DATA NOTE

The genome sequence of the Neglected Rustic moth, *Xestia castanea* (Esper, 1798) (Lepidoptera: Noctuidae)

[version 1; peer review: 2 approved]

Gavin R. Broad[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]Natural History Museum, London, England, UK

v1 **First published:** 17 Apr 2026, **11**:226
https://doi.org/10.12688/wellcomeopenres.26316.1
Latest published: 17 Apr 2026, **11**:226
https://doi.org/10.12688/wellcomeopenres.26316.1

Abstract

We present a genome assembly from an individual female *Xestia castanea* (Neglected Rustic; Arthropoda; Insecta; Lepidoptera; Noctuidae). The assembly contains two haplotypes with total lengths of 721.78 megabases and 685.53 megabases. Most of haplotype 1 (99.81%) is scaffolded into 32 chromosomal pseudomolecules, including the W, and Z sex chromosomes and a supernumerary chromosome B_1. Most of haplotype 2 (99.63%) is scaffolded into 32 chromosomal pseudomolecules, including a supernumerary chromosome B_2. The mitochondrial genome has also been assembled, with a length of 15.36 kilobases. Gene annotation of this assembly on Ensembl identified 15 258 protein-coding genes. 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

Xestia castanea moth, Neglected Rustic, genome sequence, chromosomal, Lepidoptera

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 17 Apr 2026	✓ view	✓ view

1. **Vahap Eldem**, Istanbul University, Istanbul, Turkey

2. **Shangzhe Zhang**, University of California Riverside (Ringgold ID: 8790), Riverside, UK

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

This article is included in the Tree of Life gateway.

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

Author roles: Broad GR: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing;

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: Broad GR, Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding Collective *et al.* **The genome sequence of the Neglected Rustic moth, *Xestia castanea* (Esper, 1798) (Lepidoptera: Noctuidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2026, **11**:226 https://doi.org/10.12688/wellcomeopenres.26316.1

First published: 17 Apr 2026, **11**:226 https://doi.org/10.12688/wellcomeopenres.26316.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Amphiesmenoptera; Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Noctuoidea; Noctuidae; Noctuinae; Noctuini; *Xestia*; *Xestia castanea* (Esper, 1798) (NCBI:txid988045).

Background

Xestia castanea, Neglected Rustic, is a noctuid moth which inhabits heaths, moorland and bogs, where the larvae feed on heathers (*Erica* species) and other low-growing plants, such as Bog-myrtle (*Myrica gale*) and Cowberry (*Vaccinium vitis-idaea*) (Henwood *et al.*, 2020). The adults are rather plain with some fine black markings, most noticeably on the lower and inner edge of the reniform stigma. Greyer individuals are more frequent in the south of Britain, reddish in the north, with some other local colour variants (South, 1907; Waring *et al.*, 2017). Larvae feed at night from October through to May, pupating in the ground with adults then on the wing mainly in August and September, when they are attracted to light. Some individuals carry significant pollen loads (Devoto *et al.*, 2011).

There has been a substantial loss of range (69% decline in distribution since 1970) and decline in abundance (down 14%) in Britain and Ireland (Randle *et al.*, 2019), where *X. castanea* is now mainly a species of the north and west, although still with populations on the heaths of southern England. South (1907) records *X. castanea* as occurring on 'the larger tracts of heathery ground throughout the British Isles', but many of these tracts have been lost, and *X. castanea* is probably one of a range of moth species declining as a result of land use change and increased nitrogen deposition (Fox *et al.*, 2014). This decline is not an exclusively British phenomenon; in Belgium, for example, *X. castanea* is noted as 'very rare and declining', regionally extinct in Flanders.

Xestia castanea is found almost entirely in Europe. Nedumpally *et al.* (2025), in a phylogenetic analysis including a large proportion of the northern European *Xestia*, found *X. castanea* to be sister species to *X. subrosea*, formerly classified in the genus *Coenophila.* Although genital morphology had previously obscured the close relationship between these species, they overlap in habitat and foodplants, although *X. subrosea* is much more localised. The genomes of an

Figure 1. Photograph of the *Xestia castanea* (ilXesCast1) specimen used for genome sequencing.

increasing number of *Xestia* species, and related Noctuini genera, will help in understanding the diversification of a predominantly northern radiation of Noctuidae.

