DATA NOTE

The genome sequence of the Eastern Pale Clouded Yellow, *Colias erate* (Esper, 1805) (Lepidoptera: Pieridae)

[version 1; peer review: 2 approved]

Vlad Dincă[1], Daniel Linke[2], Dominik Schmitz[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]Grigore Antipa National Museum of Natural History, Bucharest, Romania
[2]Biology Centre of the Czech Academy of Sciences, Institute of Entomology, České Budějovice, Czech Republic
[3]Independent researcher, Berlin, Germany
[4]Tree of Life Programme, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 09 Feb 2026, **11**:102
https://doi.org/10.12688/wellcomeopenres.25891.1
Latest published: 09 Feb 2026, **11**:102
https://doi.org/10.12688/wellcomeopenres.25891.1

Abstract

We present a genome assembly from a male specimen of *Colias erate* (Eastern Pale Clouded Yellow; Arthropoda; Insecta; Lepidoptera; Pieridae). The assembly contains two haplotypes with total lengths of 324.54 megabases and 323.68 megabases. Most of haplotype 1 (99.67%) is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 15.19 kilobases. This work is part of Project Psyche, a collaborative programme generating genomes for European butterflies and moths.

Keywords

Colias erate; Eastern Pale Clouded Yellow; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 09 Feb 2026	✓ view	✓ view

1. **Emily Hornett**, University of Liverpool, Liverpool, UK

2. **Xue Qing**, Nanjing Agricultural University, Nanjing, China

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: Dincă V: Investigation, Resources, Supervision, Writing – Review & Editing; **Linke D**: Writing – Original Draft Preparation; **Schmitz D**: Data Curation; **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.
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: Dincă V, Linke D, Schmitz D *et al.* **The genome sequence of the Eastern Pale Clouded Yellow, *Colias erate* (Esper, 1805) (Lepidoptera: Pieridae) [version 1; peer review: 2 approved]** Wellcome Open Research 2026, **11**:102 https://doi.org/10.12688/wellcomeopenres.25891.1

First published: 09 Feb 2026, **11**:102 https://doi.org/10.12688/wellcomeopenres.25891.1

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; Papilionoidea; Pieridae; Coliadinae; *Colias*; *Colias erate* (Esper, 1805) (NCBI:txid72852)

Background

Colias erate, commonly referred to as the Eastern Pale Clouded Yellow, is a butterfly species in the Pieridae family. It is widely distributed, ranging from its western limit in central and south-eastern Europe, across Ukraine, southern European Russia, Caucasus, Transcaucasia, temperate and subtropical parts of Asia, parts of north-eastern Africa (e.g. Ethiopia, Somalia) (Slamka, 2004; Tolman & Lewington, 2009; Tshikolovets, 2011).

Colias erate is a migrant with strong dispersal abilities (Settele *et al.*, 2008) that has expanded its range towards western Europe and parts of the Balkans during the last decades of the 20th century (Tolman & Lewington, 2009). It was recently removed from the checklist of Cyprus (Hutsebaut & Couckuyt, 2021; John *et al.*, 2006). *Colias erate* typically occurs in open habitats such as meadows, flower-rich fields, scrublands, cultivated areas and favours drier environments (Dzurinka *et al.*, 2022), occurring at elevations up to 2 000 m (Tshikolovets, 2011).

The wingspan of *C. erate* is approximately 46–52 mm. In males, the dorsal side of the wings is typically lemon-yellow with a continuous black border, while females have the black border interrupted and display both yellow and whitish forms. *C. erate* is difficult to distinguish from other *Colias* species, especially *Colias crocea*, due to high phenotypic plasticity (Dzurinka *et al.*, 2021; Ellers & Boggs, 2002), DNA barcode sharing (Dincă *et al.*, 2011), and frequent hybridisation (Alberti, 1943; Dzurinka *et al.*, 2022). Sexes differ in colour reception, with females showing an extended red spectrum that may aid in host-plant detection for oviposition or play a role in courtship (Ogawa *et al.*, 2013). According to Hutsebaut and Couckuyt (2021), males of *C. erate* can be reliably separated from males of *C. crocea* due to different UV reflectance patterns (i.e. reflection absent in *C. erate* and present in *C. crocea*). The absence of an androconial patch towards the base of the hindwing upperside also seems indicative of *C. erate* (Hutsebaut & Couckuyt, 2021). The androconial patch is absent in the specimen analysed here, originating from south-eastern Romania (Figure 1).

