The genome sequence of a tachinid fly, *Nowickia ferox* (Panzer, 1809) [version 1; peer review: awaiting peer review]

Steven Falk¹, Chris Raper²,
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, Darwin Tree of Life Consortium

¹Independent researcher, Kenilworth, England, UK
²Natural History Museum, London, England, UK

**Abstract**
We present a genome assembly from an individual female *Nowickia ferox* (a tachinid fly; Arthropoda; Insecta; Diptera; Tachinidae). The genome sequence is 670.7 megabases in span. Most of the assembly is scaffolded into 6 chromosomal pseudomolecules, including the X sex chromosome. The mitochondrial genome has also been assembled and is 17.19 kilobases in length. Gene annotation of this assembly on Ensembl identified 27,893 protein coding genes.

**Keywords**
Nowickia ferox, a tachinid fly, genome sequence, chromosomal, Diptera

This article is included in the Tree of Life gateway.
Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: Falk S: Investigation, Resources; Raper C: 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) and the Darwin Tree of Life Discretionary Award (218328). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2023 Falk S 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: Falk S, Raper C, University of Oxford and Wytham Woods Genome Acquisition Lab et al. The genome sequence of a tachinid fly, Nowickia ferox (Panzer, 1809) [version 1; peer review: awaiting peer review] Wellcome Open Research 2023, 8:275 https://doi.org/10.12688/wellcomeopenres.19575.1

First published: 23 Jun 2023, 8:275 https://doi.org/10.12688/wellcomeopenres.19575.1
Species taxonomy
Eukaryota; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Eremoneura; Cyclorrhapha; Schizophora; Calyptratae; Oestroidea; Tachinidae; Tachininae; Tachinini; Nowickia; Nowickia ferox (Panzer, 1809) (NCBI:txid613196).

Background
Nowickia ferox (Panzer, 1809), is a large tachinid fly with a maximum body length of ~18 mm. It is common south of a line drawn between South Wales and the Wash, with a few scattered records further north. It is most likely to be confused with Tachina fera (Linnaeus, 1761) but N. ferox is dark black and orange with black legs; while T. fera is brown and orange with orange/brown legs. The host for this species is usually the Dark Arches moth, Apamea monoglypha (Hufnagel, 1766) (Belshaw, 1993). The adults nectar on a wide range of flowers and can be seen from mid-June to late September, on heathland, woodland margins and gardens, in one brood (Tschorsnig & Herting, 1994).

There has been a move (O’Hara, 2020) to classify Nowickia as subgenus of Tachina, but European taxonomists are yet to adopt this reorganisation.

The genome of Nowickia ferox was sequenced as part of the Darwin Tree of Life Project, a collaborative effort to sequence all named eukaryotic species in the Atlantic Archipelago of Britain and Ireland. Here we present a chromosomally complete genome sequence for Nowickia ferox, based on one female specimen from Wytham Woods.

Genome sequence report
The genome was sequenced from one female Nowickia ferox (Figure 1) collected from Wytham Woods, Oxfordshire (biological vice-county Berkshire), UK (51.77, –1.33). A total of 41-fold coverage in Pacific Biosciences single-molecule HiFi long reads was generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 79 missing joins or mis-joins and removed 7 haplotypic duplications, increasing the assembly length by 2.4% and the scaffold number by 22.56%, and increasing the scaffold N50 by 7.23%.

The final assembly has a total length of 670.7 Mb in 103 sequence scaffolds with a scaffold N50 of 116.6 Mb (Table 1). Most (99.95%) of the assembly sequence was assigned to 6 chromosomal-level scaffolds, representing 5 autosomes and the X sex chromosome. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 2–Figure 5; Table 2). 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 65.5 with k-mer completeness of 100%, and the assembly has a BUSCO v5.3.2 completeness of 98.4% (single = 98.1%, duplicated = 0.3%), using the diptera_odb10 reference set (n = 3,285).

Metadata for specimens, spectral estimates, sequencing runs, contaminants and pre-curation assembly statistics can be found at https://links.tol.sanger.ac.uk/species/613196.

Genome annotation report
The Nowickia ferox genome assembly (GCA_936439885.1) was annotated using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Nowickia_ferox_GCA_936439885.1/Info/Index). The resulting annotation includes 28,730 transcribed mRNAs from 27,893 protein-coding genes.

