DATA NOTE

The genome sequence of the Portuguese Dappled White, *Euchloe tagis* (Hübner, 1804) (Lepidoptera: Pieridae)

[version 1; peer review: 4 approved]

Mattia Menchetti[1], Joan Carles Hinojosa[1], Roger Vila[1], Charlotte J. Wright[2], Joana I. Meier[2], Mark L. Blaxter[2],
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]Institut de Biologia Evolutiva (CSIC-UPF), Barcelona, Catalonia, Spain
[2]Tree of Life Programme, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 09 Feb 2026, **11**:100
https://doi.org/10.12688/wellcomeopenres.25883.1
Latest published: 09 Feb 2026, **11**:100
https://doi.org/10.12688/wellcomeopenres.25883.1

Abstract

We present a genome assembly from a male specimen of *Euchloe tagis* (Portuguese Dappled White; Arthropoda; Insecta; Lepidoptera; Pieridae). The assembly contains two haplotypes with total lengths of 309.40 megabases and 308.42 megabases. Most of haplotype 1 (99.74%) 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 15.31 kilobases. Gene annotation of this assembly on Ensembl identified 11 716 protein-coding genes. This work is part of Project Psyche, a collaborative programme generating genomes for European butterflies and moths.

Keywords

Euchloe tagis; Portuguese Dappled White; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ✓ ✓

	1	2	3	4
version 1 09 Feb 2026	✓ view	✓ view	✓ view	✓ view

1. **Melody Clark**, Natural Environment Research Council, Cambridge, UK
2. **Temitope Opeyemi Oriowo**, Leibniz Institute for the Analysis of Biodiveristy Change, Germany, Germany
3. **Gurinder Kaur Walia**, Punjabi University, Patiala, India
4. **Tianzhu Xiong**, Cornell University, Ithaca, USA

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: Menchetti M: Investigation, Resources, Writing – Original Draft Preparation; **Hinojosa JC**: Investigation, Resources; **Vila R**: Investigation, Resources, 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). RV was funded by grant PID2022-139689NB-I00 (MICIU/ AEI/ 10.13039/501100011033 and ERDF, EU) and grant 2021-SGR-00420 (Departament de Recerca i Universitats, Generalitat de Catalunya).
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: Menchetti M, Hinojosa JC, Vila R *et al.* **The genome sequence of the Portuguese Dappled White, *Euchloe tagis* (Hübner, 1804) (Lepidoptera: Pieridae) [version 1; peer review: 4 approved]** Wellcome Open Research 2026, **11**:100 https://doi.org/10.12688/wellcomeopenres.25883.1

First published: 09 Feb 2026, **11**:100 https://doi.org/10.12688/wellcomeopenres.25883.1

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; Pierinae; Pierini; *Euchloe*; *Euchloe*; *Euchloe tagis* (Hübner, 1804) (NCBI:txid415341)

Background

The Portugese Dappled White, *Euchloe tagis* (Hübner, 1804), is a western Mediterranean butterfly, present in north-west Italy, France, the Iberian Peninsula and parts of north Africa (Morrocco and Algeria) (Marabuto *et al.*, 2020). The species inhabits xerothermophilous rocky limestone areas on dense Mediterranean shrubland (Marabuto *et al.*, 2020). It is a small Pieridae that resembles the more common *Euchloe crameri* Butler, 1896, from which it can be distinguished morphologically (having a curved hindwing costa) and via DNA barcoding (Dincă *et al.*, 2017). It belongs to the subgenus *Iberochloe*, which some authors consider a distinct genus (Back *et al.*, 2008). The species is univoltine and larvae feed on flowers and young seed capsules of various species of candytufts (genus *Iberis*, Brassicaceae). It hibernates at the pupal stage, with adults typically flying between February and May (García-Barros *et al.*, 2013).

