DATA NOTE

The genome sequence of the Dusky Grizzled Skipper, *Pyrgus cacaliae* (Rambur, 1839) (Lepidoptera: Hesperiidae)

[version 1; peer review: 2 approved, 1 approved with reservations]

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:** 11 Aug 2025, **10**:424
https://doi.org/10.12688/wellcomeopenres.24672.1
Latest published: 11 Aug 2025, **10**:424
https://doi.org/10.12688/wellcomeopenres.24672.1

Abstract

We present a genome assembly from a female specimen of *Pyrgus cacaliae* (Dusky Grizzled Skipper; Arthropoda; Insecta; Lepidoptera; Hesperiidae). The assembly contains two haplotypes with total lengths of 808.81 megabases and 696.79 megabases. Most of haplotype 1 (94.63%) is scaffolded into 31 chromosomal pseudomolecules, including the W and Z sex chromosomes. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 15.43 kilobases.

Keywords

Pyrgus cacaliae, Dusky Grizzled Skipper, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ?

	1	2	3
version 1 11 Aug 2025	✓ view	✓ view	? view

1. **Arun Arumugaperumal**, Rajalakshmi Engineering College, Chennai, India
2. **Annabel Whibley**, BRI, Lincoln, New Zealand
3. **Yu-Feng Huang**, Yuan Ze University (Ringgold ID: 34895), Zhongli District, Taiwan

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 Dusky Grizzled Skipper, *Pyrgus cacaliae* (Rambur, 1839) (Lepidoptera: Hesperiidae) [version 1; peer review: 2 approved, 1 approved with reservations]** Wellcome Open Research 2025, **10**:424 https://doi.org/10.12688/wellcomeopenres.24672.1

First published: 11 Aug 2025, **10**:424 https://doi.org/10.12688/wellcomeopenres.24672.1

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 cacaliae* (Rambur, 1839) (NCBI:txid876070)

Background

The Dusky Grizzled Skipper *Pyrgus cacaliae* is a butterfly of the family Hesperiidae, endemic to Europe. Its main distribution is in the Alps, but there are also isolated populations in the Pyrenees, the Rila, Pirin and Stara Planina mountains in Bulgaria, the Dinarid mountain chain in Bosnia-Herzegovina and the central Carpathians in Romania (Kolev, 2010; Kudrna & Lux, 2015; Tshikolovets, 2011). Its current distribution has been suggested to reflect glacial refugia after a more continuous range prior to the last glaciation that extended from northern Europe to the glaciers of the Alps (De Jong, 1972), but formal biogeographic studies are still missing.

Pyrgus cacaliae is very similar to other species in the genus. However, the white spots on the upper forewings are greatly reduced on the upper forewings and absent on the hindwings, and the background colour is rather dull. The inner margin of the underside of the hindwing is marked with two white spots forming an exclamation mark, half on a dark and half on a light background, and there is no whitish mark in the basal area of the hindwing discal cell. In case of doubt, examination of the male and female genitalia will give a reliable identification (Higgins, 1975; LSPN, 1999).

The species is univoltine and flies from June to August, depending on weather conditions. It mainly colonises wet meadows and pastures, usually near streams, along swamps and resurgences. Dry grasslands are sometimes used. The species is found above 1 700 m (Tshikolovets, 2011; Wagner, 2006), although there are some records at lower altitudes. Males are territorial.

The eggs are laid individually on the leaves of the host plants, namely *Potentilla erecta*, *P. crantzii*, *P. aurea*, *Geum montanum* and possibly *Sibbaldia procumbens* (Hernández-Roldán *et al.*, 2011; LSPN, 1999; Wagner, 2003). The caterpillar lives in a shelter made of different parts of leaves connected by silk threads. Its life cycle is still poorly understood, but it seems that the larval development lasts two years, with the first overwintering during the first or second larval stage and the second overwintering at the pupal stage (Lafranchis & Kan, 2015; Wagner, 2003; Wagner, 2006). Locally, the two-year life cycle may mean that adults fly only every other year, or at least at different abundances (Wagner, 2024).