Larvae feed on various species of Fabaceae, such as *Medicago*, *Trifolium*, *Onobrychis*, *Vicia* and *Melilotus* Clarke (2024). The butterfly is multivoltine, flying from May to October in two to four generations.

We present a chromosome-level genome sequence for the Eastern Pale Clouded Yellow, *Colias erate*, sequenced as part of Project Psyche (Wright *et al.*, 2025).

Figure 1. Voucher photograph of the *Colias erate* (ilColErat1) specimen used for genome sequencing.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult male *Colias erate* (specimen ID SAN28000227, ToLID ilColErat1; Figure 1), collected from Negureni, Constanta, Romania (latitude 44.089, longitude 27.748) on 2023-06-27. The specimen was collected and identified by Vlad Dinca (University of Oulu).

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 ilColErat1 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 21.82 ng/μL and a yield of 1 025.54 ng, with a fragment size of 13.4 kb.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), according to the manufacturer's instructions. The kit includes

reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit. The sample was sequenced on a Revio instrument (Pacific Biosciences). The prepared library was normalised to 2 nM, and 15 µL was used for making complexes. Primers were annealed and polymerases bound to generate circularised complexes, following the manufacturer's instructions. Complexes were purified using 1.2X SMRTbell beads, then diluted to the Revio loading concentration (200–300 pM) and spiked with a Revio sequencing internal control. The sample was sequenced on a Revio 25M SMRT cell. The SMRT Link software (Pacific Biosciences), a web-based workflow manager, was used to configure and monitor the run and to carry out primary and secondary data analysis.

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

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the head of the ilColErat1 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) 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

Table 1. Specimen and sequencing data for BioProject PRJEB84273.

Platform	PacBio HiFi	Hi-C
ToLID	ilColErat1	ilColErat1
Specimen ID	SAN28000227	SAN28000227
BioSample (source individual)	SAMEA115811660	SAMEA115811660
BioSample (tissue)	SAMEA116115751	SAMEA116115750
Tissue	thorax	head
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR14151105	ERR14171994
Read count total	2.11 million	657.92 million
Base count total	16.42 Gb	99.35 Gb

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 six breaks and 19 joins. 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 *Colias erate* specimen generated 16.42 Gb (gigabases) from 2.11 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 332.95 Mb, with a heterozygosity of 1.79% and repeat content of 28.84% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 48× coverage. Hi-C sequencing produced 99.35 Gb from 657.92 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 324.54 Mb in 86 scaffolds, with 47 gaps, and a scaffold N50 of 11.34 Mb (Table 2).

Most of the assembly sequence (99.67%) was assigned to 31 chromosomal-level scaffolds, representing 30 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 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 chromosome was identified based on BUSCO gene painting with ancestral Merian elements.

The mitochondrial genome was also assembled (length 15.19 kb, OZ228901.1). 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 66.7, and for haplotype 2, 66.1. When the two haplotypes are combined, the assembly achieves an estimated QV of 66.4. The *k*-mer completeness is 67.87% for haplotype 1, 67.86% for haplotype 2, and 96.56% for the combined haplotypes (Figure 5). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.9% of the expected gene set (single = 98.6%, duplicated = 0.3%) in haplotype 1. For haplotype 2,

