



DATA NOTE

The genome sequence of the Dune Robberfly, *Philonicus*

albiceps (Meigen, 1820)

[version 1; peer review: 3 approved]

Olga Sivell ¹, Erica McAlister¹, Ryan Mitchell²,
Natural History Museum Genome Acquisition Lab,
Darwin Tree of Life Barcoding collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory
team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

¹Natural History Museum, London, England, UK

²Independent researcher, Sligo, Sligo County, Ireland

V1 First published: 14 Jul 2025, 10:349
<https://doi.org/10.12688/wellcomeopenres.24346.1>
Latest published: 14 Jul 2025, 10:349
<https://doi.org/10.12688/wellcomeopenres.24346.1>

Abstract

We present a genome assembly from a female specimen of *Philonicus albiceps* (Dune Robberfly; Arthropoda; Insecta; Diptera; Asilidae). The genome sequence has a total length of 253.92 megabases. Most of the assembly (98.91%) is scaffolded into 6 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 16.9 kilobases. Gene annotation of this assembly on Ensembl identified 11,887 protein-coding genes.

Keywords

Philonicus albiceps, Dune Robberfly, genome sequence, chromosomal, Diptera



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

Open Peer Review

Approval Status

	1	2	3
version 1			
14 Jul 2025	view	view	view

1. **Darren Obbard** , The University of Edinburgh, Edinburgh, UK
2. **Laurence Despres** , Laboratoire d'Ecologie Alpine (Ringgold ID: 56837), Grenoble, France
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: **Sivell O:** Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **McAlister E:** Investigation, Resources; **Mitchell R:** 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 Sivell O *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: Sivell O, McAlister E, Mitchell R *et al.* **The genome sequence of the Dune Robberfly, *Philonicus albiceps* (Meigen, 1820) [version 1; peer review: 3 approved]** Wellcome Open Research 2025, 10:349 <https://doi.org/10.12688/wellcomeopenres.24346.1>

First published: 14 Jul 2025, 10:349 <https://doi.org/10.12688/wellcomeopenres.24346.1>

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Asiloidea; Asilidae; Asilinae; *Philonicus*; *Philonicus albiceps* (Meigen, 1820) (NCBI:txid468783)

Background

Philonicus albiceps (Meigen, 1820) is a large, stout fly, with a body length of 13–18 mm and wing length 10–13 mm. It is quite dark, greyish or greyish yellow brown without a tessellated pattern on abdomen and has black legs with mostly white bristles. It is a habitat specific and morphologically very distinctive species, readily identifiable by the single pair of white bristles on scutellum. Other similar looking species have black bristles on the legs and scutellum or a larger number of white bristles on the scutellum. The male and female genitalia are also distinctive (Stubbs & Drake, 2014).

Philonicus albiceps has a Palearctic distribution. It is common and widespread on coast in England, Wales and Ireland, and in Scotland up to Ayr and the Firth of Forth. *P. albiceps* is common on sand dunes, usually close to the shore where it can be spotted sitting on bare sand and among marram grass. This species is occasionally collected from inland heathland near the coast of Norfolk and Suffolk. The adults appear from mid-May to mid-October, peaking from mid-June to mid-July (Stubbs & Drake, 2014).

The eggs are laid in the sand among marram grass roots. Once the female finishes ovipositing, she sweeps the sand back to cover the hole with a “brush” of short spines and two long upturned bristles on the anal lamellae and with the hardened tip of the eighth sternite (Oldroyd, 1969). The larva and pupa of *P. albiceps* have been described by Melin (1923). Adult asilids are venomous and inject their prey with specially adapted mouthparts leading to quick death (Drukewitz *et al.*, 2018; Whitfield, 1925). They hunt flying insects, mainly fairly large species such as calyptate flies, hoverflies, crane flies and other robberflies. Prey records also include a grasshopper, a shield bug, a beetle, solitary bees and the Common Blue Damsel fly *Enallagma cyathigerum* (Stubbs & Drake, 2014; Yarrow, 1937).

The high-quality genome of *Philonicus albiceps* was sequenced from a single specimen (NHMUK014438339, SAMEA11024992) from Penhale Dunes, England (Figure 1). The genome 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 Asilidae, as well as biology and ecology of the species.

