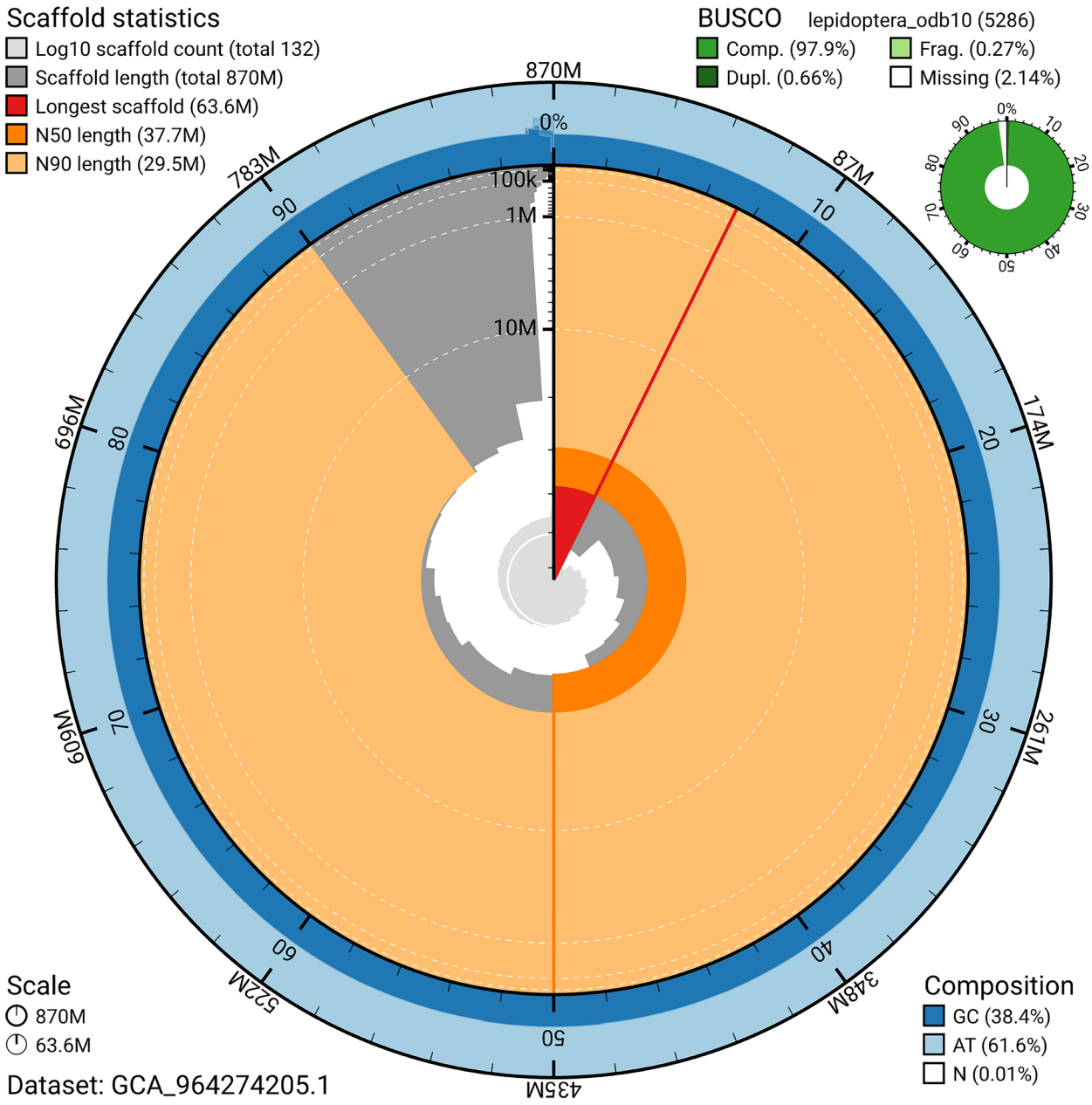

Figure 5. Assembly metrics for ilSatPrun1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