The sequenced specimen (Figure 1) was collected in Beinn Eighe National Nature Reserve in the far Northwest of Scotland, where *X. castanea* is numerous, as part of a collaboration between Darwin Tree of Life and Nature Scot.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult female *Xestia castanea* (specimen ID NHMUK015060365, ToLID ilXesCast1; Figure 1), collected from Beinn Eighe, Scotland, United Kingdom (latitude 57.61, longitude −5.31) on 2022-08-23. The specimen was collected and identified by Gavin Broad.

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

Nucleic acid extraction

Detailed protocols for nucleic acid extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The ilXesCast1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the head and thorax was homogenised by powermashing using a PowerMasher II tissue disruptor.

High molecular weight (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 38.6 ng/μL and a yield of 1 814.20 ng, with a fragment size of 13.7 kb. The Genomic Quality Number (GQN) was 6.5.

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 25 M 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 head and thorax tissue from the ilXesCast1 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, 2022), producing two haplotypes. Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*,

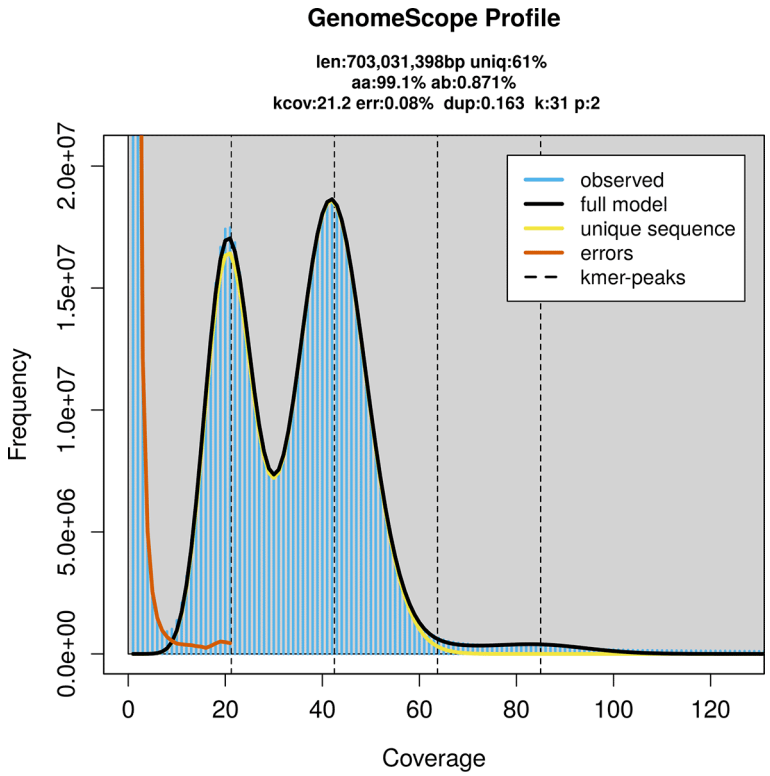

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

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 MerquryFK (Rhie *et al.*, 2020).

The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023).

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included 27 breaks and 198 joins. This reduced the scaffold count by 53.0%. The curation process is described at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The MerquryFK 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

Table 1. Specimen and sequencing data for BioProject PRJEB83541.

Platform	PacBio HiFi	Hi-C
ToLID	ilXesCast1	ilXesCast1
Specimen ID	NHMUK015060365	NHMUK015060365
BioSample (source individual)	SAMEA115574579	SAMEA115574579
BioSample (tissue)	SAMEA115599601	SAMEA115599601
Tissue	head and thorax	head and thorax
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR14106135	ERR14075572
Read count total	3.46 million	323.76 million
Base count total	30.73 Gb	97.77 Gb

Table 2. Genome assembly statistics.

