DATA NOTE

The genome sequence of the Ribwort slender, *Aspilapteryx tringipennella* (Zeller, 1839) (Lepidoptera: Gracillariidae)

[version 1; peer review: 2 approved]

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

[1]Syngenta International Research Station, Jealott's Hill, Berkshire, England, UK
[2]University of Exeter, Penryn, Cornwall, England, UK
[3]University of Oxford, Oxford, England, UK

v1 **First published:** 21 Oct 2025, **10**:588
https://doi.org/10.12688/wellcomeopenres.25049.1
Latest published: 21 Oct 2025, **10**:588
https://doi.org/10.12688/wellcomeopenres.25049.1

Abstract

We present a genome assembly from an individual male *Aspilapteryx tringipennella* (Ribwort slender; Arthropoda; Insecta; Lepidoptera; Gracillariidae). The genome sequence has a total length of 261.71 megabases. Most of the assembly (99.02%) is scaffolded into 30 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled, with a length of 16.86 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

Aspilapteryx tringipennella; Ribwort slender; genome sequence; chromosomal; Lepidoptera

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 21 Oct 2025	✓ view	✓ view

1. **Danilo Trabudo do Amaral**, Universidade Federal do ABC Centro de Ciencias Naturais e Humanas, Santo André, Brazil

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

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: Sims I: Investigation, Resources; **Hutchinson F**: Investigation, Resources; **Crowley LM**: Investigation, Resources;

Competing interests: No competing interests were disclosed.

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

How to cite this article: Sims I, Hutchinson F, Crowley LM *et al.* **The genome sequence of the Ribwort slender, *Aspilapteryx tringipennella* (Zeller, 1839) (Lepidoptera: Gracillariidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, **10**:588 https://doi.org/10.12688/wellcomeopenres.25049.1

First published: 21 Oct 2025, **10**:588 https://doi.org/10.12688/wellcomeopenres.25049.1

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; Tineoidea; Gracillariidae; Gracillariinae; *Aspilapteryx*; *Aspilapteryx tringipennella* (Zeller, 1839) (NCBI:txid1594445)

Background

Aspilapteryx tringipennella (Ribwort Stilt) is a distinctive micro-moth with a smooth-scaled head and pale yellowish-brown forewings, which are pale greyish along the costa towards the apex (Sterling *et al.*, 2023). The costal half is speckled with numerous small dark dots arranged in rows from the base. It has a wingspan of 10–13 mm and, like its close relatives in the genus *Caloptilia*, it often rests with the forepart of its body raised on the front legs (Kimber, 2025).

The species has two generations annually, flying from April to October (Sterling *et al.*, 2023). The larvae feed on Ribwort Plantain (*Plantago lanceolata*), forming pale blotch mines on the upper surface of leaves that cause the edges to roll inwards (Sterling *et al.*, 2023). The second generation overwinters as larvae (Kimber, 2025).

The Ribwort Stilt occurs in grasslands, old quarries, vegetated shingle, and waste ground where Ribwort Plantain is present (Sterling *et al.*, 2023). In Britain, it is widespread (NBN Atlas Partnership, 2025). Globally, all GBIF occurrences are from Europe. The species has been recorded in several countries, most frequently in the United Kingdom, the Netherlands, Norway, Sweden and France (GBIF Secretariat, 2023).

This assembly is the first genome for the genus *Aspilapteryx* and one of seven genomes available for the family Gracillariidae as of September 2025 (data obtained via NCBI datasets, O'Leary *et al.*, 2024). The assembly was produced using the Tree of Life pipeline from a specimen collected in Hartslock Nature Reserve, England, UK (Figure 1), as part of the Darwin Tree of Life Project.

Figure 1. Photograph of the *Aspilapteryx tringipennella* (ilAspTrin1) specimen used for genome sequencing.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Aspilapteryx tringipennella* (specimen ID NHMUK013805904, ToLID ilAspTrin1; Figure 1), collected from Hartslock Nature Reserve, England, United Kingdom (latitude 51.51, longitude –1.11) on 2021-07-29. The specimen was collected by Ian Sims and identified by Ian Sims and David Lees. A second specimen was used for Hi-C sequencing (specimen ID Ox003079, ToLID ilAspTrin2). It was collected from Wytham Woods, Oxfordshire, United Kingdom (latitude 51.772, longitude –1.338) on 2022-07-22. The specimen was collected by Finley Hutchinson and Liam Crowley and identified by Finley Hutchinson. For the Darwin Tree of Life sampling and metadata approach, refer to 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 ilAspTrin1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the whole organism was homogenised by powermashing using a PowerMasher II tissue disruptor.

