DATA NOTE

The genome sequence of the Clay Triple-lines, *Cyclophora linearia* (Hübner, 1799) (Lepidoptera: Geometridae)

[version 1; peer review: 2 approved]

Laura Sivess[1], Gavin R. Broad[1], Stephanie Holt[1], Liam M. Crowley[2],
Natural History Museum Genome Acquisition Lab,
University of Oxford and Wytham Woods Genome Acquisition Lab,
Darwin Tree of Life Barcoding Collective,
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, Darwin Tree of Life Consortium

[1]Natural History Museum, London, England, UK
[2]University of Oxford, Oxford, England, UK

v1 **First published:** 19 Dec 2025, **10**:679
https://doi.org/10.12688/wellcomeopenres.25389.1
Latest published: 19 Dec 2025, **10**:679
https://doi.org/10.12688/wellcomeopenres.25389.1

Abstract

We present a genome assembly from an individual male *Cyclophora linearia* (Clay Triple-lines; Arthropoda; Insecta; Lepidoptera; Geometridae). The assembly contains two haplotypes with total lengths of 280.22 megabases and 284.23 megabases. Most of haplotype 1 (99.9%) 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 16.54 kilobases. Gene annotation of this assembly on Ensembl identified 11 742 protein-coding genes. This assembly was generated as part of the Darwin Tree of Life project, which produces reference genomes for eukaryotic species found in Britain and Ireland.

Keywords

Cyclophora linearia; Clay Triple-lines; genome sequence; chromosomal; Lepidoptera

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 19 Dec 2025	✓ view	✓ view

1. **Mohadeseh Tahami**, University of Jyväskylä, Jyväskylä, Finland

2. **Annabel Whibley**, Bragato Research Institute, Clyde, New Zealand

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

This article is included in the Tree of Life gateway.

Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: Sivess L: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Broad GR**: Investigation, Resources; **Holt S**: Investigation, Resources; **Crowley LM**: Investigation, Resources;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540) and the Darwin Tree of Life Discretionary Award [218328, https://doi.org/10.35802/218328].
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: Sivess L, Broad GR, Holt S *et al.* **The genome sequence of the Clay Triple-lines, *Cyclophora linearia* (Hübner, 1799) (Lepidoptera: Geometridae) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, **10**:679 https://doi.org/10.12688/wellcomeopenres.25389.1

First published: 19 Dec 2025, **10**:679 https://doi.org/10.12688/wellcomeopenres.25389.1

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; Geometroidea; Geometridae; Sterrhinae; *Cyclophora*; *Cyclophora linearia* (Hübner, 1799) (NCBI:txid934830)

Background

Described by Hübner in 1799, *Cyclophora linearia*, commonly known as the Clay Triple-lines, belongs to the family Geometridae and is most frequent in southern Britain, where Beech (*Fagus* species) is most common. Since the early 2000s, its range has gradually expanded northwards, reaching Scotland, and is rapidly spreading North in Ireland; this spread has been accompanied by a marked increase in abundance (Randle *et al.*, 2019). Globally, *C. linearia* is widespread across Europe, with its distribution confined to the Palaearctic region, extending west to Ireland and east to Turkey and Iran (GBIF Secretariat, 2023).

In the United Kingdom, *C. linearia* flies from May to July, with a partial second generation from August to mid-September. With a wingspan of 26 to 33 mm, the Clay Triple-lines is yellowish-brown in colour. The species name *linearia* refers to the distinctive crosslines spanning both the forewings and hindwings. These total three brown lines, with the central line being the darkest and most clearly defined. The second generation differs slightly in appearance, showing orange-pink undertones and a central discal spot, closely resembling the Maiden's Blush (*Cyclophora punctaria*). The larvae are monophagous in the wild, feeding exclusively on the leaves of Beech, thereby restricting the species' habitat to Beech-rich environments such as woodlands, hedgerows, gardens, and parks. However, they have been reared successfully on Pedunculate Oak (*Quercus robur*) (Wheeler & Dickson, 2022).

