DATA NOTE

The genome sequence of the Dark Umber moth, *Philereme transversata* (Hufnagel, 1767)

[version 1; peer review: 1 approved, 2 approved with reservations]

Douglas Boyes[1+], Clare Boyes[2],
University of Oxford and Wytham Woods 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]UK Centre for Ecology & Hydrology, Wallingford, England, UK
[2]Independent researcher, Welshpool, Wales, UK

[+] Deceased author

v1 **First published:** 02 Jun 2025, **10**:300
https://doi.org/10.12688/wellcomeopenres.24265.1
Latest published: 02 Jun 2025, **10**:300
https://doi.org/10.12688/wellcomeopenres.24265.1

Abstract

We present a genome assembly from a male specimen of *Philereme transversata* (Dark Umber; Arthropoda; Insecta; Lepidoptera; Geometridae). The genome sequence has a total length of 591.75 megabases. Most of the assembly (99.1%) is scaffolded into 20 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled, with a length of 16.07 kilobases. Gene annotation of this assembly on Ensembl identified 12,207 protein-coding genes.

Keywords

Philereme transversata, Dark Umber, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ? ✓ ?

	1	2	3
version 1 02 Jun 2025	? view	✓ view	? view

1. **Yu-Feng Huang**, National Chung Hsing University, Taichung, Taiwan
2. **Annabel Whibley**, Bragato Research Institute, Lincoln, New Zealand
3. **Arun Arumugaperumal**, Rajalakshmi Engineering College, Chennai, India

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: Boyes D: Investigation, Resources; **Boyes C**: 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: Boyes D, Boyes C, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Dark Umber moth, *Philereme transversata* (Hufnagel, 1767) [version 1; peer review: 1 approved, 2 approved with reservations]** Wellcome Open Research 2025, **10**:300 https://doi.org/10.12688/wellcomeopenres.24265.1

First published: 02 Jun 2025, **10**:300 https://doi.org/10.12688/wellcomeopenres.24265.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Amphiesmenoptera; Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Geometroidea; Geometridae; Larentiinae; *Philereme*; *Philereme transversata* (Hufnagel, 1767) (NCBI:txid873512)

Background

The Dark Umber (*Philereme transversata*) is in the family Geometridae. It is a moth of woodlands and scrub, and is frequent across southern England, although rare in Devon and Cornwall. It becomes scarcer further north. It has declined in abundance since 1970 (Randle *et al.*, 2019). It is present throughout Europe (GBIF Secretariat, 2023).

The moth is a rich brown with a darker central cross band. The trailing edges of the wings are distinctively scalloped. There is one generation a year which flies between July and August. The caterpillars feed at night on Buckthorn and probably Alder Buckthorn, and then pupate in a cocoon in the ground (Waring *et al.*, 2017).

As part of the Darwin Tree of Life Project – which aims to generate high-quality reference genomes for all named eukaryotic species in Britain and Ireland to .to support research, conservation, and the sustainable management of biodiversity – we present a chromosomally complete genome sequence for the Dark Umber, *Philereme transversata*. This genome was assembled using the Tree of Life pipeline from a specimen collected in Wytham Woods, Oxfordshire, United Kingdom (Figure 1).

Genome sequence report

Sequencing data

The genome of a specimen of *Philereme transversata* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 61.42 Gb (gigabases) from 6.12 million reads, which were used to assemble the genome. GenomeScope analysis estimated the haploid genome size at 583.27 Mb, with a heterozygosity of 0.67% and repeat content of 35.69%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 103 coverage. Hi-C sequencing produced 109.02 Gb from 722.01 million reads, used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The assembly was improved by manual curation, which corrected three misjoins or missing joins and removed three haplotypic duplications. These interventions reduced the total assembly length by 0.83% and decreased the scaffold count by 3.33%. The final assembly has a total length of 591.75 Mb in 57 scaffolds, with 68 gaps, and a scaffold N50 of 29.48 Mb (Table 2).

Figure 1. Photograph of the *Philereme transversata* (ilPhiTran2) specimen used for genome sequencing.

