



DATA NOTE

The genome sequence of a tephritid fly, *Philophylla caesio*

(Harris, 1780)

[version 1; peer review: 2 approved]

James McCulloch^{1,2},

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

¹University of Oxford, Oxford, England, UK²Tree of Life, Wellcome Sanger Institute, Hinxton, England, UK**v1** First published: 09 Jul 2025, 10:338
<https://doi.org/10.12688/wellcomeopenres.24331.1>Latest published: 09 Jul 2025, 10:338
<https://doi.org/10.12688/wellcomeopenres.24331.1>

Abstract

We present a genome assembly from a male specimen of *Philophylla caesio* (tephritid fly; Arthropoda; Insecta; Diptera; Tephritidae). The genome sequence has a total length of 777.43 megabases. Most of the assembly (98.51%) is scaffolded into 6 chromosomal pseudomolecules, including the X sex chromosome. The mitochondrial genome has also been assembled, with a length of 19.03 kilobases. Gene annotation of this assembly on Ensembl identified 31,396 protein-coding genes.

Keywords

Philophylla caesio, tephritid fly, genome sequence, chromosomal, Diptera



This article is included in the [Tree of Life](#) gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1	✓	✓
09 Jul 2025	view	view

1. **Panagiotis Ioannidis** , Foundation for Research & Technology - Hellas, Crete, Greece

2. **Danilo Trabudo do Amaral** , Universidade Federal de Sao Carlos, São Carlos, Brazil

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: McCulloch J: Investigation, Resources;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (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.

Copyright: © 2025 McCulloch J *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: McCulloch J, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective *et al.* **The genome sequence of a tephritid fly, *Philophylla caesio* (Harris, 1780) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, 10:338 <https://doi.org/10.12688/wellcomeopenres.24331.1>

First published: 09 Jul 2025, 10:338 <https://doi.org/10.12688/wellcomeopenres.24331.1>

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; Acalyptrata; Tephritoidea; Tephritidae; Trypetinae; Trypetini; Trypetina; *Philophylla*; *Philophylla caesio* (Harris, 1780) (NCBI:txid472889)

Background

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 support research, conservation, and the sustainable use of biodiversity – we present a chromosomally complete genome sequence for the tephritid fly, *Philophylla caesio*. 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 *Philophylla caesio* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 37.46 Gb (gigabases) from 3.86 million reads, which were used to assemble the genome. GenomeScope analysis estimated the haploid genome size at 681.23 Mb, with a heterozygosity of 1.24% and repeat content of 37.67%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 51 coverage. Hi-C sequencing produced 85.38 Gb from 565.43 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 174 misjoins or missing joins and removed 6 haplotypic duplications. These interventions reduced



Figure 1. Photograph of the *Philophylla caesio* (idPhiCaes1) specimen used for genome sequencing.

the total assembly length by 0.51%, decreased the scaffold count by 61.24%, and increased the scaffold N50 by 26.58%. The final assembly has a total length of 777.43 Mb in 80 scaffolds, with 445 gaps, and a scaffold N50 of 124.27 Mb (Table 2).

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 (98.51%) was assigned to 6 chromosomal-level scaffolds, representing 5 autosomes and the X sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3). During curation, chromosome X was identified based on read coverage statistics. The heterogametic sex was sampled, and this individual is XO. The following regions of the assembly have uncertain order and orientation: Chromosome 1 (~49–158 Mbp), Chromosome 2 (~46–56.5 Mbp), Chromosome 3 (~79.5–85 Mbp), Chromosome 4 (~47–54 Mbp), and Chromosome 5 (~18.4–25 Mbp). 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 63.4. The *k*-mer completeness is 80.87% for the primary haplotype and 75.69% for the alternate haplotype; and 97.28% for the combined primary and alternate assemblies. BUSCO v.5.5.0 analysis using the diptera_odb10 reference set ($n = 3,285$) identified 98.6% of the expected gene set (single = 97.8%, duplicated = 0.8%).

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.Q63**.

Genome annotation report

The *Philophylla caesio* genome assembly (GCA_963971545.1) was annotated externally by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 32,282 transcribed mRNAs from 31,396 protein-coding genes. The average transcript length is 3,386.81 bp, with 2.97 exons per

Table 1. Specimen and sequencing data for *Philophylla caesio*.