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol. We used centrifuge-mediated fragmentation to produce DNA fragments in the 8–10 kb range, following the Covaris g-TUBE protocol for ultra-low input (ULI). Sheared DNA was purified by 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.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Prior to library preparation, the DNA was fragmented to ~10 kb. Ultra-low-input (ULI) libraries were prepared using the PacBio SMRTbell® Express Template Prep Kit 2.0 and gDNA Sample Amplification

Kit. Samples were normalised to 20 ng DNA. Single-strand overhang removal, DNA damage repair, and end-repair/A-tailing were performed according to the manufacturer's instructions, followed by adapter ligation. A 0.85× pre-PCR clean-up was carried out with Promega ProNex beads.

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

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

The sample was sequenced 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 ilAspTrin2 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

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

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

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 33 breaks and 37 joins. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

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

completeness and assembly quality for the primary and alternate haplotypes using the *k*-mer databases (k = 31) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

The genome was analysed using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified from the NCBI Taxonomy (Schoch *et al.*, 2020). For the three domain-level lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all outputs into a blobdir for visualisation. The BlobToolKit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), with containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

PacBio sequencing of the *Aspilapteryx tringipennella* specimen generated 44.64 Gb (gigabases) from 4.18 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 231.34 Mb, with a heterozygosity of 2.10% and repeat content of 26.92% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 176× coverage. Hi-C sequencing produced 90.85 Gb from 601.64 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The final assembly has a total length of 261.71 Mb in 143 scaffolds, with 324 gaps, and a scaffold N50 of 9.5 Mb (Table 2).

Most of the assembly sequence (99.02%) was assigned to 30 chromosomal-level scaffolds, representing 29 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size

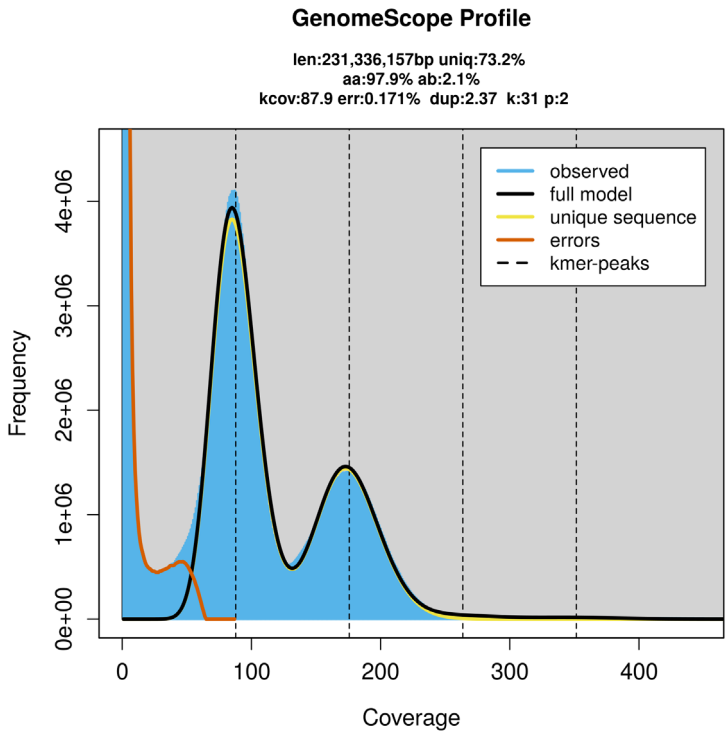

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB74968.

Platform	PacBio HiFi	Hi-C
ToLID	ilAspTrin1	ilAspTrin2
Specimen ID	NHMUK013805904	Ox003079
BioSample (source individual)	SAMEA112222179	SAMEA113425683
BioSample (tissue)	SAMEA112222240	SAMEA113425864
Tissue	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR12921316	ERR12945470
Read count total	4.18 million	601.64 million
Base count total	44.64 Gb	90.85 Gb

Table 2. Genome assembly statistics.

