Check for updates

DATA NOTE

The genome sequence of a hoverfly , *Melangyna compositarum* (Verrall, 1873) (Diptera: Syrphidae)

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

Liam M. Crowley[1], Katie J. Woodcock[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]University of Oxford, Oxford, England, UK
[2]Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 18 Mar 2026, **11**:179
https://doi.org/10.12688/wellcomeopenres.26158.1
Latest published: 18 Mar 2026, **11**:179
https://doi.org/10.12688/wellcomeopenres.26158.1

Abstract

We present a genome assembly from an individual male *Melangyna compositarum* (hoverfly; Arthropoda; Insecta; Diptera; Syrphidae). The assembly contains two haplotypes with total lengths of 999.96 megabases and 858.82 megabases. Most of haplotype 1 (91.7%) is scaffolded into 5 chromosomal pseudomolecules, including the X sex chromosome. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 16.87 kilobases. This assembly was generated as part of the Darwin Tree of Life project, which produces reference genomes for eukaryotic species found in Britain and Ireland.

Keywords

Melangyna compositarum, hoverfly, genome sequence, chromosomal, Diptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ? ✓ ✓

	1	2	3
version 1 18 Mar 2026	? view	✓ view	✓ view

1. **Jaakko Pohjoismäki**, University of Eastern Finland, Joensuu, Finland
2. **Darren Obbard**, The University of Edinburgh, Edinburgh, UK
3. **Hans-Peter Fuehrer**, University of Veterinary Medicine, Veinna, Austria

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: Crowley LM: Investigation, Resources; **Woodcock KJ**: Writing – Original Draft Preparation;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540) and the Darwin Tree of Life Discretionary Award [218328, https://doi.org/10.35802/218328].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Crowley LM *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: Crowley LM, Woodcock KJ, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of a hoverfly , *Melangyna compositarum* (Verrall, 1873) (Diptera: Syrphidae) [version 1; peer review: 2 approved, 1 approved with reservations]** Wellcome Open Research 2026, **11**:179 https://doi.org/10.12688/wellcomeopenres.26158.1

First published: 18 Mar 2026, **11**:179 https://doi.org/10.12688/wellcomeopenres.26158.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Eremoneura; Cyclorrhapha; Aschiza; Syrphoidea; Syrphidae; Syrphinae; Syrphini; *Melangyna*; *Melangyna compositarum* (Verrall, 1873) (NCBI:txid1822530).

Background

Melangyna compositarum (Verrall, 1873) is a frequent hoverfly species found across the UK and Ireland (Stubbs & Falk, 2002). Reliably distinguishing between two of the *Melangyna* species; *M. compositarum* and *Melissodes labiatarum* is notoriously difficult due to inconsistent morphological features (Stubbs & Falk, 2002; van Veen, 2010). Some debate exists around whether they truly represent separate species (Ball & Morris, 2015; Stubbs & Falk, 2002). In adult males eye hairiness is used as a morphological marker with *compositarum* eyes displaying a sparse sprinkling of hairs or none at all. Contrastingly, *labiatarum* has a dense covering of pale eye hairs (Stubbs & Falk, 2002; van Veen, 2010). Adult female *compositarum* hoverflies have a wider, whitish-grey face, while in *labiatarum* females the face is slightly narrower and yellowish-grey (Stubbs & Falk, 2002). Additionally, in *compositarum* females, the frons is undusted with a broader, shiny black anterior region, whereas in *labiatarum* it is entirely dusted (Stubbs & Falk, 2002).

