DATA NOTE

The genome sequence of *Tachina fera* Linnaeus, 1761, a tachinid fly [version 1; peer review: 1 approved, 1 approved with reservations]

University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Will Nash\(^1\), Darwin Tree of Life Consortium

\(^1\)Earlham Institute, Norwich, UK

**Abstract**

We present a genome assembly from an individual female *Tachina fera* (Arthropoda; Insecta; Diptera; Tachinidae). The genome sequence is 752 megabases in span. The majority of the assembly (99.98%) is scaffolded into 6 chromosomal pseudomolecules, with the X sex chromosome assembled. The complete mitochondrial genome was also assembled and is 17.4 kilobases in length.

**Keywords**

*Tachina fera*, genome sequence, chromosomal, Diptera

This article is included in the Tree of Life gateway.

**Open Peer Review**

Approval Status

1. Jaakko L.O. Pohjoismäki\(^2\), University of Eastern Finland, Joensuu, Finland

2. Sheina Sim\(^2\), USDA Agricultural Research Service, Hilo, USA

Any reports and responses or comments on the article can be found at the end of the article.
Species taxonomy
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Oestroidea; Tachinidae; Tachiniinae; Tachinini; Tachina; Tachina fera Linnaeus, 1761 (NCBI:txid631328).

Background
Tachina fera (Linnaeus, 1761) is one of the most striking flies commonly encountered in the UK countryside. With adults ranging between 9 and 14 mm in length, it is an easily noticeable fly. Spiky bristles, characteristic of the Tachinidae, adorn a chestnut abdomen with a dark central stripe. Tachina fera is abundant across Europe, North Africa and Asia (Tschorsnig & Herting, 1994). In the UK, T. fera is bivoltine, with adults in flight from May to June, and from July to September (Belshaw, 1993). Adults feed at a range of flowers throughout the landscape. Tachina fera has mainly been recorded emerging from Noctuid moth caterpillars (Belshaw, 1993). The method of parasitism utilised by T. fera is notable as the egg is not placed into the host by the mother but laid pre-incubated onto leaves close to it. The larva, once hatched, will make its own way to the host, stimulated by vibration (Belshaw, 1993; Stireman et al., 2006). The parasitic nature of Tachinid species such as T. fera mean they are important, but underappreciated, regulators of insect herbivory in our ecosystem (Stireman et al., 2006), as well as playing important roles in pollination (e.g. Martel et al., 2021). The chromosome-level genome assembly presented here is, to our knowledge, the first high-quality resource developed for a Tachinid and represents a key step in understanding the complex ecology of these beautiful and spiky flies.

Genome sequence report
The genome was sequenced from a single female T. fera collected from Wytham Woods, Oxfordshire (Biological vice-county: Berkshire), UK (latitude 51.770, longitude -1.338) (Figure 1). A total of 41-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 46-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 246 missing/misjoins and removed 60 haplotypic duplications, reducing the assembly size by 1.88% and the scaffold number by 94.81%, and increasing the scaffold N50 by 120.51%.

The final assembly has a total length of 752 Mb in 12 sequence scaffolds with a scaffold N50 of 142 Mb (Table 1). The majority, 99.98%, of the assembly sequence was assigned to 6 chromosomal-level scaffolds, representing 5 autosomes (numbered by sequence length), and the X sex chromosome (Figure 2–Figure 5; Table 2). The order and orientation of contigs within the centromere of chromosome 2 are not known. Lots

Table 1. Genome data for Tachina fera, idTacFera2.1.

