DATA NOTE

The genome sequence of the Small China-mark, *Cataclysta lemnata* (Linnaeus, 1758) (Lepidoptera: Crambidae)

[version 1; peer review: 6 approved, 1 approved with reservations]

David C. Lees[1], Natural History Museum 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

v1 **First published:** 21 Nov 2025, **10**:653
https://doi.org/10.12688/wellcomeopenres.25182.1
Latest published: 21 Nov 2025, **10**:653
https://doi.org/10.12688/wellcomeopenres.25182.1

Abstract

We present a genome assembly from an individual male *Cataclysta lemnata* (Small China-mark; Arthropoda; Insecta; Lepidoptera; Crambidae). The genome sequence has a total length of 432.44 megabases. Most of the assembly (99.51%) is scaffolded into 28 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled, with a length of 15.36 kilobases. 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

Cataclysta lemnata; Small China-mark; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ? ✓ ✓ ✓ ✓

	1	2	3	4	5	6
version 1 21 Nov 2025	✓ view	? view	✓ view	✓ view	✓ view	✓ view

1. **Laurence Despres**, Laboratoire d'Ecologie Alpine (Ringgold ID: 56837), Grenoble, France
2. **Christophe Klopp**, INRAE, Castanet-Tolosan, France
3. **Stephan Koblmüller**, University of Graz, Graz, Austria
4. **Xiaoguang Liu**, Henan Agricultural University, Zhengzhou, China
5. **Melody Clark**, Natural Environment Research Council, Cambridge, UK
6. **Yvonne Jing Mei Liew**, Universiti Malaya, Kuala Lumpur, Malaysia

Please see the article web page for additional

reviewers.

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

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

Author roles: Lees DC: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing;

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: Lees DC, Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding Collective *et al.* **The genome sequence of the Small China-mark, *Cataclysta lemnata* (Linnaeus, 1758) (Lepidoptera: Crambidae) [version 1; peer review: 6 approved, 1 approved with reservations]** Wellcome Open Research 2025, **10**:653
https://doi.org/10.12688/wellcomeopenres.25182.1

First published: 21 Nov 2025, **10**:653 https://doi.org/10.12688/wellcomeopenres.25182.1

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; Pyraloidea; Crambidae; Nymphulinae; *Cataclysta*; *Cataclysta lemnata* (Linnaeus, 1758) (NCBI:txid1594245)

Background

Cataclysta lemnata (Linnaeus, 1758), known in Britain as the Small China-mark, is a moth in the family Crambidae, belonging to the largely aquatic subfamily Acentropinae. It is sexually dimorphic. The male has a forewing length of 8–11 mm (Sterling *et al.*, 2023), and the wingspan is about 13–18 mm in males and 18–24 mm in females (Farahpour-Haghani *et al.*, 2017). The male forewing is glossy white with a post-median black discal spot, and the hindwing has a dark discal crescent bordered by irregular narrow brown lines and a black margin containing about 6–7 white glittery eyespots; the female is yellowish-brown, with the same ocelli (Sterling *et al.*, 2023). These eyespots are often displayed at rest when the moth adopts a shallowly tented triangular posture and has a general resemblance to the eyes of salticid spiders (Hill, 2022: 15). The moth flies from May to late September, with a peak in Britain in July and August (Norfolk Moths, 2025). The male flies crepuscularly over the water surface, while females are nocturnal (Catalogue of the Lepidoptera of Belgium, 2024). Both sexes, especially females, are attracted to light.

The Small China-mark occurs in freshwater habitats including ponds and lakes (Sterling *et al.*, 2023). The oval larval case made of leaf fragments contains bubbles which allow the larva to breathe and the white oval cocoon is attached to leaves just below the water surface (Catalogue of the Lepidoptera of Belgium, 2024; Marković *et al.*, 2024).

There are currently 3 568 records of *C. lemnata* on the NBN Atlas Partnership (2025) and it is widely distributed in England, and Ireland (Moths Ireland, 2025). In Scotland it occurs mainly in the southwest (Scottish Micro Moth Distribution Maps, 2025); it also occurs along the eastern seaboard of Ireland (Moths Ireland, 2025). There are 24 039 records (GBIF Secretariat, 2025), distributed from Ireland through Europe as far east as Lake Baikal and from southern Scandinavia to the shores of the Mediterranean, Spain, Crete and the Caspian Sea. It is one of the most abundant Palaearctic aquatic moths (Marković *et al.*, 2024).

