DATA NOTE

The genome sequence of the Carline Skipper, *Pyrgus carlinae* (Rambur, 1839) (Lepidoptera: Hesperiidae)

[version 1; peer review: 2 approved]

Yannick Chittaro[1], Daniel Linke[2], Kay Lucek[3], Charlotte J. Wright[4], Joana I. Meier[4], Mark L. Blaxter[4],
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]Biology Centre of the Czech Academy of Sciences, Institute of Entomology, České Budějovice, Czech Republic
[3]University of Neuchâtel, Neuchâtel, Switzerland
[4]Tree of Life Programme, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 20 Jan 2026, **11**:50
https://doi.org/10.12688/wellcomeopenres.25810.1
Latest published: 20 Jan 2026, **11**:50
https://doi.org/10.12688/wellcomeopenres.25810.1

Abstract

We present a genome assembly from a female specimen of *Pyrgus carlinae* (Carline Skipper; Arthropoda; Insecta; Lepidoptera; Hesperiidae). The assembly contains two haplotypes with total lengths of 808.07 megabases and 554.08 megabases. Most of haplotype 1 (99.18%) is scaffolded into 25 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. Gene annotation of this assembly on Ensembl identified 14 216 protein-coding genes. This work is part of Project Psyche, a collaborative programme generating genomes for European butterflies and moths.

Keywords

Pyrgus carlinae; Carline Skipper; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 20 Jan 2026	✓ view	✓ view

1. **Ian M Carr**, University of Leeds, Leeds, UK

2. **Laura Jimeneza Burney**, University of Liverpool, Liverpool, UK

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; **Linke D**: Writing – Original Draft Preparation; **Lucek K**: Supervision, Writing – Review & Editing; **Wright CJ**: Funding Acquisition, Project Administration, Supervision; **Meier JI**: Funding Acquisition, Project Administration, Supervision; **Blaxter ML**: Funding Acquisition, Project Administration, Supervision;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540). DL was funded by Czech Science Foundation (GACR) grant 22-35084J and Technology Agency of the Czech Republic (TACR) grant SS07010197. KL is a recipient of Swiss National Science Foundation (SNSF) Grant ID 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, Linke D, Lucek K *et al.* **The genome sequence of the Carline Skipper, *Pyrgus carlinae* (Rambur, 1839) (Lepidoptera: Hesperiidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2026, **11**:50 https://doi.org/10.12688/wellcomeopenres.25810.1

First published: 20 Jan 2026, **11**:50 https://doi.org/10.12688/wellcomeopenres.25810.1

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 carlinae* (Rambur, 1839) (NCBI:txid2505785)

Background

Pyrgus carlinae, known commonly as the Carline Skipper, is a small butterfly species in the Hesperiidae family. It is endemic to the western and south-western Alps, occurring in south-eastern France, southern Switzerland and north-western Italy (Tolman & Lewington, 2009). *P. carlinae* is absent from the eastern Alps and confined to mid- to high-elevation environments, generally between 1 000 and 2 800 m, being most common at around 2 000 m (Wagner, 2025). The upper altitudinal range is higher in the southern Alps than the northern Alps, likely due to frequent high-pressure systems and increased sun (Wagner, 2006).

This skipper inhabits dry, rocky alpine pastures, grassy clearings, and open slopes with sparse vegetation. Morphologically, *P. carlinae* closely resembles several other *Pyrgus* species and is difficult to identify in the field. It is characterised mainly by a distinctive C-shaped mark near the forewing costa and obscure white hindwing markings. Females usually show yellowish dorsal dusting with reduced spots, while males show larger spots, including more pronounced ones on the dorsal hindwing (Tolman & Lewington, 2009; Wagner, 2025). Hybrids exist between *P. carlinae* and *P. cirsii*, with intermediate external and genitalia morphology in south-eastern France (Guillaumin, 1963).

Adults are univoltine, with a single flight period from late June to early August (occasionally into September), depending on altitude and local climate. Eggs are deposited on the leaf underside of the host plant, and larvae overwinter within the eggshell and feed during April to June. In captivity, a second generation is possible when temperatures are adequately high (Wagner, 2006). The primary larval host plants are *Potentilla* spp., including *P. verna*, *P. reptans*, *P. hirta* and *P. tabernaemontani* (Tolman & Lewington, 2009).

