



DATA NOTE

The genome sequence of the Square-spot Deerfly, *Chrysops viduatus* (Fabricius, 1794) (Diptera: Tabanidae)

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

Erica McAlister¹, Ryan Mitchell², Olga Sivell¹ ¹, Judith A. Webb³,
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, County Sligo, Ireland

³Independent researcher, Kidlington, England, UK

V1 First published: 21 Nov 2025, 10:650
<https://doi.org/10.12688/wellcomeopenres.25193.1>
Latest published: 21 Nov 2025, 10:650
<https://doi.org/10.12688/wellcomeopenres.25193.1>

Abstract

We present a genome assembly from an individual female *Chrysops viduatus* (Square-spot Deerfly; Arthropoda; Insecta; Diptera; Tabanidae). The genome sequence has a total length of 312.35 megabases. Most of the assembly (87.6%) is scaffolded into 5 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 16.08 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

Chrysops viduatus; Square-spot Deerfly; genome sequence; chromosomal; Diptera



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

Open Peer Review

Approval Status ? ✓ ✓

	1	2	3
version 1	?	✓	✓
21 Nov 2025	view	view	view

1. **Stuart J.E. Baird** , Academy of Sciences of the Czech Republic, Národní, Czech Republic
2. **Chao Du**, Baotou Teachers College, Sichuan Academy of Agricultural Sciences, Baotou, China
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: **McAlister E:** Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Mitchell R:** Investigation, Resources; **Sivell O:** Investigation, Resources; **Webb JA:** 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 McAlister E *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: McAlister E, Mitchell R, Sivell O *et al.* **The genome sequence of the Square-spot Deerfly, *Chrysops viduatus* (Fabricius, 1794) (Diptera: Tabanidae) [version 1; peer review: 2 approved, 1 approved with reservations]** Wellcome Open Research 2025, 10:650 <https://doi.org/10.12688/wellcomeopenres.25193.1>

First published: 21 Nov 2025, 10:650 <https://doi.org/10.12688/wellcomeopenres.25193.1>

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Tabanomorpha; Tabanoidea; Tabanidae; Chrysopsinae; Chrysopsini; *Chrysops*; *Chrysops viduatus* (Fabricius, 1794) (NCBI:txid798133)

Background

Chrysops viduatus (Fabricius, 1794) is one of the medium-sized (7–10 mm) banded-winged species of horseflies (Tabanidae) and is commonly referred to as the Square-spot Deerfly. Both males and females have colourful eye patterns, with the eyes meeting at the top in the males, and the sexes have further differences in their abdominal patterns. Both the wing and the abdominal markings are used for identifications (Stubbs & Drake, 2014).

This species is widely distributed across the Palaearctic region, but is concentrated in wetter habitats such as wet woodlands, mires and wet meadows. In the UK, most of the records are from south of the Severn-Wash line, with the Dorset Heath and the New Forest being particularly good locations (Soldierflies and Allies Recording Scheme, 2025).

The small, oblong eggs are attached to the vegetation overhanging water bodies, and larvae develop in mud or sandy substrate at the edge of these. Unusually for horseflies, these larvae, whose development may take up to a year, are not carnivorous but have a diet predominately of decaying vegetation (Burger, 1977). The pupal stage of the genus is shorter than the other genera, again below ground. Adults can be seen from June to August (with a few outliers), with the females taking a blood meal from large mammals, humans included, for egg development. Males, and when not blood feeding, females feed on nectar and pollen, and some species are important pollinators although more research is needed (Johnson, 2025).

Only two genomes are available for the family Tabanidae, both of them for the genus *Chrysops* as of August 2025 (data obtained via NCBI datasets, O'Leary *et al.*, 2024). We present a chromosome-level genome sequence for *Chrysops viduatus*, the Square-spot Deerfly. The assembly was produced using the Tree of Life pipeline from a specimen collected in Buxton Heath, England, UK (Figure 1). This assembly was generated as part of the Darwin Tree of Life Project, which aims to generate high-quality reference genomes for all named eukaryotic species in Britain and Ireland to support research, conservation, and the sustainable use of biodiversity (Darwin Tree of Life Project Consortium, 2022).