C. lemnata larvae feed particularly on Duckweed (*Lemna* L., Araceae) (Marković *et al.*, 2024) and have been recently found to feed on Water ferns (*Azolla* Lam., Azollaceae) in Iran (Farahpour-Haghani *et al.*, 2017). They can eat other aquatic macrophytes such as *Stratiotes aloides* L., *Typha latifolia* L. and *Glyceria maxima* (Hartm.) Holmb. (van der Velde, 1988). The early stages, first studied in detail by Chapman (1905) are described and illustrated in Farahpour-Haghani *et al.* (2017).

On BOLD (accessed 14/10/2025) the DNA barcode from the mitochondrial assembly (OZ008366.1) represents the Barcode Index Number (BIN) BOLD:AAC2080, with limited haplotype variability (maximum 0.64%, $n = 53$) and identical haplotypes occurring in Europe. This cluster is remote (a *p*-distance of over 9%) from other acentropine crambids on BOLD. Hawking (2014) considered it different from seven other species of *Cataclysta*, which may not merit belonging in the same genus.

TWe present a chromosomally complete genome sequence for *C. lemnata*. The genome will be useful for investigating the phylogenetic relationships of *C. lemnata* and adaptations to aquatic life (Pabis, 2018), particularly in its early stages. Their larvae can eat holes in PVC foil, providing another example of Lepidoptera that can be used in studies of plastic waster degradation (van der Velde, 1991). It may well also be useful in further studies of the properties of underwater silk (Heckenhauer *et al.*, 2024).

The assembly was produced using the Tree of Life pipeline from a specimen collected in Winterton Dunes, England, UK (Figure 1), as part of the Darwin Tree of Life project.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Cataclysta lemnata* (specimen ID NHMUK015054648, ToLID ilCatLemn2; Figure 1), collected from Winterton Dunes, England, United Kingdom (latitude 52.73, longitude 1.69) on 2022-06-29. The specimen was collected and identified by David Lees (Natural History Museum).

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

Figure 1. Photographs of the *Cataclysta lemnata* (ilCatLemn2) specimen used for genome sequencing.

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.

Sample metadata were collected in line with the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

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 ilCatLemn2 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the head and 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. We used centrifuge-mediated fragmentation to produce DNA fragments in the 8–10 kb range, following the Covaris g-TUBE protocol for ultra-low input (ULI). 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.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Prior to library preparation, the DNA was fragmented to ~10 kb. Ultra-low-input (ULI) libraries were prepared using the PacBio SMRTbell® Express Template Prep Kit 2.0 and gDNA Sample Amplification Kit. Samples were normalised to 20 ng DNA. Single-strand overhang removal, DNA damage repair, and end-repair/A-tailing were performed according to the manufacturer's instructions, followed by adapter ligation. A 0.85× pre-PCR clean-up was carried out with Promega ProNex beads.

The DNA was evenly divided into two aliquots for dual PCR (reactions A and B), both following the manufacturer's protocol. A 0.85× post-PCR clean-up was performed with ProNex beads. DNA concentration was measured using a Qubit Fluorometer v4.0 (Thermo Fisher Scientific) with the Qubit HS Assay Kit, and fragment size was assessed on an Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit. PCR reactions A and B were then pooled, ensuring a total mass of ≥500 ng in 47.4 μl.

The pooled sample underwent another round of DNA damage repair, end-repair/A-tailing, and hairpin adapter ligation. A 1× clean-up was performed with ProNex beads, followed by DNA quantification using the Qubit and fragment size analysis using the Agilent Femto Pulse. Size selection was performed on the Sage Sciences PippinHT system, with target fragment size determined by Femto Pulse analysis (typically 4–9 kb). Size-selected libraries were cleaned with 1.0× ProNex beads and normalised to 2 nM before sequencing.

The sample was sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was in the range 40–135 pM. The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, and to perform primary and secondary analysis of the data upon completion.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen head and thorax tissue of the ilCatLemn2 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 were created. 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 6000.

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 (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). The Hi-C reads (Rao *et al.*, 2014) were mapped to the

primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019), and the contigs were scaffolded in YaHS (Zhou *et al.*, 2023) with 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 46 breaks, 88 joins, and removal of 24 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

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 the primary and alternate 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 *Cataclysta lemnata* specimen generated 29.41 Gb (gigabases) from 2.70 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 389.68 Mb, with a heterozygosity of 0.86% and repeat content of 22.63% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 71× coverage. Hi-C sequencing produced 134.27 Gb from 889.23 million reads, which were