P. carlinae is listed as Least Concern on the European Red List of Butterflies (van Swaay *et al.*, 2025). Despite its limited range, no major population declines have been reported. However, Wagner (2025) has documented localised reductions due to agricultural intensification and increased irrigation, transforming dry grasslands. Furthermore, the species may be threatened by climate warming. As with many other mountain species, the conservation of *P. carlinae* relies on maintaining traditional extensive grazing or mowing practices that preserve the open structure of alpine meadows and prevent succession.

We present a chromosome-level genome sequence for *Pyrgus carlinae*, sequenced as part of Project Psyche. The sequence data were derived from a female specimen (Figure 1) collected from Conthey, Valais, Switzerland.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult female *Pyrgus carlinae* (specimen ID SAN28000147, ToLID ilPyrCarl1; Figure 1), collected from Conthey, Valais, Switzerland (latitude 46.2872, longitude 7.3116) on 2023-08-02. The specimen was collected and identified by Yannick Chittaro.

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 ilPyrCarl1 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 63.43 ng/μL and a yield of 2 981.21 ng, with a fragment size of 15.4 kb.

Figure 1. Voucher photograph of the *Pyrgus carlinae* (ilPyrCarl1) specimen used for genome sequencing.

PacBio HiFi library preparation and sequencing

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

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

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

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen head tissue of the ilPyrCarl1 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

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

Prior to sequencing, libraries were normalised to 10 ng/µL. Normalised libraries were quantified again to create equimolar and/or weighted 2.8 nM pools. Pool 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)

Table 1. Specimen and sequencing data for BioProject PRJEB80574.

Platform	PacBio HiFi	Hi-C
ToLID	ilPyrCarl1	ilPyrCarl1
Specimen ID	SAN28000147	SAN28000147
BioSample (source individual)	SAMEA115110069	SAMEA115110069
BioSample (tissue)	SAMEA115110072	SAMEA115110071
Tissue	thorax	head
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13726004	ERR13731947
Read count total	1.71 million	813.64 million
Base count total	18.78 Gb	122.86 Gb

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 nine breaks and 142 joins. This reduced the scaffold count by 27.1%, increased the scaffold N50 by 3.9%, and increased the total assembly length by 5.0%. The curation process is described at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

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

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

We used lep_busco_painter to paint Merian elements along chromosomes (Wright *et al.*, 2024). Merian elements represent the 32 ancestral linkage groups in Lepidoptera. The painting process utilised BUSCO gene locations from the lepidoptera_odb10 set (Kriventseva *et al.*, 2019) and chromosome lengths from NCBI Datasets. Each complete BUSCO gene (both single-copy and duplicated) was assigned to a Merian element based on a reference database, then plotted along chromosomes drawn to scale.

Genome sequence report

Sequence data

PacBio sequencing of the *Pyrgus carlinae* specimen generated 18.78 Gb (gigabases) from 1.71 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 723.99 Mb, with a heterozygosity of 2.50% and repeat content of 40.98% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 25× coverage. Hi-C sequencing produced 122.86 Gb from 813.64 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 808.07 Mb in 174 scaffolds, with 285 gaps, and a scaffold N50 of 30.67 Mb (Table 2).

Most of the assembly sequence (99.18%) was assigned to 25 chromosomal-level scaffolds, representing 23 autosomes and the W and Z sex chromosomes. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Chromosome painting with Merian elements illustrates the distribution of orthologues along chromosomes and highlights patterns of chromosomal evolution relative to Lepidopteran ancestral linkage groups (Figure 4). The Z and W chromosomes were identified by read coverage. The exact order and orientation of the contigs on chromosome W (16.0–71.0 Mbp) are unknown.