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult female *Chrysops viduatus* (specimen ID NHMUK015059108, ToLID idChrVidu2; Figure 1), collected from Buxton Heath, England, United Kingdom (latitude 52.75, longitude 1.22) on 2022-07-04. The specimen was collected by Erica McAlister



Figure 1. Photograph of the *Chrysops viduatus* (idChrVidu2) specimen used for genome sequencing.

and identified by Ryan Mitchell. A second specimen was used for Hi-C sequencing (specimen ID NHMUK014036858, ToLID idChrVidu1). It was collected from Parsonage Moor, Abington, England, United Kingdom (latitude 51.6942, longitude –1.3344) on 2021-06-19. This specimen was collected by Ryan Mitchell and Olga Sivell and identified by Judith Webb. For the Darwin Tree of Life sampling and metadata approach, refer to Lawniczak *et al.* (2022).

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 idChrVidu2 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the whole organism was homogenised by powermashing using a PowerMasher II tissue disruptor.

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol. DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor®3 for LI PacBio protocol. Sheared DNA was purified by manual 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 34.03 ng/μL and a yield of 1 599.41 ng, with a fragment size of 14.1 kb.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), following the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

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

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the idChrVidul 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 and equimolar and/or weighted 2.8 nM pools were created. 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 ([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](#)) with the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats ([Formenti et al., 2022](#)), BUSCO ([Manni et al., 2021](#)) and MERQUY.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. [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 11 breaks, 28 joins, and removal of two haplotypic duplications. The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation>. PretextViewSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The Merquy.FK tool ([Rhie et al., 2020](#)) was run in a Singularity container ([Kurtzer et al., 2017](#)) 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 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 *Chrysops viduatus* specimen generated 39.60 Gb (gigabases) from 3.93 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 221.10 Mb, with a heterozygosity of 0.81% and repeat content of 27.47% ([Figure 2](#)). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 170× coverage. Hi-C sequencing produced 143.50 Gb from 950.34 million reads, which were used to scaffold the assembly. [Table 1](#) summarises the specimen and sequencing details.

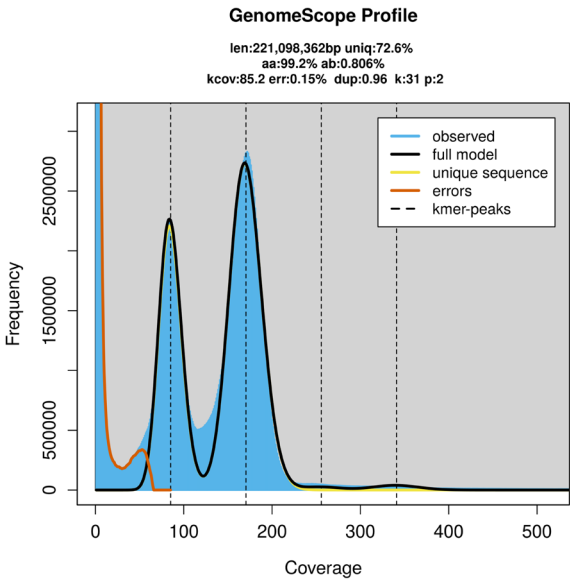


Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB67427.

Platform	PacBio HiFi	Hi-C
ToLID	idChrVidu2	idChrVidu1
Specimen ID	NHMUK015059108	NHMUK014036858
BioSample (source individual)	SAMEA112964067	SAMEA11025051
BioSample (tissue)	SAMEA112975187	SAMEA11025294
Tissue	whole organism	other somatic animal tissue
Instrument	Revio	Illumina NovaSeq 6000
Run accessions	ERR12120050	ERR12121878
Read count total	3.93 million	950.34 million
Base count total	39.60 Gb	143.50 Gb

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The final assembly has a total length of 312.35 Mb in 150 scaffolds, with 76 gaps, and a scaffold N50 of 55.79 Mb (Table 2).