Although this species is not currently considered to be globally threatened (Van Swaay *et al.*, 2010), global warming undoubtedly threatens the survival of certain populations of this very localised species, particularly outside the Alps (Settele *et al.*, 2008).

This reference genome will help to clarify the taxonomic situation and evolutionary history within the disjunct areas of the species. For instance, the Romanian populations have recently been assigned to a distinct subspecies, *hebsgaardi* (Bjerregård & Mølgaard, 2023) on the basis of limited genetic data (see also Dapporto & Vila, 2022), but also on the basis of morphological differences. The number of chromosomes (n = 30) was already highlighted by de Lesse (1960).

We present a chromosome-level, haplotype-resolved genome sequence of the Dusky Grizzled Skipper, *Pyrgus cacaliae*, sequenced as part of Project Psyche. The sequence data were derived from a female specimen (Figure 1) collected from Simplon, Valais, Switzerland.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult female *Pyrgus cacaliae* (specimen ID SAN28000144, ToLID ilPyrCaca1; Figure 1), collected from Simplon, Valais, Switzerland (latitude 46.2506, longitude 8.0273; elevation 2 000 m) on 14/07/2023. The specimen was collected and identified by Yannick Chittaro (Info Fauna, Neuchâtel, Switzerland).

Figure 1. Voucher photograph of the *Pyrgus cacaliae* (ilPyrCaca1) specimen used for genome sequencing.

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 ilPyrCaca1 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 64.4 ng/μL and a yield of 3 026.80 ng, with a fragment size of 16.2 kb. The 260/280 spectrophotometric ratio was 1.85, and the 260/230 ratio was 2.58.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Samples with an average fragment size greater than 8 kb and total mass exceeding 400 ng were eligible for the low-input SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences, California, USA), depending on genome size and required sequencing depth. Libraries were prepared using the SMRTbell Prep Kit 3.0 according to the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

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

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

Table 1. Specimen and sequencing data for BioProject.

Platform	PacBio HiFi	Hi-C
ToLID	ilPyrCaca1	ilPyrCaca1
Specimen ID	SAN28000144	SAN28000144
BioSample (source individual)	SAMEA115110095	SAMEA115110095
BioSample (tissue)	SAMEA115110098	SAMEA115110096
Tissue	thorax	head
Sequencing platform and model	Revio	Illumina NovaSeq X
Run accessions	ERR13605502	ERR13602165
Read count total	2.23 million	671.17 million
Base count total	23.93 Gb	101.35 Gb

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen head tissue of the ilPyrCaca1 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). 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 11 breaks, 43 joins, and removal of 50 haplotypic duplications. 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 cacaliae* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 23.93 Gb (gigabases) from 2.23 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 748.49 Mb, with a heterozygosity of 1.05% and repeat content of 38.33%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 31× coverage. Hi-C sequencing produced 101.35 Gb from 671.17 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 808.81 Mb in 127 scaffolds, with 203 gaps, and a scaffold N50 of 26.69 Mb (Table 2).

Most of the assembly sequence (94.63%) was assigned to 31 chromosomal-level scaffolds, representing 29 autosomes and the W and Z sex chromosomes. The Z and W sex chromosomes were identified by read coverage. Most of the W chromosome is in unlocalised contigs assigned to W, as the exact order and orientation could not be determined. The chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 2; Table 3). Chromosome painting with Merian elements illustrates the distribution of orthologues along chromosomes and highlights patterns of chromosomal evolution relative to Lepidopteran ancestral linkage groups (Figure 3).

Table 2. Genome assembly statistics.

