



DATA NOTE

The genome sequence of the Poplar Cosmet moth,

Batrachedra praeangusta (Haworth, 1828)

[version 1; peer review: 2 approved]

Gavin R. Broad ¹, Clare Boyes²,

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

¹Natural History Museum, London, England, UK

²Independent researcher, Welshpool, Wales, UK

V1 First published: 02 Jun 2025, 10:288
<https://doi.org/10.12688/wellcomeopenres.24308.1>

Latest published: 02 Jun 2025, 10:288
<https://doi.org/10.12688/wellcomeopenres.24308.1>

Abstract

We present a genome assembly from a female specimen of *Batrachedra praeangusta* (Poplar Cosmet; Arthropoda; Insecta; Lepidoptera; Batrachedridae). The genome sequence has a total length of 486.93 megabases. Most of the assembly (99.67%) is scaffolded into 29 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled, with a length of 15.21 kilobases.

Keywords

Batrachedra praeangusta, Poplar Cosmet moth, genome sequence, chromosomal, Lepidoptera

Open Peer Review

Approval Status

	1	2
version 1		
02 Jun 2025	view	view

1. Sivasankaran Kuppusamy , Loyola College, Chennai, India

2. Arun Arumugaperumal , Rajalakshmi Engineering College, Chennai, India

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: Broad GR: Investigation, Resources, Writing – Review & Editing; Boyes C: 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.

Copyright: © 2025 Broad GR *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Broad GR, Boyes C, Natural History Museum Genome Acquisition Lab *et al.* **The genome sequence of the Poplar Cosmet moth, *Batrachedra praeangusta* (Haworth, 1828) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, 10:288 <https://doi.org/10.12688/wellcomeopenres.24308.1>

First published: 02 Jun 2025, 10:288 <https://doi.org/10.12688/wellcomeopenres.24308.1>

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; Gelechioidea; Batrachedridae; Batrachedrinae; *Batrachedra*; *Batrachedra praeangusta* (Haworth, 1828) (NCBI:txid347710)

Background

Batrachedra praeangusta is a micro-moth in the family Batrachedridae. There are three species in this genus in the UK (Hall & Plant, 2022) and only about 140 worldwide, with most species in the tropics (Emmet & Langmaid, 2002). The moth is greyish, with some creamy shading, and has black and white striped legs. The narrow forewings measure between 4–8 mm and its antennae are almost as long as its wings. At rest, *B. praeangusta* looks very slender, tapering posteriorly, and sits with the head end raised. Common and widespread throughout England and Wales, but local in its distribution in Scotland and Ireland (Sterling *et al.*, 2012), *B. praeangusta* can easily be separated on colour pattern from the two similarly plain yellowish species, *B. confusella* and *B. pinicolella*.

The moth is found in a range of habitats but particularly woodlands where its caterpillar feeds on *Populus* spp. and *Salix* spp. Initially it feeds in the female catkins, before moving onto the buds of the tree where it feeds inside a protective web. After about three weeks the larva pupates, often secreting itself in the bark of the host tree. The pupal stage lasts up to two weeks before the adults emerge. They can be found by day and readily come to light, with one generation a year (Emmet & Langmaid, 2002).

The genome of the Poplar Cosmet, *Batrachedra praeangusta*, was sequenced as part of the Darwin Tree of Life Project, a collaborative effort to sequence all named eukaryotic species in the Atlantic Archipelago of Britain and Ireland.

Genome sequence report

Sequencing data

The genome of a specimen of *Batrachedra praeangusta* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 75.35 Gb from 8.23 million reads, which were used to assemble the genome. GenomeScope analysis estimated the haploid genome size at 470.66 Mb, with a heterozygosity of 1.19% and repeat content of 36.11%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 121 coverage. Hi-C sequencing produced 126.16 Gb from 835.52 million 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 assembly was improved by manual curation, which



Figure 1. Photographs of the *Batrachedra praeangusta* (ilBatPrae3) specimen used for genome sequencing.

corrected 148 misjoins or missing joins. These interventions reduced the total assembly length by 1.15% and decreased the scaffold count by 5.43%. The final assembly has a total length of 486.93 Mb in 86 scaffolds, with 125 gaps, and a scaffold N50 of 17.63 Mb (Table 2).