Notably, the geographic distribution between the two species appears distinct with *compositarum* being found principally in Northern England and Scotland (Stubbs & Falk, 2002). *M. compositarum* hoverflies display a preference for coniferous woodland habitats and can be found from May onwards peaking between mid-June to mid-August (Ball & Morris, 2000; Ball & Morris, 2015; Stubbs & Falk, 2002). Males have been observed forming small swarms (Ball & Morris, 2015). Adults are frequent flower visitors with an inclination for umbellifers including *Heracleum* and *Angelica* (Ball & Morris, 2000; Ball & Morris, 2015). Knowledge of *M. compositarum* larval development is limited. Generally *Melangyna* genus larvae are known to be oligophagous aphid predators, often feeding on a specific species of aphid or aphids found on specific trees (Ball & Morris, 2000; van Veen, 2010). The completed genome sequence for *Melangyna compositarum* provides a valuable tool to further the knowledge of this relatively obscure hoverfly species.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Melangyna compositarum* (specimen ID Ox000703, ToLID idMelComo1; Figure 1), collected from Wytham Woods, Oxfordshire, UK (latitude 51.77, longitude −1.339) on 2020-07-24. The specimen was collected and identified by Liam Crowley (University of Oxford).

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

Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The idMelComo1 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. HMW DNA was extracted using the Automated MagAttract v2 protocol. We used centrifuge-mediated fragmentation to produce DNA fragments in the 8–10 kb range, following the Covaris g-TUBE protocol for ultra-low input (ULI). Sheared DNA was purified by automated SPRI (solid-phase reversible immobilisation). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 2.02 ng/μL and a yield of 787.80 ng.

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

Figure 1. Photograph of the *Melangyna compositarum* (idMelComo1) specimen used for genome sequencing.

overhang removal, DNA damage repair, and end-repair/A-tailing were performed according to the manufacturer's instructions, followed by adapter ligation. A 0.85× pre-PCR clean-up was carried out with Promega ProNex beads.

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

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

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

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen head and thorax tissue of the idMelComo1 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.

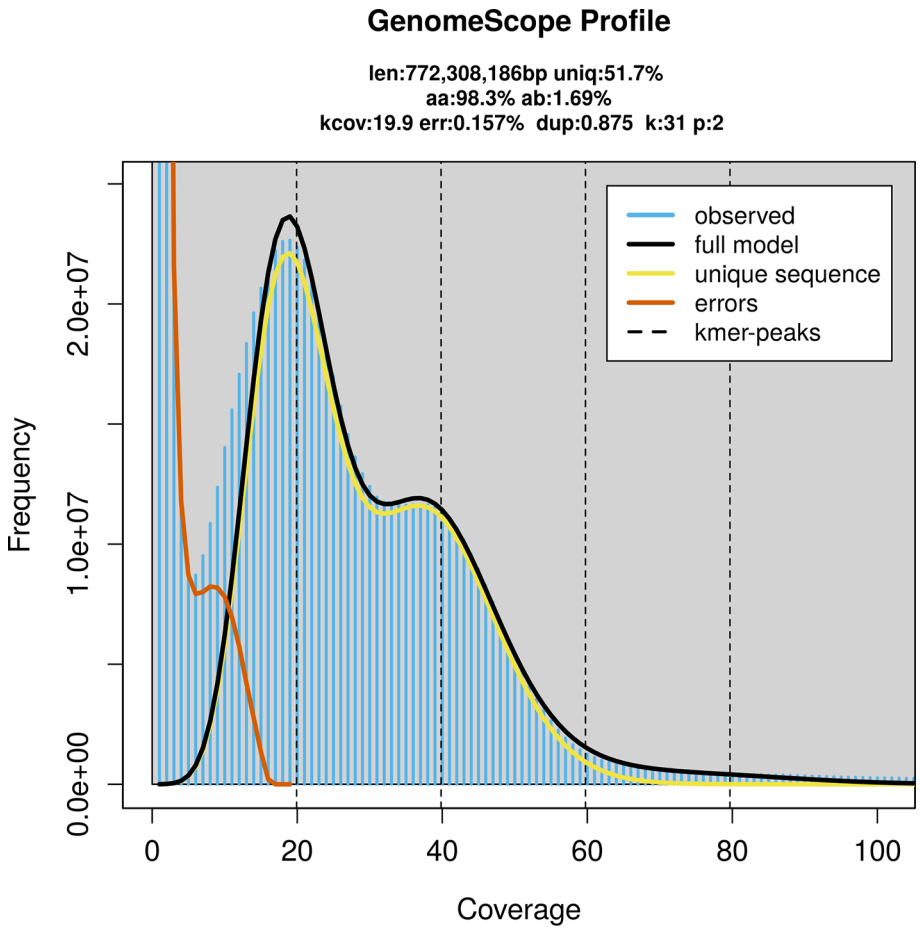

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. **Specimen and sequencing data for BioProject PRJEB83545.**

