DATA NOTE

The genome sequence of a tephritid fruit fly, *Xyphosia miliaria* (Schrank, 1781) (Diptera: Tephritidae)

[version 1; peer review: 2 approved]

Ryan Mitchell[1], Liam M. Crowley[2], James McCulloch[2,3], Olga Sivell[4], Natural History Museum Genome Acquisition Lab, University of Oxford and Wytham Woods Genome Acquisition Lab, 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]Independent researcher, Sligo, County Sligo, Ireland
[2]University of Oxford, Oxford, England, UK
[3]Wellcome Sanger Institute, Hinxton, England, UK
[4]Natural History Museum, London, England, UK

v1 **First published:** 21 Oct 2025, **10**:585
https://doi.org/10.12688/wellcomeopenres.25047.1
Latest published: 21 Oct 2025, **10**:585
https://doi.org/10.12688/wellcomeopenres.25047.1

Abstract

We present a genome assembly from an individual male *Xyphosia miliaria* (tephritid fruit fly; Arthropoda; Insecta; Diptera; Tephritidae). The assembly contains two haplotypes with total lengths of 806.98 megabases and 799.90 megabases. Most of haplotype 1 (97.34%) is scaffolded into 7 chromosomal pseudomolecules, including the X and Y sex chromosomes. Most of haplotype 2 (83.09%) is scaffolded into 5 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 19.41 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

Xyphosia miliaria; tephritid fruit fly; genome sequence; chromosomal; Diptera

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 21 Oct 2025	✓ view	✓ view

1. **Daubian Santos**, Universidade Federal do ABC, Santo André, Brazil
2. **Matsapume Detcharoen**, Division of Biological Science, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla, Thailand, Hat Yai, Thailand

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: Mitchell R: Investigation, Resources; **Crowley LM**: Investigation, Resources; **McCulloch J**: Investigation, Resources; **Sivell O**: 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: Mitchell R, Crowley LM, McCulloch J *et al.* **The genome sequence of a tephritid fruit fly, *Xyphosia miliaria* (Schrank, 1781) (Diptera: Tephritidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, **10**:585 https://doi.org/10.12688/wellcomeopenres.25047.1

First published: 21 Oct 2025, **10**:585 https://doi.org/10.12688/wellcomeopenres.25047.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Eremoneura; Cyclorrhapha; Schizophora; Acalyptratae; Tephritoidea; Tephritidae; Tephritinae; Xyphosiini; *Xyphosia*; *Xyphosia miliaria* (Schrank, 1781) (NCBI:txid381397)

Background

Xyphosia miliaria (Schrank, 1781) is a species from the Tephritidae, the largest of the picture-winged fly families in Britain. Tephritidae are also known as gall flies and, outside Britain, are often called fruit flies. This is the only species from the genus *Xyphosia* Robineau-Desvoidy, 1830 occurring in Britain (Chandler, 2025). It can be identified using characters listed and pictured in White (1988) and Clements (1990; 2020). The fly is small, with body length of about 2 mm and 3.7–6.3 mm wing length (White, 1988). The body and legs are orange; the oviscape is dark orange with a black tip. The wing pattern is distinctive and consists of three large dark spots, at the wing tip, the end of the discal cell and near the end of R_1, against a reticulated pattern of smaller dark and pale spots as figured in White (1988). White (1988) also provides a key to puparia of tephritid species found in thistles (Mill, 1754).

Xyphosia miliaria is a Palaearctic species common across Europe and in Asia is found in: Russia, China, Mongolia, Kyrgyzstan, Kazakhstan (GBIF Secretariat, 2023). It is very frequent and widespread in Britain and Ireland (GBIF Secretariat, 2023; NBN Atlas Partnership, 2025).

