






DATA NOTE

The genome sequence of the thistle gall fly, *Urophora cardui* (Linnaeus 1758) [version 1; peer review: 4 approved]

Leila Franzen ¹, Liam M. Crowley ², Nathan Medd ³,
University of Oxford and Wytham Woods Genome Acquisition Lab,
Darwin Tree of Life Barcoding collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory
team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

¹University of Bath, Bath, England, UK

²University of Oxford, Oxford, England, UK

³University of Edinburgh, Edinburgh, Scotland, UK

V1 First published: 03 Sep 2024, 9:502
<https://doi.org/10.12688/wellcomeopenres.22919.1>
Latest published: 03 Sep 2024, 9:502
<https://doi.org/10.12688/wellcomeopenres.22919.1>

Abstract

We present a genome assembly from an individual female thistle gall fly, *Urophora cardui* (Arthropoda; Insecta; Diptera; Tephritidae). The genome sequence has a total length of 837.80 megabases. Most of the assembly is scaffolded into 6 chromosomal pseudomolecules. The mitochondrial genome has also been assembled and is 20.37 kilobases in length.

Keywords





Urophora cardui, thistle gall fly, genome sequence, chromosomal, Diptera



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

Open Peer Review

Approval Status 

	1	2	3	4
version 1				
03 Sep 2024	view	view	view	view

1. **Henrique Antonioli** , Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil
2. **Jason Charamis** , Foundation for Research and Technology - Hellas, Irákleion, Greece
3. **Daniel Gebert** , University of Cambridge, England, UK
4. **Craig Wilding** , Liverpool John Moores University, Liverpool, UK

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: **Franzen L:** Investigation, Resources; **Crowley LM:** Investigation, Resources, Writing – Review & Editing; **Medd N:** 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 [206194, <https://doi.org/10.35802/206194>] 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: © 2024 Franzen L *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: Franzen L, Crowley LM, Medd N *et al.* **The genome sequence of the thistle gall fly, *Urophora cardui* (Linnaeus 1758) [version 1; peer review: 4 approved]** Wellcome Open Research 2024, 9:502 <https://doi.org/10.12688/wellcomeopenres.22919.1>

First published: 03 Sep 2024, 9:502 <https://doi.org/10.12688/wellcomeopenres.22919.1>

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; Acalyptratae; Tephritoidea; Tephritidae; Tephritinae; Myopitini; *Urophora*; *Urophora cardui* (Linnaeus 1758) (NCBI:txid503482).

Background

Urophora cardui (Linnaeus 1758), commonly known as the Canada Thistle Gall Fly or Thistle Stem Gall Fly, is a fruit fly in the family Tephritidae. The adult fly possesses a gloss black thorax, matte black abdomen, light cream to white scutellum, and a cream head dusted in rusty orange. Its wings are clear with distinct dark bands that, unlike others in the genus, fuse at the hind margin of the wing to form distinct a 'M' shape (White, 1988).

Urophora cardui occurs throughout temperate Central Europe, with records spanning from the United Kingdom in the west to Southern Russia and the shores of Lake Baikal in the east. It has also been introduced to North America as a biological control agent to manage the population of its primary host plant, the Canada or Creeping Thistle (*Cirsium arvense*), an invasive weed in that region (Peschken & Harris, 1975). Its distribution is more latitudinally constrained with most records coming from between 45° and 65° North in its native Palearctic range between 35° and 50° North in its introduced Nearctic range (GBIF.org, 2024). *U. cardui* is able to disperse relatively large distances (Eber & Brandl, 1997; Schlumprecht, 1989) and the northern extent of its range may be increasing due to warmer summers, as demonstrated by observations in Finland (Jansson, 1991).

Urophora cardui is a stem-gall-forming parasite of *Cirsium* thistles. In the central European and western parts of its range it is restricted to *C. arvense*. However, at the extremes of its range it switches host: to *C. creticum* in the eastern Mediterranean region (White & Korneyev, 1989) and to *C. setosum* in the east, between Ukraine and northern Kazakhstan (Korneyev & White, 1996). In both cases this host shift seems to reflect the relative rarity of the fly's preferred host in these regions. In the wild females oviposit around 130 to 150 eggs in their lifetime divided into small clutches of between 1 and 12 offspring per multilocular gall (Freese & Zwölfer, 1996). Eggs, laid in the vegetative shoots, have an incubation period of 6.3 days at 24 °C (Peschken & Harris, 1975). Galls develop around 15 days after oviposition and begin maturation at approximately 36 days, at which stage the larva enters 3rd instar and gains weight rapidly (Lalonde & Shorthouse, 1985). At between 60 to 100 days post-oviposition, galls contain mature larvae which diapause overwinter and pupate within the gall during spring (Peschken & Harris, 1975). Adults emerge in early summer with most UK records falling between June and September (NBN Atlas Partnership, 2024).

