



DATA NOTE

The genome sequence of the Northern Grizzled Skipper, *Pyrgus centaureae* (Rambur, 1839) (Lepidoptera: Hesperidae)

[version 1; peer review: 3 approved]

Vlad Dincă¹, Daniel Linke², Charlotte J. Wright³, Joana I. Meier³, Mark L. Blaxter³,
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

¹Grigore Antipa National Museum of Natural History, Bucharest, Romania²Biology Centre of the Czech Academy of Sciences, Institute of Entomology, České Budějovice, Czech Republic³Tree of Life Programme, Wellcome Sanger Institute, Hinxton, England, UK

V1 First published: 19 Mar 2026, 11:184
<https://doi.org/10.12688/wellcomeopenres.26169.1>
Latest published: 19 Mar 2026, 11:184
<https://doi.org/10.12688/wellcomeopenres.26169.1>

Abstract

We present a genome assembly from a female specimen of *Pyrgus centaureae* (Northern Grizzled Skipper; Arthropoda; Insecta; Lepidoptera; Hesperidae). The assembly contains two haplotypes with total lengths of 758.98 megabases and 671.41 megabases. Most of haplotype 1 (99.62%) 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 15.4 kilobases.

Keywords

Pyrgus centaureae; Northern 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			
19 Mar 2026	view	view	view

1. **Melody Clark** , Natural Environment Research Council, Cambridge, UK
2. **Pritha Dey** , National Centre for Biological Sciences-TIFR, Bengaluru, India
3. **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 authors: Charlotte J. Wright (cw22@sanger.ac.uk), Project Psyche Community (Mark.Blaxter@sanger.ac.uk)

Author roles: **Dincă V:** Investigation, Resources, Supervision, Writing – Review & Editing; **Linke D:** Writing – Original Draft Preparation; **Wright CJ:** Funding Acquisition, Project Administration, Supervision; **Meier JI:** Funding Acquisition, Project Administration, Supervision; **Blaxter ML:** Funding Acquisition, Project Administration, Supervision;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540). DL is funded by the Czech Science Foundation (GACR) grant 22-35084J and Technology Agency of the Czech Republic (TACR) grant SS07010197. MM receives funding from the Research Council of Finland, Grant 336449.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Dincă V *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Dincă V, Linke D, Wright CJ *et al.* **The genome sequence of the Northern Grizzled Skipper, *Pyrgus centaureae* (Rambur, 1839) (Lepidoptera: Hesperidae) [version 1; peer review: 3 approved]** Wellcome Open Research 2026, 11:184 <https://doi.org/10.12688/wellcomeopenres.26169.1>

First published: 19 Mar 2026, 11:184 <https://doi.org/10.12688/wellcomeopenres.26169.1>

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; Hesperidae; Pyrginae; *Pyrgus*; *Pyrgus centaureae* (Rambur, 1839) (NCBI:txid1666702).

Background

P. centaureae, commonly known as the Northern Grizzled Skipper, is a small butterfly species in the family Hesperidae with a Holarctic distribution. In Europe and Asia, it is present from northern Fennoscandia through northern Russia, the Urals, and the Altai Mountains. In the Nearctic, its range includes Canada and Alaska, extending south to New Mexico in the Rocky Mountains and North Carolina in the Appalachian Mountains (de Jong, 1975; Lotts & Naberhaus, 2025; Tolman & Lewington, 2009). It inhabits marshes, tundra, open hillsides, or shrublands at elevations from 0 to 2 600 m, while being mostly absent from coastal areas.

Adults of *P. centaureae* have a wingspan of 25–33 mm, with veins lined in white on the ventral side, which is unusual for *Pyrgus* species. Additionally, they display a distinctive white spot near the costa on the dorsal hindwing (Tolman & Lewington, 2009). Larval development takes two years in the subarctic, with eggs deposited during June to July (in southern North American populations, adults fly considerably earlier, from March to June (Chazal et al., 2004)). In the subarctic, larvae overwinter as second-instar caterpillars. In the second year, larvae complete development and overwinter as a pupa. Notably, larval host plants switch during development, at least in European subspecies: first-instar larvae prefer dwarf birch leaves (*Betula nana*) and shift to cloudberry leaves (*Rubus chamaemorus*) after the first moult (Wickman, 2012). Larvae construct shelters from folded birch or cloudberry leaves (Wickman, 2012). In addition to cloudberry, larvae may feed on other plants in the Rose family, such as wild strawberries (*Fragaria virginiana*) and cinquefoils (*Potentilla canadensis* or *P. diversifolia*) (Lotts & Naberhaus, 2025).

Genetically, *P. centaureae* is split between Palaearctic and Nearctic lineages. A recently described subspecies, *P. centaureae dzezh* Gorbunov, 2007, which occurs in eastern Siberia, has also been recorded from Alaska (Zhang et al., 2021), suggesting multiple relatively recent trans-Beringian dispersal events.

