DATA NOTE

The genome sequence of the Sage Skipper, *Muschampia proto* (Ochsenheimer, 1808) (Lepidoptera: Hesperiidae)

[version 1; peer review: 3 approved, 1 approved with reservations]

Mattia Menchetti[1], Joan Carles Hinojosa[1], Lucas Kaminski[2], Roger Vila[1], Charlotte J. Wright[3], Joana I. Meier[3], Mark L. Blaxter[3], Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team, Wellcome Sanger Institute Scientific Operations: Sequencing Operations, Wellcome Sanger Institute Tree of Life Core Informatics team, Tree of Life Core Informatics collective, Project Psyche Community

[1]Institut de Biologia Evolutiva (CSIC-UPF), Barcelona, Catalonia, Spain
[2]Universidade Federal de Alagoas, Maceió, State of Alagoas, Brazil
[3]Tree of Life Programme, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 24 Jan 2026, **11**:74
https://doi.org/10.12688/wellcomeopenres.25829.1
Latest published: 24 Jan 2026, **11**:74
https://doi.org/10.12688/wellcomeopenres.25829.1

Abstract

We present a genome assembly from a male specimen of *Muschampia proto* (Sage Skipper; Arthropoda; Insecta; Lepidoptera; Hesperiidae). The assembly contains two haplotypes with total lengths of 542.52 megabases and 541.58 megabases. Most of haplotype 1 (99.67%) is scaffolded into 30 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 15.37 kilobases. This work is part of Project Psyche, a collaborative programme generating genomes for European butterflies and moths.

Keywords

Muschampia proto; Sage Skipper; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ✓ ?

	1	2	3	4
version 1 24 Jan 2026	✓ view	✓ view	✓ view	? view

1. **Temitope Opeyemi Oriowo**, Leibniz Institute for the Analysis of Biodiveristy Change, Museum Koenig Bonn, Bonn, Germany
2. **Matthew Merkin**, University of Liverpool, Liverpool, UK
3. **Vadim Pisarenco**, Universitat of Barcelona, Barcelona, Spain
4. **Yukun Sun**, The Field Museum, Chicago, 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: Menchetti M: Investigation, Resources, Writing – Original Draft Preparation; **Hinojosa JC**: Investigation, Resources; **Kaminski L**: Investigation, Resources; **Vila R**: Investigation, Resources, Supervision, Writing – Review & Editing; **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). 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.

How to cite this article: Menchetti M, Hinojosa JC, Kaminski L *et al.* **The genome sequence of the Sage Skipper, *Muschampia proto* (Ochsenheimer, 1808) (Lepidoptera: Hesperiidae) [version 1; peer review: 3 approved, 1 approved with reservations]** Wellcome Open Research 2026, **11**:74 https://doi.org/10.12688/wellcomeopenres.25829.1

First published: 24 Jan 2026, **11**:74 https://doi.org/10.12688/wellcomeopenres.25829.1

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; Hesperioidea; Hesperiidae; Pyrginae; *Muschampia*; *Muschampia proto* (Ochsenheimer, 1808) (NCBI:txid1107688)

Background

The Sage Skipper, *Muschampia proto* (Ochsenheimer, 1808), is a small butterfly species of the family Hesperiidae found in the Iberian Peninsula, southern France, and North Africa (Morocco and Algeria) (Hinojosa *et al.*, 2021). The species has xerophilic tendencies and is often found in association with broadly Mediterranean vegetation in open habitats (García-Barros *et al.*, 2013). It typically occurs at elevations of 200–1 500 m, but has been recorded from 0–1 800 m.

Larvae feed exclusively on *Phlomis* (Lamiaceae). Initially, they construct shelters by binding the edges of leaves with silk and consuming those leaves. As they develop, they move to inflorescences and feed on them (García-Barros *et al.*, 2013). It overwinters at either the egg or the larval stage (García-Barros *et al.*, 2013).

Muschampia proto is univoltine (Fernández-Rubio, 1991) and typically flies between April and October (García-Barros *et al.*, 2013), although adults are rarely observed before June. A partial second generation has been suggested for Catalonia (Spain) (Vila *et al.*, 2018).

Muschampia proto is the type species of the genus, and is part of a complex of cryptic species that started to diverge about 2 million years ago (Hinojosa *et al.*, 2021): *Muschampia alta*, occurring in southern Italy and the Balkan Peninsula, and *Muschampia proteides*, present in the eastern Europe, the Near East and Iran. These species were, until recently, considered subspecies of *M. proto* (Hinojosa *et al.*, 2021). For most of its range, *M. proto* shows a low mitochondrial diversity without spatial structure (Dapporto *et al.*, 2022). Nevertheless, the single specimen sequenced from Algeria is remarkably divergent from other sampled individuals. The species is listed as Least Concern in the IUCN Red List of Europe (van Swaay *et al.*, 2025).

