DATA NOTE

The genome sequence of the Small Argent and Sable moth, *Epirrhoe tristata* (Linnaeus, 1758)

[version 1; peer review: 2 approved]

Marc Botham[1],
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]UK Centre for Ecology & Hydrology, Wallingford, England, UK

v1 **First published:** 20 Sep 2024, **9**:541
https://doi.org/10.12688/wellcomeopenres.23055.1
Latest published: 20 Sep 2024, **9**:541
https://doi.org/10.12688/wellcomeopenres.23055.1

Abstract

We present a genome assembly from an individual male Small Argent and Sable moth *Epirrhoe tristata* (Arthropoda; Insecta; Lepidoptera; Geometridae). The genome sequence spans 313.80 megabases. Most of the assembly is scaffolded into 30 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled and is 16.92 kilobases in length. Gene annotation of this assembly on Ensembl identified 16,469 protein-coding genes.

Keywords

Epirrhoe tristata, Small Argent and Sable moth, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 20 Sep 2024	✓ view	✓ view

1. **Erin Toffelmier**, University of California, Los Angeles, USA

2. **Alba Almazán**, Universitat de Barcelona Facultat de Biologia (Ringgold ID: 98709), Barcelona, Spain
Universitat de Barcelona Institut de Recerca de la Biodiversitat (Ringgold ID: 535620), Barcelona, Spain
Sophie Tandonnet, Universitat de Barcelona Institut de Recerca de la Biodiversitat (Ringgold ID: 535620), Barcelona, Spain
Universitat de Barcelona Facultat de Biologia (Ringgold ID: 98709), Barcelona, Spain

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: Botham M: 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: Botham M, Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team, Wellcome Sanger Institute Scientific Operations: Sequencing Operations *et al.* **The genome sequence of the Small Argent and Sable moth, *Epirrhoe tristata* (Linnaeus, 1758) [version 1; peer review: 2 approved]** Wellcome Open Research 2024, **9**:541 https://doi.org/10.12688/wellcomeopenres.23055.1

First published: 20 Sep 2024, **9**:541 https://doi.org/10.12688/wellcomeopenres.23055.1

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; Larentiinae; *Epirrhoe*; *Epirrhoe tristata* (Linnaeus, 1758) (NCBI:txid934838).

Background

Epirrhoe tristata, the Small Argent And Sable, is a moth of the genus Epirrhoe in the family Geometridae. The genome of *E. tristata* 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 chromosomally complete genome sequence for *Epirrhoe tristata*, based on a male specimen from Glen Strathfarrar, Scotland, UK.

Genome sequence report

The genome of an adult male *Epirrhoe tristata* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of 27.05 Gb (gigabases) from 2.34 million reads, providing approximately 84-fold coverage. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data, which produced 104.25 Gbp from 690.41 million reads, yielding an approximate coverage of 332-fold. Specimen and sequencing information is summarised in Table 1.

Manual assembly curation corrected four missing joins or misjoins, reducing the scaffold number by 2.94%. The final assembly has a total length of 313.80 Mb in 32 sequence scaffolds with a scaffold N50 of 11.3 Mb (Table 2). The total count of gaps in the scaffolds is 44. The snail plot in Figure 2 provides a summary of the assembly statistics, while the distribution of assembly scaffolds on GC proportion and coverage is shown in Figure 3. The cumulative assembly plot in Figure 4 shows curves for subsets of scaffolds assigned to different phyla.

Figure 1. Photograph of *Epirrhoe tristata* by AfroBrazilian (not the specimen used for genome sequencing).

Most (99.97%) of the assembly sequence was assigned to 30 chromosomal-level scaffolds, representing 29 autosomes and the Z sex chromosome. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 3). Chromosome Z was assigned by alignment to *Xanthorhoe spadicearia* (GCA_947086425.1) (Boyes *et al.*, 2024). 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.