The larvae develop in the flower heads (capitulum) of thistles *Cirsium* spp. and *Carduus* spp. (Asteraceae: Carduoideae), feeding on thistle achenes (seeds) and pupating inside the host in a cocoon formed of pappus hairs (White, 1988). *X. miliaria* can develop equally well in male or female thistle heads (unlike e.g. *Terellia* Robineau-Desvoidy, 1830 which develops only in female heads), and it shows no preference for fertilised plants (Basov, 2004; Walker *et al.*, 2008). The larva of *X. miliaria* was described by Persson (1963). In Britain there are two generations a year, and the species overwinters as a larva (White, 1988). The adults are on the wing from June to August, peaking in July (NBN Atlas Partnership, 2025). *Xyphosia miliaria* has been recorded developing in the following plant species in Britain: *Cirsium arvense*, *C. palustre*, *Centaurea nigra* (likely not a usual host), and abroad also in *Carduus acanthoides*, *C. nutans*, *Cirsium eriophorum*, *C. oleraceum* and *Sonchus* spp. (White, 1988). The flies inside the flower heads are attacked by hymenopteran parasitoids, such as *Torymus chloromerus* (Walker, 1833) (Hymenoptera: Torymidae) and *Pteromalus elevatus* (Walker, 1834) (Hymenoptera: Pteromalidae) (Walker *et al.*, 2008).

The high-quality genome of *Xyphosia miliaria* was sequenced from a single female (NHMUK015058995; SAMEA112964123) from Upton Broad and Marshes, England. The genome of *X. miliaria* presented here 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. It will aid research into phylogeny of true flies and taxonomy, biology and ecology of the species.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Xyphosia miliaria* (specimen ID NHMUK015058995, ToLID idXypMili3; Figure 1), collected from Upton Broad and Marshes, England, United Kingdom (latitude 52.67, longitude 1.52) on 2022-07-03. The specimen was collected and identified by Ryan Mitchell. A second specimen was used for Hi-C sequencing (specimen ID Ox002637, ToLID idXypMili1), collected from Wytham Woods, Oxfordshire, United Kingdom (latitude 51.772, longitude –1.338) on 2022-08-03. The specimen was collected by James McCulloch and Liam Crowley, and formally identified by James McCulloch. Sample metadata were collected in line with the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

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

Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of

Figure 1. Photograph of the *Xyphosia miliaria* (idXypMili3) specimen used for genome sequencing.

Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The idXypMili3 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 using the Automated MagAttract v2 protocol. We used centrifuge-mediated fragmentation to produce DNA fragments in the 8–10 kb range, following the Covaris g-TUBE protocol for ultra-low input (ULI). Sheared DNA was purified by 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. 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 1.96 ng/μL and a yield of 254.80 ng. The 260/280 spectrophotometric ratio was 1.96, and the 260/230 ratio was 1.26.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Prior to library preparation, the DNA was fragmented to ~10 kb. Ultra-low-input (ULI) libraries were prepared using the PacBio SMRTbell® Express Template Prep Kit 2.0 and gDNA Sample Amplification Kit. Samples were normalised to 20 ng DNA. Single-strand overhang removal, DNA damage repair, and end-repair/A-tailing were performed according to the manufacturer's instructions, followed by adapter ligation. A 0.85× pre-PCR clean-up was carried out with Promega ProNex beads.

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

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

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

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen whole organism tissue of the idXypMili1 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 to 16 PCR cycles. Post-PCR clean-up was performed with SPRISelect beads. Libraries were quantified using the AccuClear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech).

Prior to sequencing, libraries were normalised to 10 ng/μL. Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq 6000.

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 34 breaks and 100 joins. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020) was run in a Singularity container (Kurtzer *et al.*, 2017) to evaluate *k*-mer completeness and assembly quality for 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 *Xyphosia miliaria* specimen generated 59.92 Gb (gigabases) from 8.81 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 681.97 Mb, with a heterozygosity of 2.05% and repeat content of 31.47% (Figure 2). These estimates guided expectations for the assembly. Based