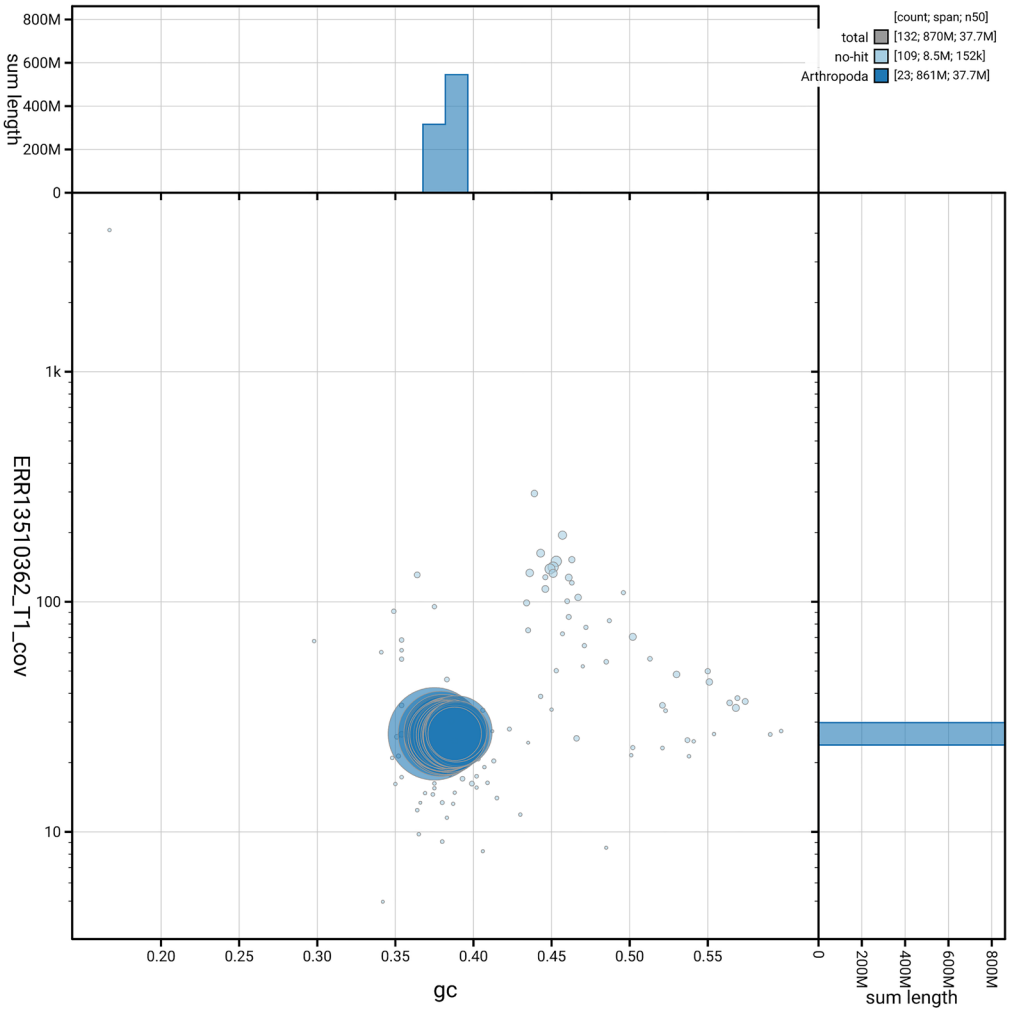

Figure 6. BlobToolKit GC-coverage plot for ilSatPrun1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Satyrium pruni* assembly.

Measure (Benchmark)	Value
EBP summary (haplotype 1)	6.C.Q59
Contig N50 length (≥ 1 Mb)	4.10 Mb
Scaffold N50 length (= chromosome N50)	37.75 Mb
Consensus quality (QV) (≥ 40)	Haplotype 1: 59.6; haplotype 2: 59.7; combined: 59.6
k-mer completeness (≥ 95%)	Haplotype 1: 87.70%; Haplotype 2: 87.71%; combined: 99.16%
BUSCO* (S > 90%; D < 5%)	C:97.9%[S:97.2%,D:0.7%],F:0.3%,M:1.9%,n:5286
Percentage of assembly assigned to chromosomes (≥ 90%)	99.02%

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: Satyrium pruni. Accession number PRJEB79118. The genome sequence is released openly for reuse. The *Satyrium pruni* 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	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
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
lep_busco_painter	1.0.0	https://github.com/charlottewright/lep_busco_painter
MerquryFK	1.1.1	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
PretextSnapshot	N/A	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

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

Butterfly Conservation: **Black hairstreak *Satyrium pruni*.** 2025.
Reference Source

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

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

Fox R, Dennis EB, Brown AF, *et al.*: **A revised red list of British butterflies.** *Insect Conserv Divers.* 2022; **15**(5): 485–95.
Publisher 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, [Accessed 2 April 2024].
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**

Silvonen K, Top-Jensen M, Fibiger M: **Suomen päivä- ja yöperhoset – maastokäsikirja: a field guide to the butterflies and moths of Finland. In Finnish.** Oestermarie: Bugbook Publishing, 2014.
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**

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 08 September 2025

https://doi.org/10.21956/wellcomeopenres.27129.r129025

✔ **Christopher J. Fields**
Carver Biotechnology Center, University of Illinois Urbana-Champaign, Urbana, 61801, USA

The report for the Black Hairstreak genome is quite comprehensive and is fairly easy to follow. I also particularly like the application of chromosome painting in this particular report, which is a nice unique analysis for this particular species adding some useful additional biological context. Also happy to see that haplotype-phased assembly was pursued; past contributions seemed to focus on only the partially phased primary assembly.

The only small concern is from some confusion regarding specifically which assembly was used for HiC scaffolding (or whether both haplotype assemblies were used). In this section (pg 4, under 'Genome assembly'):

The HiFi reads were assembled using Hifiasm in Hi-C phasing mode (Chenget 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

In the above and in the full results it's pretty obvious two haplotypes were generated as the final products, but the confusing part is this: 'Hi-C reads (Rao et al., 2014) were mapped to the **primary contigs** using bwa-mem2 (Vasimuddin et al., 2019).' To me this would mean the pre-phased assembly, the 'p_ctg' file from hifiasm, which I have seen reported in other Wellcome papers. I'm assuming this is supposed to be hap1? Or were HiC reads aligned to both and filtered (ala HapHiC) to scaffold both? Would be great to clarify this.

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: Computational 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 04 September 2025

https://doi.org/10.21956/wellcomeopenres.27129.r131172

Frederik Hendrickx
Royal Belgian institute of natural Sciences, Brussels, Belgium

This manuscript describes the genome sequence of the Black Hairstreak (*Satyrium pruni)*, a widespread though endangered butterfly from Europe. The genome, assembled from a single male using both PacBio HiFi and Hi-C data is of very high quality and contiguity. Based on the Hi-C contact map, 99% of the total contig length nicely assembles into 23 pseudomolecules, illustrating that the assembly is of very high quality. The homology with the Merian elements adds additional information of the quality and correctness of the assembly. As a minor point, I would suggest that the authors briefly explain what Merian elements are as it is a rather specialized term that is only used in butterfly genomics. At least reference to a paper that explains this term should be added. I have no further suggestions for improvement.

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: Insect evolutionary 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.