Assembly name	ilAspTrin1.1
Assembly accession	GCA_964188165.1
Alternate haplotype accession	GCA_964194445.1
Assembly level	chromosome
Span (Mb)	261.71
Number of chromosomes	30
Number of contigs	467
Contig N50	1.43 Mb
Number of scaffolds	143
Scaffold N50	9.5 Mb
Sex chromosomes	Z
Organelles	Mitochondrion: 16.86 kb

(Figure 3; Table 3). Chromosome Z was assigned based on synteny to *Euspilapteryx auroguttella* (GCA_951802225.1) (Boyes *et al.*, 2024).

The mitochondrial genome was also assembled (length 16.86 kb, OZ076590.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

The combined primary and alternate assemblies achieve an estimated QV of 59.2. The *k*-mer completeness is 70.02% for the primary assembly, 71.70% for the alternate haplotype, and 99.63% for the combined assemblies (Figure 4).

BUSCO v.5.5.0 analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 96.2% of the expected gene set (single = 95.4%, duplicated = 0.8%). 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 **6.C.Q59**, 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 within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project. Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible. The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material

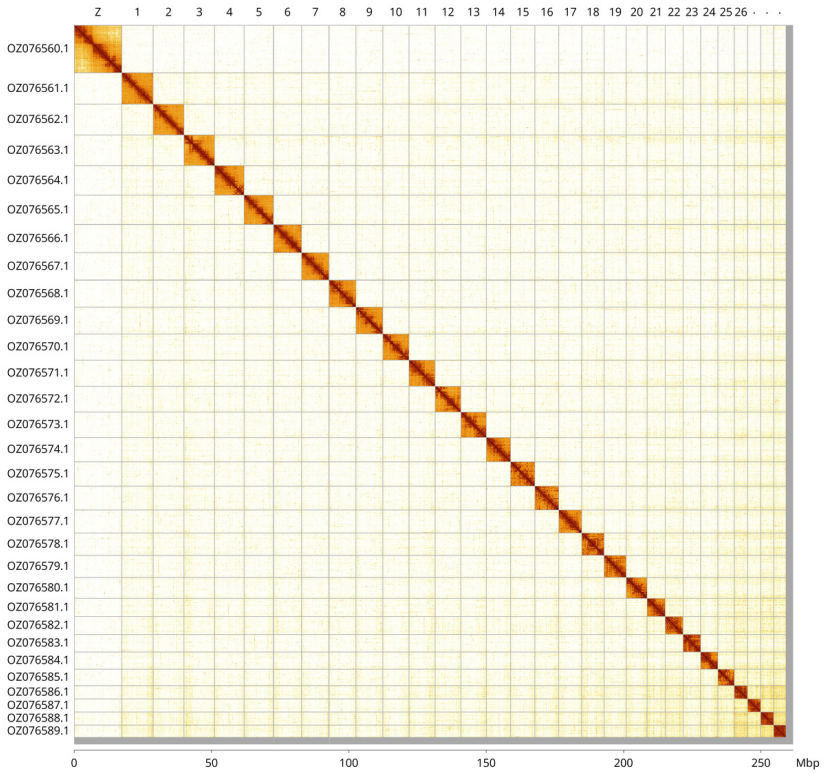

Figure 3. Hi-C contact map of the *Aspilapteryx tringipennella* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Aspilapteryx tringipennella* ilAspTrin1.

INSDC accession	Molecule	Length (Mb)	GC%
OZ076561.1	1	11.39	37.50
OZ076562.1	2	11.24	37.50
OZ076563.1	3	11.13	37.50
OZ076564.1	4	10.76	37.50
OZ076565.1	5	10.70	37.50
OZ076566.1	6	10.21	37
OZ076567.1	7	10.01	37
OZ076568.1	8	9.77	37
OZ076569.1	9	9.77	37
OZ076570.1	10	9.54	37
OZ076571.1	11	9.50	37.50
OZ076572.1	12	9.38	37
OZ076573.1	13	9.24	37.50
OZ076574.1	14	8.85	37
OZ076575.1	15	8.85	37
OZ076576.1	16	8.71	38
OZ076577.1	17	8.38	37
OZ076578.1	18	8.15	37.50
OZ076579.1	19	8	37
OZ076580.1	20	7.64	37.50
OZ076581.1	21	6.55	37.50
OZ076582.1	22	6.53	37.50
OZ076583.1	23	6.36	37.50
OZ076584.1	24	6.25	38
OZ076585.1	25	6.01	37.50
OZ076586.1	26	4.85	38
OZ076587.1	27	4.76	37.50
OZ076588.1	28	4.71	39
OZ076589.1	29	4.50	38.50
OZ076560.1	Z	17.40	37.50