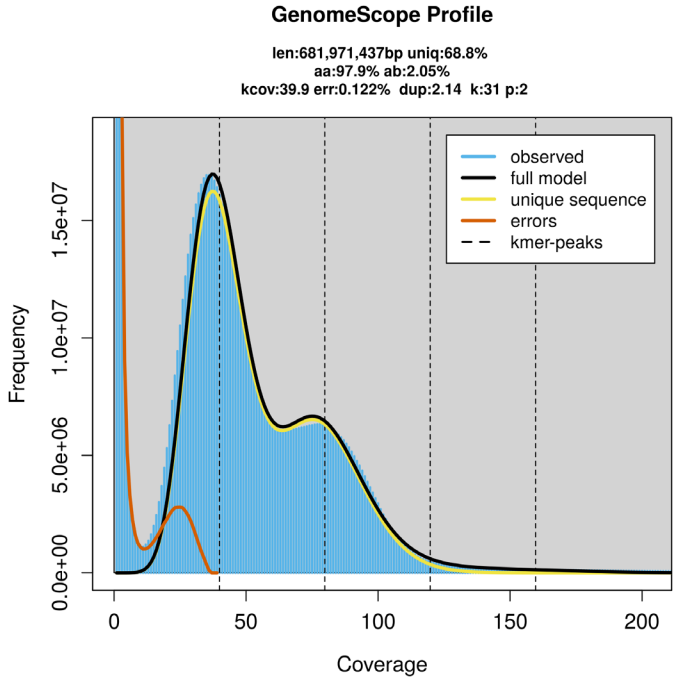

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

on the estimated genome size, the sequencing data provided approximately 80× coverage. Hi-C sequencing produced 120.83 Gb from 800.23 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 806.98 Mb in 261 scaffolds, with 509 gaps, and a scaffold N50 of 132.49 Mb (Table 2).

Most of the assembly sequence (97.34%) was assigned to 7 chromosomal-level scaffolds, representing 5 autosomes and the X and Y sex chromosomes. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). During curation, the X chromosome was identified by copy number. Y chromosome scaffolds were found, but it could not be assembled since Hi-C data came from a female specimen. Some pieces of the Y chromosome may be in the unassembled contigs.

The mitochondrial genome was also assembled. 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 59.0, and for haplotype 2, 57.9. When the two haplotypes are combined, the assembly achieves an estimated QV of 58.4. The *k*-mer completeness

Table 1. Specimen and sequencing data for BioProject PRJEB85366.

Platform	PacBio HiFi	Hi-C
ToLID	idXypMili3	idXypMili1
Specimen ID	NHMUK015058995	Ox002637
BioSample (source individual)	SAMEA112964123	SAMEA112232815
BioSample (tissue)	SAMEA112975256	SAMEA112233321
Tissue	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq 6000
Run accessions	ERR14231578	ERR14242273
Read count total	8.81 million	800.23 million
Base count total	59.92 Gb	120.83 Gb

Table 2. Genome assembly statistics.

Assembly name	idXypMili3.hap1.1	idXypMili3.hap2.1
Assembly accession	GCA_965200055.1	GCA_965200985.1
Assembly level	chromosome	chromosome
Span (Mb)	806.98	799.90
Number of chromosomes	7	5
Number of contigs	770	8 227
Contig N50	4.2 Mb	0.67 Mb
Number of scaffolds	261	6 657
Scaffold N50	132.49 Mb	129.48 Mb
Longest scaffold length (Mb)	164.07	161.6
Sex chromosomes	X and Y	-
Organelles	Mitochondrion: 19.41 kb	-