Despite its current IUCN status of Least Concern at the European level (van Swaay *et al.*, 2025), the species is distributed in small and fragmented populations, often threatened by habitat loss (e.g., in Iberia and France, Olivares & Back, 2004; OPIE-Proserpine, 2009) and overgrazing (e.g., in Morocco, Tennent, 1996). This scattered distribution is primarily due to its ecological specialisation, which is closely tied to the presence of its host plants and the geological characteristics of the area (Marabuto *et al.*, 2020).

Populations likely diversified and expanded during the Pleistocene, after having split between Africa and Europe during the Messinian Salinity Crisis (5.96–5.33 Mya) (Marabuto *et al.*, 2020). Current populations show high intraspecific mtDNA diversity (Dapporto *et al.*, 2022; Marabuto *et al.*, 2020) and more than 10 subspecies have been described (Manil *et al.*, 2021). The Algerian taxon *pechi* Staudinger, 1885 is either treated as a subspecies of *E. tagis* or as a different species (Manil *et al.*, 2021).

Euchloe tagis has a karyotype of 31 chromosomes (de Lesse, 1970). The genome sequence we present here will provide a valuable resource for reconstructing the species' intricate phylogeographic history and the taxonomic status of the highly divergent north African taxa, and to inform conservation genetics assessments.

We present a chromosome-level genome sequence for the Portuguese Dappled White, *Euchloe tagis*, sequenced as part of Project Psyche (Wright *et al.*, 2025). The sequence data were derived from a male specimen (Figure 1) collected from La Granja d'Escarp, Segrià, Catalonia, Spain.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult male *Euchloe tagis* (specimen ID SAN28000292, ToLID ilEucTagi1; Figure 1), collected from La Granja d'Escarp, Segrià, Catalonia, Spain. (latitude 41.4007, longitude 0.3583) on 2021-03-25. The specimen was collected by Roger Vila, Mattia Menchetti and Joan Carles Hinojosa and formally identified by Joan Carles Hinojosa and Roger Vila.

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 ilEucTagi1 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.

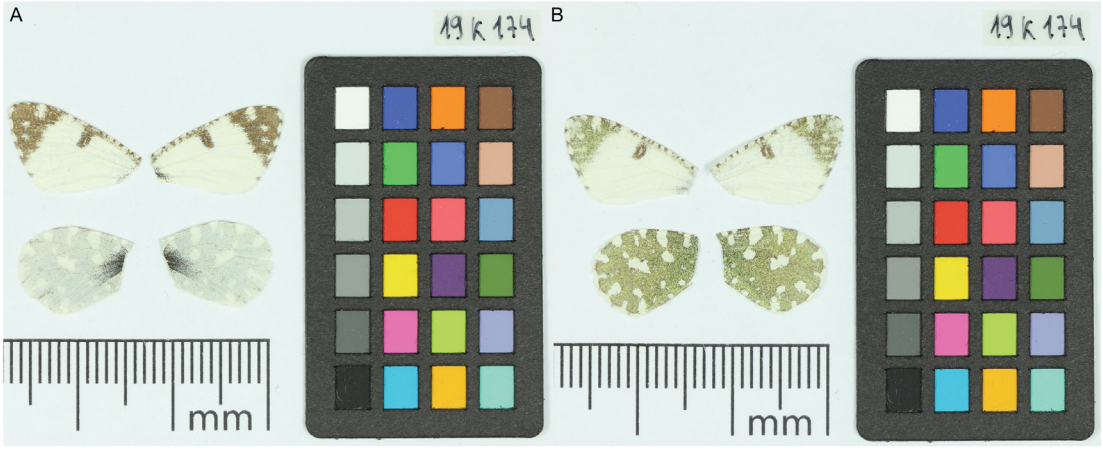

Figure 1. Voucher photographs of the *Euchloe tagis* (ilEucTagi1) specimen used for genome sequencing.

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 36.63 ng/μL and a yield of 1 721.61 ng, with a fragment size of 15.5 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 whole organism of the ilEucTagi1 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

Table 1. Specimen and sequencing data for BioProject PRJEB80941.

