DATA NOTE

The genome sequence of the Small Fan-footed Wave moth, *Idaea biselata* (Hufnagel, 1767)

[version 1; peer review: 3 approved, 1 approved with reservations]

Ian Sims[1], David C. Lees[2], Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team, Wellcome Sanger Institute Scientific Operations: Sequencing Operations, Wellcome Sanger Institute Tree of Life Core Informatics team, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]Syngenta International Research Station, Jealott's Hill, Berkshire, England, UK
[2]Natural History Museum, London, England, UK

v1 **First published:** 19 Feb 2025, **10**:78
https://doi.org/10.12688/wellcomeopenres.23706.1
Latest published: 19 Feb 2025, **10**:78
https://doi.org/10.12688/wellcomeopenres.23706.1

Abstract

We present a genome assembly from an individual male specimen of *Idaea biselata* (Small Fan-footed Wave; Arthropoda; Insecta; Lepidoptera; Geometridae). The genome sequence has a total length of 424.30 megabases. Most of the assembly (97.24%) is scaffolded into 13 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled and is 17.33 kilobases in length. Gene annotation of this assembly on Ensembl identified 11,310 protein-coding genes.

Keywords

Idaea biselata, Small Fan-footed Wave, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ? ✓ ✓

	1	2	3	4
version 1 19 Feb 2025	✓ view	? view	✓ view	✓ view

1. **Pritha Dey**, National Centre for Biological Sciences-TIFR, Bengaluru, India
2. **Yu-Feng Huang**, National Chung Hsing University, Taichung City, Taiwan
Yuan Ze University (Ringgold ID: 34895), Zhongli District, Taiwan
3. **Axel Künstner**, University of Lübeck, Lübeck, Germany
4. **Shangzhe Zhang**, University of St Andrews Library, St Andrews, UK

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: Sims I: Investigation, Resources; **Lees DC**: 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 [206194, https://doi.org/10.35802/206194] 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, Lees DC, Natural History Museum Genome Acquisition Lab *et al.* **The genome sequence of the Small Fan-footed Wave moth, *Idaea biselata* (Hufnagel, 1767) [version 1; peer review: 3 approved, 1 approved with reservations]** Wellcome Open Research 2025, **10**:78 https://doi.org/10.12688/wellcomeopenres.23706.1

First published: 19 Feb 2025, **10**:78 https://doi.org/10.12688/wellcomeopenres.23706.1

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; Geometroidea; Geometridae; Sterrhinae; *Idaea*; *Idaea biselata* (Hufnagel, 1767) (NCBI:txid104448)

Background

The genome of the Small Fan-footed Wave, *Idaea biselata*, 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 (Blaxter *et al.*, 2022). Here we present a chromosome-level genome sequence for *Idaea biselata*, based on a male specimen from Hartslock Nature Reserve, England, United Kingdom (Figure 1).

Genome sequence report

The genome of *Idaea biselata* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of 21.52 Gb (gigabases) from 2.41 million reads, providing an estimated 49-fold coverage. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data, which produced 83.54 Gb from 553.26 million reads. Specimen and sequencing details are summarised in Table 1.

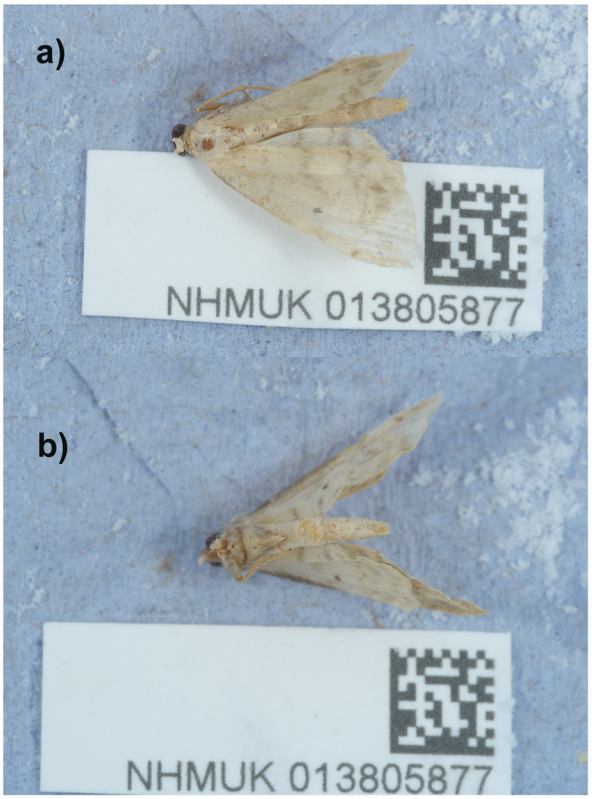

Figure 1. Photographs of the *Idaea biselata* (ilIdaBise2) specimen used for genome sequencing.

