DATA NOTE

The genome sequence of the Pearly Heath, *Coenonympha arcania* Linnaeus, 1761 (Lepidoptera: Nymphalidae)

[version 1; peer review: 3 approved]

Roger Vila[1], Sabina Vila[1], Paul Doniol-Valcroze[2], Charlotte J. Wright[3], Joana I. Meier[3], Mark L. Blaxter[3], 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-Univ. Pompeu Fabra), Barcelona, Spain
[2]CEFE, CNRS, University of Montpellier, Montpellier, France
[3]Tree of Life, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 25 Jul 2025, **10**:364
https://doi.org/10.12688/wellcomeopenres.24624.1
Latest published: 25 Jul 2025, **10**:364
https://doi.org/10.12688/wellcomeopenres.24624.1

Abstract

We present a genome assembly from a female specimen of *Coenonympha arcania* (Pearly Heath; Arthropoda; Insecta; Lepidoptera; Nymphalidae). The assembly contains two haplotypes with total lengths of 480.99 megabases and 458.33 megabases. Most of haplotype 1 (99.95%) is scaffolded into 29 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 15.28 kilobases.

Keywords

Coenonympha arcania, Pearly Heath, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 1 25 Jul 2025	✓ view	✓ view	✓ view

1. **Fedor Čiampor**, Slovak Academy of Sciences, Bratislava, Slovakia
2. **Jasmine D. Alqassar**, The George Washington University, Washington, USA
3. **Quentin Rougemont**, Université Paris Saclay, Paris, France

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: Vila R: Investigation, Resources, Supervision; **Vila S**: Investigation, Resources; **Doniol-Valcroze P**: Writing – Original Draft Preparation; **Wright CJ**: Funding Acquisition, Project Administration, Supervision; **Meier JI**: Funding Acquisition, Project Administration, Supervision; **Blaxter ML**: Funding Acquisition, Project Administration, Supervision;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540). RV was supported 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: Vila R, Vila S, Doniol-Valcroze P *et al.* **The genome sequence of the Pearly Heath, *Coenonympha arcania* Linnaeus, 1761 (Lepidoptera: Nymphalidae) [version 1; peer review: 3 approved]** Wellcome Open Research 2025, **10**:364 https://doi.org/10.12688/wellcomeopenres.24624.1

First published: 25 Jul 2025, **10**:364 https://doi.org/10.12688/wellcomeopenres.24624.1

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; Nymphalidae; Satyrinae; Satyrini; Coenonymphina; *Coenonympha*; *Coenonympha arcania* Linnaeus, 1761 (NCBI:txid111894)

Background

The Pearly Heath *Coenonympha arcania* Linnaeus, 1761 is a small nymphalid butterfly generally associated with low and mid-altitude (0–-1 600 m a.s.l.) semi-open scrublands (Bozano, 2002; Lafranchis *et al.*, 2015; Lafranchis, 2016). The species is distributed across most of the Western Palaearctic region and is considered as Least Concern (LC) by the European IUCN RedList (van Swaay *et al.*, 2025). In most of its distribution, this species generally shows a remarkably short generation – approximately one month long -– from May to July, depending on local climate, but can occasionally exhibit a partial second generation in September (Lafranchis *et al.*, 2015).

Coenonympha arcania can be recognised by its brown hindwing with a broad, indented, whitish postdiscal band and four to six black eyespots with white pupils and orange circles that are usually not aligned in a regular curve (Figure 1). The dorsal surface of the wings, with brown hindwings and orange forewings with brown margins, is seldom visible at rest (Bozano, 2002; Lafranchis, 2016).

