



## DATA NOTE

# The genome sequence of the Pale Tussock moth, *Calliteara pudibunda* (Linnaeus, 1758) (Lepidoptera: Erebidæ)

[version 1; peer review: awaiting peer review]

Laura Sivess <sup>1</sup>, Gavin R. Broad <sup>1</sup>, Stephanie Holt <sup>1</sup>, Douglas Boyes<sup>2</sup>,  
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

<sup>1</sup>Natural History Museum, London, England, UK

<sup>2</sup>UK Centre for Ecology & Hydrology, Wallingford, England, UK

**v1** First published: 21 Oct 2025, 10:587  
<https://doi.org/10.12688/wellcomeopenres.25048.1>  
Latest published: 21 Oct 2025, 10:587  
<https://doi.org/10.12688/wellcomeopenres.25048.1>

## Open Peer Review

**Approval Status** Awaiting Peer Review

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

## Abstract

We present a genome assembly from an individual male *Calliteara pudibunda* (Pale Tussock; Arthropoda; Insecta; Lepidoptera; Erebidæ). The genome sequence has a total length of 1 035.55 megabases. Most of the assembly (99.83%) is scaffolded into 88 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled, with a length of 16.72 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

*Calliteara pudibunda*; Pale Tussock; genome sequence; chromosomal; Lepidoptera



This article is included in the [Tree of Life](#) gateway.

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

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

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540) and the Darwin Tree of Life Discretionary Award [218328, <https://doi.org/10.35802/218328>]. *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2025 Sivess L *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:** Sivess L, Broad GR, Holt S *et al.* **The genome sequence of the Pale Tussock moth, *Calliteara pudibunda* (Linnaeus, 1758) (Lepidoptera: Erebidae) [version 1; peer review: awaiting peer review]** Wellcome Open Research 2025, **10**:587 <https://doi.org/10.12688/wellcomeopenres.25048.1>

**First published:** 21 Oct 2025, **10**:587 <https://doi.org/10.12688/wellcomeopenres.25048.1>

## 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; Noctuoidea; Erebidae; Lymantriinae; *Calliteara*; *Calliteara pudibunda* (Linnaeus, 1758) (NCBI:txid875926)

## Background

*Calliteara pudibunda*, commonly known as the Pale Tussock, belongs to the family Erebidae (subfamily Lymantriinae). It is widespread across England, Wales, and Northern Ireland, with a more limited distribution in Scotland (NBN Atlas Partnership, 2025). Native to the north-western Palearctic region, the species is broadly distributed throughout Europe (GBIF Secretariat, 2023), with sporadic occurrences outside its typical range, including georeferenced records from eastern Asia (GBIF Secretariat, 2023).

Adults fly in May and June and have a wingspan ranging from 40 to 60 mm. They are characterised by pale grey wings with brown markings. The species is mildly sexually dimorphic: males possess feathered, orange-brown antennae and generally exhibit darker and more defined wing patterns than the larger females. Both sexes have distinctive, furry forelegs that are forward facing when at rest (Kimber, 2025).

Larvae resemble those of the Dark Tussock (*Dicallomera fascelina*) but are distinguished by four tufts of yellow hairs and a body colour ranging from light yellow to pale green. After feeding from June to early October, the larvae overwinter as pupae. The species feeds on a variety of broadleaved trees and deciduous shrubs, including Barberry (*Berberis vulgaris*), Beech (*Fagus* spp.), Blackthorn (*Prunus spinosa*), Oak (*Quercus* spp.), and Hazel (*Corylus avellana*). Consequently, its habitats are diverse and include gardens, hedgerows, woodlands, and scrublands. Larvae fluoresce under UV light and can be found that way on foliage at night.

Historically, Pale Tussock larvae were considered pests of cultivated hop (*Humulus lupulus*), a plant introduced to Britain by the early 16th century and extensively grown for beer production; indeed, an old vernacular name in England was ‘Hop Dog’. At its peak in the 1870s, hop cultivation covered over 70 000 acres of arable land (Cornell, 2009). Today, the species is known to occasionally cause defoliation in beech forests in Europe, with outbreaks typically occurring at intervals of 20 to 30 years. These outbreaks typically persist for 2 to 3 years before collapsing, likely due to larval mortality induced by pathogens (Mazzoglio *et al.*, 2005). Notably, in 2018 and 2019, an outbreak in Bursa Province, Turkey, resulted in widespread defoliation of beech forests, marking the southernmost recorded pest outbreak for this species. Warmer winters have been associated with increased outbreak risk, likely due to decreased pupal mortality and accelerated population growth (Ipekdal, 2022).

Strategic forest management can mitigate the pest potential of *C. pudibunda*. In particular, cultivating mixed forest stands significantly reduces outbreak likelihood, with studies showing relative abundance of this species declining as the spruce-to-beech ratio increases (Heiermann & Schütz, 2008).

A genome sequence of *C. pudibunda* was produced as part of the Darwin Tree of Life project. The specimen used for sequencing (Figure 1) was collected from Gilbert White’s House, Selborne, United Kingdom, during a targeted field sampling effort. This assembly represents the first high-quality genome for the genus *Calliteara* as of September 2025 (data obtained via NCBI datasets, O’Leary *et al.*, 2024). This genomic resource will facilitate functional and comparative genomic studies in areas such as insect neurobiology, visual system development, and host plant specialisation. Furthermore, it contributes to the growing body of genomic data available for Lepidoptera, supporting investigations into molecular evolution and adaptation across this diverse order.

