DATA NOTE

The genome sequence of the Reed Leopard moth, *Phragmataecia castaneae* Hübner, 1790 (Lepidoptera: Cossidae)

[version 1; peer review: 2 approved]

Mick Acourt[1], 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]RSPB Sutton Fen, Sutton, Norfolk, England, UK

v1 **First published:** 20 Nov 2025, **10**:639
https://doi.org/10.12688/wellcomeopenres.25089.1
Latest published: 20 Nov 2025, **10**:639
https://doi.org/10.12688/wellcomeopenres.25089.1

Abstract

We present a genome assembly from an individual male *Phragmataecia castaneae* (Reed Leopard moth; Arthropoda; Insecta; Lepidoptera; Cossidae). The assembly contains two haplotypes with total lengths of 590.72 megabases and 591.65 megabases. Most of haplotype 1 (99.8%) is scaffolded into 30 chromosomal pseudomolecules, including the Z sex chromosome. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 15.5 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

Phragmataecia castaneae; Reed Leopard 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 Nov 2025	✓ view	✓ view

1. **Lorena Layana**, Institute of Science and Technology Austria (Ringgold ID: 148492), Klosterneuburg, Austria

2. **Sophie Tandonnet**, Universitat de Barcelona Institut de Recerca de la Biodiversitat (Ringgold ID: 535620), Barcelona, Spain
Tòt Senar Serra, Universitat de Barcelona Departament de Genetica Microbiologia i Estadistica (Ringgold ID: 200148), Barcelona, Spain
Universitat de Barcelona Institut de Recerca de la Biodiversitat (Ringgold ID: 535620),

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: Acourt M: Investigation, Resources, Writing – Original Draft Preparation;

Competing interests: No competing interests were disclosed.

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

How to cite this article: Acourt M, Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding Collective *et al.* **The genome sequence of the Reed Leopard moth, *Phragmataecia castaneae* Hübner, 1790 (Lepidoptera: Cossidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, **10**:639 https://doi.org/10.12688/wellcomeopenres.25089.1

First published: 20 Nov 2025, **10**:639 https://doi.org/10.12688/wellcomeopenres.25089.1

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; Apoditrysia; Cossoidea; Cossidae; Zeuzerinae; *Phragmataecia*; *Phragmataecia castaneae* Hübner, 1790 (NCBI:txid753407)

Background

We present a chromosome-level genome sequence for *Phragmataecia castaneae*, the Reed Leopard. This moth is light grey-brown. The wings are elongate with rounded tips and scattered pale spots. The abdomen is long and extends beyond the wing tips when the moth is at rest. Males and females look similar, although females are slightly larger (Heath & Emmet, 1985; Kimber, 2025). The adults fly from June to July. *P. castaneae* larvae live inside stems of common reed, *Phragmites australis* below water level for up to 21 months, and pupation occurs within the reed stem (Heath & Emmet, 1985; Kimber, 2025).

In the UK *P. castaneae* is local and scarce. It is recorded mainly from parts of Cambridgeshire and the Norfolk Broads, and a small population in Dorset, which has not been recorded since 2023 (Kimber, 2025; Wheeler, 2025). Its wider range extends from Spain to Japan and from North Africa to southern Scandinavia. It also occurs in England, Corsica, Sardinia, Sicily and Crete (Heath & Emmet, 1985; The Butterfly Foundation, 2025). On the IUCN Red List it is reported as Near Threatened, and in Great Britain it is listed as Nationally Rare (Wheeler, 2025).

The genome assembly was produced using the Tree of Life pipeline from a specimen collected in Sutton North Fen, England, UK (Figure 1). This assembly is the first high-quality genome for the genus *Phragmataecia* and one of four genomes available for the family Cossidae as of October 2025 (data obtained via NCBI datasets, O'Leary *et al.*, 2024). This assembly is the RefSeq reference assembly for this species.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Phragmataecia castaneae* (specimen ID NHMUK014584888, ToLID ilPhrCast1; Figure 1), collected from Sutton North Fen, England, United Kingdom (latitude 52.4, longitude 0.11) on 2022-07-05. The specimen was collected and identified by Mick Acourt. 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.