Genome assembly	Haplotype 1	Haplotype 2
Assembly name	ilXesCast1.hap1.1	ilXesCast1.hap2.1
Assembly accession	GCA_965113115.1	GCA_965113175.1
Assembly level	chromosome	chromosome
Span (Mb)	721.78	685.53
Number of chromosomes	32	32
Number of contigs	318	339
Contig N50	7.09 Mb	6.6 Mb
Number of scaffolds	77	135
Scaffold N50	24.57 Mb	23.77 Mb
Longest scaffold length (Mb)	30.68	27.88
Sex chromosomes	W and Z	-
Organelles	Mitochondrion: 15.36 kb	-

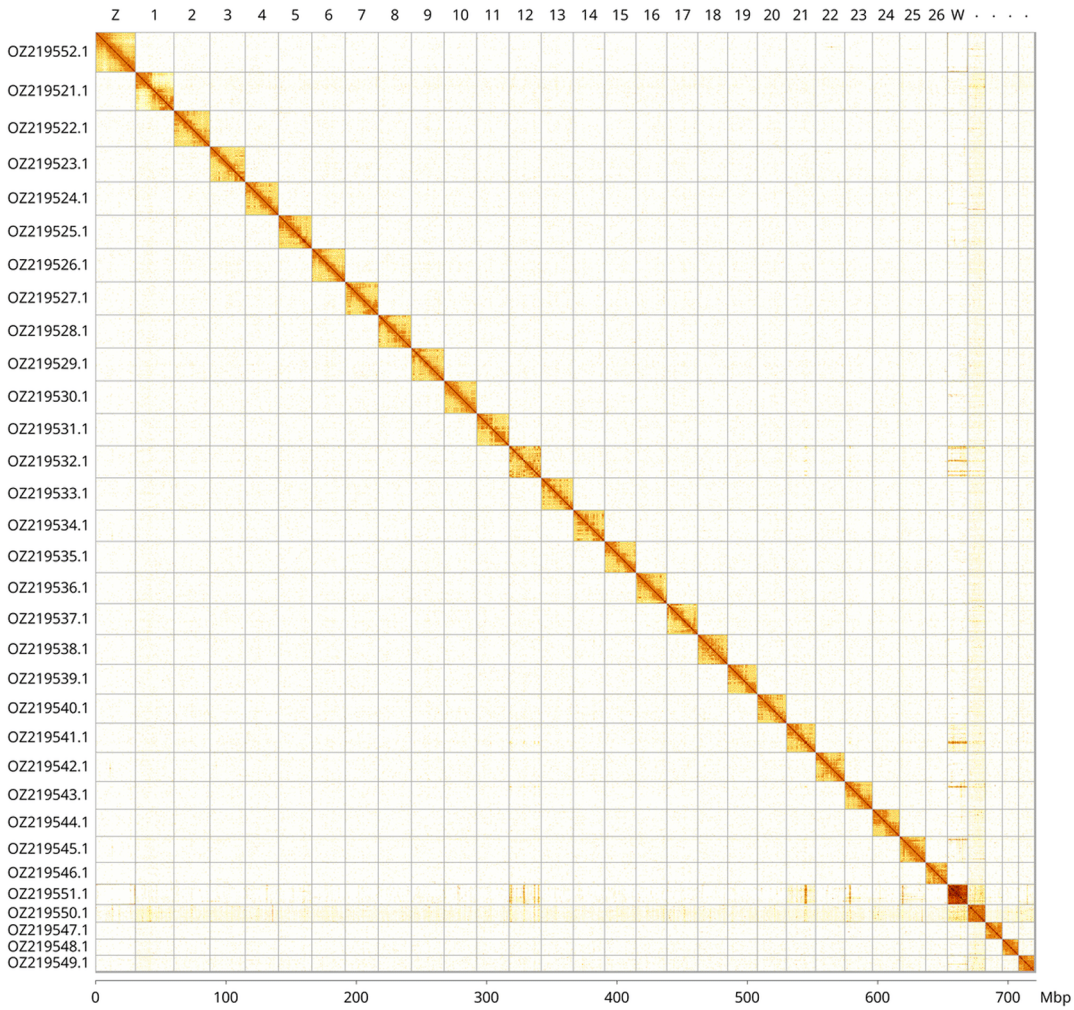

Figure 3. Hi-C contact map of the *Xestia castanea* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in both haplotypes of the genome assembly of *Xestia castanea*, ilXesCast1.