The estimated Quality Value (QV) of the final assembly is 69.4 with *k*-mer completeness of 100.0%, and the assembly has a BUSCO v5.3.2 completeness of 98.4% (single = 98.0%, duplicated = 0.4%), using the lepidoptera_odb10 reference set (n = 5,286).

Metadata for specimens, BOLD barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly statistics are given at https://links.tol.sanger.ac.uk/species/934838.

Genome annotation report

The *Epirrhoe tristata* genome assembly (GCA_951394285.1) was annotated at the European Bioinformatics Institute (EBI) on Ensembl Rapid Release. The resulting annotation includes 16,639 transcribed mRNAs from 16,469 protein-coding] genes (Table 2; https://rapid.ensembl.org/Epirrhoe_tristata_GCA_951394285.1/Info/Index). The average transcript length is 6,435.52. There are 1.01 coding transcripts per gene and 6.00 exons per transcript.

Methods

Sample acquisition

An adult male *Epirrhoe tristata* (specimen ID SAN00002622, ToLID ilEpiTris1) was collected from Glen Strathfarrar, Scotland, UK (latitude 57.41, longitude –4.73) on 2022-06-27 using a moth trap. The specimen was collected and identified by Marc Botham (UK Centre for Ecology & Hydrology) and preserved by flash freezing.

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 core procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io (Denton *et al.*, 2023b). In sample preparation, the ilEpiChri1 sample was weighed and dissected 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

Table 1. Specimen and sequencing data for *Epirrhoe tristata*.

Project information			
Study title	*Epirrhoe tristata* (small argent and sable)		
Umbrella BioProject	PRJEB61371		
Species	*Epirrhoe tristata*		
BioSample	SAMEA112198543		
NCBI taxonomy ID	934838		
Specimen information			
Technology	**ToLID**	**BioSample accession**	**Organism part**
PacBio long read sequencing	ilEpiTris1	SAMEA112198599	thorax
Hi-C sequencing	ilEpiTris1	SAMEA112198599	thorax
RNA sequencing	ilEpiTris1	SAMEA112198600	abdomen
Sequencing information			
Platform	**Run accession**	**Read count**	**Base count (Gb)**
Hi-C Illumina NovaSeq 6000	ERR11242569	6.90e+08	104.25
PacBio Sequel IIe	ERR11242145	2.34e+06	27.05
RNA Illumina NovaSeq 6000	ERR11837486	5.60e+07	8.46

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 ilEpiTris1 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.

Library preparation and sequencing

Pacific Biosciences HiFi circular consensus DNA sequencing libraries were constructed according to the manufacturers' instructions. Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit. DNA and RNA sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences Sequel IIe (HiFi) and Illumina NovaSeq 6000 (RNA-Seq) instruments.

Hi-C data were generated from frozen thorax tissue of the ilEpiTris1 sample, using the Arima-HiC v2 kit. The tissue was fixed with a TC buffer containing formaldehyde, resulting in crosslinked DNA. The crosslinked DNA was digested with a restriction enzyme master mix. The resulting 5'-overhangs were filled in and labelled with a biotinylated nucleotide. The biotinylated DNA was then fragmented, enriched, barcoded, and amplified using the NEBNext Ultra II DNA Library Prep Kit. Hi-C sequencing was performed on an Illumina NovaSeq 6000 instrument, using paired-end sequencing with a read length of 150 bp.

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. 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.

Table 2. Genome assembly data for *Epirrhoe tristata*, ilEpiTris1.1.

Genome assembly		
Assembly name	ilEpiTris1.1	
Assembly accession	GCA_951394285.1	
Accession of alternate haplotype	*GCA_951394275.1*	
Span (Mb)	313.80	
Number of contigs	77	
Contig N50 length (Mb)	6.5	
Number of scaffolds	32	
Scaffold N50 length (Mb)	11.3	
Longest scaffold (Mb)	17.63	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	69.4	*≥ 50*
k-mer completeness	100.0%	*≥ 95%*
BUSCO**	C:98.4%[S:98.0%,D:0.4%], F:0.4%,M:1.2%,n:5,286	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	99.97%	*≥ 95%*
Sex chromosomes	Z	*localised homologous pairs*
Organelles	Mitochondrial genome: 16.92 kb	*complete single alleles*
Genome annotation of assembly GCA_951394285.1 at Ensembl		
Number of protein-coding genes	16,469	
Number of gene transcripts	16,639	

* Assembly metric benchmarks are adapted from column VGP-2020 of "Table 1: Proposed standards and metrics for defining genome assembly quality" from Rhie *et al.* (2021).

