DATA NOTE

The genome sequence of the Mountain Dappled White, *Euchloe simplonia* (Boisduval, 1828) (Lepidoptera: Pieridae)

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

Yannick Chittaro[1], Daniel Linke[2], Kay Lucek[3], Charlotte J. Wright[4], Joana I. Meier[4], Mark L. Blaxter[4]

[1]Info fauna, Neuchâtel, Switzerland
[2]Institute of Entomology, Biology Centre CAS, České Budějovice, Czech Republic
[3]University of Neuchâtel, Neuchâtel, UK
[4]Tree of Life, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 15 Aug 2025, **10**:446
https://doi.org/10.12688/wellcomeopenres.24695.1
Latest published: 15 Aug 2025, **10**:446
https://doi.org/10.12688/wellcomeopenres.24695.1

Abstract

We present a genome assembly from a male specimen of *Euchloe simplonia* (Mountain Dappled White; Arthropoda; Insecta; Lepidoptera; Pieridae). The assembly contains two haplotypes with total lengths of 397.32 megabases and 398.10 megabases. Most of haplotype 1 (99.26%) 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 15.23 kilobases.

Keywords

Euchloe simplonia, Mountain Dappled White, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ? ✓ ✓

	1	2	3
version 1 15 Aug 2025	? view	✓ view	✓ view

1. **Christopher B. Cunningham**, University of Georgia, Georgia, USA
2. **Matthew Merkin**, University of Liverpool, Liverpool, UK
3. **Panagiotis Ioannidis**, Foundation for Research & Technology - Hellas, Crete, Greece

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Charlotte J. Wright (cw22@sanger.ac.uk)

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

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540). DL was supported by Czech Science Foundation (GACR) grant 22-35084J and Technology Agency of the Czech Republic (TACR) grant SS07010197, and KL was supported by the Swiss National Science Foundation (SNSF) Grant ID 202869.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Chittaro Y, Linke D, Lucek K *et al.* **The genome sequence of the Mountain Dappled White, *Euchloe simplonia* (Boisduval, 1828) (Lepidoptera: Pieridae) [version 1; peer review: 2 approved, 1 approved with reservations]** Wellcome Open Research 2025, **10**:446 https://doi.org/10.12688/wellcomeopenres.24695.1

First published: 15 Aug 2025, **10**:446 https://doi.org/10.12688/wellcomeopenres.24695.1

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 simplonia* (Boisduval, 1828) (NCBI:txid547582)

Background

Euchloe simplonia, known as the Mountain Dappled White, is a high-altitude butterfly species within the Pieridae family. Its distribution is restricted to three mountain ranges in Europe, the Cantabrian Mountains in North-Western Spain, the Pyrenees and the south-western Alps (Caritg *et al.*, 2011; Dubatolov & Kosterin, 1994; Tolman & Lewington, 2009).

The species favours alpine and subalpine meadows, rocky slopes, and grassy areas at elevations between 500 and 2 200 metres; however, it usually occurs at elevations above 1 500 metres. Adults are univoltine, flying from late April to late July, with flight periods varying based on altitude and local conditions (Tolman & Lewington, 2009).

E. simplonia is characterised by white wings with black markings on the forewings, which can be quite variable. The underside of the hindwings displays greenish marbling, providing effective camouflage among vegetation. Larval host plants include species such as *Biscutella laevigata*, *Iberis spathulata*, and *Erucastrum nasturtiifolium* (Tolman & Lewington, 2009).

E. simplonia is considered as least concern by IUCN (IUCN, 2025). However, due to its specialised habitat, limited range and disconnected populations, *E. simplonia* might be endangered in parts of its distribution. Conservation efforts should focus on preserving its alpine meadow habitats and monitoring population trends.

We present a chromosome-level, haplotype-resolved genome sequence of the Mountain Dappled White, *Euchloe simplonia*, sequenced as part of Project Psyche. The sequence data were derived from a male specimen (Figure 1) collected from Zeneggen VS, Switzerland.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult male *Euchloe simplonia* (specimen ID SAN28000120, ToLID ilEucSimp1; Figure 1), collected from Zeneggen VS, Switzerland (latitude 46.2751, longitude 7.8583; elevation 1 400 m) on 05/06/2023. The specimen was collected and identified by Yannick Chittaro (Info Fauna, Neuchâtel, Switzerland).