Platform	PacBio HiFi	Hi-C
ToLID	ilEucTagi1	ilEucTagi1
Specimen ID	SAN28000292	SAN28000292
BioSample (source individual)	SAMEA115768757	SAMEA115768757
BioSample (tissue)	SAMEA115768862	SAMEA115768862
Tissue	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13800491	ERR13802633
Read count total	1.67 million	683.64 million
Base count total	18.52 Gb	103.23 Gb

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 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 14 breaks and 32 joins. This reduced the scaffold count by 27.4% and increased scaffold N50 by 0.7%. 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 *Euchloe tagis* specimen generated 18.52 Gb (gigabases) from 1.67 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 305.57 Mb, with a heterozygosity of 2.12% and repeat content of 15.61% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 58× coverage. Hi-C sequencing produced 103.23 Gb from 683.64 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 309.40 Mb in 68 scaffolds, with 58 gaps, and a scaffold N50 of 10.61 Mb (Table 2).

Most of the assembly sequence (99.74%) 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). Chromosome Z was assigned based on synteny to the genome of *Euchloe simplonia* (GCA_964265075.1).

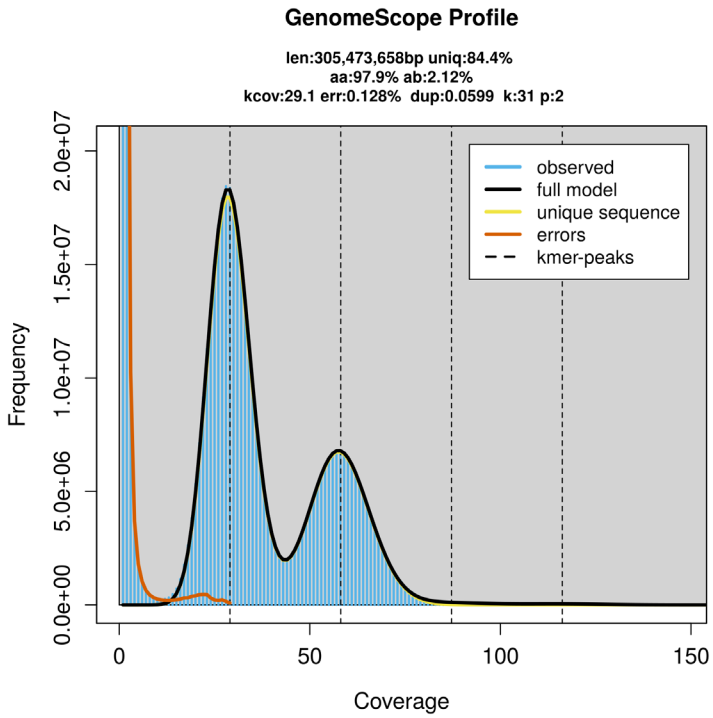

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 2. Genome assembly statistics.

Assembly name	ilEucTagi1.hap1.1	ilEucTagi1.hap2.1
Assembly accession	GCA_964277245.1	GCA_964277155.1
Assembly level	chromosome	scaffold
Span (Mb)	309.40	308.42
Number of chromosomes	31	scaffold-level
Number of contigs	126	112
Contig N50	6.68 Mb	5.64 Mb
Number of scaffolds	68	60
Scaffold N50	10.61 Mb	10.59 Mb
Longest scaffold length (Mb)	13.33	-
Sex chromosomes	Z	-
Organelles	Mitochondrion: 15.31 kb	-

The mitochondrial genome was also assembled (length 15.31 kb, OZ195029.1). 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 64.5, and for haplotype 2, 65.9. When the two haplotypes are combined, the assembly achieves an estimated QV of 65.1. The *k*-mer completeness is