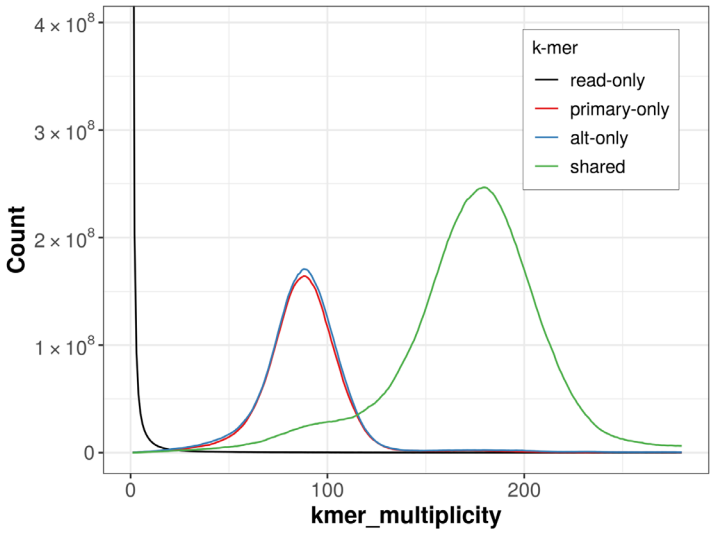

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

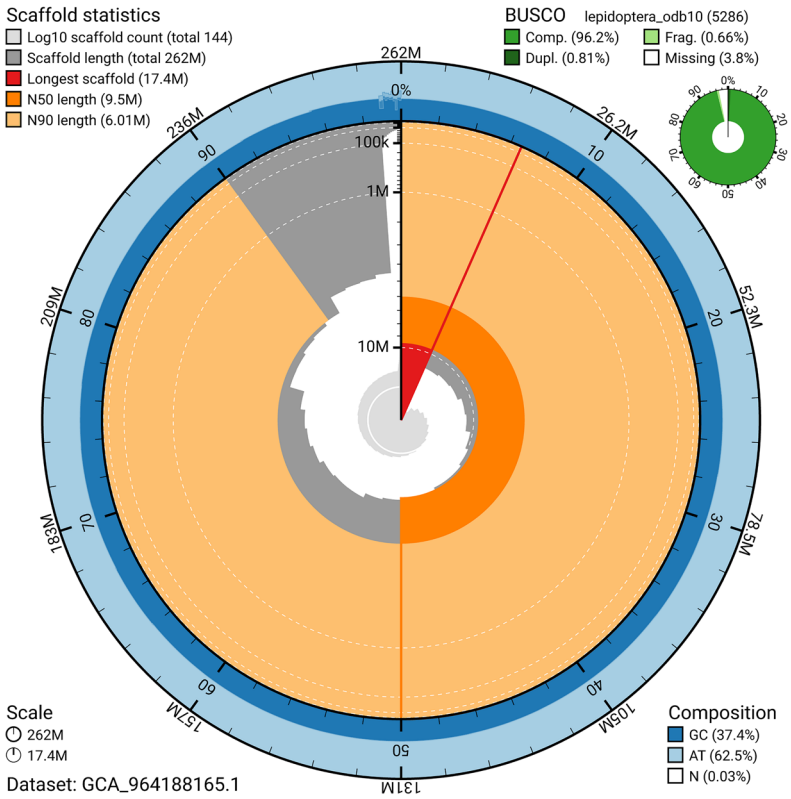

Figure 5. Assembly metrics for ilAspTrin1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue 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.

