DATA NOTE

The genome sequence of the Southern Grizzled Skipper, *Pyrgus malvoides* (Elwes & Edwards, 1897) (Lepidoptera: Hesperiidae)

[version 1; peer review: 5 approved]

Yannick Chittaro[1], Kay Lucek[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]Info fauna, Neuchâtel, Switzerland
[2]University of Neuchâtel, Neuchâtel, Switzerland
[3]Tree of Life, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 31 Jul 2025, **10**:404
https://doi.org/10.12688/wellcomeopenres.24653.1
Latest published: 31 Jul 2025, **10**:404
https://doi.org/10.12688/wellcomeopenres.24653.1

Abstract

We present a genome assembly from a male specimen of *Pyrgus malvoides* (Southern Grizzled Skipper; Arthropoda; Insecta; Lepidoptera; Hesperiidae). The assembly contains two haplotypes with total lengths of 746.71 megabases and 747.11 megabases. Most of haplotype 1 (99.8%) is scaffolded into 31 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 15.41 kilobases.

Keywords

Pyrgus malvoides, Southern Grizzled Skipper, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ✓ ✓ ✓

	1	2	3	4	5
version 1 31 Jul 2025	✓ view	✓ view	✓ view	✓ view	✓ view

1. **Olli-Pekka Smolander**, University of Helsinki, Helsinki, Finland
2. **Matthew Merkin**, University of Liverpool, Liverpool, UK
3. **Taslima Sheikh**, Baba Ghulam Shah Badshah University, Rajouri, India
4. **Luana Bataglia**, Universita Cattolica del Sacro Cuore - Campus di Piacenza e Cremona (Ringgold ID: 96985), Piacenza, Italy
5. **Gulab D Khedkar**, Dr. Babasaheb Ambedkar Marathwada University,

Aurangabad, India

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: Chittaro Y: Investigation, Resources, Writing – Original Draft Preparation; **Lucek K**: Resources, Supervision; **Wright CJ**: Funding Acquisition, Project Administration, Supervision; **Meier JI**: Funding Acquisition, Project Administration, Supervision; **Blaxter ML**: Funding Acquisition, Project Administration, Supervision;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540). KL was supported by the Swiss National Science Foundation Grant PCEFP3_202869.
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: Chittaro Y, Lucek K, Wright CJ *et al.* **The genome sequence of the Southern Grizzled Skipper, *Pyrgus malvoides* (Elwes & Edwards, 1897) (Lepidoptera: Hesperiidae) [version 1; peer review: 5 approved]** Wellcome Open Research 2025, **10**:404 https://doi.org/10.12688/wellcomeopenres.24653.1

First published: 31 Jul 2025, **10**:404 https://doi.org/10.12688/wellcomeopenres.24653.1

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; *Pyrgus*; *Pyrgus malvoides* (Elwes & Edwards, 1897) (NCBI:txid1107715)

Background

The Southern Grizzled Skipper *Pyrgus malvoides* is widespread in south-western Europe. It colonises the Iberian Peninsula, the southern half of France and Switzerland, Liechtenstein, western Austria, the extreme south of Germany, Italy (including Sicily), western Slovenia and north-western Croatia (Istria) (Aistleitner, 1996; De Jong, 1972; Fuchs & Wolf, 2006; Guillaumin, 1971; Huemer & Tarmann, 1993; Koren *et al.*, 2013; LSPN, 1999).

In the north and east, *P. malvoides* is replaced by its sister species *Pyrgus malvae*. In addition, *Pyrgus melotis* is found in Turkey and the Middle East (De Jong, 1987). In the past, these three taxa were sometimes considered to be a 'semi-species', belonging to the 'superspecies' *Pyrgus malvae* (De Jong, 1972). Other authors have considered the taxon *malvoides* to be a subspecies of *Pyrgus malvae* (Higgins, 1975; Tshikolovets, 2011). These two taxa have a parapatric distribution and are known to hybridise regularly in the contact zone (Aistleitner, 1996; Guillaumin, 1962; Guillaumin, 1971; Higgins, 1975). However, this contact zone remains narrow (Guillaumin, 1971) and constant, and since there are marked differences on male and female genitalia (De Jong, 1972), the two taxa are now generally treated as distinct species in recent references (Wiemers *et al.*, 2018).

Compared to other *Pyrgus* species occurring in the range of *P. malvoides*, this species is easily recognised by its small size and the well-defined white spots on the upper side of the hindwings. In addition, the underside of the hindwings has distinct bright veins, and the bright spots are generally very small.

