DATA NOTE

The genome sequence of the Woodland Grayling, *Hipparchia fagi* (Scopoli, 1763) (Lepidoptera: Nymphalidae)

[version 1; peer review: 2 approved]

Roger Vila[1], Alena Sucháčková Bartoňová[2,3], Charlotte J. Wright[4], Joana I. Meier[4], Mark L. Blaxter[4], 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]Institut de Biologia Evolutiva, CSIC - Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
[2]Biology Centre of the Academy of Sciences of the Czech Republic, České Budějovice, Czech Republic
[3]Senckenberg German Entomological Institute, Müncheberg, Germany
[4]Tree of Life Programme, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 16 Feb 2026, **11**:126
https://doi.org/10.12688/wellcomeopenres.25873.1
Latest published: 16 Feb 2026, **11**:126
https://doi.org/10.12688/wellcomeopenres.25873.1

Abstract

We present a genome assembly from a male specimen of *Hipparchia fagi* (Woodland Grayling; Arthropoda; Insecta; Lepidoptera; Nymphalidae). The assembly contains two haplotypes with total lengths of 380.62 megabases and 381.27 megabases. Most of haplotype 1 (99.92%) is scaffolded into 29 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 15.47 kilobases. This work is part of Project Psyche, a collaborative programme generating genomes for European butterflies and moths.

Keywords

Hipparchia fagi; Woodland Grayling; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 16 Feb 2026	✓ view	✓ view

1. **Vlad Eugen Dinca**, University of Oulu Centre of Excellence in Cell-Extracellular Matrix Research, Oulu, Finland

2. **Jose Martinez**, University of Florida, Gainesville, 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: 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 funded by VEDA FELLOWSHIPS CZ.02.01.01/00/22_010/0008117. RV was supported by grant PID2022-139689NB-I00 (MICIU/ AEI/ 10.13039/501100011033 and ERDF, EU) and by 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.

How to cite this article: Vila R, Sucháčková Bartoňová A, Wright CJ *et al.* **The genome sequence of the Woodland Grayling, *Hipparchia fagi* (Scopoli, 1763) (Lepidoptera: Nymphalidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2026, **11**:126 https://doi.org/10.12688/wellcomeopenres.25873.1

First published: 16 Feb 2026, **11**:126 https://doi.org/10.12688/wellcomeopenres.25873.1

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 fagi* (Scopoli, 1763) (NCBI:txid111910)

Background

Hipparchia fagi (Scopoli, 1763), the Woodland Grayling, is a Ponto-Mediterranean species of wooded dry grasslands in low and mid elevations. It is a large dark grey Nymphalid butterfly with a yellowish band across both forewing and hindwing. Its forewing length is 29–40 mm (Middleton-Welling *et al.*, 2020). It closely resembles *H. hermione* and *H. syriaca*, and it can be sympatric with these species (Dinca *et al.*, 2009; Pinzari & Sbordoni, 2013).

It is distributed through the western Palearctic – from northern Spain (Cantabrian Mountains) across southern to western Kazakhstan and the southern Urals. It reaches the northern limits of its range in southern Germany and southern Czech Republic. It is absent from the eastern Mediterranean including Asian Turkey. It has one generation per year, with adults starting to fly in early June in southern Europe, diapause in summer and re-appear again in September (Lepiforum eV, 2025). In central Europe, *H. fagi* flies from late June to September. On hot days, the butterflies rest in the canopy or on the tree trunks. The habitat is different types of wooded dry grasslands – margins of pine and oak woodlands, grasslands on sandy soils, steppes with scattered bushes or trees, rocky places, and coppiced woodlands. Males use the perching strategy to meet females, and scent stimuli from androconia on males' wings are involved in mate communication (Pinzari & Sbordoni, 2013).

Females lay eggs on open areas with sparse vegetation, on grass tufts or dry plant material (Mollenbeck *et al.*, 2009). Larvae feed on grasses such as *Bromus erectus*, *Festuca rubra* or *Brachypodium pinnatum* (Tolman & Lewington, 2008). The third instar larva overwinters, and it continues to feed in the spring during the nights, and it pupates in a chamber in the soil or grass tuft (Benes *et al.*, 2002; Lepiforum eV, 2025). The larval microhabitat is characterised by bare ground, rocks, stones and gravel, litter and tall and sturdy grass tufts.