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

Figure 1. Voucher photograph of the *Euchloe simplonia* (ilEucSimp1) specimen used for genome sequencing.

(Howard *et al.*, 2025). The ilEucSimp1 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 72.41 ng/μL and a yield of 3 403.27 ng, with a fragment size of 15.6 kb. The 260/280 spectrophotometric ratio was 1.9, and the 260/230 ratio was 2.68.

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), following the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed

Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

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

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

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen head tissue of the ilEucSimp1 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

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

Prior to sequencing, libraries were normalised to 10 ng/µL. Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq X.

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

Genome assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts ($k = 31$) was generated from the filtered reads using FastK. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were assembled using Hifiasm in Hi-C phasing mode (Cheng *et al.*, 2021; Cheng *et al.*, 2022), producing two haplotypes. Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). Contigs were further scaffolded with Hi-C data in YaHS (Zhou *et al.*, 2023), using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated

Table 1. Specimen and sequencing data for BioProject.

Platform	PacBio HiFi	Hi-C
ToLID	ilEucSimp1	ilEucSimp1
Specimen ID	SAN28000120	SAN28000120
BioSample (source individual)	SAMEA115109934	SAMEA115109934
BioSample (tissue)	SAMEA115109951	SAMEA115109952
Tissue	thorax	head
Sequencing platform and model	Revio	Illumina NovaSeq X
Run accessions	ERR13485714	ERR13474179
Read count total	1.93 million	653.17 million
Base count total	18.46 Gb	98.63 Gb

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 5 breaks and 20 joins. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

Chromosomal painting was performed using lep_busco_painter using Merian elements, which represent the 32 ancestral linkage groups in Lepidoptera (Wright *et al.*, 2024). Painting was based on gene locations from the lepidoptera_odb10 BUSCO analysis and chromosome lengths from the genome index produced using SAMtools faidx (Danecek *et al.*, 2021). Each complete BUSCO (including both single-copy and duplicated BUSCOs) was assigned to a Merian element using a reference database, and coloured positions were plotted along chromosomes drawn to scale.

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

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

Genome sequence report

Sequence data

The genome of a specimen of *Euchloe simplonia* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 18.46 Gb (gigabases) from 1.93 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 394.47 Mb, with a heterozygosity of 1.13% and repeat content of 33.61%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 46× coverage. Hi-C sequencing produced 98.63 Gb from 653.17 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The genome was assembled into two haplotypes using Hi-C phasing. Haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The final assembly has a total length of 397.32 Mb in 76 scaffolds, with 61 gaps, and a scaffold N50 of 13.57 Mb (Table 2).

Most of the assembly sequence (99.26%) was assigned to 31 chromosomal-level scaffolds, representing 30 autosomes and the Z sex chromosome. Chromosome Z was assigned based on synteny to the genome of *Colias crocea* (GCF_905220415.1) (Ebdon *et al.*, 2021). These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 2; Table 3). Chromosome painting with Merian elements illustrates the distribution of orthologues along chromosomes and highlights patterns of chromosomal evolution relative to Lepidopteran ancestral linkage groups (Figure 3).

The mitochondrial genome was also assembled. This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

For haplotype 1, the estimated QV is 65.0, and for haplotype 2, 65.0. When the two haplotypes are combined, the assembly achieves an estimated QV of 65.0. The *k*-mer completeness is 77.18% for haplotype 1, 77.12% for haplotype 2, and 99.64% for the combined haplotypes (Figure 4). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) (Kriventseva *et al.*, 2019) identified 98.9% of the expected gene set (single = 98.6%, duplicated = 0.3%) for haplotype 1. The snail

Table 2. Genome assembly statistics.