Project information			
Study title	Philophylla caesio		
Umbrella BioProject	PRJEB70504		
Species	Philophylla caesio		
BioSpecimen	SAMEA112232776		
NCBI taxonomy ID	472889		
Specimen information			
Technology	ToLID	BioSample accession	Organism part
PacBio long read sequencing	idPhiCaes1	SAMEA112233277	whole organism
Hi-C sequencing	idPhiCaes1	SAMEA112233277	whole organism
Sequencing information			
Platform	Run accession	Read count	Base count (Gb)
Hi-C Illumina NovaSeq 6000	ERR12321264	5.65e+08	85.38
PacBio Revio	ERR12340372	3.86e+06	37.46

Table 2. Genome assembly data for *Philophylla caesio*.

Genome assembly		
Assembly name	idPhiCaes1.1	
Assembly accession	GCA_963971545.1	
Alternate haplotype accession	GCA_963971555.1	
Assembly level for primary assembly	chromosome	
Span (Mb)	777.43	
Number of contigs	525	
Number of scaffolds	80	
Longest scaffold (Mb)	222.32	
Assembly metric	Measure	Benchmark
Contig N50 length	3.97 Mb	≥ 1 Mb
Scaffold N50 length	124.27 Mb	= chromosome N50
Consensus quality (QV)	Primary: 63.6; alternate: 63.3; combined: 63.4	≥ 40
k-mer completeness	Primary: 80.87%; alternate: 75.69%; combined: 97.28%	≥ 95%
BUSCO*	C:98.6%[S:97.8%,D:0.8%], F:0.4%,M:0.9%,n:3,285	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	98.51%	≥ 90%
Sex chromosomes	X	localised homologous pairs
Organelles	Mitochondrial genome: 19.03 kb	complete single alleles

* BUSCO scores based on the diptera_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.