The species is part of the Holarctic species-rich *Coenonympha* genus and specifically belongs to the hero clade, in which interspecific hybridisation and gene flow are pervasive (Greenwood *et al.*, 2025; Kodandaramaiah & Wahlberg, 2009). The species is further part of the *C. arcania–gardetta* species complex, in which hybrid speciation with *C. gardetta* is implied in the establishment of two hybrid species (*C. darwiniana* and *C. cephalidarwiniana*) (Capblancq *et al.*, 2024; Doniol-Valcroze *et al.*, 2024; Schmitt & Besold, 2010). Past and ongoing studies therefore aim at clarifying the role of hybridisation and gene flow in the diversification and adaptation of the clade. Various experiments have focused on the species' ecology and notably on the thermal tolerance of its different developmental stages (Doniol-Valcroze *et al.*, 2024; Nève & Després, 2020; Vrba *et al.*, 2022). Both population genomic and chemical ecology approaches have been applied to the species complex in order to decipher the traits and the origins of the isolating mechanisms at play in the species complex (Capblancq *et al.*, 2019; Capblancq *et al.*, 2024; Doniol-Valcroze, 2024; Doniol-Valcroze *et al.*, 2025). Moreover, a chromosome-level reference genome of *C. arcania* was recently published (Legeai *et al.*, 2024). Overall, resolved chromosome-level genome assemblies are important resources for ongoing studies of the species complex and for advancing our understanding of hybrid speciation, the build-up of reproductive isolation, genome evolution and ecological diversification.

Figure 1. Photograph of *Coenonympha arcania* by Jörg Hempel (not the specimen used for genome sequencing).

Methods

Sample acquisition

The specimen used for genome sequencing was an adult female *Coenonympha arcania* (specimen ID SAN28000277, ToLID ilCoeArcn1), collected from El Brull, Osona, Barcelona, Catalonia, Spain (latitude 41.812, longitude 2.294; elevation 830 m) on 23/07/2018. The specimen was collected and identified by Roger Vila and Sabina Vila (Institut De Biologia Evolutiva).

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 ilCoeArcn1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the whole organism was homogenised by powermashing using a PowerMasher II tissue disruptor.

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol. DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor®3 for LI PacBio protocol. Sheared DNA was purified by automated SPRI (solid-phase reversible immobilisation). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 53.06 ng/μL and a yield of 2 493.82 ng, with a fragment size of 13.9 kb. The 260/280 spectrophotometric ratio was 1.9, and the 260/230 ratio was 2.19.

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 whole organism tissue of the ilCoeArcn1 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.

Table 1. Specimen and sequencing data for BioProject PRJEB80653.

Platform	PacBio HiFi	Hi-C
ToLID	ilCoeArcn1	ilCoeArcn1
Specimen ID	SAN28000277	SAN28000277
BioSample (source individual)	SAMEA115768730	SAMEA115768730
BioSample (tissue)	SAMEA115768830	SAMEA115768830
Tissue	whole organism	whole organism
Sequencing platform and model	Revio	Illumina NovaSeq X
Run accessions	ERR13762697	ERR13732204
Read count total	1.47 million	734.19 million
Base count total	15.11 Gb	110.86 Gb

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 11 breaks, 17 joins, and removal of 10 haplotypic duplications. 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 *Coenonympha arcania* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 15.11 Gb (gigabases) from 1.47 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 458.25 Mb, with a heterozygosity of 2.10% and repeat content of 37.40%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 32× coverage. Hi-C sequencing produced 110.86 Gb from 734.19 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 480.99 Mb in 39 scaffolds, with 181 gaps, and a scaffold N50 of 17.68 Mb (Table 2).

Most of the assembly sequence (99.95%) was assigned to 29 chromosomal-level scaffolds, representing 28 autosomes and the Z sex chromosome. The Z chromosome was identified by coverage and alignment to *Coenonympha glycerion* (GCA_963855885.1). There was no evidence of a W chromosome in this assembly, therefore it appears that this species exhibits the ZO karyotype. The 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 65.4, and for haplotype 2, 65.4. When the two haplotypes are combined, the assembly achieves an estimated QV of 65.4. The *k*-mer completeness is 69.37% for haplotype 1, 65.99% for haplotype 2, and 99.47% for the combined haplotypes (Figure 4). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) (Kriventseva *et al.*, 2019) identified 98.2% of the expected gene set

Table 2. Genome assembly statistics.