The female lays her eggs individually, usually on the underside of host plants. The caterpillar feeds on various rosaceous plants, including *Agrimonia eupatoria*, *Comarum palustre*, *Filipendula vulgaris*, *Fragaria vesca*, many *Potentilla* and *Rubus*, *Sanguisorba minor*, and so on (Clarke, 2024; Hernández-Roldán *et al.*, 2012). As soon as it hatches, the caterpillar makes a shelter from a few silk threads and a leaf that it rolls up. As the caterpillar grows, it creates larger and larger shelters (LSPN, 1999). The caterpillars are often parasitised by parasitoid wasps of the genera *Trichogramma*, *Dolichogenoidea* and *Cotesia* (Hernández-Roldán *et al.*, 2012).

Pyrgus malvoides overwinters as a pupa, making the butterfly one of the first to fly during the year. While the species generally has two generations (April to June and July to August) in the southern part of its range (De Jong, 1972; Lafranchis *et al.*, 2015), in central Europe *Pyrgus malvoides* is generally univoltine (LSPN, 1999; Wagner, 2006), with one generation from April to August, depending on altitude.

Pyrgus malvoides colonises a wide variety of habitats, from plains to altitudes above 2 400 m, from Mediterranean scrubland to meadows and rough pastures, roadsides, forest edges, abandoned quarries and gravel pits, but also alpine meadows and pastures or marshes. The species prefers areas with patchy vegetation (LSPN, 1999). The species is common and not globally threatened (Van Swaay *et al.*, 2010).

Mitochondrial DNA supports substantial cryptic diversity for *P. malvoides*, with one main clade being associated with the Iberian Peninsula and the other with Italy and the Alps (Dapporto *et al.*, 2022). Further studies are needed to reconcile these results with the genital differences observed – in particular by de Jong (1972), who considered two subspecies within *P. malvoides* (ssp. *malvoides* in most of the range and ssp. *modestior* in peninsular Italy and Sicily). The contact zones between *P. malvoides* and *P. malvae* should also be studied in detail using the latest genetic genomic techniques to assess the degree of reproductive isolation between the two species.

The genome presented here will complement the sequence data already available for *Pyrgus malvae* (Hayward *et al.*, 2022), providing the opportunity to also perform comparative genomics within this species complex. The sequence data were derived from a male specimen (Figure 1) collected from Euseigne VS, Switzerland.

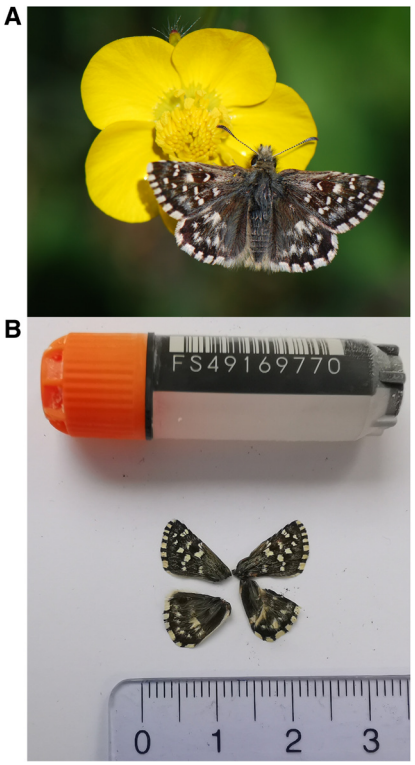

Figure 1. A. Live specimen. **B**. Voucher photograph of the *Pyrgus malvoides* (ilPyrMalo1) specimen used for genome sequencing.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult male *Pyrgus malvoides* (specimen ID SAN28000130, ToLID ilPyrMalo1; Figure 1), collected from Euseigne VS, Switzerland (latitude 46.1773, longitude 7.4159; elevation 800 m) on 26/04/2023. The specimen was collected and identified by Yannick Chittaro (Info Fauna, Neuchâtel, Switzerland).

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 ilPyrMalo1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the thorax was homogenised by powermashing using a PowerMasher II tissue disruptor.

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol. DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor®3 for LI PacBio protocol. Sheared DNA was purified by automated SPRI (solid-phase reversible immobilisation). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 35.46 ng/μL and a yield of 1 666.62 ng, with a fragment size of 13.3 kb. The 260/280 spectrophotometric ratio was 2.03, and the 260/230 ratio was 3.11.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), following the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

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

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

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen head tissue of the ilPyrMalo1 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

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

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

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

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).

Table 1. Specimen and sequencing data for BioProject PRJEB79796.

Platform	PacBio HiFi	Hi-C
ToLID	ilPyrMalo1	ilPyrMalo1
Specimen ID	SAN28000130	SAN28000130
BioSample (source individual)	SAMEA115110012	SAMEA115110012
BioSample (tissue)	SAMEA115110049	SAMEA115110050
Tissue	thorax	head
Sequencing platform and model	Revio	Illumina NovaSeq X
Run accessions	ERR13660018	ERR13654255
Read count total	3.68 million	711.95 million
Base count total	38.68 Gb	107.50 Gb

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 and 58 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 *Pyrgus malvoides* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 38.68 Gb (gigabases) from 3.68 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 741.69 Mb, with a heterozygosity of 1.65% and repeat content of 38.03%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 50× coverage. Hi-C sequencing produced 107.50 Gb from 711.95 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 746.71 Mb in 82 scaffolds, with 122 gaps, and a scaffold N50 of 27.17 Mb (Table 2).