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

Most of the assembly sequence (99.1%) was assigned to 20 chromosomal-level scaffolds, representing 19 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3). During curation, the Z chromosome was assigned based on synteny to *Philereme vetulata* (GCA_918857605.2).

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.

Assembly quality metrics

The estimated Quality Value (QV) and *k*-mer completeness metrics, along with BUSCO completeness scores, were calculated for each haplotype and the combined assembly. The QV reflects the base-level accuracy of the assembly, while *k*-mer completeness indicates the proportion of expected *k*-mers identified in the assembly. BUSCO scores provide a measure of completeness based on benchmarking universal single-copy orthologues.

The combined primary and alternate assemblies achieve an estimated QV of 65.0. The *k*-mer completeness is 84.47% for the primary haplotype and 84.18% for the alternate haplotype; and 99.64% for the combined primary and alternate assemblies. BUSCO v.5.5.0 analysis using the lepidoptera_odb10

Table 1. **Specimen and sequencing data for *Philereme transversata*.**

Project information			
Study title	Philereme transversata (dark umber)		
Umbrella BioProject	PRJEB65731		
Species	*Philereme transversata*		
BioSpecimen	SAMEA10978949		
NCBI taxonomy ID	873512		
Specimen information			
Technology	**ToLID**	**BioSample accession**	**Organism part**
PacBio long read sequencing	ilPhiTran2	SAMEA10979299	head and thorax
Hi-C sequencing	ilPhiTran1	SAMEA7701748	head and thorax
Sequencing information			
Platform	**Run accession**	**Read count**	**Base count (Gb)**
Hi-C Illumina NovaSeq 6000	ERR12035307	7.22e+08	109.02
PacBio Revio	ERR12015773	6.12e+06	61.42

Table 2. **Genome assembly data for *Philereme transversata*.**

Genome assembly		
Assembly name	ilPhiTran2.1	
Assembly accession	GCA_963924445.1	
Alternate haplotype accession	*GCA_963924475.1*	
Assembly level for primary assembly	chromosome	
Span (Mb)	591.75	
Number of contigs	125	
Number of scaffolds	57	
Longest scaffold (Mb)	41.65	
Assembly metric	**Measure**	***Benchmark***
Contig N50 length	9.52 Mb	*≥ 1 Mb*
Scaffold N50 length	29.48 Mb	*= chromosome N50*
Consensus quality (QV)	Primary: 65.4; alternate: 64.8; combined: 65.0	*≥ 40*
k-mer completeness	Primary: 84.47%; alternate: 84.18%; combined: 99.64%	*≥ 95%*
BUSCO*	C:98.4%[S:97.8%,D:0.6%], F:0.3%,M:1.3%,n:5,286	*S > 90%; D < 5%*
Percentage of assembly assigned to chromosomes	99.1%	*≥ 90%*
Sex chromosomes	Z	*localised homologous pairs*
Organelles	Mitochondrial genome: 16.07 kb	*complete single alleles*

* BUSCO scores based on the lepidoptera_odb10 BUSCO set using version 5.5.0. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison.