Various aspects of this species ecology and life history have been studied due to its potential use as a classical biological control agent in North America (Peschken & Derby, 1997; Peschken & Harris, 1975; Zwölfer *et al.*, 1970). Since its first release in British Columbia in 1974 subsequent releases in Canada and the USA have established some seemingly stable populations which are thought to partially contribute to the ongoing control of *C. arvense* through stunting and reduced seed production (De Clerck-Floate & Cárcamo, 2011; Peschken *et al.*, 1982; Winston *et al.*, 2016).

More recently this species, and its hymenopteran parasites, have been studied for a range of different topics: gall evolution (Zwölfer, 1979), biogeographic dynamics (Eber & Brandl, 1996), host-parasitoid interactions (Johannesen & Seitz, 2003; Zwölfer *et al.*, 2007; Zwölfer & Arnold-Rinehart, 1994), and population genetics (Eber & Brandl, 1997; Johannesen *et al.*, 2010; Seitz & Komma, 1984; Steinmetz *et al.*, 2004). It is hoped that the high-quality reference genome presented here will be of great value to all these fields, especially the latter.

Genome sequence report

The genome of an adult female *Urophora cardui* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of 27.86 Gb (gigabases) from 2.05 million reads, providing approximately 31-fold coverage. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data, which produced 134.28 Gbp from 889.27 million reads, yielding an approximate coverage of 160-fold. Specimen and sequencing information is summarised in Table 1.

Manual assembly curation corrected 124 missing joins or mis-joins and five haplotypic duplications, reducing the scaffold number by 43.7%, and decreasing the scaffold N50



Figure 1. Photograph of the *Urophora cardui* (idUroCard1) specimen used for genome sequencing.

Table 1. Specimen and sequencing data for *Urophora cardui*.

Project information			
Study title	Urophora cardui (thistle gall fly)		
Umbrella BioProject	PRJEB62614		
Species	<i>Urophora cardui</i>		
BioSample	SAMEA112232592		
NCBI taxonomy ID	503482		
Specimen information			
Technology	ToLID	BioSample accession	Organism part
PacBio long read sequencing	idUroCard1	SAMEA112233058	Whole organism
Hi-C sequencing	idUroCard1	SAMEA112233058	Whole organism
Sequencing information			
Platform	Run accession	Read count	Base count (Gb)
Hi-C Illumina NovaSeq 6000	ERR11496088	8.89e+08	134.28
PacBio Sequel IIE	ERR11483519	2.05e+06	27.86

by 0.5%. The final assembly has a total length of 837.80 Mb in 66 sequence scaffolds with a scaffold N50 of 163.0 Mb (Table 2). The total count of gaps in the scaffolds is 575. The snail plot in Figure 2 provides a summary of the assembly statistics, while Figure 3 shows the distribution of base coverage against position for sequences in each chromosome. The cumulative assembly plot in Figure 4 shows curves for subsets of scaffolds assigned to different phyla. Most (99.78%) of the assembly sequence was assigned to 6 chromosomal-level scaffolds. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 3). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial genome was also assembled and can be found as a contig within the multifasta file of the genome submission.

The estimated Quality Value (QV) of the final assembly is 60.8 with *k*-mer completeness of 100.0%, and the assembly has a BUSCO v5.3.2 completeness of 96.5% (single = 96.1%, duplicated = 0.4%), using the diptera_odb10 reference set ($n = 3,285$).

Metadata for specimens, BOLD barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly statistics are given at <https://links.tol.sanger.ac.uk/species/503482>.

Methods

Sample acquisition

An adult female *Urophora cardui* (specimen ID Ox002373, ToLID idUroCard1) was netted in Wytham Woods, Oxfordshire, UK (latitude 51.77, longitude -1.34) on 2022-05-28. The specimen was collected by Leila Franzen and Liam Crowley (University of Oxford), identified by Liam Crowley (University of Oxford) and preserved on dry ice.

The initial species 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 specimens and stored in ethanol, while the remaining parts of the specimen were shipped on dry ice to the Wellcome Sanger Institute (WSI). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding have been deposited on protocols.io (Beasley *et al.*, 2023).

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life

Table 2. Genome assembly data for *Urophora cardui*, idUroCard1.1.

Genome assembly		
Assembly name	idUroCard1.1	
Assembly accession	GCA_960531455.1	
Accession of alternate haplotype	GCA_960531475.1	
Span (Mb)	837.80	
Number of contigs	642	
Contig N50 length (Mb)	3.0	
Number of scaffolds	66	
Scaffold N50 length (Mb)	163.0	
Longest scaffold (Mb)	171.21	
Assembly metrics*		Benchmark
Consensus quality (QV)	60.8	≥ 50
k-mer completeness	100.0%	≥ 95%
BUSCO**	C:96.5%[S:96.1%,D:0.4%], F:0.9%,M:2.6%,n:3,285	C ≥ 95%
Percentage of assembly mapped to chromosomes	99.78%	≥ 95%
Sex chromosomes	Not identified	localised homologous pairs
Organelles	Mitochondrial genome: 20.37 kb	complete single alleles

* Assembly metric benchmarks are adapted from column VGP-2020 of “Table 1: Proposed standards and metrics for defining genome assembly quality” from [Rhie et al. \(2021\)](#).