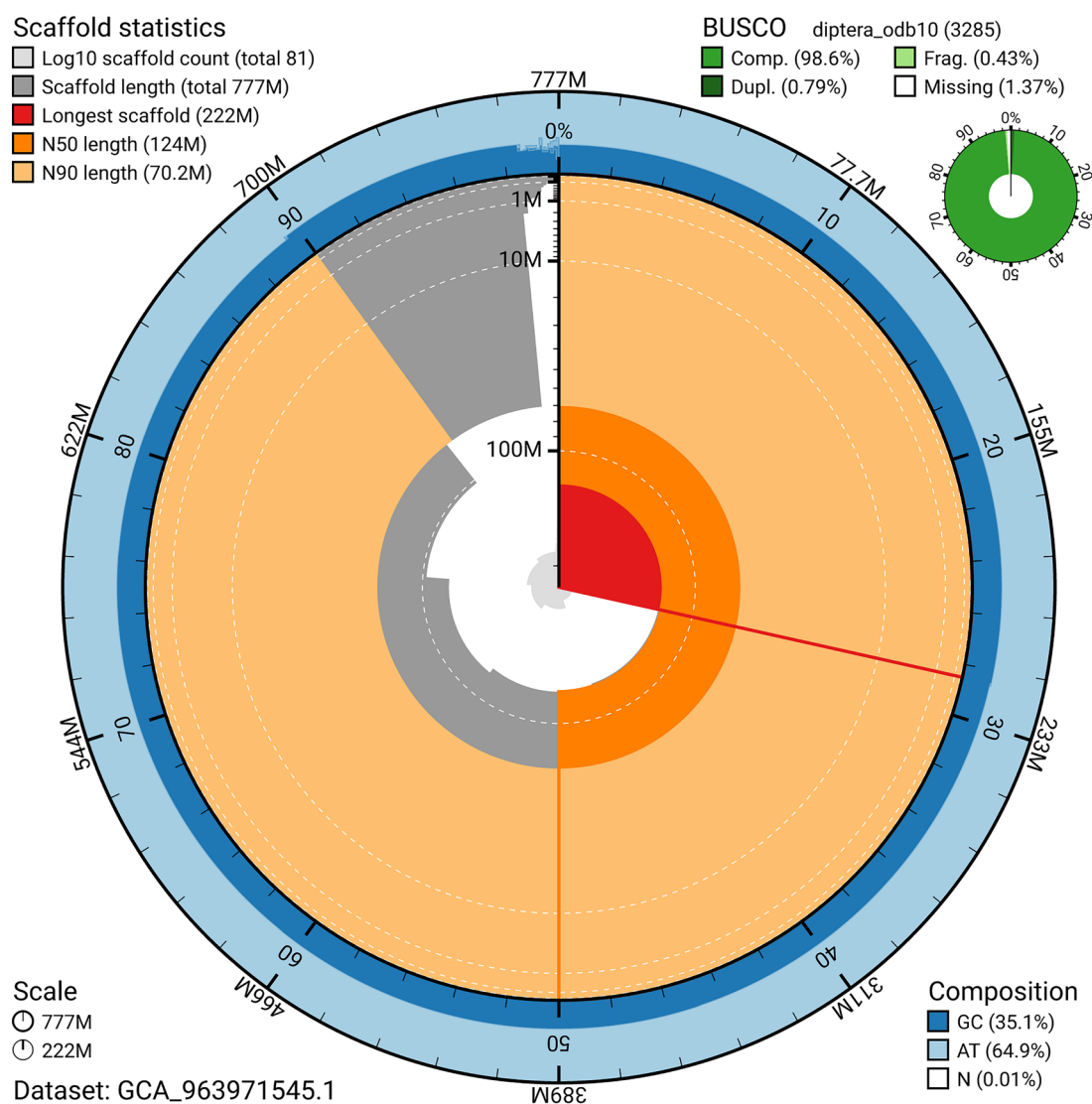


Figure 2. Genome assembly of *Philophylla caesio*, idPhiCaes1.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. 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 diptera_odb10 set is presented at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963971545.1/dataset/GCA_963971545.1/snail.

transcript. For further information about the annotation, please refer to <https://beta.ensembl.org/species/f9b5431c-2506-4b4b-be38-92000fb439db>.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Philophylla caesio* (specimen ID Ox002589, ToLID idPhiCaes1), collected from Wytham Woods, Oxfordshire, United Kingdom (latitude 51.772, longitude -1.338) on 2022-07-29 by netting.

The specimen was collected and identified by James McCulloch (University of Oxford) 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 the 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