## Methods

### Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Calliteara pudibunda* (specimen ID NHMUK014536854, ToLID ilCalPudi2; Figure 1), collected from Selborne, Gilbert



**Figure 1.** Photograph of the *Calliteara pudibunda* (ilCalPudi2) specimen used for genome sequencing.

White's House, England, UK (latitude 51.09, longitude -0.94) on 2021-06-10. The specimen was collected by Gavin Broad, Stephanie West and Laura Sivess and formally identified by Gavin Broad. A second specimen was used for Hi-C sequencing (specimen ID Ox000406, ToLID ilCalPudi1). It was collected from Wytham Woods, Oxfordshire, UK (latitude 51.771, longitude -1.337) on 2020-05-22. The specimen was collected and identified by Douglas Boyes. Sample metadata were collected in line with the Darwin Tree of Life project standards described by [Lawniczak et al. \(2022\)](#).

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

#### Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on [protocols.io](#) ([Howard et al., 2025](#)). The ilCalPudi2 sample was weighed and [triaged](#) to determine the appropriate extraction protocol. Tissue from the head was homogenised by [powermashing](#) using a PowerMasher II tissue disruptor.

HMW DNA was extracted in the WSI Scientific Operations core using the [Automated MagAttract v2](#) protocol. DNA was sheared into an average fragment size of 12–20 kb following the [Megaruptor®3 for LI PacBio](#) protocol. Sheared DNA was purified by [manual SPRI](#) (solid-phase reversible immobilisation). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 9.12 ng/μL and a yield of 410.40 ng, with a fragment size of 12.6 kb. The 260/280 spectrophotometric ratio was 1.7, and the 260/230 ratio was 1.03.

RNA was extracted from thorax tissue of ilCalPudi2 in the Tree of Life Laboratory at the WSI using the [RNA Extraction: Automated MagMax™ mirVana protocol](#). The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

#### PacBio HiFi library preparation and sequencing

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

following the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

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

#### Hi-C

##### *Sample preparation and crosslinking*

The Hi-C sample was prepared from 20–50 mg of frozen head tissue from the ilCalPudi1 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 Diagenode Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

##### *Hi-C library preparation and sequencing*

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

Prior to sequencing, libraries were normalised to 10 ng/μL. Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High

Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq 6000.

### RNA library preparation and sequencing

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

### Genome assembly

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

The HiFi reads were assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. 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 10 breaks and 17 joins. The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation>. PretextViewSnapshot 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 *Callithea pudibunda* specimen generated 84.06 Gb (gigabases) from 9.24 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 1005.48 Mb, with a heterozygosity of 0.77% and repeat content of 32.36% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 81× coverage. Hi-C sequencing produced 128.40 Gb from 850.35 million reads, which were used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. Table 1 summarises the specimen and sequencing details.

### Assembly statistics

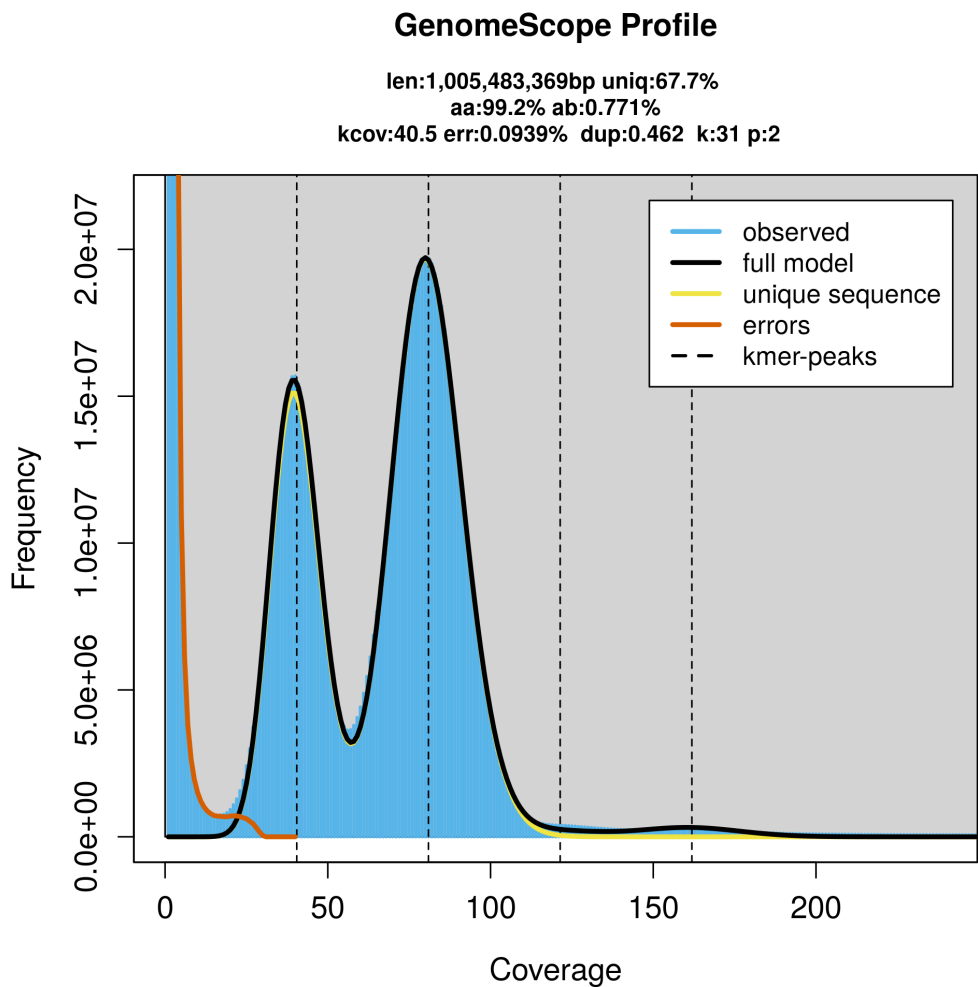
The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The final assembly has a total length of 1 035.55 Mb in 110 scaffolds, with 356 gaps, and a scaffold N50 of 10.85 Mb (Table 2).