The snail plot in Figure 2 provides a summary of the assembly statistics, indicating the distribution of scaffold lengths and other assembly metrics. Figure 3 shows the distribution of scaffolds by GC proportion and coverage. Figure 4 presents a cumulative assembly plot, with separate curves representing different scaffold subsets assigned to various phyla, illustrating the completeness of the assembly.

Most of the assembly sequence (99.67%) was assigned to 29 chromosomal-level scaffolds, representing 28 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3). During curation, the Z chromosome was identified based on BUSCO gene painting with ancestral Merian elements (Wright *et al.*, 2024).

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

Assembly quality metrics

The estimated Quality Value (QV) and *k*-mer completeness metrics, along with BUSCO completeness scores, were calculated

Table 1. Specimen and sequencing data for *Batrachedra praeangusta*.

Project information			
Study title	Batrachedra praeangusta (poplar cosmet)		
Umbrella BioProject	PRJEB76773		
Species	Batrachedra praeangusta		
BioSpecimen	SAMEA111458703		
NCBI taxonomy ID	347710		
Specimen information			
Technology	ToLID	BioSample accession	Organism part
PacBio long read sequencing	ilBatPrae3	SAMEA111458790	abdomen
Hi-C sequencing	ilBatPrae2	SAMEA10979333	whole organism
Sequencing information			
Platform	Run accession	Read count	Base count (Gb)
Hi-C Illumina NovaSeq 6000	ERR13317836	8.36e+08	126.16
PacBio Sequel II	ERR13301420	5.74e+05	9.5
PacBio Revio	ERR13304157	7.65e+06	65.86

Table 2. Genome assembly data for *Batrachedra praeangusta*.

Genome assembly		
Assembly name	ilBatPrae3.1	
Assembly accession	GCA_964656335.1	
Alternate haplotype accession	GCA_964656285.1	
Assembly level for primary assembly	chromosome	
Span (Mb)	486.93	
Number of contigs	211	
Number of scaffolds	86	
Longest scaffold (Mb)	38.33	
Assembly metric	Measure	Benchmark
Contig N50 length	6.78 Mb	≥ 1 Mb
Scaffold N50 length	17.63 Mb	= chromosome N50
Consensus quality (QV)	Primary: 63.0; alternate: 59.2; combined: 60.1	≥ 40
k-mer completeness	Primary: 77.36%; alternate: 76.47%; combined: 99.69%	$\geq 95\%$
BUSCO*	C:97.9%;[S:97.4%,D:0.5%], F:0.3%,M:1.7%,n:5,286	$S > 90\%$; $D < 5\%$
Percentage of assembly mapped to chromosomes	99.67%	$\geq 90\%$
Sex chromosomes	Z	localised homologous pairs
Organelles	Mitochondrial genome: 15.21 kb	complete single alleles

* BUSCO scores based on the lepidoptera_odb10 BUSCO set using version 5.5.0. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison.