H. fagi is considered of "Least Concern" in the European Red List of butterflies (van Swaay *et al.*, 2025). Although its situation is probably stable in southern Europe, it is disappearing from many localities north of the Alps due to land use abandonment and subsequent vegetation overgrowth, as the larvae are dependent on disturbed habitats (Benes *et al.*, 2002; Lepiforum eV, 2025).

We present a chromosome-level, haplotype-resolved genome sequence for *H. fagi*, sequenced as part of Project Psyche. The sequence data were derived from a male specimen (Figure 1) collected from Turó de Tagamanent, Barcelona, Catalonia, Spain.

Figure 1. Voucher photographs of the *Hipparchia fagi* (ilHipFagi1) specimen used for genome sequencing.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult male *Hipparchia fagi* (specimen ID SAN28000310, ToLID ilHipFagi1; Figure 1), collected from Turó de Tagamanent, Barcelona, Catalonia, Spain (latitude 41.7402, longitude 2.2932) on 2018-09-11. 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 ilHipFagi1 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.

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 30.28 ng/μL and a yield of 1 423.16 ng, with a fragment size of 14.0 kb. The 260/280 spectrophotometric ratio was 1.91, and the 260/230 ratio was 2.06.

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 ilHipFagi1 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).

Table 1. Specimen and sequencing data for BioProject PRJEB82274.

Platform	PacBio HiFi	Hi-C
ToLID	ilHipFagi1	ilHipFagi1
Specimen ID	SAN28000310	SAN28000310
BioSample (source individual)	SAMEA115768738	SAMEA115768738
BioSample (tissue)	SAMEA115768882	SAMEA115768882
Tissue	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13946498	ERR13947526
Read count total	2.34 million	606.24 million
Base count total	25.03 Gb	91.54 Gb

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

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

Genome assembly

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

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

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

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 12 breaks, 27 joins, and removal of 16 haplotypic duplications. This reduced the scaffold count by 41.7%. The curation process is described at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

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

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

We used lep_busco_painter 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 fagi* specimen generated 25.03 Gb (gigabases) from 2.34 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 384.90 Mb, with a heterozygosity of 1.17% and repeat content of 26.77% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 63× coverage. Hi-C sequencing produced 91.54 Gb from 606.24 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 380.62 Mb in 41 scaffolds, with 81 gaps, and a scaffold N50 of 14.35 Mb (Table 2).

Most of the assembly sequence (99.92%) was assigned to 29 chromosomal-level scaffolds, representing 28 autosomes and the Z sex chromosome. 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 Z chromosome was assigned based on Merian elements. During