Assembly name	ilEucSimp1.hap1.1	ilEucSimp1.hap2.1
Assembly accession	GCA_964265075.1	GCA_964265035.1
Assembly level	chromosome	scaffold
Span (Mb)	397.32	398.10
Number of chromosomes	31	N/A
Number of contigs	137	149
Contig N50	6.89 Mb	6.84 Mb
Number of scaffolds	76	102
Scaffold N50	13.57 Mb	13.44 Mb
Longest scaffold length (Mb)	17.93	N/A
Sex chromosomes	Z	N/A
Organelles	Mitochondrial genome: 15.23 kb	N/A

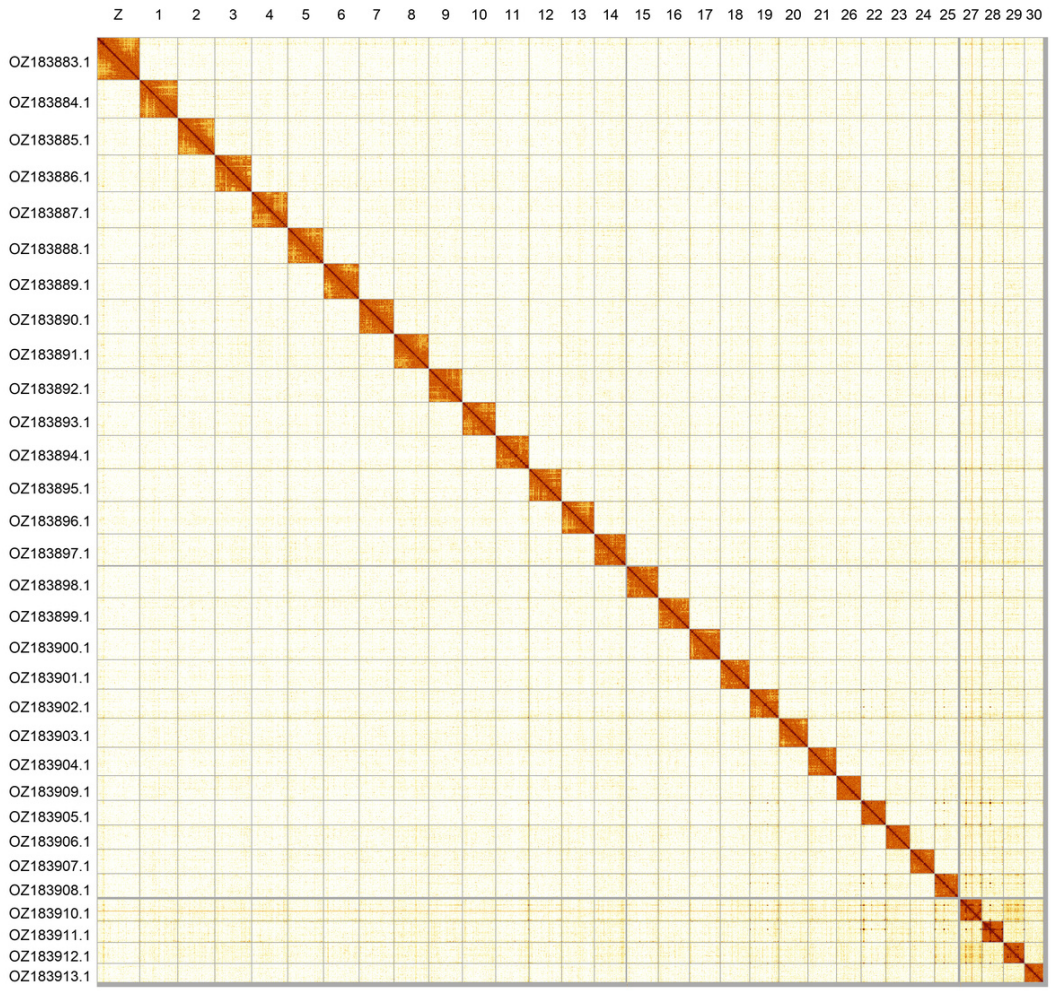

Figure 2. Hi-C contact map of the *Euchloe simplonia* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Euchloe simplonia* ilEucSimp1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ183884.1	1	15.90	34	M2
OZ183885.1	2	15.46	34	M1
OZ183886.1	3	15.32	34	M17;M20
OZ183887.1	4	15.15	34	M3
OZ183888.1	5	14.97	34	M8
OZ183889.1	6	14.82	33.50	M9
OZ183890.1	7	14.59	34	M5
OZ183891.1	8	14.54	34	M7
OZ183892.1	9	14.02	34	M12
OZ183893.1	10	13.94	34	M6