Most of the assembly sequence (99.8%) was assigned to 31 chromosomal-level scaffolds, representing 30 autosomes and the Z sex chromosome. 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

Table 2. Genome assembly statistics.

Assembly name	ilPyrMalo1.hap1.1	ilPyrMalo1.hap2.1
Assembly accession	GCA_964264845.1	GCA_964264935.1
Assembly level	chromosome	scaffold
Span (Mb)	746.71	747.11
Number of chromosomes	31	N/A
Number of contigs	204	185
Contig N50	9.82 Mb	7.94 Mb
Number of scaffolds	82	57
Scaffold N50	27.17 Mb	26.88 Mb
Longest scaffold length (Mb)	33.5	N/A
Sex chromosomes	Z	N/A
Organelles	Mitochondrial genome: 15.41 kb	N/A

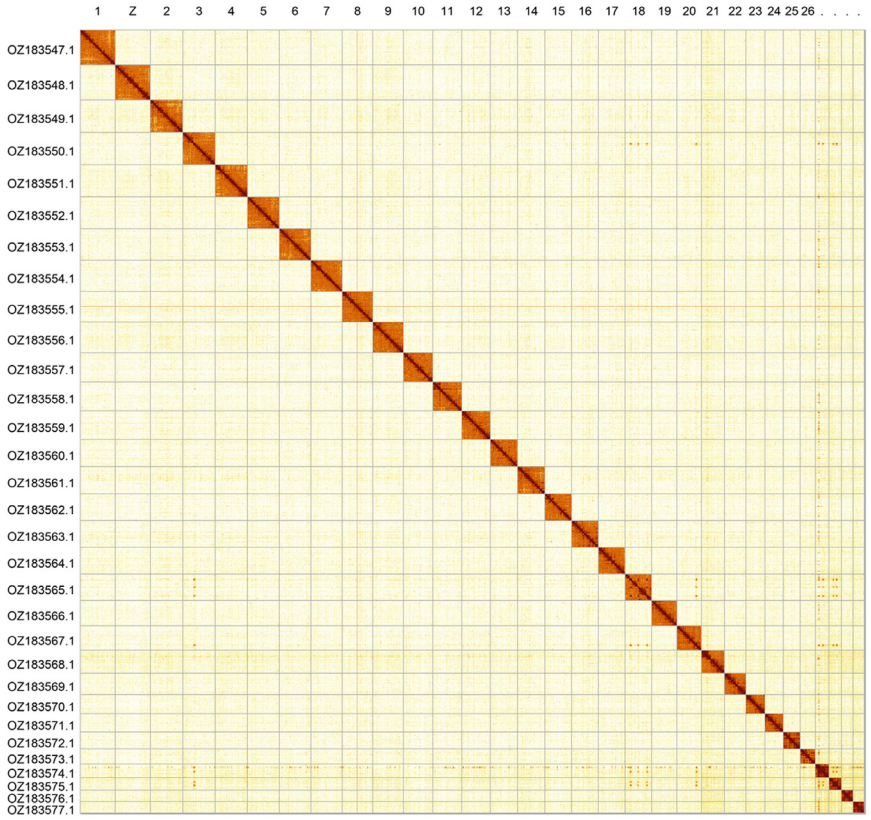

Figure 2. Hi-C contact map of the *Pyrgus malvoides* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Pyrgus malvoides* ilPyrMalo1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ183547.1	1	33.50	36.50	M14;M2
OZ183549.1	2	30.94	36.50	M1
OZ183550.1	3	30.77	36.50	M7
OZ183551.1	4	30.73	36.50	M17;M20
OZ183552.1	5	30.11	36.50	M8
OZ183553.1	6	30.09	36.50	M9
OZ183554.1	7	29.73	36.50	M12
OZ183555.1	8	29.16	36.50	M5
OZ183556.1	9	28.95	36.50	M3
OZ183557.1	10	27.90	36.50	M18
OZ183558.1	11	27.59	36.50	M16
OZ183559.1	12	27.17	36.50	M6
OZ183560.1	13	25.89	37	M22
OZ183561.1	14	25.83	37	M4
OZ183562.1	15	25.51	36.50	M21
OZ183563.1	16	25.44	37	M11
OZ183564.1	17	25.36	37	M15
OZ183565.1	18	25.20	37.50	M23
OZ183566.1	19	24.07	37	M10
OZ183567.1	20	23.26	37	M13
OZ183568.1	21	21.99	38.50	M19
OZ183569.1	22	20.25	37.50	M14
OZ183570.1	23	18.05	37.50	M26
OZ183571.1	24	17.49	37.50	M24
OZ183572.1	25	16.08	37.50	M28
OZ183573.1	26	14.37	37.50	M27
OZ183574.1	27	12.93	40.50	M30
OZ183575.1	28	11.80	38.50	M31
OZ183576.1	29	11.15	38	M25
OZ183577.1	30	10.62	38	M29
OZ183548.1	Z	33.30	36.50	MZ