** BUSCO scores based on the diptera_odb10 BUSCO set using version 5.3.2. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/idUroCard1_1/dataset/idUroCard1_1/busco.

Core Laboratory includes a sequence of core procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols developed by the WSI Tree of Life laboratory are publicly available on protocols.io ([Denton et al., 2023b](#)).

In sample preparation, the idUroCard1 sample was weighed and dissected on dry ice ([Jay et al., 2023](#)). Tissue from the whole organism was homogenised using a PowerMasher II tissue disruptor ([Denton et al., 2023a](#)). HMW DNA was extracted using the Automated MagAttract v1 protocol ([Sheerin et al., 2023](#)). DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 30 ([Todorovic et al., 2023](#)). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA ([Strickland et al., 2023](#)). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High

Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

Sequencing

Pacific Biosciences HiFi circular consensus DNA sequencing libraries were constructed according to the manufacturers’ instructions. DNA sequencing was performed by the Scientific Operations core at the WSI on a Pacific Biosciences Sequel IIe instrument. Hi-C data were also generated from whole organism tissue of idUroCard1 using the Arima-HiC v2 kit. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on the Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

Assembly

The HiFi reads were first assembled using Hifiasm ([Cheng et al., 2021](#)) with the --primary option. Haplotypic duplications were identified and removed using purge_dups ([Guan et al., 2020](#)).

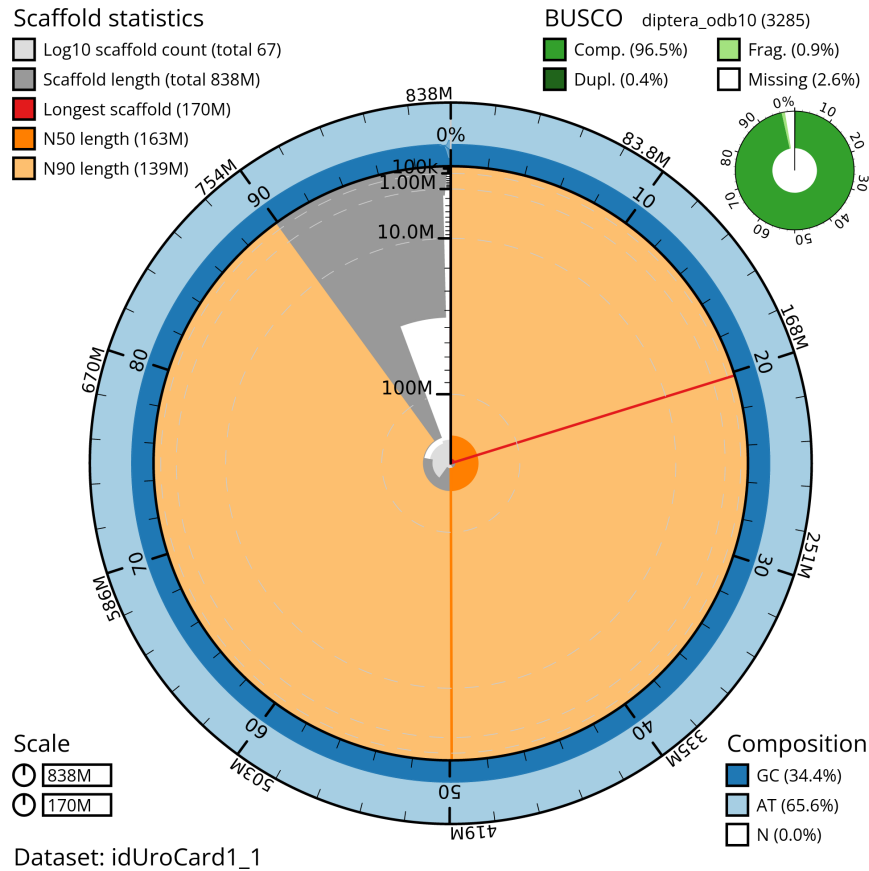


Figure 2. Genome assembly of *Urophora cardui*, idUroCard1.1: metrics. The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 837,836,207 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (169,694,088 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (162,960,411 and 139,447,831 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the diptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/idUroCard1_1/dataset/idUroCard1_1/snail.

The Hi-C reads were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). The contigs were further scaffolded using the provided Hi-C data (Rao *et al.*, 2014) in YaHS (Zhou *et al.*, 2023) using the --break option. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline (article in

preparation). Flat files and maps used in curation were generated in TreeVal (Pointon *et al.*, 2023). Manual curation was primarily conducted using PretextView (Harry, 2022), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023) and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were corrected, and duplicate sequences were tagged and removed. The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation> (article in preparation).