** BUSCO scores based on the lepidoptera_odb10 BUSCO set using version 5.3.2. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/ilEpiTris1_1/dataset/ilEpiTris1_1/busco.

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 (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. Sex chromosomes were identified by synteny analysis. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation (article in preparation).

Evaluation of the final assembly

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). To assess the assembly metrics, the *k*-mer completeness and QV consensus quality values were calculated in Merqury (Rhie *et al.*, 2020). This work was done using the "sanger-tol/readmapping" (Surana *et al.*, 2023a) and "sanger-tol/genomenote" (Surana *et al.*, 2023b) pipelines. The genome readmapping pipelines were developed using the nf-core tooling (Ewels *et al.*, 2020), use MultiQC (Ewels *et al.*, 2016), and make extensive use of the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), and the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions. 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.

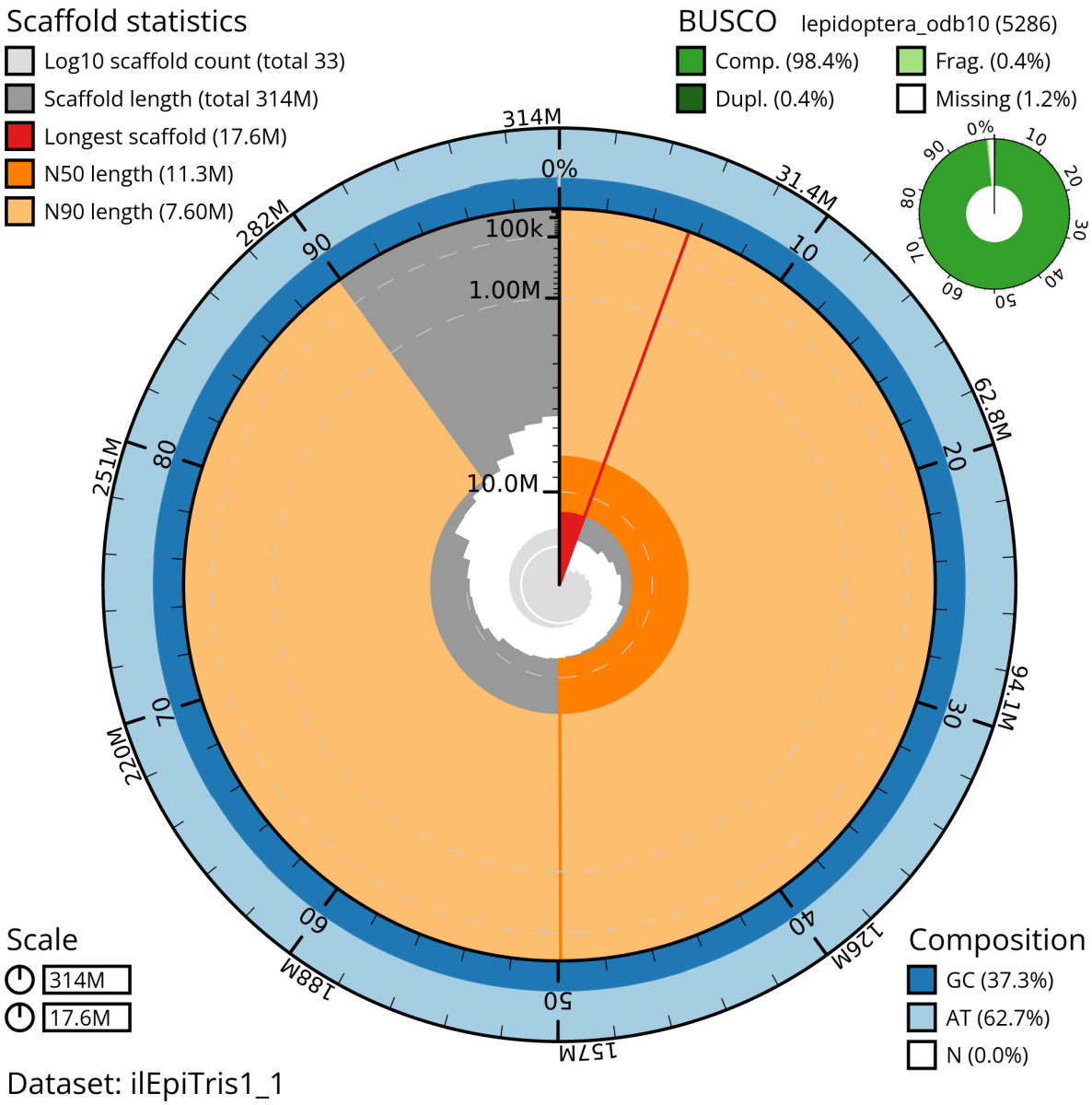

Figure 2. Genome assembly of *Epirrhoe tristata*, ilEpiTris1.1: metrics. The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 313,821,882 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (17,631,079 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (11,321,105 and 7,599,468 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the lepidoptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilEpiTris1_1/dataset/ilEpiTris1_1/snail.

Genome annotation

The BRAKER2 pipeline (Brůna *et al.*, 2021) was used in the default protein mode to generate annotation for the *Epirrhoe tristata* assembly (GCA_951394285.1) in Ensembl Rapid Release at the EBI.