Assembly name	ilCoeArcn1.hap1.1	ilCoeArcn1.hap2.1
Assembly accession	GCA_964275175.1	GCA_964275225.1
Assembly level	chromosome	scaffold
Span (Mb)	480.99	458.33
Number of chromosomes	29	N/A
Number of contigs	220	280
Contig N50	4.11 Mb	3.69 Mb
Number of scaffolds	39	119
Scaffold N50	17.68 Mb	17.26 Mb
Longest scaffold length (Mb)	25.64	N/A
Sex chromosomes	Z	N/A
Organelles	Mitochondrial genome: 15.28 kb	N/A

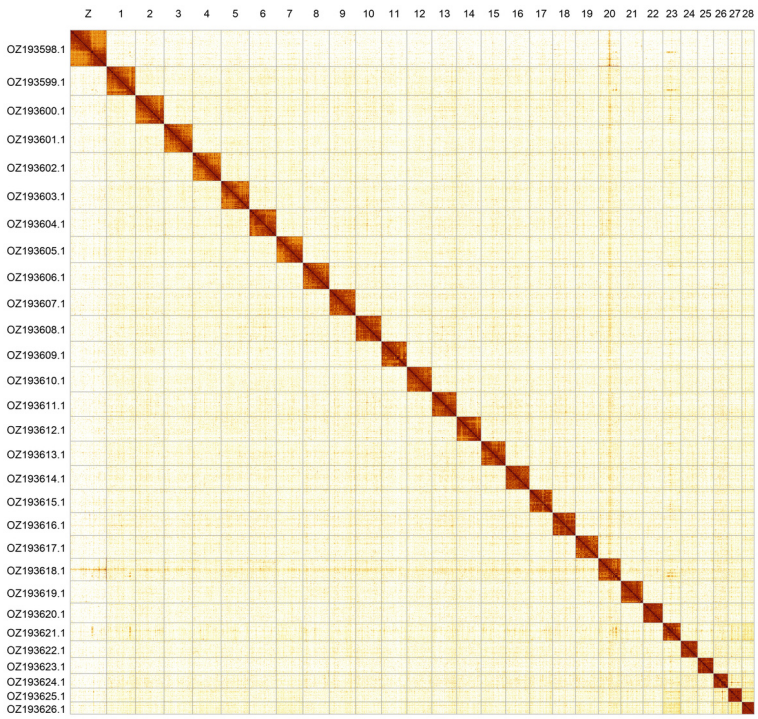

Figure 2. Hi-C contact map of the *Coenonympha arcania* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Coenonympha arcania* ilCoeArcn1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ193599.1	1	20.24	38	M19;M26
OZ193600.1	2	20.22	38	M2
OZ193601.1	3	20.06	38	M1
OZ193602.1	4	20.01	38.50	M17;M20
OZ193603.1	5	19.86	38.50	M3
OZ193604.1	6	18.95	37.50	M9
OZ193605.1	7	18.64	39	M14;M29
OZ193606.1	8	18.63	38	M8
OZ193607.1	9	18.51	38	M5
OZ193608.1	10	18.04	38	M12
OZ193609.1	11	17.86	37.50	M18
OZ193610.1	12	17.68	38.50	M7
OZ193611.1	13	17.33	38	M16
OZ193612.1	14	17.26	38	M4
OZ193613.1	15	17.25	38.50	M6
OZ193614.1	16	16.64	38.50	M22
OZ193615.1	17	16.25	38.50	M15
OZ193616.1	18	16.18	38	M21
OZ193617.1	19	15.92	39	M11
OZ193618.1	20	15.91	39	M23
OZ193619.1	21	15.69	38.50	M10
OZ193620.1	22	13.77	38	M13
OZ193621.1	23	12.66	41.50	M30
OZ193622.1	24	11.83	39	M24
OZ193623.1	25	11.19	39	M28
OZ193624.1	26	10.24	39	M27
OZ193625.1	27	9.69	39.50	M25
OZ193626.1	28	8.60	41	M31
OZ193598.1	Z	25.64	37.50	MZ