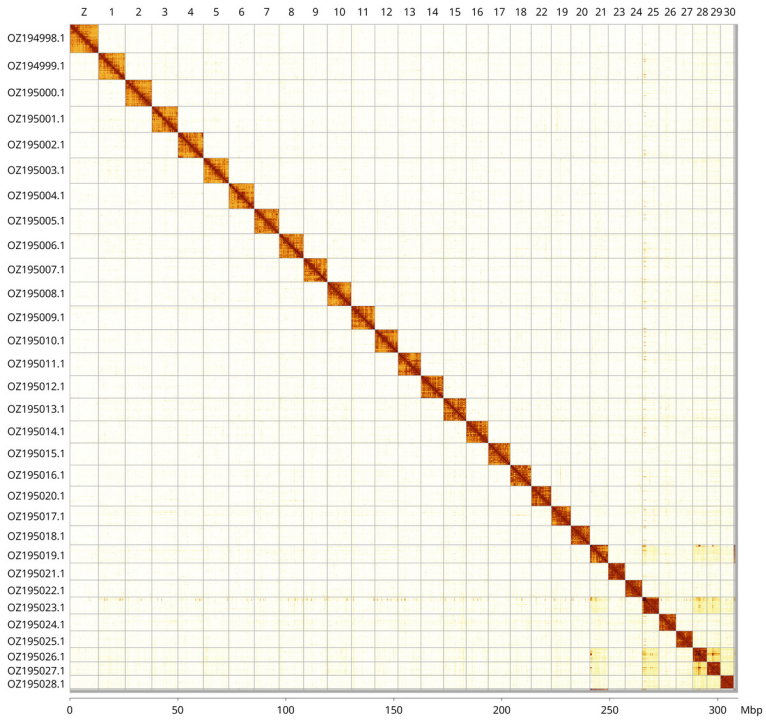

Figure 3. Hi-C contact map of the *Euchloe tagis* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Euchloe tagis* ilEucTagi1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ194999.1	1	12.47	33.50	M1
OZ195000.1	2	12.24	33	M2
OZ195001.1	3	12.05	33.50	M3
OZ195002.1	4	11.82	33	M9
OZ195003.1	5	11.79	33.50	M17;M20
OZ195004.1	6	11.78	33	M8
OZ195005.1	7	11.51	33.50	M5
OZ195006.1	8	11.29	33.50	M16
OZ195007.1	9	11.06	33	M18
OZ195008.1	10	11.05	33.50	M7
OZ195009.1	11	10.97	33	M12
OZ195010.1	12	10.66	33.50	M4
OZ195011.1	13	10.61	33.50	M6
OZ195012.1	14	10.46	33	M21
OZ195013.1	15	10.46	34	M22
OZ195014.1	16	10.27	33.50	M15
OZ195015.1	17	10.17	33.50	M10
OZ195016.1	18	9.74	33.50	M11
OZ195017.1	19	9.24	33.50	M13
OZ195018.1	20	9.08	34	M14
OZ195019.1	21	8.84	33.50	M23
OZ195020.1	22	9.62	34	M27
OZ195021.1	23	7.88	33.50	M28
OZ195022.1	24	7.87	34	M24
OZ195023.1	25	7.85	37.50	M29
OZ195024.1	26	7.82	33.50	M26
OZ195025.1	27	7.70	34	M19
OZ195026.1	28	6.67	34.50	M31
OZ195027.1	29	6.22	34.50	M30
OZ195028.1	30	6.08	34	M25
OZ194998.1	Z	13.33	33.50	MZ