Genome sequence report

Sequencing data

The genome of a specimen of *Philonicus albiceps* (Figure 1) was sequenced using Pacific Biosciences single-molecule

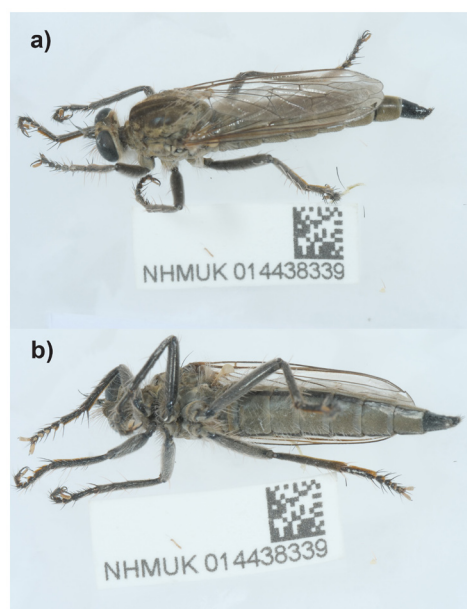


Figure 1. Photograph of the *Philonicus albiceps* (idPhiAlbi1) specimen used for genome sequencing.

HiFi long reads, generating 25.14 Gb (gigabases) from 2.69 million reads, which were used to assemble the genome. GenomeScope analysis estimated the haploid genome size at 250.80 Mb, with a heterozygosity of 0.54% and repeat content of 9.87%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 91× coverage. Hi-C sequencing produced 110.71 Gb from 733.19 million reads, used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. 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 37 misjoins or missing joins and removed four haplotypic duplications. These interventions decreased the scaffold count by 40.82%. The final assembly has a total length of 253.92 Mb in 28 scaffolds, with 201 gaps, and a scaffoldN50 of 57.53 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 (98.91%) was assigned to 6 chromosomal-level scaffolds. These chromosome-level scaffolds,

Table 1. Specimen and sequencing data for *Philonicus albiceps*.**Table 2. Genome assembly data for *Philonicus albiceps*.**

Genome assembly		
Assembly name	idPhiAlbi1.1	
Assembly accession	GCA_963969385.1	
Alternate haplotype accession	GCA_963969375.1	
Assembly level for primary assembly	chromosome	
Span (Mb)	253.92	
Number of contigs	229	
Number of scaffolds	28	
Longest scaffold (Mb)	75.08	
Assembly metric	Measure	Benchmark
Contig N50 length	2.5 Mb	≥ 1 Mb
Scaffold N50 length	57.53 Mb	= chromosome N50
Consensus quality (QV)	Primary: 63.7; alternate: 63.8; combined: 63.7	≥ 40
k-mer completeness	Primary: 88.89%; alternate: 75.90%; combined: 99.34%	≥ 95%
BUSCO*	C:96.3%[S:95.8%,D:0.5%], F:0.9%,M:2.7%,n:3,285	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	98.91%	≥ 90%
Sex chromosomes	Not identified	localised homologous pairs
Organelles	Mitochondrial genome: 16.9 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.