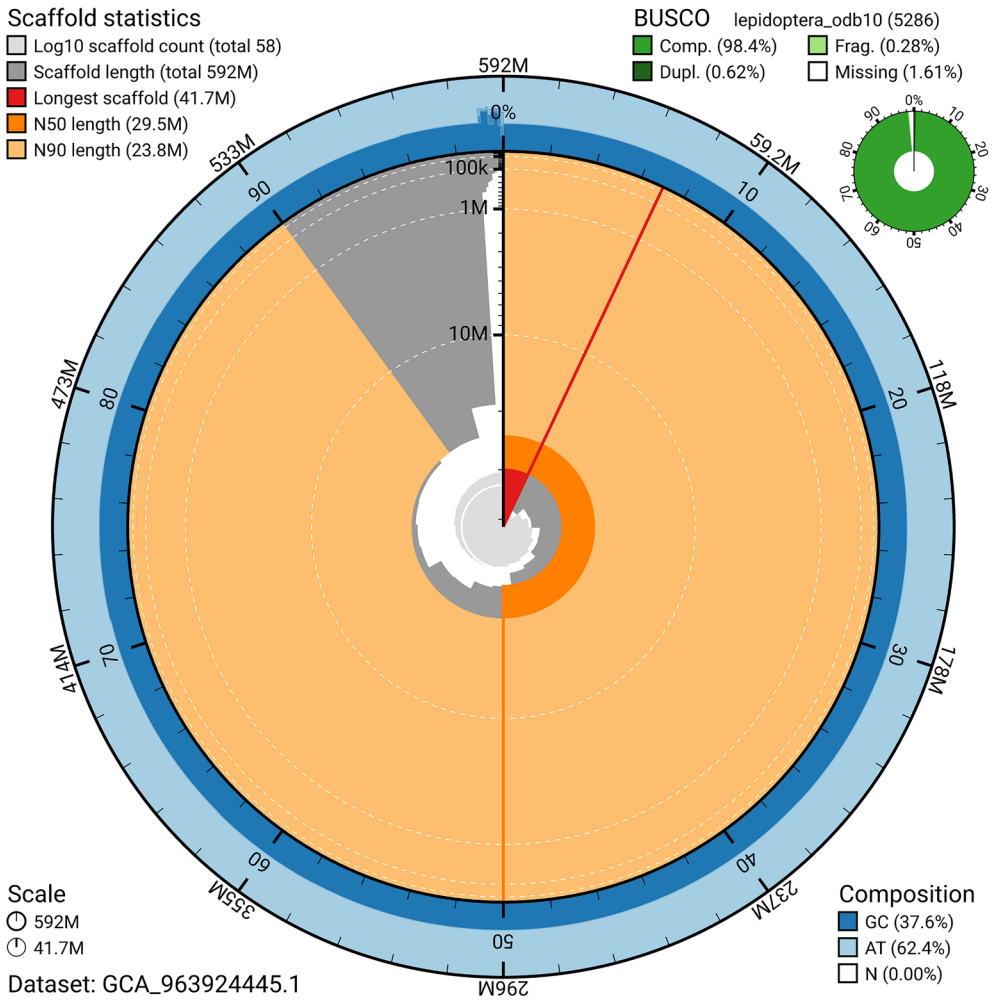

Figure 2. Genome assembly of *Philereme transversata*, ilPhiTran2.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the lepidoptera_odb10 set is presented at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963924445.1/dataset/GCA_963924445.1/snail.

reference set (n = 5,286) identified 98.4% of the expected gene set (single = 97.8%, duplicated = 0.6%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The primary assembly achieves the EBP reference standard of **6.C.Q65**.

Genome annotation report

The *Philereme transversata* genome assembly (GCA_963924445.1) was annotated externally by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 21,451 transcribed mRNAs from 12,207 protein-coding and 1,934 non-coding genes. The average transcript length is 18,457.27 bp. There are 1.52 coding transcripts per gene and 7.48 exons per transcript. For further information about the annotation, please refer to https://beta.ensembl.org/species/5ce8e749-aef4-4d48-93df-a5d729e2e8cc.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Philereme transversata* (specimen ID Ox001682, ToLID ilPhiTran2), collected from Wytham Woods, Oxfordshire, United Kingdom (latitude 51.772, longitude -1.338) on 2021-07-17, using a light trap. A second specimen was used for Hi-C