<table>
<thead>
<tr>
<th>Project accession data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly identifier</td>
</tr>
<tr>
<td>Species</td>
</tr>
<tr>
<td>Specimen</td>
</tr>
<tr>
<td>NCBI taxonomy ID</td>
</tr>
<tr>
<td>BioProject</td>
</tr>
<tr>
<td>BioSample ID</td>
</tr>
<tr>
<td>Isolate information</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Raw data accessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>PacificBiosciences SEQUEL II</td>
</tr>
<tr>
<td>10X Genomics Illumina</td>
</tr>
<tr>
<td>Hi-C Illumina</td>
</tr>
<tr>
<td>PolyA RNA-Seq Illumina</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genome assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly accession</td>
</tr>
<tr>
<td>Accession of alternate haplotype</td>
</tr>
<tr>
<td>Span (Mb)</td>
</tr>
<tr>
<td>Number of contigs</td>
</tr>
<tr>
<td>Contig N50 length (Mb)</td>
</tr>
<tr>
<td>Number of scaffolds</td>
</tr>
<tr>
<td>Scaffold N50 length (Mb)</td>
</tr>
<tr>
<td>Longest scaffold (Mb)</td>
</tr>
<tr>
<td>BUSCO* genome score</td>
</tr>
</tbody>
</table>

*BUSCO scores based on the diptera_odb10 BUSCO set using v5.1.2. 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.genomeweb.org/view/idTacFera2.1/dataset/CAJMZS01/busco.

Figure 1. Image of the Tachina fera specimen taken during preservation and processing.
of apparent haplotypic duplication was excised from this region owing to a divergent Hi-C pattern and seeming low coverage (which was somewhat ambiguous due to read coverage levels in this repetitive region).

The assembly has a BUSCO v5.1.2 (Manni et al., 2021) completeness of 98.4% (single 97.9%, duplicated 0.5%) using the diptera_odb10 reference set (n=3285). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

**Methods**

**Sample acquisition and DNA extraction**

One female *T. fera* sample (idTacFera2), and a second sample of unknown sex (idTacFera1) was collected from Wytham Woods, Oxfordshire (Biological vice-county: Berkshire), UK (latitude 51.770, longitude -1.338) by Liam Crowley, University of Oxford, on 15 June 2020. The specimen was caught in woodland with a net, identified by the same individual, snap-frozen on dry ice and stored using a CoolRack.

DNA was extracted from the head/thorax of idTacFera2 at the Wellcome Sanger Institute (WSI) Scientific Operations core using the Qiagen MagAttract HMW DNA kit, according to the manufacturer’s instructions. RNA (from the abdomen of idTacFera1) was extracted in the Tree of Life Laboratory at the WSI using TRIZol, according to the manufacturer’s instructions. RNA was then eluted in 50 μl RNAse-free water and its concentration RNA assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit RNA Broad-Range (BR) Assay kit. Analysis of the integrity of the RNA was done...
Figure 3. Genome assembly of Tachina fera, idTacFera2.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/idTacFera2.1/dataset/CAJMZS01/blob.

Sequencing
Pacific Biosciences HiFi circular consensus and 10X Genomics Chromium read cloud sequencing libraries were constructed according to the manufacturers’ instructions. Sequencing was performed by the Scientific Operations core at the Wellcome Sanger Institute on Pacific Biosciences SEQUEL II (HiFi), Illumina HiSeq X (10X) and Illumina HiSeq 4000 (RNA-Seq) instruments. Hi-C data were generated in the Tree of Life laboratory from remaining head/thorax tissue of idTacFera2 using the Arima v2 kit and sequenced on a HiSeq X instrument.

Genome assembly
Assembly was carried out with HiFiasm (Cheng et al., 2021); haplotypic duplication was identified and removed with purge_dups (Guan et al., 2020). One round of polishing was performed by aligning 10X Genomics read data to the assembly with longranger align, calling variants with freebayes (Garrison & Marth, 2012). The assembly was then scaffolded with Hi-C data (Rao et al., 2014) using SALSA2 (Ghurye et al., 2019). The assembly was checked for contamination and corrected using gEVAL (Chow et al., 2016) as described previously (Howe et al., 2021). Manual curation was performed using gEVAL, HiGlass (Kerpedjiev et al., 2018) and Pretext. The mitochondrial genome was assembled using MitoHiFi.
**Figure 4.** Genome assembly of *Tachina fera*, idTacFera2.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/idTacFera2.1/dataset/CAJMZS01/cumulative.