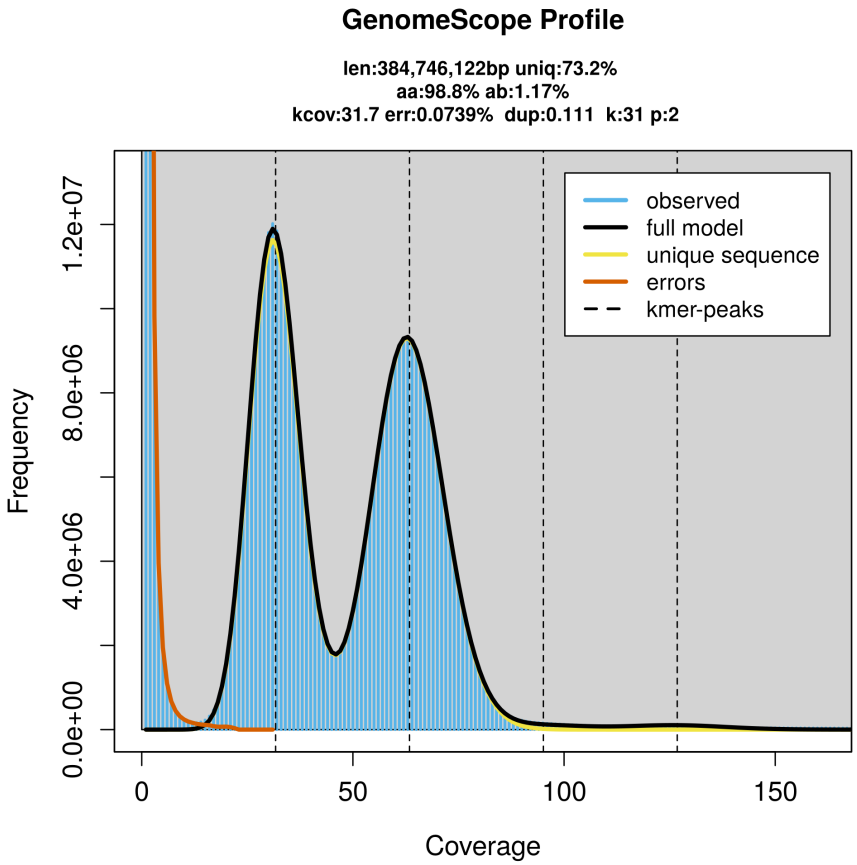

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 2. Genome assembly statistics.

Assembly name	ilHipFagi1.hap1.1	ilHipFagi1.hap2.1
Assembly accession	GCA_964396745.1	GCA_964396605.1
Assembly level	chromosome	scaffold
Span (Mb)	380.62	381.27
Number of chromosomes	29	0
Number of contigs	122	113
Contig N50	6.92 Mb	6.71 Mb
Number of scaffolds	41	45
Scaffold N50	14.35 Mb	14.61 Mb
Longest scaffold length (Mb)	17.18	-
Sex chromosomes	Z	-
Organelles	Mitochondrion: 15.47 kb	-

