DATA NOTE

The genome sequence of the Nut-tree Tussock moth, *Colocasia coryli* (Linnaeus, 1758) (Lepidoptera: Noctuidae)

[version 1; peer review: 2 approved]

Liam M. Crowley[1], Finley Hutchinson[2], Clare Boyes[3],
University of Oxford and Wytham Woods Genome Acquisition Lab,
Darwin Tree of Life Barcoding Collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]University of Oxford, Oxford, England, UK
[2]University of Exeter, Penryn, England, UK
[3]Independent researcher, Welshpool, Wales, UK

v1 **First published:** 23 Dec 2025, **10**:695
https://doi.org/10.12688/wellcomeopenres.25394.1
Latest published: 23 Dec 2025, **10**:695
https://doi.org/10.12688/wellcomeopenres.25394.1

Abstract

We present a genome assembly from an individual male *Colocasia coryli* (Nut-tree Tussock; Arthropoda; Insecta; Lepidoptera; Noctuidae). The genome sequence has a total length of 768.61 megabases. Most of the assembly (99.73%) is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled, with a length of 15.31 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

Colocasia coryli; Nut-tree Tussock; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 23 Dec 2025	✓ view	✓ view

1. **David EK Ferrier**, University of St Andrews, Scotland, UK
2. **Yash Gupta**, Naresuan University, Phitsanulok, Thailand

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

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

Author roles: Crowley LM: Investigation, Resources; **Hutchinson F**: Investigation, Resources; **Boyes C**: Writing – Original Draft Preparation;

Competing interests: No competing interests were disclosed.

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

How to cite this article: Crowley LM, Hutchinson F, Boyes C *et al.* **The genome sequence of the Nut-tree Tussock moth, *Colocasia coryli* (Linnaeus, 1758) (Lepidoptera: Noctuidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, **10**:695 https://doi.org/10.12688/wellcomeopenres.25394.1

First published: 23 Dec 2025, **10**:695 https://doi.org/10.12688/wellcomeopenres.25394.1

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; Noctuidae; Pantheinae; *Colocasia*; *Colocasia coryli* (Linnaeus, 1758) (NCBI:txid987903)

Background

Colocasia coryli, or the Nut-tree tussock (Figure 1), is a moth in the family Noctuidae. In Britain and Ireland it is common in southern England and relatively common in Scotland and Ireland; however it occurs locally in the north of England. It has increased in abundance and range in recent years (Randle *et al.*, 2019). It occurs throughout Europe, although it is less frequent in southern Europe with records extending to Russia (GBIF Secretariat, 2025).

The moth occurs in broad-leaved woodland where many deciduous tree species are used as larval foodplants. The adult moth (forewing length 14–17 mm) is usually grey-brown in the basal half and lighter grey or brown in the outer half, although there is variation in parts of the range. The male has feathery antennae. It has two generations each year, flying in April and May, and again in July and August (Waring *et al.*, 2017). The egg is laid singly on a leaf and hatches in four weeks (Heath & Emmet, 1983). The larvae are variable in colour but are easily recognisable by the long, forward-facing hair tufts on the side of the second segment, which are orangey brown or black; as well as orange tufts along the dorsum of segments one, two and eight (Henwood *et al.*, 2020).

The genome of *Colocasia coryli* 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. Here we present a chromosomally complete genome sequence *Colocasia coryli* based on one specimen from Wytham Woods, Oxfordshire, UK.

Figure 1. Photograph of *Colocasia coryli* (Photograph by Janet Graham.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Colocasia coryli* (specimen ID Ox003070, ToLID ilColCory3), collected from Wytham Woods, Oxfordshire, UK (latitude 51.772, longitude –1.338) on 2022-07-22. The specimen was collected by Finley Hutchinson and Liam Crowley, and formally identified by Finley Hutchinson. A second specimen was used for Hi-C sequencing (specimen ID SAN00002555, ToLID ilColCory2). It was collected from Carrifran Wildwood, Moffat Hills, Southern Uplands, Scotland, UK (latitude 55.3911, longitude –3.3279) on 2022-05-22. The specimen was collected and identified by Dorothy Lyle. The same specimen was used for RNA sequencing.