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'**,

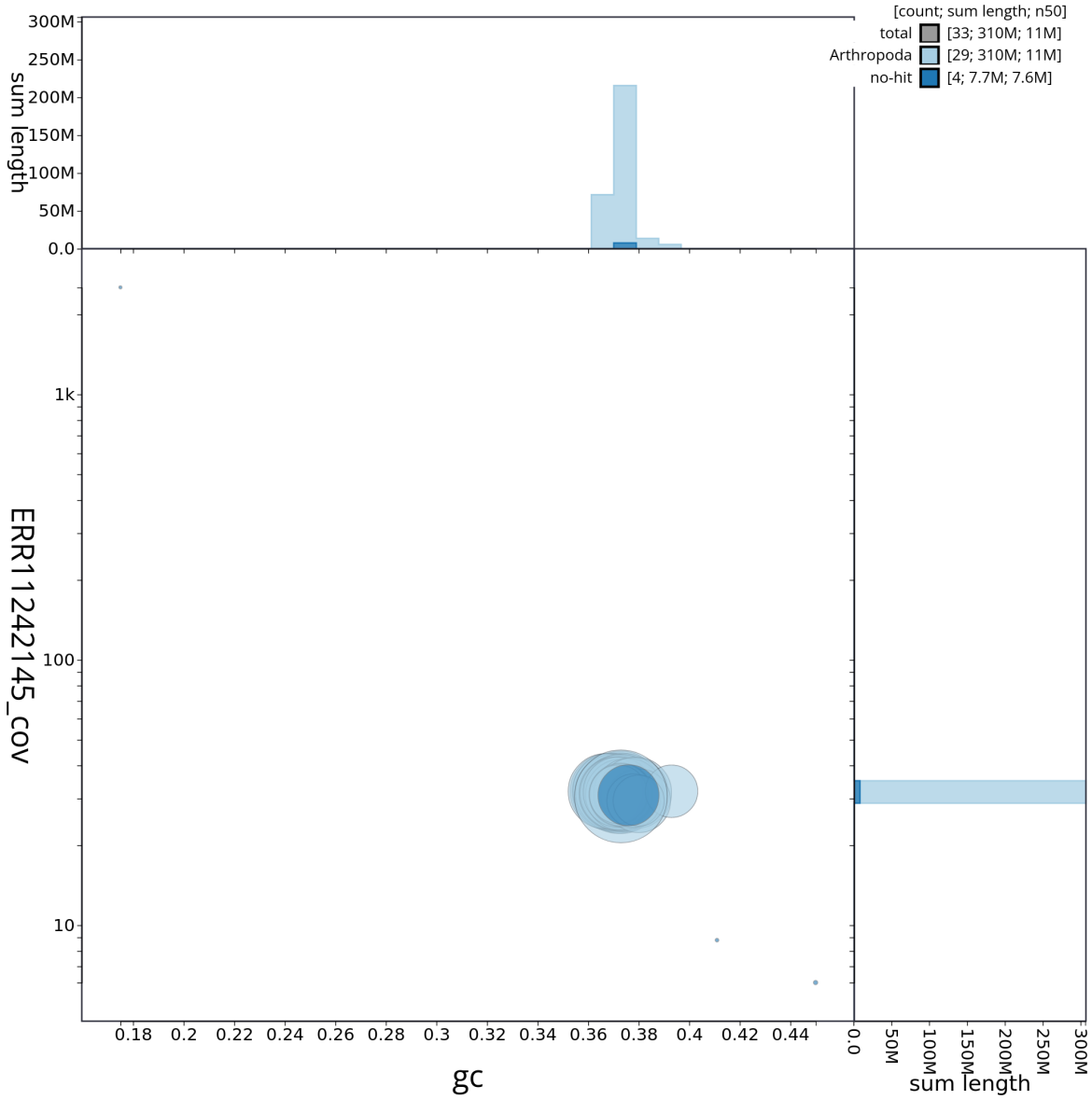

Figure 3. Genome assembly of *Epirrhoe tristata*, ilEpiTris1.1: BlobToolKit GC-coverage plot. Sequences are coloured by phylum. Circles are sized in proportion to sequence length. Histograms show the distribution of sequence length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilEpiTris1_1/dataset/ilEpiTris1_1/blob.