Haplotype 1				Haplotype 2			
INSDC accession	**Name**	**Length (Mb)**	**GC%**	**INSDC accession**	**Name**	**Length (Mb)**	**GC%**
OZ219521.1	1	29.55	39	OZ219682.1	1A	16.24	38.50
OZ219522.1	2	27.75	38	OZ219683.1	1B	12.98	39
OZ219523.1	3	26.91	38.50	OZ219684.1	2	27.88	38.50
OZ219524.1	4	25.71	38.50	OZ219685.1	3	26.87	38.50
OZ219525.1	5	25.57	38	OZ219686.1	4	25.58	38
OZ219526.1	6	25.55	38.50	OZ219687.1	5	25.39	38
OZ219527.1	7	25.31	38	OZ219688.1	6	25.45	38.50
OZ219528.1	8	25.30	38	OZ219689.1	7	25.14	38
OZ219529.1	9	25.18	38.50	OZ219690.1	8	25.53	38
OZ219530.1	10	25.02	38.50	OZ219691.1	9	25.11	38.50

Table 3. *Continued*

Haplotype 1				Haplotype 2			
INSDC accession	**Name**	**Length (Mb)**	**GC%**	**INSDC accession**	**Name**	**Length (Mb)**	**GC%**
OZ219531.1	11	24.80	38	OZ219692.1	10	25.20	38.50
OZ219532.1	12	24.74	38.50	OZ219693.1	11	25.04	38
OZ219533.1	13	24.57	38	OZ219694.1	12	24.67	38.50
OZ219534.1	14	24.05	38	OZ219695.1	13	24.32	38
OZ219535.1	15	24	38.50	OZ219696.1	14	23.77	38
OZ219536.1	16	23.80	38.50	OZ219697.1	15	24.24	38.50
OZ219537.1	17	23.69	38.50	OZ219698.1	16	23.43	38.50
OZ219538.1	18	22.88	38	OZ219699.1	17	23.48	38.50
OZ219539.1	19	22.80	38.50	OZ219700.1	18	23	38.50
OZ219540.1	20	22.39	39	OZ219701.1	19	22.98	38.50
OZ219541.1	21	22.34	38.50	OZ219702.1	20	22.45	38.50
OZ219542.1	22	22.31	38.50	OZ219703.1	21	23	39
OZ219543.1	23	21.20	38.50	OZ219704.1	22	22.38	38.50
OZ219544.1	24	20.91	38.50	OZ219705.1	23	21.50	38.50
OZ219545.1	25	19.83	39.50	OZ219706.1	24	20.88	38.50
OZ219546.1	26	16.77	38.50	OZ219707.1	25	19.46	39
OZ219547.1	27	12.96	38.50	OZ219708.1	26	16.96	38.50
OZ219548.1	28	12.44	39	OZ219709.1	27	13.04	39
OZ219549.1	29	12.31	39.50	OZ219710.1	28	12.53	39
OZ219551.1	W	15.72	42	OZ219711.1	29	12.37	40
OZ219552.1	Z	30.68	38	OZ219712.1	B2	13.61	39.50
OZ219550.1	B1	13.40	39	OZ219713.1	B3	8.53	39.50