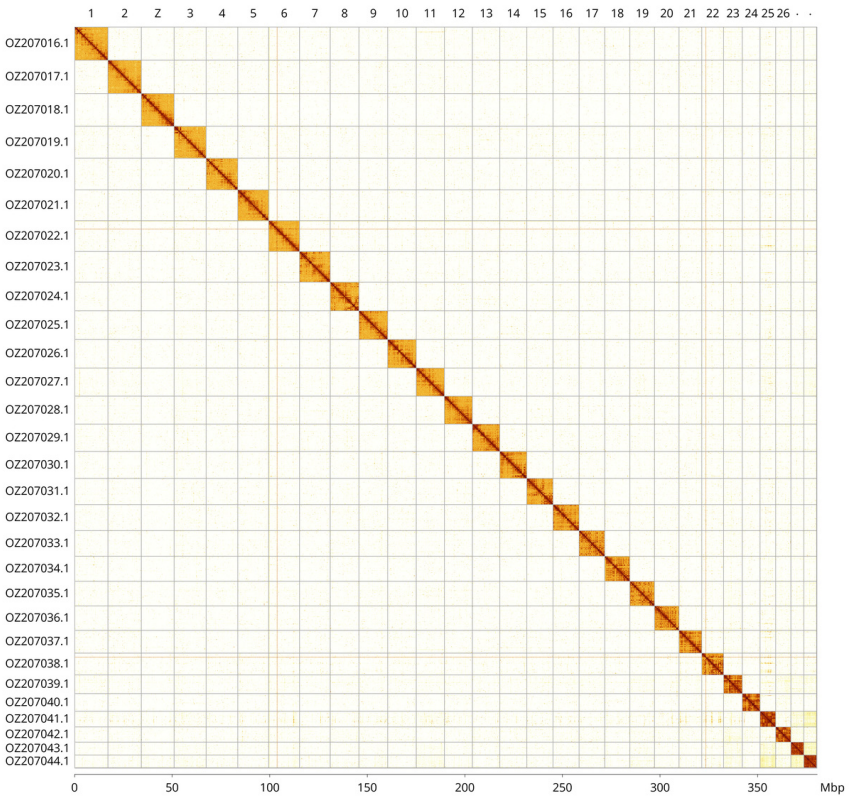

Figure 3. Hi-C contact map of the *Hipparchia fagi* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Hipparchia fagi* ilHipFagi1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ207016.1	1	17.18	37	M1
OZ207017.1	2	17.11	36.50	M2
OZ207019.1	3	16.46	36.50	M17;M20
OZ207020.1	4	16.19	37	M3
OZ207021.1	5	15.85	36.50	M8
OZ207022.1	6	15.79	36.50	M19;M26
OZ207023.1	7	15.70	36.50	M9
OZ207024.1	8	14.74	36.50	M18
OZ207025.1	9	14.72	36.50	M5
OZ207026.1	10	14.61	36.50	M12
OZ207027.1	11	14.48	37	M14;M29
OZ207028.1	12	14.35	37	M7
OZ207029.1	13	14.04	36.50	M16
OZ207030.1	14	13.80	37	M4
OZ207031.1	15	13.45	37	M6
OZ207032.1	16	13.40	36.50	M21
OZ207033.1	17	13.10	37	M22
OZ207034.1	18	12.87	37	M15
OZ207035.1	19	12.64	37.50	M11
OZ207036.1	20	12.56	37	M10
OZ207037.1	21	11.61	37	M13
OZ207038.1	22	11.23	37.50	M23
OZ207039.1	23	9.60	38	M24
OZ207040.1	24	9.07	37.50	M28
OZ207041.1	25	8.04	39	M30
OZ207042.1	26	7.79	37.50	M27
OZ207043.1	27	6.63	38	M25
OZ207044.1	28	6.55	38	M31
OZ207018.1	Z	16.75	36.50	MZ

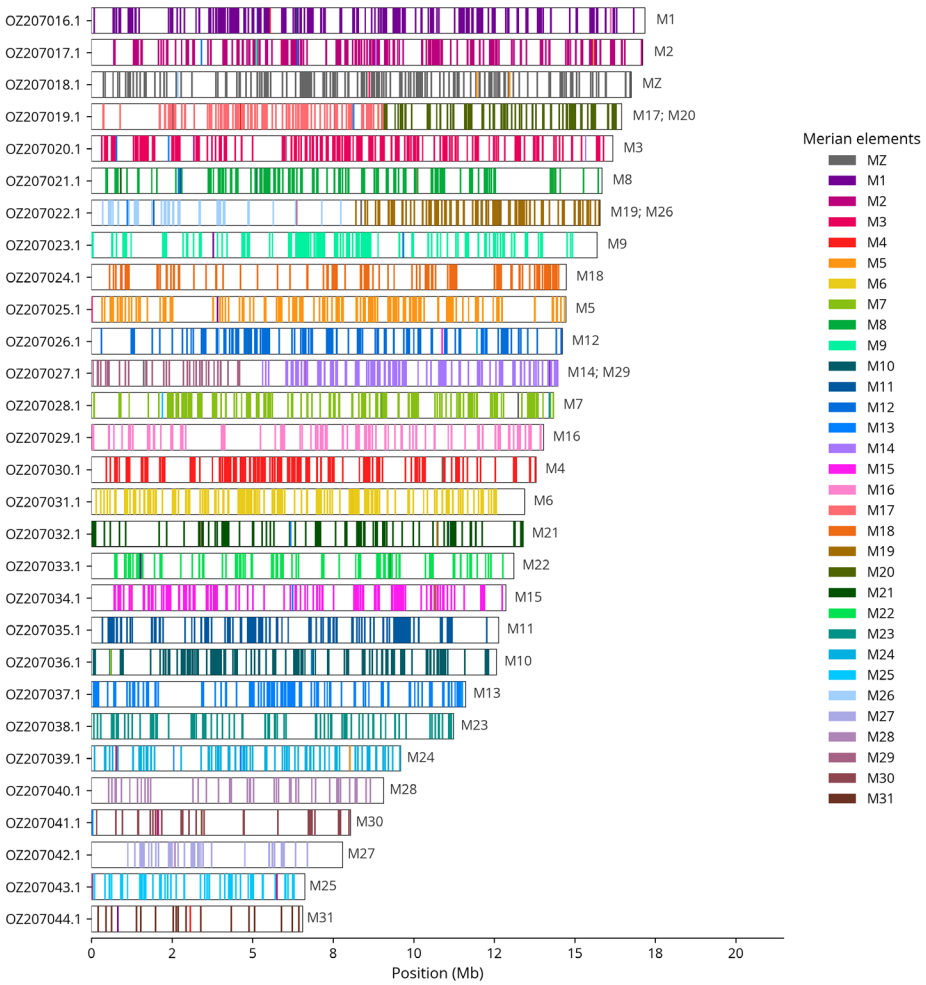

Figure 4. Merian elements painted across chromosomes in the ilHipFagi1.hap1.1 assembly of *Hipparchia fagi*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

curation, we noted that scaffolds on chromosome 23 in the region ~2.79-3.35 Mb have uncertain order and orientation.