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

Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The ilColCory3 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the head and thorax was homogenised by powermashing using a PowerMasher II tissue disruptor. HMW DNA was extracted using the Automated MagAttract v2 protocol. DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor®3 for LI PacBio protocol. Sheared DNA was purified by automated SPRI (solid-phase reversible immobilisation). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 9.16 ng/μL and a yield of 3 572.40 ng.

RNA was extracted from abdomen tissue of ilColCory2 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol. The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), following the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

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

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the head and thorax of the ilColCory2 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

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

Prior to sequencing, libraries were normalised to 10 ng/µL. Normalised libraries were quantified again to create equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq 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 6000, generating paired-end reads.

Genome assembly

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

The HiFi reads were assembled using Hifiasm (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).

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 18 breaks and 21 joins. This reduced the scaffold count by 1.6%, increased the scaffold N50 by 3.0%, and reduced the total assembly length by 0.8%. The curation process is described at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020) was run in a Singularity container (Kurtzer *et al.*, 2017) to evaluate *k*-mer completeness and assembly quality for 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 *Colocasia coryli* specimen generated 30.78 Gb (gigabases) from 3.31 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 769.36 Mb, with a heterozygosity of 0.83% and repeat content of 31.56% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 39× coverage. Hi-C sequencing produced 160.86 Gb from 1 065.32 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 768.61 Mb in 62 scaffolds, with 19 gaps, and a scaffold N50 of 26.69 Mb (Table 2).

Most of the assembly sequence (99.73%) was assigned to 31 chromosomal-level scaffolds, representing 30 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). The Z chromosome was assigned by BUSCO gene painting with ancestral Merian elements (Wright *et al.*, 2024).

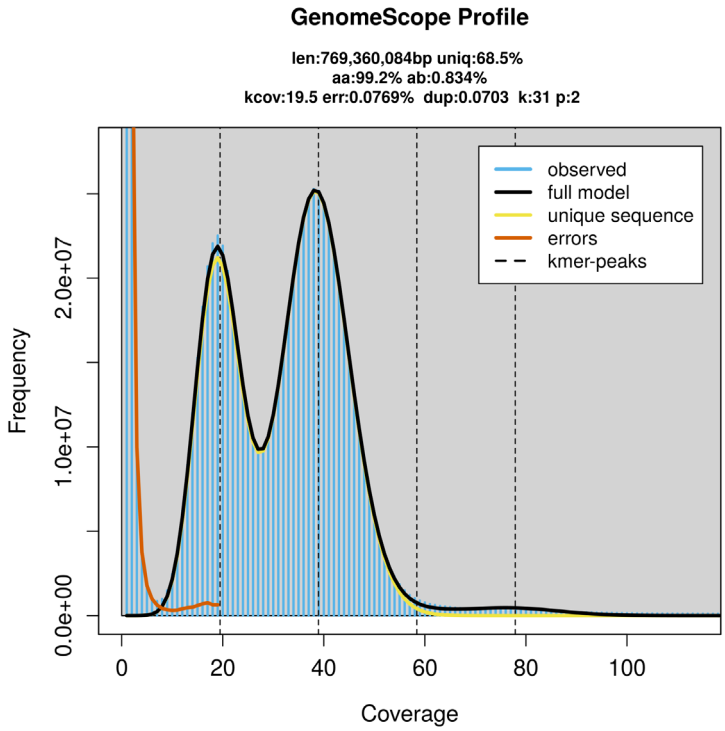

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. **Specimen and sequencing data for BioProject PRJEB89322.**