Platform	PacBio HiFi	Hi-C
ToLID	idMelComo1	idMelComo1
Specimen ID	Ox000703	Ox000703
BioSample (source individual)	SAMEA7701564	SAMEA7701564
BioSample (tissue)	SAMEA7701762	SAMEA7701761
Tissue	abdomen	head and thorax
Instrument	Revio	Illumina NovaSeq 6000
Run accessions	ERR14105730; ERR14104859	ERR14075574
Read count total	3.70 million	716.46 million
Base count total	33.63 Gb	108.19 Gb

Table 2. **Genome assembly statistics.**

Assembly name	idMelComo1.hap1.1	idMelComo1.hap2.1
Assembly accession	GCA_965663475.1	GCA_965663505.1
Assembly level	chromosome	scaffold
Span (Mb)	999.96	858.82
Number of chromosomes	5	scaffold-level
Number of contigs	5 195	4 759
Contig N50	0.37 Mb	0.36 Mb
Number of scaffolds	1 699	1 638
Scaffold N50	253.18 Mb	224.68 Mb
Longest scaffold length (Mb)	327.33	-
Sex chromosomes	X	-
Organelles	Mitochondrion: 16.87 kb	-

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 146 breaks and 238 joins. This reduced the scaffold count by 5.0%, increased the scaffold N50 by 1.1%, and reduced the total assembly length by 1.3%. 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 databases ($k = 31$) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