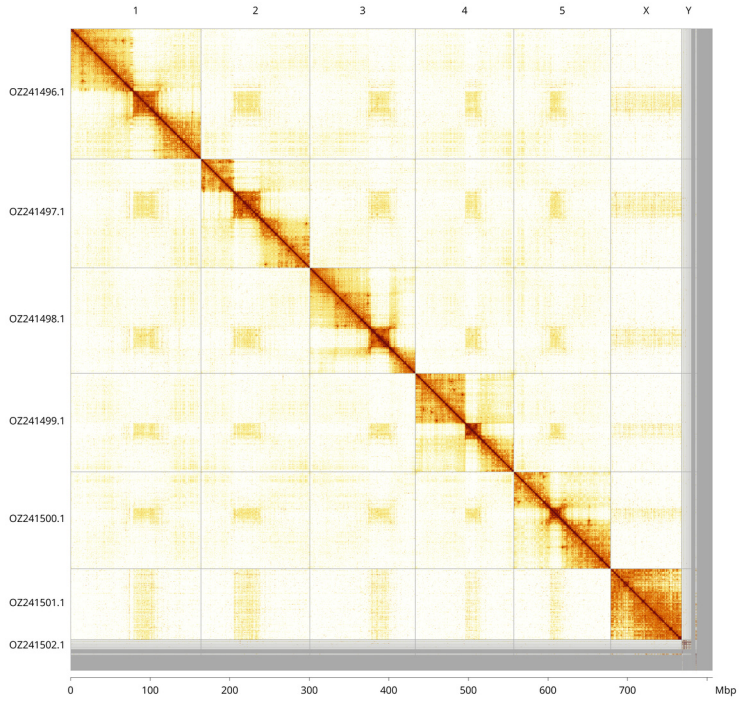

Figure 3. Hi-C contact map of the *Xyphosia miliaria* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in both haplotypes of the genome assembly of *Xyphosia miliaria*, idXypMili3.

Haplotype 1				Haplotype 2			
INSDC accession	**Name**	**Length (Mb)**	**GC%**	**INSDC accession**	**Name**	**Length (Mb)**	**GC%**
OZ241496.1	1	164.07	33	OZ241814.1	1	161.60	33
OZ241497.1	2	136.81	33	OZ241815.1	2	133.38	33
OZ241498.1	3	132.49	33	OZ241816.1	3	129.48	33
OZ241499.1	4	123.79	32.50	OZ241817.1	4	119.16	32.50
OZ241500.1	5	121.77	33	OZ241818.1	5	120.99	33
OZ241501.1	X	89.47	35				
OZ241502.1	Y	17.08	30.50				

is 72.49% for haplotype 1, 68.14% for haplotype 2, and 98.93% for the combined haplotypes (Figure 4).

BUSCO analysis using the diptera_odb10 reference set (n = 3 285) identified 98.6% of the expected gene set (single = 98.2%, duplicated = 0.4%) 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 for the haplotype 1, is **6.C.Q59**, meeting the recommended reference standard.