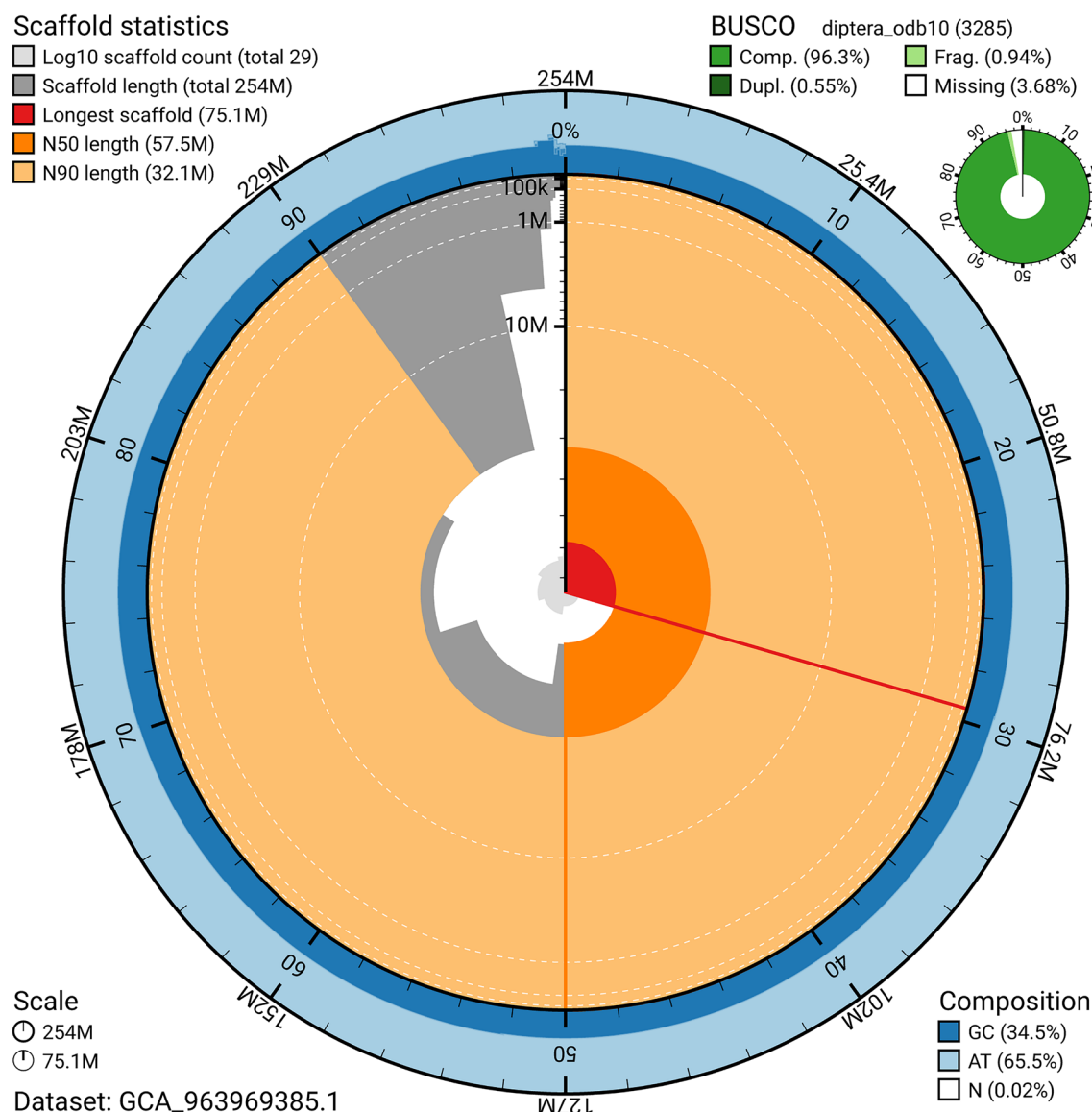


Figure 2. Genome assembly of *Philonicus albiceps*, idPhiAlbi1.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. 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_963969385.1/dataset/GCA_963969385.1/snail.

confirmed by Hi-C data, are named according to size (Figure 5; Table 3). During curation, we noted that read coverage suggests this is the homogametic sex, but we did not identify the sex chromosome(s) as sequence data from the heterogametic sex was not available and homology is unreliable for sex chromosome identification in Diptera due to frequent sex chromosome turnover (Vicoso & Bachtrog, 2015).

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.