The mitochondrial genome was also assembled (length 15.43 kb, OZ194705.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

For haplotype 1, the estimated QV is 63.9, and for haplotype 2, 63.7. When the two haplotypes are combined, the assembly achieves an estimated QV of 63.8. The *k*-mer completeness is 68.45% for haplotype 1, 50.92% for haplotype 2, and 94.32%

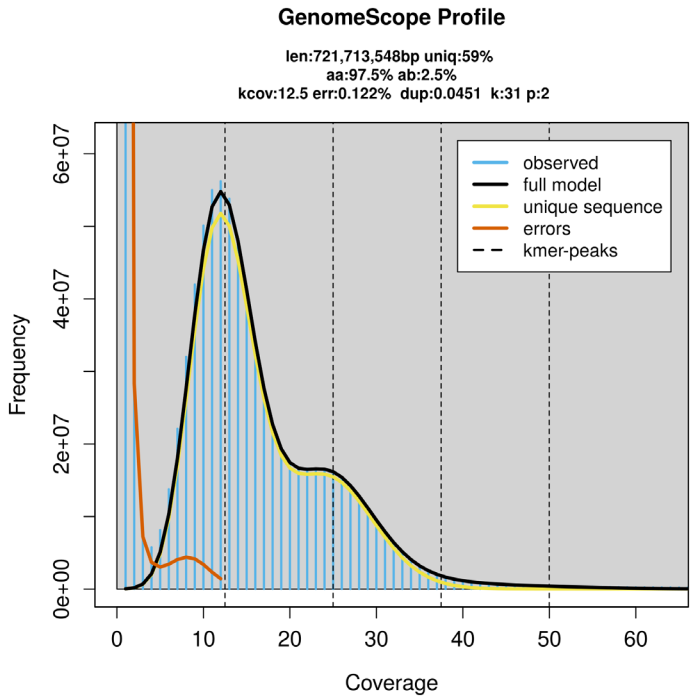

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 2. Genome assembly statistics.

Assembly name	ilPyrCarl1.hap1.1	ilPyrCarl1.hap2.1
Assembly accession	GCA_964276835.1	GCA_964276845.1
Assembly level	chromosome	scaffold
Span (Mb)	808.07	554.08
Number of chromosomes	25	scaffold-level
Number of contigs	459	346
Contig N50	6.1 Mb	5.43 Mb
Number of scaffolds	174	173
Scaffold N50	30.67 Mb	29.1 Mb
Longest scaffold length (Mb)	71.19	-
Sex chromosomes	W and Z	-
Organelles	Mitochondrion: 15.43 kb	-

for the combined haplotypes (Figure 5). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.4% of the expected gene set (single = 96.7%, duplicated = 1.6%) in haplotype 1. For haplotype 2, BUSCO analysis identified 74.9% of the expected gene set (single = 74.5%, duplicated = 0.4%). The snail plot in Figure 6 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 7 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) the Earth BioGenome Project Report on