**Figure 5.** Genome assembly of *Tachina fera*, idTacFera2.1: Hi-C contact map. Hi-C contact map of the idTacFera2.1 assembly, visualised in HiGlass. Chromosomes are arranged in size order from left to right and top to bottom.
(Uliano-Silva et al., 2021), which performs annotation using MitoFinder (Allio et al., 2020). The genome was analysed and BUSCO scores generated within the BlobToolKit environment (Challis et al., 2020). Table 3 contains a list of all software tool versions used, where appropriate.

### Ethics/compliance issues
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. 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. 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

The genome sequence is released openly for reuse. The *T. fera* 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 the 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.

### Author information
Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.5746938.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.5744972.

Members of the Wellcome Sanger Institute Tree of Life programme are listed here: https://doi.org/10.5281/zenodo.6125027.

Members of Wellcome Sanger Institute Scientific Operations; DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.5746904.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.6125046.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.5638618.

---

### Table 2. Chromosomal pseudomolecules in the genome assembly of *Tachina fera*, idTacFera2.1.

<table>
<thead>
<tr>
<th>INSDC accession</th>
<th>Chromosome</th>
<th>Size (Mb)</th>
<th>GC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR999963.1</td>
<td>1</td>
<td>191.82</td>
<td>30.1</td>
</tr>
<tr>
<td>LR999964.1</td>
<td>2</td>
<td>150.60</td>
<td>30.7</td>
</tr>
<tr>
<td>LR999965.1</td>
<td>3</td>
<td>142.00</td>
<td>30.2</td>
</tr>
<tr>
<td>LR999966.1</td>
<td>4</td>
<td>127.83</td>
<td>30.2</td>
</tr>
<tr>
<td>LR999967.1</td>
<td>5</td>
<td>125.18</td>
<td>30.0</td>
</tr>
<tr>
<td>LR999968.1</td>
<td>X</td>
<td>14.18</td>
<td>31.7</td>
</tr>
<tr>
<td>LR999969.1</td>
<td>MT</td>
<td>0.02</td>
<td>19.9</td>
</tr>
<tr>
<td>-</td>
<td>Unplaced</td>
<td>0.11</td>
<td>30.2</td>
</tr>
</tbody>
</table>

---

### Table 3. Software tools used.

<table>
<thead>
<tr>
<th>Software tool</th>
<th>Version</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hiiasm</td>
<td>0.12</td>
<td>Cheng et al., 2021</td>
</tr>
<tr>
<td>purge_dups</td>
<td>1.2.3</td>
<td>Guan et al., 2020</td>
</tr>
<tr>
<td>SALSA2</td>
<td>2.2</td>
<td>Ghurye et al., 2019</td>
</tr>
<tr>
<td>longranger align</td>
<td>2.2.2</td>
<td><a href="https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines">https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines</a></td>
</tr>
<tr>
<td>freebayes</td>
<td>1.3.1-17-gaa2ace8</td>
<td>Garrison &amp; Marth, 2012</td>
</tr>
<tr>
<td>MitoHiFi</td>
<td>1.0</td>
<td>Uliano-Silva et al., 2021</td>
</tr>
<tr>
<td>gEVAL</td>
<td>N/A</td>
<td>Chow et al., 2016</td>
</tr>
<tr>
<td>HiGlass</td>
<td>1.11.6</td>
<td>Kerpedjiev et al., 2018</td>
</tr>
<tr>
<td>PretextView</td>
<td>0.1.x</td>
<td><a href="https://github.com/wtsi-hpag/PretextView">https://github.com/wtsi-hpag/PretextView</a></td>
</tr>
<tr>
<td>BlobToolKit</td>
<td>3.0.5</td>
<td>Challis et al., 2020</td>
</tr>
</tbody>
</table>
References


Open Peer Review

Current Peer Review Status: 

Version 1

Reviewer Report 08 August 2022

https://doi.org/10.21956/wellcomeopenres.19657.r51184

© 2022 Sim 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.