Assembly errors were corrected by manual curation, including 75 missing joins or mis-joins and 13 haplotypic duplications. This reduced the assembly length by 0.51%, the scaffold number by 5.69%, and increased the scaffold N50 by 1.71%. The final assembly has a total length of 424.30 Mb in 380 sequence scaffolds, with 1,283 gaps, and a scaffold N50 of 31.0 Mb (Table 2).

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

Most of the assembly sequence (97.24%) was assigned to 13 chromosomal-level scaffolds, representing 12 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by the Hi-C data, are named in order of size (Figure 5; Table 3). During manual curation, chromosome Z was assigned by synteny to the genomes of *Idaea straminata* (GCA_951213275.1) and *Idaea dimidiata* (GCA_949358125.1).

While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial genome was also assembled and can be found as a contig within the multifasta file of the genome submission, and as a separate fasta file.

The final assembly has a Quality Value (QV) of 57.4. The *k*-mer completeness for the primary assembly is 81.19%, for the alternate haplotype is 80.26%, and for the combined assemblies is 98.58%. BUSCO (v5.4.3) analysis using the lepidoptera_odb10 reference set (n = 5,286) indicated a completeness score of 94.6% (single = 93.9%, duplicated = 0.7%). The assembly achieves the Earth Biogenome Project (EBP) reference standard of 5.C.Q57. Other quality metrics are given in Table 2.

Genome annotation report

The *Idaea biselata* genome assembly (GCA_958496205.1) was annotated at the European Bioinformatics Institute (EBI) on Ensembl Rapid Release. The resulting annotation includes 20,126 transcribed mRNAs from 11,310 protein-coding and 1,435 non-coding genes (Table 2; https://rapid.ensembl.org/Idaea_biselata_GCA_958496205.1/Info/Index). The average transcript length is 12,678.94. There are 1.58 coding transcripts per gene and 6.59 exons per transcript.

Methods

Sample acquisition and DNA barcoding

An adult male specimen of *Idaea biselata* (specimen ID NHMUK013805877, ToLID ilIdaBise2) was collected from Hartslock Nature Reserve, England, United Kingdom (latitude 51.51, longitude –1.11) on 2021-07-29, using a light trap. The specimen was collected by Ian Sims (British Entomological and Natural History Society and identified by Ian Sims and

Table 1. Specimen and sequencing data for *Idaea biselata*.

Project information			
Study title	Idaea biselata (small fan-footed wave)		
Umbrella BioProject	PRJEB61335		
Species	*Idaea biselata*		
BioSample	SAMEA111458426		
NCBI taxonomy ID	104448		
Specimen information			
Technology	**ToLID**	**BioSample accession**	**Organism part**
PacBio long read sequencing	ilIdaBise2	SAMEA111458471	thorax
Hi-C sequencing	ilIdaBise2	SAMEA111458501	head
RNA sequencing	ilIdaBise2	SAMEA111458521	abdomen
Sequencing information			
Platform	**Run accession**	**Read count**	**Base count (Gb)**
Hi-C Illumina NovaSeq 6000	ERR11242543	5.53e+08	83.54
PacBio Sequel IIe	ERR11242126	2.41e+06	21.52
RNA Illumina NovaSeq 6000	ERR12245555	6.46e+07	9.76

David Lees (Natural History Museum) and preserved by dry freezing at –80 °C.

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 specimens and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding have been deposited on protocols.io (Beasley *et al.*, 2023).

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io (Denton *et al.*, 2023b).

The ilIdaBise2 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the thorax was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a).

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023). The DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Bates *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Strickland *et al.*, 2023). 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.

RNA was extracted from abdomen tissue of ilIdaBise2 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol (do Amaral *et al.*, 2023). 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.

Hi-C preparation

Tissue from the head of the ilIdaBise2 sample was processed at the WSI Scientific Operations core, using the Arima-HiC v2 kit. Tissue (stored at –80 °C) was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde. After crosslinking, the tissue was homogenised using the Diagnocine Power Masher-II and BioMasher-II tubes and pestles. Following the kit manufacturer's instructions, crosslinked DNA

Table 2. Genome assembly data for *Idaea biselata*, ilIdaBise2.1.