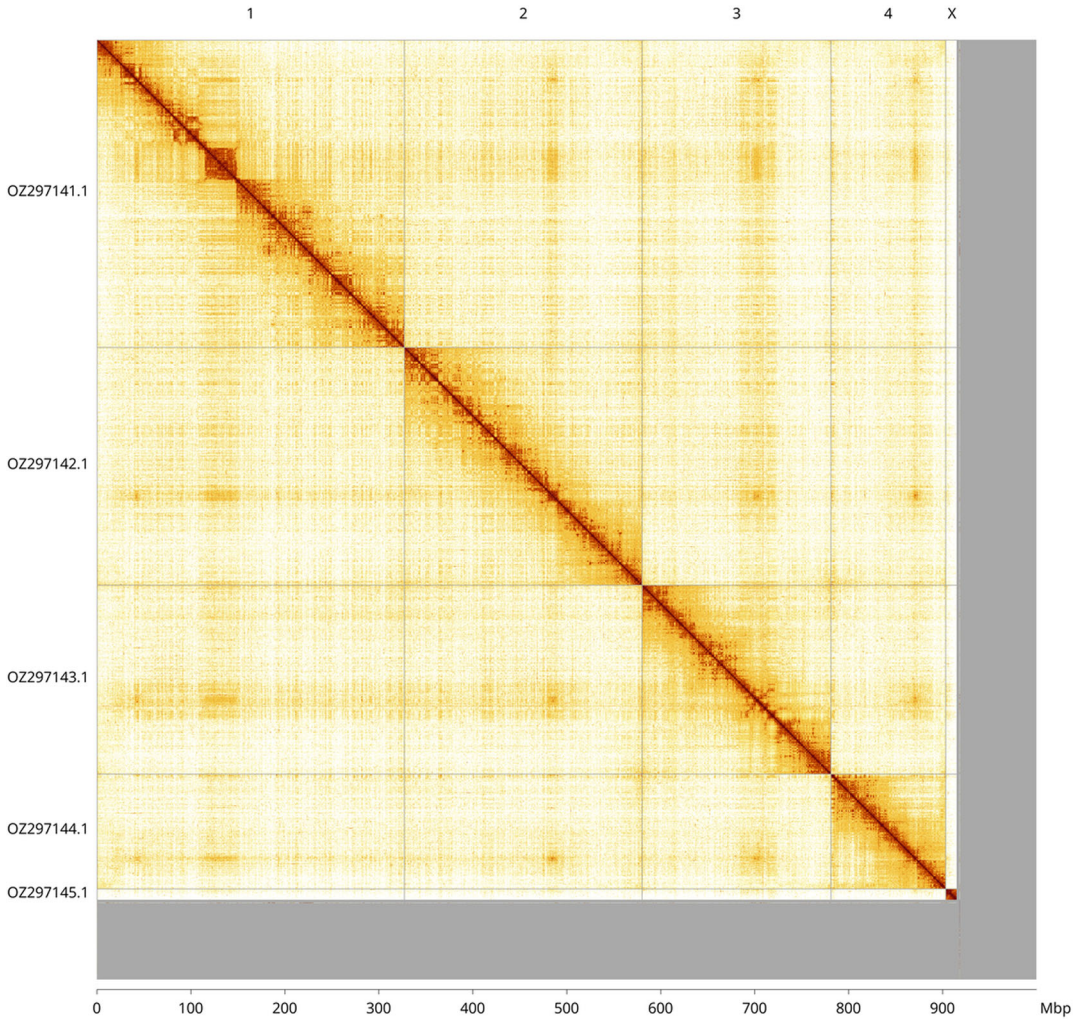

Figure 3. Hi-C contact map of the *Melangyna compositarum* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Melangyna compositarum* idMelComo1.

INSDC accession	Molecule	Length (Mb)	GC%
OZ297141.1	1	327.33	33
OZ297142.1	2	253.18	32.50
OZ297143.1	3	200.80	32.50
OZ297144.1	4	122.04	32.50
OZ297145.1	X	13.58	33.50

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

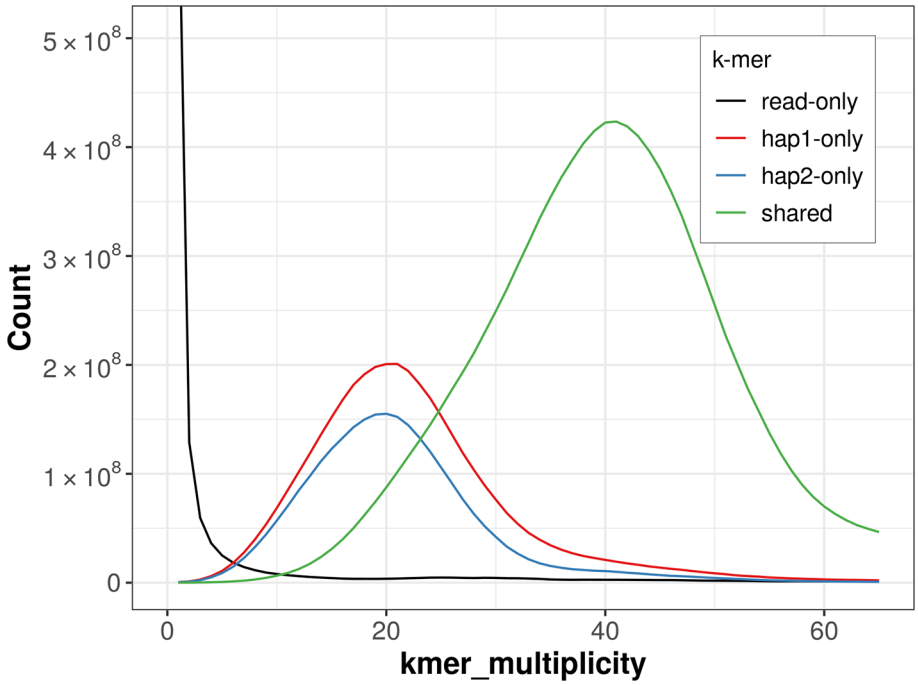

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

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 *Melangyna compositarum* specimen generated 33.63 Gb (gigabases) from 3.70 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 775.63 Mb, with a heterozygosity of 1.69% and repeat content of 48.33% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 40× coverage. Hi-C sequencing produced 108.19 Gb from 716.46 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 999.96 Mb in 1 699 scaffolds, with 3 496 gaps, and a scaffold N50 of 253.18 Mb (Table 2).