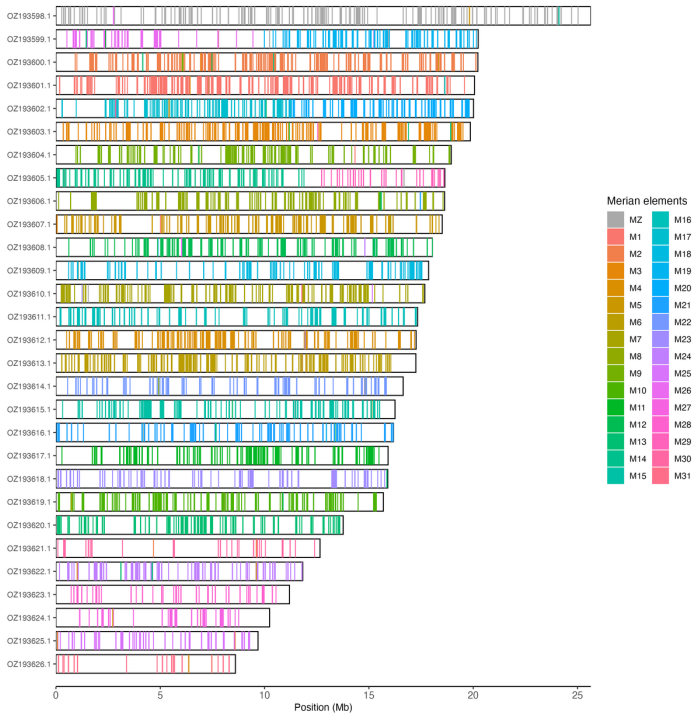

Figure 3. Merian elements painted across chromosomes in the ilCoeArcn1.hap1.1 assembly of *Coenonympha arcania*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

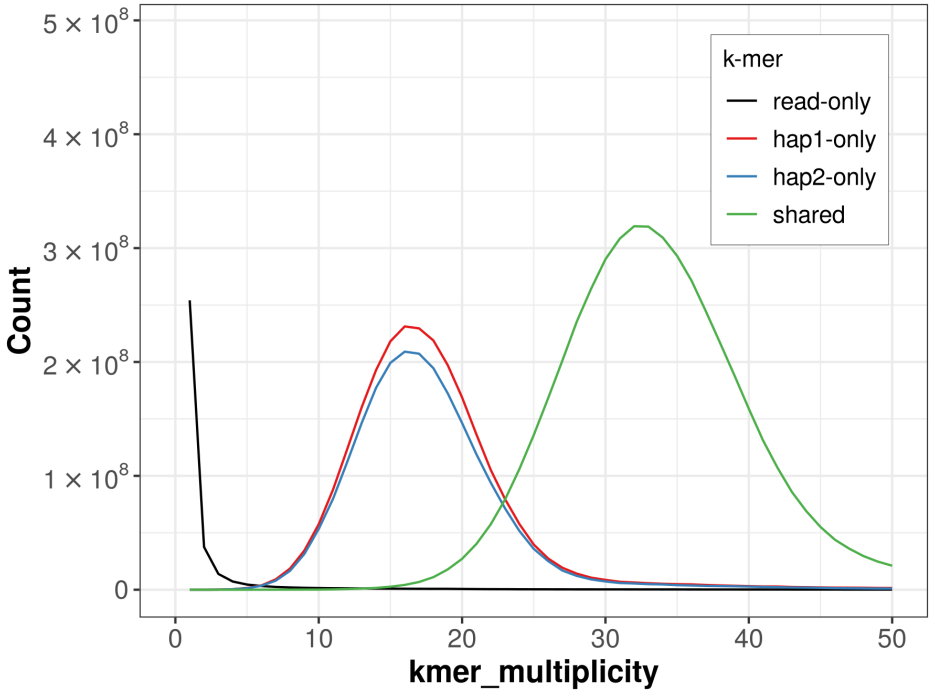

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers 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.