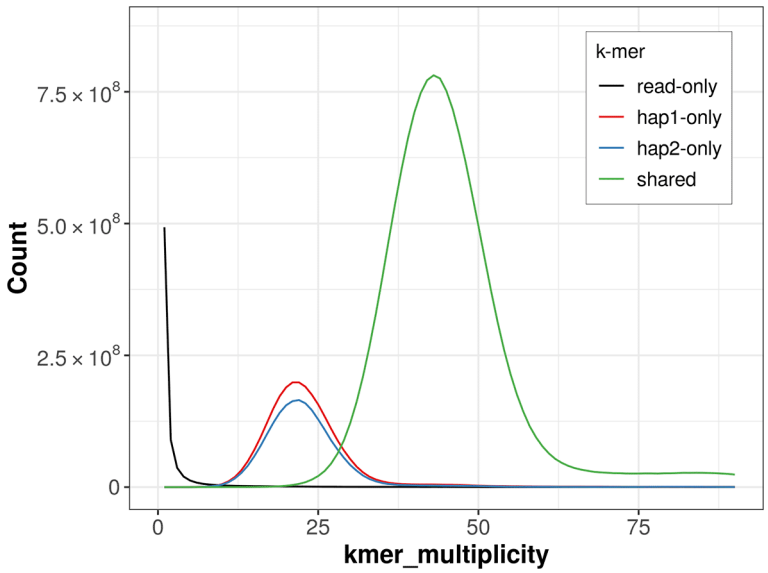

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

(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 *Xestia castanea* specimen generated 30.73 Gb (gigabases) from 3.46 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 703.03 Mb, with a heterozygosity of 0.87% and repeat content of 39.15% (Figure 2). These estimates guided expectations for the assembly.

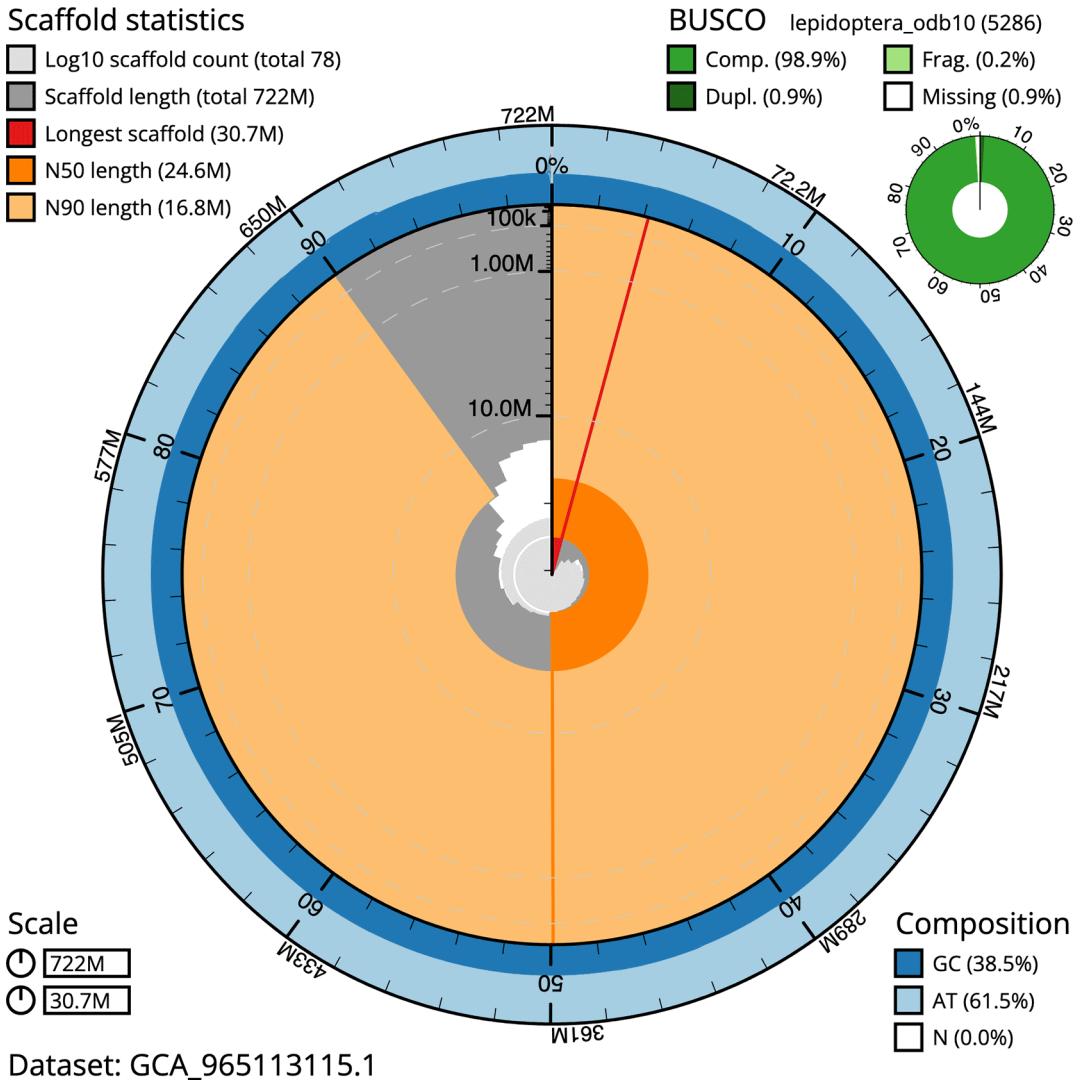

Figure 5. Assembly metrics for ilXesCast1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost 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.