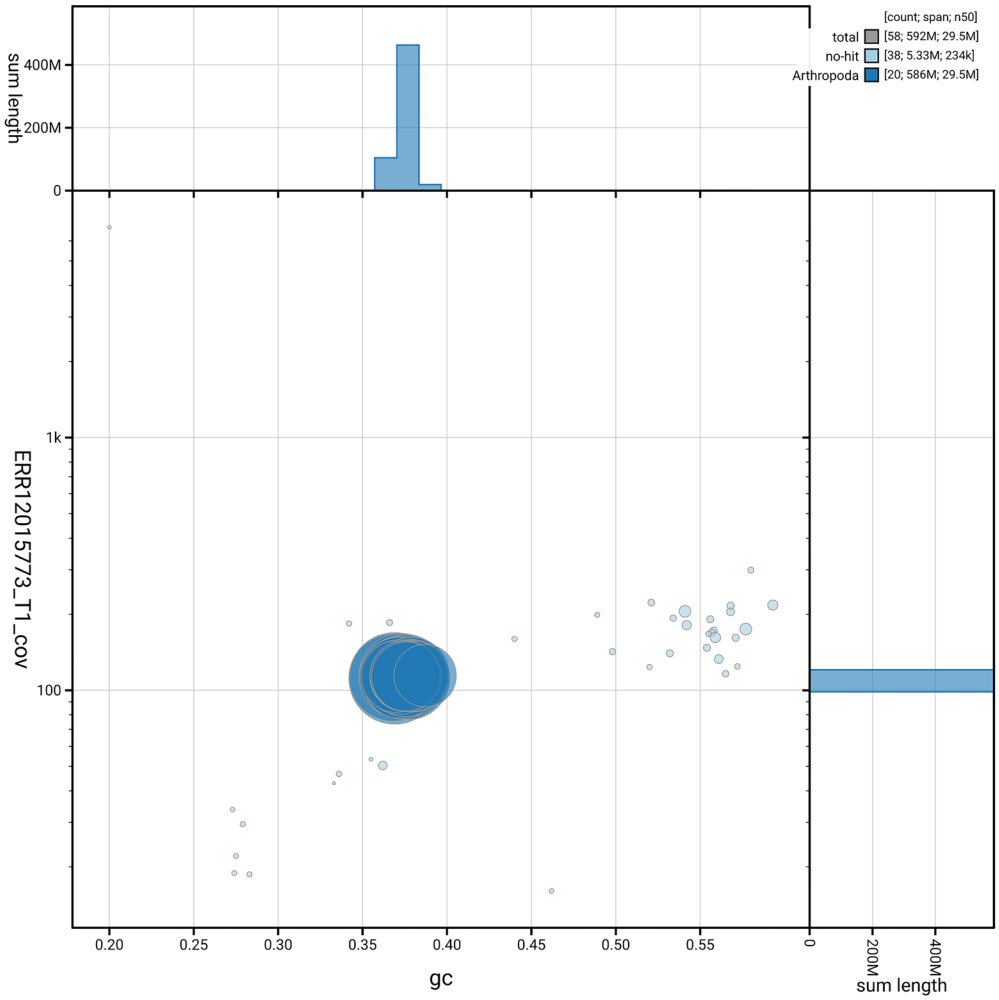

Figure 3. Genome assembly of *Philereme transversata*, ilPhiTran2.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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 at https://blobtoolkit.genomehubs.org/view/GCA_963924445.1/dataset/GCA_963924445.1/blob.

sequencing (specimen ID Ox000695, ToLID ilPhiTran1). This specimen was collected from the same location on 2020-07-20, using a light trap. Both specimens were collected and identified by Douglas Boyes and preserved on dry ice.

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 each specimen and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI) (Pereira *et al.*, 2022). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding have been deposited on protocols.io (Beasley *et al.*, 2023).

Metadata collection for samples adhered to the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification (Howard *et al.*, 2025). Detailed protocols are available on protocols.io (Denton *et al.*, 2023b). The ilPhiTran2 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the head