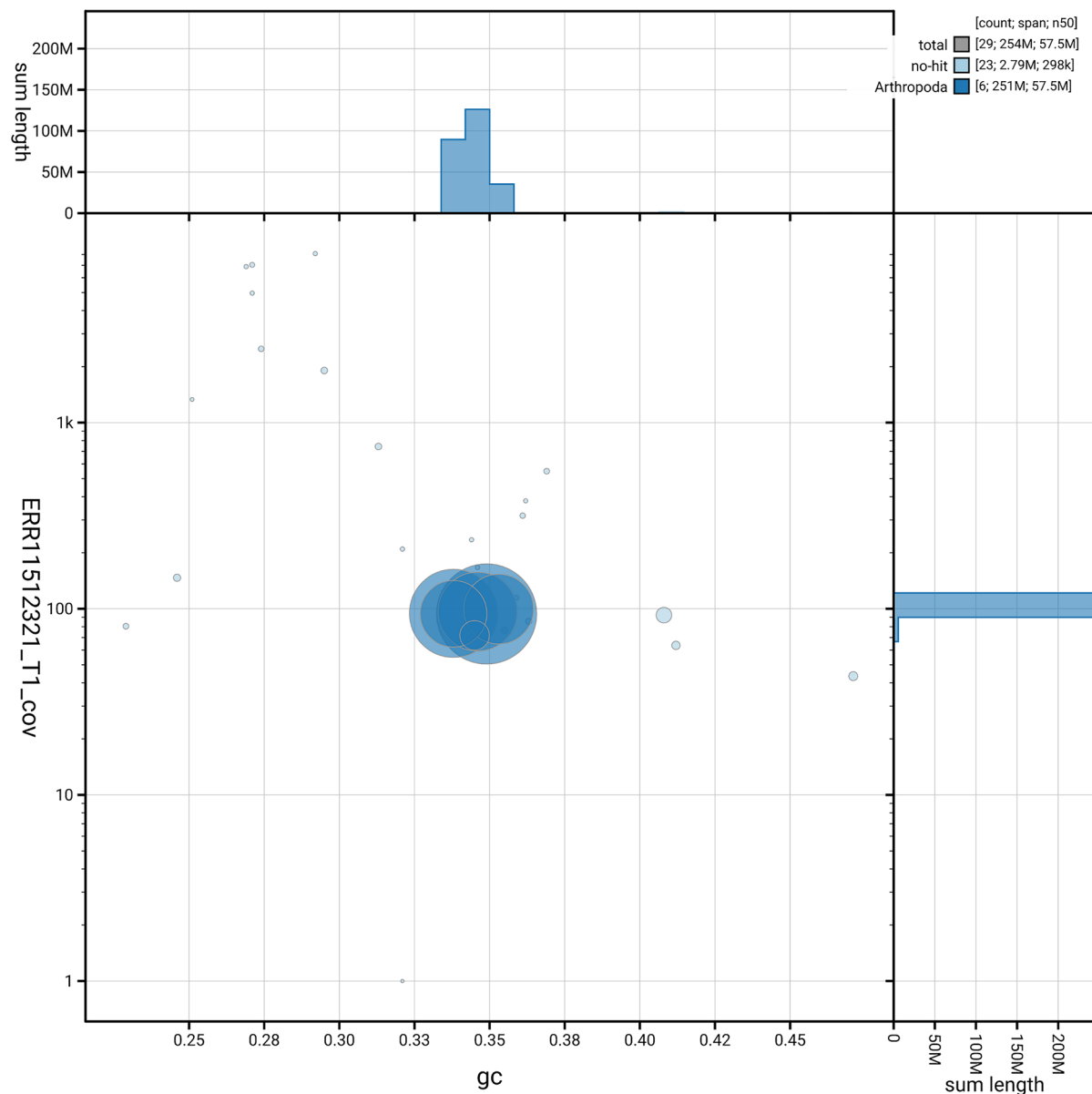


Figure 3. Genome assembly of *Philonicus albiceps*, idPhiAlbi1.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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_963969385.1/dataset/GCA_963969385.1/blob.

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 63.7. The k -mer completeness is 88.89%

for the primary haplotype and 75.90% for the alternate haplotype; and 99.34% for the combined primary and alternate assemblies. BUSCO v.5.5.0 analysis using the diptera_odb10 reference set ($n = 3,285$) identified 96.3% of the expected gene set (single = 95.8%, duplicated = 0.5%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on