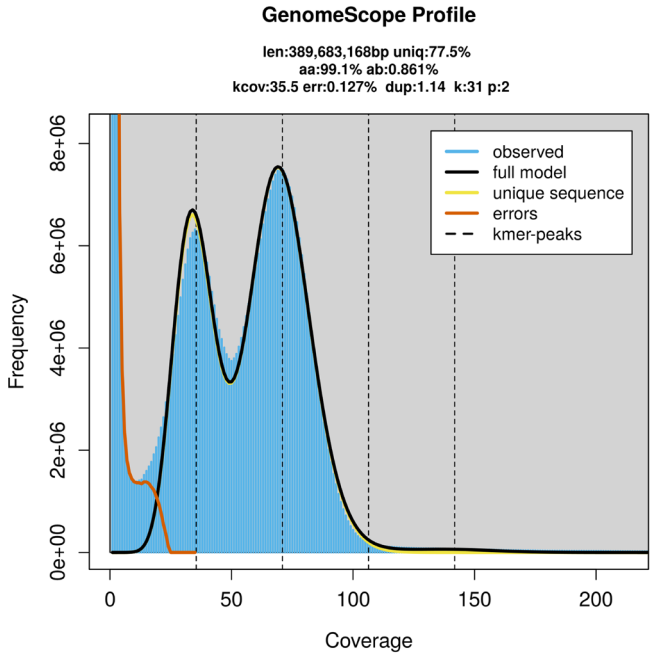

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The final assembly has a total length of 432.44 Mb in 137 scaffolds, with 874 gaps, and a scaffold N50 of 15.95 Mb (Table 2).

Most of the assembly sequence (99.51%) was assigned to 28 chromosomal-level scaffolds, representing 27 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Z chromosome identified based on alignment with the genome of *Nymphula nitidulata* (GCA_947347705.1).

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.

The combined primary and alternate assemblies achieve an estimated QV of 59.9. The *k*-mer completeness is 85.17% for the primary assembly, 84.70% for the alternate haplotype, and 99.36% for the combined assemblies (Figure 4).

BUSCO v.5.5.0 analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 96.6% of the expected gene set (single = 96.2%, duplicated = 0.3%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for the primary assembly. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage.

Table 1. Specimen and sequencing data for BioProject PRJEB72119.

Platform	PacBio HiFi	Hi-C
ToLID	ilCatLemn2	ilCatLemn2
Specimen ID	NHMUK015054648	NHMUK015054648
BioSample (source individual)	SAMEA112963103	SAMEA112963103
BioSample (tissue)	SAMEA112963199	SAMEA112963199
Tissue	head and thorax	head and thorax
Instrument	Sequel IIe	Illumina NovaSeq 6000
Run accessions	ERR12535459	ERR12543000
Read count total	2.70 million	889.23 million
Base count total	29.41 Gb	134.27 Gb

Table 2. Genome assembly statistics.

Assembly name	ilCatLemn2.1
Assembly accession	GCA_963931975.1
Alternate haplotype accession	GCA_963931965.1
Assembly level	chromosome
Span (Mb)	432.44
Number of chromosomes	28
Number of contigs	1 011
Contig N50	0.99 Mb
Number of scaffolds	137
Scaffold N50	15.95 Mb
Sex chromosomes	Z
Organelles	Mitochondrion: 15.36 kb