(single = 97.7%, duplicated = 0.5%) for haplotype 1. The snail 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.

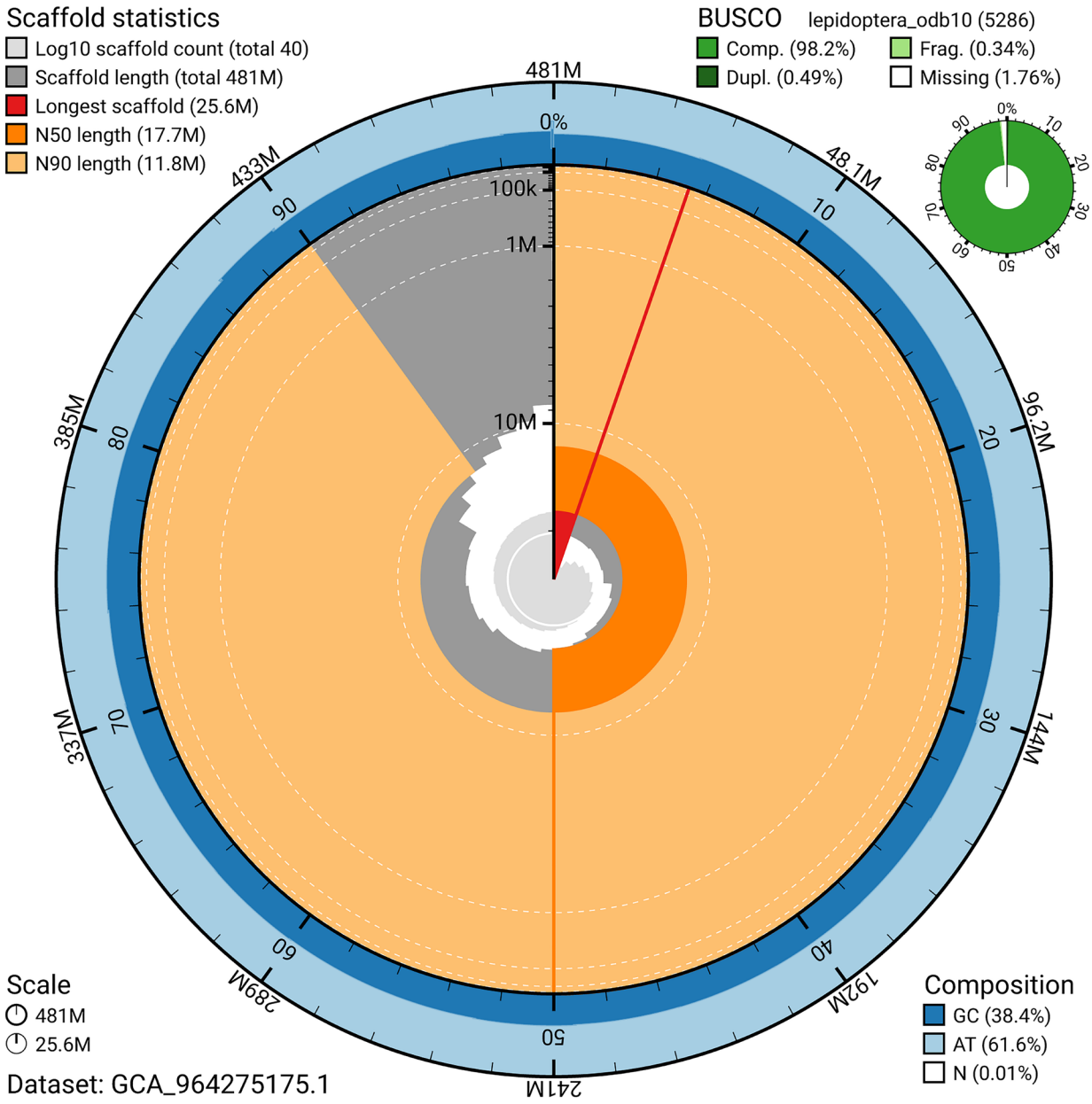

Figure 5. Assembly metrics for ilCoeArcn1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