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 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)

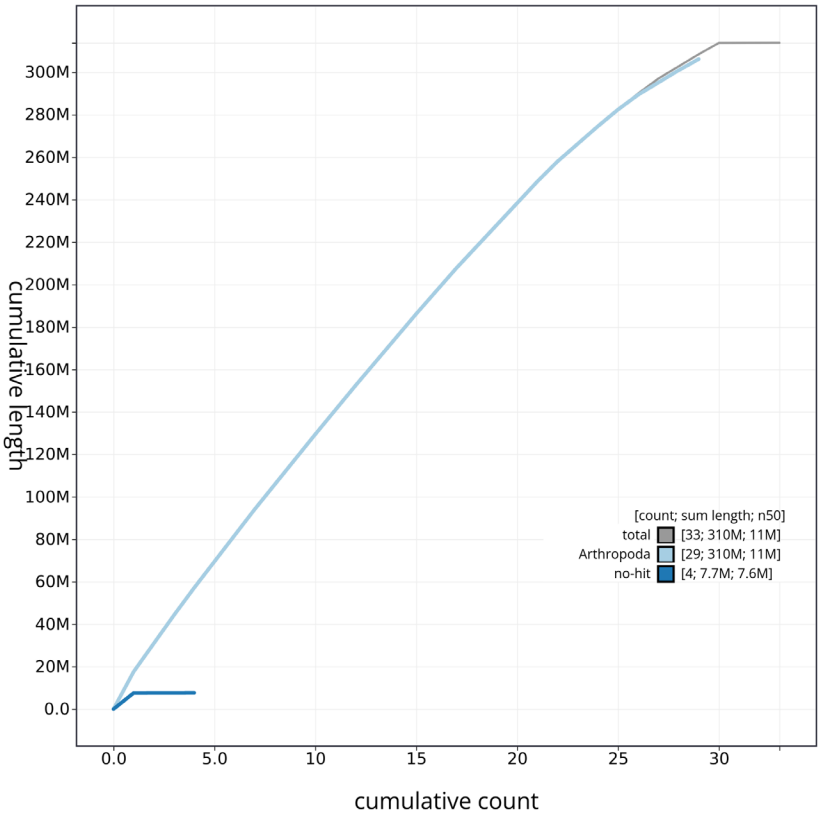

Figure 4. Genome assembly of *Epirrhoe tristata* ilEpiTris1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all sequences. Coloured lines show cumulative lengths of sequences assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilEpiTris1_1/dataset/ilEpiTris1_1/cumulative.