Muschampia proto has a karyotype of 30 chromosomes (De Lesse, 1970). The first reference genome for *M. proto* presented here offers an valuable resource for investigating the origins of cryptic diversity and the speciation process of *M. proto* and related taxa. The sequence data were derived from a male specimen (Figure 1) collected from Orcera, Jaén, Andalusia, Spain.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult male *Muschampia proto* (specimen ID SAN28000344, ToLID ilMusProo1; Figure 1), collected from Orcera, Jaén, Andalusia, Spain (latitude 38.3091, longitude –2.5825) on 2019-06-06. The specimen was collected by Roger Vila, Mattia Menchetti, Joan Carles Hinojosa, Lucas Kaminski and identified by Roger Vila.

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 ilMusProo1 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

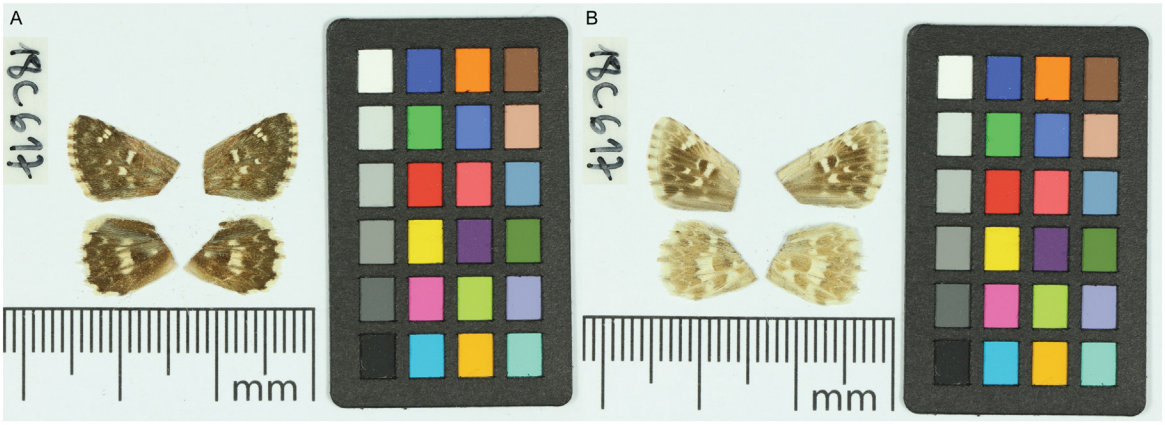

Figure 1. Voucher photographs of the *Muschampia proto* (ilMusProo1) specimen used for genome sequencing.

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 68.86 ng/μL and a yield of 3 236.42 ng, with a fragment size of 15.2 kb.

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 ilMusProo1 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

Table 1. Specimen and sequencing data for BioProject PRJEB81610.

Platform	PacBio HiFi	Hi-C
ToLID	ilMusProo1	ilMusProo1
Specimen ID	SAN28000344	SAN28000344
BioSample (source individual)	SAMEA115769899	SAMEA115769899
BioSample (tissue)	SAMEA115769919	SAMEA115769919
Tissue	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13900447	ERR13907225
Read count total	3.88 million	859.80 million
Base count total	36.91 Gb	129.83 Gb

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 curation reduced the scaffold count by 18.3%. 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 painted Merian elements, which represent the 32 ancestral linkage groups in Lepidoptera (Wright *et al.*, 2024), along chromosomes using lep_busco_painter. Painting used BUSCO gene locations from the lepidoptera_odb10 set (Kriventseva *et al.*, 2019) and chromosome lengths from NCBI Datasets. Each complete BUSCO (including both single-copy and duplicated BUSCOs) was assigned to a Merian element using a reference database. Positions were coloured and plotted along chromosomes drawn to scale.

Genome sequence report

Sequence data

PacBio sequencing of the *Muschampia proto* specimen generated 36.91 Gb (gigabases) from 3.88 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 530.91 Mb, with a heterozygosity of 1.93% and repeat content of 23.81% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 66× coverage. Hi-C sequencing produced 129.83 Gb from 859.80 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 542.52 Mb in 84 scaffolds, with 88 gaps, and a scaffold N50 of 19.8 Mb (Table 2).