OZ183894.1	11	13.94	34	M18
OZ183895.1	12	13.64	34	M16
OZ183896.1	13	13.57	34	M4
OZ183897.1	14	13.47	34	M22
OZ183898.1	15	13.23	33.50	M21
OZ183899.1	16	13.04	34	M15
OZ183900.1	17	12.84	34	M10
OZ183901.1	18	12.38	34	M11
OZ183902.1	19	12.15	34.50	M23
OZ183903.1	20	12.13	34.50	M14
OZ183904.1	21	11.89	34	M13
OZ183905.1	22	10.37	34	M19
OZ183906.1	23	10.24	34	M28
OZ183907.1	24	10.17	34.50	M26
OZ183908.1	25	10.16	34	M24
OZ183909.1	26	10.65	34	M27
OZ183910.1	27	9.15	35.50	M30
OZ183911.1	28	9.05	36	M29
OZ183912.1	29	8.73	35	M31
OZ183913.1	30	7.90	34.50	M25
OZ183883.1	Z	17.93	34	MZ

plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.Q65**, 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

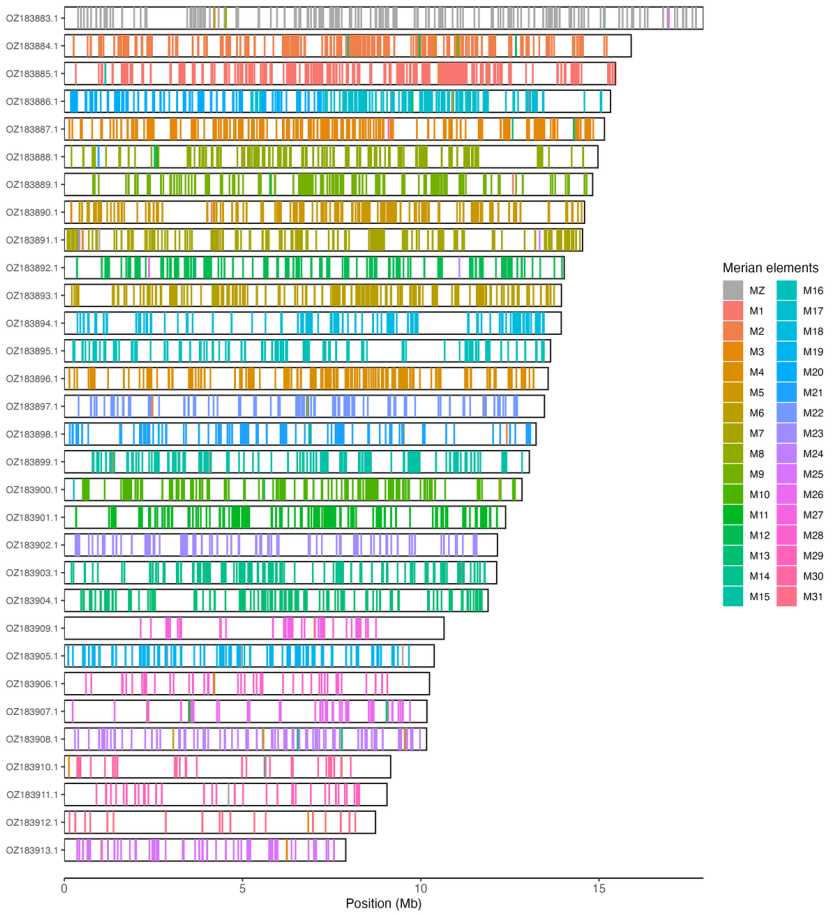

Figure 3. Merian elements painted across chromosomes in the ilEucSimp1.hap1.1 assembly of *Euchloe simplonia*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