The genome of *C. linearia* was generated within the Darwin Tree of Life project, using a specimen light-trapped at Gilbert White's House, Selborne, United Kingdom. This resource will facilitate functional and comparative genomic research on topics such as host–plant interactions, visual system development, and insect neurobiology. It contributes to the expanding body of Lepidoptera genomic data, enhancing studies of molecular evolution and adaptation within this diverse order.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was a male *Cyclophora linearia* (specimen ID NHMUK014536923, ToLID ilCycLine1; Figure 1), collected from Selborne, Gilbert White's House, England, UK (latitude 51.09, longitude –0.94) on 2021-06-10. The specimen was collected by Laura Sivess, Stephanie Holt and Gavin Broad, and formally identified by Gavin Broad. A second specimen was used for Hi-C sequencing (specimen ID Ox003714, ToLID ilCycLine2). It was collected from Wytham Woods, Oxfordshire, UK (latitude 51.772, longitude –1.338) on 2023-06-20. The specimen was collected and identified by Liam Crowley. The same specimen was used for RNA sequencing.

Figure 1. Photographs of the *Cyclophora linearia* (ilCycLine1) specimen used for genome sequencing.

The initial identification was verified by an additional DNA barcoding process according to the framework developed by Twyford *et al.* (2024). A small sample was dissected from the specimen and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI) (see the protocol). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding are available on protocols.io.

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 ilCycLine1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the abdomen was homogenised by powermashing using a PowerMasher II tissue disruptor. HMW DNA was extracted 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 8.36 ng/μL and a yield of 1 086.80 ng.

RNA was extracted from whole organism tissue of ilCycLine2 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol. The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), 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.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen whole organism tissue of the ilCycLine2 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

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

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

RNA library preparation and sequencing

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

Genome assembly

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

The HiFi reads were assembled using Hifiasm in Hi-C phasing mode (Cheng *et al.*, 2021; 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).

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 39 breaks and 106 joins. The curation process is described at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020) was run in a Singularity container (Kurtzer *et al.*, 2017) 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 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).

Genome sequence report

Sequence data

PacBio sequencing of the *Cyclophora linearia* specimen generated 15.71 Gb (gigabases) from 1.55 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 277.40 Mb, with a heterozygosity of 1.27% and repeat content of 17.80% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 55× coverage. Hi-C sequencing produced 96.09 Gb from 636.33 million reads, which were used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The genome was assembled into two haplotypes using Hi-C phasing. Haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The final assembly

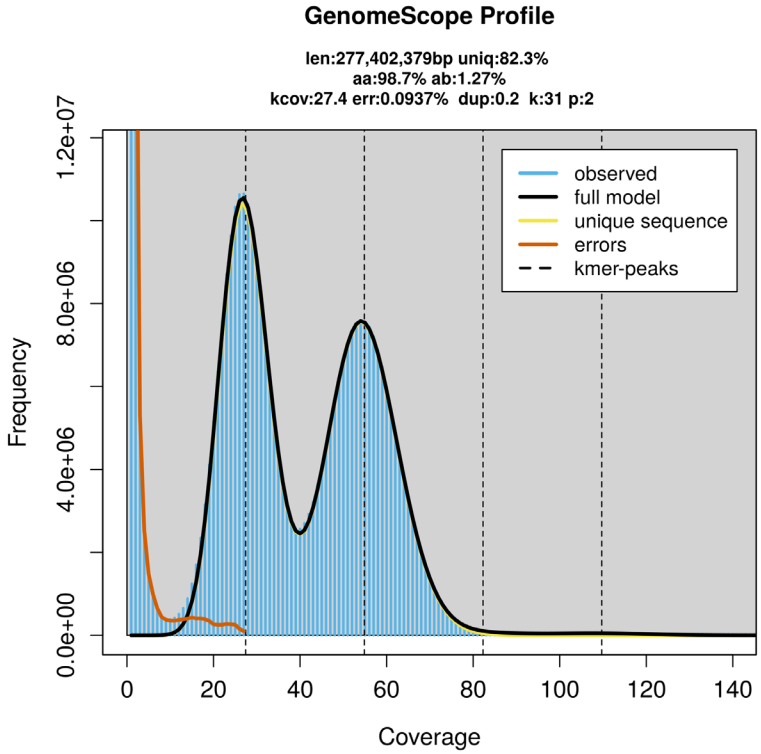

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB81921.