Most of the assembly sequence (99.83%) was assigned to 88 chromosomal-level scaffolds, representing 87 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). The Z chromosome was 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.

The combined primary and alternate assemblies achieve an estimated QV of 54.0. The  $k$ -mer completeness is 83.35% for the primary assembly, 84.44% for the alternate haplotype, and 99.46% for the combined assemblies (Figure 4).

BUSCO v.5.8.3 analysis using the lepidoptera\_odb10 reference set ( $n = 5\ 286$ ) identified 99.2% of the expected gene set (single = 98.2%, duplicated = 1.0%). The snail plot in Figure 5 summarises the scaffold length distribution and other



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

**Table 1.** Specimen and sequencing data for BioProject PRJEB76359.

| Platform                      | PacBio HiFi     | Hi-C                  | RNA-seq            |
|-------------------------------|-----------------|-----------------------|--------------------|
| ToLID                         | ilCalPudi2      | ilCalPudi1            | ilCalPudi2         |
| Specimen ID                   | NH Muk014536854 | Ox000406              | NH Muk014536854    |
| BioSample (source individual) | SAMEA111457921  | SAMEA7520528          | SAMEA111457921     |
| BioSample (tissue)            | SAMEA111457991  | SAMEA7520623          | SAMEA111458076     |
| Tissue                        | head            | head                  | thorax             |
| Instrument                    | Revio           | Illumina NovaSeq 6000 | Illumina NovaSeq X |
| Run accessions                | ERR13245287     | ERR13248957           | ERR14792833        |
| Read count total              | 9.24 million    | 850.35 million        | 99.54 million      |
| Base count total              | 84.06 Gb        | 128.40 Gb             | 15.03 Gb           |

Table 2. Genome assembly statistics.

|                               |                         |
|-------------------------------|-------------------------|
| Assembly name                 | ilCalPudi2.1            |
| Assembly accession            | GCA_965615895.1         |
| Alternate haplotype accession | GCA_965615845.1         |
| Assembly level                | chromosome              |
| Span (Mb)                     | 1 035.55                |
| Number of chromosomes         | 88                      |
| Number of contigs             | 466                     |
| Contig N50                    | 3.93 Mb                 |
| Number of scaffolds           | 110                     |
| Scaffold N50                  | 10.85 Mb                |
| Sex chromosomes               | Z                       |
| Organelles                    | Mitochondrion: 16.72 kb |