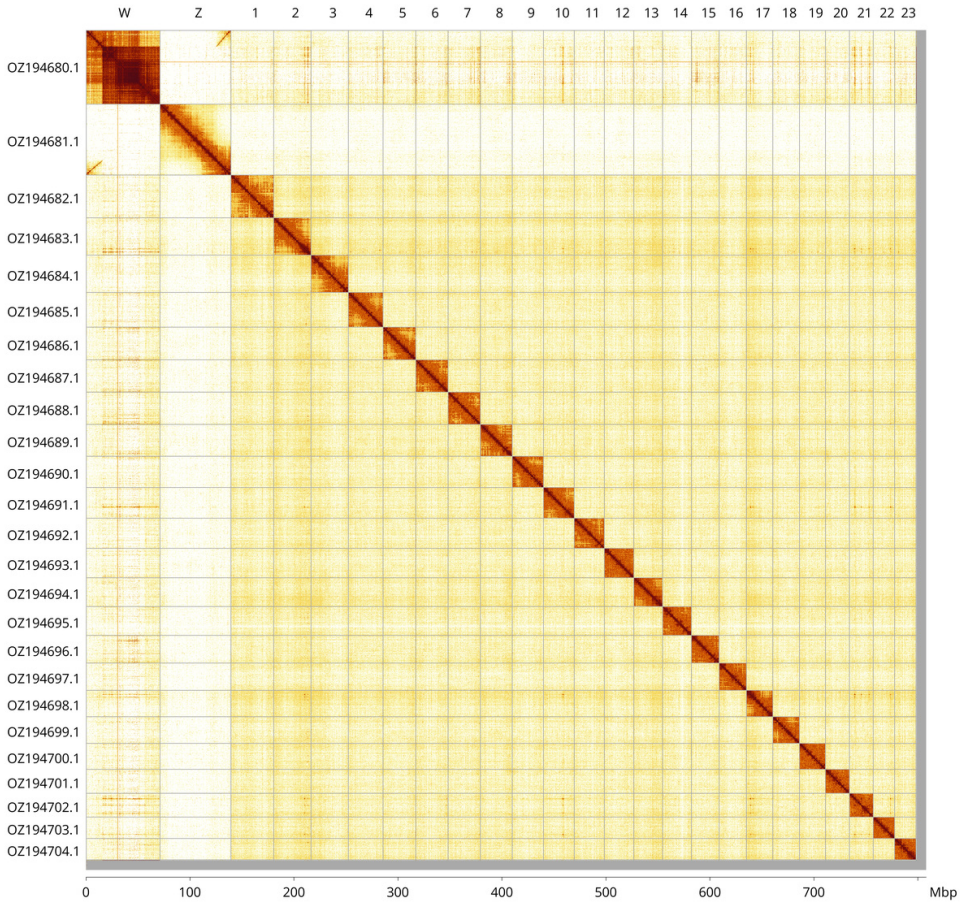

Figure 3. Hi-C contact map of the *Pyrgus carlinae* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Pyrgus carlinae* ilPyrCarl1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ194682.1	1	41.09	36.50	M15;M24
OZ194683.1	2	35.95	37.50	M10;M30
OZ194684.1	3	35.91	36.50	M21;M25
OZ194685.1	4	33.41	36.50	M14;M2
OZ194686.1	5	31.36	36.50	M1
OZ194687.1	6	31.09	36.50	M8
OZ194688.1	7	31.02	36.50	M12
OZ194689.1	8	30.67	36.50	M17;M20
OZ194690.1	9	30.14	36	M9
OZ194691.1	10	29.53	36.50	M7
OZ194692.1	11	28.99	36.50	M3
OZ194693.1	12	28.30	36.50	M5

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ194694.1	13	27.81	37	M26;M29
OZ194695.1	14	27.69	36	M18
OZ194696.1	15	26.55	36.50	M6
OZ194697.1	16	26.15	36.50	M16
OZ194698.1	17	25.56	37.50	M28;M31
OZ194699.1	18	25.51	36.50	M4
OZ194700.1	19	25.11	36.50	M22
OZ194701.1	20	23.09	37	M11
OZ194702.1	21	22.74	37	M23
OZ194703.1	22	20.73	36.50	M13
OZ194704.1	23	20.53	37	M14
OZ194680.1	W	74.23	37	M27
OZ194681.1	Z	68.28	36.50	M19;M27;MZ