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	ilCycLine1	ilCycLine2	ilCycLine2
Specimen ID	NHMUK014536923	Ox003714	Ox003714
BioSample (source individual)	SAMEA111458034	SAMEA114644766	SAMEA114644766
BioSample (tissue)	SAMEA111458108	SAMEA114645395	SAMEA114645395
Tissue	abdomen	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq X	Illumina NovaSeq X
Run accessions	ERR13917231	ERR13925953	ERR14792855
Read count total	1.55 million	636.33 million	114.77 million
Base count total	15.71 Gb	96.09 Gb	17.33 Gb

has a total length of 280.22 Mb in 36 scaffolds, with 40 gaps, and a scaffold N50 of 10.11 Mb (Table 2).

Most of the haplotype 1 assembly sequence (99.9%) 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). The Z chromosome was assigned by BUSCO gene painting with ancestral Merian elements (Wright *et al.*, 2024).

The mitochondrial genome was also assembled (length 16.54 kb, OZ218285.1). 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 64.9, and for haplotype 2, 64.1. When the two haplotypes are combined, the assembly achieves an estimated QV of 64.5. The *k*-mer completeness is 75.56% for haplotype 1, 75.35% for haplotype 2, and 99.39% for the combined haplotypes (Figure 4).

BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.1% of the expected gene set (single = 97.8%, duplicated = 0.3%) in haplotype 1. For haplotype 2, BUSCO v.5.5.0 analysis identified 98.0% of the expected gene set (single = 96.7%, duplicated = 1.3%). 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) and 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 *Cyclophora linearia* genome assembly (GCA_965112375.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 20 868 transcribed mRNAs from 11 742 protein-coding and 1 890 non-coding genes. The average transcript length is 11 956.23 bp, with an average of 1.53 coding transcripts per gene and 7.55 exons per transcript. For further information about the annotation, 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 Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the **'Darwin Tree of Life Project Sampling Code of Practice'**, which can be found in full on the Darwin Tree of Life website. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project. Further, 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 further undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Darwin Tree of Life Partner,

Table 2. Genome assembly statistics.

Assembly name	ilCycLine1.hap1.1	ilCycLine1.hap2.1
Assembly accession	GCA_965112375.1	GCA_965112355.1
Assembly level	chromosome	scaffold
Span (Mb)	280.22	284.23
Number of chromosomes	31	scaffold-level
Number of contigs	76	464
Contig N50	7.52 Mb	2.84 Mb
Number of scaffolds	36	238
Scaffold N50	10.11 Mb	10.05 Mb
Longest scaffold length (Mb)	14.49	-
Sex chromosomes	Z	-
Organelles	Mitochondrion: 16.54 kb	-