Assembly name	ilPyrCaca1.hap1.1	ilPyrCaca1.hap2.1
Assembly accession	GCA_964261425.1	GCA_964261415.1
Assembly level	chromosome	scaffold
Span (Mb)	808.81	696.79
Number of chromosomes	31	N/A
Number of contigs	330	229
Contig N50	6.72 Mb	6.11 Mb
Number of scaffolds	127	50
Scaffold N50	26.69 Mb	26.79 Mb
Longest scaffold length (Mb)	58.08	N/A
Sex chromosomes	W and Z	N/A
Organelles	Mitochondrial genome: 15.43 kb	N/A

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 64.8, and for haplotype 2, 65.7. When the two haplotypes are combined, the assembly achieves an estimated QV of 65.2. The *k*-mer completeness is 80.82% for haplotype 1, 74.41% for haplotype 2, and 99.11% 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 = 98.1%, duplicated = 0.5%) 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.Q64**, meeting the recommended reference standard.

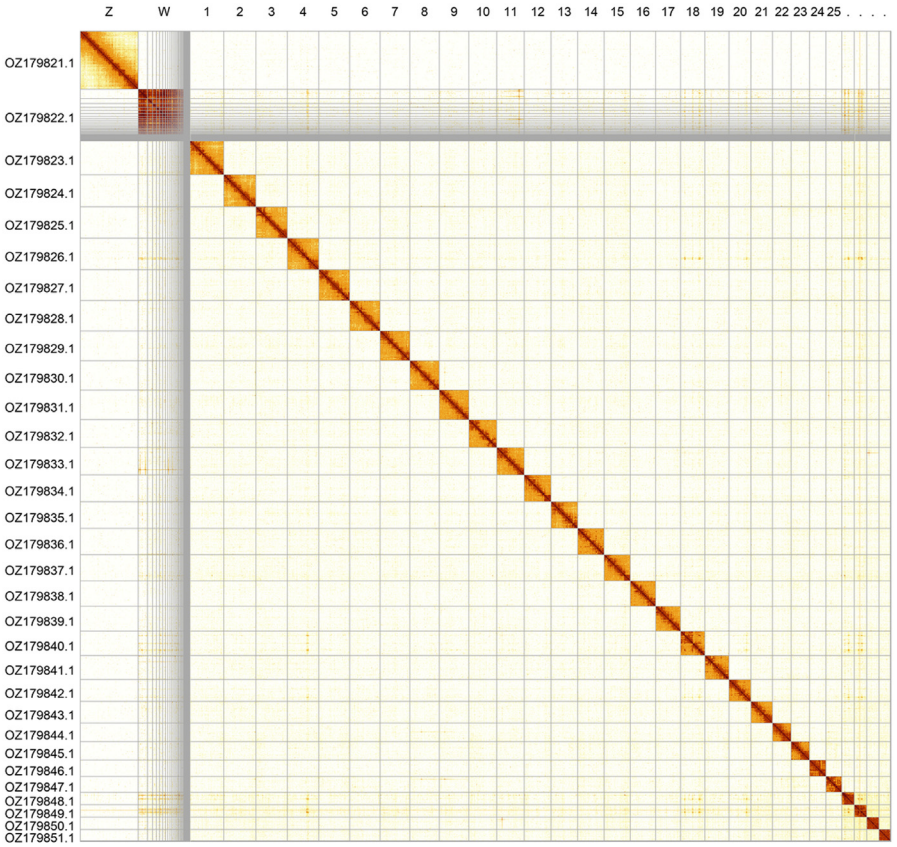

Figure 2. Hi-C contact map of the *Pyrgus cacaliae* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Pyrgus cacaliae* ilPyrCaca1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ179823.1	1	33.66	36.50	M14;M2