The IUCN assesses *P. centaureae* as Vulnerable, with decreasing populations (van Swaay et al., 2025a; van Swaay et al., 2025b). In North America, some subspecies require conservation action due to fragmented ranges, such as *P. c. wyandot* inhabiting the Appalachian Mountains (Chazal et al., 2004). Outside Europe, populations are generally secure, although the species may be rare in parts of its range, especially at the periphery (Lotts & Naberhaus, 2025).

We present a chromosome-level genome sequence of *P. centaureae*, sequenced as part of Project Psyche (Wright et al., 2025). The sequence data were derived from a female specimen (Figure 1) collected from Rauhala, Lapland, Finland.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult female *P. centaureae* (specimen ID SAN28000256, ToLID ilPyrCent2; Figure 1), collected from Rauhala, Lapland, Finland (latitude 67.8829, longitude 24.2031; elevation 265 m) on 2020-06-22. The specimen was collected and identified by Vlad Dincă.

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 ilPyrCent2 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the whole organism was homogenised by powermashing using a PowerMasher II tissue disruptor.

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol. DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor@3 for LI PacBio protocol. Sheared DNA was purified by automated SPRI (solid-phase reversible immobilisation). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 41.23 ng/μL and a yield of 1 937.81 ng, with a fragment size of 16.2 kb.



Figure 1. Voucher photograph of the *Pyrgus centaureae* (ilPyrCent2) specimen used for genome sequencing.

RNA was extracted from whole organism tissue of ilPyrCent2 in the Tree of Life Laboratory at the WSI using the [RNA Extraction: Automated MagMax™ mirVana protocol](#). The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

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 25 M 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 PRJEB80080.

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	ilPyrCent2	ilPyrCent2	ilPyrCent2
Specimen ID	SAN28000256	SAN28000256	SAN28000256
BioSample (source individual)	SAMEA115575572	SAMEA115575572	SAMEA115575572
BioSample (tissue)	SAMEA115575600	SAMEA115575600	SAMEA115575600

Table 1. *Continued*

Platform	PacBio HiFi	Hi-C	RNA-seq
Tissue	whole organism	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq X	Illumina NovaSeq X
Run accessions	ERR13660092	ERR13670051	ERR14986714
Read count total	2.22 million	677.22 million	132.97 million
Base count total	23.75 Gb	102.26 Gb	20.08 Gb

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the whole organism of the ilPyrCent2 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](#).

RNA-seq library preparation and sequencing

Libraries were prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs), following the manufacturer's instructions. Poly(A) mRNA in the total RNA solution was isolated using oligo (dT) beads, converted to cDNA, and uniquely indexed; 14 PCR cycles were performed. Libraries were size-selected to produce fragment lengths of 100–300 bp. Libraries were quantified, normalised, pooled to a final concentration of 2.8 nM, and diluted to 150 pM for loading. Sequencing was carried out on the Illumina NovaSeq X to generate 150-bp paired-end reads.

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

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

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included 12 breaks and 30 joins. This reduced the scaffold count by 11.7%, reduced scaffold N50 by 0.6%, and increased the total assembly length by 1.1%. The curation process is described at <https://gitlab.com/wtsi-grit/rapid-curation>. PretextView 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).

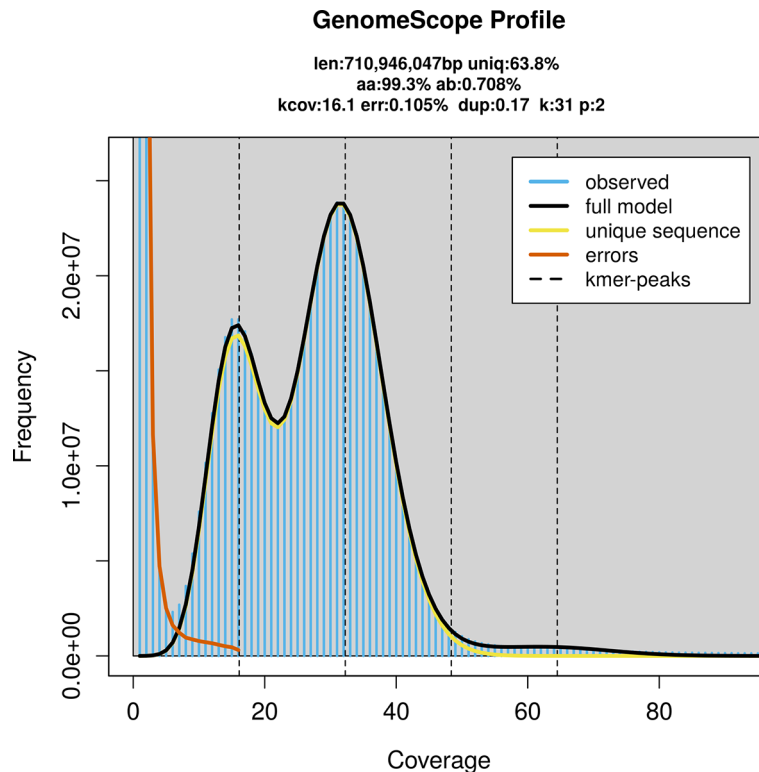


Figure 2. Frequency distribution of k -mers generated using GenomeScope2. The plot shows observed and modelled k -mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 2. Genome assembly statistics.