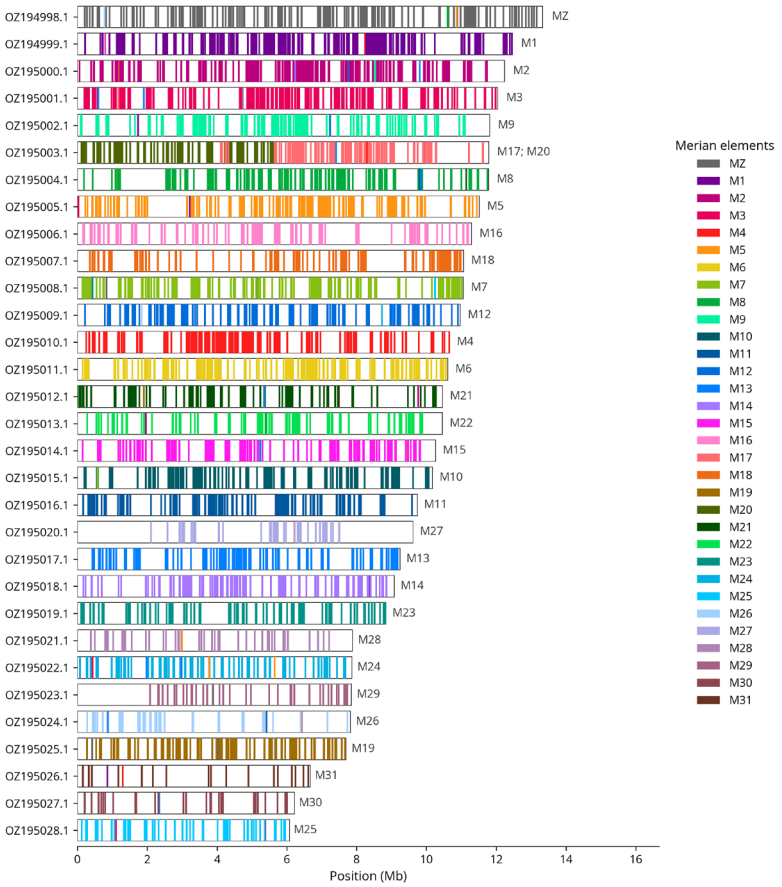

Figure 4. Merian elements painted across chromosomes in the ilEucTagi1.hap1.1 assembly of *Euchloe tagis*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

67.30% for haplotype 1, 67.29% for haplotype 2, and 99.69% for the combined haplotypes (Figure 5). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.8% of the expected gene set (single = 98.6%, duplicated = 0.2%) in haplotype 1. For haplotype 2, BUSCO analysis identified 98.9% of the expected gene set (single = 98.7%, duplicated = 0.2%). 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 Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.Q64**, meeting the recommended reference standard.

Genome annotation report

The *Euchloe tagis* genome assembly (GCA_964277245.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 21 335 transcribed mRNAs from 11 716 protein-coding and 1 631 non-coding genes. The average transcript length is 12 200.23 bp, with an average of 1.60 coding transcripts per gene and 7.34 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).