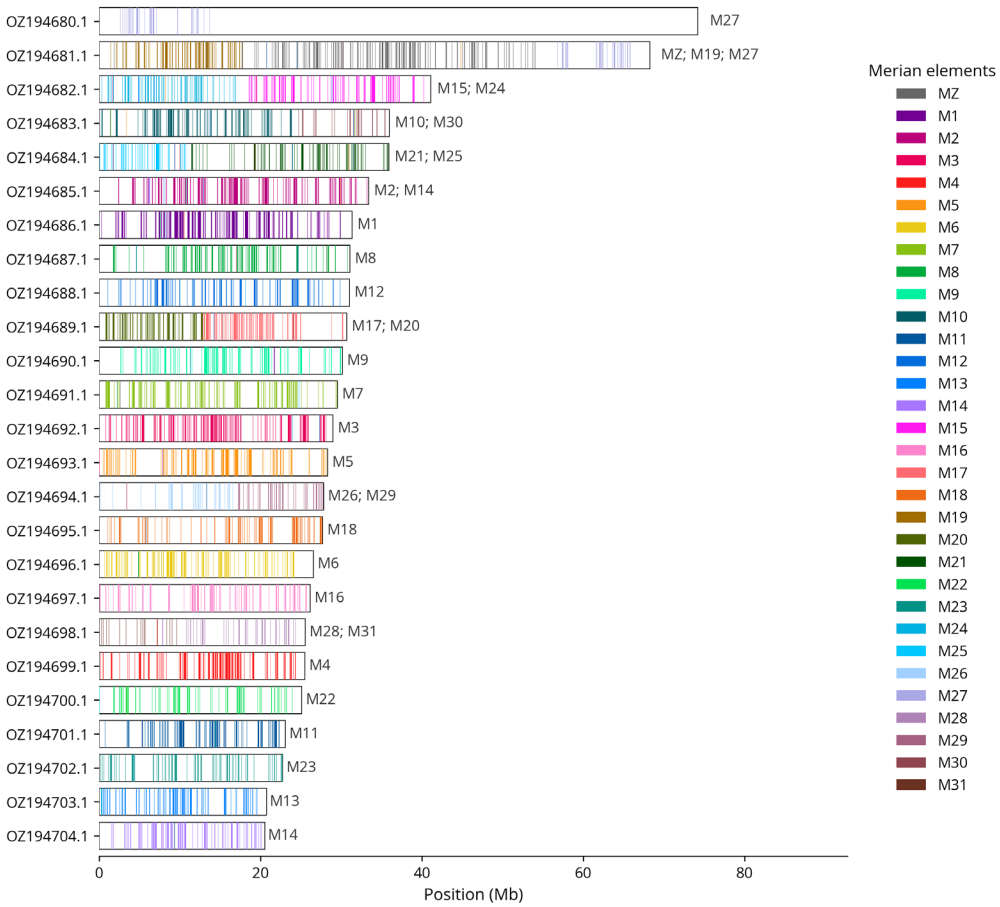

Figure 4. Merian elements painted across chromosomes in the ilPyrCarl1.hap1.1 assembly of *Pyrgus carlinae*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

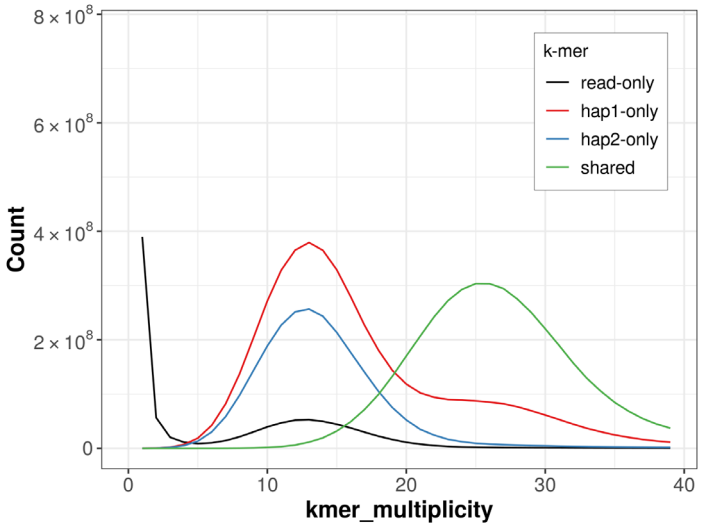

Figure 5. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve (the homozygous peak) corresponds to *k*-mers shared by both haplotypes and the red and blue curves (the heterozygous peaks) show *k*-mers found only in one of the haplotypes.