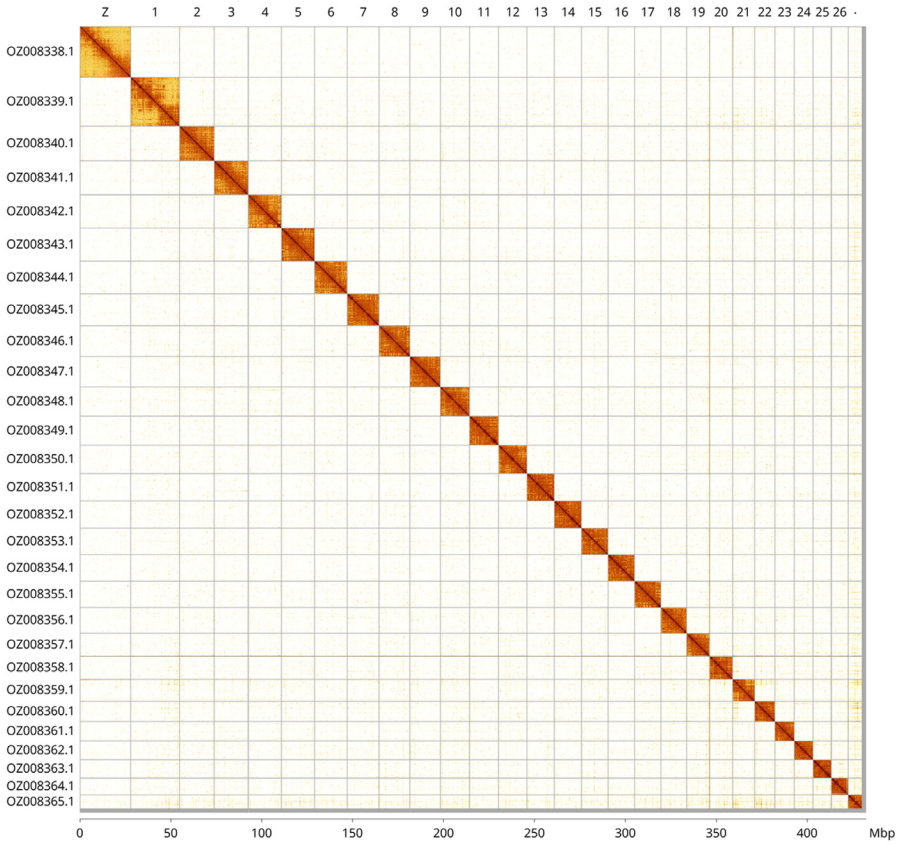

Figure 3. Hi-C contact map of the *Cataclysta lemnata* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Cataclysta lemnata* ilCatLemn2.

INSDC accession	Molecule	Length (Mb)	GC%
OZ008339.1	1	26.92	37.50
OZ008340.1	2	19.10	37.50
OZ008341.1	3	18.70	37.50
OZ008342.1	4	18.24	38
OZ008343.1	5	18.14	37.50
OZ008344.1	6	18.04	37.50
OZ008345.1	7	17.54	37.50
OZ008346.1	8	16.96	37.50
OZ008347.1	9	16.72	37.50
OZ008348.1	10	16.04	38
OZ008349.1	11	15.95	38
OZ008350.1	12	15.55	38
OZ008351.1	13	15.14	37.50
OZ008352.1	14	14.96	38
OZ008353.1	15	14.60	38
OZ008354.1	16	14.56	38
OZ008355.1	17	14.46	38.50
OZ008356.1	18	14.14	38
OZ008357.1	19	12.71	39.50
OZ008358.1	20	12.55	38
OZ008359.1	21	12.18	39
OZ008360.1	22	11.11	39
OZ008361.1	23	10.66	39
OZ008362.1	24	10.34	39
OZ008363.1	25	10.09	39.50
OZ008364.1	26	9.20	39.50
OZ008365.1	27	7.74	40
OZ008338.1	Z	27.99	37