Most of the assembly sequence (99.67%) was assigned to 30 chromosomal-level scaffolds, representing 29 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). Chromosome Z was identified by synteny to GCA_905147235.1.

The mitochondrial genome was also assembled (length 15.37 kb, OZ200009.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

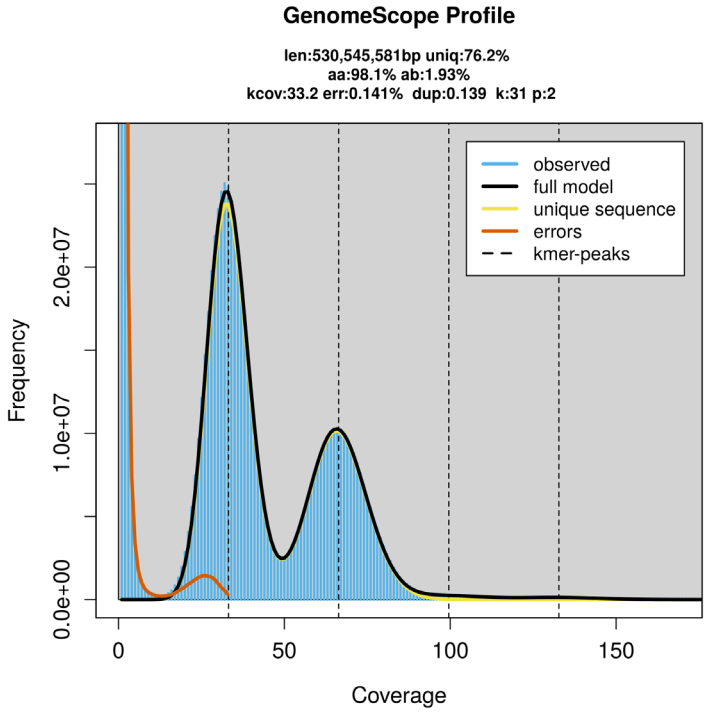

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 2. Genome assembly statistics.

Assembly name	ilMusProo1.hap1.1	ilMusProo1.hap2.1
Assembly accession	GCA_964304755.1	GCA_964304745.1
Assembly level	chromosome	scaffold
Span (Mb)	542.52	541.58
Number of chromosomes	30	scaffold-level
Number of contigs	172	164
Contig N50	7.67 Mb	7.77 Mb
Number of scaffolds	84	77
Scaffold N50	19.8 Mb	19.78 Mb
Longest scaffold length (Mb)	40.53	-
Sex chromosomes	Z	-
Organelles	Mitochondrion: 15.37 kb	-

Assembly quality metrics

For haplotype 1, the estimated QV is 65.0, and for haplotype 2, 65.2. When the two haplotypes are combined, the assembly achieves an estimated QV of 65.1. The *k*-mer completeness is 68.93% for haplotype 1, 68.89% for haplotype 2, and 99.78% 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 = 98.0%, duplicated = 0.5%) in haplotype 1. For haplotype 2, BUSCO analysis identified 98.6% of the expected gene set