Figure 1. Photograph of the *Phragmataecia castaneae* (ilPhrCast1) specimen used for genome sequencing.

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 ilPhrCast1 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. DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor®3 for LI PacBio protocol. Sheared DNA was purified by automated SPRI (solid-phase reversible immobilisation). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 73 ng/μL and a yield of 3 431.00 ng, with a fragment size of 13.6 kb.

PacBio HiFi library preparation and sequencing

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

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

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the ilPhrCast1 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

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

Prior to sequencing, libraries were normalised to 10 ng/µL. Normalised libraries were quantified again to create equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq 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 in Hi-C phasing mode (Cheng *et al.*, 2021; Cheng *et al.*, 2022), producing two haplotypes. Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). Contigs were further scaffolded with Hi-C data 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. 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 41 breaks and 72 joins, reducing the scaffold count by 29.0%. 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 both 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 *Phragmataecia castaneae* specimen generated 32.30 Gb (gigabases) from 3.33 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 599.17 Mb, with a heterozygosity of 0.36% and repeat content of 33.29% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 53× coverage. Hi-C sequencing produced 151.79 Gb from 1 005.26 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The genome was assembled into two haplotypes using Hi-C phasing. Haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The final assembly has a total length of 590.72 Mb in 65 scaffolds, with 152 gaps, and a scaffold N50 of 20.82 Mb (Table 2).

Most of the assembly sequence (99.8%) 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 3; Table 3). The sex chromosome Z was assigned by homology to the genome of *Zeuzera pyrina* (GCA_907165235.1) (Boyes *et al.*, 2023).

The mitochondrial genome was also assembled (length 15.5 kb, OZ195577.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

For haplotype 1, the estimated QV is 63.9, and for haplotype 2, 63.7. When the two haplotypes are combined, the assembly achieves an estimated QV of 63.8. The *k*-mer completeness is 90.29% for haplotype 1, 90.37% for haplotype 2, and 99.44% for the combined haplotypes (Figure 4).

BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.4% of the expected gene set (single = 97.7%, duplicated = 0.7%) for haplotype 1. The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

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

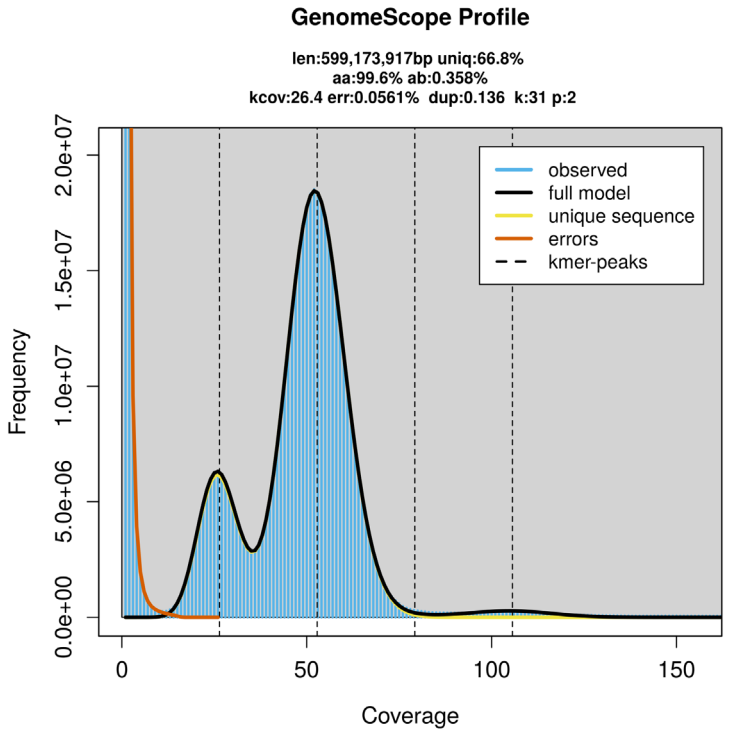

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. **Specimen and sequencing data for BioProject PRJEB81188.**

Platform	PacBio HiFi	Hi-C
ToLID	ilPhrCast1	ilPhrCast1
Specimen ID	NHMUK014584888	NHMUK014584888
BioSample (source individual)	SAMEA114805613	SAMEA114805613
BioSample (tissue)	SAMEA114805732	SAMEA114805732
Tissue	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13800655	ERR13802786
Read count total	3.33 million	1 005.26 million
Base count total	32.30 Gb	151.79 Gb