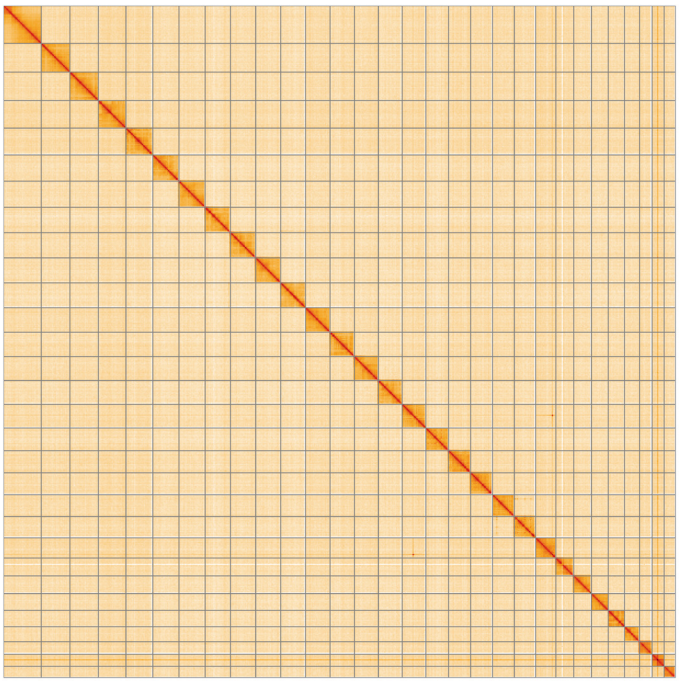

Figure 5. Genome assembly of *Epirrhoe tristata* ilEpiTris1.1: Hi-C contact map of the ilEpiTris1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=Wq3YluJmRzynlqosrwyeaA.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Epirrhoe tristata*, ilEpiTris1.

INSDC accession	Name	Length (Mb)	GC%
OX596137.1	1	13.35	37.0
OX596138.1	2	13.3	37.5
OX596139.1	3	12.85	37.5
OX596140.1	4	12.52	37.5
OX596141.1	5	12.24	37.0
OX596142.1	6	12.23	37.0
OX596143.1	7	11.83	37.0
OX596144.1	8	11.79	37.0
OX596145.1	9	11.69	37.0
OX596146.1	10	11.58	36.5
OX596147.1	11	11.46	37.5
OX596148.1	12	11.32	37.0
OX596149.1	13	11.24	37.0
OX596150.1	14	11.1	37.0
OX596151.1	15	11.04	37.5
OX596152.1	16	10.61	37.0
OX596153.1	17	10.31	38.0
OX596154.1	18	10.26	37.5
OX596155.1	19	10.14	37.0
OX596156.1	20	9.95	37.5
OX596157.1	21	9.43	38.0
OX596158.1	22	8.32	38.0
OX596159.1	23	8.26	37.5
OX596160.1	24	7.89	37.5
OX596161.1	25	7.6	37.5
OX596162.1	26	7.0	37.0
OX596163.1	27	5.85	38.0
OX596164.1	28	5.62	39.5
OX596165.1	29	5.31	38.0
OX596136.1	Z	17.63	37.5
OX596166.1	MT	0.02	17.5