Most of the haplotype 1 assembly sequence (91.7%) was assigned to 5 chromosomal-level scaffolds, representing 4 autosomes and the X sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Chromosome X was identified by copy number in the diploid assembly. No Y chromosome could be confidently identified, but it may be represented in the unassigned scaffolds in the assembly.

The mitochondrial genome was also assembled (length 16.87 kb, OZ297146.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

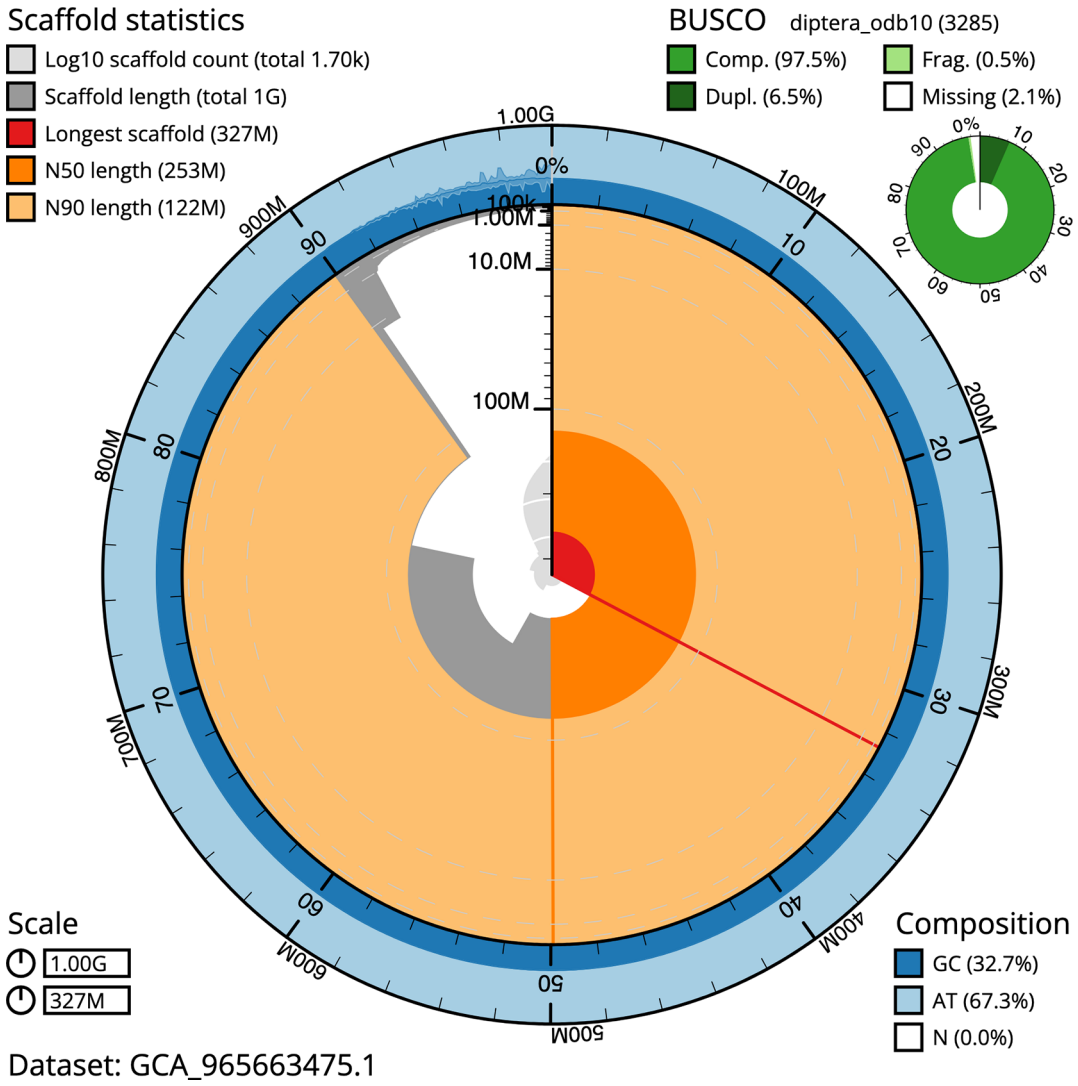

Figure 5. Assembly metrics for idMelComo1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost 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.

Assembly quality metrics

For haplotype 1, the estimated QV is 57.3, and for haplotype 2, 57.1. When the two haplotypes are combined, the assembly achieves an estimated QV of 57.2. The *k*-mer completeness is 77.26% for haplotype 1, 70.16% for haplotype 2, and 98.19% for the combined haplotypes (Figure 4).

BUSCO analysis using the endopterygota_odb10 reference set (n = 2 124) identified 98.6% of the expected gene set (single = 90.6%, duplicated = 8.0%) in haplotype 1. For haplotype 2, BUSCO v.6.0.0 analysis identified 94.7% of the expected gene set (single = 90.1%, duplicated = 4.6%). 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.