Assembly name	ilPyrCent2.hap1.1	ilPyrCent2.hap2.1
Assembly accession	GCA_964275275.1	GCA_964275205.1
Assembly level	chromosome	scaffold
Span (Mb)	758.98	671.41
Number of chromosomes	31	scaffold-level
Number of contigs	401	328
Contig N50	4.49 Mb	4.54 Mb
Number of scaffolds	127	89
Scaffold N50	26.32 Mb	25.96 Mb
Longest scaffold length (Mb)	55.72	-
Sex chromosomes	W and Z	-
Organelles	Mitochondrion: 15.4 kb	-

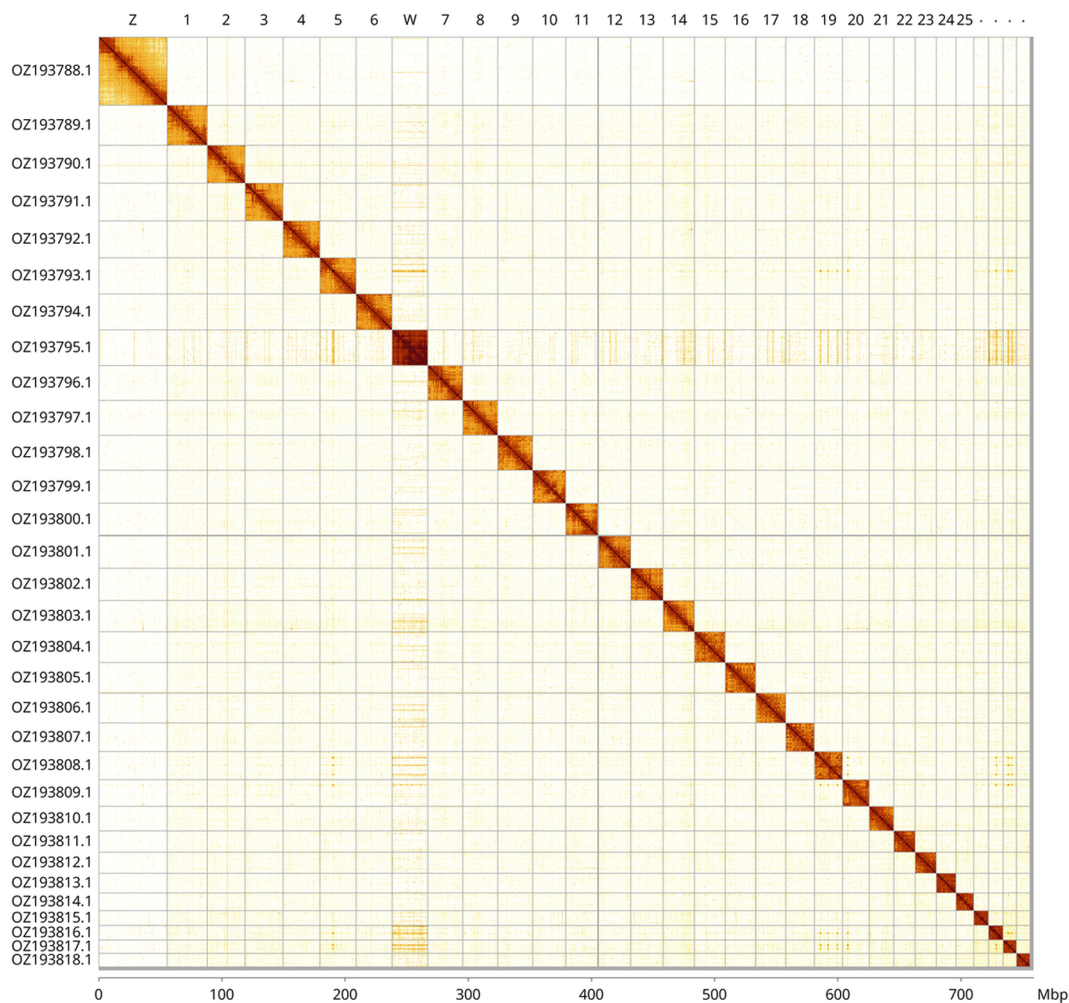


Figure 3. Hi-C contact map of the *Pyrgus centaureae* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

We used `lep_buscoPainter` 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 *P. centaureae* specimen generated 23.75 Gb (gigabases) from 2.22 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 711.77 Mb, with a heterozygosity of 0.71% and repeat content of 36.20% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 32× coverage. Hi-C sequencing produced 102.26 Gb from 677.22 million reads, which were used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. 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 758.98 Mb in 127 scaffolds, with 274 gaps, and a scaffold N50 of 26.32 Mb (Table 2).