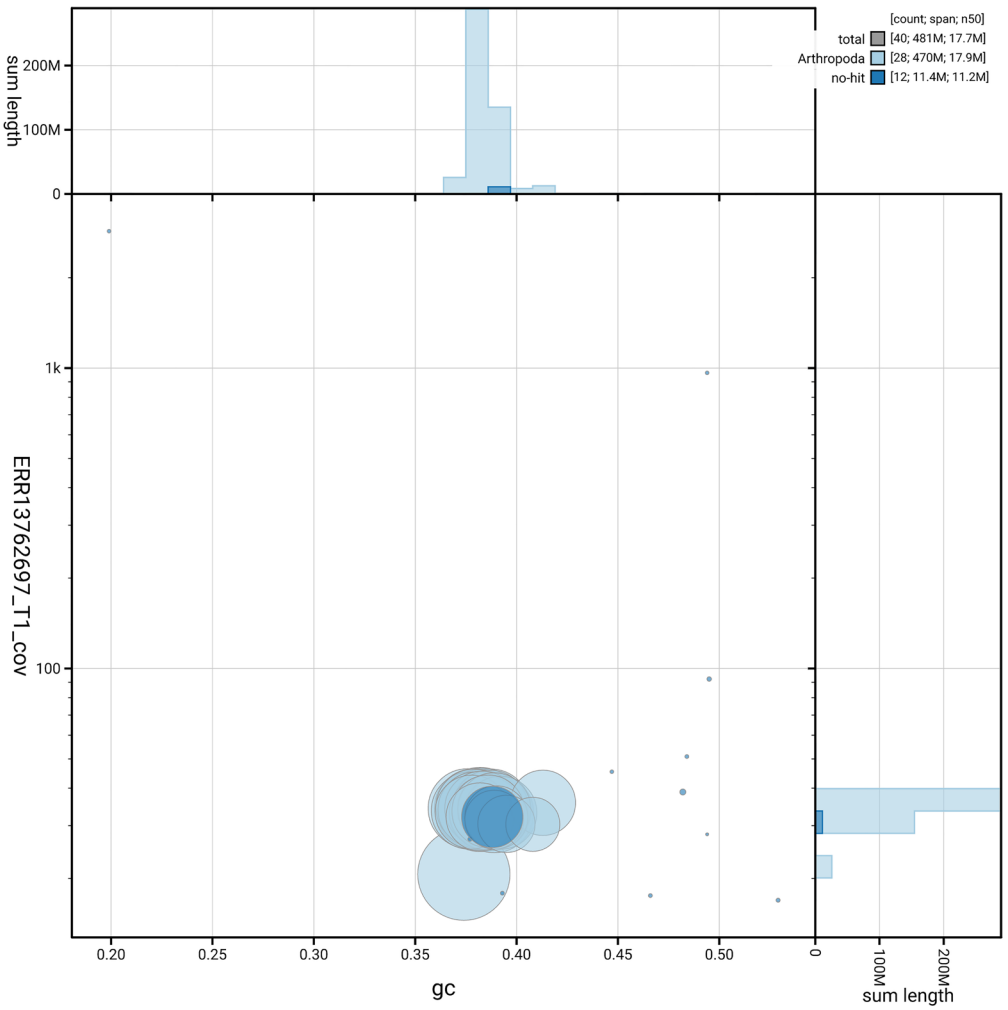

Figure 6. BlobToolKit GC-coverage plot for ilCoeArcn1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Coenonympha arcania* assembly.