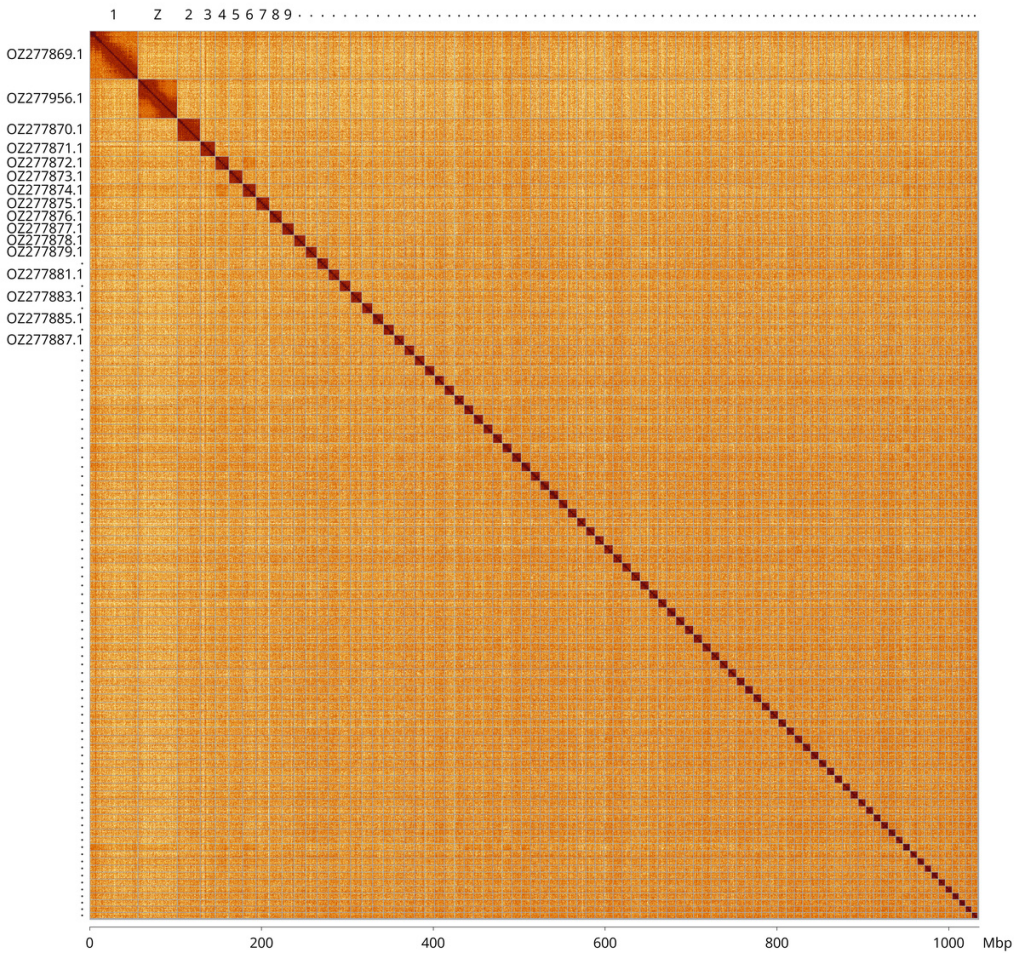


Figure 3. Hi-C contact map of the *Calliteara pudibunda* genome assembly. Assembled chromosomes are shown in order of decreasing size and labelled along the axes, with a megabase scale shown below. For legibility, small chromosomes are labelled with a dot. The plot was generated using PretextSnapshot.

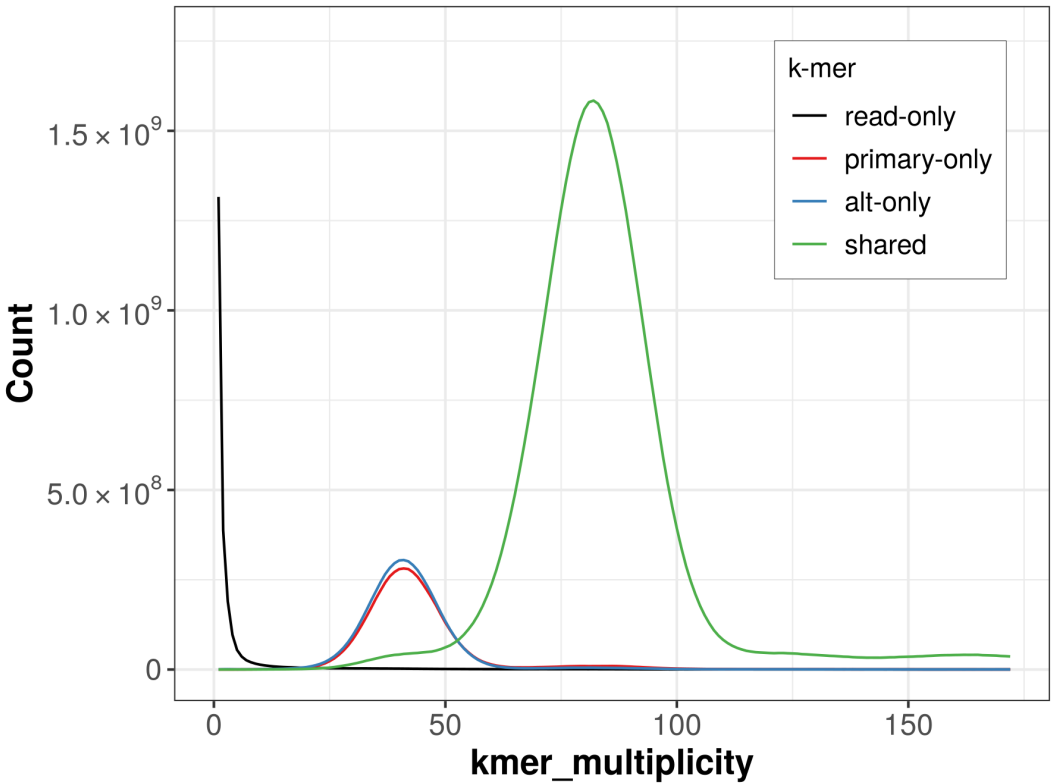
**Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Calliteara pudibunda* iCalPudi2.**