OZ179824.1	2	32.03	37	M1
OZ179825.1	3	31.43	36.50	M17;M20
OZ179826.1	4	31.39	36.50	M7
OZ179827.1	5	30.72	36.50	M8
OZ179828.1	6	30.42	36.50	M9
OZ179829.1	7	29.52	36.50	M3
OZ179830.1	8	29.40	36.50	M12
OZ179831.1	9	29.38	36.50	M5
OZ179832.1	10	27.97	36.50	M18
OZ179833.1	11	27.36	36.50	M6
OZ179834.1	12	26.69	36.50	M16
OZ179835.1	13	26.44	37	M4
OZ179836.1	14	26.32	36.50	M22
OZ179837.1	15	26.17	37	M10
OZ179838.1	16	25.28	36.50	M21
OZ179839.1	17	24.90	36.50	M15
OZ179840.1	18	24.29	37	M23
OZ179841.1	19	23.90	37	M11
OZ179842.1	20	22.04	37	M13
OZ179843.1	21	21.42	37.50	M14
OZ179844.1	22	18.56	37	M24
OZ179845.1	23	18.31	37	M26
OZ179846.1	24	16.33	37	M28
OZ179847.1	25	15.83	37	M27
OZ179848.1	26	12.58	39.50	M30
OZ179849.1	27	12.27	38	M31
OZ179850.1	28	12.24	38	M25
OZ179851.1	29	11.32	38	M29
OZ179822.1	W	51.55	37.50	N/A
OZ179821.1	Z	58.08	36.50	M19;MZ

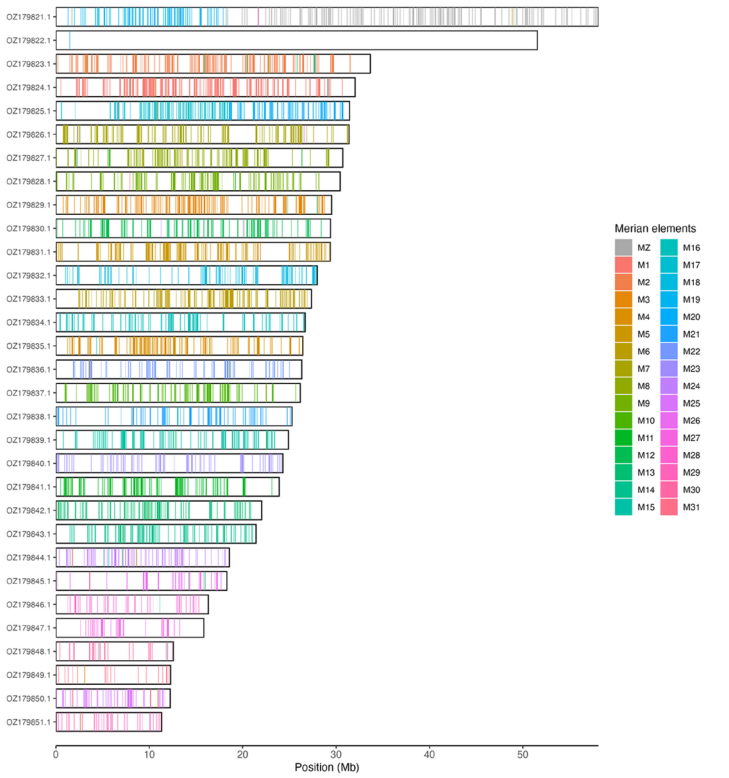

Figure 3. Merian elements painted across chromosomes in the ilPyrCaca1.hap1.1 assembly of *Pyrgus cacaliae*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