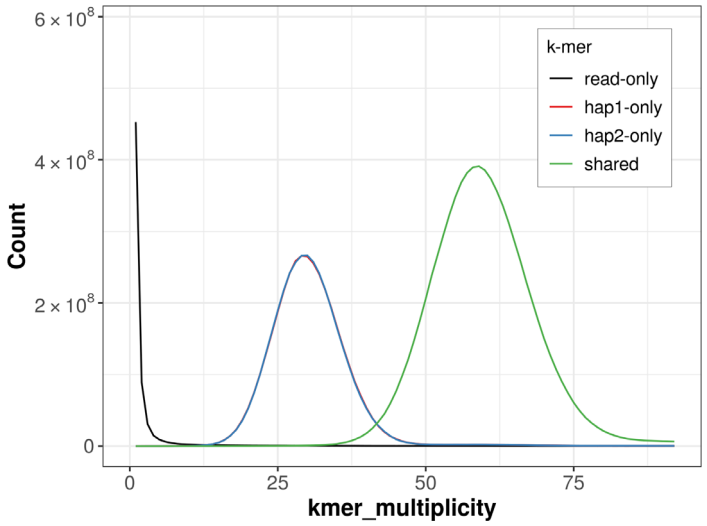

Figure 5. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

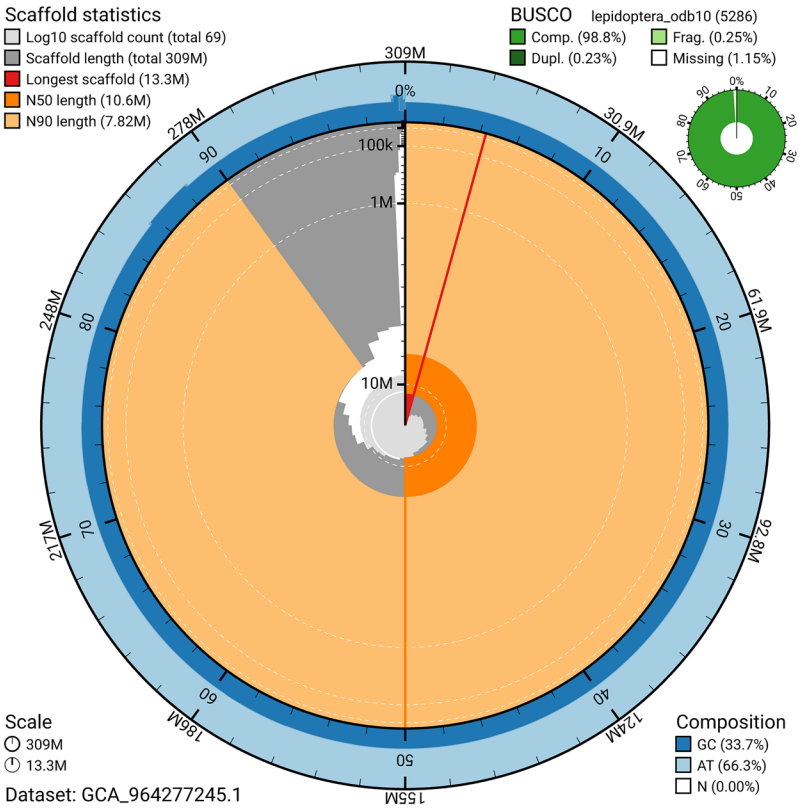

Figure 6. Assembly metrics for ilEucTagi1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