Most of the assembly sequence (99.62%) was assigned to 31 chromosomal-level scaffolds, representing 29 autosomes and the W and Z sex chromosomes. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3 and 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).

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Pyrgus centaureae* ilPyrCent2.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ193789.1	1	32.42	36.50	M14;M2
OZ193790.1	2	30.82	37	M1
OZ193791.1	3	30.77	37	M17;M20
OZ193792.1	4	29.94	36.50	M8
OZ193793.1	5	29.45	36.50	M7
OZ193794.1	6	29.21	36.50	M9
OZ193796.1	7	28.51	36.50	M3
OZ193797.1	8	28.41	36.50	M5
OZ193798.1	9	28.27	36.50	M12
OZ193799.1	10	26.95	36.50	M18
OZ193800.1	11	26.49	37	M4
OZ193801.1	12	26.32	36.50	M16
OZ193802.1	13	26.06	36.50	M6
OZ193803.1	14	25.41	37	M10
OZ193804.1	15	25.14	36.50	M22
OZ193805.1	16	24.88	36.50	M21
OZ193806.1	17	24.26	37	M15
OZ193807.1	18	23.12	37	M11
OZ193808.1	19	22.90	37	M23
OZ193809.1	20	21.69	37	M13
OZ193810.1	21	20.02	37	M14

Table 3. Continued

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ193811.1	22	17.30	37.50	M24
OZ193812.1	23	17.24	37	M26
OZ193813.1	24	15.80	37.50	M28
OZ193814.1	25	14.50	37.50	M27
OZ193815.1	26	11.99	38	M25
OZ193816.1	27	11.78	39.50	M30
OZ193817.1	28	10.94	38	M31
OZ193818.1	29	10.92	38	M29
OZ193795.1	W	28.85	37.50	-
OZ193788.1	Z	55.72	36.50	M19;MZ