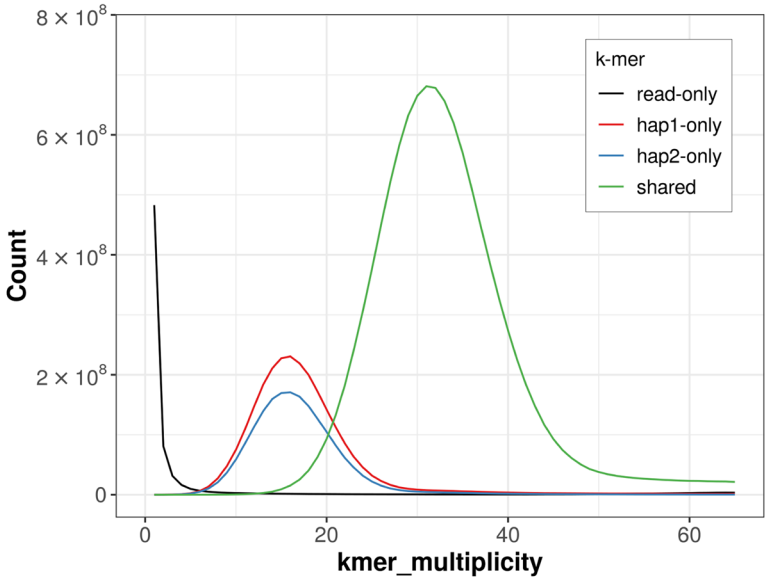

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

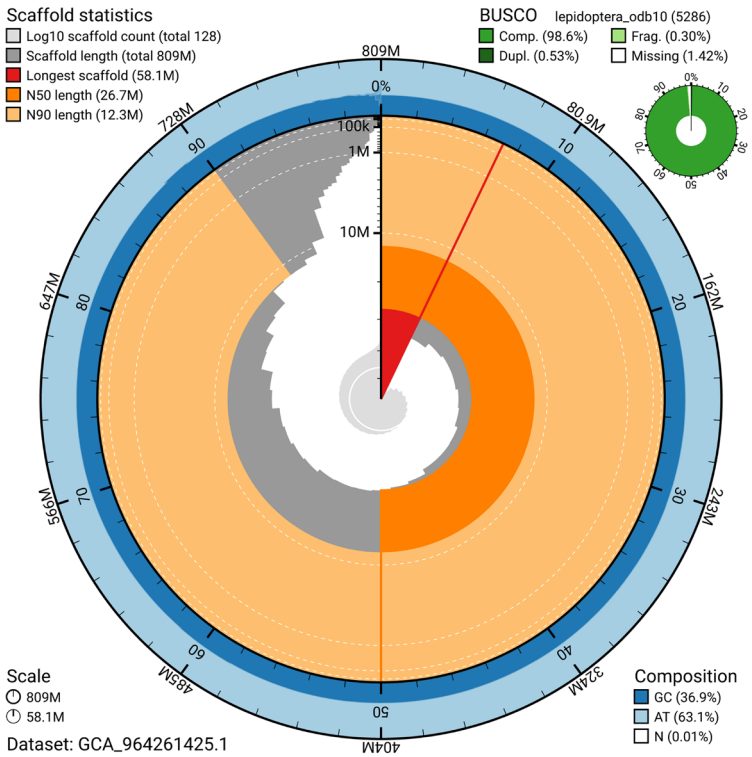

Figure 5. Assembly metrics for ilPyrCaca1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

