DATA NOTE

The genome sequence of a muscid fly, *Phaonia angelicae* (Scopoli, 1763) (Diptera: Muscidae)

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

Steven Falk[1], Andrzej Grzywacz[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]Independent researcher, Kenilworth, England, UK
[2]Department of Ecology and Biogeography, Nicolaus Copernicus University in Toruń, Toruń, Poland

v1 **First published:** 22 Apr 2026, **11**:236
https://doi.org/10.12688/wellcomeopenres.26330.1
Latest published: 22 Apr 2026, **11**:236
https://doi.org/10.12688/wellcomeopenres.26330.1

Abstract

We present a genome assembly from an individual female *Phaonia angelicae* (muscid fly; Arthropoda; Insecta; Diptera; Muscidae). The assembly contains two haplotypes with total lengths of 1 593.88 megabases and 1 575.57 megabases. Most of haplotype 1 (97.48%) is scaffolded into 5 chromosomal pseudomolecules. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 17.82 kilobases. Gene annotation of this assembly on Ensembl identified 13 923 protein-coding genes. This assembly was generated as part of the Darwin Tree of Life project, which produces reference genomes for eukaryotic species found in Britain and Ireland.

Keywords

Phaonia angelicae, muscid fly, genome sequence, chromosomal, Diptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ? ✓ ✓

	1	2	3
version 1 22 Apr 2026	? view	✓ view	✓ view

1. **Rajib Deb**, ICAR-National Research Centre, Guwahat, India
2. **Jonathan Badger**, National Cancer Institute Center for Cancer Research (Ringgold ID: 272101), Bethesda, USA
3. **Stuart J.E. Baird**, Academy of Sciences of the Czech Republic, Národní, Czech Republic

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: Falk S: Investigation, Resources; **Grzywacz A**: Investigation, 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].
Andrzej Grzywacz received supported from the National Science Centre of Poland (grant no. 2019/33/B/NZ8/02316).
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: Falk S, Grzywacz A, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of a muscid fly, *Phaonia angelicae* (Scopoli, 1763) (Diptera: Muscidae) [version 1; peer review: 2 approved, 1 approved with reservations]** Wellcome Open Research 2026, **11**:236 https://doi.org/10.12688/wellcomeopenres.26330.1

First published: 22 Apr 2026, **11**:236 https://doi.org/10.12688/wellcomeopenres.26330.1

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; Calyptratae; Muscoidea; Muscidae; Phaoniinae; *Phaonia*; *Phaonia angelicae* (Scopoli, 1763) (NCBI:txid1511665).

Background

Phaonia angelicae (Scopoli, 1763) is a representative of the cosmopolitan and species-rich genus *Phaonia* Robineau-Desvoidy, 1830 (Diptera: Muscidae). Females typically lay eggs close to suitable breeding substrates such as moss, soil, decaying or damaged parts of trees or herbaceous plants, and in some cases also carrion and dung (Skidmore, 1985). Most probably in all species a third instar larva hatches from the egg and subsequently becomes an active predator of various small invertebrates (Skidmore, 1985).

Phaonia angelicae occurs throughout the Palaearctic region (Gregor *et al.*, 2002), and is distributed widely across Europe, extending eastwards to Siberia and the Korean Peninsula (Pont & Evenhuis, 2026). The species is frequently recorded across the the United Kingdom and Ireland, where adults are most commonly observed from June to October (d'Assis Fonseca, 1968, https://nbnatlas.org/). The species occurs in a variety of habitats, including meadows, clearings, rural habitats, forest edges, scrub and parks. Adults are among the anthophilous representatives of *Phaonia* (Pont, 1993) and frequently visit flowers. *Phaonia angelicae* has been recognised as an effective pollinator of plants such as *Heracleum sphondylium* L. (hogweed), although its contribution to pollination efficiency shows considerable seasonal variation (Zych, 2007). Additionally, adults are attracted to honeydew (Skidmore, 1985), and may also visit carrion and decomposing animal tissues (Grzywacz *et al.*, 2017). Larvae develop beneath moss and, based on larval morphology, are considered obligatory predators, yet the precise prey preferences and ecology of the immature stages remain unknown (Skidmore, 1985).