Most of the assembly sequence (87.6%) 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). All chromosomes are named in size order. 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). During curation we observed that contigs in the following regions have uncertain order and orientation: on chromosome 2 between 40.19 Mb and 41.9 Mb and on chromosome 4 from 29.9 Mb to the end.

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.

The combined primary and alternate assemblies achieve an estimated QV of 62.2. The *k*-mer completeness is 86.01% for

Table 2. Genome assembly statistics.

Assembly name	idChrVidu2.1
Assembly accession	GCA_964274935.1
Alternate haplotype accession	GCA_964274965.1
Assembly level	chromosome
Span (Mb)	312.35
Number of chromosomes	5
Number of contigs	226
Contig N50	6.3 Mb
Number of scaffolds	150
Scaffold N50	55.79 Mb
Sex chromosomes	not identified
Organelles	Mitochondrion: 16.08 kb

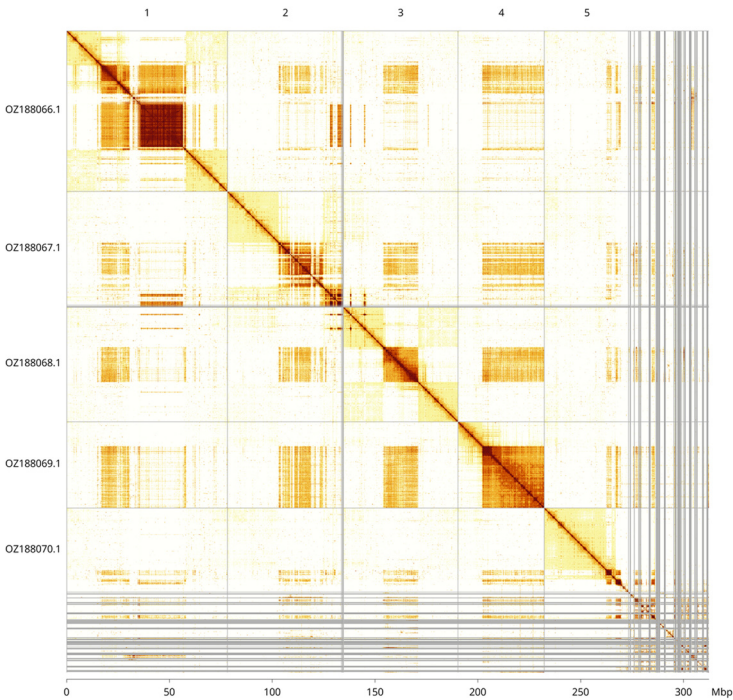


Figure 3. Hi-C contact map of the *Chrysops viduatus* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

the primary assembly, 83.22% for the alternate haplotype, and 99.54% for the combined assemblies (Figure 4).

BUSCO v.5.5.0 analysis using the diptera_odb10 reference set ($n = 3\,285$) identified 93.2% of the expected gene set (single = 92.8%, duplicated = 0.5%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for the primary assembly. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage.

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 primary assembly, is **6.7.Q62**, meeting the recommended reference standard.

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.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Chrysops viduatus* idChrVidu2.