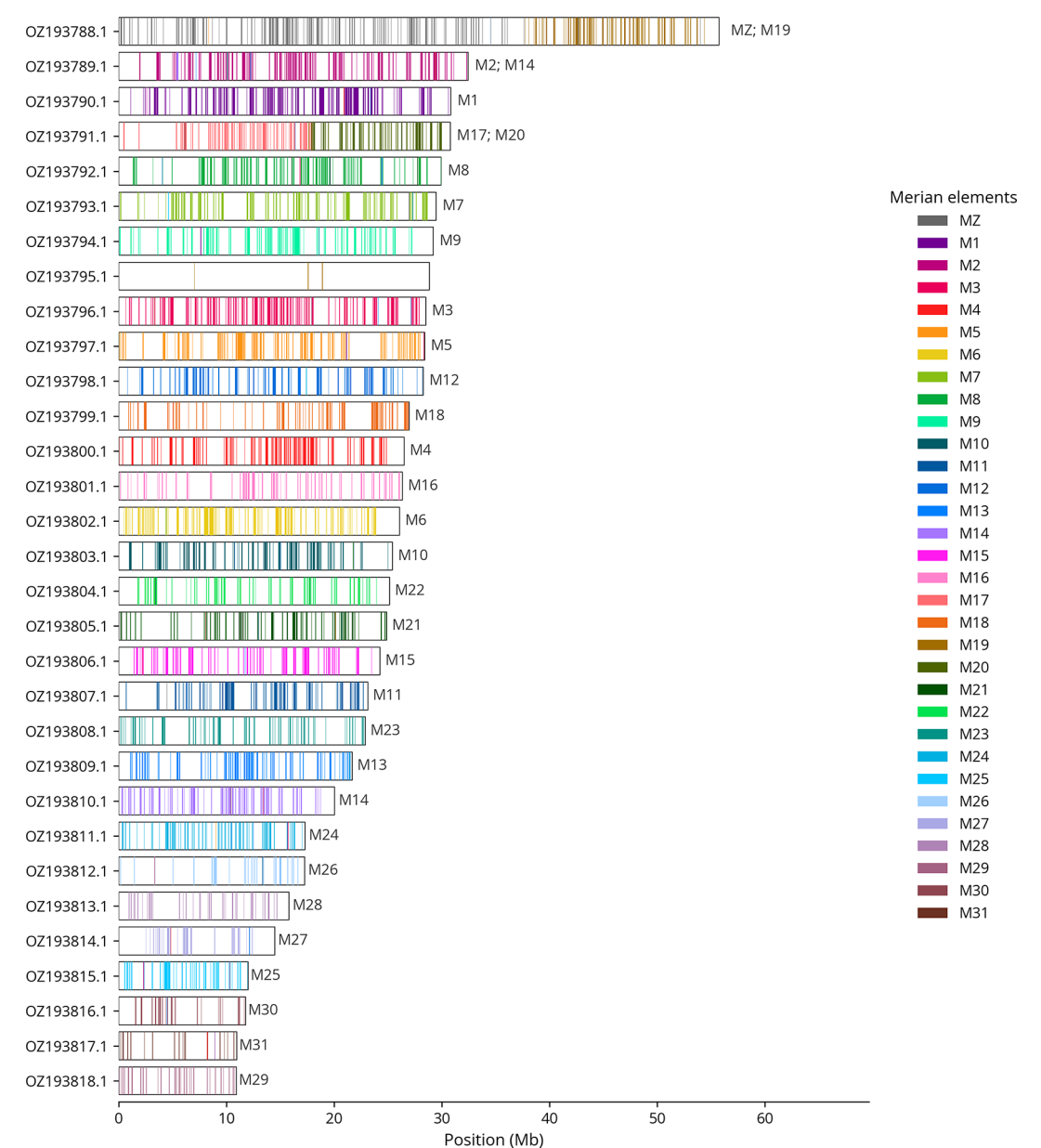


Figure 4. Merian elements painted across chromosomes in the ilPyrCent2.hap1.1 assembly of *Pyrgus centaureae*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

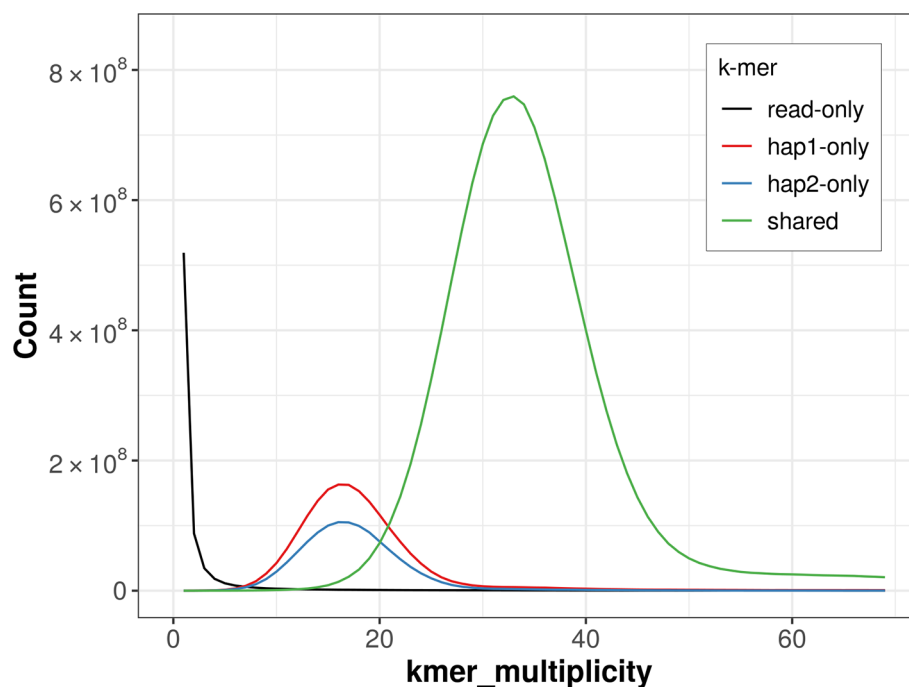


Figure 5. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

The mitochondrial genome was also assembled (length 15.4 kb, OZ193819.1). 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 62.4, and for haplotype 2, 62.0. When the two haplotypes are combined, the assembly achieves an estimated QV of 62.2. The *k*-mer completeness is 86.99% for haplotype 1, 80.25% for haplotype 2, and 99.31% for the combined haplotypes (Figure 5).

BUSCO analysis using the *lepidoptera_odb10* reference set ($n = 5\,286$) identified 98.5% of the expected gene set (single = 97.8%, duplicated = 0.6%) in haplotype 1. For haplotype 2, BUSCO analysis identified 91.8% of the expected gene set (single = 91.5%, duplicated = 0.3%).

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) and the Earth BioGenome Project Report on Assembly Standards January 2026. The EBP metric, calculated for the haplotype 1, is **6.C.Q62**, meeting the recommended reference standard.