Methods
Sample acquisition and nucleic acid extraction
A female Nowickia ferox (idNowFero1) was collected from Wytham Woods, Oxfordshire (biological vice-county Berkshire), UK (latitude 51.77, longitude –1.33) on 2020-08-04 by netting. The specimen was collected and identified by Steven Falk (University of Oxford) and snap-frozen on dry ice. DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute (WSI). The idNowFero1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. Thorax tissue was cryogenically disrupted to a fine powder using a Covaris cryoPREP Automated Dry Pulveriser, receiving multiple impacts. High molecular weight (HMW) DNA was extracted using the Qiagen MagAttract HMW DNA extraction kit. HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 30. Sheared DNA was purified by solid-phase reversible immobilisation using AMPure PB beads with a 1.8X ratio of beads to sample to remove the shorter DNA.
Table 1. Genome data for *Nowickia ferox*, idNowFero1.1.

<table>
<thead>
<tr>
<th>Project accession data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly identifier</td>
<td>idNowFero1.1</td>
</tr>
<tr>
<td>Species</td>
<td><em>Nowickia ferox</em></td>
</tr>
<tr>
<td>Specimen</td>
<td>idNowFero1</td>
</tr>
<tr>
<td>NCBI taxonomy ID</td>
<td>613196</td>
</tr>
<tr>
<td>BioProject</td>
<td>PRJEB50974</td>
</tr>
<tr>
<td>BioSample ID</td>
<td>SAMEA7746479</td>
</tr>
<tr>
<td>Isolate information</td>
<td>idNowFero1: thorax (DNA sequencing), head (Hi-C scaffolding)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assembly metrics*</th>
<th>Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus quality (QV)</td>
<td>65.6</td>
</tr>
<tr>
<td>≥ 50</td>
<td></td>
</tr>
<tr>
<td>k-mer completeness</td>
<td>100%</td>
</tr>
<tr>
<td>≥ 95%</td>
<td></td>
</tr>
<tr>
<td>BUSCO**</td>
<td>C:98.4%{S:98.1%,D:0.3%}, F:0.7%,M:0.9%,n:3,285</td>
</tr>
<tr>
<td>C ≥ 95%</td>
<td></td>
</tr>
<tr>
<td>Percentage of assembly mapped to chromosomes</td>
<td>99.95%</td>
</tr>
<tr>
<td>≥ 95%</td>
<td></td>
</tr>
<tr>
<td>Sex chromosomes</td>
<td>X chromosome</td>
</tr>
<tr>
<td>localised homologous pairs</td>
<td></td>
</tr>
<tr>
<td>Organelles</td>
<td>Mitochondrial genome assembled</td>
</tr>
<tr>
<td>complete single alleles</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Raw data accessions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PacificBiosciences SEQUEL II</td>
<td>ERR8705881</td>
</tr>
<tr>
<td>Hi-C Illumina</td>
<td>ERR8702814</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genome assembly</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly accession</td>
<td>GCA_936439885.1</td>
</tr>
<tr>
<td>Accession of alternate haplotype</td>
<td>GCA_936446695.1</td>
</tr>
<tr>
<td>Span (Mb)</td>
<td>670.7</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>241</td>
</tr>
<tr>
<td>Contig N50 length (Mb)</td>
<td>35.2</td>
</tr>
<tr>
<td>Number of scaffolds</td>
<td>103</td>
</tr>
<tr>
<td>Scaffold N50 length (Mb)</td>
<td>116.6</td>
</tr>
<tr>
<td>Longest scaffold (Mb)</td>
<td>154.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genome annotation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of protein-coding genes</td>
<td>27,893</td>
</tr>
<tr>
<td>Number of gene transcripts</td>
<td>28,730</td>
</tr>
</tbody>
</table>

* 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 v5.3.2. 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/idNowFero1.1/dataset/CAKZFG01/busc.

fragments and concentrate the DNA sample. The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.
Figure 2. Genome assembly of *Nowickia ferox*, idNowFero1.1: metrics. The BlobToolKit Snailplot 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 670,715,075 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 (154,429,712 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (116,629,194 and 104,111,281 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/idNowFero1.1/dataset/CAKZFG01/snail.