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 67.7, and for haplotype 2, 66.8. When the two haplotypes are combined, the assembly achieves an estimated QV of 67.2. The *k*-mer completeness is 71.37% for haplotype 1, 71.43% for haplotype 2, and 99.52% for the combined haplotypes (Figure 4). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) (Kriventseva *et al.*, 2019) identified 98.6% of the expected gene set (single = 98.2%, duplicated = 0.4%) for haplotype 1. The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.Q67**, meeting the recommended reference standard.

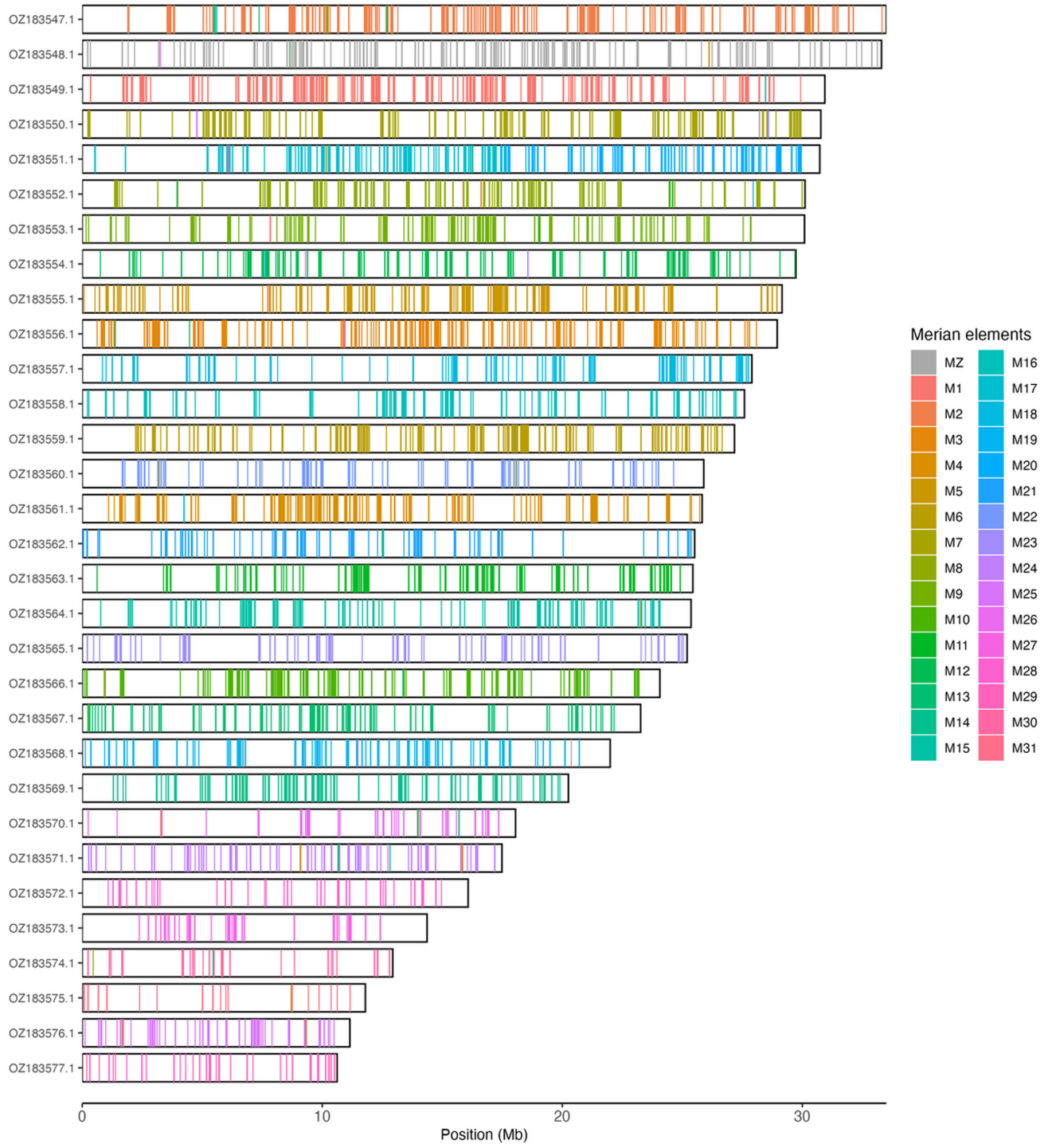

Figure 3. Merian elements painted across chromosomes in the ilPyrMalo1.hap1.1 assembly of *Pyrgus malvoides*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