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	ilColCory3	ilColCory2	ilColCory2
Specimen ID	Ox003070	SAN00002555	SAN00002555
BioSample (source individual)	SAMEA112775018	SAMEA112198366	SAMEA112198366
BioSample (tissue)	SAMEA112775081	SAMEA112198398	SAMEA112198399
Tissue	head and thorax	head and thorax	abdomen
Instrument	Revio	Illumina NovaSeq 6000	Illumina NovaSeq 6000
Run accessions	ERR14955677	ERR14986753; ERR14986752	ERR14986754
Read count total	3.31 million	1 065.32 million	64.30 million
Base count total	30.78 Gb	160.86 Gb	9.71 Gb

Table 2. **Genome assembly statistics.**

Assembly name	ilColCory3.1
Assembly accession	GCA_965285835.1
Alternate haplotype accession	GCA_965285995.1
Assembly level	chromosome
Span (Mb)	768.61
Number of chromosomes	31
Number of contigs	81
Contig N50	24.07 Mb
Number of scaffolds	62
Scaffold N50	26.69 Mb
Sex chromosomes	Z
Organelles	Mitochondrion: 15.31 kb

The mitochondrial genome was also assembled (length 15.31 kb, OZ260524.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

The combined primary and alternate assemblies achieve an estimated QV of 66.5. The *k*-mer completeness is 81.42% for the primary assembly, 79.53% for the alternate haplotype, and 99.70% for the combined assemblies (Figure 4).

BUSCO v.5.7.1 analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.9% of the expected gene set (single = 98.3%, duplicated = 0.6%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for the primary assembly. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage.

Table 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 **7.C.Q68**, 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

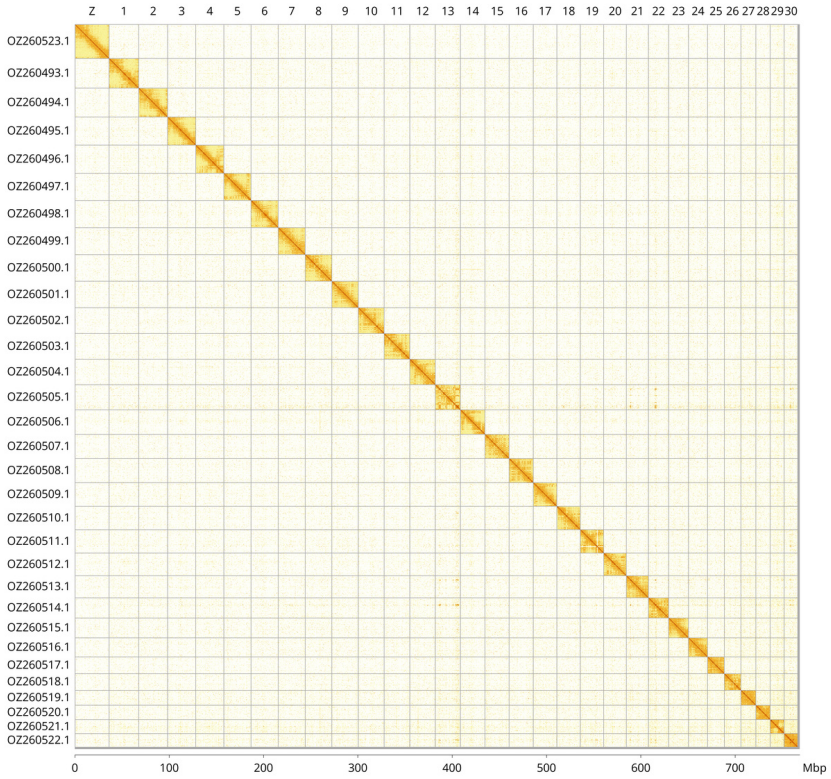

Figure 3. Hi-C contact map of the *Colocasia coryli* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Colocasia coryli* ilColCory3.