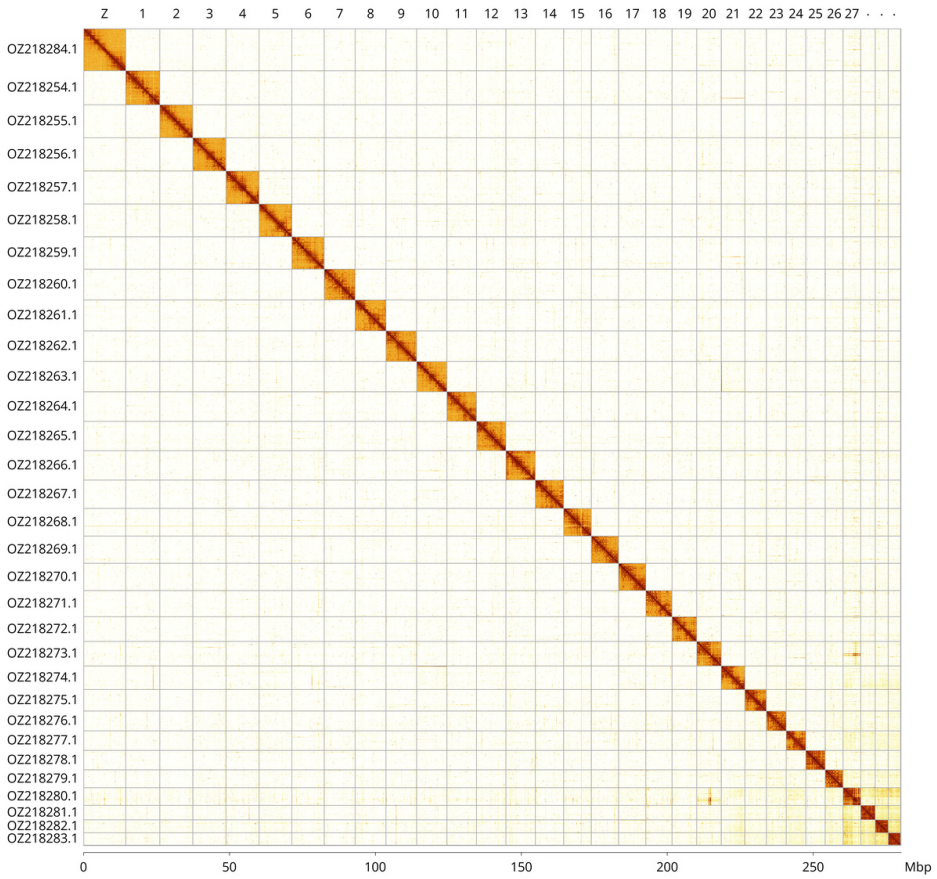

Figure 3. Hi-C contact map of the *Cyclophora linearia* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Cyclophora linearia* ilCycLine1.

INSDC accession	Molecule	Length (Mb)	GC%
OZ218254.1	1	11.64	36.50
OZ218255.1	2	11.38	37
OZ218256.1	3	11.35	36.50
OZ218257.1	4	11.30	37
OZ218258.1	5	11.22	37
OZ218259.1	6	11.14	36
OZ218260.1	7	10.59	36
OZ218261.1	8	10.53	36
OZ218262.1	9	10.52	36.50
OZ218263.1	10	10.38	36
OZ218264.1	11	10.12	37
OZ218265.1	12	10.11	36.50
OZ218266.1	13	10.05	36.50
OZ218267.1	14	9.68	36.50
OZ218268.1	15	9.48	37
OZ218269.1	16	9.39	36.50
OZ218270.1	17	9.34	37
OZ218271.1	18	8.96	37.50
OZ218272.1	19	8.48	36.50
OZ218273.1	20	8.37	37
OZ218274.1	21	8.07	38
OZ218275.1	22	7.39	37
OZ218276.1	23	6.76	37
OZ218277.1	24	6.72	38.50
OZ218278.1	25	6.65	37.50
OZ218279.1	26	6.18	37.50
OZ218280.1	27	6.04	39
OZ218281.1	28	4.97	38.50
OZ218282.1	29	4.41	40.50
OZ218283.1	30	4.24	39
OZ218284.1	Z	14.49	37