The genome presented here will provide a valuable resource for phylogenomic analyses within Muscidae and for comparative genomic studies investigating evolutionary diversification and ecological adaptations in *Phaonia* species and related dipterans.

We present a chromosome-level genome sequence for *Phaonia angelicae*, generated via the Tree of Life pipeline using a specimen collected from Wytham Woods, Oxfordshire, UK (Figure 1).

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult female *Phaonia angelicae* (specimen ID Ox000739, ToLID idPhaAnge1; Figure 1), collected from Wytham Woods, Oxfordshire, UK (latitude 51.766, longitude −1.309) on 2020-08-03. The specimen was collected and identified by Steven Falk.

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

Nucleic acid extraction

Detailed protocols for nucleic acid extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The idPhaAnge1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the abdomen was homogenised by powermashing using a PowerMasher II tissue disruptor.

High molecular weight (HMW) DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol. We used centrifuge-mediated fragmentation to produce DNA fragments in the 8–10 kb range, following the Covaris g-TUBE protocol for ultra-low input (ULI). Sheared DNA was purified by manual SPRI (solid-phase reversible immobilisation).

Figure 1. Photograph of the *Phaonia angelicae* (idPhaAnge1) specimen used for genome sequencing.

The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 29.8 ng/μL and a yield of 1 341.00 ng, with a fragment size of 12.3 kb. The Genomic Quality Number (GQN) was 6.3.

PacBio HiFi library preparation and sequencing

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

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

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

The sample was sequenced 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, and to perform primary and secondary analysis of the data upon completion.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen head tissue from the idPhaAnge1 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer

(Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

Biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size selected to 400–600 bp using SPRISelect beads. DNA was enriched with Arima-HiC v2 kit Enrichment beads. End repair, A-tailing, and adapter ligation were carried out with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs), following a modified protocol where library preparation occurs while DNA remains bound to the Enrichment beads. Library amplification was performed using KAPA HiFi HotStart mix and a custom Unique Dual Index (UDI) barcode set (Integrated DNA Technologies). Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, libraries were amplified with 10–16 PCR cycles. Post-PCR clean-up was performed with SPRISelect beads. Libraries were quantified using the AccuClear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech).

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

Genome assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts ($k = 31$) was generated from the filtered reads using FastK. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were assembled using Hifiasm in Hi-C phasing mode (Cheng *et al.*, 2021, 2022), producing two haplotypes. Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). Contigs were further scaffolded with Hi-C data in YaHS (Zhou *et al.*, 2023), using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MerquryFK (Rhie *et al.*, 2020).

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

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included 36 breaks, 55 joins, and removal of 36 haplotypic duplications. This reduced the scaffold count by 5.4%, increased the scaffold N50 by 22.7%, and reduced the total assembly length by 1.1%. The curation process is described at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

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

The genome was analysed using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified from the NCBI Taxonomy (Schoch *et al.*, 2020). For the three domain-level lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all outputs into a blobdir for visualisation. The BlobToolKit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), with containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

PacBio sequencing of the *Phaonia angelicae* specimen generated 113.13 Gb (gigabases) from 11.54 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 1 228.20 Mb, with a heterozygosity of 1.88% and repeat content of 48.40% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 82× coverage. Hi-C sequencing produced 140.04 Gb from 463.72 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The genome was assembled into two haplotypes using Hi-C phasing. Haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The final assembly has a total length of 1 593.88 Mb in 926 scaffolds, with 637 gaps, and a scaffold N50 of 315.03 Mb (Table 2).