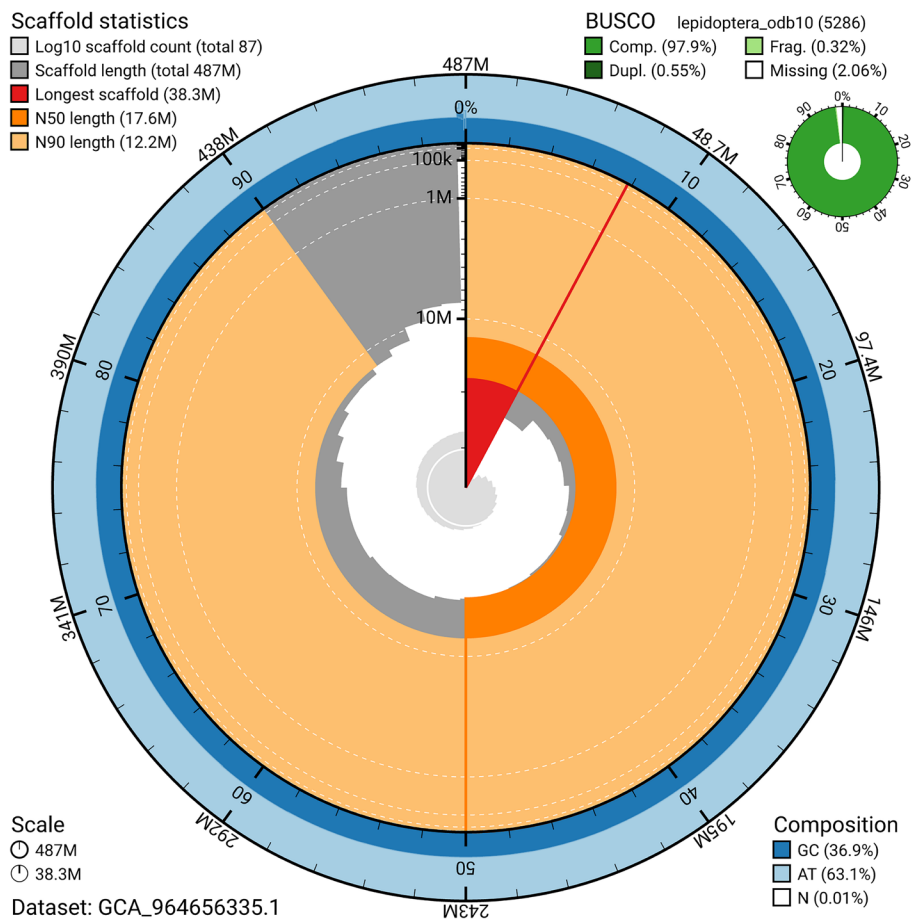


Figure 2. Genome assembly of *Batrachedra praeangusta*, ilBatPrae3.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the 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 is available at https://blobtoolkit.genomehubs.org/view/GCA_964656335.1/dataset/GCA_964656335.1/snail.

for each haplotype and the combined assembly. The QV reflects the base-level accuracy of the assembly, while *k*-mer completeness indicates the proportion of expected *k*-mers identified in the assembly. BUSCO scores provide a measure of completeness based on benchmarking universal single-copy orthologues.

The combined primary and alternate assemblies achieve an estimated QV of 60.1. The *k*-mer completeness is 77.36% for the primary haplotype and 76.47% for the alternate haplotype; and 99.69% for the combined primary and alternate assemblies. BUSCO v.5.5.0 analysis using the lepidoptera_odb10 reference set (*n* = 5,286) identified 97.9% of the expected gene set (single = 97.4%, duplicated = 0.5%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on

Assembly Standards September 2024. The primary assembly achieves the EBP reference standard of **6.C.Q63**.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult female *Batrachedra praeangusta* (specimen ID NHMUK010636412, ToLID ilBatPrae3), collected from Tonbridge, England, United Kingdom (latitude 51.19, longitude 0.29) on 2021-07-13, using actinic light. The specimen was collected and identified by Gavin Broad (Natural History Museum) and preserved by dry freezing (−80 °C).

A second specimen was used for Hi-C sequencing (specimen ID Ox001708, ToLID ilBatPrae2). This was an adult specimen collected from Middletown, Powys, Wales, United Kingdom