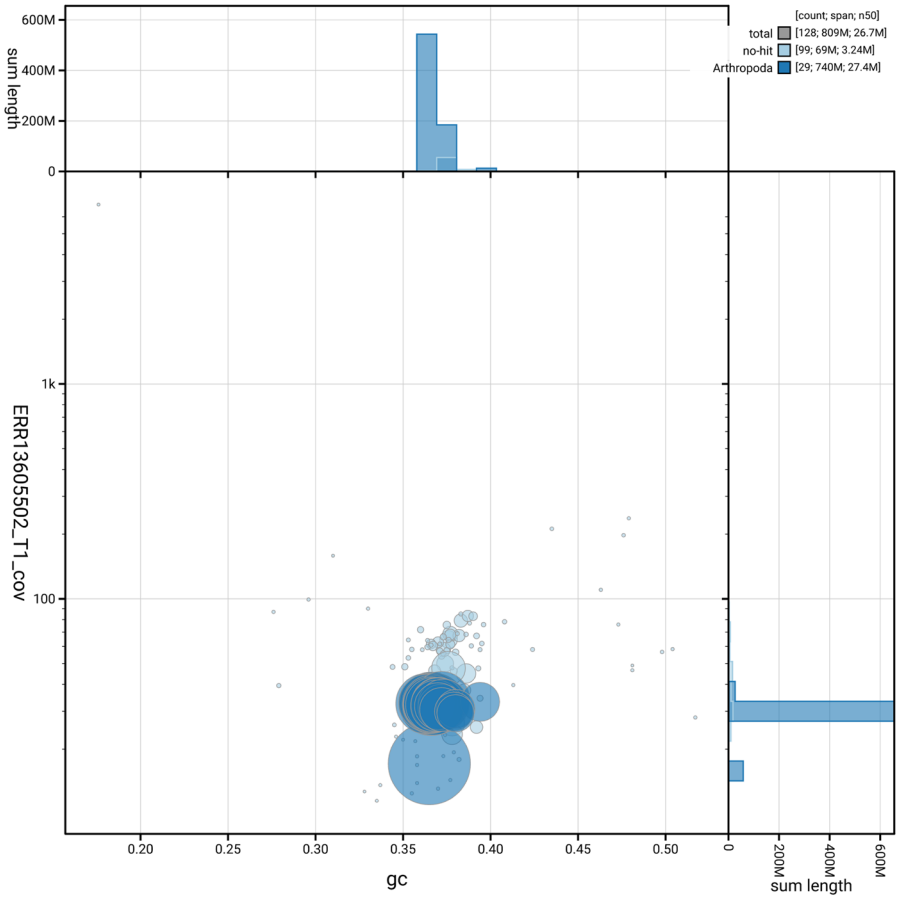

Figure 6. BlobToolKit GC-coverage plot for ilPyrCaca1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 cacaliae* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q64	6.C.Q40
Contig N50 length	6.72 Mb	≥ 1 Mb
Scaffold N50 length	26.69 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 64.8; haplotype 2: 65.7; combined: 65.2	≥ 40
k-mer completeness	Haplotype 1: 80.82%; Haplotype 2: 74.41%; combined: 99.11%	≥ 95%
BUSCO	C:98.6%[S:98.1%,D:0.5%],F:0.3%,M:1.1%,n:5286	S > 90%; D & 5%
Percentage of assembly assigned to chromosomes	99.87%	≥ 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 cacaliae (dusky grizzled skipper). Accession number PRJEB79178. The genome sequence is released openly for reuse. The *Pyrgus cacaliae* 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
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC

Software	Version	Source
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextSnapshot	N/A	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.6.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

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

Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text

Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the universal protein knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531.
PubMed Abstract | Publisher Full Text | Free Full Text

Bjerregård EB, Mølgaard MS: **A new taxon in the genus *Pyrgus* (Lepidoptera: Hesperiidae) from the Carpathian Mountains in Romania.** *Phegea.* 2023; **51**(1): 42–46.
Publisher 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

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 ML, Vila R: **The atlas of mitochondrial genetic diversity for western palearctic butterflies.** *Glob Ecol Biogeogr.* 2022; **31**(11): 2184–2190.
Publisher Full Text

de Lesse H: **Spéciation et variation chromosomiques chez les Lépidoptères Rhopalocères.** *Annales de Sciences naturelles, Zoologie (12).* 1960; **2**(1): 1–223.
Reference Source

De Jong R: **Systematics and geographic history of the genus *Pyrgus* in the palaearctic region (Lepidoptera, Hesperiidae).** *Tijdschrift voor Entomologie.* 1972; **115**: 1–121.
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

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

Hernández-Roldán JL, Vicente JC, Munguira ML: **Historia natural de *Pyrgus cacaliae* (Rambur, 1839) en la Península Ibérica (Lepidoptera: Hesperiidae).** *SHILAP Revista de lepidopterología.* 2011; **39**(153): 61–73.
Reference Source