| INSDC accession | Molecule | Length (Mb) | GC%   |
|-----------------|----------|-------------|-------|
| OZ277869.1      | 1        | 56.33       | 36    |
| OZ277870.1      | 2        | 26.63       | 36    |
| OZ277871.1      | 3        | 17.54       | 35.50 |
| OZ277872.1      | 4        | 15.93       | 36    |
| OZ277873.1      | 5        | 15.80       | 36    |
| OZ277874.1      | 6        | 15.77       | 36.50 |
| OZ277875.1      | 7        | 15.44       | 36.50 |
| OZ277876.1      | 8        | 14.53       | 35.50 |
| OZ277877.1      | 9        | 14.10       | 36.50 |
| OZ277878.1      | 10       | 13.42       | 36    |
| OZ277879.1      | 11       | 13.33       | 36    |
| OZ277880.1      | 12       | 13.12       | 35    |
| OZ277881.1      | 13       | 13.07       | 36    |
| OZ277882.1      | 14       | 13.01       | 36    |
| OZ277883.1      | 15       | 12.97       | 36    |
| OZ277884.1      | 16       | 12.67       | 35.50 |
| OZ277885.1      | 17       | 12.57       | 36    |
| OZ277886.1      | 18       | 12.32       | 36    |
| OZ277887.1      | 19       | 11.98       | 35.50 |
| OZ277888.1      | 20       | 11.91       | 35.50 |
| OZ277889.1      | 21       | 11.74       | 36    |
| OZ277890.1      | 22       | 11.63       | 36    |
| OZ277891.1      | 23       | 11.42       | 36    |
| OZ277892.1      | 24       | 11.38       | 36    |
| OZ277893.1      | 25       | 11.32       | 36    |
| OZ277894.1      | 26       | 11.30       | 36.50 |
| OZ277895.1      | 27       | 11.22       | 35.50 |
| OZ277896.1      | 28       | 11.11       | 35.50 |
| OZ277897.1      | 29       | 11.10       | 36    |
| OZ277898.1      | 30       | 11.06       | 36    |
| OZ277899.1      | 31       | 10.95       | 35.50 |
| OZ277900.1      | 32       | 10.91       | 37    |
| OZ277901.1      | 33       | 10.85       | 36    |
| OZ277902.1      | 34       | 10.81       | 36    |
| OZ277903.1      | 35       | 10.79       | 36.50 |
| OZ277904.1      | 36       | 10.64       | 36    |
| OZ277905.1      | 37       | 10.64       | 35.50 |
| OZ277906.1      | 38       | 10.60       | 35    |
| OZ277907.1      | 39       | 10.55       | 36.50 |

| INSDC accession | Molecule | Length (Mb) | GC%   |
|-----------------|----------|-------------|-------|
| OZ277908.1      | 40       | 10.52       | 35.50 |
| OZ277909.1      | 41       | 10.50       | 35    |
| OZ277910.1      | 42       | 10.49       | 35.50 |
| OZ277911.1      | 43       | 10.47       | 35.50 |
| OZ277912.1      | 44       | 10.47       | 35.50 |
| OZ277913.1      | 45       | 10.43       | 36    |
| OZ277914.1      | 46       | 10.41       | 36.50 |
| OZ277915.1      | 47       | 10.36       | 35.50 |
| OZ277916.1      | 48       | 10.35       | 36    |
| OZ277917.1      | 49       | 10.27       | 36.50 |
| OZ277918.1      | 50       | 10.24       | 35.50 |
| OZ277919.1      | 51       | 10.23       | 36    |
| OZ277920.1      | 52       | 10.19       | 35.50 |
| OZ277921.1      | 53       | 9.99        | 36    |
| OZ277922.1      | 54       | 9.92        | 35.50 |
| OZ277923.1      | 55       | 9.84        | 35.50 |
| OZ277924.1      | 56       | 9.80        | 35.50 |
| OZ277925.1      | 57       | 9.74        | 35.50 |
| OZ277926.1      | 58       | 9.67        | 36    |
| OZ277927.1      | 59       | 9.67        | 35.50 |
| OZ277928.1      | 60       | 9.61        | 36.50 |
| OZ277929.1      | 61       | 9.55        | 35.50 |
| OZ277930.1      | 62       | 9.55        | 36    |
| OZ277931.1      | 63       | 9.41        | 36.50 |
| OZ277932.1      | 64       | 9.38        | 36    |
| OZ277933.1      | 65       | 9.35        | 36    |
| OZ277934.1      | 66       | 9.33        | 35.50 |
| OZ277935.1      | 67       | 9.15        | 36    |
| OZ277936.1      | 68       | 9.14        | 36    |
| OZ277937.1      | 69       | 9.09        | 36.50 |
| OZ277938.1      | 70       | 9.08        | 35.50 |
| OZ277939.1      | 71       | 8.92        | 36.50 |
| OZ277940.1      | 72       | 8.89        | 36.50 |
| OZ277941.1      | 73       | 8.82        | 36    |
| OZ277942.1      | 74       | 8.55        | 36    |
| OZ277943.1      | 75       | 8.50        | 36.50 |
| OZ277944.1      | 76       | 8.48        | 36.50 |
| OZ277945.1      | 77       | 8.42        | 36.50 |
| OZ277946.1      | 78       | 8.31        | 36.50 |
| OZ277947.1      | 79       | 8.25        | 36    |
| OZ277948.1      | 80       | 8.16        | 36.50 |

| INSDC accession | Molecule | Length (Mb) | GC%   |
|-----------------|----------|-------------|-------|
| OZ277949.1      | 81       | 8.13        | 36    |
| OZ277950.1      | 82       | 8.09        | 37    |
| OZ277951.1      | 83       | 8.06        | 36.50 |
| OZ277952.1      | 84       | 7.80        | 36.50 |

| INSDC accession | Molecule | Length (Mb) | GC%   |
|-----------------|----------|-------------|-------|
| OZ277953.1      | 85       | 7.55        | 36    |
| OZ277954.1      | 86       | 7.34        | 36.50 |
| OZ277955.1      | 87       | 6.94        | 37    |
| OZ277956.1      | Z        | 45.89       | 35.50 |