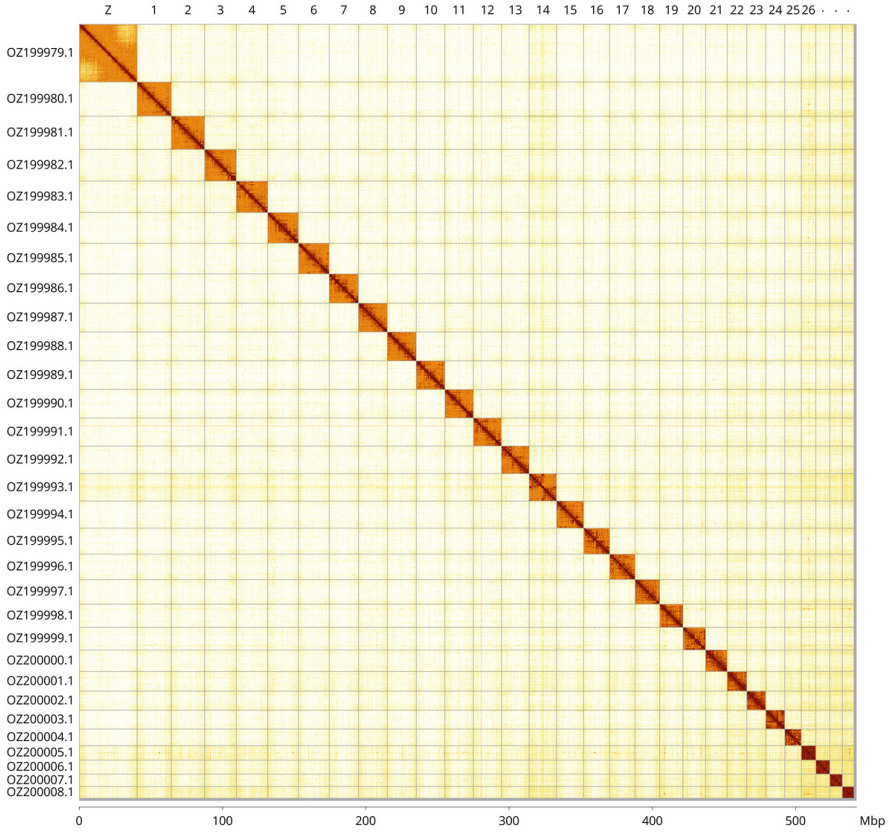

Figure 3. Hi-C contact map of the *Muschampia proto* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Muschampia proto* ilMusProo1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ199980.1	1	23.81	35	M14;M2
OZ199981.1	2	23.21	35	M1
OZ199982.1	3	22.09	35	M17;M20
OZ199983.1	4	22.03	35	M3
OZ199984.1	5	21.47	35	M8
OZ199985.1	6	21.34	34.50	M9
OZ199986.1	7	20.46	34.50	M12
OZ199987.1	8	20.23	35	M7
OZ199988.1	9	20.13	35	M5
OZ199989.1	10	20	34.50	M18
OZ199990.1	11	19.80	35	M16
OZ199991.1	12	19.62	35	M4
OZ199992.1	13	19.33	35	M6
OZ199993.1	14	19.21	36	M22
OZ199994.1	15	18.75	35	M21
OZ199995.1	16	18.21	35	M15
OZ199996.1	17	17.78	35.50	M11
OZ199997.1	18	17.18	35	M10
OZ199998.1	19	16.18	35	M13
OZ199999.1	20	15.85	35.50	M23
OZ200000.1	21	15.07	35.50	M14
OZ200001.1	22	13.71	35.50	M26
OZ200002.1	23	13.24	36	M19
OZ200003.1	24	13.15	35.50	M28
OZ200004.1	25	11.66	35.50	M27
OZ200005.1	26	10.02	38.50	M30
OZ200006.1	27	9.83	36.50	M25
OZ200007.1	28	8.73	36.50	M29
OZ200008.1	29	8.12	37	M31
OZ199979.1	Z	40.53	34.50	M24;MZ