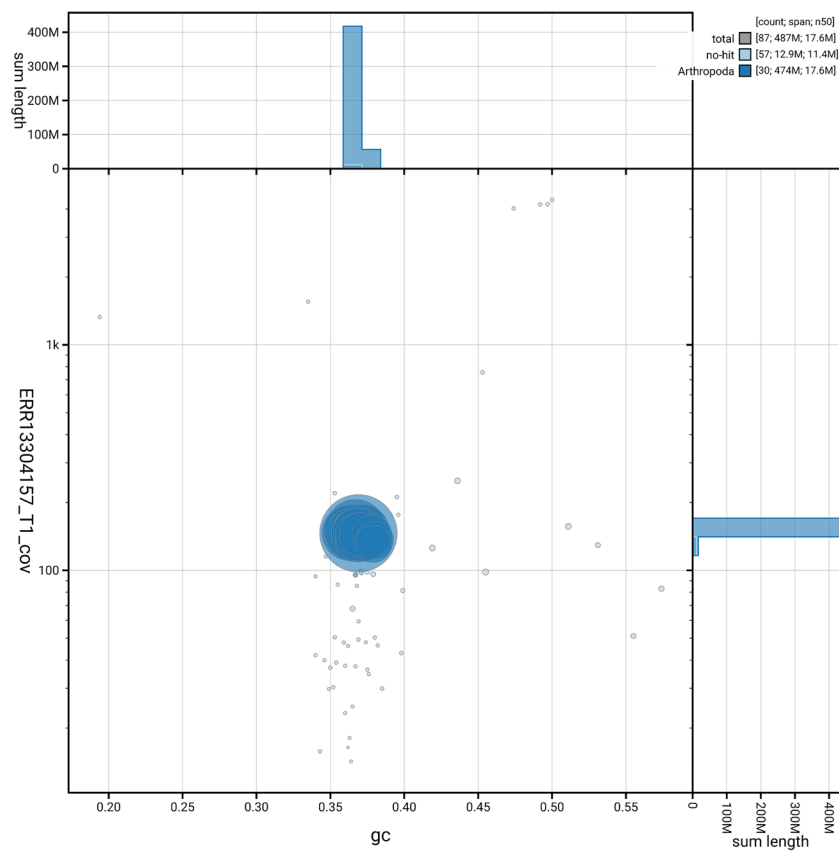


Figure 3. Genome assembly of *Batrachedra praeangusta*, ilBatPrae3.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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 at https://blobtoolkit.genomehubs.org/view/GCA_964656335.1/dataset/GCA_964656335.1/blob.

(latitude 52.702, longitude -3.03) on 2021-07-12, using a light trap. The specimen was collected and identified by Clare Boyes (independent researcher) and preserved on dry ice.

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 each specimen and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI) (Pereira *et al.*, 2022). 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 have been deposited on protocols.io (Beasley *et al.*, 2023).

Metadata collection for samples adhered to the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io (Howard *et al.*, 2025). The ilBatPrae3 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the abdomen was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023). HMW DNA was extracted using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023a). For ultra-low input (ULI) PacBio sequencing, DNA was fragmented using the Covaris g-TUBE method (Oatley *et al.*, 2023c). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Oatley *et al.*, 2023b). 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.