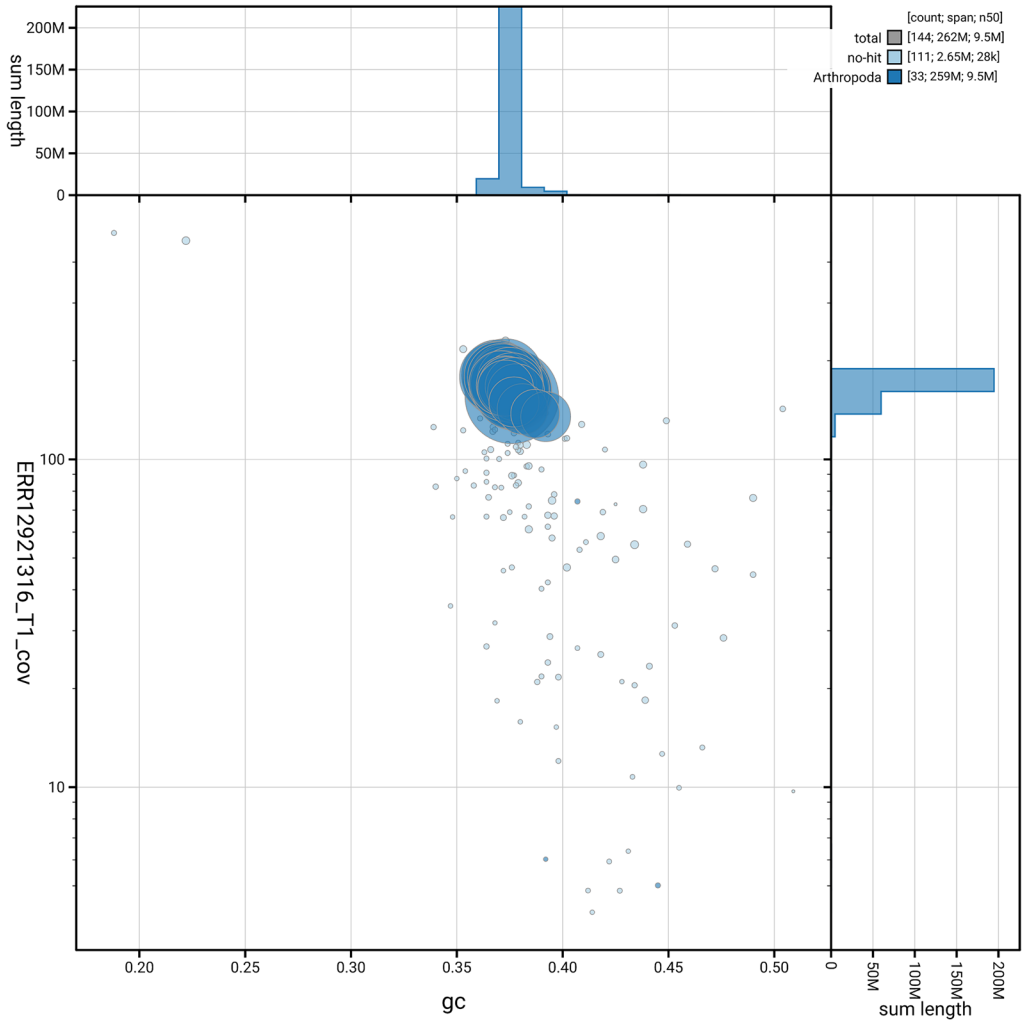

Figure 6. BlobToolKit GC-coverage plot for ilAspTrin1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display 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 *Aspilapteryx tringipennella* assembly.

Measure	Value	Benchmark
EBP summary (primary)	6.C.Q59	6.C.Q40
Contig N50 length	1.43 Mb	≥ 1 Mb
Scaffold N50 length	9.50 Mb	= chromosome N50
Consensus quality (QV)	Primary: 59.5; alternate: 58.9; combined: 59.2	≥ 40
k-mer completeness	Primary: 70.02%; alternate: 71.70%; combined: 99.63%	≥ 95%
BUSCO	C:96.2% [S:95.4%; D:0.8%]; F:0.7%; M:3.1%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.02%	≥ 90%