INSDC accession	Molecule	Length (Mb)	GC%
OZ188066.1	1	78.30	32
OZ188067.1	2	56.35	32
OZ188068.1	3	55.79	34.50
OZ188069.1	4	42.17	36
OZ188070.1	5	41	31

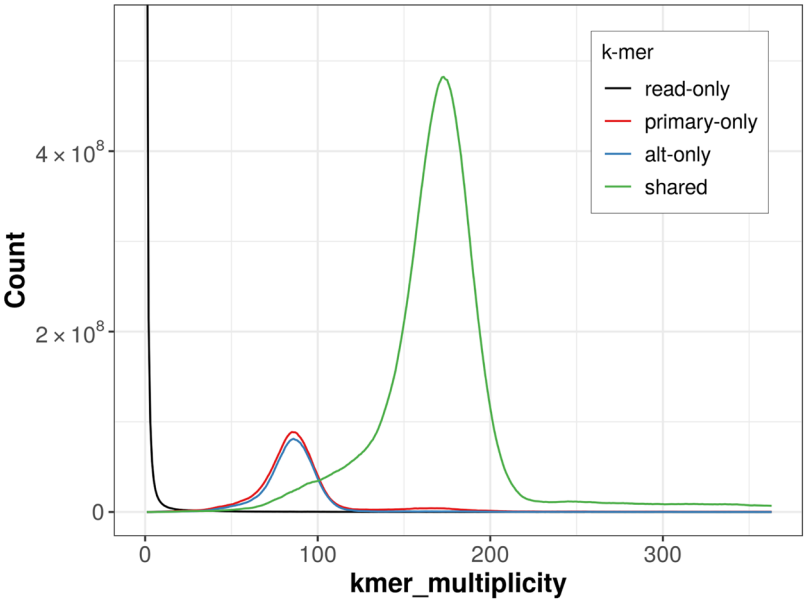


Figure 4. Evaluation of k -mer completeness using MerquyFK. This plot illustrates the recovery of k -mers from the original read data in the final assemblies. The horizontal axis represents k -mer multiplicity, and the vertical axis shows the number of k -mers. The black curve represents k -mers that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