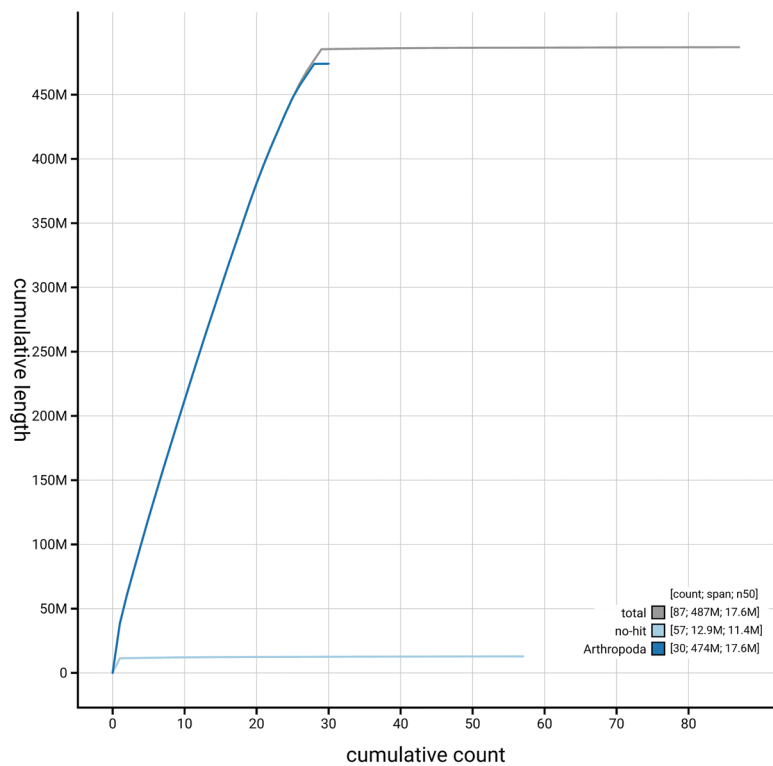


Figure 4. Genome assembly of *Batrachedra praeangusta*, ilBatPrae3.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the busco genes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_964656335.1/dataset/GCA_964656335.1/cumulative.

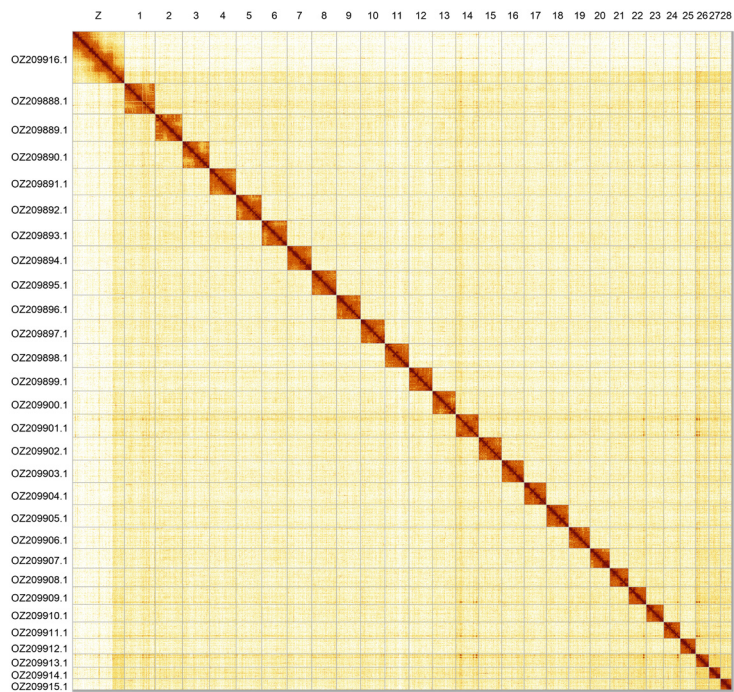


Figure 5. Genome assembly of *Batrachedra praeangusta*: Hi-C contact map of the ilBatPrae3.1 assembly, produced in PretextView. Chromosomes are shown in order of size and labelled.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Batrachedra praeangusta*, ilBatPrae3.

INSDC accession	Name	Length (Mb)	GC%
OZ209888.1	1	22.68	36.5
OZ209889.1	2	20.14	37
OZ209890.1	3	19.87	37
OZ209891.1	4	19.62	37
OZ209892.1	5	18.82	36.5
OZ209893.1	6	18.68	37
OZ209894.1	7	18.16	36.5
OZ209895.1	8	18.13	36.5
OZ209896.1	9	17.93	36.5
OZ209897.1	10	17.8	36.5
OZ209898.1	11	17.63	36.5
OZ209899.1	12	17.41	36.5
OZ209900.1	13	17.01	37
OZ209901.1	14	16.99	37
OZ209902.1	15	16.9	36.5
OZ209903.1	16	16.47	37
OZ209904.1	17	16.46	37
OZ209905.1	18	16.45	37
OZ209906.1	19	15.64	37
OZ209907.1	20	14.61	37
OZ209908.1	21	13.74	37
OZ209909.1	22	13.01	37
OZ209910.1	23	12.96	37
OZ209911.1	24	12.16	37
OZ209912.1	25	11.35	37
OZ209913.1	26	9.73	38
OZ209914.1	27	8.42	38
OZ209915.1	28	8.25	38
OZ209916.1	Z	38.33	37
OZ209917.1	MT	0.02	20