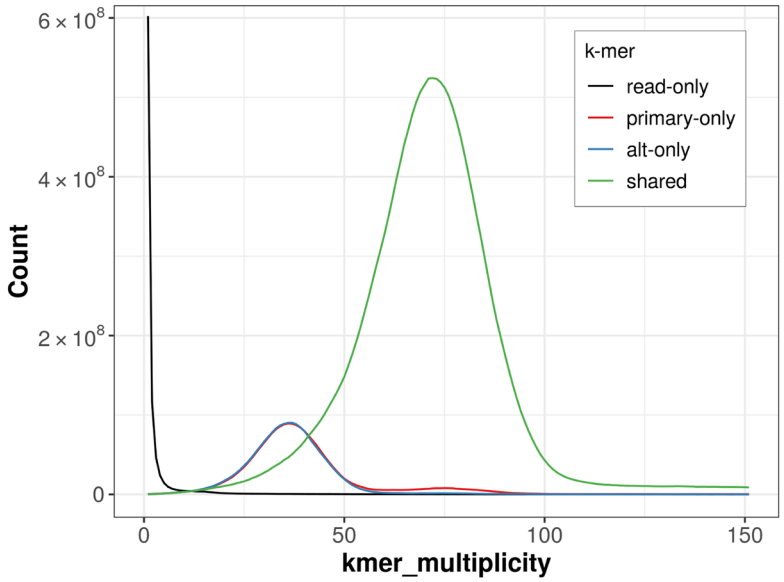

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

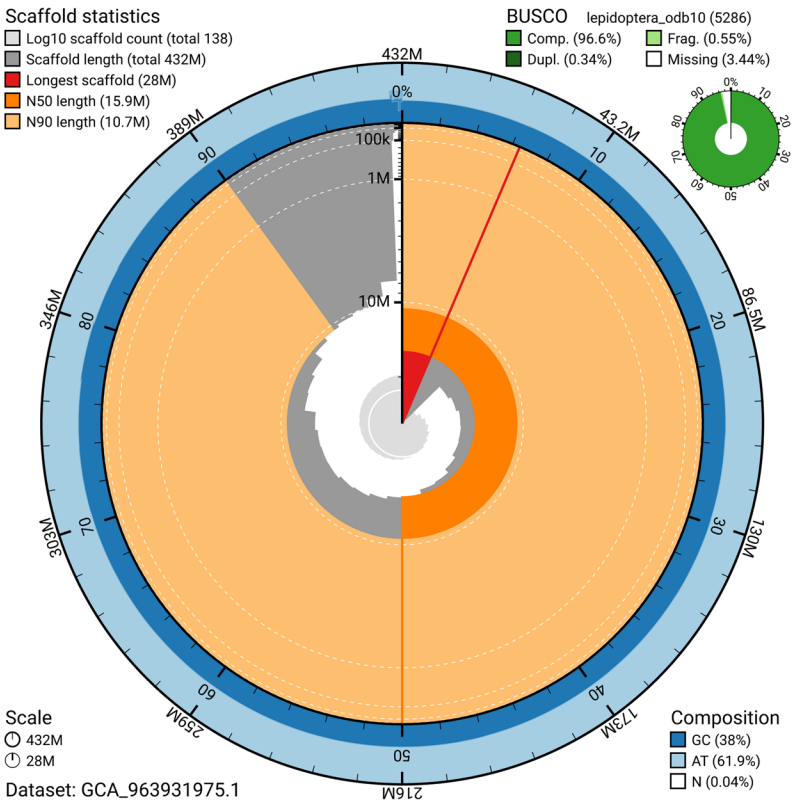

Figure 5. Assembly metrics for ilCatLemn2.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue 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 lepidoptera_odb10 set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

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 for the primary assembly is **5.C.Q59**.

Genome annotation report

The *Cataclysta lemnata* genome assembly (GCA_963931975.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation was done using BRAKER2

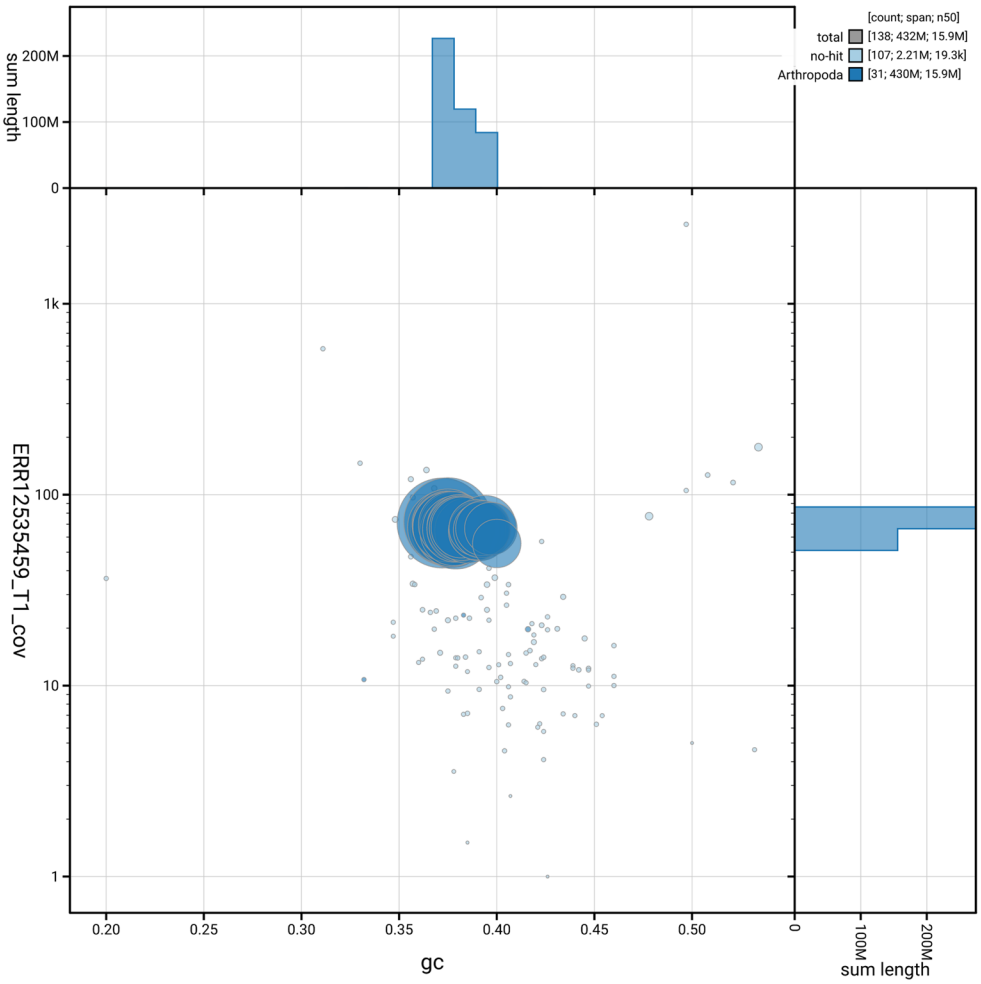

Figure 6. BlobToolKit GC-coverage plot for ilCatLemn2.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display 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 *Cataclysta lemnata* assembly.