Evaluation of the final assembly

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). To assess the assembly metrics, the *k*-mer completeness and QV consensus quality values were calculated in Merqury (Rhie *et al.*, 2020). This work was done

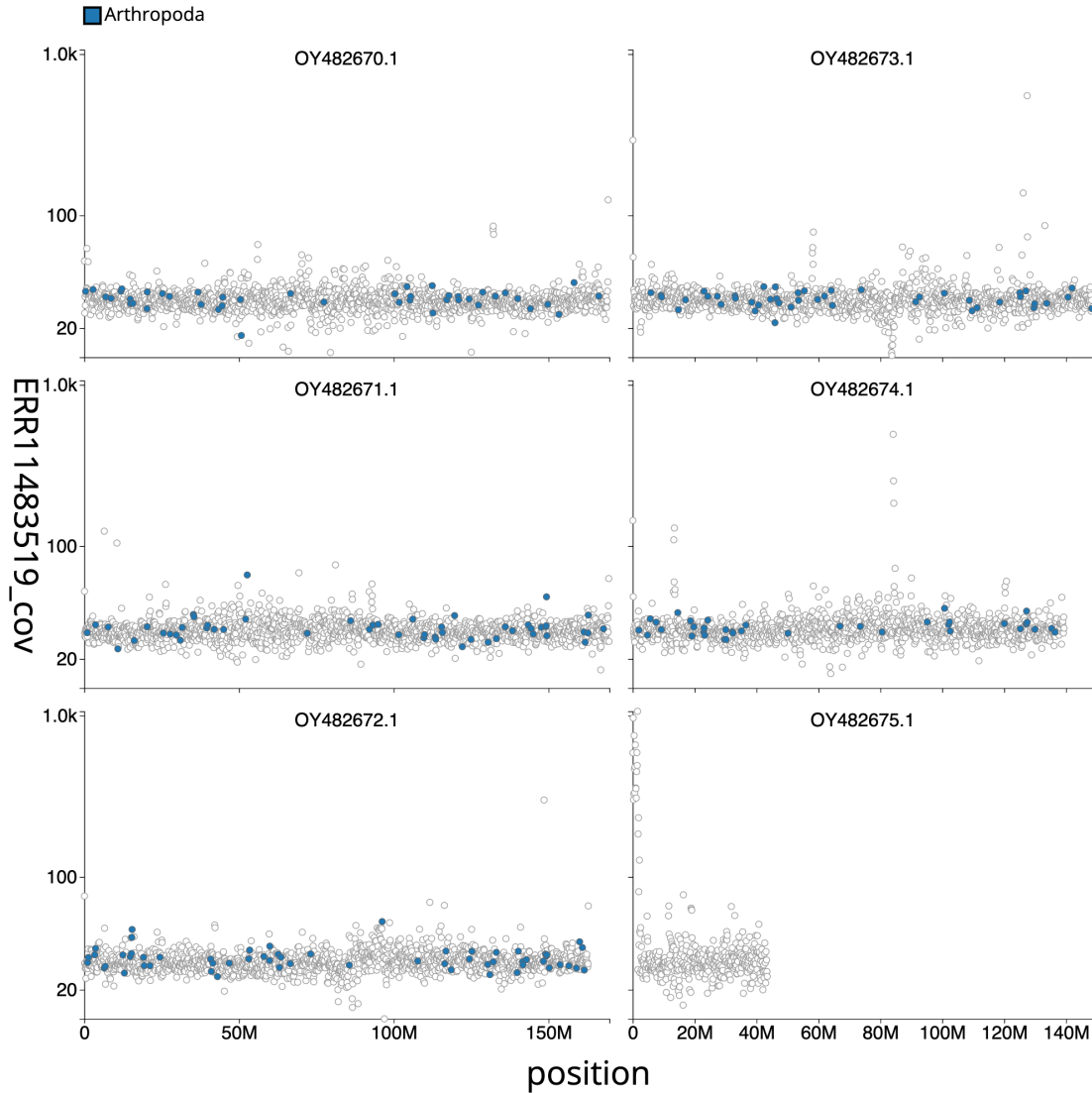


Figure 3. Genome assembly of *Urophora cardui*, idUroCard1.1: Distribution plot of base coverage in ERR11483519 against position for sequences in the assembly. Windows of 100 kb are coloured by phylum. The assembly has been filtered to exclude sequences with length < 2,550,000. An interactive version of this figure is available [here](#).

using the “sanger-tol/readmapping” (Surana *et al.*, 2023a) and “sanger-tol/genomenote” (Surana *et al.*, 2023b) pipelines. The genome readmapping pipelines were developed using the nf-core tooling (Ewels *et al.*, 2020), use MultiQC (Ewels *et al.*, 2016), and make extensive use of the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), and the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions. The genome was also analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021; Simão *et al.*, 2015) were calculated.