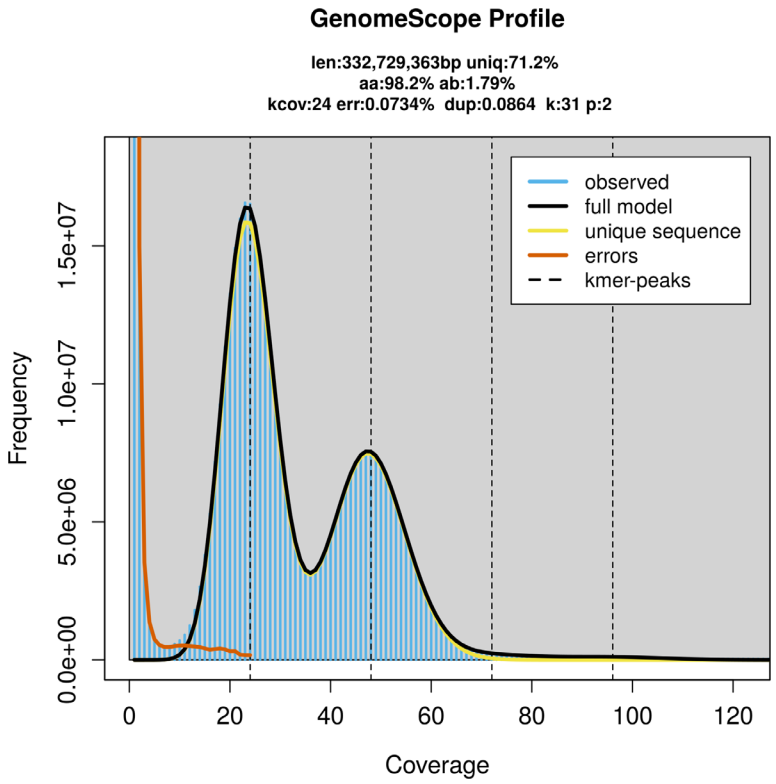

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 2. Genome assembly statistics.

Assembly name	ilColErat1.hap1.1	ilColErat1.hap2.1
Assembly accession	GCA_965153205.1	GCA_965153295.1
Assembly level	chromosome	scaffold
Span (Mb)	324.54	323.68
Number of chromosomes	31	scaffold-level
Number of contigs	133	112
Contig N50	6.15 Mb	6.67 Mb
Number of scaffolds	86	60
Scaffold N50	11.34 Mb	11.37 Mb
Longest scaffold length (Mb)	17.33	-
Sex chromosomes	Z	-
Organelles	Mitochondrion: 15.19 kb	-

BUSCO analysis identified 98.9% of the expected gene set (single = 98.6%, 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.