Genome assembly		
Assembly name	ilIdaBise2.1	
Assembly accession	GCA_958496205.1	
Accession of alternate haplotype	*GCA_958496235.1*	
Span (Mb)	424.30	
Number of contigs	1,664	
Number of scaffolds	380	
Longest scaffold (Mb)	60.78	
Assembly metrics*		***Benchmark***
Contig N50 length (Mb)	0.6	*≥ 1 Mb*
Scaffold N50 length (Mb)	31.0	*= chromosome N50*
Consensus quality (QV)	57.4	*≥ 40*
k-mer completeness	Primary: 81.19%; alternate: 80.26%; combined: 98.58%.	*≥ 95%*
BUSCO**	C:94.6%[S:93.9%,D:0.7%], F:1.5%,M:3.9%,n:5,286	*S > 90%, D < 5%*
Percentage of assembly mapped to chromosomes	97.24%	*≥ 90%*
Sex chromosomes	Z	*localised homologous pairs*
Organelles	Mitochondrial genome: 17.33 kb	*complete single alleles*
Genome annotation of assembly GCA_958496205.1 at Ensembl		
Number of protein-coding genes	11,310	
Number of non-coding genes	1,435	
Number of gene transcripts	20,126	

* Assembly metric benchmarks are adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024.

** BUSCO scores based on the lepidoptera_odb10 BUSCO set using version 5.4.3. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison.

was digested using a restriction enzyme master mix. The 5'-overhangs were then filled in and labelled with biotinylated nucleotides and proximally ligated. An overnight incubation was carried out for enzymes to digest remaining proteins and for crosslinks to reverse. A clean up was performed with SPRIselect beads prior to library preparation.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Pacific Biosciences HiFi circular consensus DNA sequencing libraries were prepared using the PacBio Express Template Preparation Kit v2.0 (Pacific Biosciences, California, USA) as per the manufacturer's instructions. The kit includes the reagents required for removal of single-strand overhangs, DNA damage repair, end repair/A-tailing, adapter ligation, and nuclease treatment. Library preparation also included a library purification step using AMPure PB beads (Pacific Biosciences, California, USA) and size selection step to remove templates shorter than 3 kb using AMPure PB modified SPRI. DNA concentration was quantified using the Qubit Fluorometer v2.0 and Qubit HS Assay Kit and the final library fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument and gDNA 165kb gDNA and 55kb BAC analysis kit. Samples were sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was in the range 40–135 pM. The SMRT link software, a PacBio web-based