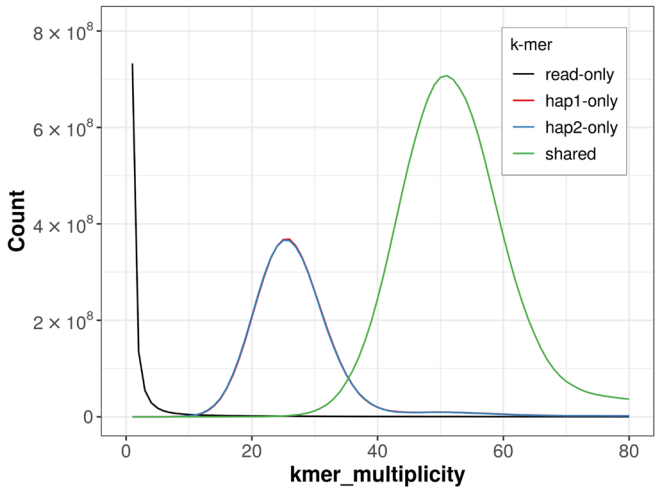

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

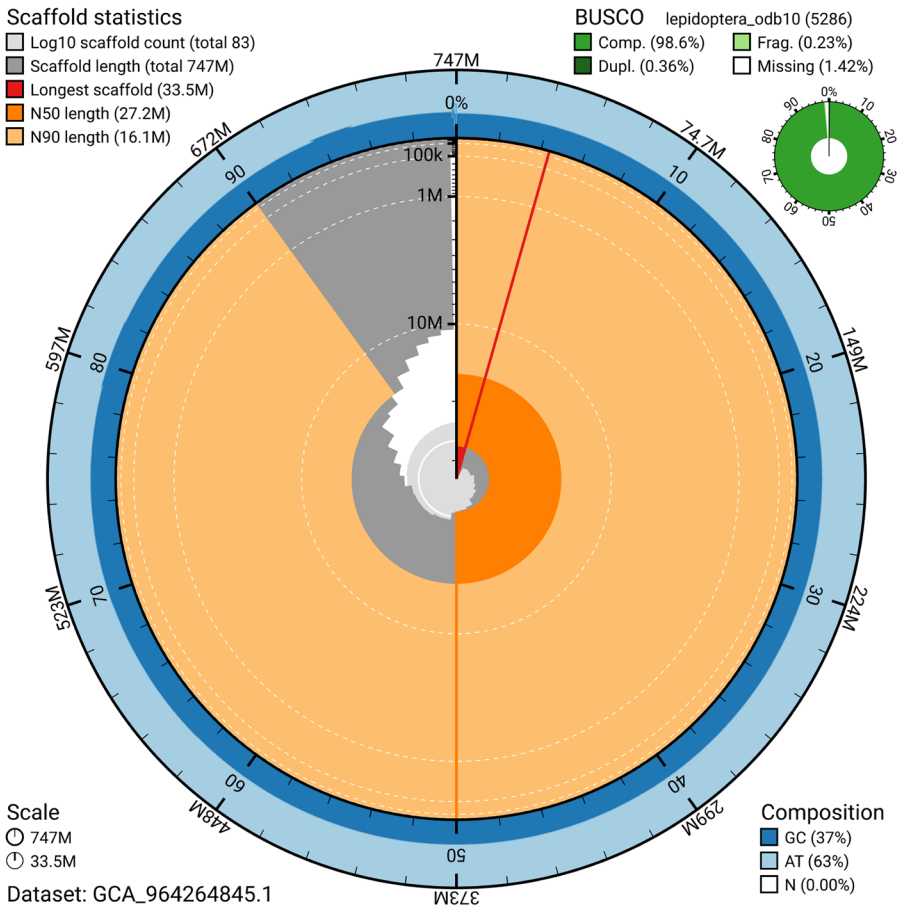

Figure 5. Assembly metrics for ilPyrMalo1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