Hi-C sample preparation and crosslinking

Hi-C data were generated from 20–50 mg of frozen tissue from the ilBatPrae2 sample using the Arima-HiC v2 kit (Arima Genomics). As per manufacturer's instructions, tissue was fixed, and the DNA crosslinked using a TC buffer with 22%

formaldehyde concentration, and a final formaldehyde concentration of 2%. The tissue was then homogenised using the Diagenode Power Masher-II. The crosslinked DNA was digested using a restriction enzyme master mix, then biotinylated and ligated. A clean up was performed with SPRIselect beads prior to library preparation. DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit, and sample biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core.

PacBio HiFi (ULI)

The sample requires Covaris g-TUBE shearing to approximately 10 kb prior to library preparation. Ultra-low input libraries were prepared using PacBio SMRTbell® Express Template Prep Kit 2.0 and PacBio SMRTbell® gDNA Sample Amplification Kit. To begin, samples were normalised to 20 ng of DNA. Initial removal of single-strand overhangs, DNA damage repair, and end repair/A-tailing were performed per manufacturer's instructions. From the SMRTbell® gDNA Sample Amplification Kit, amplification adapters were then ligated. A 0.85X pre-PCR clean-up was performed with Promega ProNex beads and the sample was then divided into two for a dual PCR. PCR reactions A and B each followed the PCR programs as described in the manufacturer's protocol. A 0.85X post-PCR clean-up was performed with ProNex beads for PCR reactions A and B and DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit and fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and gDNA 55kb BAC analysis kit. PCR reactions A and B were then pooled, ensuring the total mass was ≥ 500 ng in 47.4 μ L. The pooled sample then repeated the process for DNA damage repair, end repair/A-tailing and additional hairpin adapter ligation. A 1X clean-up was performed with ProNex beads and DNA concentration was quantified using the Qubit and fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies). Size selection was performed using Sage Sciences' PippinHT system with target fragment size determined by analysis from the Femto Pulse, usually a value between 4000 and 9000 bp. Size selected libraries were then cleaned-up using 1.0X ProNex beads and normalised to 2 nM before proceeding to sequencing.

Samples were sequenced on a Revio instrument (Pacific Biosciences, California, USA). Prepared libraries were normalised to 2 nM, and 15 μ L was used for making complexes. Primers were annealed and polymerases were bound to create circularised complexes according to manufacturer's instructions. The complexes were purified with the 1.2X clean up with SMRTbell beads. The purified complexes were then diluted to the Revio loading concentration (in the range 200–300 pM), and spiked with a Revio sequencing internal control. Samples were sequenced on Revio 25M SMRT cells (Pacific Biosciences,

California, USA). The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, as well as perform primary and secondary analysis of the data upon completion.

Hi-C

For Hi-C library preparation, the biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size-selected to 400–600 bp using SPRISelect beads. DNA was then enriched using Arima-HiC v2 Enrichment beads. The NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) was used for end repair, A-tailing, and adapter ligation, following a modified protocol in which library preparation is carried out while the DNA remains bound to the enrichment beads. PCR amplification was performed using KAPA HiFi HotStart mix and custom dual-indexed adapters (Integrated DNA Technologies) in a 96-well plate format. Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, samples were amplified for 10–16 PCR cycles. Post-PCR clean-up was carried out using SPRISelect beads. The libraries were quantified using the Accuclear Ultra High Sensitivity dsDNA Standards Assay kit (Biotium) and normalised to 10 ng/μL before sequencing. Hi-C sequencing was performed on the Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