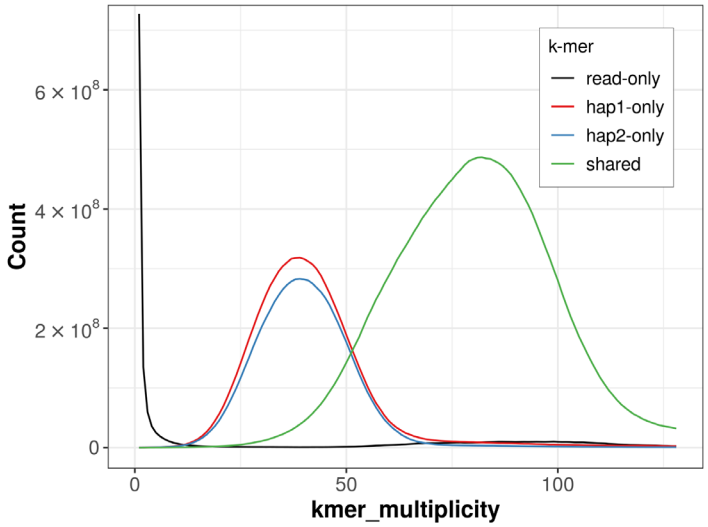

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

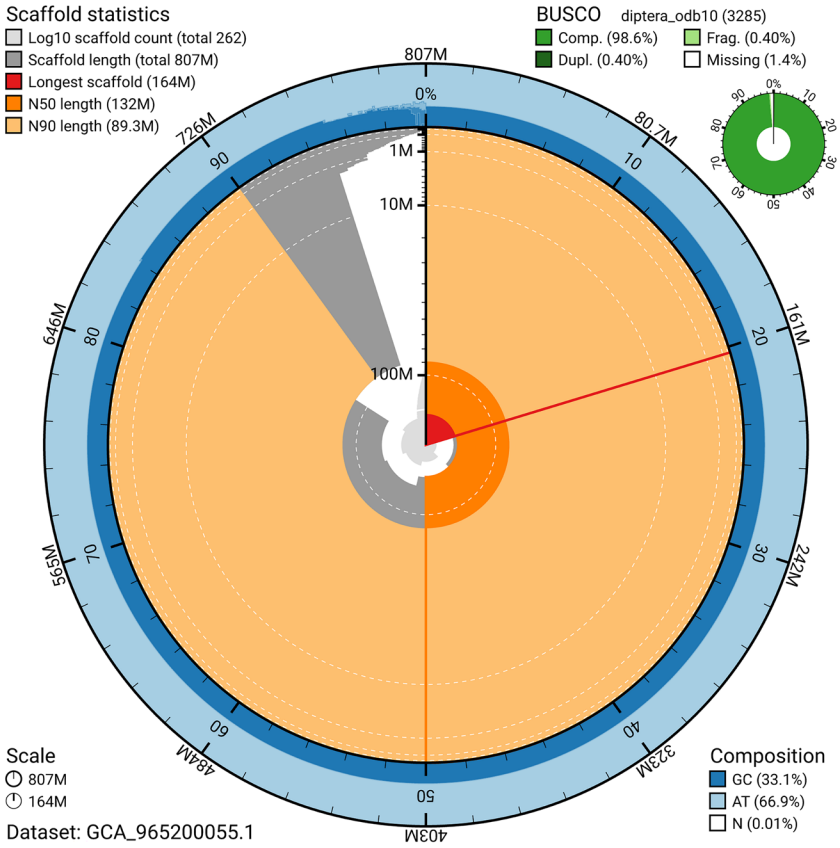

Figure 5. Assembly metrics for idXypMili3.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost 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.

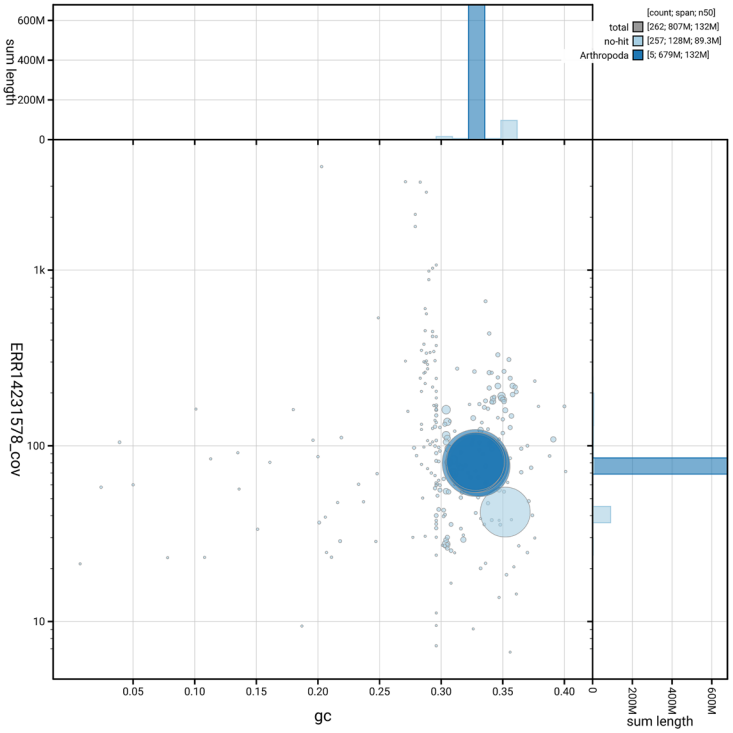

Figure 6. BlobToolKit GC-coverage plot for idXypMili3.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Xyphosia miliaria* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q59	6.C.Q40
Contig N50 length	4.20 Mb	≥ 1 Mb
Scaffold N50 length	132.49 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 59.0; haplotype 2: 57.9; combined: 58.4	≥ 40
k-mer completeness	Haplotype 1: 72.49%; Haplotype 2: 68.14%; combined: 98.93%	≥ 95%
BUSCO	C:98.6% [S:98.2%; D:0.4%]; F:0.4%; M:1.0%; n:3 285	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	97.34%	≥ 90%

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: Xyphosia miliaria. Accession number PRJEB85366. The genome sequence is released openly for reuse. The *Xyphosia miliaria* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665) and the Sanger Institute Tree of Life Programme (PRJEB43745). 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

Acknowledgements

Thank you to Duncan Sivell for comments on the draft of the Background.