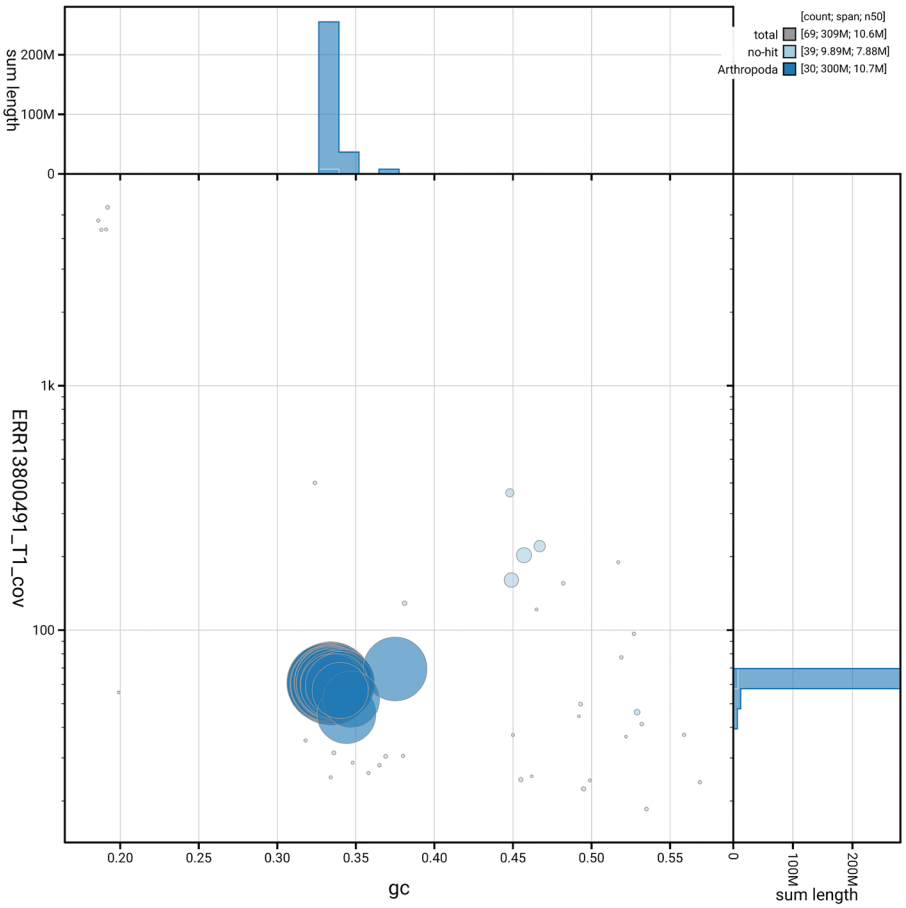

Figure 7. BlobToolKit GC-coverage plot for ilEucTagi1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Euchloe tagis* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q64	6.C.Q40
Contig N50 length	6.68 Mb	≥ 1 Mb
Scaffold N50 length	10.61 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 64.5; haplotype 2: 65.9; combined: 65.1	≥ 40
k-mer completeness	Haplotype 1: 67.30%; Haplotype 2: 67.29%; combined: 99.69%	≥ 95%
BUSCO	C:98.8% [S:98.6%; D:0.2%]; F:0.2%; M:0.9%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.74%	≥ 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

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: Euchloe tagis (Portugese dappled white). Accession number PRJEB80941 (https://identifiers.org/ena.embl/PRJEB80941). The genome sequence is released openly for reuse. The *Euchloe tagis* 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	4.3.9	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

Software	Version	Source
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.6.0	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

Back W, Knebelsberger TH, Miller MA: **Molekularbiologische untersuchungen und systematik der palaearktischen arten von *Euchloe* Hübner, [1819] (Lepidoptera: Pieridae).** *Entomologische Zeitschrift.* 2008; **118**(4): 151–69.
Reference Source

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

Dapporto L, Menchetti M, Vodă R, *et al.*: **The atlas of mitochondrial genetic diversity for Western palaearctic butterflies.** *Glob Ecol Biogeogr.* 2022; **31**(11): 2184–90.
Publisher Full Text

de Lesse H: **Formules chromosomiques de quelques Rhopalocères paléarctiques [lep.].** *Bull Soc Entomol France.* 1970; **75**(7): 214–16.
Publisher Full Text

Dincă V, Talavera G, Vila R: **First record of *Euchloe tagis* (Hübner, 1804) in the province of Tarragona (Catalonia) based on morphology and DNA data (Lepidoptera: Pieridae).** *Buttl Soc Catalana Lepidopterol.* 2017; **107**: 7–15.
Reference Source

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

García-Barros E, Munguira ML, Stefanescu C, *et al.*: **Lepidoptera. Papilionoidea**. Madrid: Museo Nacional de Ciencias Naturales, 2013; **37**.
Reference Source

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

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

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

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): D807–11.
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

Manil L, Lalanne-Cassou B, Casini PM: ***Euchloe (Iberochloe) tagis* (Hübner, 1804) distribution and differentiation in France and in Italy and description of two new subspecies (Lepidoptera: Pieridae).** *Lépidopteres.* 2021; **30**(78): 20–42.
Reference Source

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | Publisher Full Text | Free Full Text

Marabuto E, Pina-Martins F, Rebelo MT, *et al.*: **Ancient divergence, a crisis of salt and another of ice shaped the evolution of the west mediterranean butterfly *Euchloe tagis*.** *Biol J Linn Soc.* 2020; **131**(3): 487–504.
Publisher Full Text

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

Olivares J, Back W: ***Euchloe tagis* (Hübner, [1804]) (Lepidoptera, Pieridae) en Andalousie et description d'une sous-espèce nouvelle.** *Linneana Belgica.* 2004; **19**: 229–40.

OPIE-Proserpine: **Atlas Des Papillons de Jour de provence-alpes-côte d'Azur**. Turriers: Naturalia Publications, 2009.