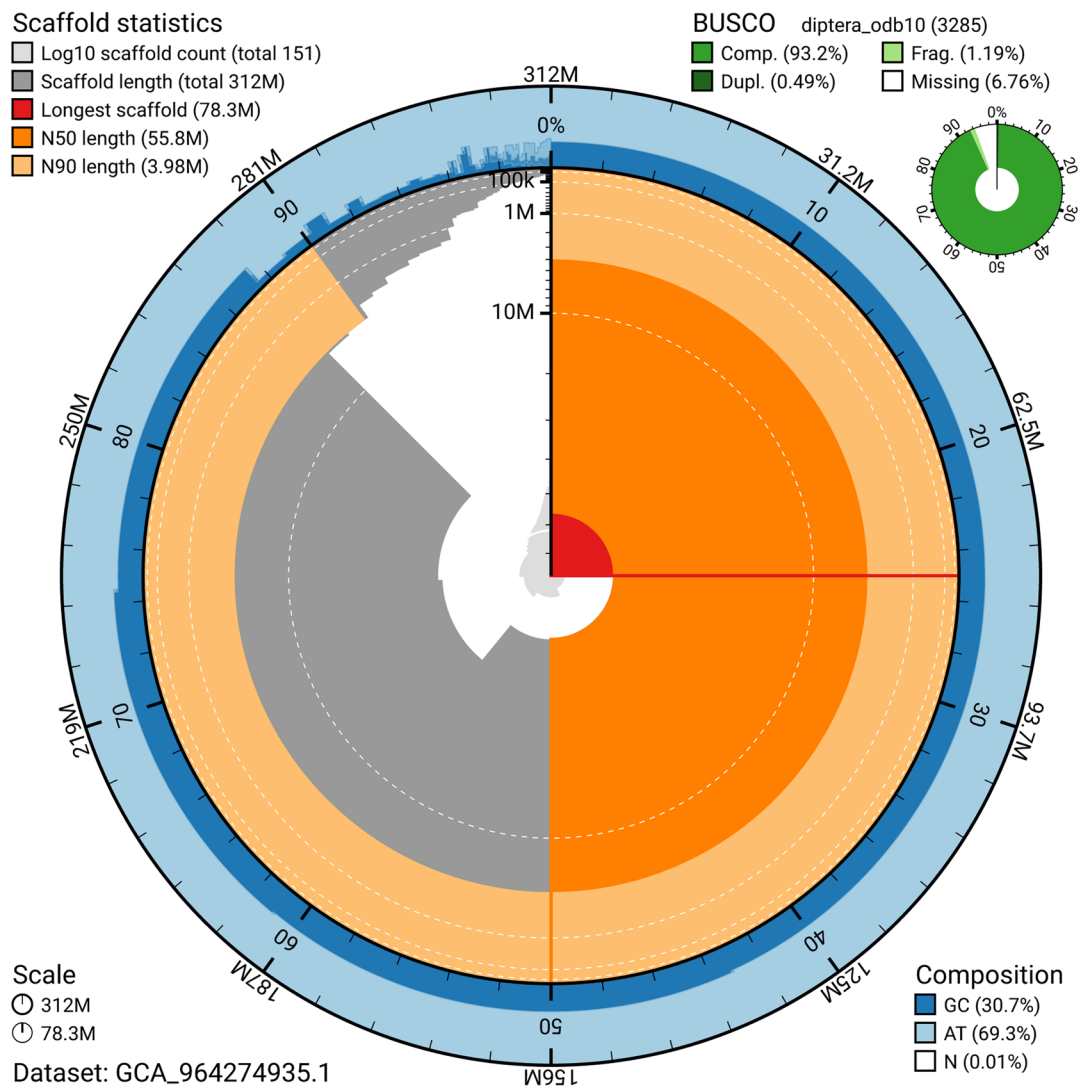


Figure 5. Assembly metrics for idChrVidu2.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue 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 diptera_odb10 set is presented at the top right. An interactive version of this figure can be accessed on the [BlobToolKit viewer](#).

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)