The mitochondrial genome was also assembled (length 15.47 kb, OZ207045.1). 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, 66.0. When the two haplotypes are combined, the assembly achieves an estimated QV of 66.1. The *k*-mer completeness is 76.57% for haplotype 1, 76.58% for haplotype 2, and 99.37% 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 98.6% of the expected gene set (single = 98.1%, duplicated = 0.4%). The snail plot in Figure 6 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 7 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.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 themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of

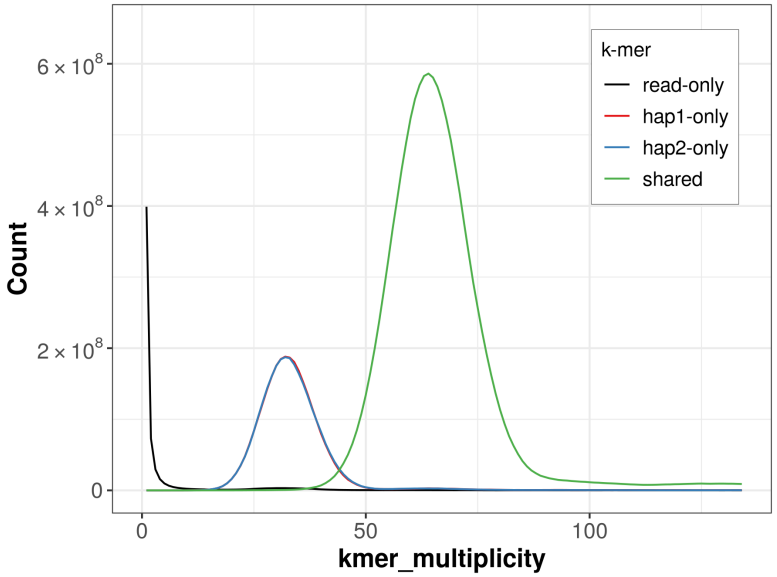

Figure 5. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

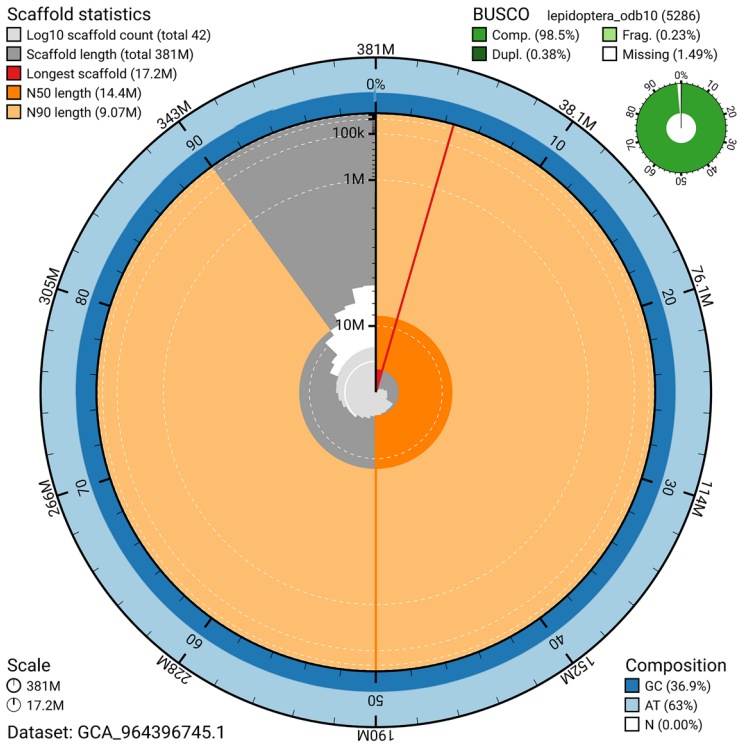

Figure 6. Assembly metrics for ilHipFagi1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