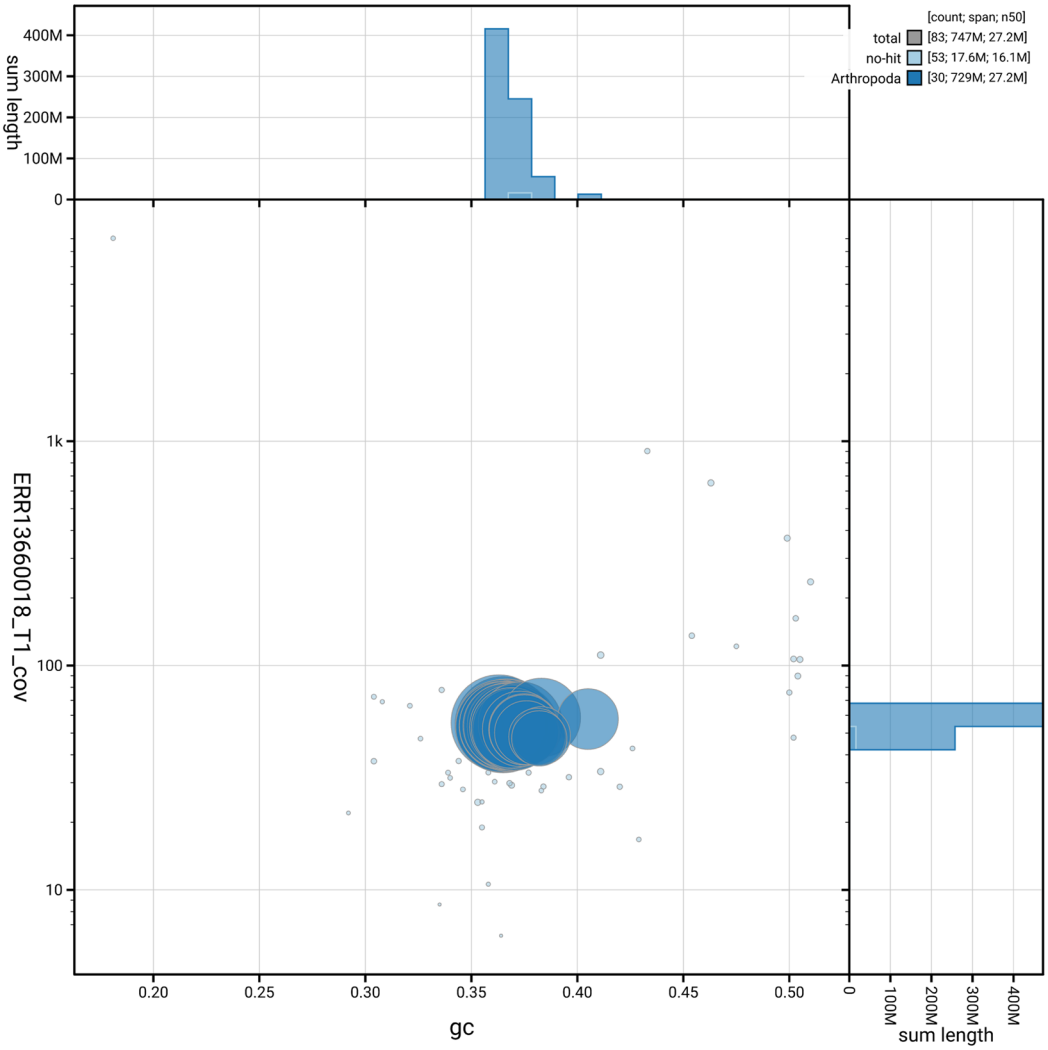

Figure 6. BlobToolKit GC-coverage plot for ilPyrMalo1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Pyrgus malvoides* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.7.Q67	6.C.Q40
Contig N50 length	9.82 Mb	≥ 1 Mb
Scaffold N50 length	27.17 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 67.7; haplotype 2: 66.8; combined: 67.2	≥ 40
k-mer completeness	Haplotype 1: 71.37%; Haplotype 2: 71.43%; combined: 99.52%	≥ 95%
BUSCO	C:98.6%[S:98.2%,D:0.4%], F:0.2%,M:1.2%,n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.80%%	≥ 90%

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: Pyrgus malvoides (southern grizzled skipper). Accession number PRJEB79796. The genome sequence is released openly for reuse. The *Pyrgus malvoides* 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	1.1.2	https://github.com/thegenemyers/MERQURY.FK

Software	Version	Source
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextSnapshot	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

Aistleitner E: **Die Arealgrenzen der beiden Dickkopffalter-Arten *Pyrgus malvae* L. und *Pyrgus malvoides* Elw. & Edw. (Lepidoptera Hesperiidae) in Vorarlberg (Österreich) und Liechtenstein.** *Vorarlb Naturschau.* 1996; **1**: 335–44.
Reference Source

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, 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

Clarke HE: **A checklist of European butterfly larval foodplants.** *Ecol Evol.* 2024; **14**(1): e10834.
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 Palearctic butterflies.** *Glob Ecol Biogeogr.* 2022; **31**(11): 2184–2190.
Publisher Full Text

De Jong R: **Systematics and geographic history of the genus *Pyrgus* in the Palaearctic region.** *Tijdschrift voor Entomologie.* 1972; **115**: 1–121.
Reference Source

De Jong R: **Superspecies *Pyrgus malvae* (Lepidoptera: Hesperiidae) in the east Mediterranean, with notes on phylogenetic and biological relationships.** *Zool Meded.* 1987; **61**(34): 483–500.
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

Fuchs G, Wolf W: ***Pyrgus malvoides* Elwes & Edwards, 1897 auch in den bayerischen Alpen – Erstnachweise für Deutschland (Insecta: Lepidoptera: Hesperiidae).** 2006.