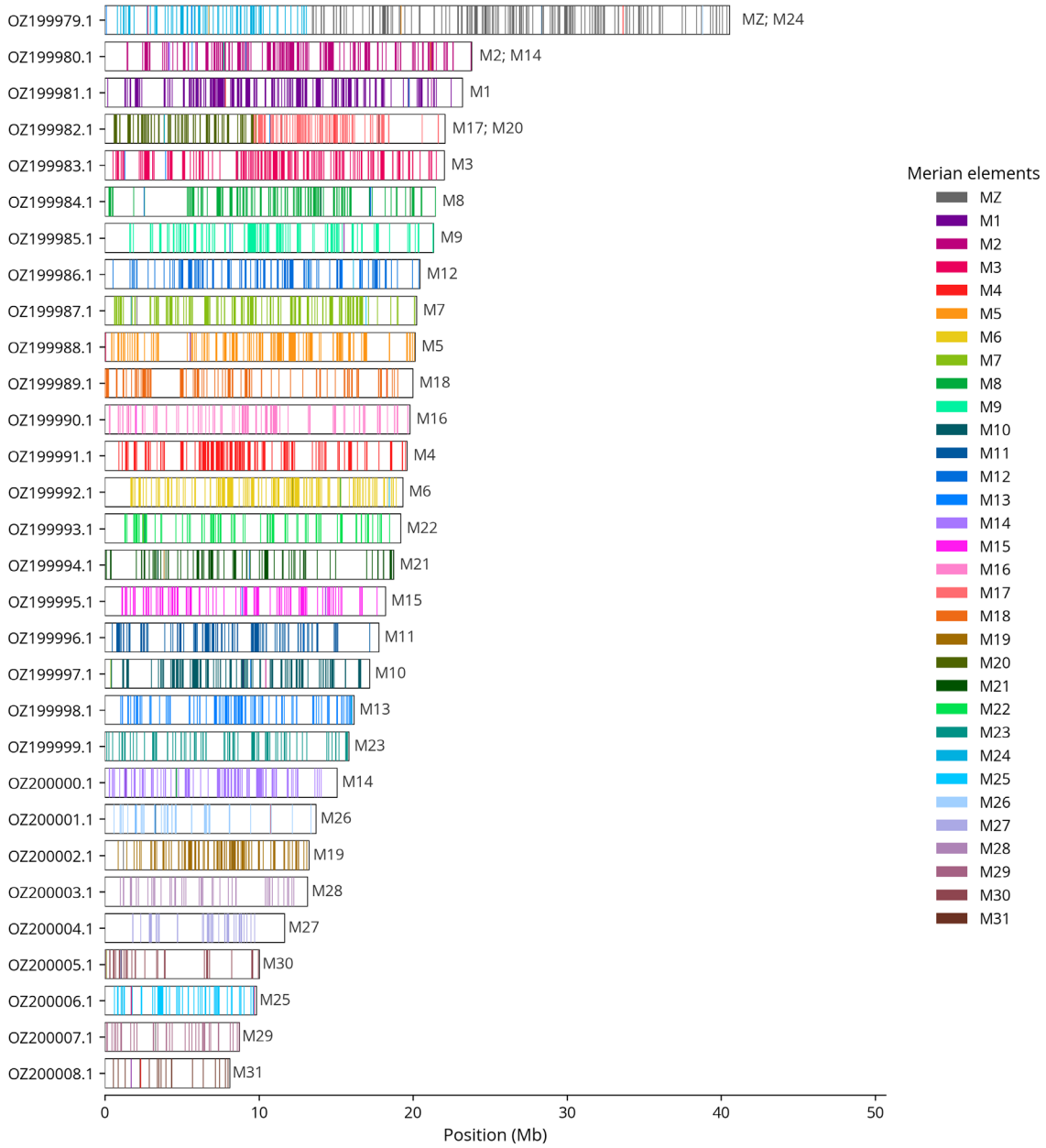

Figure 4. Merian elements painted across chromosomes in the ilMusProo1.hap1.1 assembly of *Muschampia proto*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

(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.Q65**, meeting the recommended reference standard.

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Tree of Life collaborator. The Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing