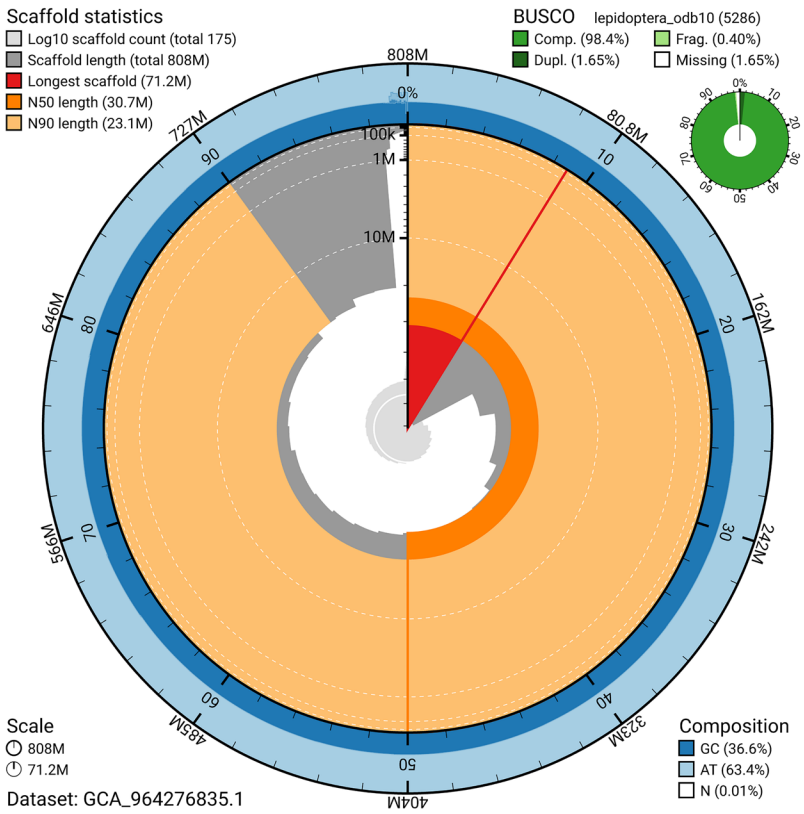

Figure 6. Assembly metrics for ilPyrCarl1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

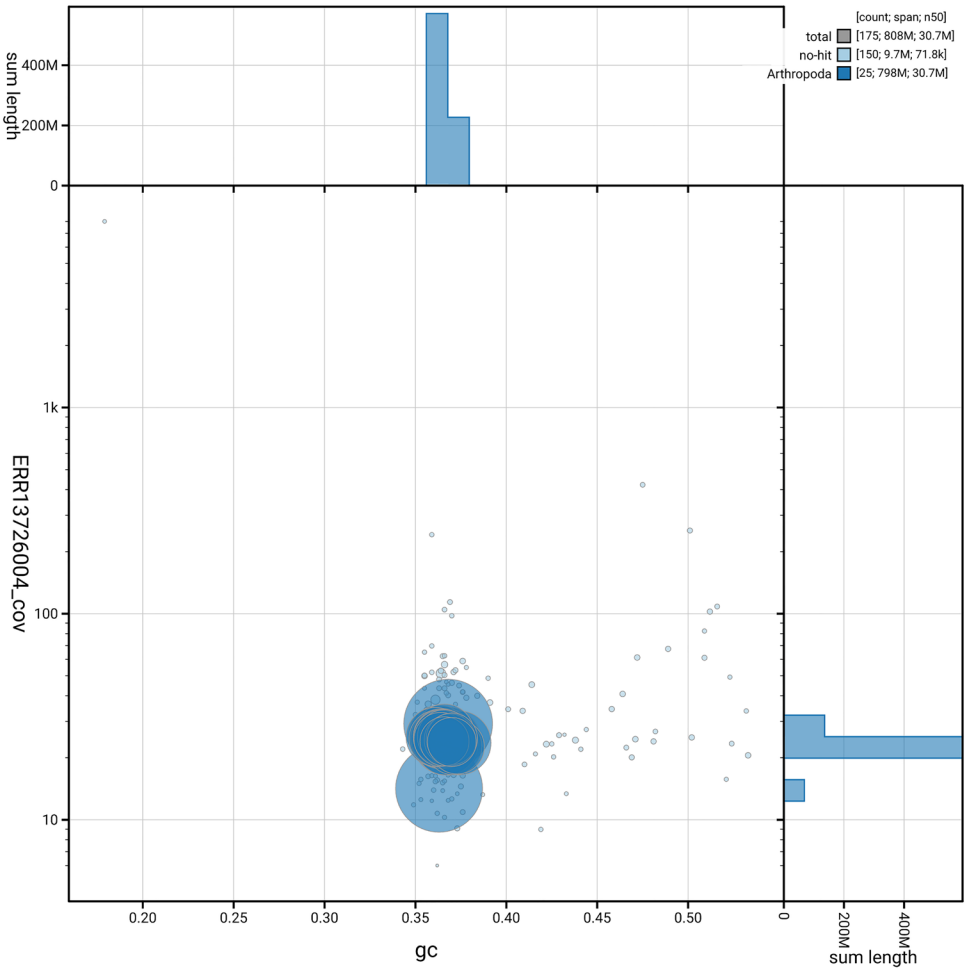

Figure 7. BlobToolKit GC-coverage plot for ilPyrCarl1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 carlinae* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q63	6.C.Q40
Contig N50 length	6.10 Mb	≥ 1 Mb
Scaffold N50 length	30.67 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 63.9; haplotype 2: 63.7; combined: 63.8	≥ 40
k-mer completeness	Haplotype 1: 68.45%; Haplotype 2: 50.92%; combined: 94.32%	≥ 95%
BUSCO	C:98.4% [S:96.7%; D:1.6%]; F:0.4%; M:1.2%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.18%	≥ 90%