Table 4 contains a list of relevant software tool versions and sources.

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the ‘**Darwin Tree of Life Project Sampling Code of Practice**’, which can be found in full on the Darwin Tree of Life website [here](#). By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project.

Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature

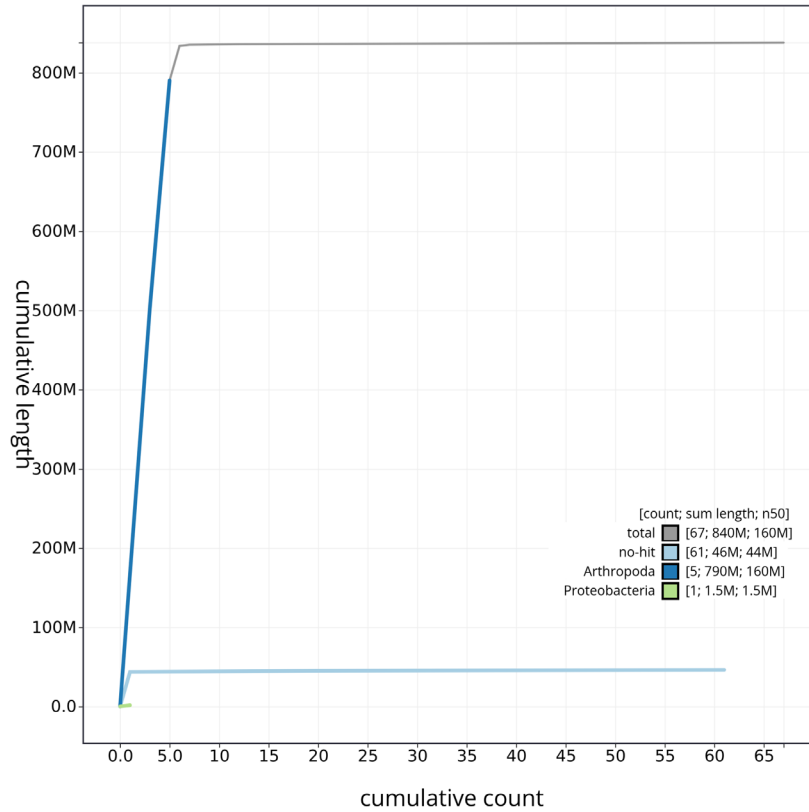


Figure 4. Genome assembly of *Urophora cardui* idUroCard1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all sequences. Coloured lines show cumulative lengths of sequences assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/idUroCard1_1/dataset/idUroCard1_1/cumulative.

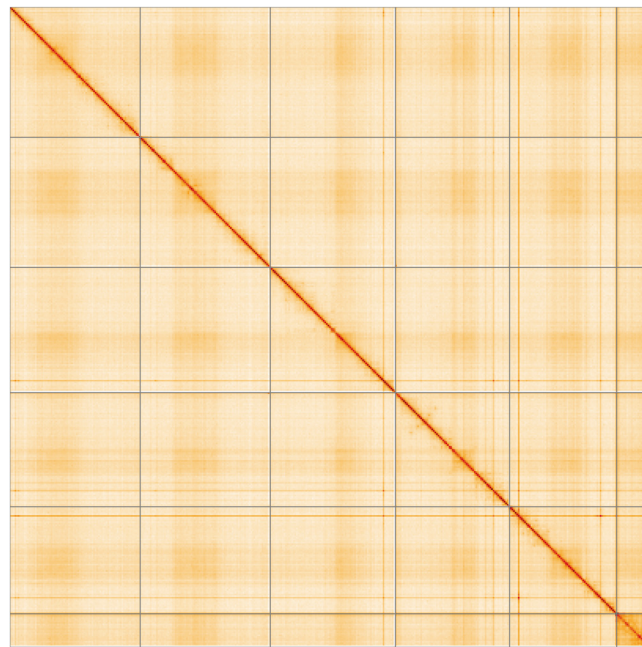


Figure 5. Genome assembly of *Urophora cardui* idUroCard1.1: Hi-C contact map of the idUroCard1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at <https://genome-note-higlass.tol.sanger.ac.uk/l/?d=L-VARUgZTPCe69o4zihIHA>.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Urophora cardui*, idUroCard1.

INSDC accession	Name	Length (Mb)	GC%
OY482670.1	1	169.38	34.5
OY482671.1	2	169.69	34.5
OY482672.1	3	162.96	34.5
OY482673.1	4	148.57	34.5
OY482674.1	5	139.45	34.5
OY482675.1	6	43.56	34.5
OY482676.1	MT	0.02	17.0

Table 4. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	4.2.1	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.3.2	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Hifiasm	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	1.11.6	https://github.com/higlass/higlass
Mercury.FK	d00d98157618f4e8d1a9190026b19b471055b22e	https://github.com/thegenemyers/MERQURY.FK
MitoHiFi	2	https://github.com/marcelauliano/MitoHiFi
PretextView	0.2	https://github.com/wtsi-hpag/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
sanger-tol/genomenote	v1.0	https://github.com/sanger-tol/genomenote
sanger-tol/readmapping	1.1.0	https://github.com/sanger-tol/readmapping/tree/1.1.0
YaHS	yahs-1.1.91eebc2	https://github.com/c-zhou/yahs

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: *Urophora cardui* (thistle gall fly). Accession number PRJEB62614; <https://identifiers.org/ena.embl/PRJEB62614> (Wellcome Sanger Institute, 2024). The

genome sequence is released openly for reuse. The *Urophora cardui* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project. 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**.

Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: <https://doi.org/10.5281/zenodo.12157525>.

Members of the Darwin Tree of Life Barcoding collective are listed here: <https://doi.org/10.5281/zenodo.12158331>

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: <https://doi.org/10.5281/zenodo.12162482>.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: <https://doi.org/10.5281/zenodo.12165051>.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: <https://doi.org/10.5281/zenodo.12160324>.

Members of the Tree of Life Core Informatics collective are listed here: <https://doi.org/10.5281/zenodo.12205391>.

Members of the Darwin Tree of Life Consortium are listed here: <https://doi.org/10.5281/zenodo.4783558>.

References

- Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 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](#)
- Beasley J, Uhl R, Forrest LL, et al.: **DNA barcoding SOPs for the Darwin Tree of Life project.** *protocols.io.* 2023; [Accessed 25 June 2024]. [Publisher Full Text](#)
- 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](#)
- da Veiga Leprevost F, Grüning BA, Alves Aflitos S, et al.: **BioContainers: an open-source and community-driven framework for software standardization.** *Bioinformatics.* 2017; **33**(16): 2580–2582. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- De Clerck-Floate R, Cárcamo H: **Biocontrol arthropods: new denizens of Canada's grassland agroecosystems.** *Arthropods Can Grassl.* 2011; **2**: 291–321. [Reference Source](#)
- Denton A, Oatley G, Cornwell C, et al.: **Sanger Tree of Life sample homogenisation: PowerMash.** *protocols.io.* 2023a. [Publisher Full Text](#)
- Denton A, Yatsenko H, Jay J, et al.: **Sanger Tree of Life wet laboratory protocol collection V.1.** *protocols.io.* 2023b. [Publisher Full Text](#)
- Diesch C, Stevens GJ, Xie P, et al.: **JBrowse 2: a modular genome browser with views of synteny and structural variation.** *Genome Biol.* 2023; **24**(1): 74. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Eber S, Brandl R: **Metapopulation dynamics of the tephritid fly *Urophora cardui*: an evaluation of incidence-function model assumptions with field data.** *J Anim Ecol.* 1996; **65**(5): 621–630. [Publisher Full Text](#)
- Eber S, Brandl R: **Genetic differentiation of the tephritid fly *Urophora cardui* in Europe as evidence for its biogeographical history.** *Mol Ecol.* 1997; **6**(7): 651–660. [Publisher 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](#)
- Freese G, Zwölfer H: **The problem of optimal clutch size in a tritrophic system: the oviposition strategy of the thistle gallfly *Urophora cardui* (Diptera, Tephritidae).** *Oecologia.* 1996; **108**(2): 293–302. [PubMed Abstract](#) | [Publisher Full Text](#)
- GBIF.org: **GBIF occurrence download.** GBIF, 2024; [Accessed 16 July 2024]. [Publisher Full Text](#)
- Grüning B, Dale R, Sjödin A, et al.: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Guan D, McCarthy SA, Wood J, et al.: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Harry E: **PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps.** 2022; [Accessed 19 October 2022]. [Reference Source](#)
- Howe K, Chow W, Collins J, et al.: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): g1aa153. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Jansson A: **Distribution and dispersal of *Urophora cardui* (Diptera, Tephritidae) in Finland in 1985–1991.** *Entomol Fenn.* 1991; **2**(4): 211–216. [Publisher Full Text](#)
- Jay J, Yatsenko H, Narváez-Gómez JP, et al.: **Sanger Tree of Life sample preparation: triage and dissection.** *protocols.io.* 2023. [Publisher Full Text](#)
- Johannesen J, Drüeke U, Seitz A: **Parapatric diversification after post-glacial range expansion in the gall fly *Urophora cardui* (Tephritidae).** *J Biogeogr.* 2010; **37**(4): 635–646. [Publisher Full Text](#)
- Johannesen J, Seitz A: **Larval distributions of the ectoparasitoid wasp *Eurytoma robusta* relative to the host tephritid gall fly *Urophora cardui*.** *Entomol Exp Appl.* 2003; **107**(1): 47–55. [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](#)
- Korneyev V, White I: **Fruit flies of the genus *Urophora* R.-D. (Diptera, Tephritidae) of Eastern Palaearctics. II. Review of species of the subgenus *Urophora* s. str. Communication 3.** *Entomol Rev.* 1996; **76**: 499–513.
- 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](#)

Lalonde RG, Shorthouse JD: **Growth and development of larvae and galls of *Urophora cardui* (Diptera, Tephritidae) on *Cirsium arvense* (Compositae).** *Oecologia*. 1985; **65**(2): 161–165.