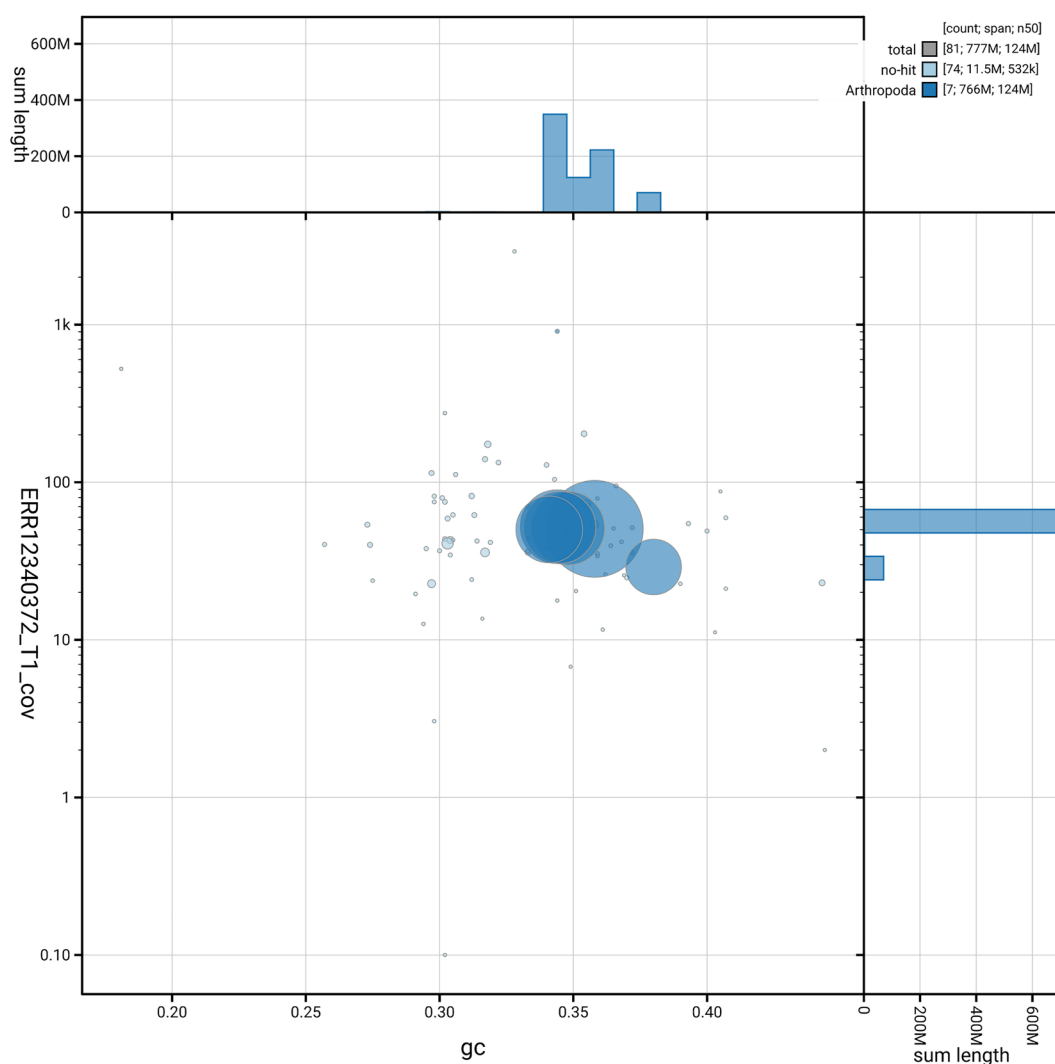


Figure 3. Genome assembly of *Philophylla caesio*, idPhiCaes1.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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_963971545.1/dataset/GCA_963971545.1/blob.

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 idPhiCaes1 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the whole organism 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.*, 2023a). For ultra-low input (ULI) PacBio sequencing, DNA was fragmented using the Covaris g-TUBE method (Oatley *et al.*, 2023b). Sheared DNA was purified by solid-phase reversible