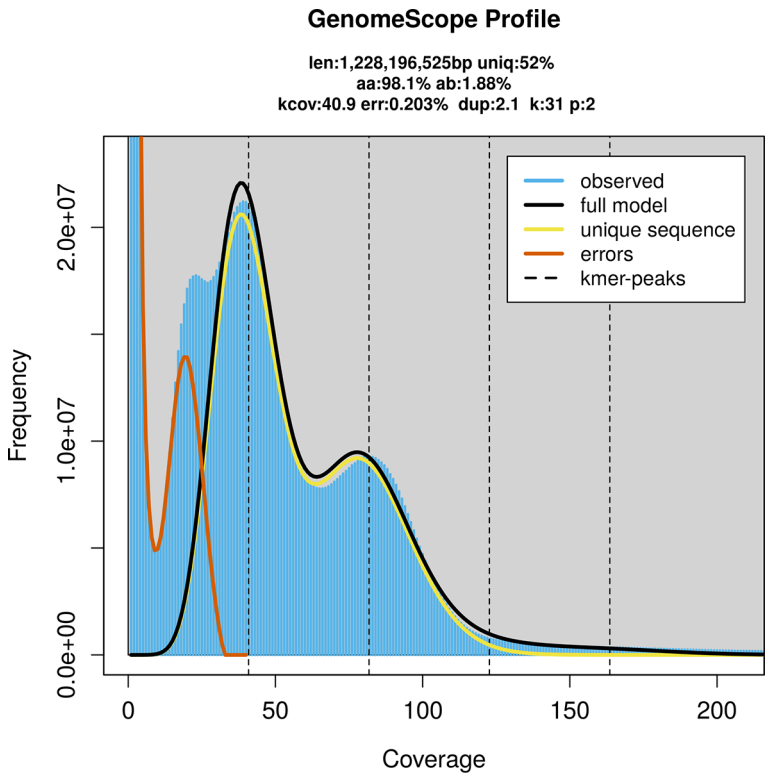

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB79289.

Platform	PacBio HiFi	Hi-C
ToLID	idPhaAnge1	idPhaAnge1
Specimen ID	Ox000739	Ox000739
BioSample (source individual)	SAMEA7746455	SAMEA7746455
BioSample (tissue)	SAMEA7746522	SAMEA7746520
Tissue	abdomen	head
Instrument	Sequel IIe	Illumina NovaSeq 6000
Run accessions	ERR13605525; ERR13605524; ERR13605526	ERR13602186
Read count total	11.54 million	463.72 million
Base count total	113.13 Gb	140.04 Gb

Table 2. **Genome assembly statistics.**

Genome assembly	Haplotype 1	Haplotype 2
Assembly name	idPhaAnge1.hap1.1	idPhaAnge1.hap2.1
Assembly accession	GCA_965111845.1	GCA_965111885.1
Assembly level	chromosome	scaffold
Span (Mb)	1 593.88	1 575.57
Number of chromosomes	5	scaffold-level
Number of contigs	1 563	1 853
Contig N50	3.92 Mb	4.35 Mb
Number of scaffolds	926	1 287
Scaffold N50	315.03 Mb	253.91 Mb
Longest scaffold length (Mb)	390.05	-
Organelles	Mitochondrion: 17.82 kb	-