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 II (HiFi) instrument. Hi-C data were also generated from head tissue of idNowFero1 using the Arima2 kit and sequenced on the Illumina NovaSeq 6000 instrument.
Genome assembly, curation and evaluation

Assembly was carried out with Hifiasm (Cheng et al., 2021) and haplotypic duplication was identified and removed with purge_dups (Guan et al., 2020). The assembly was then scaffolded with Hi-C data (Rao et al., 2014) using YaHS (Zhou et al., 2023). The assembly was checked for contamination and corrected as described previously (Howe et al., 2021). Manual curation was performed using HiGlass (Kerpedjiev et al., 2018) and Pretext (Harry, 2022). The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva et al., 2022), which runs MitoFinder (Allio et al., 2020) or MITOS (Bernt et al., 2013) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.
Figure 4. Genome assembly of *Nowickia ferox*, idNowFero1.1: 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/idNowFero1.1/dataset/CAKZFG01/cumulative.

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 using Nextflow (Di Tommaso et al., 2017) DSL2 pipelines “sanger-tol/readmapping” (Surana et al., 2023a) and “sanger-tol/genomenote” (Surana et al., 2023b). The genome
Figure 5. Genome assembly of *Nowickia ferox*, idNowFero1.1: Hi-C contact map of the idNowFero1.1 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=T-57VXMWSXyPyNtsheUXrQ.

### Table 2. Chromosomal pseudomolecules in the genome assembly of *Nowickia ferox*, idNowFero1.

<table>
<thead>
<tr>
<th>INSDC accession</th>
<th>Chromosome</th>
<th>Length (Mb)</th>
<th>GC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>OW387023.1</td>
<td>1</td>
<td>154.43</td>
<td>31.0</td>
</tr>
<tr>
<td>OW387024.1</td>
<td>2</td>
<td>124.23</td>
<td>31.5</td>
</tr>
<tr>
<td>OW387025.1</td>
<td>3</td>
<td>116.63</td>
<td>30.5</td>
</tr>
<tr>
<td>OW387026.1</td>
<td>4</td>
<td>114.33</td>
<td>31.0</td>
</tr>
<tr>
<td>OW387027.1</td>
<td>5</td>
<td>104.11</td>
<td>30.5</td>
</tr>
<tr>
<td>OW387028.1</td>
<td>X</td>
<td>15.13</td>
<td>31.5</td>
</tr>
<tr>
<td>OW387029.1</td>
<td>MT</td>
<td>0.02</td>
<td>19.5</td>
</tr>
</tbody>
</table>

was analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021; Simão *et al.*, 2015) were calculated.

Table 3 contains a list of relevant software tool versions and sources.

### Genome annotation

The BRAKER2 pipeline (Brüna *et al.*, 2021) was used in the default protein mode to generate annotation for the *Nowickia ferox* assembly (GCA_936439885.1) in Ensembl Rapid Release.

### 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 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: *Nowickia ferox*. Accession
number PRJEB50974; https://identifiers.org/ena.embl/PRJEB50974. (Wellcome Sanger Institute, 2022)

The genome sequence is released openly for reuse. The
*Nowickia ferox* 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.
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.4789928.

Members of the Darwin Tree of Life Barcoding collective are

Members of the Wellcome Sanger Institute Tree of Life
programme are listed here: https://doi.org/10.5281/zenodo.4783585.

Members of Wellcome Sanger Institute Scientific Operations:
DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.4790455.

Members of the Tree of Life Core Informatics collective are

Members of the Darwin Tree of Life Consortium are listed

### References

Abdennur N, Mirny LA: *Cooler: Scalable storage for Hi-C data and other

automated large-scale extraction of mitogenomic data in target

Belshaw R. *Tachinid Flies. Diptera: Tachinidae*. In: *Handbooks for the*

Reference Source


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


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


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


Reference Source


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


Publisher Full Text


Publisher Full Text

Torschinski HP, Herbing B: Die Raupenfliegen (Diptera: Tachinidae)


506: 1–170.

Reference Source


Publisher Full Text

Wellcome Sanger Institute: The genome sequence of a tachinid fly Nwickyia ferox (Panzer, 1809). European Nucleotide Archive. [dataset], accession number PRJEB50974, 2022.

Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text