Measure (Benchmark)	Value
EBP summary (haplotype 1)	6.C.Q65
Contig N50 length (≥ 1 Mb)	4.11 Mb
Scaffold N50 length (= chromosome N50)	17.68 Mb
Consensus quality (QV) (≥ 40)	Haplotype 1: 65.4; haplotype 2: 65.4; combined: 65.4
k-mer completeness (≥ 95%)	Haplotype 1: 69.37%; Haplotype 2: 65.99%; combined: 99.47%
BUSCO* (S > 90%; D < 5%)	C:98.2%[S:97.7%,D:0.5%],F:0.3%,M:1.4%,n:5286
Percentage of assembly assigned to chromosomes (≥ 90%)	99.95%

Wellcome Sanger Institute – Legal and Governance

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

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international).

Each transfer of samples is undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Tree of Life collaborator, Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances, other Tree of Life collaborators.

Data availability

European Nucleotide Archive: Coenonympha arcania (pearly heath). Accession number PRJEB80653. The genome sequence is released openly for reuse. The *Coenonympha arcania* 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	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
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
lep_busco_painter	1.0.0	https://github.com/charlottewright/lep_busco_painter
MerquryFK	1.1.1	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2

Software	Version	Source
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	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

Bozano GC: **Guide to the butterflies of the palearctic region. Satyrinae, Part III. Tribe Satyrini, subtribes Melanargiina and Coenonymphina, Melanargia, Coenonympha, Sinonympha, Triphysa.** Milano: Omnes Partes, 2002.

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

Capblancq T, Mavarez J, Rioux D, *et al.*: **Speciation with gene flow: evidence from a complex of alpine butterflies (*Coenonympha*, Satyridae).** *Ecol Evol.* 2019; **9**(11): 6444–6457.
PubMed Abstract | Publisher Full Text | Free Full Text

Capblancq T, Roux C, Boyer F, *et al.*: **Untangling the contribution of adaptive *versus* non-adaptive processes in the evolution of reproductive isolation between *Coenonympha* butterflies.** 2024.
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

Doniol-Valcroze P: **Hybrid speciation and adaptation in alpine butterflies in the genus Coenonympha.** 2024.

Doniol-Valcroze P, Després L, Joron M: **Strong differences in egg thermal tolerance between closely related lowland and alpine butterfly species.** *Ecol Entomol.* 2024; **49**(4): 598–603.
Publisher Full Text

Doniol-Valcroze P, Develay Nguyen L, Buatois B, *et al.*: **Non-random sorting of parental chemical compounds during hybrid speciation.** *J Evol Biol.* 2025; **38**(5): 559–571.
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

Greenwood MP, Capblancq T, Wahlberg N, *et al.*: **Whole genome data confirm pervasive gene discordance in the evolutionary history of *Coenonympha* (Nymphalidae) butterflies.** *Mol Phylogenet Evol.* 2025; **202**: 108222.
PubMed Abstract | Publisher 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

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

Kodandaramaiah U, Wahlberg N: **Phylogeny and biogeography of *Coenonympha* butterflies (Nymphalidae: Satyrinae) - patterns of colonization in the holarctic.** *Syst Entomol.* 2009; **34**(2): 315–23.
Publisher Full Text

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

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

Lafranchis T: **Papillons de France - Guide de détermination des Papillons Diurnes.** France: Diatheo, 2016.
Reference Source

Lafranchis T, Jutzeler D, Guillosson JY, *et al.*: **La Vie Des Papillons.** France: Diatheo, 2015.
Reference Source

Legeai F, Romain S, Capblancq T, *et al.*: **Chromosome-level assembly and annotation of the pearly heath *Coenonympha arcania* butterfly genome.** *Genome Biol Evol.* 2024; **16**(3): evae055.
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, [Accessed 2 April 2024].
Reference Source

Nève G, Després L: **Cold adaptation across the elevation gradient in an alpine butterfly species complex.** *Ecol Entomol.* 2020; **45**(5): 997–1003.
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**

Schmitt T, Besold J: **Upslope movements and large scale expansions: the taxonomy and biogeography of the *Coenonympha arcania-C. darwiniana-C. gardetta* butterfly species complex.** *Zool J Linn Soc.* 2010; **159**(4): 890–904.
Publisher Full Text

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

van Swaay C, Ellis S, Warren M: ***Coenonympha arcania* (Europe assessment).** 2025.