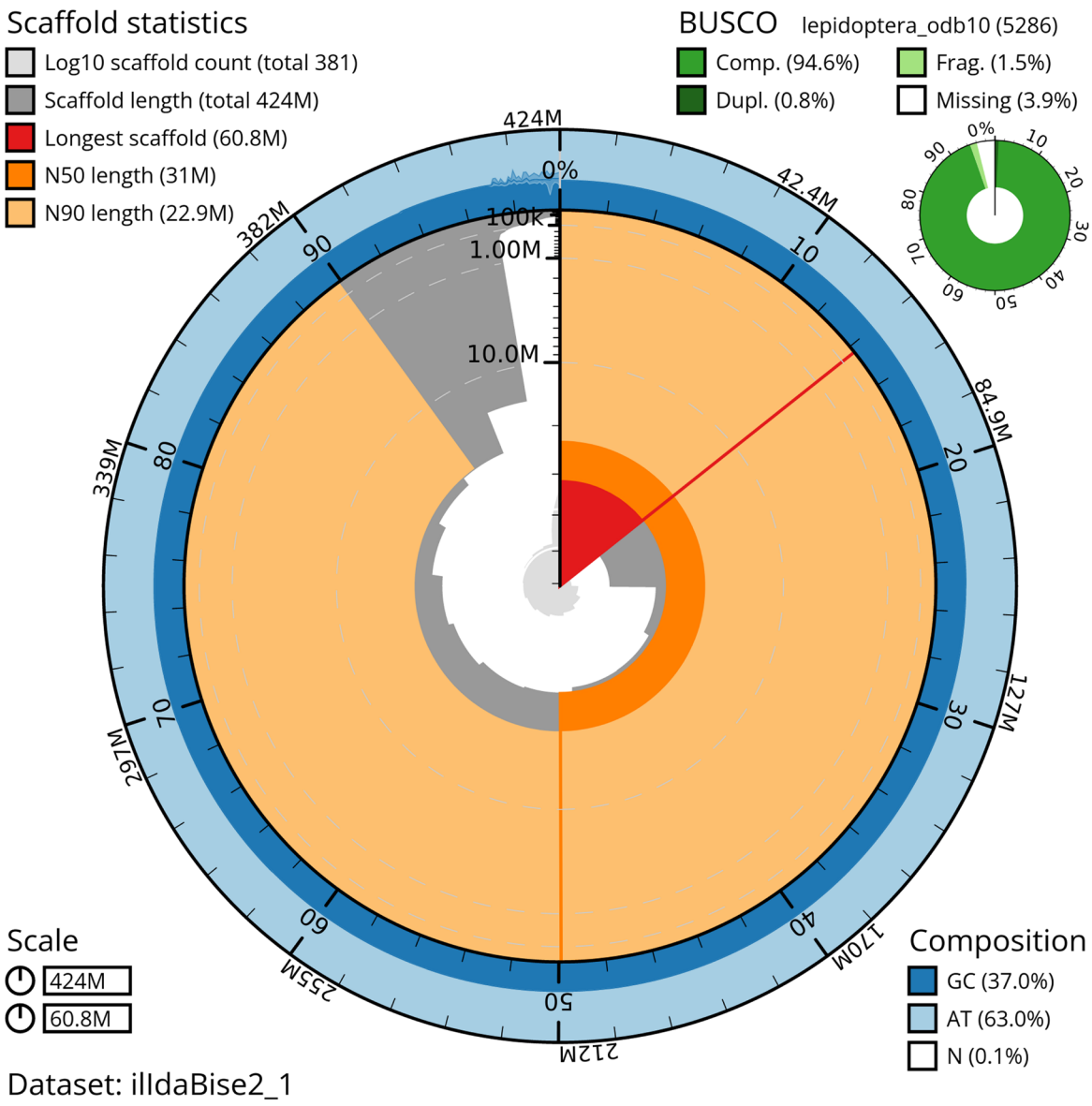

Figure 2. Genome assembly of *Idaea biselata*, ilIdaBise2.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the lepidoptera_odb10 set is presented at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilIdaBise2_1/dataset/ilIdaBise2_1/snail.

end-to-end workflow manager, was used to set-up and monitor the run, as well as perform primary and secondary analysis of the data upon completion.

For Hi-C library preparation, DNA was fragmented to a size of 400 to 600 bp using a Covaris E220 sonicator. The DNA was then enriched, barcoded, and amplified using the NEBNext Ultra II DNA Library Prep Kit following manufacturers' instructions. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on an Illumina NovaSeq 6000 instrument.

Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit, following the manufacturer's instructions. RNA sequencing was performed on the Illumina NovaSeq 6000 instrument.