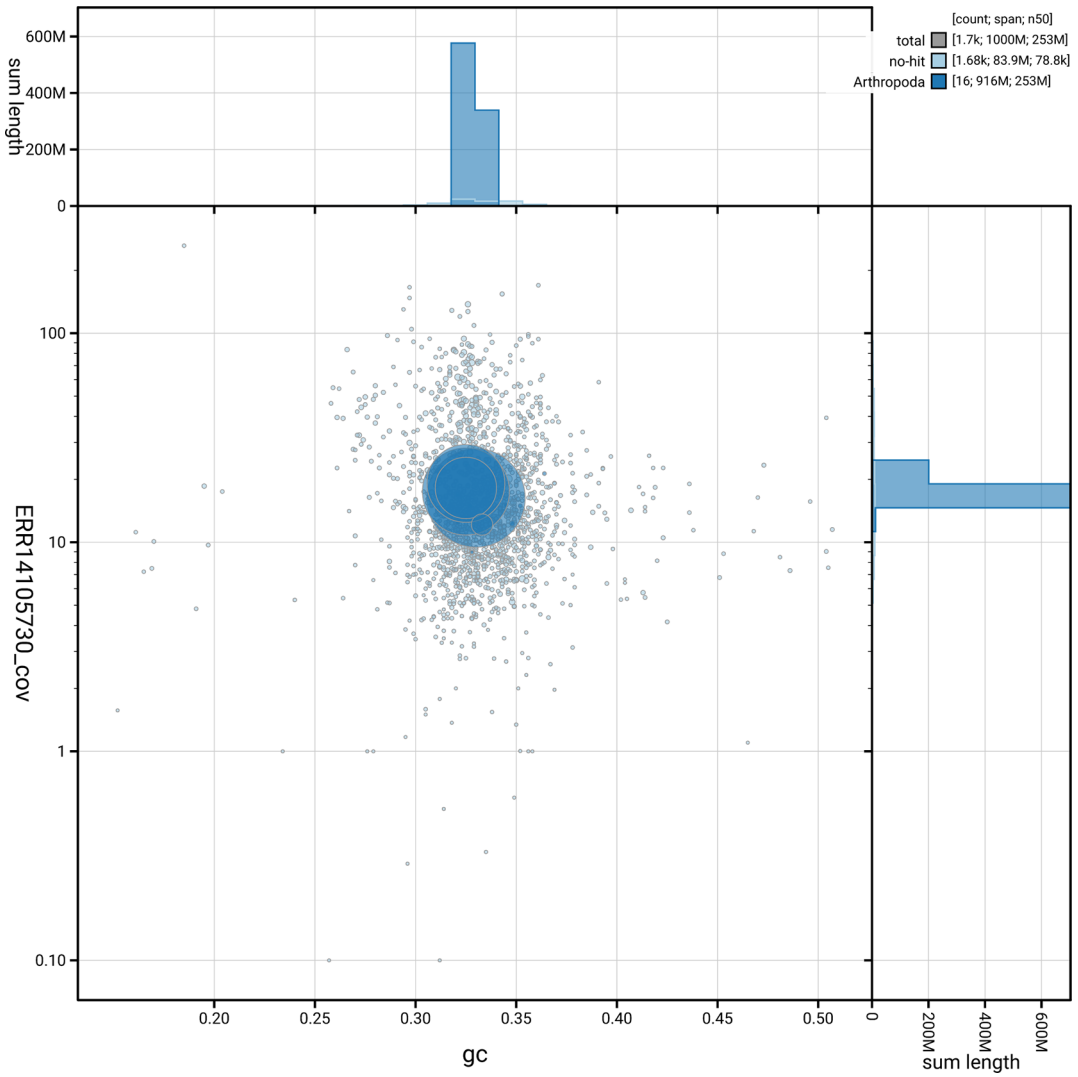

Figure 6. BlobToolKit blob plot for idMelComo1.hap1.1. The plot shows base coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the 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 *Melangyna compositarum* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	5.C.Q57	6.C.Q40
Contig N50 length	0.37 Mb	≥ 1 Mb
Scaffold N50 length	253.18 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 57.3; haplotype 2: 57.1; combined: 57.2	≥ 40
k-mer completeness	Haplotype 1: 77.26%; Haplotype 2: 70.16%; combined: 98.19%	≥ 95%
BUSCO	C:98.6% [S:90.6%, D:8.0%], F:0.5%, M:0.9%, n:2 124	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	91.70%	≥ 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.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **5.C.Q57**.

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: Melangyna compositarum (matt-backed melangyna). Accession number PRJEB83545. The genome sequence is released openly for reuse. The *Melangyna compositarum* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665) and the Sanger Institute Tree of Life Programme (PRJEB43745). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in 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	4.4.6	https://github.com/blobtoolkit/blobtoolkit

Table 5. *Continued*

Software	Version	Source
BUSCO	6.0.0	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	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	2.28-r1209	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	24.10.4	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView
samtools	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.9.0	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	1.2.2	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

Ball S, Morris R: *Britain's Hoverflies: A Field Guide - Revised and Updated Second Edition.* Princeton University Press; 2015.

Ball SG, Morris RKA: *A Provisional Atlas of British Hoverflies (Diptera, Syrphidae).* Biological Records Centre, Centre for Ecology and Hydrology; 2000.

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.
Publisher Full Text

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

Cheng H, Jarvis ED, Fedrigo O, *et al.*: **Haplotype-resolved assembly of diploid genomes without parental data.** *Nat. Biotechnol.* 2022; **40**(9): 1332–1335.
Publisher 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.
Publisher Full Text

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): 3047–3048.
Publisher 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.
Publisher Full Text

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

bioRxiv. 2025.
Publisher Full Text

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

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

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

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

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

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

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

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1).
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**: baa062.
Publisher Full Text

Stubbs AE, Falk SJ: *British Hoverflies: An Illustrated Identification Guide.* British Entomological and Natural History Society; 2002.

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.
Publisher 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.
Publisher Full Text

van Veen MP: *Hoverflies of Northwest Europe: Identification Keys to the Syrphidae.* BRILL; 2010.

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; 314–324.
Publisher Full Text

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

Open Peer Review

Current Peer Review Status: ? ✓ ✓

Version 1

Reviewer Report 10 April 2026

https://doi.org/10.21956/wellcomeopenres.28807.r151905

© **2026 Fuehrer H.** 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.