Sheina Sim

Daniel K. Inouye US Pacific Basin Agricultural Research Center, Tropical Crop and Commodity Protection Research Unit, USDA Agricultural Research Service, Hilo, HI, USA

The Tachina fera genome was sequenced and assembled to a chromosome length and represents the highest quality genome for a tachinid fly. The work described is appropriate for the goal and is well written, though missing some details.

Major comments:
- The details for some of the analysis methods were not defined. I see that there is a blobplot provided, and that requires additional analyses to determine the taxonomic identifications and depth of coverage for each fragment, but those details were not conveyed.
- Please describe how the X chromosome was identified as those methods were not included in the manuscript.
- The reason for conducting this genome assembly effort was not well articulated. What can you learn from this assembly? What will you use the assembly for? You describe the species as an under appreciated regulator of herbivory, how will the genome improve your understanding of the species?

Minor comments:
- The snail plot is divided into 100 (not 1000) size-ordered bins

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

Are the protocols appropriate and is the work technically sound?
Yes

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

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

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Insect genomics, population genetics, and bioinformatics.

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.

---

**Jaakko L.O. Pohjoismäki**  
Department of Environmental and Biological Sciences, University of Eastern Finland, Joensuu, Finland

Thank you for the opportunity to review the data note manuscript by Will Nash et al. on the genome sequence of *Tachina fera* (Diptera: Tachinidae).

The approach taken by DToL can be seen as a benchmark for many other genome sequencing initiatives for providing a state-of-the-art, chromosomally scaffolded reference genome for the species. It is great to see reference genomes from the magnificent and important group of true flies.

I have only few minor suggestions to improve the content of the manuscript.

**Title:**
The species author’s name needs to be in parentheses - *Tachina fera* (Linnaeus, 1761) [originally *Musca fera* Linnaeus, 1761]. For clarity, the insect order and family should be added to the title (Diptera, Tachinidae).

**Species taxonomy:**
ibid for the author’s name.

**Background:**
Very nice introduction. The abdomen is maybe more of orange than chestnut, although this can vary, and the females are often darker than the males. There is another very similar species in the
UK (although apparently until now only on the Channel Islands), *Tachina magnicornis* (Zetterstedt, 1844). This is a bit problematic for the story as the females are difficult to tell apart with certainty. It is very unlikely that you would have happened to sample the rarer species, but you might need to make this reservation in the introduction.

Figure 1 is quite poor when it comes to specimen details. For vouchering, I would recommend taking more detailed images. I would for example think (based on the shape of the abdomen and apparently narrow frons, although this is poorly visible) that this is a male specimen. Or can it be that the image is from the other specimen used for the RNA-seq?

**Genome sequence report:**
Looks very good. To be sure of the specimen, you could check the sex by (most calyptrate flies follow XY-system of sex determination) by looking at the existence of the dominant male-determining factor in your sequence data (e.g. Vicoso & Bachtrog, 2015\(^1\)). Was the X chromosome present as diploid or haploid sequence count?

However, I see no presentation of the RNA-seq data? How many transcripts, what coverage etc?

**Sample acquisition & nucleic acids extraction:**
Were the whole specimens destroyed in the DNA and RNA extractions or is there some reference tissue left? Where (and how) are these stored and how can they be located? If the reference specimens are still existing, describe all associated labels in detail. For later morphological analysis it would be great to preserve at least the abdomens and legs as a voucher. The voucher should be placed in a public collection. I am sure you have some established practice within DToL but it needs to be described here.

Also: “The specimen was caught in […]” -> The specimens **were** caught ... (there was two).

**RNA extraction:**
Was there no poly-A purification? How was the rRNA depleted before the sequencing?

**Suggestion for the future:**
Depending on your RNA-seq results, it might make sense to extract RNA from the head+thorax (especially when there are more than one specimen). The RNA yield is often poor from the abdomen (on average) due to high levels of digestive enzymes, mass of gut content, fat etc. Also the head+thorax could give a better overview of the gene expression.

**References**

**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:** Molecular biology, genetics, population genetics, evolution, Diptera 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.