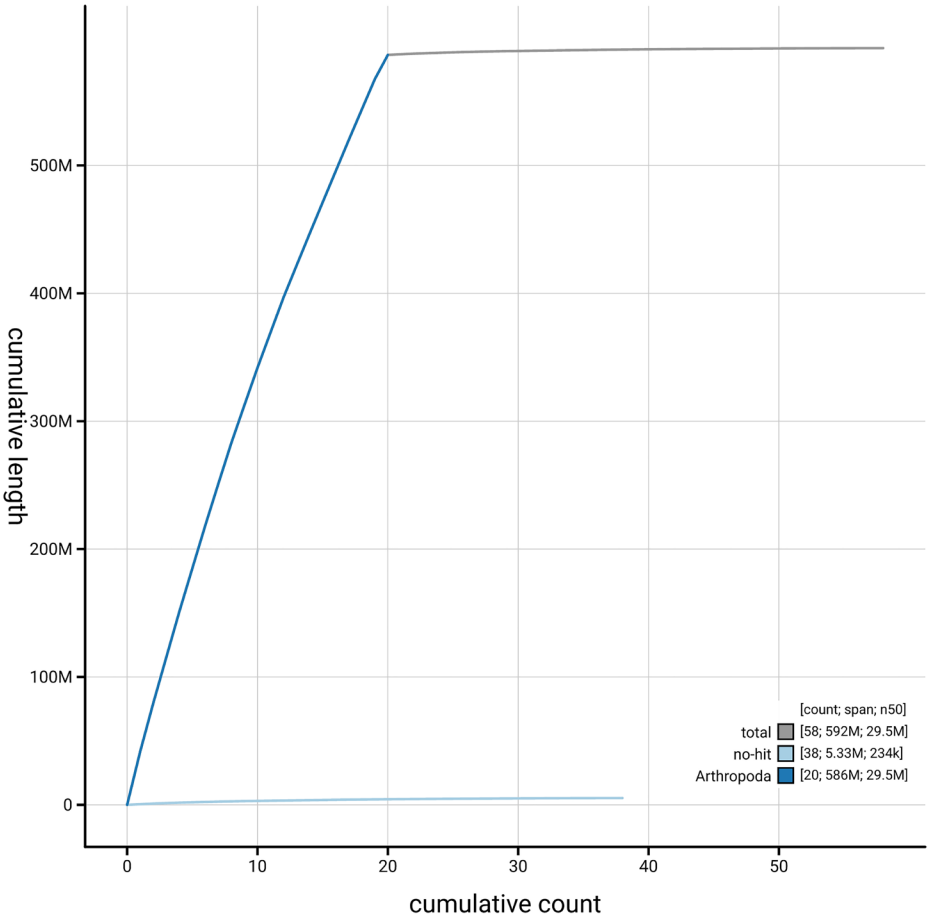

Figure 4. Genome assembly of *Philereme transversata*, ilPhiTran2.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963924445.1/dataset/GCA_963924445.1/cumulative.

and thorax was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a).

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023). The DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Bates *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Strickland *et al.*, 2023). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

Hi-C sample preparation and crosslinking

Hi-C data were generated from the head and thorax of the ilPhiTran1 sample using the Arima-HiC v2 kit (Arima Genomics) with 20–50 mg of frozen tissue (stored at –80 °C). As per manufacturer's instructions, tissue was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde concentration, and a final formaldehyde concentration of 2%. The tissue was then homogenised using the Diagnocine Power Masher-II. The crosslinked DNA was digested using a restriction enzyme master mix, then biotinylated and ligated. A clean up was performed with SPRIselect beads prior to library preparation. DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit, and sample biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core.

PacBio HiFi

At a minimum, samples were required to have an average fragment size exceeding 8 kb and a total mass over 400 ng to proceed to the low-input SMRTbell Prep Kit 3.0 protocol