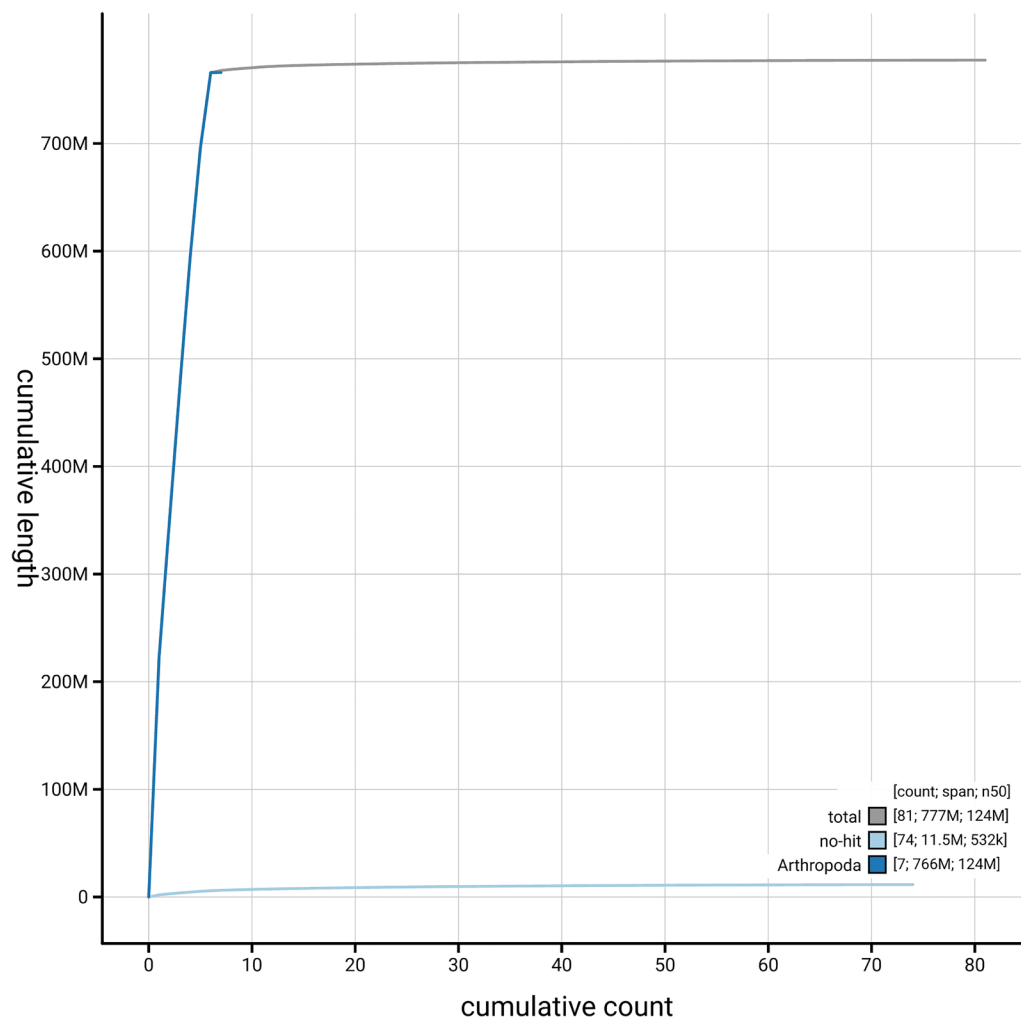


Figure 4. Genome assembly of *Philophylla caesio*, idPhiCaes1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963971545.1/dataset/GCA_963971545.1/cumulative.

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 whole organism of the idPhiCaes1 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

A ULI library was prepared using PacBio SMRTbell® Express Template Prep Kit 2.0 and PacBio SMRTbell® gDNA Sample Amplification Kit. To begin, samples were normalised to 20 ng of DNA. Initial removal of single-strand overhangs, DNA damage repair, and end repair/A-tailing were

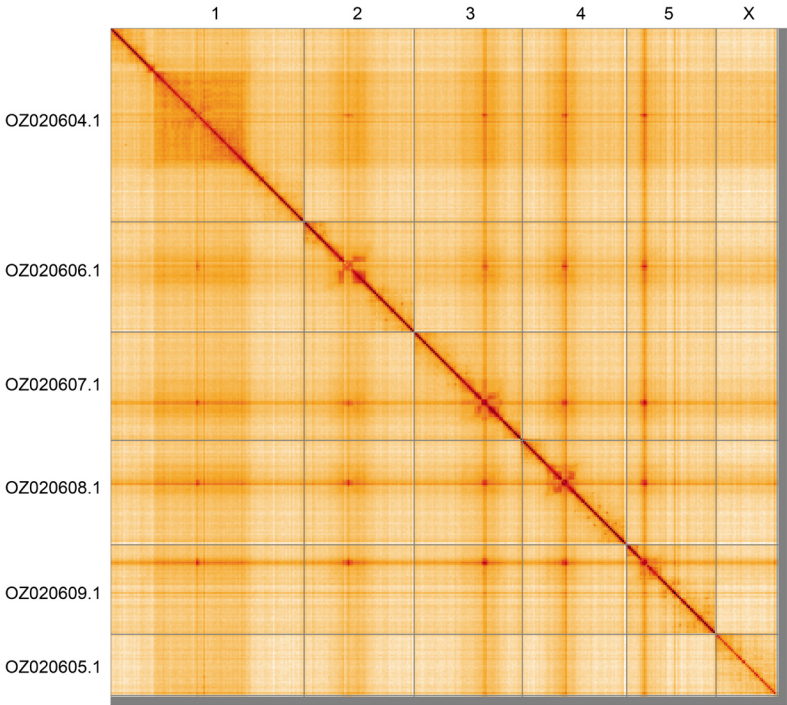


Figure 5. Genome assembly of *Philophylla caesio*: Hi-C contact map of the idPhiCaes1.1 assembly, generated using HiGlass Chromosomes are shown in order of size and labelled along the axes. An interactive version is available here: https://genome-note-higlass.tol.sanger.ac.uk/l/?d=K7cLj8rT_WDPDVssux0Yw.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Philophylla caesio*, idPhiCaes1.