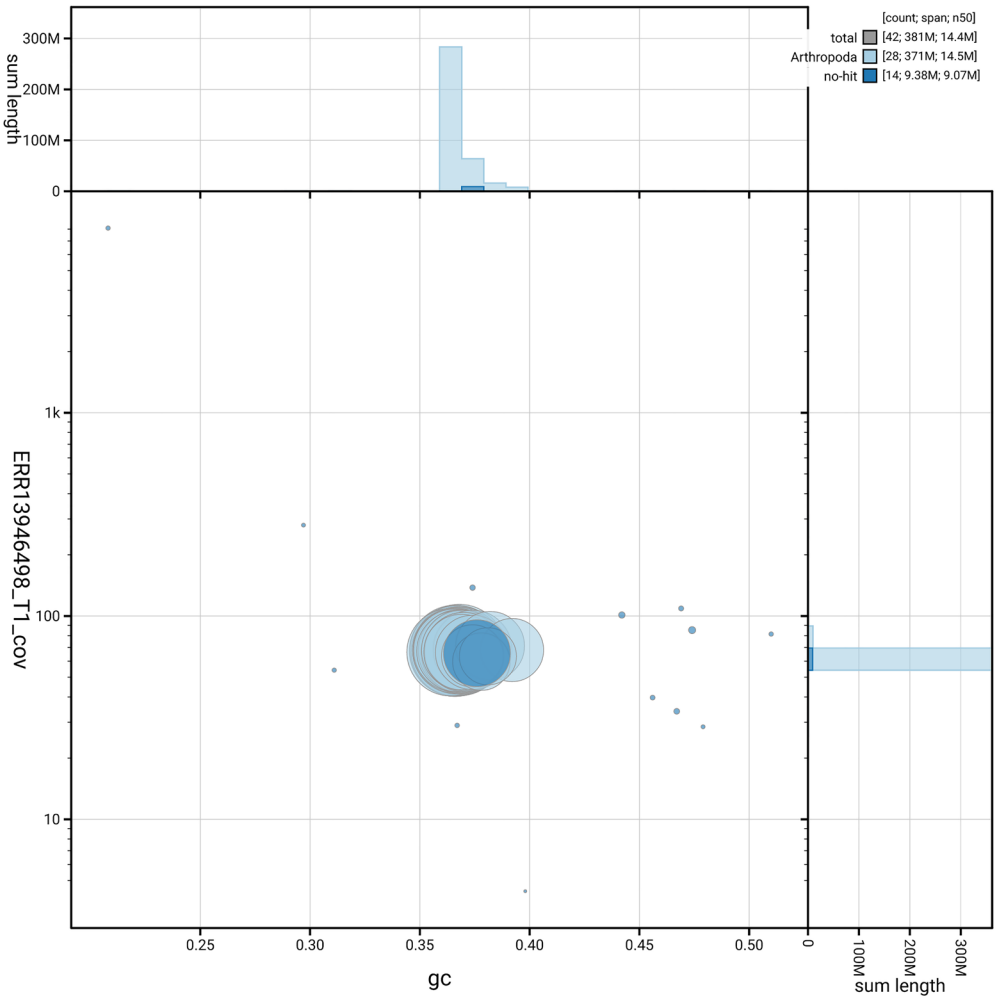

Figure 7. BlobToolKit GC-coverage plot for ilHipFagi1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 fagi* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q66	6.C.Q40
Contig N50 length	6.92 Mb	≥ 1 Mb
Scaffold N50 length	14.35 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 66.3; haplotype 2: 66.0; combined: 66.1	≥ 40
k-mer completeness	Haplotype 1: 76.57%; Haplotype 2: 76.58%; combined: 99.37%	≥ 95%
BUSCO	C:98.5% [S:98.1%; D:0.4%]; F:0.2%; M:1.3%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.92%	≥ 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

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 fagi (woodland grayling). Accession number PRJEB82274. The genome sequence is released openly for reuse. The *Hipparchia fagi* 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/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
lep_busco_painter	1.0.0	https://github.com/charlottewright/lep_busco_painter
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14, 1.17 and 1.18	https://github.com/MultiQC/MultiQC

Software	Version	Source
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.6.0	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

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

Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–10.
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

Benes J, Konvicka M, Dvorak J, *et al.*: **Motyli Ceske Republiky: Rozsireni a Ochrana i, II.** Spolecnost pro ochranu motylu, 2002.
Reference Source

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

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

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

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

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

Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.
PubMed Abstract | Publisher Full Text

Dinca V, Cuvelier S, Szekely L, *et al.*: **New data on the Rhopalocera (Lepidoptera) of Dobrogea (south-eastern Romania).** *Phegea.* 2009; **37**(1): 1–21.
Reference Source

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): 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