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

Guillaumin M: **Etude des formes intermédiaires entre *Pyrgus malvae* L. et *P. malvoides* Elw. & Edw. (Lep. Hesperiidae).** *Bull de la Soc Fr Entomol.* 1962; **67**: 168–173.
Reference Source

Guillaumin M: **Étude de la variabilité morphologique et biométrique des populations naturelles de *Pyrgus malvae* L. et *P. malvoides* Elw. et Edw. dans leur zone de contact (Lep. Hesperiidae).** *Vie et Milieu.* 1971; **22**: 91–151.
Reference Source

Hayward A, Vila R, Lohse K, *et al.*: **The genome sequence of the grizzled skipper, *Pyrgus malvae* (Linnaeus, 1758) [version 1; peer review: 2 approved, 1 approved with reservations].** *Wellcome Open Res.* 2022; **7**: 114.
PubMed Abstract | Publisher Full Text | Free Full Text

Hernández-Roldán JL, Munguira ML, Wagner W, *et al.*: **Comparative analysis**

and taxonomic use of the morphology of immature stages and natural history traits in European species of *Pyrgus* Hübner (Lepidoptera: Hesperiidae, Pyrginae). *Zootaxa.* 2012; **3470**(1): 1–71.
Publisher Full Text

Higgins LG: **The classification of European butterflies.** London: Collins, 1975.
Reference Source

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

Huemer P, Tarmann G: **Die Schmetterlinge Österreichs (Lepidoptera).** Innsbruck: Tiroler Landesmuseum Ferdinandeum, 1993.
Reference Source

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

Koren T, Beretta S, Črne M, *et al.*: **On the distribution of *Pyrgus malvoides* (Elwes & Edwards, 1897) (Lepidoptera: Hesperiidae) at the eastern part of its range.** *Entomol Gaz.* 2013; **64**(4): 225–34.
Reference Source

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

Lafranchis T, Jutzeler D, Guillosson JY, *et al.*: **La vie des papillons. Écologie, biologie et comportement des Rhopalocères de France.** Diatheo, 2015.
Reference Source

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

LSPN: **Les papillons et leurs biotopes. Espèces, Dangers qui les menacent, Protection. Suisse et régions limitrophes. Tome 2. Hesperiidae, Psychidae, Heterogynidae, Zygaenidae, Syntomidae, Limacodidae, Drepanidae, Thyatiridae, Sphingidae**. Fotorotar SA, 1999.
Reference Source

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

Tshikolovets VV: **Butterflies of Europe & the Mediterranean area.** Tshikolovets Publications, 2011.
Reference Source

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

Van Swaay C, Cuttelod A, Collins S, *et al.*: **European red list of butterflies.** Luxembourg: Publications Office of the European Union, 2010.
Reference Source

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

Wagner W: **The genus *Pyrgus* in Central Europe and its ecology – Larval habitats, host plants and life cycles.** In: T. Fartmann and G. Hermann (eds), *Larval Ecology of Butterflies and Burnets in Central Europe, Vol. 68.* Abhandlungen aus dem Westfälischen Museum für Naturkunde, 2006; 83–122.

Wiemers M, Balletto E, Dincă V, *et al.*: **An updated checklist of the European Butterflies (Lepidoptera, Papilionoidea).** *Zookeys.* 2018; **811**: 9–45.
PubMed Abstract | Publisher Full Text | Free 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 11 June 2026

https://doi.org/10.21956/wellcomeopenres.27167.r136210

✔ **Gulab D Khedkar**
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India

Manuscript review report

This manuscript presents a high-quality chromosome-level genome assembly of *Pyrgus malvoides*, generated using state-of-the-art PacBio HiFi and Hi-C sequencing technologies. The study provides a valuable genomic resource for a species of considerable evolutionary and taxonomic interest, particularly given its close relationship with *Pyrgus malvae* and the ongoing questions surrounding species divergence and hybridization.

The work is technically rigorous and demonstrates excellent assembly quality, with nearly complete chromosomal assignment, high BUSCO completeness, and outstanding consensus accuracy. The authors have employed a comprehensive assembly, curation, and validation workflow that meets and exceeds current Earth BioGenome Project standards. The manuscript is clearly written, methodologically transparent, and provides unrestricted access to all underlying datasets and analytical resources, thereby maximizing its value to the scientific community.