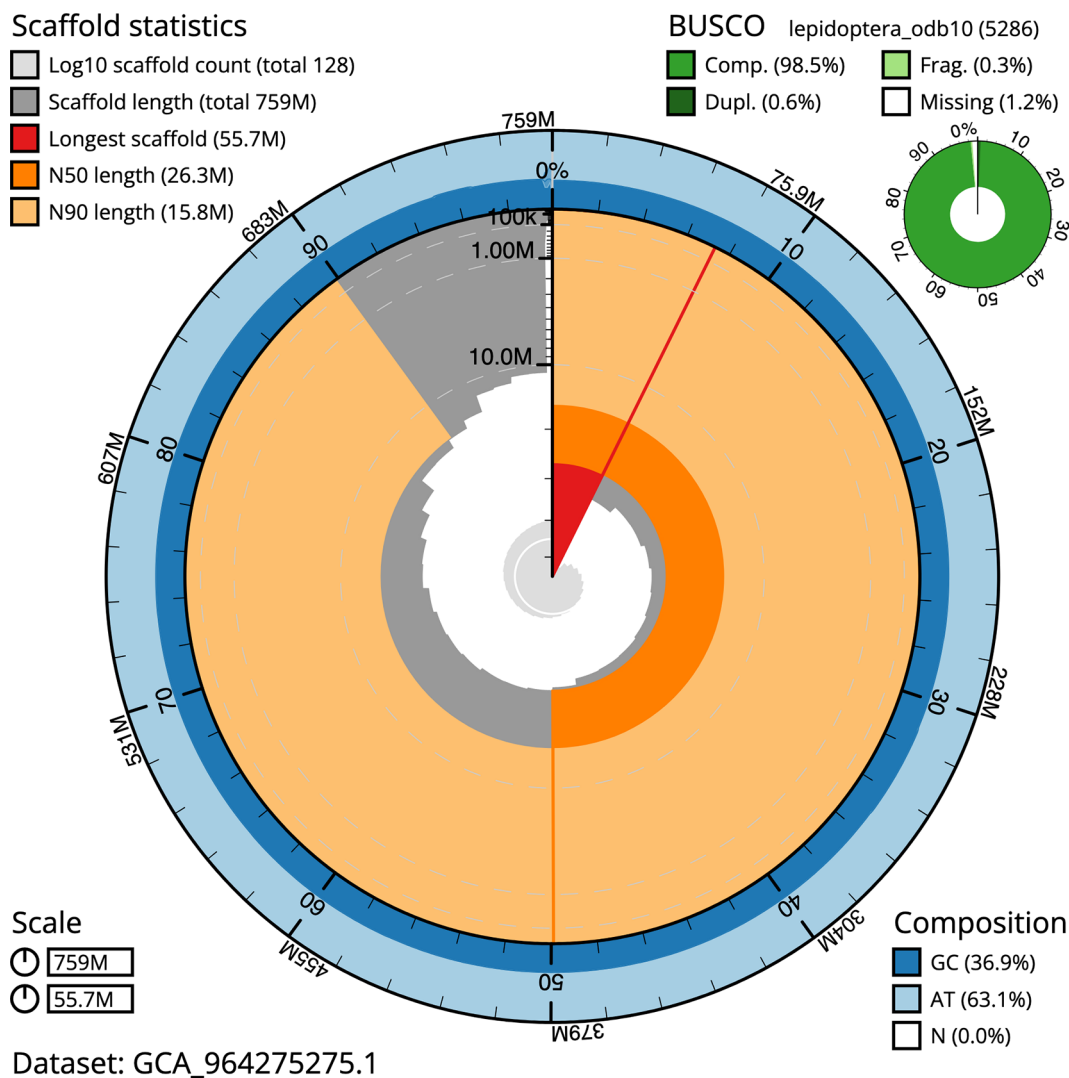


Figure 6. Assembly metrics for ilPyrCent2.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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](#).