Table 4. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	4.2.1	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.3.2	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
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Hifiasm	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	1.11.6	https://github.com/higlass/higlass
Merqury.FK	d00d98157618f4e8d1a9190026b19b47 1055b22e	https://github.com/thegenemyers/MERQURY.FK
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
PretextView	0.2	https://github.com/wtsi-hpag/PretextView
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
sanger-tol/ genomenote	v1.0	https://github.com/sanger-tol/genomenote
sanger-tol/ readmapping	1.1.0	https://github.com/sanger-tol/readmapping/ tree/1.1.0
Singularity	3.9.0	https://github.com/sylabs/singularity
YaHS	1.2a.2	https://github.com/c-zhou/yahs

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: *Epirrhoe tristata* (small argent and sable). Accession number PRJEB61371; https://identifiers.org/ena.embl/PRJEB61371 (Wellcome Sanger Institute, 2023). The genome sequence is released openly for reuse. The *Epirrhoe tristata* 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 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.* Oxford University Press, 2020; **36**(1): 311–316.
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.* Blackwell Publishing Ltd, 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

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

Boyes D, Lewis OT, University of Oxford and Wytham Woods Genome Acquisition Lab, *et al.*: **The genome sequence of the Red Twin-spot Carpet, *Xanthorhoe spadicearia* (Denis & Schiffermüller, 1775) [version 1; peer review: 1 approved, 1 approved with reservations].** *Wellcome Open Res.* 2024; **9**: 68.
Publisher Full Text

Brůna T, Hoff KJ, Lomsadze A, *et al.*: **BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database.** *NAR Genom Bioinform.* Oxford University Press, 2021; **3**(1): lqaa108.
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).* Genetics Society of America, 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.** *Nature Methods.* Nature Research, 2021; **18**(2): 170–175.
PubMed Abstract | Publisher Full Text | Free Full Text

da Veiga Leprevost F, Grüning BA, Alves Aflitos S, *et al.*: **BioContainers: an open-source and community-driven framework for software standardization.** *Bioinformatics.* 2017; **33**(16): 2580–2582.
PubMed Abstract | Publisher Full Text | Free Full Text

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, Bates A, Denton A, *et al.*: **Sanger Tree of Life RNA extraction: automated MagMax™ mirVana.** *protocols.io.* 2023.
Publisher 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.** *Nature 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

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
PubMed Abstract | Publisher Full Text | Free Full Text

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* Oxford University Press, 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.* Oxford University Press, 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.* BioMed Central Ltd, 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**

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

Oatley G, Denton A, Howard C: **Sanger Tree of Life HMW DNA extraction: automated MagAttract v.2.** *protocols.io.* 2023.
Publisher 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.* Cell Press, 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.* Nature Research, 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.* BioMed Central Ltd, 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

Surana P, Muffato M, Qi G: **sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo.* 2023a.
Publisher Full Text

Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev).** *Zenodo.* 2023b.
Publisher 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**. *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.
Publisher Full Text

Wellcome Sanger Institute: **The genome sequence of the Small Argent and Sable moth, *Epirrhoe tristata* (Linnaeus, 1758).** European Nucleotide Archive. [dataset], accession number PRJEB61371, 2023.

Zhou C, McCarthy SA, Durbin R: **YaHS: yet another Hi-C scaffolding tool.** edited by Alkan, C. *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 September 2025

https://doi.org/10.21956/wellcomeopenres.25391.r132276

✔ **Alba Almazán**
[1] Departament de Genetica Microbiologia i Estadistica, Universitat de Barcelona Facultat de Biologia (Ringgold ID: 98709), Barcelona, Catalonia, Spain
[2] Universitat de Barcelona Institut de Recerca de la Biodiversitat (Ringgold ID: 535620), Barcelona, Catalonia, Spain

Sophie Tandonnet
[1] Universitat de Barcelona Institut de Recerca de la Biodiversitat (Ringgold ID: 535620), Barcelona, Catalonia, Spain
[2] Departament de Genetica Microbiologia i Estadistica, Universitat de Barcelona Facultat de Biologia (Ringgold ID: 98709), Barcelona, Catalonia, Spain