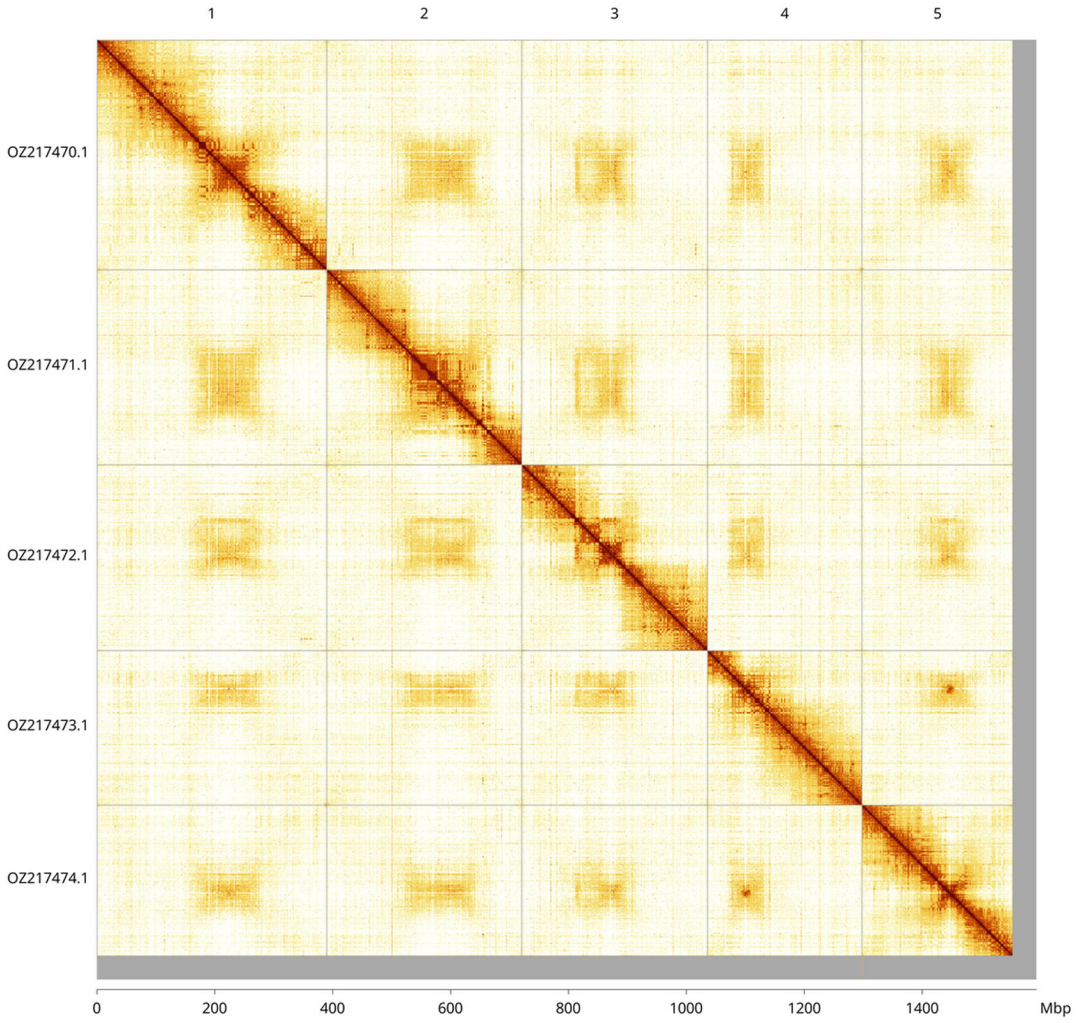

Figure 3. **Hi-C contact map of the *Phaonia angelicae* genome assembly.** Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Phaonia angelicae* idPhaAnge1.

INSDC accession	Molecule	Length (Mb)	GC%
OZ217470.1	1	390.05	35
OZ217471.1	2	331.11	35
OZ217472.1	3	315.03	35
OZ217473.1	4	262.10	35
OZ217474.1	5	255.49	35

Most of the haplotype 1 assembly sequence (97.48%) was assigned to 5 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). 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 (length 17.82 kb, OZ217475.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

For haplotype 1, the estimated QV is 62.9, and for haplotype 2, None. When the two haplotypes are combined, the assembly achieves an estimated QV of None. The *k*-mer completeness is 76.35% for haplotype 1, 75.94% for haplotype 2, and 98.43% for the combined haplotypes (Figure 4).

BUSCO analysis using the diptera_odb10 reference set (n = 3 285) identified 98.5% of the expected gene set (single = 97.1%, duplicated = 1.4%) in haplotype 1. For haplotype 2, BUSCO v.5.5.0 analysis identified 98.3% of the expected gene set (single = 96.7%, duplicated = 1.6%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