INSDC accession	Molecule	Length (Mb)	GC%
OZ260493.1	1	31.51	38.50
OZ260494.1	2	30.45	38
OZ260495.1	3	29.95	38.50
OZ260496.1	4	29.77	38.50
OZ260497.1	5	28.83	38.50
OZ260498.1	6	28.76	38.50
OZ260499.1	7	28.65	38.50
OZ260500.1	8	28.36	38.50
OZ260501.1	9	27.92	38.50
OZ260502.1	10	27.56	38.50
OZ260503.1	11	27.06	38.50
OZ260504.1	12	27.02	38.50
OZ260505.1	13	26.69	38.50
OZ260506.1	14	25.92	38.50
OZ260507.1	15	25.80	38.50
OZ260508.1	16	25.57	38.50
OZ260509.1	17	24.92	38.50
OZ260510.1	18	24.90	38.50
OZ260511.1	19	24.77	38.50
OZ260512.1	20	24.07	38.50
OZ260513.1	21	23.34	38.50
OZ260514.1	22	21.38	39
OZ260515.1	23	21.15	39
OZ260516.1	24	20.23	39
OZ260517.1	25	17.80	39
OZ260518.1	26	17.47	38.50
OZ260519.1	27	15.89	39.50
OZ260520.1	28	15.32	39
OZ260521.1	29	14.73	39.50
OZ260522.1	30	14.38	40.50
OZ260523.1	Z	36.33	38