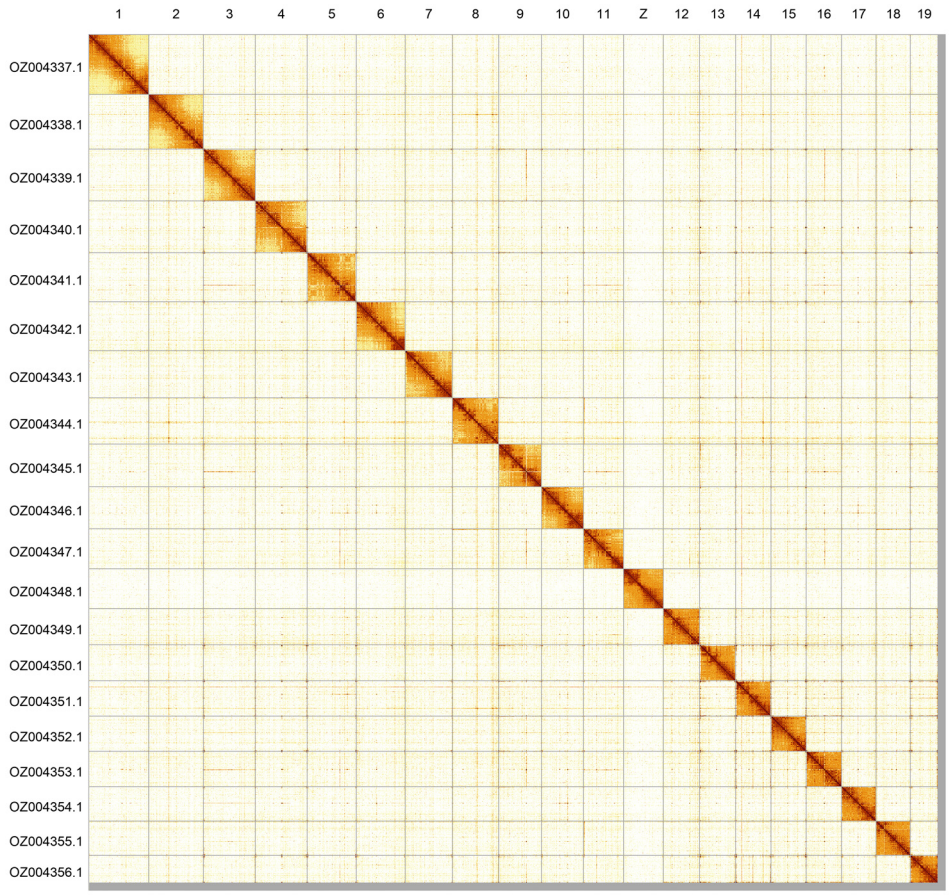

Figure 5. Genome assembly of *Philereme transversata*: Hi-C contact map of the ilPhiTran2.1 assembly, generated using PretextSnapshot. Chromosomes are shown in order of size and labelled with chromosome numbers (top) and chromosome accession numbers (left).

Table 3. Chromosomal pseudomolecules in the genome assembly of *Philereme transversata*, ilPhiTran2.

INSDC accession	Name	Length (Mb)	GC%
OZ004337.1	1	41.65	37
OZ004338.1	2	37.73	37
OZ004339.1	3	35.9	37.5
OZ004340.1	4	35.67	37.5
OZ004341.1	5	33.95	37.5
OZ004342.1	6	33.78	37.5
OZ004343.1	7	32.56	37
OZ004344.1	8	31.97	37
OZ004345.1	9	29.48	37.5
OZ004346.1	10	29.14	37.5
OZ004347.1	11	27.59	38
OZ004349.1	12	25.0	37
OZ004350.1	13	24.82	38
OZ004351.1	14	24.5	38
OZ004352.1	15	24.37	38
OZ004353.1	16	24.32	37.5
OZ004354.1	17	23.83	37.5
OZ004355.1	18	23.77	37.5
OZ004356.1	19	18.97	38.5
OZ004348.1	Z	27.43	37
OZ004357.1	MT	0.02	21

(Pacific Biosciences), depending on genome size and sequencing depth required. Libraries were prepared using the SMRTbell Prep Kit 3.0 as per the manufacturer's instructions. The kit

includes the reagents required for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead cleanup, and nuclease treatment. Size-selection and clean-up were carried out using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using the Qubit Fluorometer v4.0 (ThermoFisher Scientific) with Qubit 1X dsDNA HS assay kit and the final library fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and the gDNA 55kb BAC analysis kit.

Samples were sequenced on a Revio instrument (Pacific Biosciences, California, USA). Prepared libraries were normalised to 2 nM, and 15 µL was used for making complexes. Primers were annealed and polymerases were bound to create circularised complexes according to manufacturer's instructions. The complexes were purified with the 1.2X clean up with SMRTbell beads. The purified complexes were then diluted to the Revio loading concentration (in the range 200–300 pM), and spiked with a Revio sequencing internal control. Samples were sequenced on Revio 25M SMRT cells (Pacific Biosciences, California, USA). The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, as well as perform primary and secondary analysis of the data upon completion.