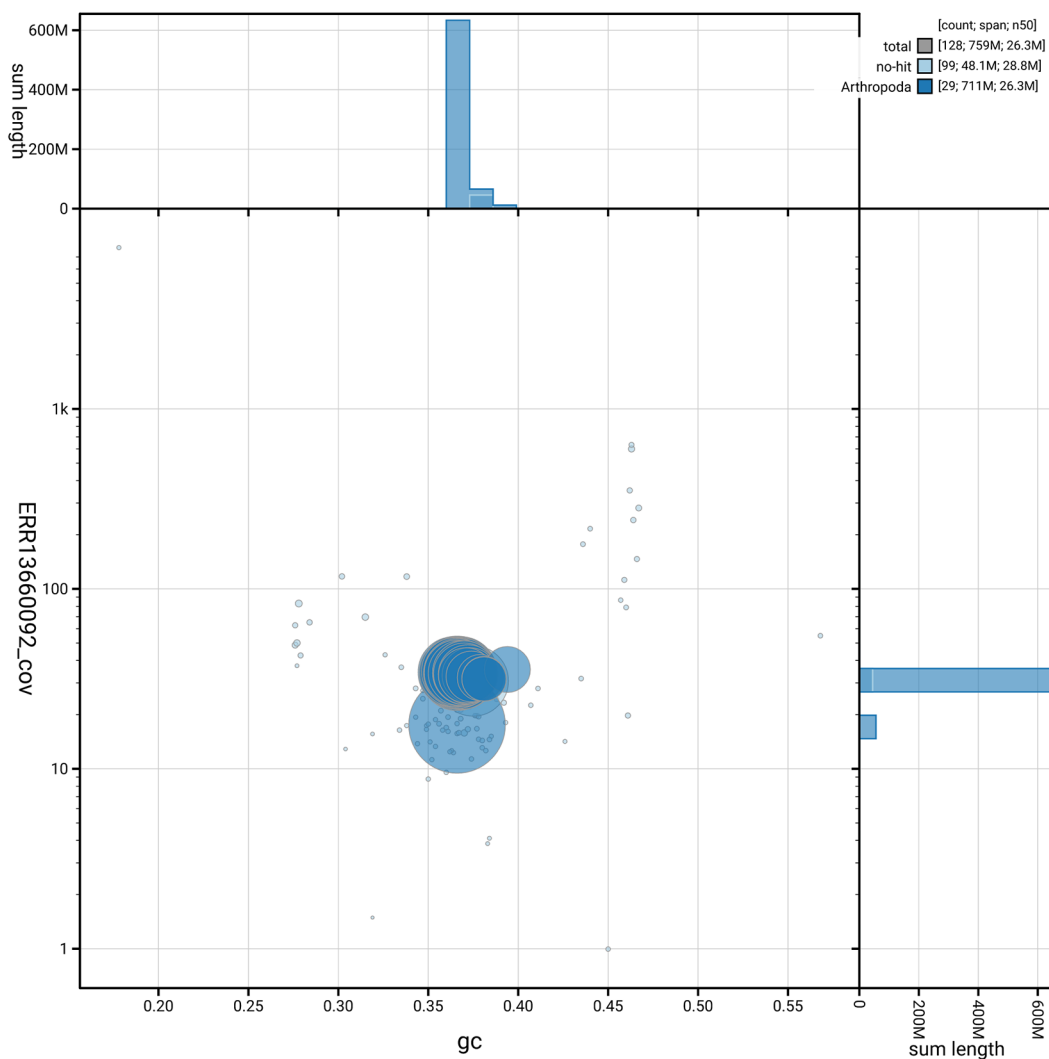


Figure 7. BlobToolKit GC-coverage plot for ilPyrCent2.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 centaureae* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q62	6.C.Q40
Contig N50 length	4.49 Mb	≥ 1 Mb
Scaffold N50 length	26.32 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 62.4; haplotype 2: 62.0; combined: 62.2	≥ 40
k-mer completeness	Haplotype 1: 86.99%; Haplotype 2: 80.25%; combined: 99.31%	≥ 95%
BUSCO	C:98.5% [S:97.8%; D:0.6%]; F:0.3%; M:1.2%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.62%	≥ 90%

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

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](#)

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: *P. centaureae*. Accession number [PRJEB80080](#). The genome sequence is released openly for reuse. The *P. centaureae* 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 [Tables 1](#) and [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.

Table 5. Software versions and sources.

Software	Version	Source
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
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

Table 5. *Continued*

Software	Version	Source
lep_buscoPainter	1.0.0	https://github.com/charlottewright/lep_buscoPainter
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
PretextView	0.0.5	https://github.com/sanger-tol/PretextView
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. [Publisher 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 M-J, 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 H-G: **Sensitive protein alignments at tree-of-life scale using DIAMOND.** *Nat. Methods.* 2021; **18**(4): 366–368. [Publisher Full Text](#)
- Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 Genes | Genomes | Genetics.* 2020; **10**(4): 1361–1374. [Publisher Full Text](#)
- Chazal AC, Roble SM, Hobson CS, *et al.*: **Status of the Appalachian grizzled skipper (*Pyrgus centaureae wyandoti*) in Virginia.** *Banisteria: a journal devoted to the natural history of Virginia.* 2004; **24**: 15–22.
- 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. [Publisher 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. [Publisher Full Text](#)
- Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2). [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- deJong R: **Notes on the genus *Pyrgus* (Lepidoptera, Hesperidae).** 1975; 1–11.
- Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat. Biotechnol.* 2017; **35**(4): 316–319. [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. [Publisher 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. [Publisher Full Text](#)
- Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: Lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025. [Publisher Full Text](#)
- Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1). [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. [Publisher 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. [Publisher Full Text](#)
- Li H: **Minimap2: Pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100. [Publisher Full Text](#)
- Lotts K, Naberhaus T: **Grizzled skipper, *Pyrgus centaureae* (Rambur, [1842]).** 2025. Butterflies; Moths of North America. [Reference Source](#)
- Manni M, Berkeley MR, Seppay 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. [Publisher Full Text](#)
- Merkel D: **Docker: Lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239). [Publisher Full Text](#)

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

[Publisher 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.

[Publisher 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.

[Publisher Full Text](#)

Rhie A, Walenz BP, Koren S, *et al.*: **Mercury: Reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1).

[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.* 2020; **2020**: baaa062.

[Publisher Full Text](#)

Tolman T, Lewington R: *Collins Butterfly Guide: The Most Complete Guide to the Butterflies of Britain and Europe.* HarperCollins Publishers; 2009.

Uliano-Silva M, Ferreira JGRN, Krashenninnikova K, *et al.*: **MitoHiFi: A Python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.

[Publisher Full Text](#)

van Swaay C, Ellis S, Warren M: ***Pyrgus carlinae*.** 2025a.

[Publisher Full Text](#)

van Swaay C, Warren M, Ellis S, *et al.*: **European Red List of butterflies: Measuring the pulse of European biodiversity.** 2025b.