Based on the estimated genome size, the sequencing data provided approximately 42× coverage. Hi-C sequencing produced 97.77 Gb from 323.76 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 721.78 Mb in 77 scaffolds, with 241 gaps, and a scaffold N50 of 24.57 Mb (Table 2).

Most of the haplotype 1 assembly sequence (99.81%) was assigned to 32 chromosomal-level scaffolds, representing 29 autosomes and the W and Z sex chromosomes, plus a supernumerary B_1 chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). The Z chromosome was assigned based on BUSCO gene painting with ancestral Merian elements (Wright *et al.*, 2024). The Z and W chromosomes were identified based on read coverage analysis and its single-copy status within a merged diploid map. The B chromosome is present in multiple copies of varying length in the haplotype 2 assembly.

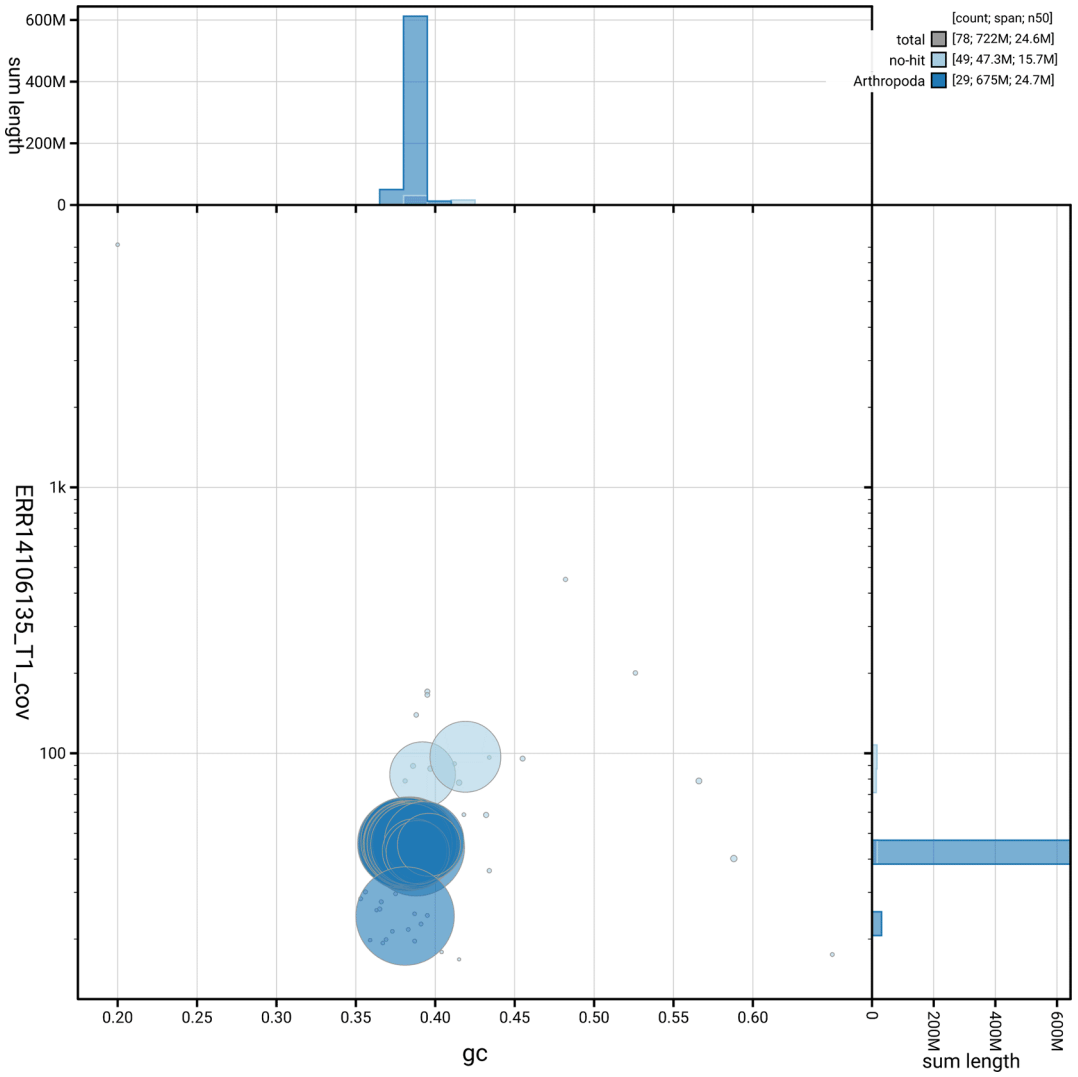

Figure 6. BlobToolKit blob plot for ilXesCast1.hap1.1. The plot shows base coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the 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 *Xestia castanea* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q63	6.C.Q40
Contig N50 length	7.09 Mb	≥ 1 Mb
Scaffold N50 length	24.57 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 63.9; haplotype 2: 63.1; combined: 63.5	≥ 40
k-mer completeness	Haplotype 1: 82.91%; Haplotype 2: 79.22%; combined: 99.51%	≥ 95%
BUSCO	C:98.9% [S:98.0%, D:0.9%], F:0.2%, M:0.9%, n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.81%	≥ 90%

Notes: The EBP summary uses log10(Contig N50); chromosome-level (C) or log10(Scaffold N50); Q (Merqury QV). BUSCO: C = complete; S = single-copy; D = duplicated; F = fragmented; M = missing; n = orthologues.

The mitochondrial genome was also assembled (length 15.36 kb, OZ219553.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

For haplotype 1, the estimated QV is 63.9, and for haplotype 2, 63.1. When the two haplotypes are combined, the assembly achieves an estimated QV of 63.5. The *k*-mer completeness is 82.91% for haplotype 1, 79.22% for haplotype 2, and 99.51% for the combined haplotypes (Figure 4).

BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.9% of the expected gene set (single = 98.0%, duplicated = 0.9%) in haplotype 1. For haplotype 2, BUSCO v.5.5.0 analysis identified 94.6% of the expected gene set (single = 93.8%, duplicated = 0.9%). 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 January 2026. The EBP metric, calculated for the haplotype 1, is **6.C.Q63**, meeting the recommended reference standard.