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 first 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 using YaHS (Zhou *et al.*, 2023) using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. Flat files and maps used in curation were generated via the TreeVal pipeline (Pointon *et al.*, 2023). Manual curation was conducted primarily in PretextView (Harry, 2022) and HiGlass (Kerpedjiev *et al.*, 2018), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination,

missed joins, and mis-joins were amended, and duplicate sequences were tagged and removed. Sex chromosomes were identified by BUSCO gene painting with ancestral Merian elements. The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation>.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020), run in a Singularity container (Kurtzer *et al.*, 2017), was used to evaluate k -mer completeness and assembly quality for 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 in the blobtoolkit pipeline, a Nextflow (Di Tommaso *et al.*, 2017) port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads in SAMtools (Danecek *et al.*, 2021) and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoT database (Challis *et al.*, 2023) to identify all matching BUSCO lineages to run BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineages, the pipeline aligns the BUSCO genes to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) with DIAMOND blastp (Buchfink *et al.*, 2021). The genome is also divided into chunks according to the density of the BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database using DIAMOND blastx. Genome sequences without a hit are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The blobtools suite combines all these 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), relying on the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), as well as the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

Table 4 contains a list of relevant software tool versions and sources.

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 [here](https://www.darwintreeoflife.org/). 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

Table 4. Software tools: versions and sources.

Software tool	Version	Source
BLAST	2.14.0	http://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/-
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	666652151335353eef2fcd58880bcef5bc2928e1	https://github.com/thegenemyers/FASTK
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Goat CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhyllp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84aa44357826c0b6753eb28de	https://github.com/higlass/higlass
MerquryFK	d00d98157618f4e8d1a9190026b19b471055b22e	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
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.6.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

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: *Batrachedra praeangusta* (popular cosmet). Accession number PRJEB76773; <https://identifiers.org/ena.embl/PRJEB76773>. The genome sequence is released openly for reuse. The *Batrachedra praeangusta* 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. The genome will be annotated using available RNA-Seq data and presented through the [Ensembl](#) pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in [Table 1](#) and [Table 2](#).

Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: <https://doi.org/10.5281/zenodo.12157525>.

Members of the Natural History Museum Genome Acquisition Lab are listed here: <https://doi.org/10.5281/zenodo.12159242>.

Members of the Darwin Tree of Life Barcoding collective are listed here: <https://doi.org/10.5281/zenodo.12158331>.

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: <https://doi.org/10.5281/zenodo.12162482>.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: <https://doi.org/10.5281/zenodo.12165051>.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: <https://doi.org/10.5281/zenodo.12160324>.

Members of the Tree of Life Core Informatics collective are listed here: <https://doi.org/10.5281/zenodo.12205391>.

Members of the Darwin Tree of Life Consortium are listed here: <https://doi.org/10.5281/zenodo.4783558>.

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](#)
- Beasley J, Uhl R, Forrest LL, *et al.*: **DNA barcoding SOPs for the Darwin Tree of Life project.** *protocols.io.* 2023; [Accessed 25 June 2024]. [Publisher 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, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 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](#)
- da Veiga Leprevost F, Grünig BA, Alves Afritas S, *et al.*: **BioContainers: an open-source and community-driven framework for software standardization.** *Bioinformatics.* 2017; **33**(16): 2580–2582. [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): gja008. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Denton A, Oatley G, Cornwell C, *et al.*: **Sanger Tree of Life sample homogenisation: PowerMash.** *protocols.io.* 2023. [Publisher Full Text](#)
- Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319. [PubMed Abstract](#) | [Publisher Full Text](#)
- Diesch C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of syntenic and structural variation.** *Genome Biol.* 2023; **24**(1): 74. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Emmett AM, Langmaid JR: **The moths and butterflies of great Britain and Ireland. Gelechiidae.** Colchester: Harley Books, 2002. [Reference Source](#)
- 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](#)
- Grünig B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 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](#)
- Hall P, Plant CW: ***Batrachedra confusella* Berggren, Aarvik, Huemer, Lee & Mutanen, 2022 (Lep.: Batrachedridae) in Britain.** *Entomol Rec J Var.* 2002; **134**(2): 53–57. [Reference Source](#)
- Harry E: **PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps.** 2022. [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): gaa153. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life sample preparation: triage and dissection.** *protocols.io.* 2023. [Publisher 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, Pereira-da-Conceicao LL, Blaxter ML, *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](#)