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

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

Kolev Z: **A significant range extension for *Pyrgus cacaliae* (Rambur, 1839) with the first record from the western Balkan Peninsula (Hesperiidae).** *Nota lepid.* 2010; **33**(1): 107–13.
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

Kudrna PO, Lux K: **Distribution Atlas of European butterflies and skippers**. Wissenschaftlicher Verlag PEKS, 2015; 632.
Reference Source

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, Kan B: **La vie des Papillons. Écologie, biologie et comportement**

des Rhopalocères de France. Diatheo, 2015; 751.
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; 667.
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.
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**

Settele J, Kudrna O, Harpke A, *et al.*: **Climatic risk Atlas of European butterflies.** *BioRisk.* 2008; **1**: 1–712.
Publisher Full Text

Tshikolovets VV: **Butterflies of Europe & the mediterranean area.** Tshikolovets Publications. 2011; 544.
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: **Beobachtungen zur Biologie von *Pyrgus andromedae* (Wallengren, 1853) und *Pyrgus cacaliae* (Rambur, 1840) in den Alpen (Lepidoptera: Hesperiidae).** *Entomologische Zeitschrift.* 2003; **113**(12): 346–56.
Reference Source

Wagner W: **The genus *Pyrgus* in central Europe and its ecology – Larval habitats, host plants and life cycles.** *Abhandlungen aus dem Westfälischen Museum für Naturkunde.* 2006; **68**(3/4): 83–122.

Wagner W: **European Lepidoptera and their ecology: Pyrgus cacaliae.** 2024.
Reference Source

Wright CJ, Stevens L, Mackintosh A, *et al.*: **Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera.** *Nat Ecol Evol.* 2024; **8**(4): 777–90.
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 29 December 2025

https://doi.org/10.21956/wellcomeopenres.27187.r129621

?

Yu-Feng Huang
Computer Science and Engineering, Yuan Ze University (Ringgold ID: 34895), Zhongli District, Taoyuan City, Taiwan

This manuscript presents a chromosome-level, haplotype-resolved genome assembly of *Pyrgus cacaliae*, the Dusky Grizzled Skipper, generated under the Wellcome Sanger Institute's Tree of Life (ToL) and Project Psyche programs. The assembly achieves excellent metrics - scaffold N50 of 26.69 Mb, BUSCO completeness of 98.6%, and QV > 64 - meeting Earth BioGenome Project (EBP) reference genome standards (6.C.Q64).

The paper is concise, methodologically rigorous, and consistent with ToL data note conventions. It provides a high-quality reference genome for *Pyrgus*, with clear relevance for comparative Lepidopteran genomics and conservation studies. The sequencing, assembly, and curation pipelines follow standardized and well-documented workflows, ensuring long-term reproducibility.

The manuscript is technically sound and well written. However, several clarifications would further strengthen completeness and transparency.

Major Comments

1. Haplotype Representation and Explanation

The genome includes two haplotypes - haplotype 1 assembled to chromosome level and haplotype 2 assembled to scaffold level - yet the manuscript does not explain the basis for this difference in assembly contiguity. Please clarify whether the reduced continuity and completeness of haplotype 2 (N50 = 26.79 Mb; k-mer completeness = 74.41%) reflects biological heterozygosity, reduced read coverage, or curation thresholds applied during assembly polishing. A short explanation would help readers interpret the relative reliability and intended usage of each haplotype.

2. Mitochondrial Genome Description

The mitochondrial genome (15.43 kb) is mentioned only briefly. It should be explicitly stated

whether the mitogenome is complete and circularized, as circularity is a key quality criterion in mitochondrial assemblies. Providing a succinct annotation summary (e.g., 13 PCGs, 22 tRNAs, 2 rRNAs) and any comparison with existing *Pyrgus* mitogenomes would strengthen this section. Additionally, an INSDC accession number for the mitochondrial genome is missing and should be included.