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

Manni M, Berkeley MR, Seppely M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol*. 2021; **38**(10): 4647–4654.

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

Merkel D: **Docker: lightweight linux containers for consistent development and deployment.** *Linux J*. 2014; **2014**(239): 2, [Accessed 2 April 2024].

[Reference Source](#)

NBN Atlas Partnership: ***Urophora cardui* (Linnaeus, 1758).** NBN Atlas, 2024; [Accessed 16 July 2024].

[Reference Source](#)

Peschken D, Derby J: **Establishment of *Urophora cardui* (Diptera: Tephritidae) on Canada thistle, *Cirsium arvense* (Asteraceae), and colony development in relation to habitat and parasitoids in Canada.** In: *Vertical Food Web Interactions: Evolutionary Patterns and Driving Forces*. Springer, 1997; 53–66.

[Publisher Full Text](#)

Peschken DP, Finnamore DB, Watson AK: **Biocontrol of the weed Canada thistle (*Cirsium arvense*): releases and development of the gall fly *Urophora cardui* (Diptera: Tephritidae) in Canada.** *Can Entomol*. 1982; **114**(4): 349–357.

[Publisher Full Text](#)

Peschken DP, Harris P: **Host specificity and biology of *Urophora cardui* (Diptera: Tephritidae). A biocontrol agent for Canada thistle (*Cirsium arvense*).** *Can Entomol*. 1975; **107**(10): 1101–1110.

[Publisher Full Text](#)

Pointon DL, Eagles W, Sims Y, *et al.*: **sanger-tol/treeval v1.0.0 – Ancient Atlantis.** 2023.

[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): 245.

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

Schlumprecht H: **Dispersal of the thistle gallfly *Urophora cardui* and its endoparasitoid *Eurytoma serraiulae* (Hymenoptera: Eurytomidae).** *Ecol Entomol*. 1989; **14**(3): 341–348.

[Publisher Full Text](#)

Seitz A, Komma M: **Genetic polymorphism and its ecological background in *Tephritid* populations (Diptera: Tephritidae).** In: *Population Biology and Evolution*. Springer, 1984; 143–158.

[Publisher Full Text](#)

Sheerin E, Sampaio F, Oatley G, *et al.*: **Sanger Tree of Life HMW DNA extraction: automated MagAttract v.1.** *protocols.io*. 2023; [Accessed 21 November 2023].

[Publisher Full Text](#)

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics*. 2015; **31**(19): 3210–3212.

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

Steinmetz R, Johannesen J, Seitz A: **Clinal genetic variation and the 'rare**

allele phenomenon' in random mating populations of *Urophora cardui* (Diptera: Tephritidae). *Genetica*. 2004; **122**(3): 277–290.

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

Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *protocols.io*. 2023.

[Publisher Full Text](#)

Surana P, Muffato M, Qi G: **sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo*. 2023a.

[Publisher Full Text](#)

Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev).** *Zenodo*. 2023b.

[Publisher Full Text](#)

Todorovic M, Sampaio F, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor#3 for PacBio HiFi.** *protocols.io*. 2023.

[Publisher Full Text](#)

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life Project [version 1; peer review: awaiting peer review].** *Wellcome Open Res*. 2024; **9**: 339.

[Publisher Full Text](#)

Uliano-Silva M, Ferreira JGRN, Krashennikova 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](#)

Wellcome Sanger Institute: **The genome sequence of the thistle gall fly, *Urophora cardui* (Linnaeus 1758).** European Nucleotide Archive. [dataset], accession number PRJEB62614, 2024.

White IM: **Tephritid flies: Diptera: Tephritidae, 1.** In: *Handbooks for the Identification of British Insects*. London: Royal Entomological Society of London, 1988.

[Reference Source](#)

White IM, Korneyev VA: **A revision of the western Palaearctic species of *Urophora* Robineau-Desvoidy (Diptera: Tephritidae).** *Syst Entomol*. 1989; **14**(3): 327–374.

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

Winston R, Bell C, De Clerck-Floate R, *et al.*: **Biological control of weeds in the northwest.** The Forest Health Technology Enterprise Team, 2016.

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](#)

Zwölfer H: **Strategies and counterstrategies in insect population systems competing for space and food in flower heads and plant galls.** *Fortschr D Zool*. 1979; **25**(2–3): 331–353.

[Reference Source](#)

Zwölfer H, Arnold-Rinehart J: **The evolution of interactions and diversity in plant-insect systems: the *Urophora-Eurytoma* food web in galls on palearctic cardueae.** In: *Biodiversity and Ecosystem Function*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994; 211–233.