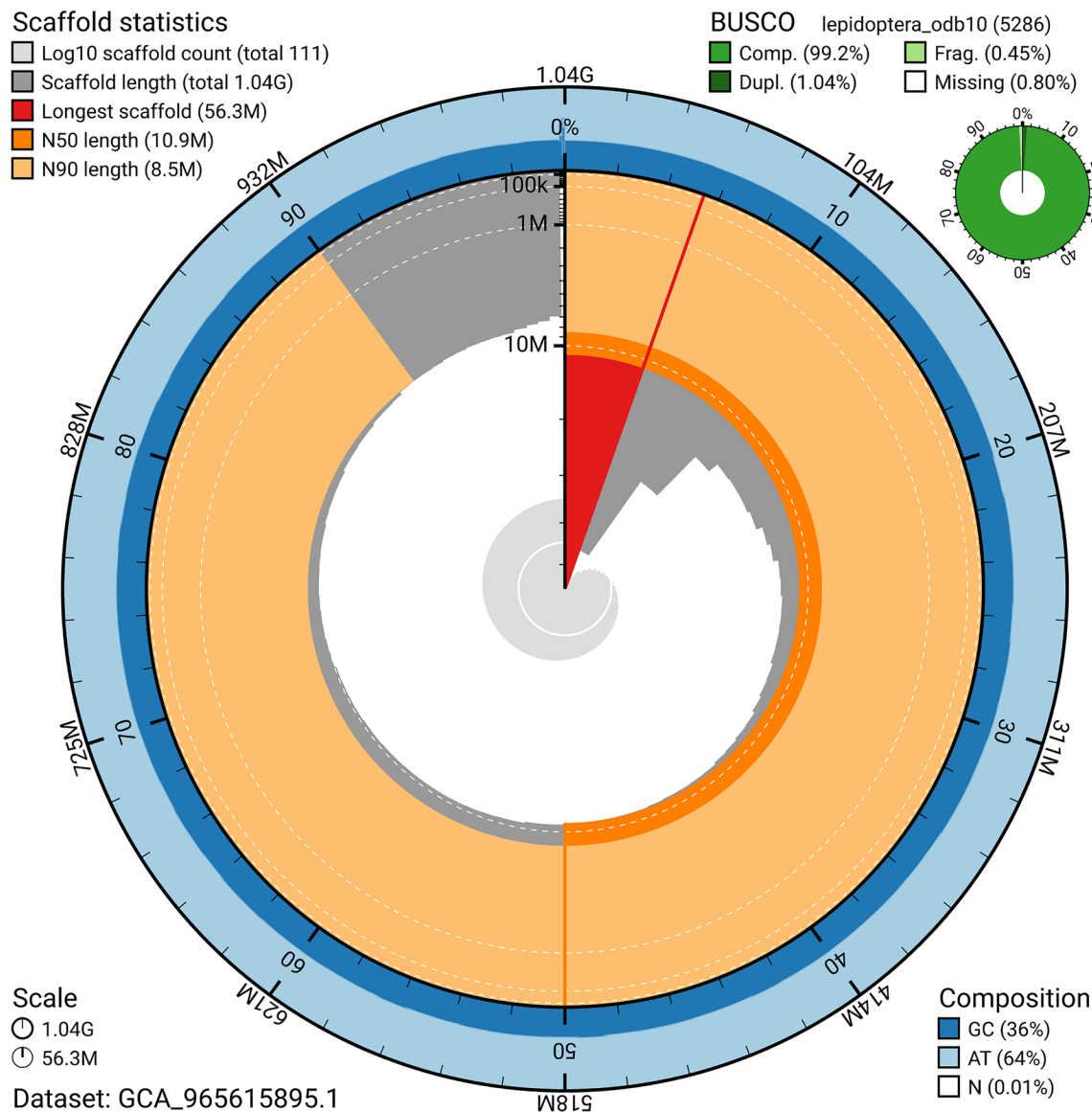


**Figure 4. Evaluation of *k*-mer completeness using MerquryFK.** This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

assembly statistics for the primary assembly. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the primary assembly, is **6.C.Q54**, meeting the recommended reference standard.