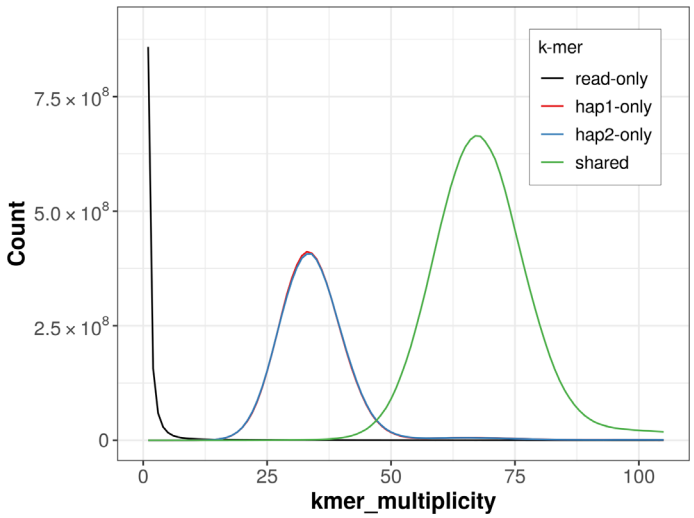

Figure 5. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

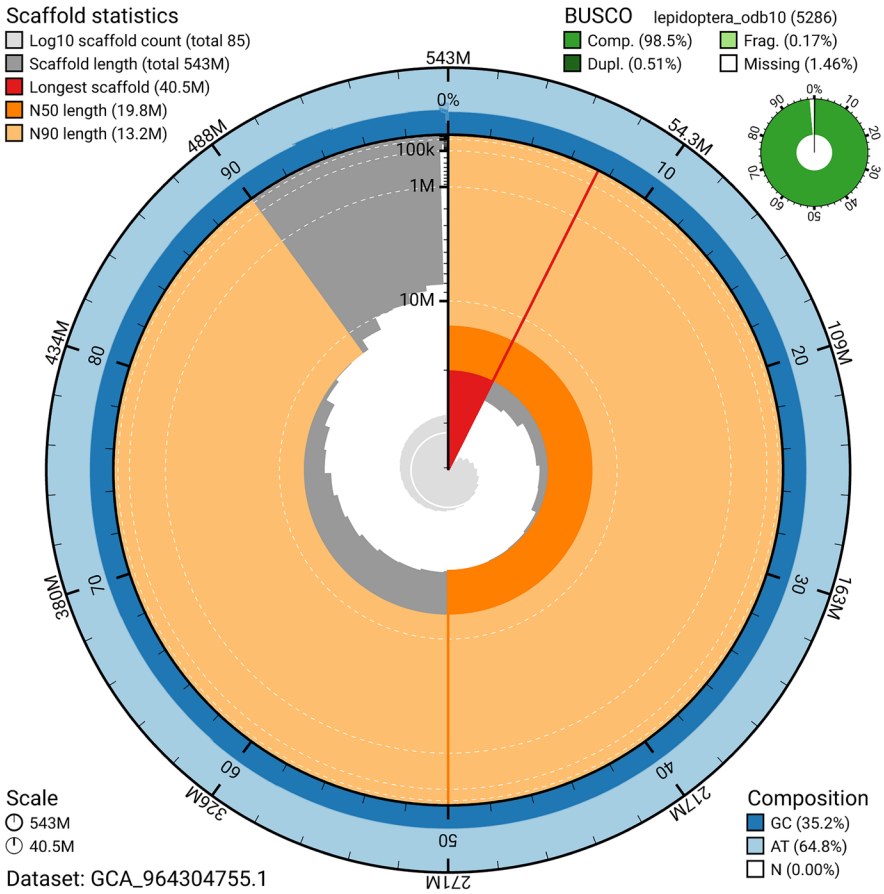

Figure 6. Assembly metrics for ilMusProo1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