Genome annotation report

The *Xestia castanea* genome assembly (GCA_965113115.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 28 018 transcribed mRNAs from 15 258 protein-coding and 3 995 non-coding genes. The average transcript length is 18 672.29 bp, with an average of 1.46 coding transcripts per gene and 6.29 exons per transcript. Further details of this annotation are available from the Ensembl annotation page.

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

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 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: Xestia castanea. Accession number PRJEB83541. The genome sequence is released openly for reuse. The *Xestia castanea* 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. Raw data and assembly accession identifiers are reported in Tables 1 and 2.

Production code used in genome assembly at the WSI Tree of Life is available at https://github.com/sanger-tol. Table 5 lists software versions used in this study.

Table 5. Software versions and sources.

Software	Version	Source
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
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	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView

Table 5. *Continued*

Software	Version	Source
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	1.2.2	https://github.com/c-zhou/yahs

References

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

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

Buchfink B, Reuter K, Drost H-G: **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 Genes|Genomes|Genetics.* 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.** *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).
PubMed Abstract | Publisher Full Text | Free Full Text

Devoto M, Bailey S, Memmott J: **The "night shift": Nocturnal pollen-transport networks in a boreal pine forest.** *Ecological Entomology.* 2011; **36**: 25–35.
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.** *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

Fox R, Oliver TH, Harrower C, *et al.*: **Long-term changes to the frequency of occurrence of British moths are consistent with opposing and synergistic effects of climate and land-use changes.** *J. Appl. Ecol.* 2014; **51**: 949–957.
Publisher Full Text

Henwood B, Sterling P, Lewington R: *Field Guide to the Caterpillars of Great Britain and Ireland.* Bloomsbury Wildlife; 2020.