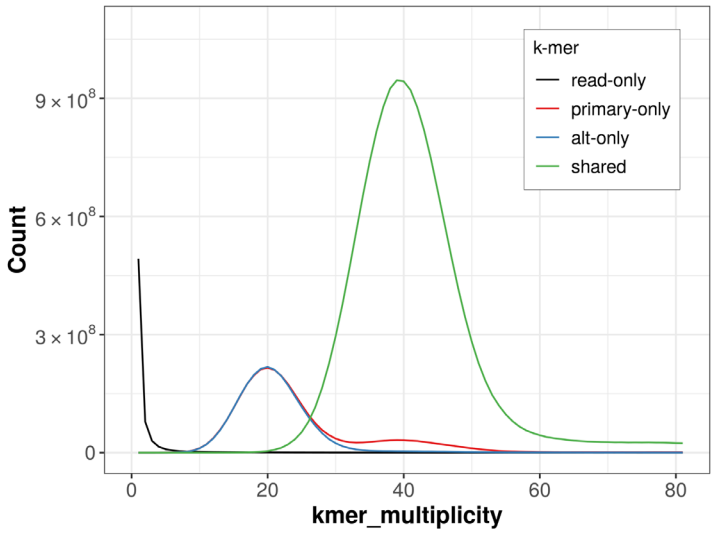

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

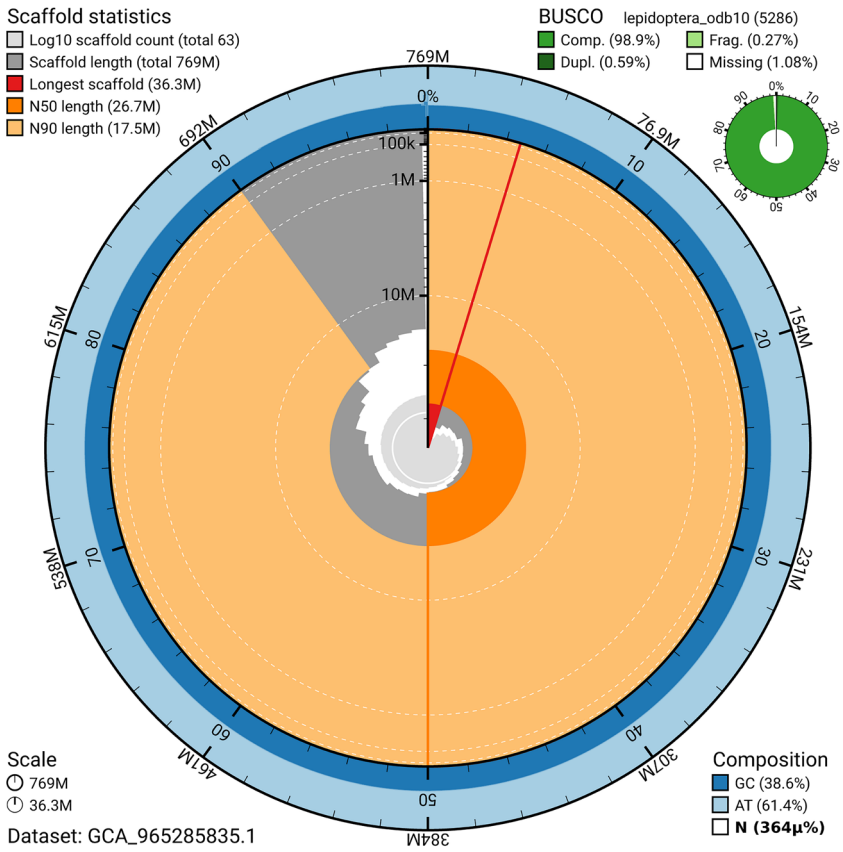

Figure 5. Assembly metrics for ilColCory3.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue 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.

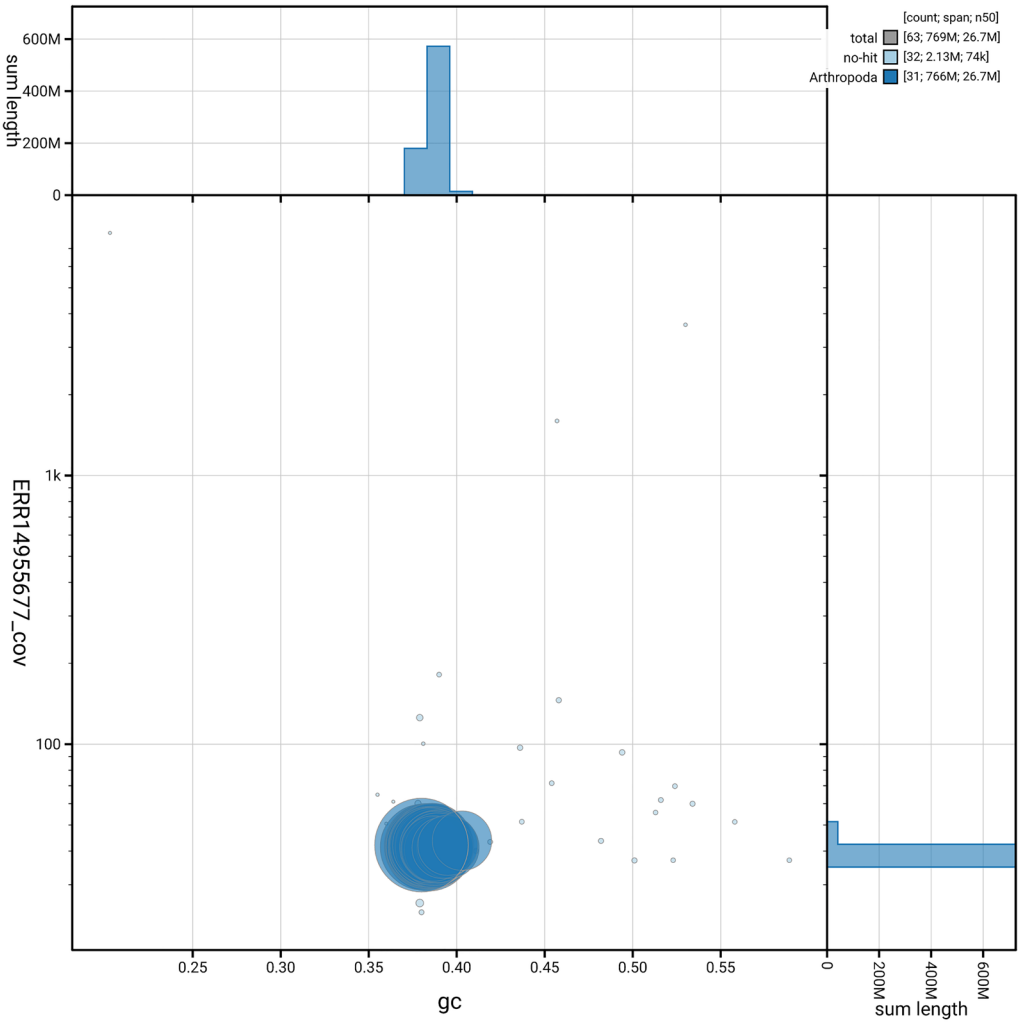

Figure 6. BlobToolKit GC-coverage plot for ilColCory3.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display 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 *Colocasia coryli* assembly.