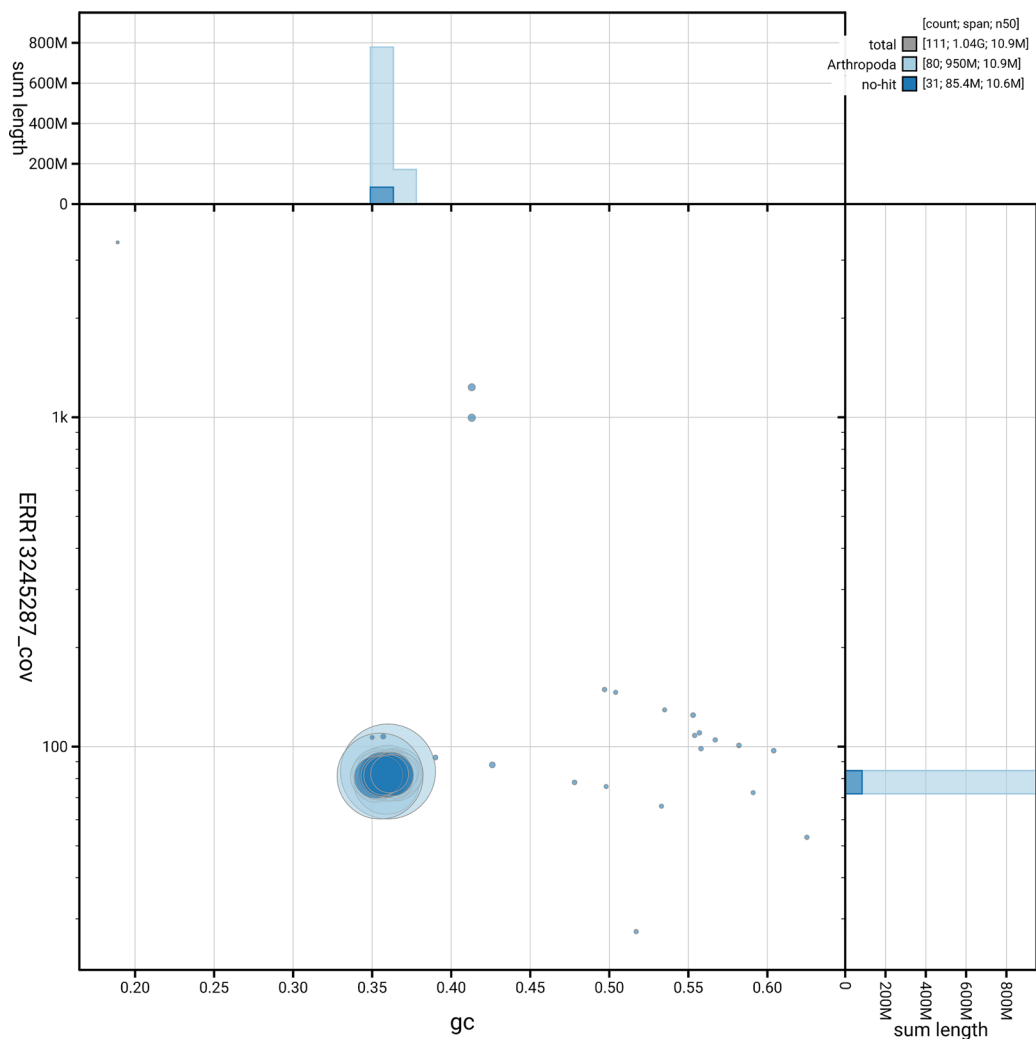
**Wellcome Sanger Institute – Legal and Governance**  
The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the ‘**Darwin Tree of Life Project Sampling Code of Practice**’, 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



**Figure 5. Assembly metrics for iCalPudi2.1.** 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](#).

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



**Figure 6. BlobToolKit GC-coverage plot for ilCalPudi2.1.** 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 *Calliteara pudibunda* assembly.**

| Measure                                        | Value                                                | Benchmark        |
|------------------------------------------------|------------------------------------------------------|------------------|
| EBP summary (primary)                          | 6.C.Q54                                              | 6.C.Q40          |
| Contig N50 length                              | 3.93 Mb                                              | ≥ 1 Mb           |
| Scaffold N50 length                            | 10.85 Mb                                             | = chromosome N50 |
| Consensus quality (QV)                         | Primary: 54.4; alternate: 53.9; combined: 54.0       | ≥ 40             |
| k-mer completeness                             | Primary: 83.35%; alternate: 84.44%; combined: 99.46% | ≥ 95%            |
| BUSCO                                          | C:99.2% [S:98.2%; D:1.0%]; F:0.5%; M:0.3%; n:5 286   | S > 90%; D < 5%  |
| Percentage of assembly assigned to chromosomes | 99.83%                                               | ≥ 90%            |

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: *Callithea pudibunda* (pale tussock). Accession number [PRJEB76359](#). The genome sequence is released openly for reuse. The *Callithea pudibunda* 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](#).

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

Author information

Contributors are listed at the following links:

- Members of the [University of Oxford and Wytham Woods Genome Acquisition Lab](#)
- Members of the [Natural History Museum Genome Acquisition Lab](#)
- Members of the [Darwin Tree of Life Barcoding collective](#)
- Members of the [Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team](#)
- Members of [Wellcome Sanger Institute Scientific Operations – Sequencing Operations](#)
- Members of the [Wellcome Sanger Institute Tree of Life Core Informatics team](#)
- Members of the [Tree of Life Core Informatics collective](#)
- Members of the [Darwin Tree of Life Consortium](#)

Table 5. Software versions and sources.

| Software       | Version     | Source                                                                                                                  |
|----------------|-------------|-------------------------------------------------------------------------------------------------------------------------|
| BEDTools       | 2.30.0      | <a href="https://github.com/arq5x/bedtools2">https://github.com/arq5x/bedtools2</a>                                     |
| BLAST          | 2.14.0      | <a href="ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/">ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/</a> |
| BlobToolKit    | 4.4.6       | <a href="https://github.com/blobtoolkit/blobtoolkit">https://github.com/blobtoolkit/blobtoolkit</a>                     |
| BUSCO          | 5.8.3       | <a href="https://gitlab.com/ezlab/busco">https://gitlab.com/ezlab/busco</a>                                             |
| bwa-mem2       | 2.2.1       | <a href="https://github.com/bwa-mem2/bwa-mem2">https://github.com/bwa-mem2/bwa-mem2</a>                                 |
| Cooler         | 0.8.11      | <a href="https://github.com/open2c/cooler">https://github.com/open2c/cooler</a>                                         |
| DIAMOND        | 2.1.8       | <a href="https://github.com/bbuchfink/diamond">https://github.com/bbuchfink/diamond</a>                                 |
| fasta_windows  | 0.2.4       | <a href="https://github.com/tolkit/fasta_windows">https://github.com/tolkit/fasta_windows</a>                           |
| FastK          | 1.1         | <a href="https://github.com/thegenemyers/FASTK">https://github.com/thegenemyers/FASTK</a>                               |
| GenomeScope2.0 | 2.0.1       | <a href="https://github.com/tbenavi1/genomescope2.0">https://github.com/tbenavi1/genomescope2.0</a>                     |
| Gfastats       | 1.3.6       | <a href="https://github.com/vgl-hub/gfastats">https://github.com/vgl-hub/gfastats</a>                                   |
| GoaT CLI       | 0.2.5       | <a href="https://github.com/genomehubs/goat-cli">https://github.com/genomehubs/goat-cli</a>                             |
| Hifiasm        | 0.19.8-r603 | <a href="https://github.com/chhy123/hifiasm">https://github.com/chhy123/hifiasm</a>                                     |
| HiGlass        | 1.13.4      | <a href="https://github.com/higlass/higlass">https://github.com/higlass/higlass</a>                                     |
| MercuryFK      | 1.1.2       | <a href="https://github.com/thegenemyers/MERQURY.FK">https://github.com/thegenemyers/MERQURY.FK</a>                     |
| Minimap2       | 2.28-r1209  | <a href="https://github.com/lh3/minimap2">https://github.com/lh3/minimap2</a>                                           |