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).
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

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

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

Merkel D: **Docker: Lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239).
Publisher Full Text

Nedumpally V, Zilli A, Yapar E, *et al.*: **Elaborating the phylogeny of Noctuidae by focusing on relationships between northern European taxa.** *Syst. Entomol.* 2025; **51**: e70010.
Publisher 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

Randle Z, Evans-Hill LJ, Parsons MS, *et al.*: *Atlas of Britain & Ireland's Larger Moths.* Pisces Publications; 2019.

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).
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.* 2020; **2020**: baaa062.
PubMed Abstract | Publisher Full Text | Free Full Text

South R: *The Moths of the British Isles.* London: Frederick Warne & Co; 1907.

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life Project.** *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.** *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE; 2019; 314–324.
Publisher Full Text

Waring P, Townsend M, Lewington R: *Field Guide to the Moths of Great Britain and Ireland.* Bloomsbury Publishing; 2017.

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

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

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 15 May 2026

https://doi.org/10.21956/wellcomeopenres.28981.r153687

✔ **Shangzhe Zhang**
Department of Entomology, University of California Riverside (Ringgold ID: 8790), Riverside, California, UK

The authors presented genomic assembly for *Xestia castanea*, which constitutes important input to the ToL project. I do think the protocol and results are rigorous, am just listing a few points that were confusing and can be improved.

1. I was very interested in the supernumerary B chromosomes reported, however I could not find any method description on how the authors defined them. In the result section, I don't see plots to enlighten me what are the B chromosomes look like.

2. In Figure 6, I believe the axis labels are understandable, but it would be neat to make them clearer to wider group of readers.

3. The EBP metric is quite a new concept for me, it would be helpful to brief how it works and when to tell the assembly is in good quality in the methods section.

4. From the aspect of using the genome as a reference in future works, it's good to have Z and W haplotypes being allocated to one haplotype but this needs some clarification.

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, genomic structural variants, 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 08 May 2026

https://doi.org/10.21956/wellcomeopenres.28981.r153685

Vahap Eldem
Istanbul University, Istanbul, Turkey

The manuscript presents a chromosome-scale, haplotype-resolved genome assembly for the Neglected Rustic moth, Xestia castanea. The study describes the collection and taxonomic verification of a female specimen from Beinn Eighe, Scotland, followed by PacBio HiFi sequencing, Hi-C scaffolding, mitochondrial genome assembly, manual curation, quality assessment, and gene annotation. The final assembly is reported as two haplotypes with total lengths of 721.78 Mb and 685.53 Mb, with high BUSCO completeness for haplotype 1 and chromosome-level scaffolding for most of the sequence. The manuscript also provides sequencing metadata, assembly statistics, chromosome/pseudomolecule tables, k-mer and BUSCO quality metrics. Data availability is clearly provided through ENA accessions. Overall, the article represents a useful genomic resource for X. castanea and for comparative studies of Noctuidae and Lepidoptera more broadly.

Several points in the description of the haplotype-resolved assembly would benefit from clarification to ensure consistency between the abstract, tables, and main text.

The abstract states that haplotype 2 is mostly scaffolded into 32 chromosomal pseudomolecules, and table 2 lists the assembly level of haplotype 2 as chromosomes. However, the assembly statistics section states that haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level. This appears inconsistent and may require editing so that the abstract, table 2, and main text describe the status of haplotype 2 consistently.

The abstract mentions a supernumerary chromosome B2 in haplotype 2, whereas table 3 lists both B2 and B3 under haplotype 2. It would be helpful to clarify whether both B2 and B3 are interpreted as supernumerary chromosomes, B-chromosome-derived scaffolds, or alternative/partial representations of the same supernumerary element.

Table 2 reports 32 chromosomes for haplotype 2, while Table 3 lists 1A, 1B, chromosomes 2–29, B2, and B3, with no W or Z chromosome. Although this gives 32 named pseudomolecules, the

biological interpretation is not fully clear. In particular, whether 1A and 1B represent a split chromosome 1, two independent chromosomal scaffolds, or another structural interpretation may need clarification. The absence of W/Z designations from haplotype 2 may also be explained briefly.

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, Bioinformatics

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