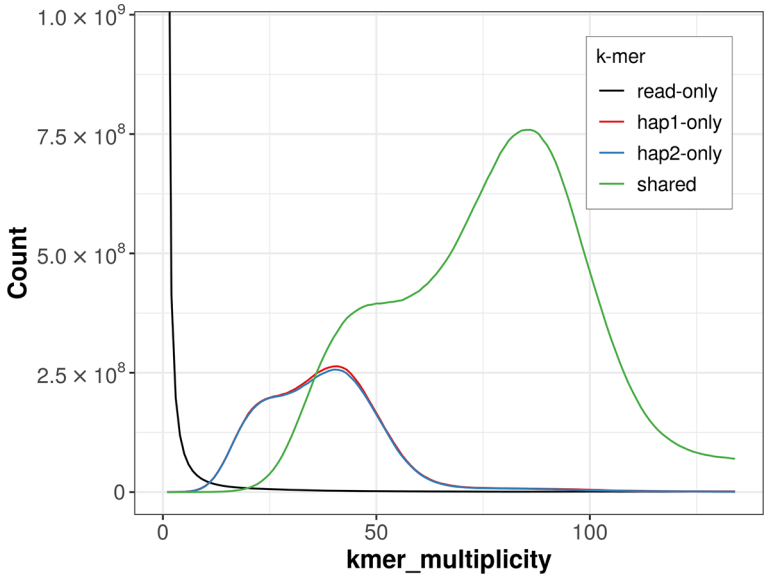

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve corresponds to *k*-mers shared by both haplotypes, and the red and blue curves show *k*-mers found only in one of the haplotypes.

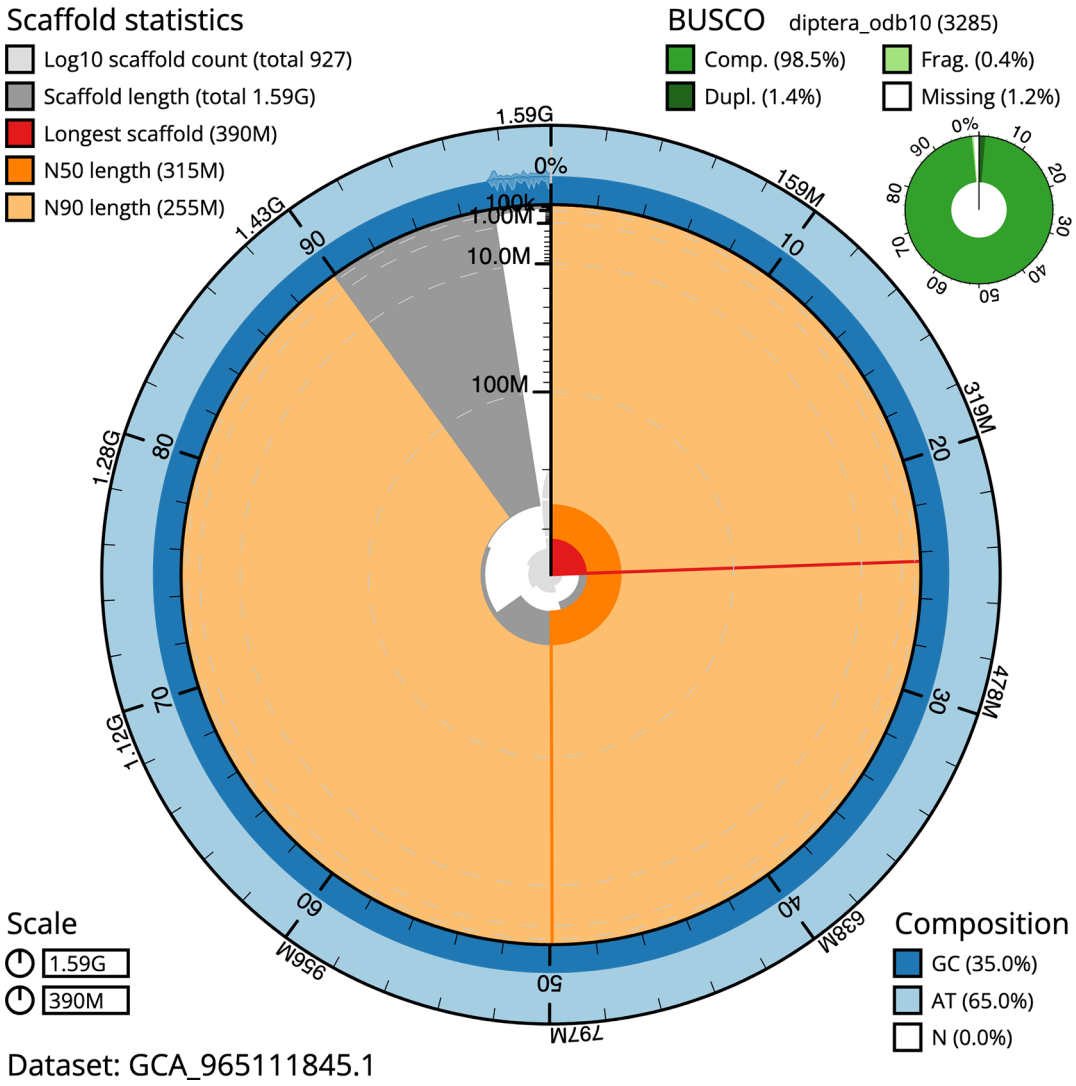

Figure 5. Assembly metrics for idPhaAnge1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards January 2026. The EBP metric, calculated for the haplotype 1, is **6.C.Q62**, meeting the recommended reference standard.