Notes: EBP summary uses log10(Contig N50); chromosome-level (C) or log10(Scaffold N50); Q (Merqury QV). BUSCO: C=complete; S=single-copy; D=duplicated; F=fragmented; M=missing; n=orthologues

Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.Q63**, meeting the recommended reference standard.

Genome annotation report

The *Pyrgus carlinae* genome assembly (GCA_964276835.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 30 370 transcribed mRNAs from 14 216 protein-coding and 4 773 non-coding genes. The average transcript length is 20 934.66 bp, with an average of 1.60 coding transcripts per gene and 6.87 exons per transcript. For further information, please refer to the Ensembl annotation page.

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 carlinae (Carlinae skipper). Accession number PRJEB80574. The genome sequence is released openly for reuse. The *Pyrgus carlinae* 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. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

Pipelines used for genome assembly at the WSI Tree of Life are available at https://pipelines.tol.sanger.ac.uk/pipelines. Table 5 lists software versions used in this study.

Author information

Contributors are listed at the following links:

- Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team
- Members of Wellcome Sanger Institute Scientific Operations – Sequencing Operations
- Members of the Wellcome Sanger Institute Tree of Life Core Informatics team
- Members of the Tree of Life Core Informatics collective
- Members of the Project Psyche Community

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.4.5	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.7.1	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass

Software	Version	Source
lep_busco_painter	1.0.0	https://github.com/charlottewright/lep_busco_painter
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.28-r1209	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	24.10.4	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView
samtools	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.7.1	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

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

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

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

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

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

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

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

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

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

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
PubMed Abstract | Publisher Full Text

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

Guillaumin M: **Les hybrides naturels de *Pyrgus carlinae* Rbr. Et *Pyrgus cirsii* Rbr. (Lep. Hesperiidae).** *Les hybrides naturels de.* 1963; **88**: 600–603.

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

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

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

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

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

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

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic

coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat Commun.* 2020; **11**(1): 1432.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

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

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI taxonomy: a comprehensive update on curation, resources and tools.** *Database (Oxford).* 2020; **2020**: baaa062.
PubMed Abstract | Publisher Full Text | Free Full Text

Tolman T, Lewington R: **Collins butterfly guide: the most complete guide to the butterflies of Britain and Europe**. HarperCollins Publishers, 2009.
Reference Source

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

van Swaay C, Ellis S, Warren M: **Pyrgus carlinae.** 2025.
Publisher Full Text

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

Wagner W: **Die Gattung Pyrgus in Mitteleuropa und ihre ökologie – Larvalhabitate, Nährpflanzen und Entwicklungszyklen.** In: T. Fartmann and G. Hermann (eds), *Larvalökologie von Tagfaltern Und Widderchen in Mitteleuropa.* 2006; **68**: 83–122.

Wagner W: **Pyrgus carlinae.** 2025.
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): 777–790.
PubMed Abstract | Publisher Full Text | Free Full Text

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 14 March 2026

https://doi.org/10.21956/wellcomeopenres.28411.r149479

✔ **Laura Jimeneza Burney**
University of Liverpool, Liverpool, UK

This report describes the high-quality-level genome assembly for a female *Pyrgus carlinae* (Carline Skipper), a Lepidoptera species endemic to the western and south-western Alps. The primary assembly (Haplotype 1) spans 808.07 Mb and has been scaffolded into 25 chromosomal pseudomolecules, which include both the W and Z sex chromosomes. This work, conducted as part of Project Psyche, utilized PacBio HiFi long-read sequencing and Hi-C chromatin conformation capture to achieve high continuity and accuracy.