✓ **Hans-Peter Fuehrer**
University of Veterinary Medicine, Veinna, Austria

The authors analyzed the genome of a hoverfly, Melangyna compositarum (Verrall, 1873) (Diptera, Syrphidae). The biology of the species and the difficulties involved in morphological identification are discussed. The manuscript is of good quality and only minor modifications are recommended.

The genus is known as “Spot-tails.”
How was the specimen collected?

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: Medical entomology

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 08 April 2026

https://doi.org/10.21956/wellcomeopenres.28807.r152479

© **2026 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 Melangyna compositarum as part of the “Darwin Tree of Life” programme. In common with other data notes from this research effort, the reporting is standardised 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 syrphids and other diptera, or on the comparative or evolutionary genomics of insects more generally.

My minor suggestions are:

(1) There is an extreme typo / copy-paste error near the start, when Melangyna labiatarum is written as Melissodes labiatarum (that would be a bee, I think?)
(2) In addition to the sequenced specimen itself, it would be nice to have an ‘in life’ photograph and/or perhaps both sexes. Perhaps even a good photo of the (subtle!) differences between M. compositarum and M. labiatarum.
(3) I think the reporting of the taxonomic status is a little too brief. It is interesting to know that “Some debate exists around whether they truly represent separate species (Ball & Morris, 2015; Stubbs & Falk, 2002).” But those two references are of little value to most readers, who (not being based in the UK) are unlikely to have access to UK field guides. What is the gist of those argument? What genetic data have been used to support either side, or is it purely morphological? Does the debate extend to taxonomists outside of the UK? I think this warrants at least another 3 sentences, especially given the potential value of a genome in resolving the argument

(4) I am surprised that no mention is made of the potential value of this genome in resolving the taxonomic debate. Obviously, a genome might tell us nothing – but equally, it could resolve the question instantly! It is a shame that DToL do not have labiatarum …

(5) has anyone proposed common names for this species? If so, what?

(6) Style: Please avoid starting a sentence with “M.”

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 Diptera 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.

Reviewer Report 08 April 2026

https://doi.org/10.21956/wellcomeopenres.28807.r151910

© 2026 Pohjoismäki J. 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.

?

Jaakko Pohjoismäki
University of Eastern Finland, Joensuu, Finland

The presented work by Liam M. Crowley, Katie J. Woodcok and the DToL consortium presents the assembly and basic characterization of the hoverfly *Melangyna compositarum* genome, following established high-throughput sequencing and assembly pipelines of the Darwin Tree of Life Project. Overall, the manuscript is clear, technically sound, and meets the expected standards.

The genome assembly is of high quality and continuity. The combination of PacBio long-read sequencing and Hi-C scaffolding has produced a highly contiguous assembly with a chromosome-scale scaffold N50 (253 Mb) and a high proportion (91.7%) of sequence assigned to chromosomes. Consensus quality values (QV ~57) and BUSCO completeness (98.6%) indicate a near-complete and accurate representation of the gene space. The sequencing depth (40x) is more than sufficient given the genome size and complexity. Also, the haplotypes have been resolved.

I have only a couple of minor points:
The duplicated BUSCO proportion (8.0%) is a bit above the recommended benchmark (<5%), which may reflect residual haplotypic duplication or assembly artifacts; a brief comment on this would improve transparency. Similarly, the relatively low k-mer completeness for individual haplotypes (77% and 70%) compared to the combined assembly suggests some fragmentation. The absence of an identifiable Y chromosome is noted. Sex chromosomes appear to be tricky but have been identified for other brachyceran flies by read coverage, heterochromatic region size and BUSCO (diptera_odb12) gene count. These points could be briefly discussed, but this is not absolutely essential as future comparative analyses might want to address such issues more universally.

However, there is a bit awkward error in the Background section that should be corrected: *Melissodes* is a Nearctic bee genus. *Melangyna labiatarum* does exists and fits the context here, so easy to fix.

Otherwise, this is a solid and valuable genome resource that will be useful for future studies on hoverfly genetics and comparative genomics.

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: cell & molecular biology, biodiversity genomics, taxonomy

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.