Hi-C

For Hi-C library preparation, the biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size-selected to 400–600 bp using SPRISelect beads. DNA was then enriched using Arima-HiC v2 Enrichment beads. The NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) was used for end repair, A-tailing, and adapter ligation, following a modified protocol in which library preparation is carried out while the DNA remains bound to the enrichment beads. PCR amplification was performed using KAPA HiFi HotStart mix and custom dual-indexed adapters (Integrated DNA Technologies) in a 96-well plate format. Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, samples were amplified for 10–16 PCR cycles. Post-PCR clean-up was carried out using SPRISelect beads. The libraries were quantified using the Accuclear Ultra High Sensitivity dsDNA Standards Assay kit (Biotium) and normalised to 10 ng/µL before sequencing. Hi-C sequencing was performed on the Illumina NovaSeq 6000 instrument using 150 bp paired-end reads.

Genome assembly, curation and evaluation

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 first assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). The Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019), and the contigs were scaffolded 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. Flat files and maps used in curation were generated via the TreeVal pipeline (Pointon *et al.*, 2023). Manual curation was conducted primarily in PretextView (Harry, 2022) and HiGlass (Kerpedjiev *et al.*, 2018), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were amended, and duplicate sequences were tagged and removed. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation.

Assembly quality assessment

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

The genome was analysed in the blobtoolkit pipeline, a Nextflow (Di Tommaso *et al.*, 2017) port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads in SAMtools (Danecek *et al.*, 2021) and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoaT database (Challis *et al.*, 2023) to identify all matching BUSCO lineages to run BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineages, the pipeline aligns the BUSCO genes to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) with DIAMOND blastp (Buchfink *et al.*, 2021). The genome is also divided into chunks according to the density of the BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database using DIAMOND blastx. Genome sequences without a hit are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The blobtools suite combines all these 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), relying on the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure

(da Veiga Leprevost *et al.*, 2017), as well as the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

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

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

Table 4. Software tools: versions and sources.

Software tool	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	666652151335353eef2fcd58880bcef5bc2928e1	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	0.19.5-r587	https://github.com/chhylp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84aa44357826c0b6753eb28de	https://github.com/higlass/higlass
MerquryFK	d00d98157618f4e8d1a9190026b19b471055b22e	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	23.04.1	https://github.com/nextflow-io/nextflow
PretextSnapshot	-	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.4.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

website here. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project.

Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible. The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international)

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: Philereme transversata (dark umber). Accession number PRJEB65731; https://identifiers.org/ena.embl/PRJEB65731. The genome sequence is released openly for reuse. The *Philereme transversata* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665), 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 Table 1 and Table 2.

Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.12157525.

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

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

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

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: https://doi.org/10.5281/zenodo.14870789.

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

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

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

References

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

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

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

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, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24.
PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374.
PubMed Abstract | Publisher Full Text | Free Full Text

Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

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

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

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

Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.
PubMed Abstract | **Publisher Full Text**

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

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: **Philereme transversata**. *GBIF Backbone Taxonomy*. 2023; [Accessed 6 May 2025].
Reference Source

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

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

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

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**

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

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

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

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. [Accessed 2 April 2024].
Reference Source

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

Pereira L, Sivell O, Sivess L, *et al.*: **DToL Taxon-specific Standard Operating Procedure for the terrestrial and freshwater arthropods working group**. 2022.
Publisher Full Text

Pointon DL, Eagles W, Sims Y, *et al.*: **sanger-tol/treeval v1.0.0 – Ancient Atlantis.** 2023.
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 and Ireland's larger Moths.** Newbury: Pisces Publications, 2019.
Reference Source

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

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

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2024; **9**: 339.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

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

Waring P, Townsend M, Lewington R: **Field guide to the Moths of Great Britain and Ireland: third edition.** Bloomsbury Wildlife Guides, 2017.
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 21 July 2025

https://doi.org/10.21956/wellcomeopenres.26762.r124826

?