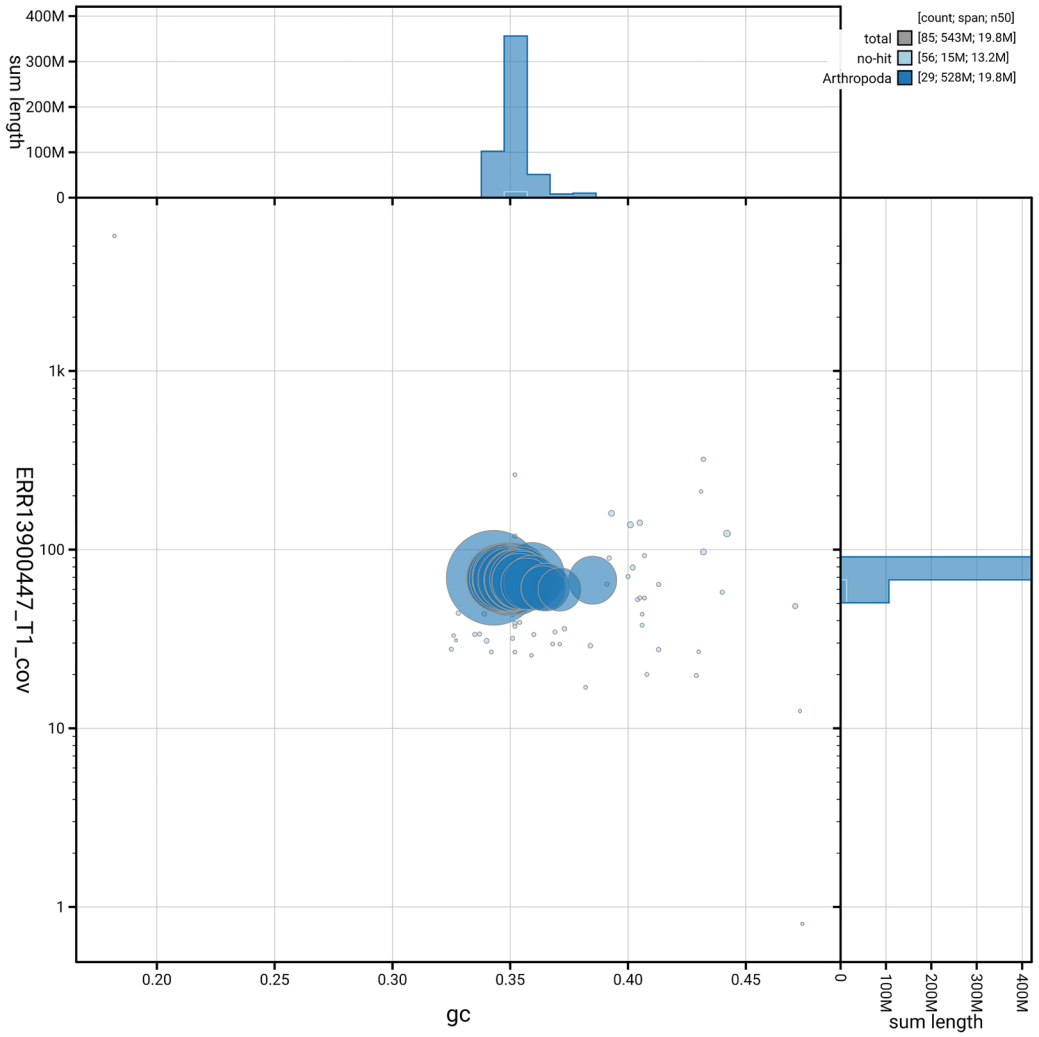

Figure 7. BlobToolKit GC-coverage plot for ilMusProo1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Muschampia proto* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q65	6.C.Q40
Contig N50 length	7.67 Mb	≥ 1 Mb
Scaffold N50 length	19.80 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 65.0; haplotype 2: 65.2; combined: 65.1	≥ 40
k-mer completeness	Haplotype 1: 68.93%; Haplotype 2: 68.89%; combined: 99.78%	≥ 95%
BUSCO	C:98.5% [S:98.0%; D:0.5%]; F:0.2%; M:1.3%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.67%	≥ 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

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: Muschampia proto. Accession number PRJEB81610. The genome sequence is released openly for reuse. The *Muschampia proto* 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
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow

Software	Version	Source
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): 403–410.
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

Dapporto L, Menchetti M, Vodă R, *et al.*: **The atlas of mitochondrial genetic diversity for Western Palaearctic butterflies.** *Glob Ecol Biogeogr.* 2022; **31**(11): 2184–2190.
Publisher Full Text

De Lesse H: **Formules chromosomiques de quelques rhopalocères paléarctiques [Lep].** *Bull Soc Entomol France.* 1970; **75**(7): 214–16.
Publisher 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

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

Fernández-Rubio F: **Guía de Las Mariposas Diurnas de La Península Ibérica, Canarias, Azores y Madeira. I y II.** Madrid: Pirámide S.A., 1991.
Reference Source

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

García-Barros E, Munguira ML, Stefanescu C, *et al.*: **Lepidoptera. Papilionoidea.** Madrid: Museo Nacional de Ciencias Naturales, 2013.

Hinojosa JC, Dapporto L, Brockmann E, *et al.*: **Overlooked cryptic diversity in *Muschampia* (Lepidoptera: Hesperiidae) adds two species to the European butterfly fauna.** *Zool J Linn Soc.* 2021; **193**(3): 847–859.
Publisher Full Text

Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025.
Publisher Full Text

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): giaa153.
PubMed Abstract | Publisher Full Text | Free Full Text

Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HiGlass: web-based visual exploration and analysis of genome interaction maps.** *Genome Biol.* 2018; **19**(1): 125.
PubMed Abstract | Publisher Full Text | Free Full Text

Kriventseva EV, Kuznetsov D, Tegenfeldt F, *et al.*: **OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs.** *Nucleic Acids Res.* 2019; **47**(D1): D807–D811.
PubMed Abstract | Publisher Full Text | Free Full Text

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

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

Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat*

Commun. 2020; **11**(1): 1432.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

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**

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: ***Muschampia proto* (Europe assessment)**. 2025.
Publisher Full Text

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

Vila R, Stefanescu C, Sesma JM: **Guia de Les Papallones Diürnes de Catalunya.** Barcelona: Lynx Edicions, 2018.

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 14 April 2026

https://doi.org/10.21956/wellcomeopenres.28430.r151364

?

Yukun Sun
The Field Museum, Chicago, Illinois, USA

This Data Note presents a reference genome for the Sage Skipper, *Muschampia proto*, from a male specimen collected in Spain. The paper reports two haplotypes. Haplotype 1 is scaffolded into 30 chromosomal pseudomolecules, including the Z chromosome. Haplotype 2 stays at the scaffold level. The mitochondrial genome is also assembled. The reason for creating the dataset is clear. The note links this genome to future work on cryptic diversity and speciation in *Muschampia* and related taxa.
The methods are described well for this article type. The paper gives the specimen details, DNA extraction steps, sequencing methods, assembly workflow, curation steps, and software versions. The accession numbers and software table also make the workflow easier to follow and repeat. On this basis, the study looks technically sound.
The data are presented in a clear and usable way. The paper gives the BioProject, sample, run, and assembly accessions, and it states that the raw data and assemblies are deposited in INSDC databases for reuse. The main quality values are also easy to find. Haplotype 1 has a span of 542.52 Mb, a scaffold N50 of 19.8 Mb, 99.67% of the sequence assigned to chromosomes, BUSCO completeness of 98.5%, and QV 65.0. The Hi-C contact map also supports chromosome-scale scaffolding.
I do have some points that should be addressed in revision.