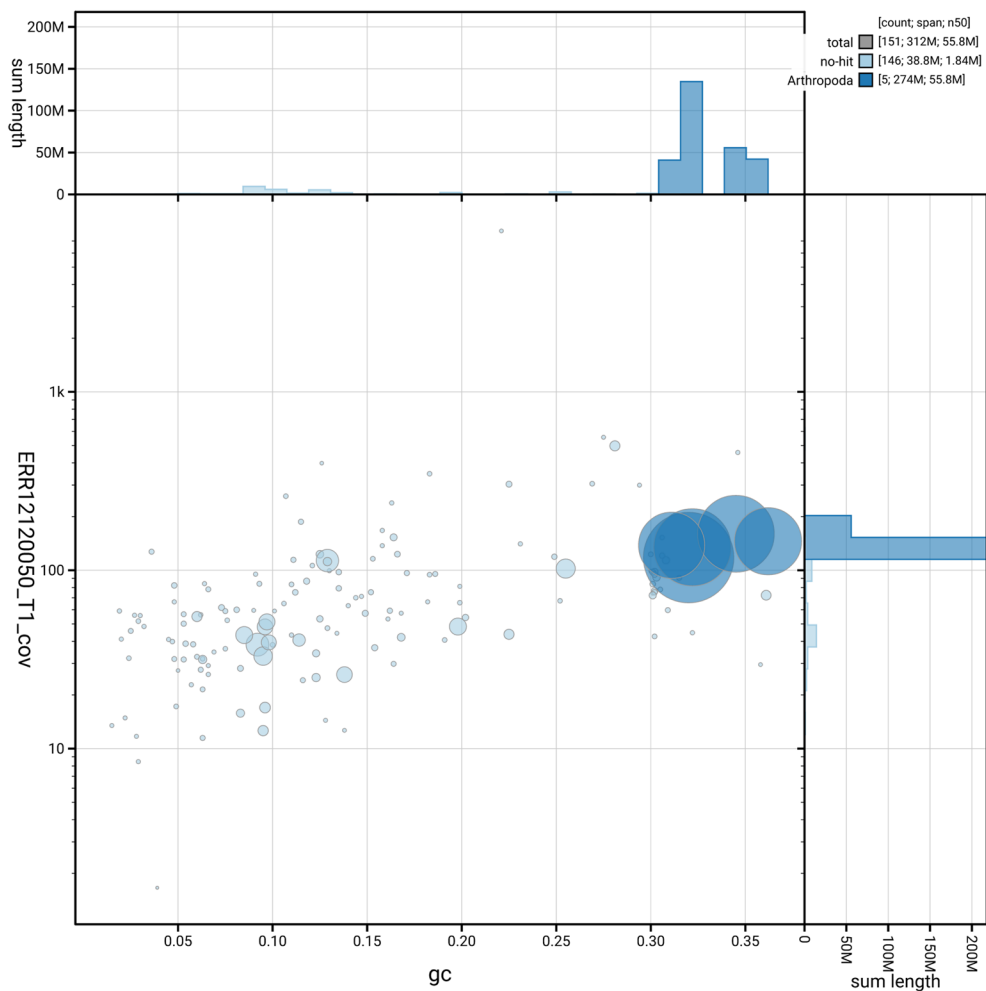


Figure 6. BlobToolKit GC-coverage plot for idChrVidu2.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the [BlobToolKit viewer](#).

Table 4. Earth Biogenome Project summary metrics for the *Chrysops viduatus* assembly.

Measure	Value	Benchmark
EBP summary (primary)	6.7.Q62	6.C.Q40
Contig N50 length	6.30 Mb	≥ 1 Mb
Scaffold N50 length	55.79 Mb	= chromosome N50
Consensus quality (QV)	Primary: 62.0; alternate: 62.4; combined: 62.2	≥ 40
k-mer completeness	Primary: 86.01%; alternate: 83.22%; combined: 99.54%	≥ 95%
BUSCO	C:93.2% [S:92.8%; D:0.5%]; F:1.2%; M:5.6%; n:3 285	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	87.60%	≥ 90%

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: *Chrysops viduatus*. Accession number [PRJEB67427](#). The genome sequence is released openly for reuse. The *Chrysops viduatus* 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 [Table 1](#) and [Table 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.

Author information

Contributors are listed at the following links:

- [Members of the Natural History Museum 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](#)

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
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
Cooler	0.8.11	https://github.com/open2c/cooler
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	0.19.5-r587	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MercuryFK	1.1.2	https://github.com/thegenemyers/MERQUERY.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	0.0.4	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView

Software	Version	Source
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	0.6.0	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

- 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](#)
- 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](#)
- Burger JF: **The biosystematics of Immature Arizona Tabanidae (Diptera)**. *Trans Am Entomol Soc (1890)*. 1977; **103**(1): 145–258.
[Reference Source](#)
- 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): gjab008.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Darwin Tree of Life Project Consortium: **Sequence locally, think globally: the Darwin Tree of Life project**. *Proc Natl Acad Sci U S A.* 2022; **119**(4): e2115642118.
[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](#)
- 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](#)
- 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): gja153.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Johnson SD: **Long-proboscid horseflies (Philolichae: Tabanidae) as pollinators of co-adapted plants in Africa and Asia**. *J Appl Entomol.* 2025; **149**(4): 466–77.
[Publisher Full Text](#)
- Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HiGlass: web-based visual exploration and analysis of genome interaction maps**. *Genome Biol.* 2018; **19**(1): 125.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute**. *PLoS One.* 2017; **12**(5): e0177459.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lawnczak MKN, Davey RP, Rajan J, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project [version 1; peer review: 2 approved with reservations]**. *Wellcome Open Res.* 2022; **7**: 187.
[Publisher Full Text](#)
- Li H: **Minimap2: pairwise alignment for nucleotide sequences**. *Bioinformatics.* 2018; **34**(18): 3094–3100.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Manni M, Berkeley MR, Seppely M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes**. *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Merkel D: **Docker: lightweight Linux containers for consistent development and deployment**. *Linux J.* 2014; **2014**(239): 2.
[Reference Source](#)
- O’Leary NA, Cox E, Holmes JB, *et al.*: **Exploring and retrieving sequence and metadata for species across the Tree of Life with NCBI datasets**. *Sci Data.* 2024; **11**(1): 732.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free 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](#)

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI taxonomy: a comprehensive update on curation, resources and tools.** *Database (Oxford).* 2020; **2020**: baaa062.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Soldierflies and Allies Recording Scheme: **Soldierflies and Allies Recording Scheme – data verified via iRecord.** Occurrence dataset, 2025.