| Software                   | Version             | Source                                                                                                    |
|----------------------------|---------------------|-----------------------------------------------------------------------------------------------------------|
| MitoHiFi                   | 3                   | <a href="https://github.com/marcelauliano/MitoHiFi">https://github.com/marcelauliano/MitoHiFi</a>         |
| MultiQC                    | 1.14; 1.17 and 1.18 | <a href="https://github.com/MultiQC/MultiQC">https://github.com/MultiQC/MultiQC</a>                       |
| Nextflow                   | 24.10.4             | <a href="https://github.com/nextflow-io/nextflow">https://github.com/nextflow-io/nextflow</a>             |
| PretextSnapshot            | -                   | <a href="https://github.com/sanger-tol/PretextSnapshot">https://github.com/sanger-tol/PretextSnapshot</a> |
| PretextView                | 0.2.5               | <a href="https://github.com/sanger-tol/PretextView">https://github.com/sanger-tol/PretextView</a>         |
| samtools                   | 1.21                | <a href="https://github.com/samtools/samtools">https://github.com/samtools/samtools</a>                   |
| sanger-tol/ascc            | 0.1.0               | <a href="https://github.com/sanger-tol/ascc">https://github.com/sanger-tol/ascc</a>                       |
| sanger-tol/blobtoolkit     | v0.8.0              | <a href="https://github.com/sanger-tol/blobtoolkit">https://github.com/sanger-tol/blobtoolkit</a>         |
| sanger-tol/curationpretext | 1.4.2               | <a href="https://github.com/sanger-tol/curationpretext">https://github.com/sanger-tol/curationpretext</a> |
| Seqtk                      | 1.3                 | <a href="https://github.com/lh3/seqtk">https://github.com/lh3/seqtk</a>                                   |
| Singularity                | 3.9.0               | <a href="https://github.com/sylabs/singularity">https://github.com/sylabs/singularity</a>                 |
| TreeVal                    | 1.4.0               | <a href="https://github.com/sanger-tol/treeval">https://github.com/sanger-tol/treeval</a>                 |
| YaHS                       | 1.2a.2              | <a href="https://github.com/c-zhou/yahs">https://github.com/c-zhou/yahs</a>                               |

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](#)

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](#)

Cornell M: **A short history of hops.** 2009. [Reference Source](#)

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial arthropod fauna [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 123. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278. [PubMed Abstract](#) | [Publisher Full Text](#)

Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

GBIF Secretariat: ***Callitarea pudibunda* (Linnaeus, 1758) in GBIF backbone taxonomy.** 2023. [Reference Source](#)

Heiermann J, Schütz S: **The effect of the tree species ratio of European beech (*Fagus sylvatica* L.) and Norway spruce (*Picea abies* (L.) Karst.) on polyphagous and monophagous pest species – *Lymantria monacha* L. and *Callitarea pudibunda* L. (Lepidoptera: Lymantriidae) as an example.** *For Ecol Manage.* 2008; **255**(3–4): 1161–66. [Publisher Full Text](#)

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

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): giab153. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Ipekda K: **Türkiye’de yaşanan kızıl kuyruklu kayın tırtılı epidemisine kuramsal bir yaklaşım.** *Turkish Journal of Forestry Türkiye Ormançılık Dergisi.* 2022; **23**(3): 212–17. [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](#)

Kimber I: **Pale Tussock *Callitarea pudibunda* (Linnaeus, 1758) in UKMoths.** 2025. [Reference Source](#)

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, Seppely 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](#)

Mazzoglio PJ, Paoletta M, Patetta A, *et al.*: ***Calliteara pudibunda* (Lepidoptera: Lymantriidae) in Northwest Italy.** *Bull Insectology.* 2005; **58**(1): 25–34.

[Reference Source](#)

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

[Reference Source](#)

NBN Atlas Partnership: ***Calliteara pudibunda*: Pale Tussock** | NBN Atlas. 2025.

[Reference Source](#)

O'Leary NA, Cox E, Holmes JB, *et al.*: **Exploring and retrieving sequence and metadata for species across the Tree of Life with NCBI datasets.** *Sci Data.* 2024; **11**(1): 732.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free 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](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Uliano-Silva M, Ferreira JGRN, 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): 777–90.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)