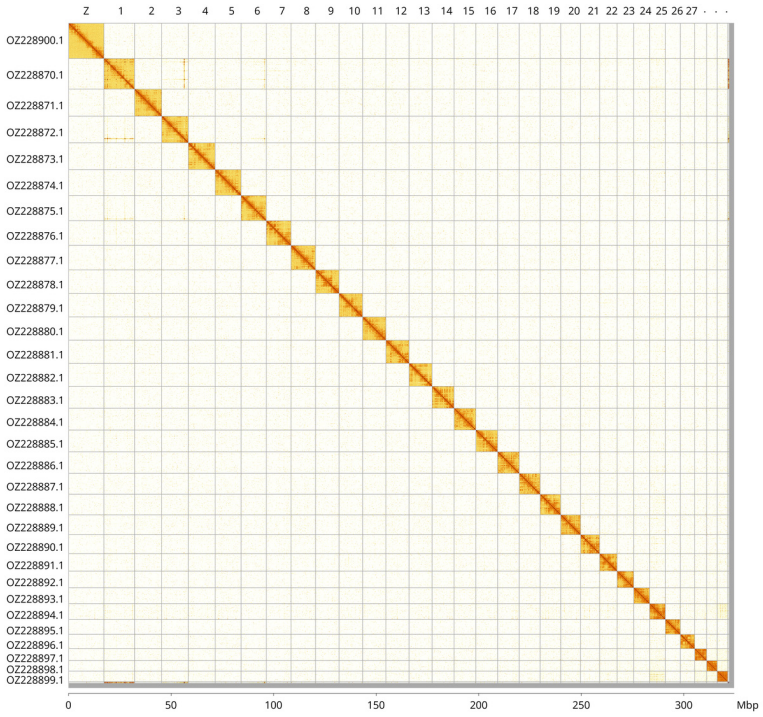

Figure 3. Hi-C contact map of the *Colias erate* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Colias erate* ilColErat1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ228870.1	1	16.58	34	M2
OZ228871.1	2	13.24	34	M1
OZ228872.1	3	13.03	34	M3
OZ228873.1	4	13.03	33.50	M17;M20
OZ228874.1	5	12.68	33.50	M8
OZ228875.1	6	12.30	33	M9
OZ228876.1	7	12	33	M12
OZ228877.1	8	11.99	33.50	M5
OZ228878.1	9	11.52	33.50	M18
OZ228879.1	10	11.44	33.50	M16
OZ228880.1	11	11.35	33.50	M7
OZ228881.1	12	11.34	33.50	M6
OZ228882.1	13	11.21	33.50	M4
OZ228883.1	14	10.67	33	M21
OZ228884.1	15	10.67	33.50	M10
OZ228885.1	16	10.59	33.50	M15
OZ228886.1	17	10.57	33.50	M22
OZ228887.1	18	10.15	34	M11
OZ228888.1	19	10.02	33.50	M14
OZ228889.1	20	9.73	33.50	M13
OZ228890.1	21	9.26	33.50	M23
OZ228891.1	22	8.54	33.50	M19
OZ228892.1	23	8.07	36	M30
OZ228893.1	24	8.05	33	M24
OZ228894.1	25	7.75	33	M26
OZ228895.1	26	7.33	33	M28
OZ228896.1	27	6.95	33	M27
OZ228897.1	28	5.80	33.50	M25
OZ228898.1	29	5.20	33.50	M29
OZ228899.1	30	5.09	34.50	M31
OZ228900.1	Z	17.33	34	MZ

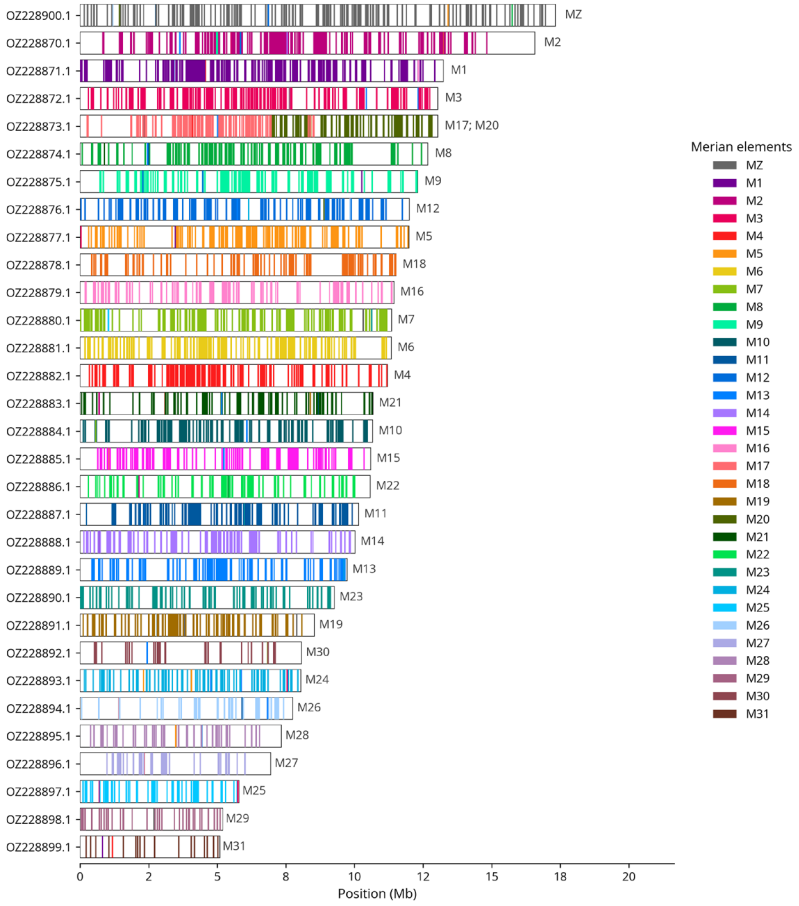

Figure 4. Merian elements painted across chromosomes in the ilColErat1.hap1.1 assembly of *Colias erate*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