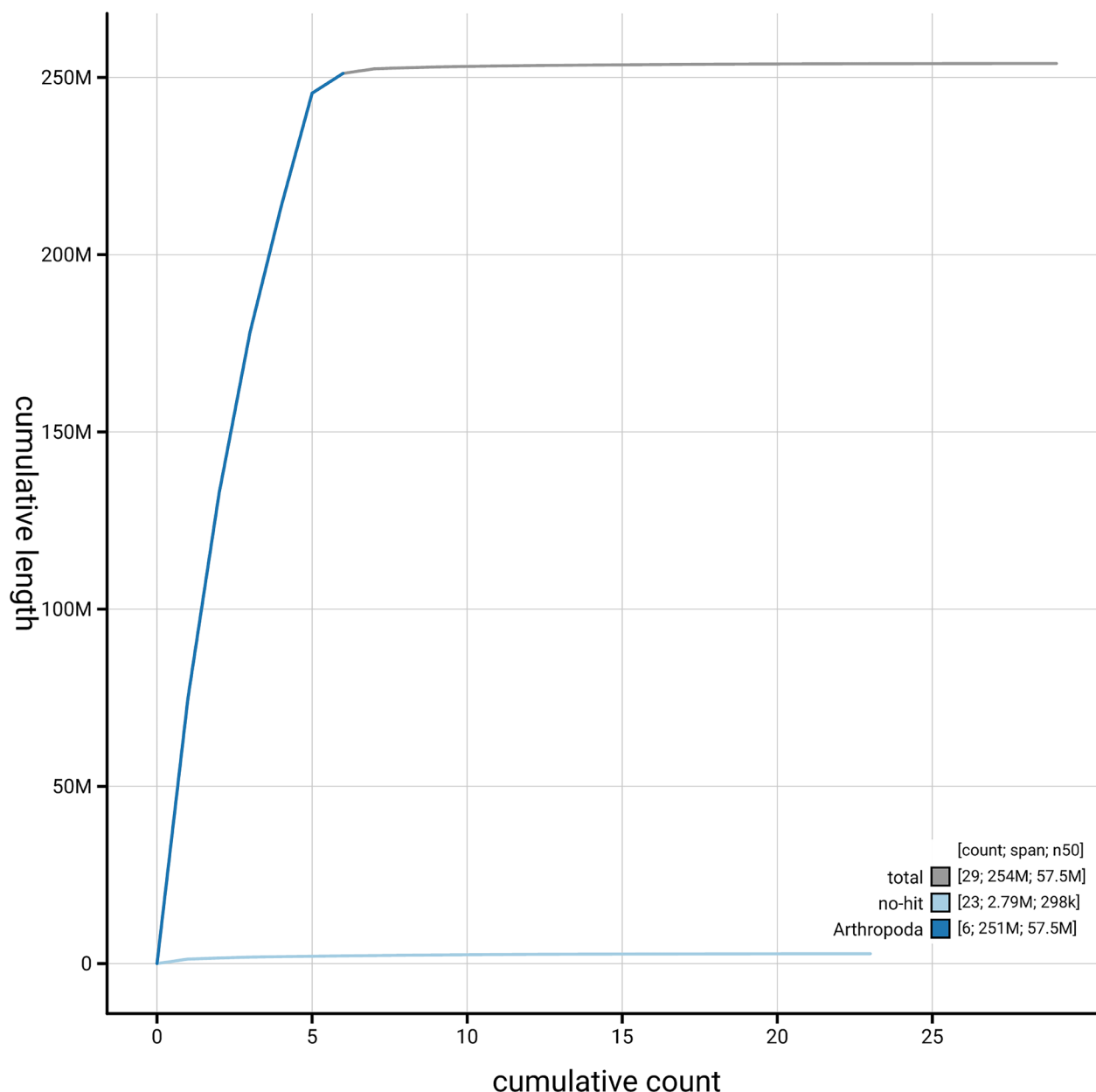


Figure 4. Genome assembly of *Philonicus albiceps*, idPhiAlbi1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963969385.1/dataset/GCA_963969385.1/cumulative.

Assembly Standards [September 2024](#). The primary assembly achieves the EBP reference standard of **6.C.Q63**.

Genome annotation report

The *Philonicus albiceps* genome assembly (GCA_963969385.1) was annotated externally by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 20,109 transcribed mRNAs from 11,887 protein-coding and 392 non-coding genes. The average transcript length is 10,937.85 bp.