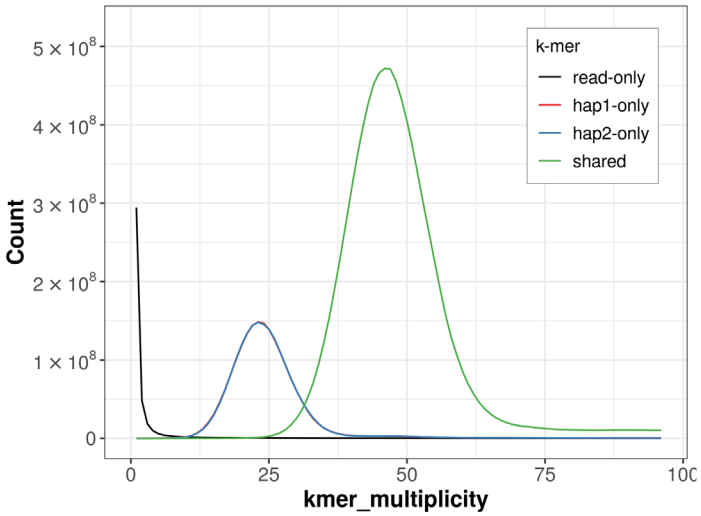

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve (the homozygous peak) corresponds to *k*-mers shared by both haplotypes and the red and blue curves (the heterozygous peaks) show *k*-mers found only in one of the haplotypes.