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

Vrba P, Sucháčková Bartoňová A, Andres M, *et al.*: **Exploring cold hardiness within a butterfly clade: supercooling ability and polyol profiles in European Satyrinae.** *Insects.* 2022; **13**(4): 369.
PubMed Abstract | **Publisher Full Text** | **Free 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**

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 25 September 2025

https://doi.org/10.21956/wellcomeopenres.27135.r132655

✔ **Quentin Rougemont**
Université Paris Saclay, Paris, France

Review of 'The genome sequence of the Pearly Heath, *Coenonympha arcania* Linnaeus, 1761 (Lepidoptera: Nymphalidae)
This is a high-quality genome assembly for the Pearly Heath, providing providing valuable ressources for future studies, especially given the existence of other genome in this species complex.
The genome assemblies was easily accessible online, which is great.
The whole pipeline follows standard practice and the report of the results meet the quality requirement. The pipeline is also easilly findable online following the provided links.
The merian plot was really nice.
I am a bit surprised by the very low k-mer completeness of each haplotype assembly separately. Is this due to a high number of structural variation between each haplotype ?
Minor comments:
It would be nice to specify whether only default parameters were used for each tools or if some parameters had to be set to some specific values.
Why not showing busco scores from the 'lepidoptera_odb12', moreover using the BUSCO6.0.0 seems to provide more accurate identification of busco genes, compared with the BUSCO5.5.0 used here.
Was there any evidence of contamination from the ASCC pipleine?

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

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

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

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: population 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 24 September 2025

https://doi.org/10.21956/wellcomeopenres.27135.r132654

Jasmine D. Alqassar
The George Washington University, Washington, District of Columbia, USA

In this article Vila et al. present a high-quality chromosome-level reference genome for *Coenonympha arcania*, the pearly heath butterfly. BUSCO analysis demonstrates the genome assembly is highly contiguous and complete (BUSCO score: C: 98.2%[S:97.7, D:0.5%], F: 0.34%, M:1.76% n=5286). This genome resource will serve as a valuable resource for the lepidopteran genomics community.

Although it is possible that this species does not have a W chromosome and instead uses the ancestral ZZ/ZO sex determination system, a lack of a W chromosome in a genome assembly is only partial evidence of this possibility because of the known difficulties associated with sequencing lepidopteran W chromosomes. Further cytological analysis would be needed to confirm the true absence of a W chromosome. Therefore, I ask that the authors amend the latter half of the following statement in the “Assembly statistics” section “There was no evidence of a W chromosome in this assembly, therefore it appears that this species exhibits the ZO karyotype.” to instead state that “this suggests that this species may exhibit the ZO karyotype; however this possibility must be confirmed by further cytological analysis”.

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

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

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

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics and evolutionary developmental biology

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 09 September 2025

https://doi.org/10.21956/wellcomeopenres.27135.r130506

Fedor Čiampor
Slovak Academy of Sciences, Bratislava, Slovakia

This report focuses on generating the genome of Pearly Heath *Coenonympha arcania* Linnaeus, 1761, a small butterfly from the Nymphalidae family. The reason for sequencing the genome of this species is to study its position in the species complex and also the role of hybridization and gene flow in the diversification and adaptation of the clade.
Although I am not very proficient in WGS, in my opinion, the methods and outputs were described very thoroughly and accurately.
The only comment I have on the report is why did the authors use a female? In insect taxonomy, it is very common that females are very difficult to determine, and species are very often clearly delimited based only on male genitalia. Furthermore, if the sequenced species belongs to a complex in which species hybridize with each other, how can the authors be 100% sure that the species they sequenced is *Coenonympha arcania*? Perhaps in this case there is evidence that the female used does indeed belong to the species mentioned, but it would be good to mention it.
And, please check the sentence "Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools.". It doesn't sound quite right to me.

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: DNA (meta)barcoding, biodiversity, taxonomy, phylogeny insects, mostly freshwater beetles

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.