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	4.4.5	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.7.1	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK

Software	Version	Source
Minimap2	2.28-r1209	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	24.10.4	https://github.com/nextflow-io/nextflow
PretextSnapshot	-	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
samtools	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.7.1	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

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

Basov VM: **Ecological peculiarities of *Xyphosia miliaria* (Schrank, 1781) (Diptera: Tephritidae) in Eastern Europe.** *News Kharkov Entomol Soc.* 2004; **11**(1–2): 177–81.
Reference Source

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

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

Chandler P: **An update of the 1998 checklist of Diptera of the British Isles**. 2025.
Reference Source

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

Clements DK: **Provisional keys to the Otitidae and Platystomatidae of the British Isles.** *Dipter Digt.* 1990; **6**: 32–41.
Reference Source

Clements DK: **Keys to British picture-wing flies (Diptera: Tephritidae, Ulidiidae, Platystomatidae, Pallopteridae and Opomyzidae)**. 2020.

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial arthropod fauna [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 123.
PubMed Abstract | Publisher Full Text | Free Full Text

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
PubMed Abstract | Publisher Full Text

Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
PubMed Abstract | Publisher Full Text | Free Full Text

GBIF Secretariat: ***Xyphosia miliaria* (Schrank, 1781) in GBIF Backbone Taxonomy**. 2023.
Publisher Full Text

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

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

Lawniczak MKN, Davey RP, Rajan J, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project [version 1; peer review: 2 approved with reservations].** *Wellcome*

Open Res. 2022; **7**: 187.
Publisher Full Text

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

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

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

NBN Atlas Partnership: ***Xyphosia miliaria* (Schrank, 1781)**. 2025.
Reference Source

Persson PI: **Studies on the biology and larval morphology of some Trypetidae (Dipt.).** *Opuscula Entomol.* 1963; **28**: 33–69.

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**

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

Walker M, Hartley SE, Jones TH: **The relative importance of resources and natural enemies in determining herbivore abundance: thistles, tephritids and parasitoids.** *J Anim Ecol.* 2008; **77**(5): 1063–71.
PubMed Abstract | **Publisher Full Text**

White IM: ***Tephritid flies*.** London: The Royal Entomological Society, 1988; **10**.
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 29 December 2025

https://doi.org/10.21956/wellcomeopenres.27613.r140485

✔ **Matsapume Detcharoen**
Division of Biological Science, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla, Thailand, Hat Yai, Thailand

The paper presents a high quality, phased reference genome for the tephritid fruit fly *Xyphosia miliaria*. The authors assembled two haplotypes of roughly 807 Mb and 800 Mb, with most of haplotype 1 assigned to seven chromosome scale scaffolds that include the X and Y chromosomes, and most of haplotype 2 assigned to five autosomal chromosomes. Assembly quality is very high and 98.6 percent complete BUSCO. The authors highlight that this resource will support studies of fly phylogeny, host plant use, parasitoid interactions, and broader biodiversity genomics efforts.

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: Molecular 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.

Reviewer Report 25 November 2025

https://doi.org/10.21956/wellcomeopenres.27613.r140480

Daubian Santos
Universidade Federal do ABC, Santo André, State of São Paulo, Brazil

The article is very good and a part of the excellent Darwin Tree of Life Project.
Good writing and background well stablished. The methodology is also clear and efficient.
I only said two small comments:
- Tephritidae is the largest of the picture-winged fly families in general, not only in Britain.
- Although it is not obligated, the authors may include the reference of species authors original descriptions cited in the text (for example, Robineau-Desvoidy, 1830) in the references.

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: Taxonomy, systematics, protocols of extracting genetic data

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.