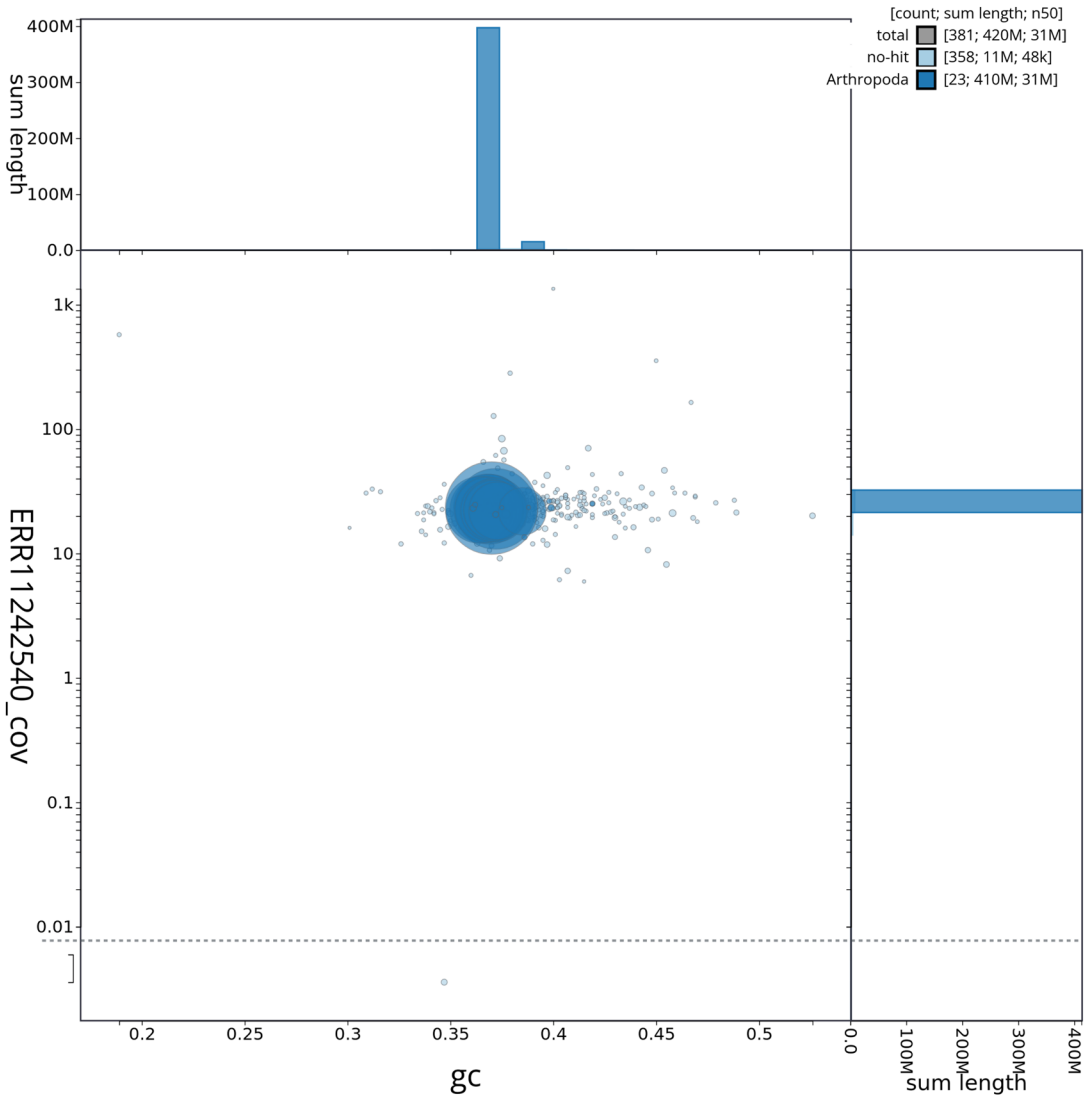

Figure 3. Genome assembly of *Idaea biselata*, ilIdaBise2.1: BlobToolKit GC-coverage plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilIdaBise2_1/dataset/ilIdaBise2_1/blob.

Genome assembly, curation and evaluation

Assembly

The HiFi reads were first assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). The Hi-C reads were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). The contigs were further scaffolded using the provided Hi-C data (Rao *et al.*, 2014) in YaHS (Zhou *et al.*, 2023) using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

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

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline (article in preparation). Manual curation was primarily conducted using PretextView (Harry, 2022), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023) and HiGlass