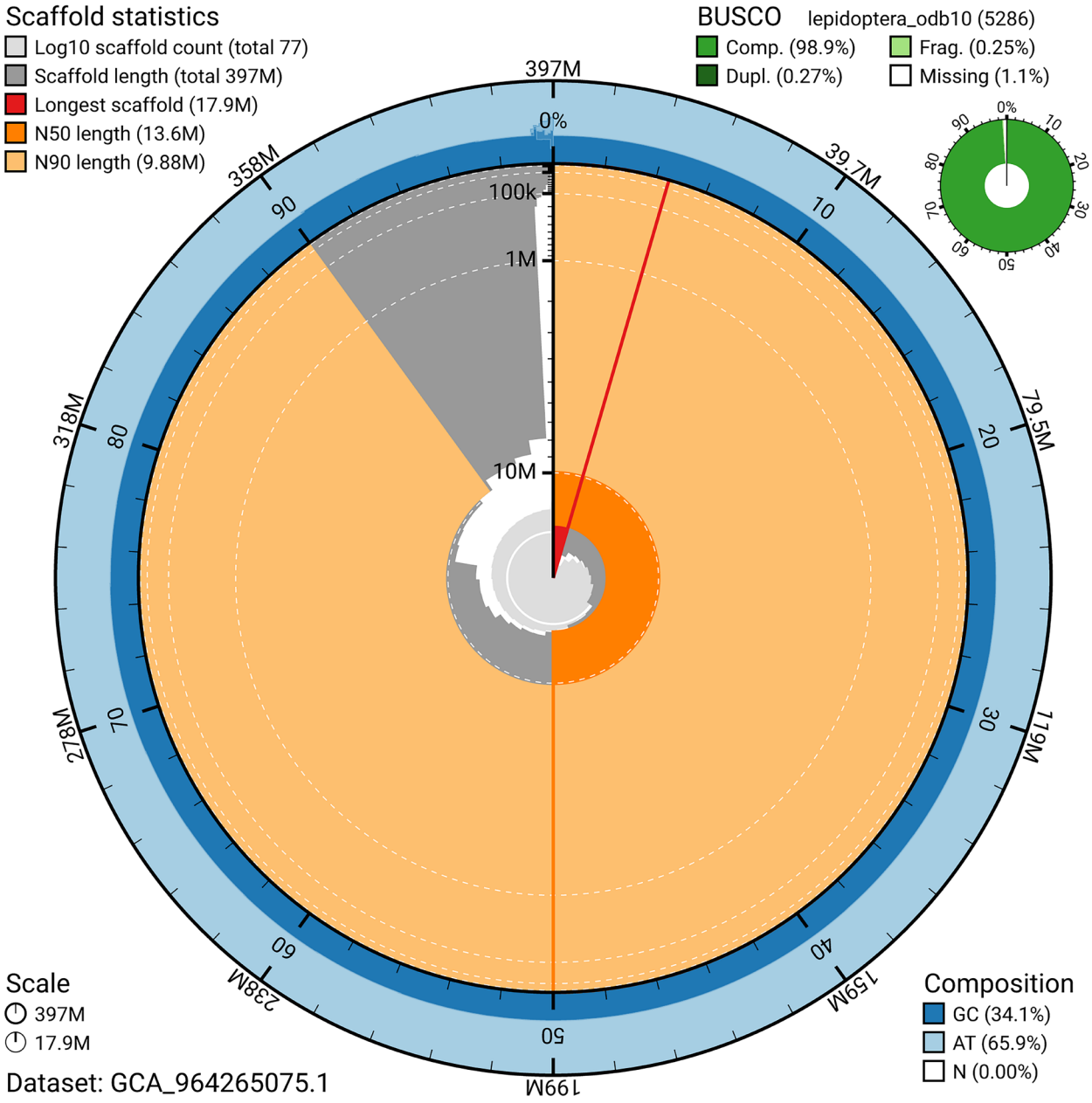

Figure 5. Assembly metrics for ilEucSimp1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