Genome annotation report

The *Phaonia angelicae* genome assembly (GCA_965111845.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 23 626 transcribed mRNAs from 13 923 protein-coding and 3 002 non-coding genes. The average transcript length is 19 263.38 bp, with an average of 1.40 coding transcripts per gene and 4.58 exons per transcript. Further details of this annotation are available from the Ensembl annotation page.

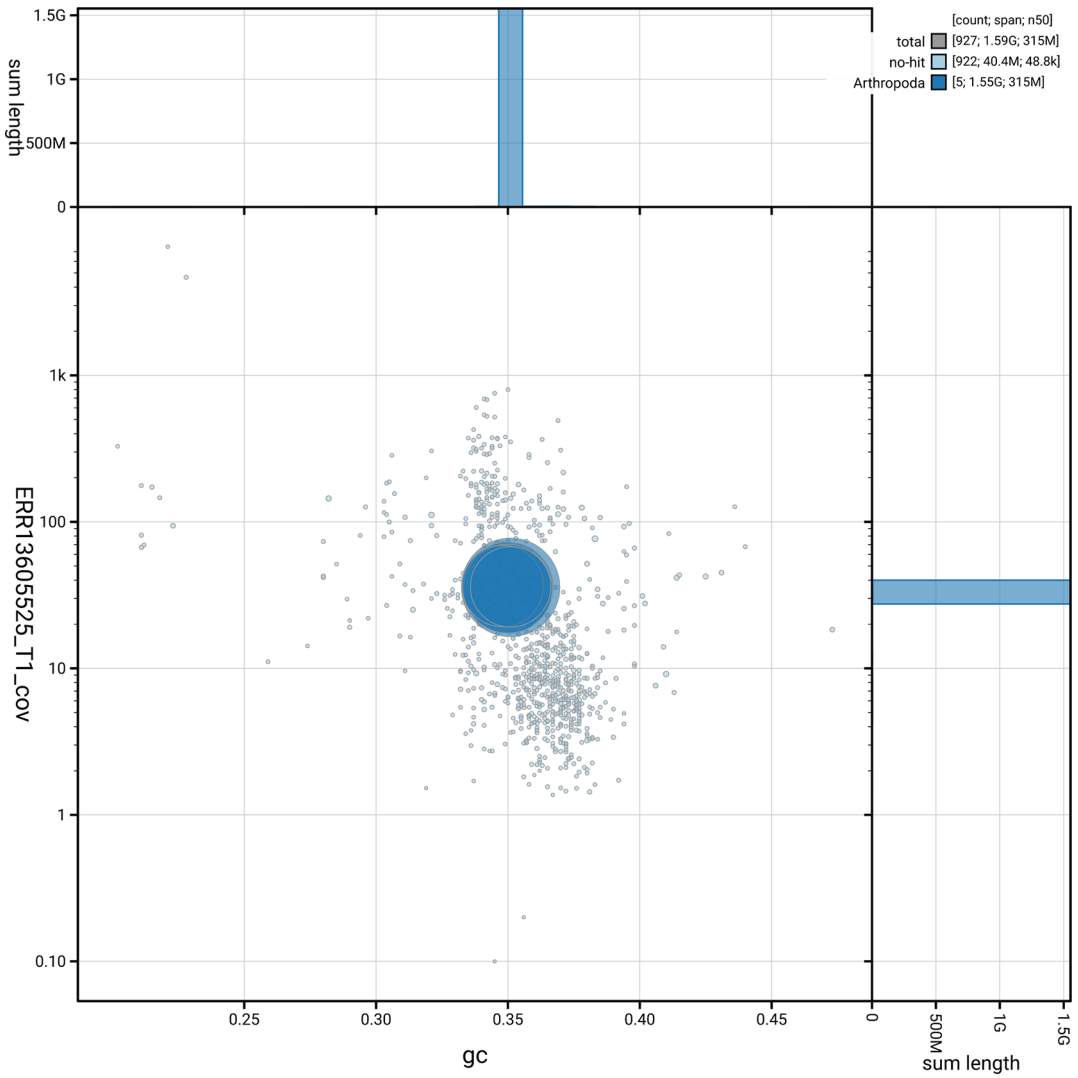

Figure 6. BlobToolKit blob plot for idPhaAnge1.hap1.1. The plot shows base coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the BlobToolKit viewer.