Measure	Value	Benchmark
EBP summary (primary)	5.C.Q59	6.C.Q40
Contig N50 length	0.99 Mb	≥ 1 Mb
Scaffold N50 length	15.95 Mb	= chromosome N50
Consensus quality (QV)	Primary: 59.9; alternate: 60.8; combined: 59.9	≥ 40
k-mer completeness	Primary: 85.17%; alternate: 84.70%; combined: 99.36%	≥ 95%
BUSCO	C:96.6% [S:96.2%; D:0.3%]; F:0.5%; M:2.9%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.51%	≥ 90%

and identified 17 004 protein-coding genes. 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, Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances, other Darwin Tree of Life collaborators.

Data availability

European Nucleotide Archive: Cataclysta lemnata (small china-mark). Accession number PRJEB72119. The genome sequence is released openly for reuse. The *Cataclysta lemnata* 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 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

Software	Version	Source
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	23.04.1	https://github.com/nextflow-io/nextflow
PretextSnapshot	-	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.4.0	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

Catalogue of the Lepidoptera of Belgium: **Cataclysta lemnata (Linnaeus, 1758)**. 2024.
Reference Source

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

Chapman TA: **The earlier stages of *Cataclysta lemnata*, L.** *Entomologist*. 1905; **38**: 1–5, 38–43, 90.

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

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

Farahpour-Haghani A, Hassanpour M, Alinia F, *et al.*: **Water ferns *Azolla* spp. (Azollaceae) as new host plants for the small China-mark moth, *Cataclysta lemnata* (Linnaeus, 1758) (Lepidoptera, Crambidae, Acentropinae).** *Nota Lepidopterologica.* 2017; **40**(1): 1–13.
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: ***Cataclysta lemnata* Linnaeus, 1758**. 2025.
Reference Source

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898.
PubMed Abstract | Publisher Full Text | Free Full Text

Hawking JH: **Systematics and ecology of the Australian aquatic moths, Acentropinae (Insecta: Lepidoptera)**. 2014.
Reference Source

Heckenhauer J, Plotkin D, Martinez JI, *et al.*: **Genomic resources of aquatic

Lepidoptera, *Elophila obliteralis* and *Hyposmocoma kahamanoa*, reveal similarities with Trichoptera in amino acid composition of major silk genes. *G3 (Bethesda).* 2024; **14**(9): jkae093.
PubMed Abstract | Publisher Full Text | Free Full Text

Hill DE: **A salticid archetype for salticid spiders.** *Peckhamia.* 2022; **275**(1): 1–39.
Reference Source

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

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

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

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

Lawniczak MKN, Davey RP, Rajan J, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project [version 1; peer review: 2 approved with reservations].** *Wellcome Open Res.* 2022; **7**: 187.
Publisher 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

Marković V, Novaković B, Božanić M, *et al.*: **The small China-mark *Cataclysta lemnata* (Linnaeus, 1758) (Lepidoptera: Crambidae: Acentropinae) in Serbia.** *Acta Entomologica Serbica.* 2024; **29**(1): 63–70.
Publisher Full Text

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

Moths Ireland: **63.116 Cataclysta lemnata (small China-mark)**. 2025.
Reference Source

NBN Atlas Partnership: **Cataclysta lemnata (Linnaeus, 1758) Small China-mark**. 2025.
Reference Source

Norfolk Moths: **Cataclysta lemnata Small China-mark**. 2025.
Reference Source

Pabis K: **What is a moth doing under water? Ecology of aquatic and semi-aquatic Lepidoptera.** *Knowl Manag Aquat Ecosyst.* 2018; **419**: 42.
Publisher Full Text

Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat Commun.* 2020; **11**(1): 1432.
PubMed Abstract | Publisher Full Text | Free Full Text

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI taxonomy: a comprehensive update on curation, resources and tools.** *Database (Oxford).* 2020; **2020**: baaa062.
PubMed Abstract | Publisher Full Text | Free Full Text

Scottish Micro Moth Distribution Maps: **63.116 *Cataclysta lemnata***. 2025.
Reference Source

Sterling P, Parsons M, Lewington R, *et al.*: **Field guide to the Micro-Moths of Great Britain and Ireland**. London: Bloomsbury Publishing, 2023.
Reference Source

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

van der Velde G: ***Cataclysta lemnata* L. (Lepidoptera, Pyralidae) can survive for several years consuming macrophytes other than Lemnaceae.** *Aquat Bot.* 1988; **31**(1–2): 183–89.
Publisher Full Text

van der Velde G: **Rupsen van *Cataclysta lemnata* (L.) (Lepidoptera, Pyralidae) kunnen PVC vijverfolie perforeren.** *Entomologische Berichten.* 1991; **51**(7): 94–95.
Reference Source

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

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 February 2026

https://doi.org/10.21956/wellcomeopenres.27757.r144133

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

This paper describes the sequencing and assembly of the genome of the lepidopteran species, Cataclysta lemnata.