[Publisher Full Text](#)

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE; 2019; 314–24.

[Publisher Full Text](#)

Wickman P-O: **Värdväxtbyte hos myrvisslaren, *Pyrgus centaureae* (Lepidoptera: Hesperidae).** *Entomol Tidskr.* 2012; **133**(3): 93–100.

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](#)

Wright CJ, Wahlberg N, Vila R, *et al.*: **Project Psyche: Reference genomes for all Lepidoptera in Europe.** *Trends Ecol. Evol.* 2025; **40**(12): 1234–1250.

[PubMed Abstract](#) | [Publisher Full Text](#)

Zhang J, Cong Q, Shen J, *et al.*: **Genomics-guided refinement of butterfly taxonomy.** *Taxon. Rep. Int. Lepidoptera Surv.* 2021; **9**(3): 1–55.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#) | [Reference Source](#)

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C scaffolding tool.** *Bioinformatics.* 2023; **39**(1).

[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:   

Version 1

Reviewer Report 01 June 2026

<https://doi.org/10.21956/wellcomeopenres.28821.r152560>

© 2026 Khedkar G. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Gulab D Khedkar 

Dr Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India

Summary of evaluation

The manuscript presents a high-quality genome assembly and comprehensive genomic resource for *Pyrgus centaureae* with strong technical rigor and well-documented analytical workflows. The study is timely and valuable, particularly considering the ecological and conservation relevance of this species. The assembly quality, inclusion of haplotype-resolved genomes, and integration of mitochondrial genome data substantially enhance the utility of the dataset for future evolutionary, ecological, and comparative genomic investigations in Lepidoptera. The manuscript is clearly written, technically sound, and supported by robust methodologies and appropriate bioinformatic pipelines.

The rationale for generating the genomic resource is adequately described, and the datasets are presented in a transparent and accessible manner. The methodological details are sufficient to ensure reproducibility. The use of PacBio HiFi sequencing, Hi-C scaffolding, and multiple quality assessment tools has resulted in a highly reliable genome assembly. In particular, the chromosomal-level scaffolding and incorporation of sex chromosomes significantly strengthen the overall quality and long-term utility of the resource.

A few minor suggestions could further improve the manuscript. First, the authors may consider expanding the discussion on how this genome resource could specifically contribute to future research in Lepidoptera systematics, adaptation, phylogeography, conservation genomics, and population-level analyses. A brief comparison with currently available skipper butterfly genomes may also help contextualize the significance of this assembly within HesperIIDae genomics.

Additionally, a concise note on possible applications of the dataset for climate resilience, habitat fragmentation studies, or comparative chromosome evolution would further broaden the impact of the work.

Overall, this is a valuable and well-executed genomic resource paper that will serve as an

important reference for future biodiversity genomics studies. The manuscript meets the standards for indexing, and I recommend its acceptance after minor revision.

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, conservation and 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 24 April 2026

<https://doi.org/10.21956/wellcomeopenres.28821.r152567>

© 2026 Dey P. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Pritha Dey 

National Centre for Biological Sciences-TIFR, Bengaluru, India

This manuscript presents a genome assembly for the Northern Grizzled Skipper, including both haplotypes and the mitochondrial genome. The study provides a valuable genomic resource for this species, particularly with the high level of completeness in haplotype 1, where the majority of the assembly is scaffolded into chromosomal pseudomolecules, including both sex chromosomes.

The assembly appears to be of good quality, and the inclusion of both haplotypes is a notable strength. The authors also added the IUCN assessment of the species which adds value to this work.

However, how this resource advances research in Lepidoptera or Hesperidae is not clearly highlighted. Expanding on potential applications of the genome (e.g., evolutionary, ecological, or conservation studies) would also strengthen the impact of the 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: Evolutionary ecology

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 23 April 2026

<https://doi.org/10.21956/wellcomeopenres.28821.r152569>

© 2026 Clark M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Melody Clark 

Natural Environment Research Council, Cambridge, UK

This Data Note describes the sequencing and assembly of a female specimen of the Northern Grizzled Skipper, *Pyrgus centaureae*. This species has a Holarctic distribution with the biogeography and feeding preferences of larvae well described. The North American population is labelled as "Vulnerable" due to habitat fragmentation, which provides added urgency for the need to sequence the genome. A haplotype level assembly is described with haplotype 1 scaffolded onto 31 pseudochromosomes including the W and Z sex chromosomes; haplotype 2 is to scaffold level. The mitochondrial genome has also been assembled. Thus, the genome is of very high quality. All the methods of DNA and RNA extraction are described and rely on well-known commercial kits, the methodologies are described and follow standard Sanger pipelines which are all in the public domain. PacBio and Hi-C were used to produce the final genome and the bioinformatics pipelines are again well described and follow standard Sanger practice, as do the descriptive analyses of the haplotypes. Application of Merian elements further annotates the genome relative to the ancestral Lepidopteran lineage.

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: transcriptomics and genomics of non-model species

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.