Lepiforum eV: ***Hipparchia fagi* (Scopoli, 1763).** 2025.
Reference Source

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text | Free Full Text

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | Publisher Full Text | Free Full Text

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2.
Reference Source

Middleton-Welling J, Dapporto L, Garcia-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

Mollenbeck V, Hermann G, Fartmann T: **Does prescribed burning mean a threat to the rare satyrine butterfly Hipparchia fagi? Larval-habitat**

preferences give the answer. *J Insect Conserv.* 2009; **13**(1): 77–87.
Publisher Full Text

Pinzari M, Sbordoni V: **Species and mate recognition in two sympatric grayling butterflies: *Hipparchia fagi* and *H. hermione genava* (Lepidoptera).** *Ethol Ecol Evol.* 2013; **25**(1): 28–51.
Publisher 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.** HarperCollins, 2008.
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, Ellis S, Warren M: ***Leptotes pirithous* (Europe assessment).** 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

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 30 April 2026

https://doi.org/10.21956/wellcomeopenres.28477.r151655

✔ **Jose Martinez**
University of Florida, Gainesville, Florida, USA

Hipparchia fagi (Scopoli, 1763), commonly known as the Woodland Grayling, has been the subject of a recent genome assembly by Vila et al. This assembly shows a high degree of completeness, reflected in a BUSCO score of 98.6%, and is structured into 29 chromosomal pseudomolecules, including a fully assembled Z sex chromosome. The use of well-established and rigorous methodologies has produced a dependable and high-quality genomic resource. The value of this dataset is further enhanced by its clear organization and comprehensive documentation. The purpose of the dataset is explicitly stated, and the methods and materials are described in sufficient detail to ensure reproducibility. Additionally, the data are standardized, supported by clear visual representations, and made readily accessible, making this resource broadly applicable for diverse research purposes.

In addition to its intrinsic scientific value, the genome of *H. fagi* holds notable practical significance. While the preservation of natural ecosystems remains a central priority in conservation, this species is particularly sensitive to land-use abandonment and the subsequent encroachment of vegetation, which can alter or degrade its habitat. In this context, the availability of a high-quality genomic resource provides an important foundation for both fundamental and applied research. It enables a deeper understanding of the species' biology, population dynamics, and adaptive potential, thereby supporting the development of more informed and effective conservation and management strategies.

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: Evolutionary Biology, Ecology, IPM, Integrative Biology, Entomology, 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 15 April 2026

https://doi.org/10.21956/wellcomeopenres.28477.r152251

Vlad Eugen Dinca
University of Oulu Centre of Excellence in Cell-Extracellular Matrix Research, Oulu, Northern Ostrobothnia, Finland

In this note, the authors have successfully generated a chromosome-level genome assembly of a male *Hipparchia fagi* (Lepidoptera: Nymphalidae). The mitochondrial genome has been assembled as well.

Although I am not an expert in several of the methodological approaches used, as far as I can tell, the manuscript is technically sound and uses methodologies and pipelines that are fairly well-established through use by the Wellcome Sanger Institute.

The genome appears to be of high quality and scaffolded into 29 chromosomal pseudomolecules (including the Z sex chromosome).

A few additional comments are included below.

You are mentioning that it closely resembles other *Hipparchia species*. It would be good to briefly mention whether it can be reliably distinguished from those based on external characters, genitalia and/or DNA barcodes.

“from northern Spain (Cantabrian Mountains) across southern to western Kazakhstan and the southern Urals.” – In northern Spain, this species occurs outside the Cantabrian Mts. as well (e.g. the sequenced specimen originates from NE Spain). Perhaps this sentence could be slightly rephrased. A suggestion:

"….. from northern Spain across southern Europe to the southern Urals and western Kazakhstan."

"It has one generation per year, with adults starting to fly in early June in southern Europe, diapause in summer and re-appear again in September"- Slight rephrasing suggested: "It has one generation per year, with adults starting to fly in early June in southern Europe, diapausing in summer and re-appearing again in September"

Figure 1 – The upper image shows forewings (ventral) and hindwings (dorsal and upside down); the lower image shows the forewings (dorsal) and hindwings (dorsal and upside down). This should ideally be corrected to first show all wings dorsally and in the correct position, followed by all wings shown ventrally in the correct position.

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: Lepidoptera speciation, phylogeography, conservation.

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.