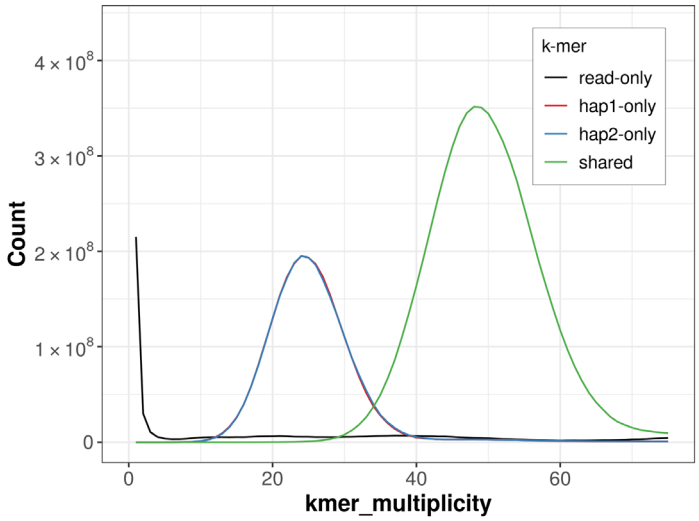

Figure 5. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

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.Q66**, meeting the recommended reference standard.

Wellcome Sanger Institute – Legal and Governance

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

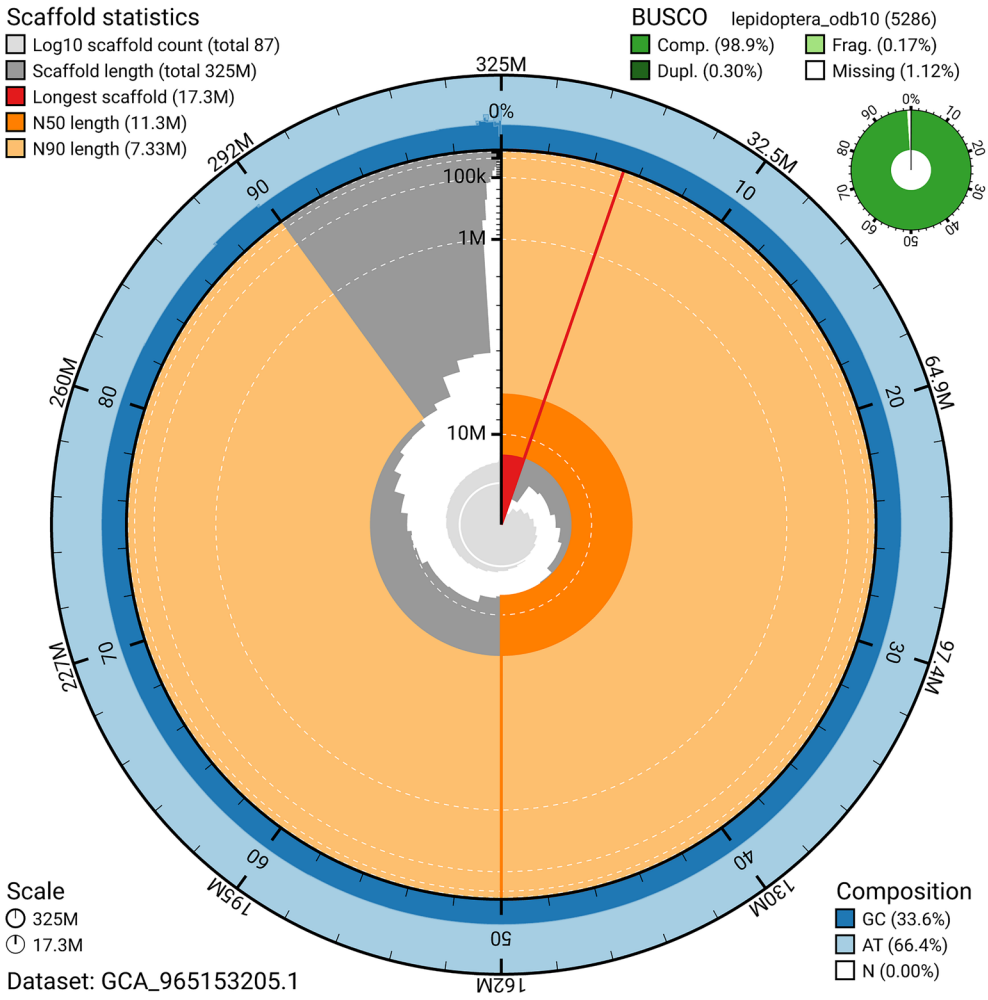

Figure 6. Assembly metrics for ilColErat1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