Table 2. **Genome assembly statistics.**

Metric	Haplotype 1	Haplotype 2
Assembly name	ilPhrCast1.hap1.1	ilPhrCast1.hap2.1
Assembly accession	GCA_964277475.1	GCA_964277415.1
Assembly level	chromosome	scaffold
Span (Mb)	590.72	591.65
Number of chromosomes	30	scaffold-level
Number of contigs	217	228
Contig N50	6.05 Mb	5.86 Mb
Number of scaffolds	65	56
Scaffold N50	20.82 Mb	20.85 Mb
Longest scaffold length (Mb)	34.25	-
Sex chromosomes	Z	-
Organelles	Mitochondrion: 15.5 kb	-

for the haplotype 1, is **6.C.Q63**, 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
- 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

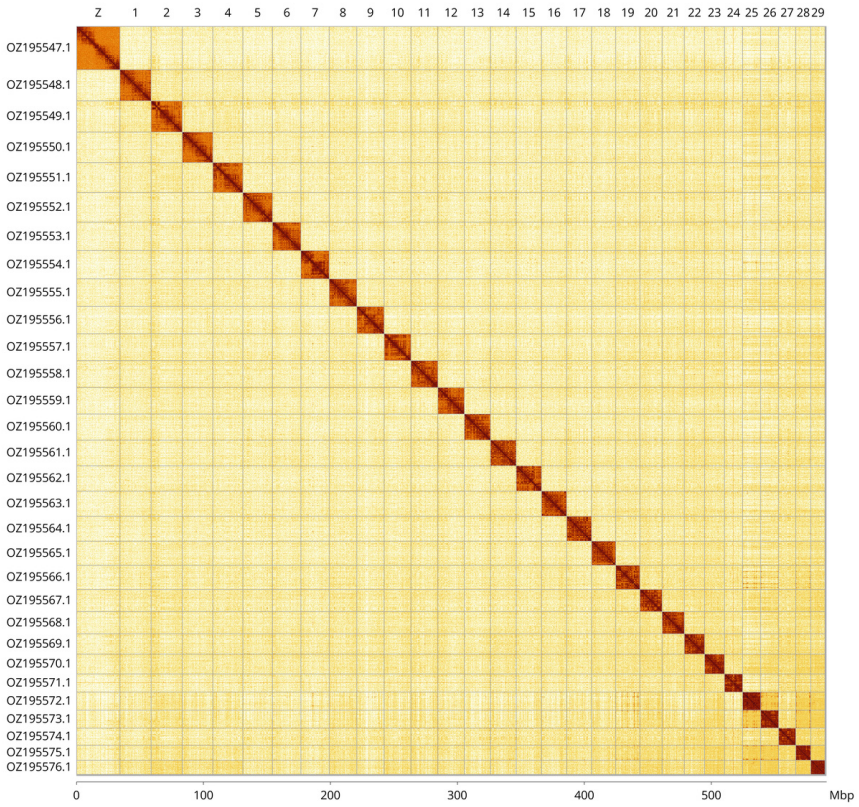

Figure 3. Hi-C contact map of the *Phragmataecia castaneae* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Phragmataecia castaneae* ilPhrCast1.

INSDC accession	Molecule	Length (Mb)	GC%
OZ195548.1	1	24.63	35
OZ195549.1	2	24.44	35
OZ195550.1	3	24.21	35
OZ195551.1	4	23.56	35
OZ195552.1	5	23.30	34.50
OZ195553.1	6	22.49	34.50
OZ195554.1	7	22.21	34.50
OZ195555.1	8	21.73	35
OZ195556.1	9	21.46	34.50
OZ195557.1	10	21.22	35
OZ195558.1	11	21.16	35
OZ195559.1	12	20.82	35
OZ195560.1	13	20.65	34.50
OZ195561.1	14	20.18	35
OZ195562.1	15	20.04	35
OZ195563.1	16	19.80	35
OZ195564.1	17	19.61	35.50
OZ195565.1	18	19.02	35
OZ195566.1	19	18.99	35.50
OZ195567.1	20	17.57	35.50
OZ195568.1	21	17.50	35
OZ195569.1	22	15.94	35.50
OZ195570.1	23	15.58	35.50
OZ195571.1	24	14.35	35.50
OZ195572.1	25	14.24	37
OZ195573.1	26	14.09	36
OZ195574.1	27	13.44	35.50
OZ195575.1	28	11.86	36
OZ195576.1	29	11.18	36.50
OZ195547.1	Z	34.25	34.50