Table 4. Earth biogenome project summary metrics for the *Phaonia angelicae* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q62	6.C.Q40
Contig N50 length	3.92 Mb	≥ 1 Mb
Scaffold N50 length	315.03 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 62.9; haplotype 2:; combined:	≥ 40
k-mer completeness	Haplotype 1: 76.35%; Haplotype 2: 75.94%; combined: 98.43%	≥ 95%
BUSCO	C:98.5% [S:97.1%, D:1.4%], F:0.4%, M:1.2%, n:3 285	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	97.48%	≥ 90%

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

Author information

Contributors are listed at the following links:

- Members of the University of Oxford and Wytham Woods Genome Acquisition Lab
- Members of the Darwin Tree of Life Barcoding collective
- Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team
- Members of Wellcome Sanger Institute Scientific Operations – Sequencing Operations
- Members of the Wellcome Sanger Institute Tree of Life Core Informatics team
- Members of the Tree of Life Core Informatics collective
- Members of the Darwin Tree of Life Consortium

Wellcome sanger institute – legal and governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the **'Darwin Tree of Life Project Sampling Code of Practice'**, which can be found in full on the Darwin Tree of Life website. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project. Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible. The overarching areas of consideration are:

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

Each transfer of samples is further undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Darwin Tree of Life Partner, Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances, other Darwin Tree of Life collaborators.

Data availability

European Nucleotide Archive: Phaonia angelicae. Accession number PRJEB79289. The genome sequence is released openly for reuse. The *Phaonia angelicae* 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 Tables 1 and 2.

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

Table 5. Software versions and sources.

Software	Version	Source
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond

Table 5. *Continued*

Software	Version	Source
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.6.0	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

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

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

Buchfink B, Reuter K, Drost H-G: **Sensitive protein alignments at tree-of-life scale using DIAMOND.** *Nat. Methods.* 2021; **18**(4): 366–368.
PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3: Genes, Genomes, Genetics.* 2020; **10**(4): 1361–74.
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): 170–75.
PubMed Abstract | Publisher Full Text | Free Full Text

Cheng H, Jarvis ED, Fedrigo O, *et al.*: **Haplotype-resolved assembly of diploid genomes without parental data.** *Nat. Biotechnol.* 2022; **40**(9): 1332–35.
PubMed Abstract | Publisher Full Text | Free Full Text

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial Arthropod fauna.** *Wellcome Open Res.* 2023; **8**: 123.
PubMed Abstract | Publisher Full Text | Free Full Text

d'Assis Fonseca ECM: **Diptera Cyclorrhapha Calyptrata section (b) Muscidae.** *Handbooks for the Identification of British Insects.* 1968; **10**: 1–118.

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

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): 3047–48.
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): 4214–16.
PubMed Abstract | Publisher Full Text | Free Full Text

Gregor F, Rozkošný R, Barták M, *et al.*: **The Muscidae (Diptera) of Central Europe.** *Folia Facultatis Scientiarum Naturalium Universitatis Masarykianae Brunensis, Biologia.* 2002; **107**: 1–280.

Grzywacz A, Hall MJR, Pape T, *et al.*: **Muscidae (Diptera) of forensic importance—an identification key to third instar larvae of the western Palaearctic region and a catalogue of the muscid carrion community.** *Int. J. Legal Med.* 2017; **131**: 855–866.
PubMed Abstract | Publisher Full Text | Free Full Text

Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: Lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025.
Publisher Full Text

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1).
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

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

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic**

coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Pont AC: **Observations on anthophilous Muscidae and other Diptera (Insecta) in Abisko National Park, Sweden.** *J. Nat. Hist.* 1993; **27**: 631–643.
Publisher Full Text

Pont AC, Evenhuis NL: **Catalogue of the Muscidae of the world (Insecta: Diptera).** *Bishop Museum Bulletin in Entomology.* 2026; **18**: 1–1341.