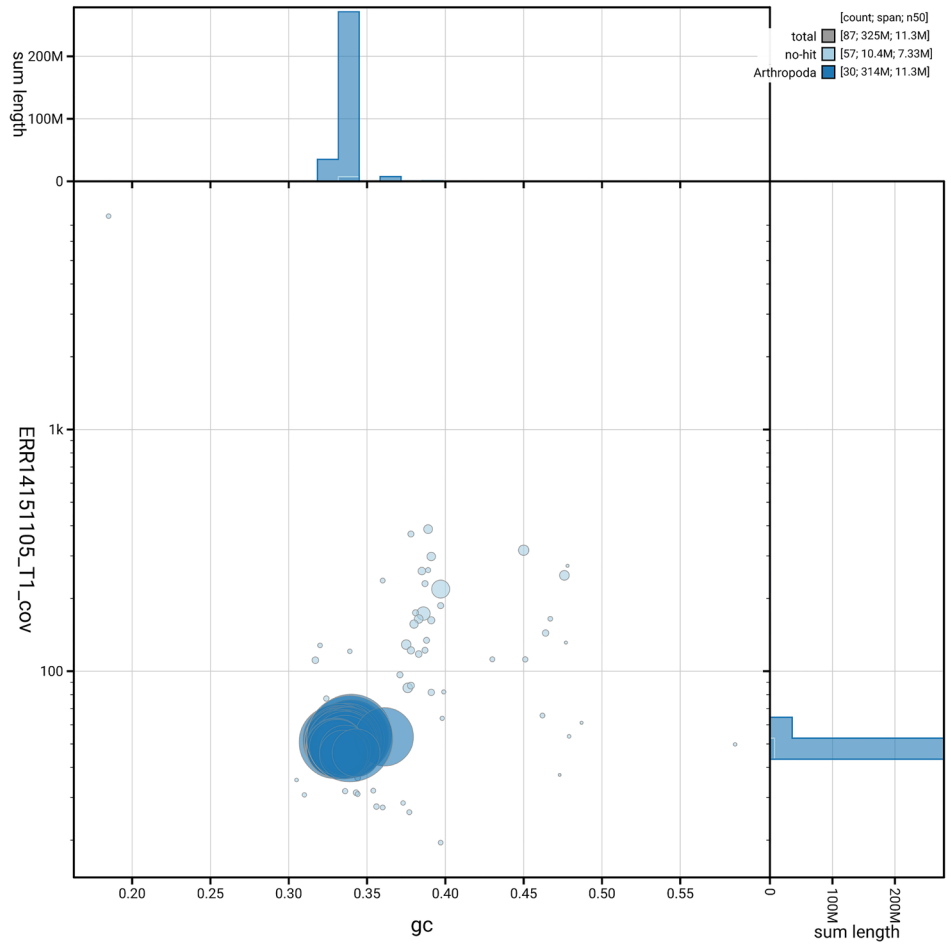

Figure 7. BlobToolKit GC-coverage plot for ilColErat1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Colias erate* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q66	6.C.Q40
Contig N50 length	6.15 Mb	≥ 1 Mb
Scaffold N50 length	11.34 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 66.7; haplotype 2: 66.1; combined: 66.4	≥ 40
k-mer completeness	Haplotype 1: 67.87%; Haplotype 2: 67.86%; combined: 96.56%	≥ 95%
BUSCO	C:98.9% [S:98.6%; D:0.3%]; F:0.2%; M:0.9%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.67%	≥ 90%

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

Data availability

European Nucleotide Archive: Colias erate (eastern pale clouded yellow). Accession number PRJEB84273. The genome sequence is released openly for reuse. The *Colias erate* genome sequencing initiative is part of the Sanger Institute Tree of Life Programme (PRJEB43745) and Project Psyche (PRJEB71705). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through Ensembl at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

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

Author information

Contributors are listed at the following links:

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

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
lep_busco_painter	1.0.0	https://github.com/charlottewright/lep_busco_painter
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.6.0	https://github.com/sanger-tol/blobtoolkit

Software	Version	Source
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	1.2.2	https://github.com/c-zhou/yahs

References

Alberti B: **Zur Frage der Hybridisation zwischen *Colias erate* Esp., *hyale* L. Und *edusa* F. Und *ueber* Umgrenzung der 3 Arten.** *Mitt Munch Entomol Ges.* 1943; **33**: 606–25.