3. Assembly Curation Transparency
The authors report 11 breaks and 43 joins during manual curation, but do not specify whether these occurred at the scaffold or chromosome level. Clarifying the type, rationale, and impact of these corrections (e.g., resolving chimeric joins, removing haplotypic duplications) would enhance transparency.
Furthermore, given that the rapid curation protocol was used, the authors should briefly indicate whether any evaluation method was applied to assess potential missing or misassembled regions following curation.

4. BUSCO and k-mer Completeness Discussion
Although BUSCO completeness (98.6%) and combined k-mer completeness (99.11%) are excellent, the substantial differences between haplotypes (80.82% vs. 74.41%) warrant explanation. A short interpretative note regarding heterozygosity, haplotype collapse, or assembly limitations would assist users in understanding genome representation and completeness.

5. Numerical or Reporting Consistency
There is a discrepancy between the main text and Table 4: the text reports that “94.63% of the assembly sequence was assigned to chromosomes,” while Table 4 lists “99.87%.” This inconsistency should be corrected for clarity and accuracy.

Minor Comments
1. Expression of Units
Replace “15.43 kb” with the exact value (e.g., 15,431 bp or 15,439 bp) when referring to the mitochondrial genome. Given the small size of mitogenomes, reporting precise base pair length is preferable to rounded kilobase values.

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, Mitogenome, Machine learning, Cancer 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 23 September 2025

https://doi.org/10.21956/wellcomeopenres.27187.r129625

Annabel Whibley
Grapevine Improvement, BRI, Lincoln, New Zealand

In this Data Note, Chittaro and colleagues report the genome sequence of the Dusky Grizzled Skipper, *Pyrgus cacaliae.* Sequenced as part of project Psyche, the introduction sets out the natural history of this species and identifies some knowledge gaps that a genome sequence may help to address.

Nearly 95% of the assembly could be scaffolded into 31 pseudomolecules, and the use of a female specimen means that both Z and W chromosomes are reported in this assembly, although the resolution of the W is hampered by its highly repetitive nature. The sequencing and assembly pipeline, and subsequent evaluations, follow standard DToL procedures and templating. The low-input library prep route was used. The assembly quality is excellent, with a QV >64 and very complete BUSCO and k-mer statistics. Public data accessions are working.

Minor comments that should be addressed: In the Genome sequence report section, the statement about coverage for the assembly backbone is little ambiguous ("HiFi sequence data" rather than "sequencing data" would be clearer). I am a little sceptical that the in vivo photograph in Figure 1 is the same individual as in the voucher panel below, but perhaps the legend to this figure could specify. There is also, I think, an error in the last row of Table 4. Both the abstract and main text state 94.63% of the assembly sequence is assigned to chromosomes, whereas this table states "99.87").

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, bioinformatics

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 11 September 2025

https://doi.org/10.21956/wellcomeopenres.27187.r129626

Arun Arumugaperumal
Department of Biotechnology, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India

The data note describes the genome sequencing of a butterfly species *Pyrgus cacaliae*. The butterfly specimen has been neatly documented as a photograph in figure 1. This is the first report of the high-quality genome sequence of the organism. The assembly presented is of size 808.81 Mb spread along 31 chromosomes. The butterfly follows ZW sex system and both chromosomes were scaffolded since this was a female specimen. The mitogenome has also been sequenced which is of size 15.43 kb. High-end sequencing technology and standard protocols were used to arrive at the high-quality genome of this butterfly. There are several other Pyrgus genomes sequenced as part of Darwin tree of life and has been published. The genome size of *P. cacaliae* is found to be comparable to those genome sizes. The BUSCO completeness of 98.6% indicates that the genome sequence provided is a near-complete one. The links for accessing the data have been provided except for the Ensembl annotation. The article can be indexed.

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

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.