Measure	Value	Benchmark
EBP summary (primary)	7.C.Q68	6.C.Q40
Contig N50 length	24.07 Mb	≥ 1 Mb
Scaffold N50 length	26.69 Mb	= chromosome N50
Consensus quality (QV)	Primary: 68.6; alternate: 65.7; combined: 66.5	≥ 40
k-mer completeness	Primary: 81.42%; alternate: 79.53%; combined: 99.70%	≥ 95%
BUSCO	C:98.9% [S:98.3%; D:0.6%]; F:0.3%; M:0.8%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.73%	≥ 90%

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

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

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

Data availability

European Nucleotide Archive: Colocasia coryli (nut-tree tussock). Accession number PRJEB89322. The genome sequence is released openly for reuse. The *Colocasia coryli* 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.

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.4.5	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.7.1	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.28-r1209	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	24.10.4	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.5	https://github.com/sanger-tol/PretextSnapshot

Software	Version	Source
PretextView	1.0.3	https://github.com/sanger-tol/PretextView
samtools	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.7.1	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2.2	https://github.com/c-zhou/yahs

Author information

Contributors are listed at the following links:

- Members of the University of Oxford and Wytham Woods 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

References

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

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

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

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

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

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

Darwin Tree of Life Project Consortium: **Sequence locally, think globally: the Darwin Tree of Life Project.** *Proc Natl Acad Sci U S A.* 2022; **119**(4): e2115642118.
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: ***Colocasia coryli* (Linnaeus, 1758).** 2025.
Reference Source

Heath J, Emmet AM: **The moths and butterflies of Great Britain and Ireland: Noctuidae (Part II) and Agaristidae.** Colchester: Harley Books, 1983.
Reference Source

Henwood B, Stirling P, Lewington R: **Field guide to the Caterpillars of Great Britain and Ireland.** London: Bloomsbury, 2020.
Reference Source

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

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

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

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

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text | Free Full Text

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

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

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

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

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

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

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

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

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

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

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

Waring P, Townsend M, Lewington R: **Field guide to the moths of Great Britain and Ireland.** London, UK: Bloomsbury, 2017.
Reference Source

Wright CJ, Stevens L, Mackintosh A, *et al.*: **Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera.** *Nat Ecol Evol.* 2024; **8**(4): 777–790.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 30 January 2026

https://doi.org/10.21956/wellcomeopenres.27978.r143401

✔ **Yash Gupta**
Department of Biology, Faculty of Science, Naresuan University, Phitsanulok, Thailand

This is a high-quality genome data note that clearly meets the standards expected for Darwin Tree of Life (DToL) outputs. The rationale is well articulated, methodologies are robust and appropriate, and the resulting datasets are accessible and reusable.

The manuscript is technically sound and suitable for indexing as a reference genome resource.

References

1. INCHOEDCHAY K, HOMCHAN S, GUPTA Y: Phylogenetic and mitochondrial genome analysis of a putative new cave cricket species (Rhaphidophoridae) from a Thai subterranean habitat. *Biodiversitas Journal of Biological Diversity*. 2025; **26** (8). Publisher Full Text

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

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

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

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I am a bioinformatician working with insect genomes.

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 19 January 2026

https://doi.org/10.21956/wellcomeopenres.27978.r143397

David EK Ferrier
University of St Andrews, Scotland, UK

The rationale is clear. This moth has been sequenced as part of the DToL initiative.

The various protocols are certainly appropriate and the work is of a high technical standard.

The materials and methods descriptions meet accepted standards for these data notes, and the datasets such as the assemblies of the two haplotypes are accessible and useable.

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.

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.