The quality of the assembly is very high, as shown by the various metrics provided by the authors. Firstly, virtually the entire assembly is assigned to one of the 28 chromosome-level scaffolds. Secondly, k-mer completeness is >99%. Thirdly, QV = 59.9. Finally, complete BUSCOs = 96.6%, even though other genomes have a slightly better score (>99%), this is an acceptable BUSCO score as well.

The blob plot (Figure 6) shows that there are some scaffolds of lower coverage and lower size than the chromosome-level scaffolds. What are these scaffolds? Could the authors comment on this?

I also saw that the authors provide (through Ensembl) a gene set for this genome. However, no evaluation of this gene set is provided. I would at least expect a BUSCO evaluation in order to compare it with the BUSCO evaluation of the assembly; this way the interested researcher can tell how good the gene set actually is. Perhaps, for example, it is missing conserved genes. Or maybe all conserved genes are predicted correctly, but it's missing other, less conserved genes which are, nevertheless, essential for the organism.

However, without a proper evaluation, one cannot know if he/she should use the provided gene set or if it's more worth investing time to generate a new (and more correct) gene set. Ensembl should definitely provide such an evaluation, but if it doesn't then it's up to the authors to do it. Running BUSCO is not such a difficult thing to do...

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 03 February 2026

https://doi.org/10.21956/wellcomeopenres.27757.r145054

Yvonne Jing Mei Liew
Universiti Malaya, Kuala Lumpur, Kuala Lumpur, Malaysia

Background
- Please correct the following sentences.

-In Scotland, it occurs mainly in the southwest
-TWe present a chromosomally complete genome sequence for C. lemnata.
-example of Lepidoptera that can be used in studies of plastic waste degradation
- Please verify the subfamily of C. lemnata. In the Background section Acentropinae was used but in the Species taxonomy Nymphulinae was used.
- In this statement, an old Dutch reference from 1991 was cited but there is actually newer reference available from year 2025. Please consider the provided reference.

Their larvae can eat holes in PVC foil, providing another example of Lepidoptera that can be used in studies of plastic waster degradation (van der Velde, 1991).
Gallitelli, L., Ceschin, S., Mariani, F., Pietrelli, L., & Scalici, M. (2025). Preliminary Observations on the Use of Microplastics by Aquatic Larvae of the Moth Cataclysta lemnata (Linnaeus, 1758). Environments, 12(3), 80. https://doi.org/10.3390/environments12030080
- chromosomally complete genome sequence

Please replace “chromosomally complete genome sequence” with appropriate terms such as chromosomal-level scaffolds. This is because although 99.51% of the assembly is assigned to 28 chromosomal scaffolds, the final assembly still contains 137 scaffolds and 874 gaps. The description chromosomally complete would imply gap-free chromosomes.

Methods - Sample acquisition and DNA barcoding
Please elaborate on where the tissue sample was dissected in the following section.
-A small sample was dissected from the specimen and stored in ethanol

Genome sequence report – Table 4
The reported contig N50 length of 0.99 Mb does not meet the stated benchmark (≥1 Mb). Please provided an explanation if this is acceptable in the main text.

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: Biochemistry

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 03 February 2026

https://doi.org/10.21956/wellcomeopenres.27757.r145052

Melody Clark
Natural Environment Research Council, Cambridge, UK

This article describes the genome sequencing of a male specimen of the small China-mark moth. The biology and distribution of this species is well described, justifying it's inclusion in the Darwin Tree of Life sequencing project. The potential uses of the genome sequence, beyond taxonomy, including the fact that the larvae can eat holes in PVC and that the genome could be used to understand the properties of underwater silk were also explored. The procedures for isolating very small amounts of DNA, the production of libraries, sequencing strategies and assemblies are well described and follow standard Darwin Tree of Life protocols, which are all open source and freely available. The genome is of a very high standard resulting in 28 scaffolds with 27 autosomes

and the Z sex chromosome. All sequencing data are freely available.

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 transcriptomics of invertebrates and fish.