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): 737–46.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI taxonomy: A comprehensive update on curation, resources and tools.** *Database.* 2020; **2020**: baaa062.
PubMed Abstract | Publisher Full Text | Free Full Text

Skidmore P: **The Biology of the Muscidae of the *World.*** 1985; **29**.

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life Project.** *Wellcome Open Res.* 2024; **9**: 339.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE; 2019; 314–24.
Publisher Full Text

Vicoso B, Bachtrog D: **Numerous transitions of sex chromosomes in Diptera.** *PLoS Biol.* 2015; **13**(4): 1–22.
Publisher Full Text

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

Zych M: **On flower visitors and true pollinators: The case of protandrous *Heracleum sphondylium* (L.) (Apiaceae).** *Plant Syst. Evol.* 2007; **263**: 159–179.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✓ ✓

Version 1

Reviewer Report 03 June 2026

https://doi.org/10.21956/wellcomeopenres.28998.r154693

✓ **Stuart J.E. Baird**
Academy of Sciences of the Czech Republic, Národní, Staré Město, Czech Republic

The genome report follows the excellent template of the Darwin Tree of Life Consortium, with few or no typos and grammatical errors. I can find only three issues to comment on further: 1/ "Manual corrections included 36 breaks, 55 joins, and removal of 36 haplotypic duplications." Most readers of the report would want to know more about the 36 duplications.

Do they correspond in some way to the 36 breaks? How are artifactual duplications distinguished from natural duplications? 2/ Figure 3, Hi-C contact map of the *Phaonia angelicae* genome assembly: this differs from the majority of such snapshots in an interesting way. Each chromosome shares similarity in a central/off-centre section, so the whole snapshot looks rather like a box of chocolates.

While it may be that life is like a box of chocolates, pretext snapshots less so, and I find it a shame there is zero commentary on this. 3/ "For haplotype 1, the estimated QV is 62.9, and for haplotype 2, None. When the two haplotypes are combined, the assembly achieves an estimated QV of None." A few simple words of explanation about why a log scaled measure requires an arbitrary word for the log(zero) case would be welcome.

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: Admixture 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 May 2026

https://doi.org/10.21956/wellcomeopenres.28998.r154307

Jonathan Badger
National Cancer Institute Center for Cancer Research (Ringgold ID: 272101), Bethesda, Maryland, USA

The authors describe their sequencing of the genome of Phaonia angelicae, a muscid fly found in Palaearctic regions in Europe and Asia. A single female individual was used as a source, although the authors were unable to recover sex chromosome(s), which is not unexpected in the Diptera. The metrics of the assembly and annotation look good and the data is available in the European Nucleotide Archive. The authors have provided all the required information needed for a genomic data note.

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: I work in computational genomics, primarily in a biomedical context but I have also worked in environmental genomics including 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.

Reviewer Report 06 May 2026

https://doi.org/10.21956/wellcomeopenres.28998.r154695

?

Rajib Deb
ICAR-National Research Centre, Guwahat, Assam, India

The manuscript states that the genome contributes to the Darwin Tree of Life initiative and will support phylogenomic and comparative studies.

- The rationale is somewhat generic and program-driven rather than hypothesis-driven.
- Limited discussion of why *P. angelicae* specifically is important (e.g., ecological, evolutionary, or applied relevance beyond being representative).
- No explicit identification of knowledge gaps this genome helps fill.
- Missing QV values for haplotype 2 and combined assembly (reported as “None”) reduce completeness of reporting
- Many steps rely on external protocols rather than being fully described.
- Some parameters and thresholds are not explicitly stated (e.g., filtering criteria, QC thresholds).
- Limited detail on:
  - Data preprocessing steps
  - Specific command-line parameters
  - Criteria for manual curation decisions
  - Navigation between datasets (raw vs processed) could be clarified.
  - No explicit statement of file formats or recommended usage workflows.

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

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: 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.