There are 1.64 coding transcripts per gene and 6.16 exons per transcript. For further information about the annotation, please refer to <https://beta.ensembl.org/species/8e00c00e-705a-45a5-8db3-c90a26a78799>.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was a female *Philonicus albiceps* (specimen ID NHMUK014438339, ToLID idPhiAlbi1). The specimen was collected by Charles Griffiths

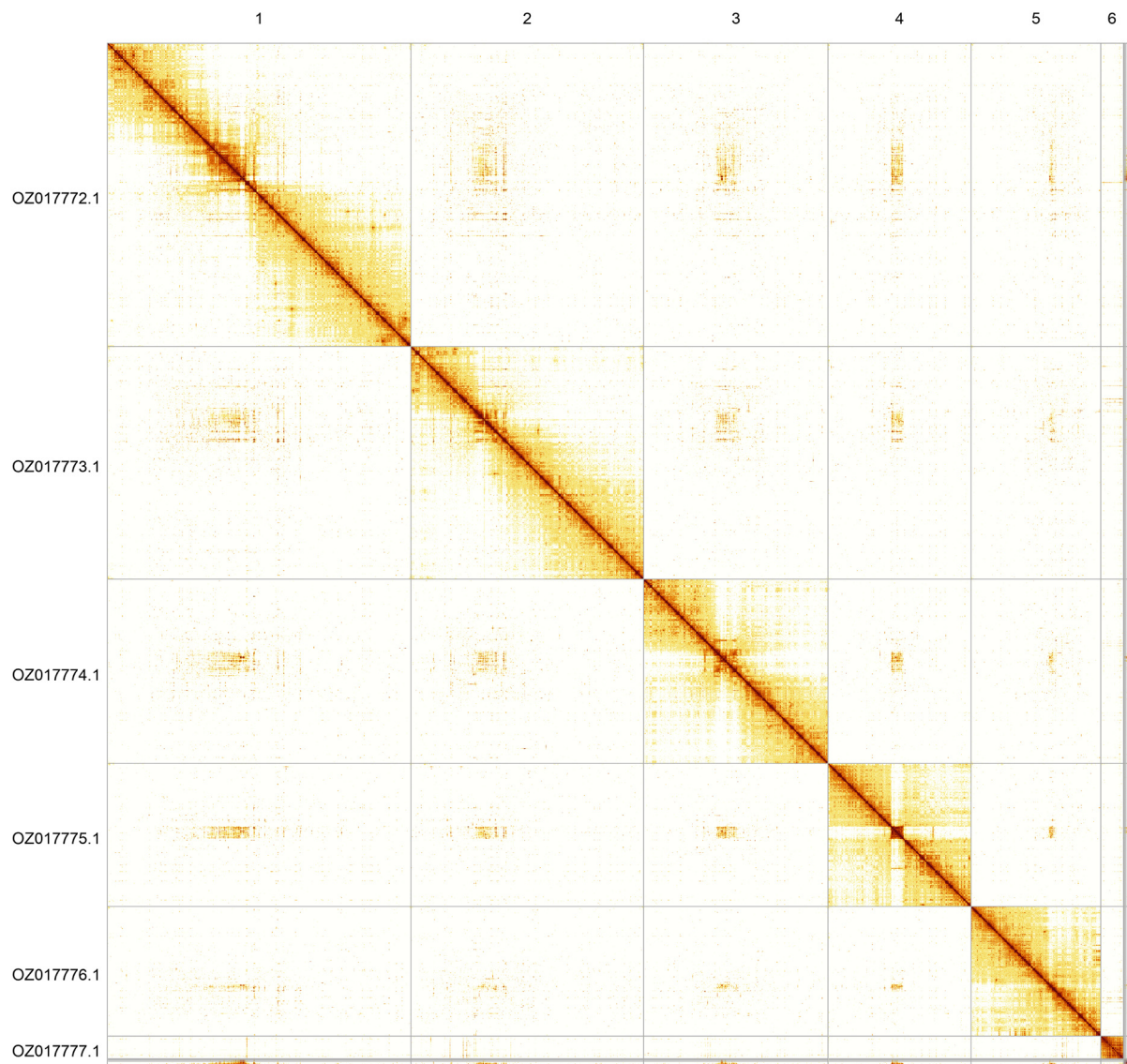


Figure 5. Genome assembly of *Philonicus albiceps*: Hi-C contact map of the idPhiAlbi1.1 assembly, generated using PretextSnapshot. Chromosomes are shown in order of size and labelled with chromosome numbers (top) and chromosome accession numbers (left).