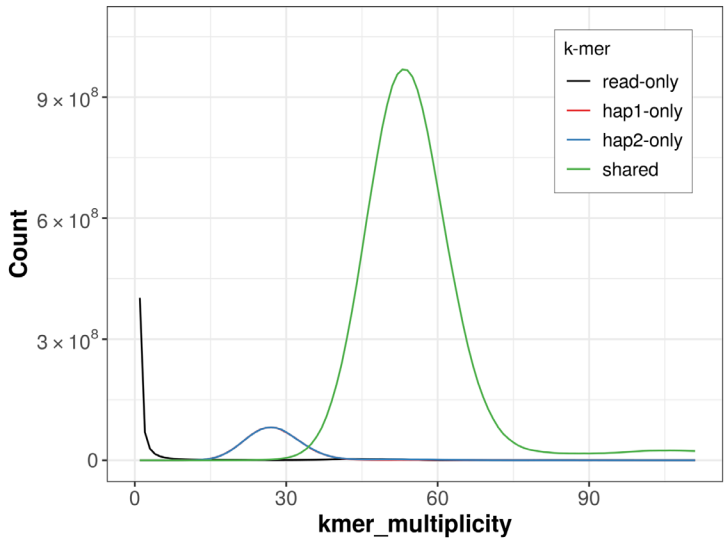

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

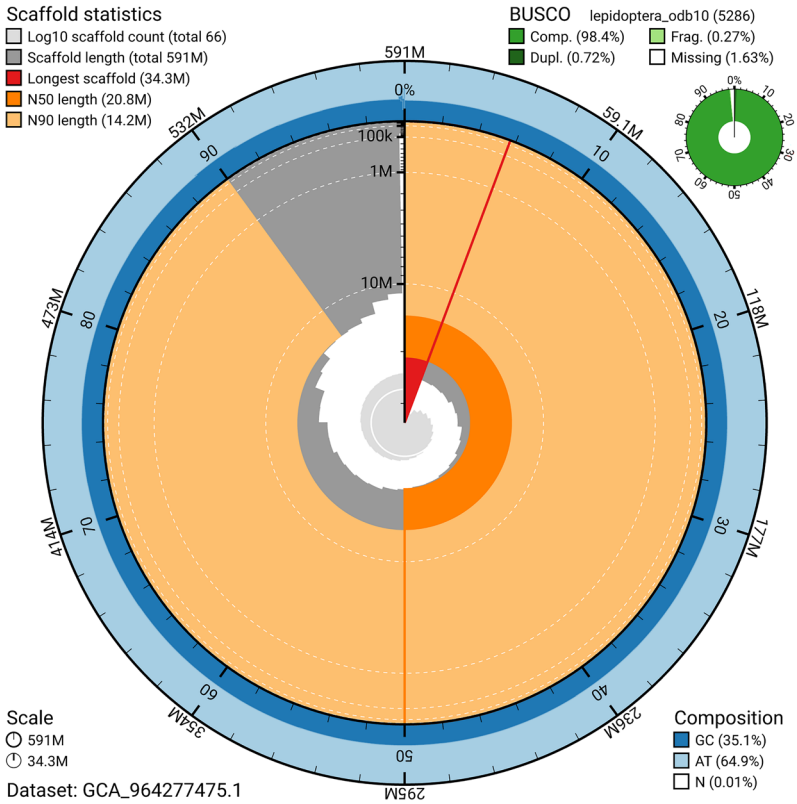

Figure 5. Assembly metrics for ilPhrCast1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost 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 set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