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**

Tennent WJ: **The butterflies of Algeria, morocco and Tunisia**. Wallingford: Gem Publishing Co., 1996.
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, Warren M, Ellis S, *et al.*: **European red list of butterflies: measuring the pulse of European biodiversity**. Brussels: European Commission, 2025.
Reference Source

Vasimuddin Md, 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

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): 1234–50.
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 23 April 2026

https://doi.org/10.21956/wellcomeopenres.28490.r152546

✔ **Tianzhu Xiong**
Cornell University, Ithaca, New York, USA

This report of the Euchloe tagis genome covers an important species of conservation value. A great amount of natural history is introduced, followed by a detailed method section on genome assembly strategies. The final assembly is at the chromosome level.

My only question is why only one haplotype is selected to be a chromosome-level assembly? It seems that hap2 has a smaller number of contigs and scaffolds, with similar N50. The total length difference between hap1 and hap2 is only 1Mb. Is there any intrinsic reason that hap2 is not called a chromosome-level assembly?

Apart from this minor point, the report is pretty complete and should be good to index.

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: Genetics 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 21 April 2026

https://doi.org/10.21956/wellcomeopenres.28490.r152543

Gurinder Kaur Walia
Punjabi University, Patiala, Punjab, India

This manuscript presents a high-quality chromosome-level genome assembly of *Euchloetagis*. The study is well-designed, and the methodologies used are robust and appropriate. The assembly metrics are excellent, with high completeness and strong chromosomal resolution, making this a valuable genomic resource for Lepidoptera research.
The manuscript is clearly written, and the data are well presented and accessible. The integration of Hi-C scaffolding and Merian element mapping further strengthens the study.
Minor suggestions:

- The manuscript primarily focuses on how the genome was sequenced and the quality of the assembly.
- However, relatively limited emphasis is given to the biological insights derived from the genome.
- A brief addition on how this genome facilitates comparative analysis of *Euchloe* with other butterfly taxa would enhance the study.
- Inclusion of insights into evolutionary history, adaptive mechanismsand host-plant associations would further strengthen the manuscript.
- Highlighting its relevance for ecological studies and conservation applications would improve the overall impact.
- Minor typographical inconsistencies should be corrected, particularly the spelling of “Portuguese,” which appears as both “Portuguese” and “Portugese” in the manuscript.

Recommendation:

- Accept with minor revisions.

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

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

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

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: molecular cytogenetics

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 21 April 2026

https://doi.org/10.21956/wellcomeopenres.28490.r152144

Temitope Opeyemi Oriowo
Leibniz Institute for the Analysis of Biodiveristy Change, Germany, Germany

- Were other K-mers tested? I always recommend trying some different kmers and then deciding which has the lowest error rate

- Overall, I think it is a well-written genome note, and aside from being a nice resource to have, especially in the face of such taxonomic uncertainty, it is technically sound too.

- All linked resources are available too

-Article approved.

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: I work in population genomics, and genome assemblies are part of what I do

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 12 March 2026

https://doi.org/10.21956/wellcomeopenres.28490.r148587

Melody Clark
Natural Environment Research Council, Cambridge, UK

This Data Note describes the genome sequencing of the Portuguese Dappled White, *Euchloe tagis* under the Project Psyche. The biology and biogeography of this butterfly is well described, as is the rationale for genome sequencing. Although this butterfly is listed as "Least Concern" under the IUCN Red List, it is threatened by habitat loss and overgrazing in some European regions. It is also highly constrained by its specificity to its host plants and the geological characteristics of the preferred area, namely rocky limestone. The genome sequence of a male specimen is of very high quality, comprising 30 pseudochromosomes and the Z sex chromosome, plus the mitochondrial genome. The DNA extraction and sequencing methodologies, plus assembly and annotation are well described and use standard protocols and tools that are in the public domain. The genome has been annotated by Ensembl. All data are in the public domain. There is also mapping data of Merian elements, relating this genome to the 32 ancestral linkage groups of Lepidoptera.

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.