(Dipterists Forum) and identified by Brian Levey (OUMNH). A second specimen used for Hi-C sequencing (specimen ID NHMUK014438340, ToLID idPhiAlbi2) was collected by Olga Sivell and Erica McAlister and identified by Ryan Mitchell. Both specimens were collected from Penhale Dunes, England, United Kingdom (latitude 50.3705, longitude -5.1373) on 2021-06-30, using an aerial net, and were preserved by dry freezing (-80°C).

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).

Table 3. Chromosomal pseudomolecules in the genome assembly of *Philonicus albiceps*, idPhiAlbi1.

INSDC accession	Name	Length (Mb)	GC%
OZ017772.1	1	75.08	35
OZ017773.1	2	57.53	34
OZ017774.1	3	45.56	34.5
OZ017775.1	4	35.33	35.5
OZ017776.1	5	32.05	34
OZ017777.1	6	5.6	34.5
OZ017778.1	MT	0.02	29

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 idPhiAlbi1 sample was prepared for DNA extraction by weighing and dissecting it on dry ice ([Jay et al., 2023](#)). Tissue from the head and thorax was homogenised using a PowerMasher II tissue disruptor ([Denton et al., 2023a](#)). HMW DNA was extracted using the Automated MagAttract v1 protocol ([Sheerin et al., 2023](#)). For ultra-low input (ULI) PacBio sequencing, DNA was fragmented using the Covaris g-TUBE method ([Oatley 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 a Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. The fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the extracted DNA had a Qubit concentration of 1.78 ng/μL and a yield of 694.20 ng. Spectrophotometric measurements indicated 260/280 and 260/230 ratios of 2.25 and 1.98, respectively.

RNA was extracted from abdomen tissue of idPhiAlbi1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ mirVana protocol ([do Amaral et al., 2023](#)). The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

Hi-C sample preparation and crosslinking

Hi-C data were generated from the head of the idPhiAlbi2 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 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

The sample requires Covaris g-TUBE shearing to approximately 10 kb prior to library preparation. Ultra-low input libraries were 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 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 Sage Sciences' PippinHT system 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.

Samples were sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library

loaded onto the Sequel IIe was in the range 40–135 pM. 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.

RNA

Poly(A) RNA-Seq libraries were prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs), following the manufacturer's instructions. Poly(A) mRNA in the total RNA solution was isolated using oligo(dT) beads, converted to cDNA, and uniquely indexed; 14 PCR cycles were performed. Libraries were size-selected to produce fragments between 100–300 bp. Libraries were quantified, normalised, pooled to a final concentration of 2.8 nM, and diluted to 150 pM for loading. Sequencing was carried out on the Illumina NovaSeq 6000 instrument.

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 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 (Grüning *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.

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	0.16.1-r375	https://github.com/chhy123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84aa44357826c0b6753eb28de	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

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 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: *Philonicus albiceps*. Accession number PRJEB62736; <https://identifiers.org/ena.embl/PRJEB62736>. The genome sequence is released openly for reuse. The *Philonicus albiceps* 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](#).

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](#)
- 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](#)
- do Amaral RJV, Bates A, Denton A, *et al.*: **Sanger Tree of Life RNA extraction: automated MagMax™ mirVana**. *protocols.io.* 2023. [Publisher Full Text](#)
- Drukewitz SH, Fuhrmann N, Undheim EAB, *et al.*: **A Dipteran's novel sucker punch: evolution of arthropod atypical venom with a neurotoxic component in robber flies (Asilidae Diptera)**. *Toxins (Basel).* 2018; **10**(1): 29. [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](#)
- Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive**

Author information

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>.</

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, Seppay 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](#)

Meigen JW: Systematische Beschreibung der bekannten europäischen zweiflügeligen Insekten. Zweiter Theil. Aachen: Forstmann, 1820. [Reference Source](#)

Melin DE: Contributions to the knowledge of the biology, metamorphosis and distribution of the Swedish asilids in relation to the whole family of asilids. Uppsala: Almqvist & Wiksells, 1923; 8. [Reference Source](#)

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, Sampaio F, Kitchin L, *et al.*: Sanger Tree of Life HMW DNA fragmentation: covaris g-TUBE for ULI PacBio. *protocols.io*. 2023; [Accessed 13 June 2024]. [Publisher Full Text](#)

Oldroyd H: Handbook for the identification of British insects: Diptera Brachycera. Sections Tabanoidea and Asiloidea. London: Royal Entomological Society, 1969. [Reference Source](#)

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](#)

Sheerin E, Sampaio F, Oatley G, *et al.*: Sanger Tree of Life HMW DNA extraction: automated MagAttract v.1. *protocols.io*. 2023. [Publisher Full Text](#)

Strickland M

Open Peer Review

Current Peer Review Status:   

Version 1

Reviewer Report 20 January 2026

<https://doi.org/10.21956/wellcomeopenres.26847.r142124>

© 2026 Arumugaperumal A. 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.