Summary:
In this genome note, the authors report the chromosomal-scale assembly of the genome of *Epirrhoe tristata,* the Small Argent and Sable moth. The assembly originates from a single male (the homogametic sex) and contains 30 pseudo-chromosomal molecules (including the Z chromosome) and 2 additional scaffolds.
The assembly is of high quality (QV 69.4), completeness (98.4% complete BUSCOs, 0% Ns), highly continuous (most of the genome in chromosome-scale scaffolds) and clean (almost no contamination). The assembly is based on Pacbio HiFi long reads and scaffolded into pseudo-chromosomal molecules using Hi-C data. The same individual (ilEpiTris1) was used for the Pacbio HiFi sequencing (thorax), the HiC (thorax) and the RNA-seq (abdomen), which is a strength of this work.
The assembly has a corresponding annotation, which we could download without problem. This annotation includes 16,469 genes. The non-coding genes were not annotated.
The paper is straightforward and easy to follow.
All in all, this note reports a new, high-quality and high-continuity genome assembly of a species not yet sequenced, along with its corresponding annotation.

Main Comments:
- How was the specimen's sex determined? Maybe add this to the report.

- Is the picture of the specimen sequenced available? If so, it could be included.
- Figure 5: it would be useful to indicate which scaffold corresponds to the Z chromosome.
- The authors report 29 autosomes + 1 Z chromosome + 2 other scaffolds.. Table 3 shows 29 autosomes, Z, and the mitochondrial genome, but that adds up to 31. Could the authors clarify what the 32nd scaffold is? If it is an unplaced nuclear scaffold, please list it explicitly in Table 3 for clarity.
- Methods, Nucleic acid extraction: The authors do not explain how they separated the thorax material for HMW DNA extraction and for HiC. Please add those details
- Methods, library preparation: which restriction enzyme was used?
- Methods, Assembly curation: the authors state that they identified the Z chromosome by synteny (to Xanthorhoe spadicearia, as stated in the report section): How conserved is the macrosynteny between E. tristata and X. spadicearia? Is there a clear 1 to 1 chromosomal correspondence? (there must some difference as the chromosome number of both species doesn't seem to be the same). Maybe the authors could show the pattern of synteny

Minor comments

- Background: this section is extremely short; do we know the range of this species? Its basic ecology/lifestyle?
- It is usually better to sequence the heterogametic sex, although we understand it is not possible to control the individuals you collect.
- The genome annotation availability link (https://rapid.ensembl.org/Epirrhoe_tristata_GCA_951394285.1/Info/Index) redirects to: https://beta.ensembl.org/species/81c60124-c548-4d22-96d6-bc1351994b2c
- As far as we could check, all the other links seem to work.
- The page https://tolqc.cog.sanger.ac.uk/darwin/insects/Epirrhoe_tristata/, compiles metadata and sequencing runs. All seems correct, however, why is there a row with "**OTHER_SOMATIC_ANIMAL_TISSUE**" in the specimen table.
- Methods: Nucleic acid extraction - " In sample preparation, the ilEpiChri1 sample..." double-check the sample code. Elsewhere it is consistently ilEpiTris1

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?

Partly

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolutionary biology, Bioinformatics, Non-model organisms, Developmental Biology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 29 September 2025

https://doi.org/10.21956/wellcomeopenres.25391.r132277

Erin Toffelmier
University of California, Los Angeles, USA

This article describes a genome assembly for *Epirrhoe tristate* as part of the Darwin Tree of Life Project. The article clearly describes the methods and the resulting high quality genome assembly.

I suggest adding assembly metrics for haplotype 2, even if it's less complete. It would also be informative to know the expected genome size and karyotype. While this is a very tidy dataset showing strong chromosome-scale scaffolding, with very few unplaced fragments, it would be beneficial to see supporting karyotypic evidence, if available. Finally, consider adding a brief rationale for using *Xanthorhoe spadicearia* in the Z-chromosome synteny analysis (i.e. it's in the same family).

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, population genomics, ecology

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.