INSDC accession	Name	Length (Mb)	GC%
OZ020604.1	1	222.32	36
OZ020606.1	2	126.41	34.5
OZ020607.1	3	124.27	35
OZ020608.1	4	120.0	34.5
OZ020609.1	5	102.65	34
OZ020605.1	X	70.21	38
OZ020610.1	MT	0.02	19

performed per manufacturer’s instructions. From the SMRTbell® gDNA Sample Amplification Kit, amplification adapters were then ligated. A 0.85X pre-PCR clean-up was performed with Promega ProNex beads and the sample was then divided into two for a dual PCR. PCR reactions A and B each followed the PCR programs as described in the manufacturer’s protocol. A 0.85X post-PCR clean-up was performed with ProNex beads for PCR reactions A and B and DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit and fragment size analysis

was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and gDNA 55kb BAC analysis kit. PCR reactions A and B were then pooled, ensuring the total mass was ≥500 ng in 47.4 µl. The pooled sample then repeated the process for DNA damage repair, end repair/A-tailing and additional hairpin adapter ligation. A 1X clean-up was performed with ProNex beads and DNA concentration was quantified using the Qubit and fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies). Size selection was performed using the PippinHT system (Sage Science) with target fragment size determined by analysis from the Femto Pulse, usually a value between 4000 and 9000 bp. Size-selected libraries were then cleaned-up using 1.0X ProNex beads and normalised to 2 nM before proceeding to 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

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. Sex

chromosomes were identified by read coverage statistics. 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 using the BlobToolKit pipeline, a Nextflow (Di Tommaso *et al.*, 2017) implementation of the earlier Snakemake BlobToolKit pipeline (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. Simultaneously, it queries the GoAT database (Challis *et al.*, 2023) to identify relevant BUSCO lineages and runs BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO 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 package management via Conda and Bioconda (Grünig *et al.*, 2018), and containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

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 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

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
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	44086069ee7d4d3f6f3f0012569789ec138f42b84a a44357826c0b6753eb28de	https://github.com/higlass/higlass
MercuryFK	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
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
PretextViewSnapshot	-	https://github.com/sanger-tol/PretextViewSnapshot
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

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: *Philophylla caesio*. Accession number PRJEB70504; <https://identifiers.org/ena.embl/PRJEB70504>. The genome sequence is released openly for reuse. The *Philophylla caesio* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665) and Sanger Institute Tree of Life Programme (PRJEB43745). 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 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](#)
- 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](#)
- 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](#)
- Diesch 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**(17): 276–278. [PubMed Abstract](#) | [Publisher Full Text](#)
- Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 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): g1aa153. [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](#)
- Lawnczak 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, Seppely 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.* 2023a.

[Publisher Full Text](#)

Oatley G, Sampaio F, Kitchin L, *et al.*: **Sanger Tree of Life HMW DNA fragmentation: covaris g-TUBE for ULI PacBio.** *protocols.io.* 2023b; [Accessed 13 June 2024].

[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](#)

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.*: **Merquy: 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, Krashenninnikova 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](#)

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 02 January 2026

<https://doi.org/10.21956/wellcomeopenres.26832.r142825>

© 2026 Amaral D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Danilo Trabudo do Amaral 

Universidade Federal de Sao Carlos, São Carlos, State of São Paulo, Brazil

This manuscript presents a high-quality, chromosomally resolved genome assembly of *Philophylla caesio*, generated using a robust and well-established PacBio HiFi and Hi-C pipeline. The assembly quality metrics (BUSCO, Merqury, BlobToolKit) indicate that the genome meets current reference standards and is suitable for downstream analyses.

As a minor but important suggestion, I recommend assessing the quality of the annotated gene set by running BUSCO on the predicted protein sequences. Given that no RNA-seq data were generated, reporting BUSCO completeness for the annotation would provide valuable information on the reliability of the gene predictions and further strengthen the utility of this dataset.

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: 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 21 August 2025

<https://doi.org/10.21956/wellcomeopenres.26832.r128244>

© 2025 Ioannidis P. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Panagiotis Ioannidis 

Foundation for Research & Technology - Hellas, Crete, Greece

This paper reports the sequencing and assembly of the genome of a tephritid fly.

The methodology used for assembling this genome is the standard one used by all DToL projects (PacBio + Hi-C). Thus, the assembly produced is of adequate quality for any downstream analyses, as it is witnessed by the standard QC analyses carried out here (BUSCO, MerquryFK).

I do have one rather important comment. A gene set was predicted for this genome assembly which resulted in ~32K genes. However, without including the BUSCO score for this gene set one cannot know the quality (and thus suitability) of this gene set for downstream analyses. And this is especially true for the current species, since no RNAseq data were generated, which could help improve the accuracy of gene prediction. In other words, if you feel "obliged" to run BUSCO for your genome assembly, then you should also feel obliged to run BUSCO for your gene set (should you generated one).

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: bioinformatics, insect 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.