Points that should be addressed:

1. The coverage value should be checked and explained. The paper reports 36.91 Gb of PacBio HiFi data and a haploid genome size estimate of 530.91 Mb. These numbers give about 69.5x coverage, not 66x. If the authors used filtered reads or another denominator, the paper should say so. This is a basic assembly number, so it should be exact.
2. The Z chromosome assignment needs one more sentence. The specimen is male, and the paper shows a Z chromosome in haplotype 1, while haplotype 2 stays at the scaffold level. The authors should state whether the second Z homolog is present in haplotype 2 but not at the chromosome level, or whether another explanation accounts for this pattern. It would also help to name the reference genome used for the synteny-based Z call, not only give the accession.

3. The contamination screen needs a result statement. The methods state that the assembly was decontaminated with ASCC, but the paper does not indicate whether contaminant or co-biont scaffolds were found, removed, or kept as separate records. One short sentence on the final outcome would clarify the QC process.

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: Bioinformatics, genomics, phylogenomics, genome assembly, and computational biology

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, however I have significant reservations, as outlined above.

Reviewer Report 08 April 2026

https://doi.org/10.21956/wellcomeopenres.28430.r151361

Vadim Pisarenco
Universitat of Barcelona, Barcelona, Spain

The authors present a high-quality, chromosome-level genome assembly of the Sage Skipper (*Muschampia proto*), which will serve as a valuable genomic resource for studying cryptic diversity and speciation in *Muschampia* genus.

The article is well-written, with an elaborate introduction on the *M. proto* biology and its scientific interest. The methods are clearly described and detailed enough to allow replication by others. I have some minor suggestions, outlined below:

1- Standardise the nomenclature of the species to *Muschampia proto* or *M. proto*.

- *M. proto* is now used occasionally throughout the manuscript.

2- In "Genome sequence report" - "Sequence data" section
- "GenomeScope2.0 analysis estimated the haploid genome size at 530.91 Mb" -> According to Figure 2, it should be 530.55 Mb.

3- In "Genome sequence report" - "Assembly quality metrics" section
- BUSCO percentages sum 100% in Table 4, but do not in Figure 6 (100,13) due to % of Missing BUSCOs (1.46% in Figure 6 vs 1.3% in Table 4).

4- Typo in "Background section"
- an valuable -> a valuable

All references have been checked.

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 and 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 03 April 2026

https://doi.org/10.21956/wellcomeopenres.28430.r151365

Matthew Merkin
University of Liverpool, Liverpool, UK

The authors present a chromosome-level assembly of the Sage Skipper (Muschampia proto) genome. A single male sample was sequenced using both PacBio HiFI and Hi-C methods for long

read assembly and scaffolding. A BUSCO completeness of 98.5% and EBP score of 6.C.Q65 suggest that this assembly is of high quality and sufficient for use as a reference genome.

The background is well researched, providing useful information on such a poorly reported species, as well as suggesting uses of the genome. However, I would like to have seen some reference to Project Psyche in the intro to explain why this genome was assembled.

I am also not convinced by scaffold OZ199980.1 being a fusion between M2 and M14 based on Figure 4. Instead, M14 appears to be a separate scaffold (OZ200000.1).

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: genome assembly, entomology, museomics, population genetics

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 09 March 2026

https://doi.org/10.21956/wellcomeopenres.28430.r147894

Temitope Opeyemi Oriowo
Centre for Molecular Biodiversity Research, Leibniz Institute for the Analysis of Biodiveristy Change, Museum Koenig Bonn, Bonn, North Rhine Westphalia, Germany

This is a well-written article about the genome assembly of *Muschampia proto*. In particular, I appreciate the detailed introduction to the species and understand that it would be a valuable addition to future research into the diversity of this cryptic genus.

The methods section is well-detailed, and while I was initially concerned about the lack of software

versions, I am satisfied to see them displayed in a table subsequently.

Assembly Statistics indicate a well-assembled haplotype-resolved genome, and benchmarking against the 'gold standard' was a nice touch. The reports of heterozygosity and repeat content would be nice to put in the context of other butterfly species. This would help us to better understand whether it is as expected or unique.

All the data, including the sequencing and the scripts, is accessible under the correct project numbers.

All references have been checked and validated.

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 and 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.