Arun Arumugaperumal 

Rajalakshmi Engineering College, Chennai, India

The authors have described the genome sequencing project of the Dune Robberfly, *Philonicus albiceps*. They have used a combination of PacBio long-read sequencing and Hi-C sequencing data to obtain a high-quality genome assembly. The assembly reported here is of size 253.92 Mb, spread among 6 chromosomes. A female specimen has been used for sequencing. But the sex chromosome was not identified. The mitochondrial genome of size 16.9 kb has also been assembled. The sequence data have been submitted to public databases. Annotations were obtained from the EMBL pipeline and 11,887 protein-coding genes have been reported. BUSCO analysis has shown that the sequences are 96.3% complete with respect to diptera_odb10 reference. The quality of the genome is evident from the number of genes predicted and the high N50 value. The links to datasets provided in the article are active.

The figures are not visible in the online version of the article.

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; 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.

Reviewer Report 31 December 2025

<https://doi.org/10.21956/wellcomeopenres.26847.r142129>

© 2025 Despres L. 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.



Laurence Despres 

Laboratoire d'Ecologie Alpine (Ringgold ID: 56837), Grenoble, Auvergne-Rhône-Alpes, France

The authors report on the chromosome-level genome assembly of a female specimen of the Dune Robberfly *Philonicus albiceps* (Diptera; Asilidae) a common sand habitat specialist found on the coasts throughout the Palaearctic. The genome sequence has a total length of 253.92 megabases. The assembly contains two haplotypes with most of haplotype 1 scaffolded into 6 chromosomes. The mitochondrial genome has also been assembled.

The quality of the assembly was further assessed by the proportion of complete BUSCO genes recovered, which is very high (96.6% of the expected gene set of the using the diptera_odb10 reference set). RNA sequencing data were also generated. The genome assembly is of good quality and was constructed using appropriate methods: Pac Bio HiFi long reads (91 X coverage) and Hi-C data to confirm scaffolding, following the high standard of the Darwin Tree of Life Project for results presentation.

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: evolutionary ecology and 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.

Reviewer Report 14 November 2025

<https://doi.org/10.21956/wellcomeopenres.26847.r138634>

© 2025 Obbard 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.

**Darren Obbard** 

The University of Edinburgh, Edinburgh, Scotland, UK

This data note reports the sequencing and assembly of the genome of *Philonicus albiceps* as part of the “Darwin Tree of Life” programme. In common with other data notes from this research effort, the reporting is standardized and quite brief. As such, I have very few comments to make.

The approach is state-of-the-art, the raw data appear to be of a suitably high quality, and the assembly methods are appropriate. The public availability of raw data and genome assembly are appropriate. The resulting genome is likely to be of very high quality, and I have no doubt that it will be of great value to any researchers working on asilids and other diptera, or on the comparative or evolutionary genomics of insects more generally.

The biological context and detail are very good, and much better than many of these genome reports.

My minor suggestions are:

- (1) In addition to the sequenced specimen itself, it would be nice to have an ‘in life’ photograph and/or perhaps both sexes. It would be especially nice to have one in which the diagnostic white bristles were visible. CC-BY images seem to be available from iNaturalist. Given the specific habitat, and habitat view might also be a nice addition.
- (2) Is Stubbs & Drake the intended reference for all of the statements in the paragraph beginning “*Philonicus albiceps* has a Palaearctic distribution ...”. The paragraph feels like it is lacking references, and this may be why – perhaps emphasize that all of the information is from a single source. Also, this paragraph contains a sentence beginning ‘P.’, i.e. starting with an abbreviation - which is generally considered bad practice
- (3) It is odd that the methods mention the individual was female, but final paragraph of the introduction does not.

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: Evolutionary genetics and genomics of invertebrates and their pathogens

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.