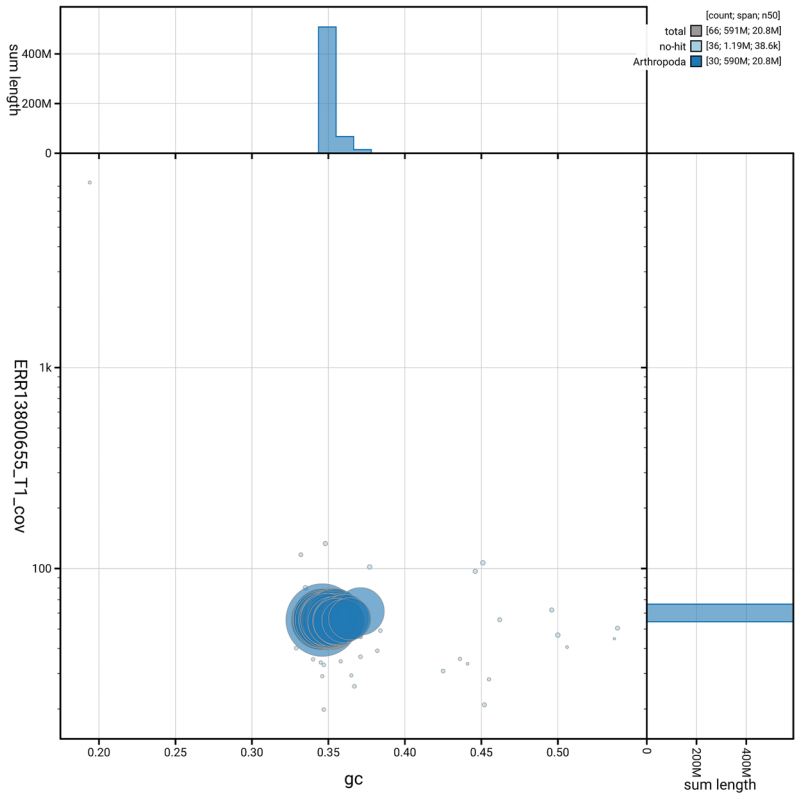

Figure 6. BlobToolKit GC-coverage plot for ilPhrCast1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Phragmataecia castaneae* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q63	6.C.Q40
Contig N50 length	6.05 Mb	≥ 1 Mb
Scaffold N50 length	20.82 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 63.9; haplotype 2: 63.7; combined: 63.8	≥ 40
k-mer completeness	Haplotype 1: 90.29%; Haplotype 2: 90.37%; combined: 99.44%	≥ 95%
BUSCO	C:98.4% [S:97.7%; D:0.7%]; F:0.3%; M:1.4%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.80%	≥ 90%

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: Phragmataecia castaneae. Accession number PRJEB81188. The genome sequence is released openly for reuse. The *Phragmataecia castaneae* 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 Natural History Museum Genome Acquisition Lab
- Members of the Darwin Tree of Life Barcoding collective
- Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team
- Members of Wellcome Sanger Institute Scientific Operations – Sequencing Operations
- Members of the Wellcome Sanger Institute Tree of Life Core Informatics team
- Members of the Tree of Life Core Informatics collective
- Members of the Darwin Tree of Life Consortium

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	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
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
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, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, *et al.*: **The genome sequence of the Leopard moth, *Zeuzera pyrina* (Linnaeus, 1761) [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 94.
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

Cheng H, Jarvis ED, Fedrigo O, *et al.*: **Haplotype-resolved assembly of diploid genomes without parental data.** *Nat Biotechnol.* 2022; **40**(9): 1332–1335.
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

Heath J, Emmet AM: **The moths and butterflies of Great Britain and Ireland, Volume 2.** Harley Books, 1985; **2**.

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, UK moths: ***Phragmataecia castaneae***. 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

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

The Butterfly Foundation: **Rietluipaard *Phragmataecia castaneae***. 2025.
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

Wheeler J, Norfolk moths: **Reed Leopard *Phragmataecia castaneae***. 2025.
Reference Source

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 07 January 2026

https://doi.org/10.21956/wellcomeopenres.27659.r141774

✔ **Sophie Tandonnet**
Universitat de Barcelona Institut de Recerca de la Biodiversitat (Ringgold ID: 535620), Barcelona, Catalonia, Spain
Tòt Senar Serra
[1] Universitat de Barcelona Departament de Genetica Microbiologia i Estadistica (Ringgold ID: 200148), Barcelona, Catalonia, Spain
[2] Universitat de Barcelona Institut de Recerca de la Biodiversitat (Ringgold ID: 535620), Barcelona, Catalonia, Spain