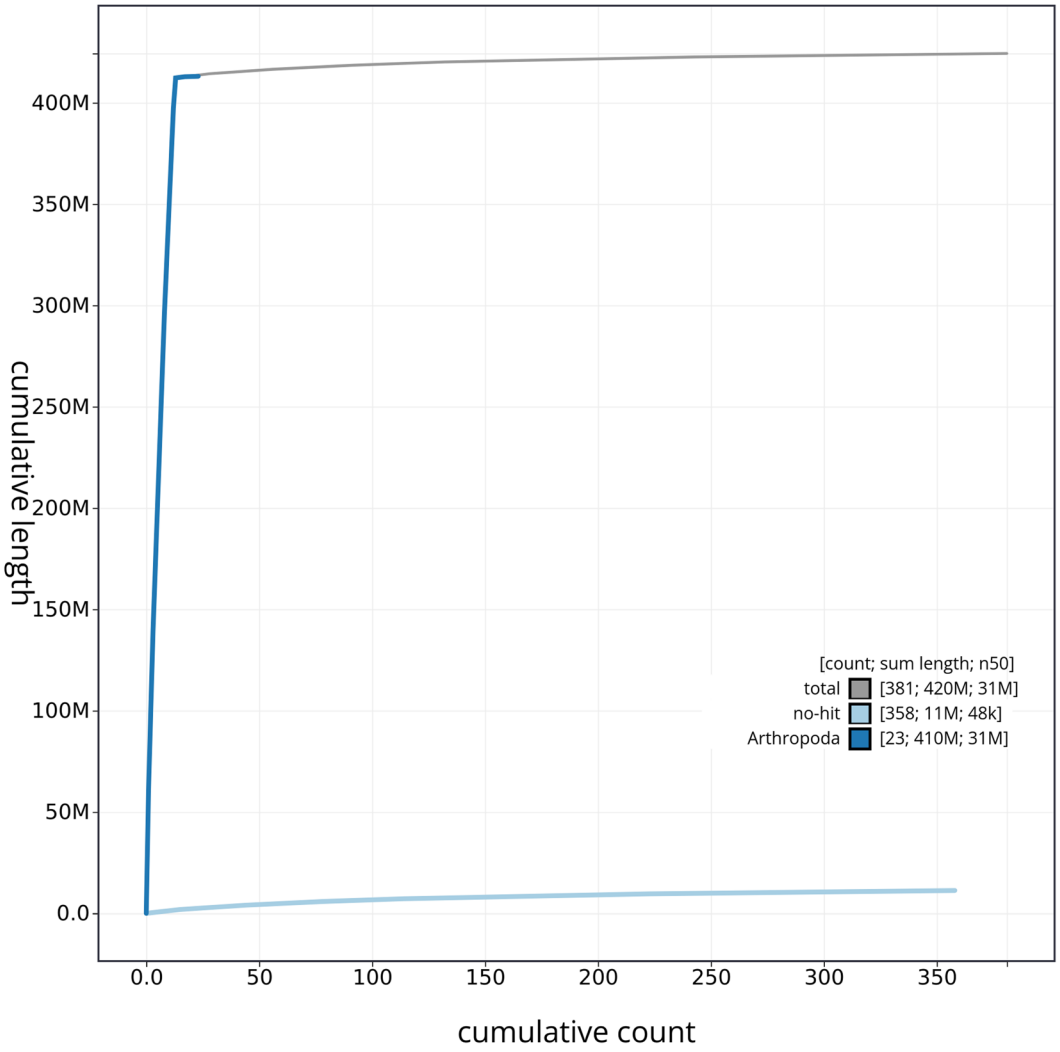

Figure 4. Genome assembly of *Idaea biselata* ilIdaBise2.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all sequences. Coloured lines show cumulative lengths of sequences assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilIdaBise2_1/dataset/ilIdaBise2_1/cumulative.

(Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were corrected, and duplicate sequences were tagged and removed. The sex chromosome was identified by synteny analysis. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation (article in preparation).

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020), run in a Singularity container (Kurtzer *et al.*, 2017), was used to evaluate *k*-mer completeness and assembly quality for the primary and alternate haplotypes using the *k*-mer databases (k = 31) that were computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

A Hi-C contact map was produced for the final version of the assembly. The Hi-C reads were aligned using bwa-mem2 (Vasimuddin *et al.*, 2019) and the alignment files were combined using SAMtools (Danecek *et al.*, 2021). The Hi-C alignments were converted into a contact map using BEDTools (Quinlan & Hall, 2010) and the Cooler tool suite (Abdennur & Mirny, 2020). The contact map was visualised in HiGlass (Kerpedjiev *et al.*, 2018).

The genome was also analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021) were calculated.

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

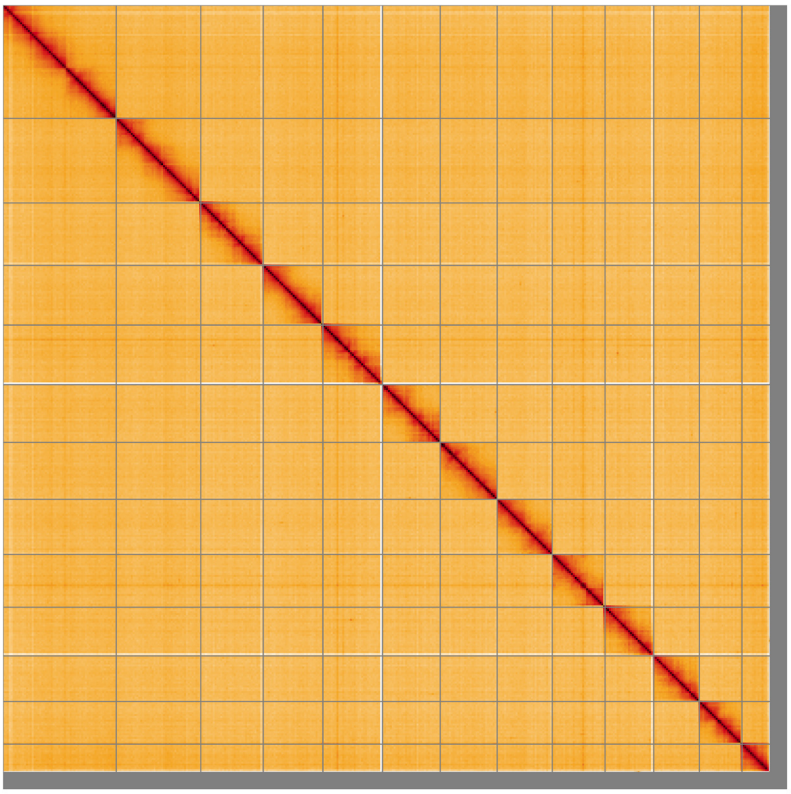

Figure 5. Genome assembly of *Idaea biselata* ilIdaBise2.1: Hi-C contact map of the ilIdaBise2.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=L_Vai7TPRBOPt5iH0NEQFw.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Idaea biselata*, ilIdaBise2.

INSDC accession	Name	Length (Mb)	GC%
OY292395.1	1	45.51	37.0
OY292396.1	2	33.5	37.0
OY292397.1	3	32.12	37.0
OY292398.1	4	32.02	37.0
OY292399.1	5	30.96	36.5
OY292400.1	6	30.63	36.5
OY292401.1	7	29.55	37.0
OY292402.1	8	28.45	37.0
OY292403.1	9	26.16	37.0
OY292404.1	10	24.54	37.0
OY292405.1	11	22.88	37.5
OY292406.1	12	15.09	38.5
OY292394.1	Z	60.78	37.0
OY292407.1	MT	0.02	19.0

Genome annotation

The Ensembl Genebuild annotation system (Aken *et al.*, 2016) was used to generate annotation for the *Idaea biselata* assembly (GCA_958496205.1) in Ensembl Rapid Release at the EBI. Annotation was created primarily through alignment of transcriptomic data to the genome, with gap filling via protein-to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 2019).