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

Clarke HE: **A checklist of European butterfly larval foodplants.** *Ecol Evol.* 2024; **14**(1): e10834.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

Dincă V, Zakharov EV, Hebert PDN, *et al.*: **Complete DNA barcode reference library for a country's butterfly fauna reveals high performance for temperate Europe.** *Proc Biol Sci.* 2011; **278**(1704): 347–55.
PubMed Abstract | Publisher Full Text | Free Full Text

Dzurinka M, Mutanen M, Šemeláková M, *et al.*: **The intraspecific variability of *Colias croceus* (Geoffroy, 1785) and *C. Erate* (Esper, 1805) (Lepidoptera, Pieridae) from the perspective of comparative morphology.** *Zoomorphology.* 2021; **140**(3): 353–363.
Publisher Full Text

Dzurinka M, Šemeláková M, Panigaj Ľ: **Taxonomy of hybridizing *Colias croceus* (Geoffroy, 1785) and *Colias erate* (Esper, 1805) (Lepidoptera, Pieridae) in light of mitochondrial and nuclear DNA, with occurrence and effects of *Wolbachia* infection.** *Zool Anz.* 2022; **299**: 73–81.
Publisher Full Text

Ellers J, Boggs CL: **The evolution of wing color in *Colias* butterflies: heritability, sex linkage, and population divergence.** *Evolution.* 2002; **56**(4): 836–40.
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

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

Hutsebaut J, Couckuyt J: **UV images greatly facilitate the taxonomy of *Colias croceus* (Geoffroy in Fourcroy, 1785) and *Colias erate* (Esper, [1805]) forms (Lepidoptera, Papilionoidea, Pieridae).** *Lambillionea.* 2021; **CXXI**(1): 68–83.

John E, Coutsis J, Makris C: **A review of records for *Colias erate* (Esper, [1805]) (Lep.: Papilionoidea Pieridae) in Cyprus: Were they all yellow forms of *Colias croceus* (Geoffroy, 1785)?** *Entomol Gaz.* 2006; **57**: 3–12.
Reference Source

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

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

Ogawa Y, Kinoshita M, Stavenga DG, *et al.*: **Sex-specific retinal pigmentation results in sexually dimorphic long-wavelength-sensitive photoreceptors in the eastern pale clouded yellow butterfly, *Colias erate*.** *J Exp Biol.* 2013; **216**(Pt 10): 1916–23.
PubMed Abstract | 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.
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**

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

Slamka F: **Die Tagfalter Mitteleuropas-Östlicher Teil: Bestimmung, Biotope Und Bionomie, Verbreitung, Gefährdung.** Slamka, 2004.
Reference Source

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

Tshikolovets VV: **Butterflies of Europe and the mediterranean area.** Pardubice: Tshikolovets Publications, 2011.
Reference Source

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

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

Watanabe M, Nakanishi Y: **Population structure and dispersals of the sulfur butterfly *Colias erate* (Lepidoptera: Pieridae) in an isolated plain.** *Jpn J Ent.* 1996; **64**(1): 17–29.

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**

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**

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 07 April 2026

https://doi.org/10.21956/wellcomeopenres.28498.r148914

✔ **Xue Qing**
Nanjing Agricultural University, Nanjing, China

This manuscript presents a high-quality, chromosome-level genome assembly for *Colias erate* . The methods are technically sound, and the assembly meets EBP reference standards. I only have one comment:

Haplotype 1 is assembled to chromosome level, while Haplotype 2 remains at scaffold level. No explanation is provided for why Hi-C data could not be used to scaffold Haplotype 2 similarly.State whether this was due to coverage, heterozygosity, or Hi-C mapping limitations. If possible, provide a Hi-C contact map or alignment statistics for Haplotype 2 in supplementary materials.

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: phylogeny, genomics, Metazoa

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 02 April 2026

https://doi.org/10.21956/wellcomeopenres.28498.r149683

Emily Hornett
University of Liverpool, Liverpool, UK

This manuscript reports the genome of the butterfly *Colias erate*, which was sequenced and assembled as part of Project Psyche, a collaborative programme aiming to sequence the genomes of European butterflies and moths.

I have two very minor comments for the introduction:

- change "while females have the black border interrupted" to "while females have the black border interrupted with yellow blotches" or similar
- add sex to figure 1 legend (male)

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: Insect and microbe genomics, symbiosis, evolution

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.