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J*. 2014; **2014**(239): 2, [Accessed 2 April 2024].
[Reference Source](#)

Oatley G, Denton A, Howard C: **Sanger Tree of Life HMW DNA extraction: automated MagAttract v.2.** *protocols.io*. 2023a.
[Publisher Full Text](#)

Oatley G, Sampaio F, Howard C: **Sanger Tree of Life fragmented DNA clean up: automated SPRI.** *protocols.io*. 2023b.
[Publisher Full Text](#)

Oatley G, Sampaio F, Kitchin L, *et al.*: **Sanger Tree of Life HMW DNA fragmentation: Covaris g-TUBE for ULI PacBio.** *protocols.io*. 2023c; [Accessed 13 June 2024].
[Publisher Full Text](#)

Pereira L, Sivell O, Sivess L, *et al.*: **DTOL taxon-specific standard operating procedure for the terrestrial and freshwater arthropods working group.** 2022.
[Publisher Full Text](#)

Pointon DL, Eagles W, Sims Y, *et al.*: **sanger-tol/treeval v1.0.0 – Ancient Atlantis.** 2023.
[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.*: **Merquy: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol*. 2020; **21**(1): 245.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Sterling P, Parsons M, Lewington R: **Field guide to the micro moths of great Britain and Ireland.** Gillingham, Dorset: British Wildlife Publishing, 2012.
[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, Krashenninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics*. 2023; **24**(1): 288.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2019; 314–324.
[Publisher Full Text](#)

Wright CJ, Stevens L, Mackintosh A, *et al.*: **Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera.** *Nat Ecol Evol*. 2024; **8**(4): 777–790.
[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 28 August 2025

<https://doi.org/10.21956/wellcomeopenres.26808.r125487>

© 2025 Arumugaperumal A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Arun Arumugaperumal 

Rajalakshmi Engineering College, Chennai, Tamil Nadu, India

The data note reports the genome sequence of the poplar cosmet moth *Batrachedra praeangusta*. It is advisable that the second author can also have ORCID linked to his name like what has been done with the first author name. The assembly reported here is of size 486.93 Mb. The mitogenome was also provided separately. The specimen collected was a female. I suppose this belongs to ZZ/ZO chromosome system. Under the heading 'Sequencing data' change '121 coverage' to '121x coverage'. The organism photograph was clearly recorded in figure 1. The accession IDs reported were working fine. The data note can be indexed.

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 14 August 2025

<https://doi.org/10.21956/wellcomeopenres.26808.r125489>

© 2025 Kuppusamy S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Sivasankaran Kuppusamy** 

Loyola College, Chennai, Tamil Nadu, India

Authors have done the chromosome-level genome sequence of *Batrachedra praeangusta* (Haworth, 1828). They produced 201.52 Gb genome. They assembled and annotated using appropriate softwares. They detected 29 chromosomal pseudomolecules through the assembly.

Comments on the manuscript

Batrachedra praeangusta (Haworth, 1828) Poplar Cosmet is a moth belonging to the family Batrachedridae. There are about 140 species in the genus *Batrachedra* worldwide. Among them, three are in the UK (Hall & Plant, 2022), with most species in the tropics.

The genome of the *B. praeangusta* was sequenced as part of the Darwin Tree of Life Project...

The author used the full name of the genus, *Batrachedra praeangusta*, throughout the manuscript. If it's possible, begin with the full genus name and then abbreviate it to *B. praeangusta*.

Authors have sequenced 201.52 Gb genome size using different sequencing platforms. They assembled and annotated the genome sequence. But they haven't included the annotation list of the genomes like a number of protein-coding genes, non-coding genes and gene transcripts.

The research article was well prepared and the manuscript meets the necessary scientific standard and is 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: Phylogenomics

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.