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the **'Darwin Tree of Life Project Sampling Code of Practice'**, which can be found in full on the Darwin Tree of Life website here. 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

Table 4. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	4.2.1	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.4.3 and 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
FastK	427104ea91c78c3b8b8b49f1a7d6bbeaa869ba1c	https://github.com/thegenemyers/FASTK
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Hifiasm	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84 aa44357826c0b6753eb28de	https://github.com/higlass/higlass
Merqury.FK	d00d98157618f4e8d1a9190026b19b471055b22e	https://github.com/thegenemyers/MERQURY.FK
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
Nextflow	23.04.0-5857	https://github.com/nextflow-io/nextflow
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
sanger-tol/ascc	-	https://github.com/sanger-tol/ascc
Singularity	3.9.0	https://github.com/sylabs/singularity
YaHS	1.2a.2	https://github.com/c-zhou/yahs

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: Idaea biselata (small fan-footed wave). Accession number PRJEB61335; https://identifiers.org/ena.embl/PRJEB61335. The genome sequence is released openly for reuse. The *Idaea biselata* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project. All raw sequence data and the assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

Author information

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

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

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

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

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

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

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

References

Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316.
PubMed Abstract | Publisher Full Text | Free Full Text

Aken BL, Ayling S, Barrell D, *et al.*: **The Ensembl gene annotation system.** *Database (Oxford).* 2016; **2016**: baw093.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor®3 for LI PacBio.** *protocols.io.* 2023.
Publisher Full Text

Beasley J, Uhl R, Forrest LL, *et al.*: **DNA barcoding SOPs for the Darwin Tree of Life project.** *protocols.io.* 2023; [Accessed 25 June 2024].
Publisher Full Text

Blaxter M, Mieszkowska N, Di Palma F, *et al.*: **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

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

Denton A, Oatley G, Cornwell C, *et al.*: **Sanger Tree of Life sample homogenisation: PowerMash.** *protocols.io.* 2023a.
Publisher Full Text

Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life wet laboratory protocol collection V.1.** *protocols.io.* 2023b.
Publisher Full Text

Diesh C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of synteny and structural variation.** *Genome Biol.* 2023; **24**(1): 74.
PubMed Abstract | Publisher Full Text | Free Full Text

do Amaral RJV, Denton A, Yatsenko H, *et al.*: **Sanger Tree of Life RNA extraction: automated MagMax™ mirVana.** *protocols.io.* 2023.
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

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

Harry E: **PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps**. 2022.
Reference Source

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

Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life sample preparation: triage and dissection.** *protocols.io.* 2023.
Publisher Full Text

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

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

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

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

Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–842.
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

Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *protocols.io.* 2023.
Publisher 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

UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–D515.
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 30 January 2026

https://doi.org/10.21956/wellcomeopenres.26154.r140286

✔ **Shangzhe Zhang**
University of St Andrews Library, St Andrews, Scotland, UK

The data note presented a high-quality genome assembly of the Small Fan-footed Wave moth, as part of the ToL project. The methods are robust and the assembly metrics are well-presented. I only have a few concerns regarding the clarity of this manuscript.

I noticed that the authors mentioned that the assembly has 380 scaffolds but all diagrams showed 381 scaffolds, is it because that the mitochondrial scaffold was not removed in these diagrams?

"While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited." Here more clarifications are needed -- the authors mentioned that the primary contig assembly was selected and used for scaffolding -- which means no phasing was performed in assembling.

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

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

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

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, Evolutionary biology

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 28 January 2026

https://doi.org/10.21956/wellcomeopenres.26154.r145303

Axel Künstner
University of Lübeck, Lübeck, Germany

This genome assembly paper for *Idaea biselata* (Small Fan-footed Wave moth) presents a high-quality chromosomal-level reference genome as part of the Darwin Tree of Life project.