- 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: Aspilapteryx tringipennella. Accession number PRJEB74968. The genome sequence is released openly for reuse. The *Aspilapteryx tringipennella* 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 Marine Biological Association Genome Acquisition Lab
- Members of the Darwin Tree of Life Barcoding collective
- Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team
- Members of Wellcome Sanger Institute Scientific Operations – Sequencing Operations
- Members of the Wellcome Sanger Institute Tree of Life Core Informatics team
- Members of the Tree of Life Core Informatics collective
- Members of the Darwin Tree of Life Consortium

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
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	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi

Software	Version	Source
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextSnapshot	-	https://github.com/sanger-tol/PretextSnapshot
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
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

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

Boyes D, Lees DC, Sims I, *et al.*: **The genome sequence of the Yellow-dotted Stilt, *Euspilapteryx auroguttella* Stephens, 1835 [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2024; **9**: 226.
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

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: ***Aspilapteryx tringipennella* (Zeller, 1839) in GBIF Backbone Taxonomy.** 2023.
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

Kimber I: ***Aspilapteryx tringipennella* (Zeller, 1839) 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, 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

NBN Atlas Partnership: ***Aspilapteryx tringipennella* (Zeller, 1839) in 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.*: **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**

Sterling P, Parsons M, Lewington R: **Field guide to the Micro Moths of Great Britain and Ireland.** Dorset: British Wildlife Publishing, 2023.
Reference Source

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

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

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

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

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 27 January 2026

https://doi.org/10.21956/wellcomeopenres.27615.r144036

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

This paper describes the sequencing and assembly of the genome of the lepidopteran Apilapteryx tringipennella.

The methodology used for genome sequencing and genome assembly is the one routinely used for all DToL projects and resulted in a very good assembly.

More specifically, virtually the entire assembly is assigned to one of the 30 chromosome-level scaffolds, thus resulting in an excellent contiguity. However, there are >100 unassigned scaffolds. These scaffolds appear to be of lower-than-expected coverage and of higher-than-expected GC content (Figure 6). It would be nice if the authors could comment on the putative origin of these scaffolds (and, of course, any other features they deem interesting).

Additionally, all the other assembly evaluation metrics are very good, with k-mer completeness at 99.63%, QV at 59.2. The fraction of complete BUSCOs is at 96.2% and even though it is a good score, it could probably be better. It would also be very nice if the authors could add a brief comment about the BUSCO results.

Finally, it is good to see that a gene set will be available, even it's not mentioned when exactly. As of today (Jan-26), there's nothing available at the Ensembl website. It is very important to point out that having a gene set is of crucial to most scientists. Genome annotation is a tedious process and not everyone can do it. So, maybe providing a gene set for these DToL assemblies should be almost "mandatory".

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: insect genomics, bioinformatics

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 12 November 2025

https://doi.org/10.21956/wellcomeopenres.27615.r138084

Danilo Trabudo do Amaral
Universidade Federal do ABC Centro de Ciencias Naturais e Humanas, Santo André, State of São Paulo, Brazil

The manuscript is very well prepared, with an excellent assembly and a clear methodological workflow. The text is concise and the results are of high quality. I commend the authors for producing a genome assembly that meets current reference standards.I have only a few recommendations that could further improve clarity and completeness:
Please include a short explanation describing the observed asymmetry between the primary and combined k-mer completeness values, in relation to haplotype resolution and the estimated heterozygosity (~2.1%). This will help readers understand why the combined completeness is close to 100% while the primary is around 70%.
Since the text already cites synteny with Euspilapteryx, consider adding one line summarizing additional evidence (for example, differential Hi-C coverage, sex-linked read depth, or mapping of male reads) to reinforce the identification of the Z chromosome.
Adding a brief paragraph about the repetitive fraction of the genome, optionally including the LTR Assembly Index (LAI), would help contextualize contiguity and N50 results.
If applicable, a short mention of whether telomeric or centromeric repeats were detected at chromosome ends would further support the assembly's chromosomal completeness.
Summarize the outcome of the ASCC step in one sentence (e.g., "No scaffolds were flagged as contaminants; X Mb removed prior to the final version") to improve transparency.
Figure 1. The specimen image currently appears of limited morphological quality. Replacing it with a clearer photo or omitting it altogether would strengthen the visual presentation.

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: Bioinformatic

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.