[Publisher Full Text](#)

Stubbs AE, Drake CM: **British soldierflies and their allies: an illustrated guide to their identification and ecology: covering all flies (Diptera) in the Families Acroceridae, Asilidae, Athericidae, Bombyliidae, Rhagionidae, Scenopinidae, Stratiomyidae, Tabanidae, Therevidae, Xylomyidae and Xylophagidae.** Entomological; Natural History Society, 2014.

[Reference Source](#)

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for**

taxonomic verification in the Darwin Tree of Life project [version 1; peer review: 2 approved]. *Wellcome Open Res.* 2024; **9**: 339.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Uliano-Silva M, Ferreira JGRN, Krashenninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.

[Publisher Full Text](#)

Vicoso B, Bachtrog D: **Numerous transitions of sex chromosomes in diptera.** *PLoS Biol.* 2015; **13**(4): e1002078.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status: ? ✓ ✓

Version 1

Reviewer Report 21 January 2026

<https://doi.org/10.21956/wellcomeopenres.27769.r142531>

© 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, Tamil Nadu, India

The authors have sequenced the genome of Square-spot Deerfly *Chrysops viduatus*. The assembly reported here is of size 312.35 Mb. They have used PacBio long-read sequencing and Hi-C sequence information to obtain a high-quality genome assembly. The genome has 5 chromosomes. The sex chromosomes have not been identified because of technical difficulty. The mitochondrial genome has also been assembled of size 16.08 kb. The quality of the assembly is evident from the high N50 values. BUSCO analysis has shown that the assembly is 93.2% complete with reference to the diptera_odb10 dataset. The data note can be indexed.

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 02 January 2026

<https://doi.org/10.21956/wellcomeopenres.27769.r142530>

© 2026 Du C. 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.



Chao Du

Baotou Teachers College, Sichuan Academy of Agricultural Sciences, Baotou, China

This article presents a well-structured and comprehensive account of the genome sequencing of *Chrysops viduatus* (Square-spot Deerfly). It provides valuable reference material for the Darwin Tree of Life project. The background section offers detailed insights into the species' classification, distribution, life habits, and developmental characteristics, which are crucial for understanding the research context. The methods described are appropriate and clearly outlined. Furthermore, the openness of the data contributes to scientific communication and verification. In conclusion, this is a high-quality DATA NOTE that provides genomic data for *Chrysops viduatus*, significantly contributing to research on this species.

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

Yes

Are the protocols appropriate and is the work technically sound?

Yes

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

Yes

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics

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 17 December 2025

<https://doi.org/10.21956/wellcomeopenres.27769.r141614>

© 2025 Baird S. 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.

**Stuart J.E. Baird** 

Academy of Sciences of the Czech Republic, Národní, Staré Město, Czech Republic

This genome report differs from previous I have reviewed in three ways: first, different individuals were used for the Hi-C and PacBio sequencing; second, there are two regions where the order and orientation of scaffolds is uncertain; third the percentage of the assembly scaffolded to chromosomes is about 10% lower than appears usual for DToL. This despite 170x PacBio coverage, and 3.5x more Gb of Hi-C than PacBio.

The genome is of course worth indexing. What is lacking in the report is some comparative commentary which would allow readers of these reports to understand the sources of variance in final outcome. eg is the repeat content of 27.47% a potential explanation? Are the unordered/unoriented regions particularly rich in repeats?

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, spatial genetics

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.