Strengths:

- **Excellent assembly quality**: Scaffold N50 of 31 Mb, QV of 57.4, and 97.24% scaffolded into 13 chromosomes meets rigorous EBP standards (5.C.Q57)
- **Comprehensive methodology**: Clear documentation of PacBio HiFi, Hi-C scaffolding, and manual curation with well-described protocols
- **Species-specific annotation**: RNA-seq data from the same specimen improves gene model accuracy (11,310 protein-coding genes)
- **Complete mitochondrial genome**: 17.33 kb mitogenome adds value for phylogenetic studies
- **Strong BUSCO scores**: 94.6% completeness validates assembly quality

Limitations:

- **Limited biological context**: As reviewer Dey notes, the paper would benefit from discussing how this genome advances geometrid moth biology, adaptation, or phylogenomics
- **Minor inconsistencies**: Reviewer Huang identifies scaffold count discrepancies (380 vs 381) between text and figures
- **Missing validation details**:
  - RNA-seq mapping rates to validate gene models not provided
  - Mitogenome circularity not explicitly confirmed
  - Software table incomplete (JBrowse2, Ensembl Genebuild missing)

Overall Assessment:
This is a technically sound resource paper with robust genomic data that fills an important gap for Geometridae genomes. The assembly quality is excellent and methodology reproducible. However, it reads as purely descriptive without contextualizing the genome's utility for addressing specific biological questions about this species or geometrid moths generally. For a data note, this is acceptable, but adding even 1-2 paragraphs on potential applications would significantly enhance impact.

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: Cancer genomics, Microbiome

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 21 January 2026

https://doi.org/10.21956/wellcomeopenres.26154.r139606

?

Yu-Feng Huang
[1] National Chung Hsing University, Taichung City, Taiwan
[2] Computer Science and Engineering, Yuan Ze University (Ringgold ID: 34895), Zhongli District, Taoyuan City, Taiwan

This Data Note presents a high-quality reference genome assembly for the Small Fan-footed Wave moth, *Idaea biselata*, generated using PacBio HiFi and Illumina Hi-C data. The assembly metrics are robust, with a scaffold N50 of 31.0 Mb and a consensus quality (QV) of 57.4. The authors have successfully assigned 97.24% of the sequence to 13 chromosomal pseudomolecules, representing the 12 autosomes and the Z sex chromosome. Notably, this study incorporates species-specific RNA-Seq data to support genome annotation, identifying 11,310 protein-coding genes.
While the data quality is high and meets Earth BioGenome Project standards, there are distinct inconsistencies regarding scaffold counts between the text and figures. Additionally, minor clarifications regarding mitogenome circularization, annotation statistics, and software documentation are required to ensure the manuscript's precision.
Major Comments
1. Discrepancy in Scaffold Counts There is a clear numerical inconsistency regarding the total number of scaffolds in the assembly. The main text, and Table 2 explicitly state that the final assembly consists of **380** sequence scaffolds. However, the legend for Figure 2 , the header for Figure 3 , and the legend for Figure 4 all list the total scaffold count as **381**. Is it possible that the

inclusion of mitochondrial genome?

Minor Comments

2. Gene Annotation Validation & Metrics This study commendably includes species-specific RNA-Seq data (ERR12245555) derived from abdomen tissue. The annotation methods state this transcriptomic data was used for alignment. While the methodology is sounder than protein-homology alone, the manuscript lacks specific metrics on the utility of this RNA-Seq data. To validate the annotation quality, please briefly include the mapping rate of the RNA-Seq reads to the final assembly. High mapping rates would empirically confirm the structural accuracy of the gene models generated.

3. Mitogenome Circularization Verification The mitochondrial genome (17.33 kb) was assembled using MitoHiFi, which the text states ensures "general quality". As noted in reviews of similar manuscripts, the text does not explicitly confirm that the assembly was verified as circular. Please confirm if the circular topology was validated (e.g., by checking for overlapping ends) and if the sequence was rotated to a standard start feature (e.g., *tRNA-Phe* or *COX1*) to ensure it represents the complete organelle genome.

4. Software Tools (Table 4) Table 4 lists the software versions used. There are minor omissions compared to the methods section. **JBrowse2**, mentioned in the curation section , and the **Ensembl Genebuild** system, mentioned in the annotation section, are not listed in the software table. Please audit Table 4 and add these missing tools with their respective versions/sources.

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

Yes

Are the protocols appropriate and is the work technically sound?

Partly

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

Yes

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Cancer Bioinformatics, Mitogenomics, Genome sequencing, assembly and gene annotation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 10 January 2026

https://doi.org/10.21956/wellcomeopenres.26154.r140289

Pritha Dey
National Centre for Biological Sciences-TIFR, Bengaluru, India

This manuscript presents a chromosomal-scale genome assembly of Idaea biselata (Lepidoptera: Geometridae), generated from a single male individual. The methods are clear and the inclusion of mitochondrial genome further enhances the utility of the dataset

The availability of a reference genome for I. biselata is a valuable contribution, particularly given the underrepresentation of Geometridae in genomic resources, and will be useful for studies in lepidopteran evolution, ecology, and comparative genomics.

However, the manuscript could be strengthened by placing the genome in a broader biological context. A short discussion on how this genome resource may advance studies on geometrid moth biology, adaptation, or phylogenomics would enhance its impact.

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.