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

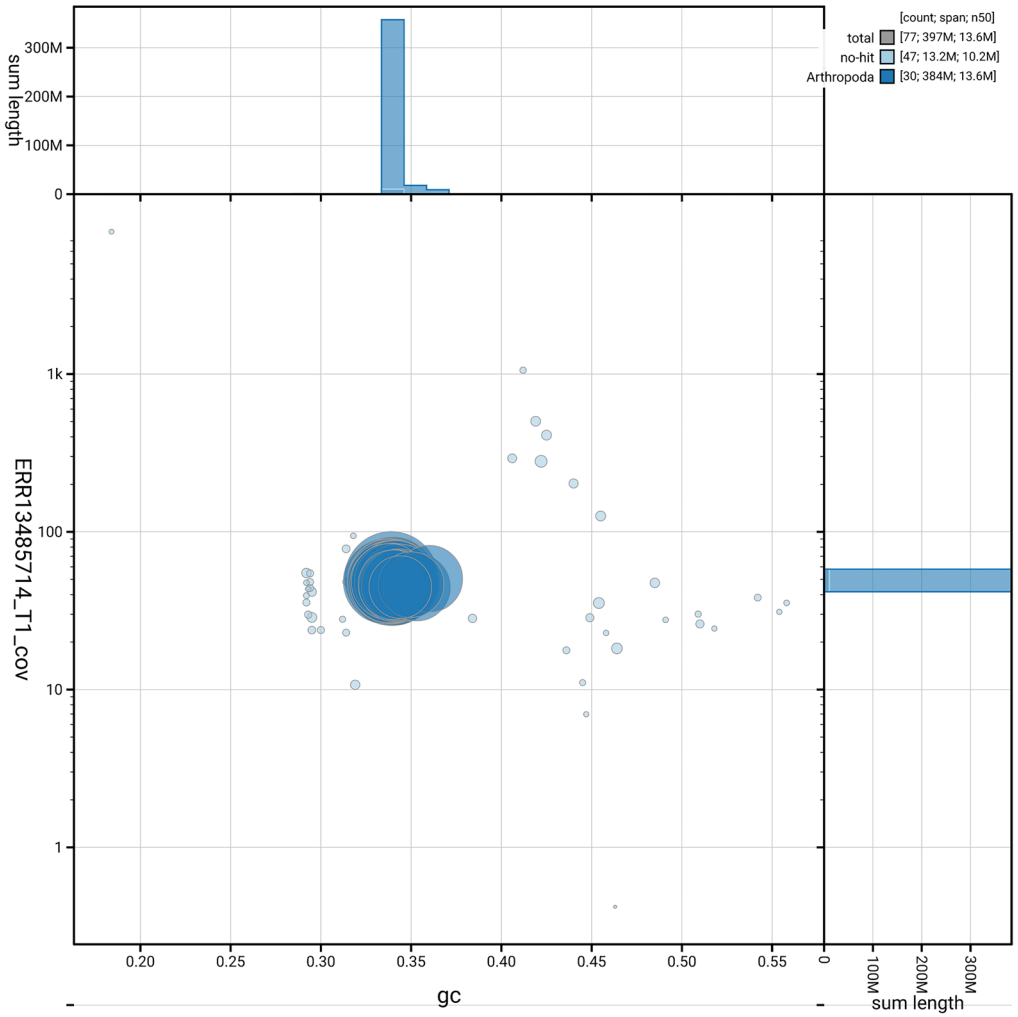

Figure 6. BlobToolKit GC-coverage plot for ilEucSimp1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 simplonia* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.7.Q65	6.C.Q40
Contig N50 length	6.89 Mb	≥ 1 Mb
Scaffold N50 length	13.57 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 65.0; haplotype 2: 65.0; combined: 65.0	≥ 40
k-mer completeness	Haplotype 1: 77.18%; Haplotype 2: 77.12%; combined: 99.64%	≥ 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	0%	≥ 90%

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: Euchloe simplonia (dappled White). Accession number PRJEB78686. The genome sequence is released openly for reuse. The *Euchloe simplonia* genome sequencing initiative is part of the Sanger Institute Tree of Life Programme (PRJEB43745) and Project Psyche (PRJEB71705). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through Ensembl at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

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

Author information

Contributors are listed at the following links:

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

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	http://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC

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

References

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

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

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

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

Caritg R, Domènech M, Dantart J, *et al.*: **Andorran butterfly monitoring scheme.** *J Insect Conserv.* 2011; **15**(1): 341–44.
Publisher Full Text

Challis R, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

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

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

Dubatolov V, Kosterin O: **A new subspecies of the *Euchloe ausonia* Hü bner species group from the highlands of the Dzhungarian Alatau (East Kazakhstan).** *Atalanta.* 1994; **25**(3/4): 513.
Reference Source

Ebdon S, Mackintosh A, Hayward A, *et al.*: **The genome sequence of the clouded yellow, *Colias* crocea (geoffroy, 1785) [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2021; **6**: 284.
PubMed Abstract | Publisher Full Text | Free 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

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
PubMed Abstract | Publisher Full Text | Free Full Text

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

IUCN: **The IUCN red list of threatened species. Version 2025-1**. 2025.
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

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**

Tolman T, Lewington R: **Collins butterfly guide**. HarperCollins Publishers, 2009.
Reference Source

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

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

Wright CJ, Stevens L, Mackintosh A, *et al.*: **Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera.** *Nat Ecol Evol.* 2024; **8**(4): 777–90.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

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

Open Peer Review

Current Peer Review Status: ? ✓ ✓

Version 1

Reviewer Report 04 November 2025

https://doi.org/10.21956/wellcomeopenres.27211.r135728

✓ **Panagiotis Ioannidis**
Foundation for Research & Technology - Hellas, Crete, Greece

This paper describes the sequencing and assembly of the lepidopteran *Euchloe simplonia*. The authors provide adequate information about the ecology of the species in the Introduction. The methodology used is the standard one, used by other DToL papers.

The genome assembly metrics are exceptional; complete BUSCOs: 98.9%, QV = 65, k-mer completeness = 99.64%, and CC ratio = 76/31 = 2.45. As such, this genome assembly is suitable for downstream analyses.

As I have mentioned in other recent reviews of mine, I believe that the usefulness of genome assemblies without the accompanying predicted gene sets (i.e. genome annotations) is limited. Especially for people with limited skills in bioinformatics such as wet-lab biologists who want to study a particular gene or gene family.

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

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

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

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics; insect 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 30 October 2025

https://doi.org/10.21956/wellcomeopenres.27211.r135726

Matthew Merkin
University of Liverpool, Liverpool, UK

Summary:

The authors present a new chromosome level assembly of a European butterfly, the Mountain Dappled White (*Euchloe simplonia*), as part of Project Psyche. DNA was extracted and sequenced from the head (HiC) and thorax (HiFi) of a single individual, followed by assembly, scaffolding and manual curation. The quality of the final assembly is high (BUSCO C:98.9%), suggesting it is suitable for use as a reference genome. The mitochondrial genome was also assembled.