- The authors report a substantial difference in span (808.07 Mb vs 554.08 Mb) and BUSCO completeness (98.4% vs 74.9%) between Haplotype 1 and Haplotype 2. While the focus is primarily on the chromosome-level Haplotype 1, the authors do not provide a technical explanation for why Haplotype 2 is significantly less complete. This information is important for understanding the assembly outcome and should be included.
- In the *Methods* section, the authors provide detailed collection data and a voucher photograph (Figure 1). However, they do not state where the physical specimen has been deposited. For the sake of long-term verification, the authors should explicitly state the public collection or museum where the voucher specimen (or its remains) is housed.
- In the *Hi-C library preparation and sequencing* section, the authors write: "create equimolar and/or weighted 2.8 nM pools". The term "weighted" is unclear. If the authors mean that the pooling was adjusted based on specific criteria such as estimated genome size or library yield, this should be stated explicitly rather than implied through the pooling description.
- With regards to genome annotation, the authors state that the average transcript length is 20,934.66 bp. If "transcript" refers to processed mRNA, this length is unusually large and more consistent with the total gene span including introns. The authors should clarify whether 20,934 bp refers to the mRNA transcript length or the gene length.
- The authors state that **MitoHiFi** was used "to ensure the general quality of the sequence". "General quality" is a vague description of the results. The authors should provide specific metrics, such as circularization status or a completeness comparison against a reference mitogenome, to substantiate the quality of the mitochondrial assembly.

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: Population genetics, evolutionary biology, conservation genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 10 February 2026

https://doi.org/10.21956/wellcomeopenres.28411.r145598

Ian M Carr
University of Leeds, Leeds, UK

The genome sequence of the Carline Skipper, *Pyrgus carlinae* (Rambur, 1839) (Lepidoptera: Hesperiidae)

- Although genome assembly is a computationally intensive process, the authors do not describe the operating system used or the computational resources (RAM, CPU cores, or runtime) required for key steps such as the assembly. This information is important for reproducibility and should be included.
- In the *PacBio HiFi library preparation and sequencing* section, the authors state: “The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment.” This is a description of the kit, not of the procedures performed. The authors should explicitly state that they carried out end repair, A-tailing, adapter ligation, etc., rather than implying it through the kit contents.
- When the Femto Pulse system is first mentioned, the manufacturer and model are not provided, but they are included on the second mention. These should be swapped so that the full details appear at first mention.
- In the *Assembly curation* section, the authors write: “The assembly was decontaminated.” The term *decontaminated* typically refers to removal of hazardous substances. A clearer phrasing would be: “Extraneous or contaminant sequences were removed.”
- The authors state: “Sequences without hits are chunked.” The term *chunked* is unclear. If

they mean that sequences were split into smaller batches (e.g., using **seqtk**) prior to BLAST searches, this should be stated explicitly.
- In the first paragraph of the *Genome assembly* section, the authors describe using k-mer analysis to estimate genome size, but they do not reference Figure 2, which displays the k-mer distribution. Figure 2 should be cited here.
- In the final paragraph of the *Assembly quality assessment* section, the authors describe the use of **lep_busco_painter**, but do not mention that Figure 4 shows the resulting plot. Figure 4 should be referenced in this paragraph.
- In Figure 6, the BUSCO categories are displayed as a circular plot from 0–100%, but the values in the key sum to 102.1%. The “Comp (98.4%)” value appears to include single-copy, duplicated, and fragmented genes, even though the latter two categories are also shown separately. “Comp” should instead reflect only the single-copy BUSCOs (96.35%). The abbreviations “Comp”, “Frag”, “dupl”, and “missing” should also be defined. Additionally, the tick marks in the lower-left scale are rendered at a single-pixel width due to monitor antialiasing, making them visually indistinguishable; consider enlarging them, referencing larger segments, or removing them.
- In the *Genome annotation report* section, the authors state that the average transcript length is 20,934.66 bp. If “transcript” refers to processed mRNA, this length is unusually large and more consistent with gene length or unprocessed pre-mRNA. The authors should clarify whether 20,934 bp refers to the mRNA transcript length or the gene length. They also state that the average gene has 1.6 transcripts; it would be clearer to say that each gene has an average of 1.6 splice variants or transcript isoforms.

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

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.