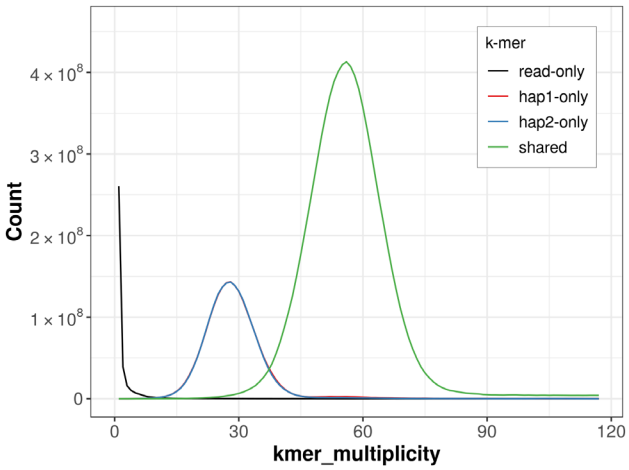

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

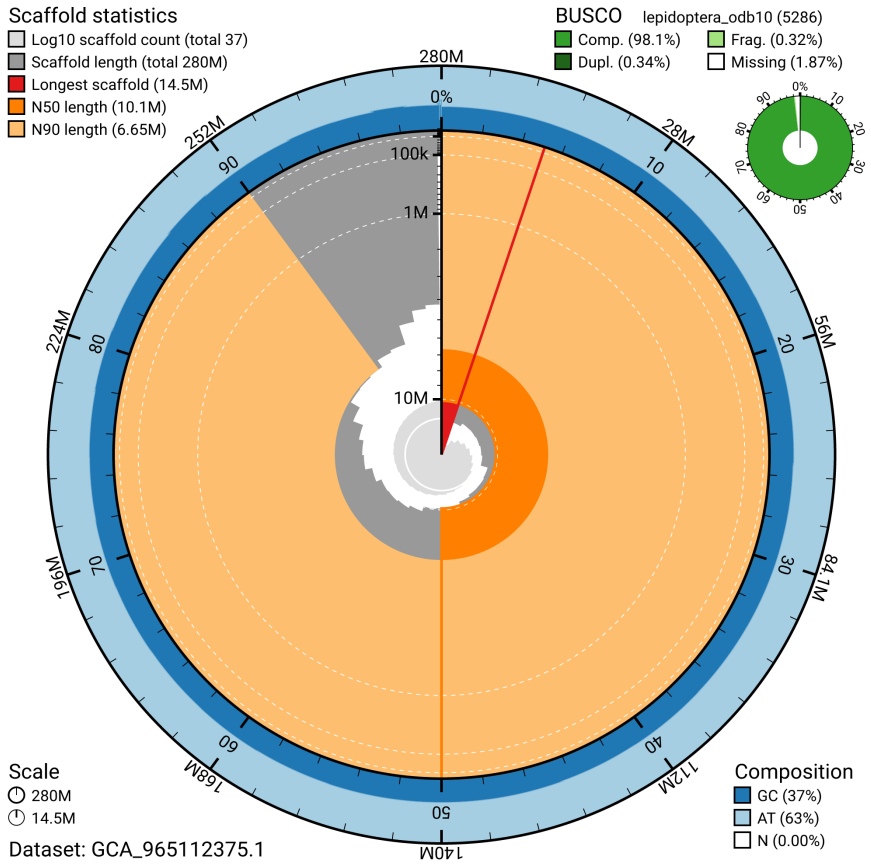

Figure 5. Assembly metrics for ilCycLine1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost 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.