In this genome note, the authors report the chromosome-scale assembly for the genome of *Phragmataecia castaneae*, the Reed Leopard. The assembly originates from a single male (the homogametic sex) was assembled in two haplotypes. Haplotype 1 was curated to chromosome level and contains 30 pseudo-chromosomal molecules (including the Z chromosome) and 35 additional scaffolds. This haplotype is of high quality (QV 63.9), completeness (98.4% complete BUSCOs) and highly continuous (99.8% of the genome in chromosome-scale scaffolds). The assembly is based on Pacbio HiFi long reads and scaffolded into pseudo-chromosomal molecules using Hi-C data. The same individual (ilPhrCast1) was used for the Pacbio HiFi sequencing and the Hi-C. The assembly doesn't have an annotation yet.
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.

Main Comments

- It is unfortunate that the homogametic sex was sequenced instead of the heterogametic sex. The W chromosome is therefore missing and no coverage analysis can be performed to identify the Z chromosome.
- Methods, Nucleic acid extraction/HiC: The authors do not explain which part of the animal was used for the DNA extraction/Pacbio and which one was used for the HiC. For both the same sample seems to have been used and table 1 states that the whole body was used for both. Please revise this information.
- Methods, Assembly curation: the authors state that they identified the Z chromosome by synteny (to *Zeuzera pyrina*, as stated in the report section): How conserved is the

macrosynteny between Z. pyrina and *P. castanae*? Is there a clear 1 to 1 chromosomal correspondence? (there must be some difference as the chromosome number of both species doesn't seem to be the same). Maybe the authors could show the pattern of macrosynteny.
- Data availability: the authors don't provide a link/Accession number for the RNA-seq data.

Minor Comments
- Background: this section is somewhat telegraphic in style and would benefit from clearer exposition.
- Methods: The methods are well written and clear, however it would be difficult to reproduce the steps easily. For instance in "Assembly quality assessment", the authors write: "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." It is unclear if all this is manual. If so, maybe add the pipeline/code to Github and reference it.

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: Evo-Devo, sex determination, sex chromosome evolution, genomics, chromosome evolution

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 30 December 2025

https://doi.org/10.21956/wellcomeopenres.27659.r141053

Lorena Layana

Institute of Science and Technology Austria (Ringgold ID: 148492), Klosterneuburg, Lower Austria, Austria

In this article, the author Acourt presents a scientific report on the genome assembly of the Lepidoptera *Phragmataecia castaneae* moth. The author assembled genomes for two haplotypes of a male *Phragmataecia castaneae* building a database of k-mer counts. Within this context, the 65 scaffolds were assembled into 30 chromosome-level scaffolds, including one Z chromosome.

The background section it's a nice introduction to highlight a few details of the *Phragmataecia castaneae* moth life cycle and their geographical location and spread. In the last paragraph, it is also specified that the assembly was made following the Tree of Life pipeline. Does the author refer to the pipeline used for this specific genome?, or is it a general one used for all assemblies in the Tree of Life project? I think a short sentence could be added here, or maybe including a link where the reader can have an overview of the available pipelines used by the Institute.

The author indicates that the Z chromosome was identified according to homology to another moth species genome *Zeuzera pyrina.* But does not explicitly mention how this homology was determined. What is more, in the reference paper for the *Zeuzera pyrina* genome assembly (Boyes *et al*. 2023), I could not easily find the method by which they assigned the sex chromosome. I suggest that the author should briefly explain the method specified in Boyes *et al.* 2023, where they called it ”localised homologous pairs”.

The quality assessment of the assembly and statistics concluded that 99.8% of the assembly (length of 590.72 Mb) was assigned to the 30 chromosome-level scaffolds, a BUSCO score of 98.4%, and a N50 of 20.82 Mb. The high quality of the sequenced data and curation allowed the author to assemble a very good genome for this Nationally rare Lepidopteran (according to Britain, Wheeler 2025).

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: sex-chromosome evolution

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.