[Publisher Full Text](#)

Zwölfer H, Böheim M, Beck E: ***Eurytoma robusta* (Hymenoptera: Eurytomidae), a local key factor in the population dynamics of *Urophora cardui* (Diptera: Tephritidae): a comparative analysis.** *Eur J Entomol*. 2007; **104**(2): 217–224.

[Publisher Full Text](#)

Zwölfer H, Englert W, Pattullo W: **Investigations on the biology, population ecology and distribution of *Urophora cardui* (L).** Weed Projects for Canada. *Prog. Rep*. 1970.

Open Peer Review

Current Peer Review Status:    

Version 1

Reviewer Report 27 September 2024

<https://doi.org/10.21956/wellcomeopenres.25236.r100659>

© 2024 Wilding 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.



Craig Wilding 

Liverpool John Moores University, Liverpool, UK

Franzen et al. present a high-quality genome sequence of the fly *Urophora cardui*. The introduction is clear and presents useful information on the biology, distribution, and use of this species in biological control.

The methods are clear. The sequence is extremely high quality - chromosomal length (6 chromosomal pseudomolecules) and with a very high BUSCO completeness.

My only comments are:

- Figure 1 - I expect the tube is used for scale, but without knowledge of the size of the tube this is not particularly useful - can a scale bar be added or the legend state the length of the tube?
- Figure 3 - I am not clear exactly what this is showing. The blue dots are Arthropoda. Does this mean these are example Arthropoda to demonstrate that the sequences from *Urophora cardui* are of a similar value or the average value of other arthropods (which does not really make sense in this plot)? If the blue dots are *Urophora cardui* (which is an arthropod) what are the other dots? This plot just needs a better explanation for those unfamiliar with such a presentation. I am also wondering why there is a peak in coverage at the start of pseudomolecule OY482675.1? This is not mentioned in the text.

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:** Genetics/genomics of mosquitoes, spiders and sea anemones**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 27 September 2024

<https://doi.org/10.21956/wellcomeopenres.25236.r99474>

© 2024 Gebert 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.

**Daniel Gebert** 

University of Cambridge, England, UK

The authors describe a genome assembly from an individual female thistle gall fly (*Urophora cardui*), which they generated using PacBio sequencing, followed by HiC scaffolding of the initially assembled contigs into chromosomes (chromosomal pseudomolecules), including the mitochondrial genome. A total of 99.78% of the assembly sequence (altogether 837.80 Mb) was assigned to six chromosome-level scaffolds, named in order of their size. However, sex chromosomes (likely the X chromosome in this female specimen) were not identified, which is unsurprising since, in close relatives such as *Bactrocera dorsalis*, which also has six chromosomes, the sex chromosomes remain unidentified as well.

The methodology is sound, and the assembly evaluation is rigorous. While the sequence report section itself is very short and could benefit from further elaboration, it contains the most important basic information for a typical DToL assembly report. The authors also provide a detailed description of the significance of this species and the value of having a high-quality genome assembly, which is very convincing.

However, it would be beneficial to state in more detail which options were used for specific purposes during assembly and scaffolding. For instance, what is the purpose of the Hifiasm --primary option, and why was it chosen? Similarly, what is the --break option in yahs, and how does it affect scaffolding? What was the mapq cutoff? Including these details would help the reader fully understand and recapitulate the methodology.

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?

Partly

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genome Biology, Evolution, Computational Biology

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 26 September 2024

<https://doi.org/10.21956/wellcomeopenres.25236.r100623>

© 2024 Charamis 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.



Jason Charamis 

Foundation for Research and Technology - Hellas, Irákleion, Greece

This work describes the genome sequence of the Canada Thistle Gall Fly. The work is technically sound, both in terms of wet lab and computational approaches. The methods and resulting assembly well described in the article and will be an important contribution to ecological genomic studies. I suggest indexing this paper.

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: Arthropod Comparative and Computational 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 25 September 2024

<https://doi.org/10.21956/wellcomeopenres.25236.r100629>

© 2024 Antonioli 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.



Henrique Antonioli 

Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil

The present study reports a highly contiguous genome assembly of *Urophora cardui*, an important fruit fly used as biological control against an invasive weed in North America. The authors bring a clear and concise introduction, providing valuable information regarding its general biology and ecology; in this sense, the rationale for creating these datasets is clearly stated throughout the text: gall evolution, biogeography, insect-plant interactions, and population genomics. Protocols are well described and provided in the methodology section as supplementary material, fully accessible in other repositories. Accession numbers for both raw and assembled data are given, enabling replication by others. Main statistics are also provided, showing that the assembly reached satisfactory levels of completeness. Although annotation of the nuclear genome was not performed, authors stated that this will be provided soon by employing RNA-seq data; annotation of the mitochondrial genome should be performed as well. The paper also lack an estimation of the repetitive content, which may be performed in future studies.

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; transposable elements

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.