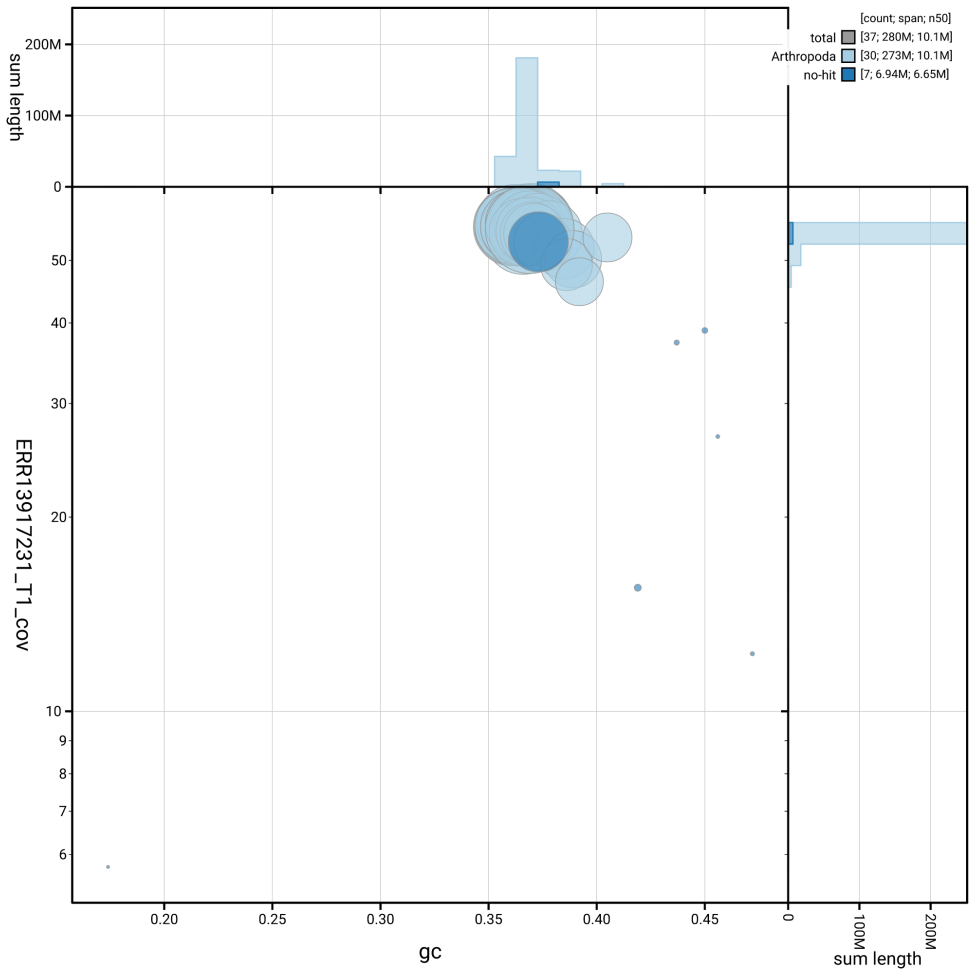

Figure 6. BlobToolKit GC-coverage plot for ilCycLine1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Cyclophora linearia* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q64	6.C.Q40
Contig N50 length	7.52 Mb	≥ 1 Mb
Scaffold N50 length	10.11 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 64.9; haplotype 2: 64.1; combined: 64.5	≥ 40
k-mer completeness	Haplotype 1: 75.56%; Haplotype 2: 75.35%; combined: 99.39%	≥ 95%
BUSCO	C:98.1% [S:97.8%; D:0.3%]; F:0.3%; M:1.6%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.90%	≥ 90%

Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances, other Darwin Tree of Life collaborators.

Data availability

European Nucleotide Archive: Cyclophora linearia. Accession number PRJEB81921. The genome sequence is released openly for reuse. The *Cyclophora linearia* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665), 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.

Production code used in genome assembly at the WSI Tree of Life is available at https://github.com/sanger-tol. Table 5 lists software versions used in this study.

Author information

Contributors are listed at the following links:

- Members of the University of Oxford and Wytham Woods Genome Acquisition Lab
- Members of the Natural History Museum Genome Acquisition Lab
- Members of the Darwin Tree of Life Barcoding collective
- 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 Darwin Tree of Life Consortium

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
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc

Software	Version	Source
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

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

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

Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

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

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial arthropod fauna [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 123.
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

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

GBIF Secretariat: ***Cyclophora linearia*, (Hübner) in GBIF Backbone Taxonomy.** 2023.
Publisher Full Text

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

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): 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

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

Randle Z, Evans-Hill LJ, Parsons MS, *et al.*: **Atlas of Britain & Ireland's larger moths.** Newbury, UK: Pisces Publications, 2019.
Reference Source

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

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2024; **9**: 339.
PubMed Abstract | Publisher Full Text | Free 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

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

Wheeler K, Dickson R: **Clay Triple-lines *Cyclophora linearia* (Hübner, [1799]) feeding on oak.** *Bulletin of the Amateur Entomologists' Society.* 2022; **81**: 183.
Reference Source

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 04 March 2026

https://doi.org/10.21956/wellcomeopenres.27973.r142866

✔ **Annabel Whibley**
Bragato Research Institute, Clyde, New Zealand

Sivess and colleagues report the genome sequence of the Clay Triple-lines, *Cyclophora linearia.* The report follows Darwin Tree of Life pipelines and templates, resulting in a high quality assembly of the two haplotypes, one scaffolded to chromosome level, and presented in comprehensive report of assembly properties,evaluation metrics, sample and sequencing metadata.
Sample identification was verified by barcoding. Two specimens were used for analysis: ilCycLine1 for the assembly backbone and ilCycLine2 for RNA and HiC. I assume that this sentence “Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI” means that the identify of the second sample was confirmed via checking of the barcode sequence in the new data tranches, but perhaps the verification of the identity of additional specimens could be made more explicit.

By virtue of having no high coverage contigs, the blobplot is quite compressed towards the top of the y-axis. Perhaps it could be manually rescaled to improve the visualisation?

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, 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 28 January 2026

https://doi.org/10.21956/wellcomeopenres.27973.r143294

Mohadeseh Tahami
University of Jyväskylä, Jyväskylä, Finland

The manuscript presents a high-quality, chromosome-scale genome assembly for *Cyclophora linearia*, generated using state-of-the-art PacBio, Hi-C, and RNA-seq approaches. The methods are described very well in details, ensuring reproducibility. The sequencing and assembly statistics are provided in details; the genome assembly includes two haplotypes with total lengths of 280.22 Mb (haplotype 1) and 284.23 Mb (haplotype 2), assembly completeness by BUSCO is estimated 98%, 31 chromosomes were scaffolded including the sex chromosome, and the complete mitochondrial genome is assembled with 16.54 kb length. Biological species background are well described with the rationale of the genome's value for ecological and evolutionary studies. Minor improvements could include a concise comparative context with related Geometridae genomes. Overall, this is a solid and valuable genome report suitable for indexing.

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

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

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

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: evolutionary 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.