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

Reviewer Report 02 February 2026

https://doi.org/10.21956/wellcomeopenres.27757.r145051

Xiaoguang Liu
Henan Agricultural University, Zhengzhou, China

As a key component of the Darwin Tree of Life Project, this study achieves the first chromosome-level genome assembly of *Cataclysta lemnata*, effectively addressing the genomic resource gap for this species. The research background is articulated with clarity, detailing the species' taxonomic status, morphological characteristics, ecological habits, and distribution range, while underscoring the scientific value of genomic data in elucidating its aquatic adaptation, phylogenetic relationships, and unique biological traits. The experimental design is rigorous, employing PacBio HiFi long-read sequencing in conjunction with Hi-C chromatin conformation capture technology, ensuring sufficient sequencing depth, standardized assembly procedures, and comprehensive data analysis. The overall technical approach adheres to the high standards of contemporary genomics research. The genome assembly exhibits excellent quality, with a chromosome anchoring rate of 99.51% and BUSCO completeness reaching 96.6%. All core metrics conform to the standards for eukaryotic reference genomes, providing robust data support for subsequent related research. The paper presents a well-structured framework and clear data presentation, complying with open science research indexing standards, and demonstrates significant scientific value and application prospects.

It is recommended that the authors supplement the details of their genome annotation. The manuscript currently only mentions the protein-coding gene annotation conducted using BRAKER2, which predicts a total of 17,004 protein-coding genes. However, it lacks comprehensive annotation information, including functional annotation results (such as Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Pfam), as well as non-coding RNA annotations (e.g., microRNA (miRNA) and transfer RNA (tRNA)). To enhance the utility of the genomic resources presented, it is advisable to provide a complete genome annotation report that details the annotation pipeline and the core databases utilized.

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: Physiological Toxicology, Insect Genomics and Evolutionary Research.

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 January 2026

https://doi.org/10.21956/wellcomeopenres.27757.r144127

Stephan Koblmüller
University of Graz, Graz, Graz, Austria

This data note by Lees et al. presents the genome sequence of the small China-mark, *Cataclysta lemnata*. The authors provide comprehensive species background information and detailed methodologies covering specimen collection, sequencing, assembly, and quality assessments. I have no major concerns but have only two minor comments:

1. Background - Barcode / BIN: What exactly do you mean to say here; that the same haplotype as the one you sequenced is found across Europe, or that this applies to several haplotypes (not just the one you sequenced)? So, maybe clarify this by slightly rephrasing this sentence. And, in general, isn't this a result of the study that actually belongs into a

Results section?
2. Background: Change “TWe present a chromosomally …” to “We present a chromosomally …”

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: phylogenetics/-genomics, population genenetics

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 27 January 2026

https://doi.org/10.21956/wellcomeopenres.27757.r145049

?

Christophe Klopp
INRAE, Castanet-Tolosan, France

The authors present the first genome assembly of the Small China-mark, *Cataclysta lemnata.* The sequencing protocol and bioinformatic pipelines used are standard.

The authors should include the Hi-C data in the hifiasm assembly as this is best practice to produce phased contigs and chromosomes.

The authors should provide hap2 in chromosomes this will ease the work of biologists interested in the species, for example to find variations between haplotypes.

Figure 4. (Evaluation of *k*-mer completeness using MerquryFK.) show that some haplotype share kmers are missing in the primary assembly. This should be discussed.

Why are the genomescope2 and merqury kmer plot not representing the same proportions

between specific and shared kmers? There seem to be much less specific kmers found by merqury.

"TWe present a" should be replaced by "We present a"

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: genome assembly and annotation

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.

Reviewer Report 22 January 2026

https://doi.org/10.21956/wellcomeopenres.27757.r145050

Laurence Despres
Laboratoire d'Ecologie Alpine (Ringgold ID: 56837), Grenoble, Auvergne-Rhône-Alpes, France

The authors report on the chromosome-level genome assembly of a male specimen of the Small China-mark, *Cataclysta lemnata* (Lepidoptera; Crambidae) a common moth found in freshwater habitats including ponds and lakes throughout the Palaearctic. The genome sequence has a total length of 432.44 megabases. The assembly contains two haplotypes with most of haplotype 1 scaffolded into 28 chromosomes, including the Z sex chromosome. The mitochondrial genome has also been assembled.
The quality of the assembly was further assessed by the proportion of complete BUSCO genes recovered, which is very high : 96.6% of the expected gene set of lepidoptera_odb10 reference set (n = 5 286). The genome assembly is of good quality and was constructed using appropriate

methods: Pac Bio HiFi long reads (71 X coverage) and Hi-C data to confirm scaffolding, following the high standard of the Darwin Tree of Life Project for results presentation. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness.

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.