Overall, this study represents a significant contribution to Lepidopteran genomics and will serve as an important reference for future investigations in evolutionary biology, comparative genomics, taxonomy, and conservation genetics

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: Genomics, 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 10 January 2026

https://doi.org/10.21956/wellcomeopenres.27167.r142801

Luana Bataglia
Universita Cattolica del Sacro Cuore - Campus di Piacenza e Cremona (Ringgold ID: 96985), Piacenza, Emilia-Romagna, Italy

The article reports the genome sequencing and assembly of the species *Pyrgus malvoides.* The article is sufficient as a data note, providing complete technical information. I have some suggestions for complementing the study as follows:

1) The background is a bit difficult to follow because there is no linearity in the ideas. I suggest that the authors focus less on biological aspects and more on the importance of the sequencing for the species, as well as the contribution of the study.

2) There is no information about gene annotation. I strongly suggest that the authors perform it and include it in the article to make it more complete.

3) The authors mention in the background the possibility of two subspecies for *P. malvoides* (ssp. *malvoides* and ssp. *modestior*) supported by mitochondrial DNA analysis, but in the results, they don't mention to which subspecies the sequenced specimen belongs, and if the resulting sequencing supports this division. This data could be interesting to report.

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: Molecular Biology, Genetics, Epigenetics, Entomology, Biotechnology.

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 08 January 2026

https://doi.org/10.21956/wellcomeopenres.27167.r142796

Taslima Sheikh
Baba Ghulam Shah Badshah University, Rajouri, Jammu and Kashmir, India

This manuscript presents a high-quality, chromosome-level genome assembly of *Pyrgus malvoides* using PacBio HiFi and Hi-C data. All methodological, analytical, ethical, and data-availability criteria are fully satisfied, and the results are scientifically sound and well supported. No points require correction for scientific validity; only minor editorial revisions are suggested. Authors have done good work

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

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 07 January 2026

https://doi.org/10.21956/wellcomeopenres.27167.r142800

Matthew Merkin
University of Liverpool, Liverpool, UK

The authors present a chromosome-level genome assembly of the European butterfly *Pyrgus malvoides* (The Southern Grizzled Skipper) as part of Project Psyche. An adult male specimen was collected in Switzerland and sequenced to obtain both HiFi and HiC reads for assembly and scaffolding. Manual correction and quality assessment then took place, with the primary haplotype possessing an EBP score of 6.7.Q67 and BUSCO completeness of 98.6%; these suggest that the genome is of a high standard and suitable for use as a reference genome.

This report follows the same format as all Project Psyche genome notes, so I only have two minor comments:

The introduction is of outstanding quality, which clearly highlights the importance of producing a genome assembly for this species being created and a useful suggestion of how the genome can be used. However, there are many irrelevant details, such as a long description of their sheltering behaviour. Whilst these details are certainly interesting, they make the introduction unnecessarily long and less focussed.

Table 3 has assigned the largest chromosome (OZ183547.1) as a fusion between M2 and M14. However, the busco painting in Figure 3 suggests that it is only M2, with M14 being on the 22nd largest chromosome (OZ183569.1).

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: comparative genomics, entomology, genome assembly, museomics

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 10 November 2025

https://doi.org/10.21956/wellcomeopenres.27167.r135243

Olli-Pekka Smolander
University of Helsinki, Helsinki, Finland

This is an exemplary, high-quality data note that is logical, consistent, and scientifically rigorous. The rationale for generating the genome is clearly articulated: *Pyrgus malvoides* has a complex taxonomic relationship with its sister species, *Pyrgus malvae*, with which it is known to hybridize. This genome provides an important and high-quality resource for comparative genomics to investigate these evolutionary questions and to explore existing evidence of cryptic diversity within the species.

The work is technically exceptional, employing a state-of-the-art methodology combining high-accuracy PacBio HiFi long reads for the primary assembly with Hi-C data for scaffolding and phasing. The authors' attention to detail is evident, the work resulting in a fully phased, diploid assembly.

The primary haplotype is 746.71 Mb, with 99.8% of the sequence anchored into 31 chromosomal pseudomolecules. The assembly's quality is demonstrated by a very high contig N50 (9.82 Mb), high accuracy (QV score of 67.7), and high completeness (98.6% complete BUSCOs). This work meets and exceeds the rigorous Earth BioGenome Project (EBP) reference standard.

The methods are documented with commendable transparency, providing sufficient detail on all materials, specimen data, and laboratory protocols to ensure replication. This commitment to open science extends to their computational methods, with links provided to their curation process in a public Git repository and to the pipelines used. The inclusion of a complete list of all software and their specific versions further highlights the project's rigorous approach. Furthermore, the presentation of results is enhanced by links to interactive versions of the key quality plots, such as those from BlobToolKit Finally, the datasets are clearly presented and made fully accessible in public repositories (ENA project PRJEB79796) and are slated for presentation in the Ensembl genome browser, ensuring their widespread and long-term utility for the scientific community.

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

Are the protocols appropriate and is the work technically sound?
Yes

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

Yes

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, multiomics, bioinformatics, computational biology, machine learning, microbiome

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.