Arun Arumugaperumal
Rajalakshmi Engineering College, Chennai, Tamil Nadu, India

This is the first report of the high quality genome sequence of *Philereme transversata* also known as the Dark Umber moth. The assembly reported herewith is of size 591.75 Mb and the mitogenome is of size 16.07 kb. The authors also annotated a total of 12,207 protein coding genes. The photograph included as figure 1 is very clear and will be very helpful for readers.

"As part of the Darwin Tree of Life Project – which aims to generate high-quality reference genomes for all named eukaryotic species in Britain and Ireland to .to". Remove the extra 'to'.

The link given in figure 3 & 4 and on the same page. The authors can correct the link at this stage of the publication process.

The manual curation has brought the assembly length to 591.75 Mb (after 3.33% decrease). But what was the change in length after decontamination using ASCC pipeline?

Did the authors use transcriptomic data to improve the assembly?

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: Bioinformatics; Genomics

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

Reviewer Report 28 June 2025

https://doi.org/10.21956/wellcomeopenres.26762.r124818

Annabel Whibley
Grapevine Improvement, Bragato Research Institute, Lincoln, Lincoln, New Zealand

Boyes and colleagues report the genome assembly and genome annotation of the Dark Umber moth (*Philereme transversata*). Overall, this is a high quality assembly (achieving EBP standard 6.C.Q65). Over 99% of the assembly has been scaffolded into 20 pseudochromosomes with excellent QV and completeness. Specimen ID has been validated by barcode sequencing. The Data Note follows Darwin Tree of Life project protocols, assembly and annotation pipelines and reporting templates. Appropriate methods are used, metadata is comprehensive and public accession links are functional.

There is a minor typo in the background section (insertion of " .to") in this sentence: "... high-quality reference genomes for all named eukaryotic species in Britain and Ireland to .to support research..."

I would also query the reporting of the average transcript length as >18kb. This seems unfeasibly high. Could the meaning of this statistic be clarified in the template?

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, Evolution, 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.

Reviewer Report 28 June 2025

https://doi.org/10.21956/wellcomeopenres.26762.r124822

?

Yu-Feng Huang
National Chung Hsing University, Taichung, Taiwan

The manuscript presents a chromosomally complete genome assembly of the Dark Umber moth, *Phlereme transversata*, as part of the Darwin Tree of Life Project. The study outlines the collection, identification, sequencing, assembly, and annotation processes, employing both PacBio HiFi long-read sequencing and Illumina Hi-C short-read data. The resulting assembly spans 591.75 Mb, with 99.1% of the sequence scaffolded into 20 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome was also assembled. Annotation performed by Ensembl identified 12,207 protein-coding genes and 1,934 non-coding genes. The manuscript provides a detailed account of laboratory methods, quality control, and data availability, offering a valuable resource for Lepidoptera genomics and biodiversity research.
However, several issues should be addressed to improve the manuscript's clarity and scientific rigor:
Typographical and Formatting Issues:

- **Background Section**:
  - Paragraph 1: The sentence “It is present throughout Europe.” is overly brief and lacks context. Consider expanding to clarify ecological relevance or distribution.
  - Paragraph 3: Typo in “to .to support research…” — remove the duplicated "to".
- **Methods – Assembly Quality Assessment**:
  - Paragraph 2: The sentence “It aligns the PacBio reads in SAMtools (Danecek et al., 2021) and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size.” should be corrected to:
    “The PacBio reads were aligned using Minimap2 (Li, 2018), and SAMtools (Danecek et al., 2021) was used to compute coverage tracks for fixed-size genomic windows.”

Unit Consistency:

- Units such as Mb and Gb should be used consistently and appropriately throughout the manuscript. In particular, the size of the mitochondrial genome should be reported in **base pairs (bp)**, not in megabases or gigabases, to reflect its smaller size accurately.

Addressing the issues above will enhance the manuscript's clarity, readability, and scientific value.

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

Yes

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

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

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics, next-generation sequencing, genome assambly and gene annotation

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