The inclusion of Table 1 (sequencing information) and Figure 3 (assignment of Merian elements) greatly improves this genome note by allowing the reader to more easily access sequencing data and identify equivalent chromosomes across species.

Minor comments:

-There is little rationale for why this species was sequenced. An additional sentence expanding on Project Psyche and its goal to sequence lepidopterans would be useful here, especially if it can be related to this sentence from the roadmap on their website "*We will prioritise sequencing for phylogenetic diversity across the order and deliver genomes for agricultural pests, ecological indicator species and threatened species.*" - this could tie in to your speculation that this species is endangered in some areas, or just that there are comparatively few Pierid genomes.

-Is there a reason why a male sample was used or no annotation was produced? The W chromosome and protein information provided in the DToL genome notes was highly useful. I see there is a commitment in the Data Availability section to utilise RNA-seq data later on, in which case the publication of this genome note seems premature.

-The inclusion of benchmark values in Table 4 is highly useful for statistics that are unfamiliar to many, such as EBP summary. However, it can be hard to compare the value and benchmark in three rows:

1) The EBP summary in the table (6.7.Q65) differs from the main text (6.C.Q65) in the central scaffold NG50 field. Although the values are explained briefly in the linked articles, it may also be worth including an additional sentence explaining them to an unfamiliar reader.

2) The benchmark for the scaffold N50 is "= chromosome N50". It would help to list the actual value of the chromosome N50 here as well.
3) "Percentage of assembly assigned to chromosomes" has a value of 0%, which does not seem correct. The sum of chromosome lengths in Table 3 is 395.34, which when divided by the assembly span in Table 2 (397.32), and multiplied by 100 gives a value of ~99.5%.

Is the rationale for creating the dataset(s) clearly described?
No

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: Comparative genomics, genetics, entomology, ancient DNA, 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 08 October 2025

https://doi.org/10.21956/wellcomeopenres.27211.r131315

?

Christopher B. Cunningham
University of Georgia, Georgia, USA

The project continues to produce quality resources for the insect genomics community. The project for the Mountain Dappled White, *Euchloe simplonia*, continue that. The methods are sound and relatively well documented. The article needs has enough details to be reasonably reproducible with the newer addition of pipeline documentation. The data supporting the publication are publicly available.

Abstract. There is no justification within the abstract; just results. At least a sentence that this was part of a large survey of insects is needed. And one sentence for its possible value as a member of an under sampled group or ecology or something to motivate reader interest.

Keywords. There are all repeated from the title or abstract. They have little value to the reader. Keywords are used by Google to index the work so should reflect its broader use.

Background. Background. The stated motivation for this study in the introduction is fine – part of a very large survey. However, it would be nice for the reader to understand if this is the first of its genus, is of particular interest for ecological or genetics studies, or something unusual about its biology. Helping the reader understand what this genome might help with investigation in. Genus names at the beginning of sentences should be spelled out.

Results. Table 1 is not really needed. They are not details that the reader needs to under the work and they can all be found under the NCBI BioProject ID or are repeated from the methods paragraph.

BUSCO completeness needs to be assessed with the updated OrthoDB version 12. Difference from 10 to 12 have made a big difference for several insects I work with and colleagues have reported the same. The paper should not be indexed until this is updated. The new version has been for 11 months.

Figure 4. Fig 4 has very little information beyond what is found elsewhere in the manuscript and can easily be summarized as a sentence in the results.

Figure 3 & 5. Suggest lots of assembly dead-ends. I appreciate that they have no hits to anything other than Arthropoda but still the variation in coverage, variation in GC content, and strong Hi-C links among themselves is very odd. It suggests the QC of the genome draft assembly might be incomplete.

Genome annotation is not presented. This greatly limits the use of this resource.

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

No

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: Insects, genetics, genomics, epigenetics, behavior, reproduction.

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

significant reservations, as outlined above.