DATA NOTE

The genome sequence of the dotted bee-fly, *Bombylius discolor* (Mikan, 1796) [version 1; peer review: 2 approved]

Gavin R. Broad, Natural History Museum 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

Abstract

We present a genome assembly from an individual female *Bombylius discolor* (the dotted bee-fly; Arthropoda; Insecta; Diptera; Bombyliidae). The genome sequence is 280 megabases in span. Most of the assembly (99.93%) is scaffolded into six chromosomal pseudomolecules, with the X sex chromosome assembled. The mitochondrial genome has also been assembled and is 16.7 kilobases in length. Genome annotation identified 10,411 protein-coding genes.

Keywords

*Bombylius discolor*, dotted bee-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: Broad GR: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing;

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: © 2022 Broad GR 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: Broad GR, Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding collective et al. The genome sequence of the dotted bee-fly, Bombylius discolor (Mikan, 1796) [version 1; peer review: 2 approved] Wellcome Open Research 2022, 7:306 https://doi.org/10.12688/wellcomeopenres.18614.1

First published: 21 Dec 2022, 7:306 https://doi.org/10.12688/wellcomeopenres.18614.1
Species taxonomy
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Asiloidea; Bombyliidae; Bombylius; Bombylius discolor (Mikan, 1796) (NCBI txid:2741128).

Background
The dotted bee-fly, Bombylius discolor, is a charismatic fly of early spring with a mainly southern distribution in England and Wales, up into the Midlands (National Biodiversity Atlas (NBN) Atlas, no date), and it appears to be increasing its range. B. discolor resembles the more common B. major, but is darker, with spotted wings, and females have a distinctive line of fuzzy white spots down the mid-line of the abdomen (Stubbs & Drake, 2014). Excellent resources exist for the identification of B. discolor and other bombyliids, including Stubbs and Drake, 2014, Steven Falk’s flickr pages and a photo ID guide associated with Bee-fly Watch, a recording initiative in Britain under the auspices of the Soldierflies and Allies Recording Scheme.

This species is widespread across southern and central Europe, and into central Asia. Mainly a species of open ground, B. discolor larvae are parasitoids of mining bees of the genus Andrena, particularly A. flavipes and A. cineraria (Ismay, 1999). Eggs are flicked backwards into the entrances of bee nest burrows, although they are not always accurate and frequently oviposit on non-target substrates (Boesi et al., 2009). Bee-flies are obligate flower visitors as the females require pollen to mature their eggs.

The genome of B. discolor 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.

Genome sequence report
The genome was sequenced from an individual female B. discolor (Figure 1) collected from a garden in Tonbridge, Kent, UK. A total of 24-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 47 missing/misjoins and removed six haplotypic duplications, reducing the assembly length by 0.07% and the scaffold number by 50% and increasing the scaffold N50 by 31.1%.

The final assembly has a total length of 280 Mb in 17 sequence scaffolds with a scaffold N50 of 53.2 Mb (Table 1). Most (99.93%) of the assembly sequence was assigned to six chromosomal-level scaffolds, representing five autosomes and the X sex chromosome (Figure 2–Figure 5; Table 2). The assembly has a BUSCO 5.3.2 (Manni et al., 2021) completeness of 94.6% using the diptera_odb10 reference set.

Table 1. Genome data for Bombylius discolor, idBomDisc1.1.

<table>
<thead>
<tr>
<th>Project accession data</th>
<th>Assembly identifier</th>
<th>idBomDisc1.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>Bombylius discolor</td>
<td></td>
</tr>
<tr>
<td>Specimen</td>
<td>idBomDisc1</td>
<td></td>
</tr>
<tr>
<td>NCBI taxonomy ID</td>
<td>2741128</td>
<td></td>
</tr>
<tr>
<td>BioProject</td>
<td>PRJEB50790</td>
<td></td>
</tr>
<tr>
<td>BioSample ID</td>
<td>SAMEA7524252</td>
<td></td>
</tr>
<tr>
<td>Isolate information</td>
<td>Female, whole organism</td>
<td></td>
</tr>
</tbody>
</table>

| Raw data accessions    | PacificBiosciences SEQUEL I | ERR9527499 |
|                       | Hi-C Illumina               | ERR8571692 |
|                       | PolyA RNA-Seq Illumina      | ERR10123673 |

| Genome assembly        | Assembly accession         | GCA_939192795.1 |
|                       | Accession of alternate haplotype | GCA_939192785.1 |
|                       | Span (Mb)                    | 280            |
|                       | Number of contigs            | 81             |
|                       | Contig N50 length (Mb)       | 7.4            |
|                       | Number of scaffolds          | 17             |
|                       | Scaffold N50 length (Mb)     | 53.2           |
|                       | Longest scaffold (Mb)        | 56.7           |
|                       | BUSCO* genome score          | C:95.3%;S:94.6%,D:0.7%,F:1.2%,M:3.4%,n:3,285 |

| Genome annotation      | Number of protein-coding genes | 10,411 |

* 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/CALNDU01/dataset/CALNDU01/busco.
While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

**Genome annotation report**

The idBomDisc1.1 genome was annotated using the Ensembl rapid annotation pipeline (Table 1: https://rapid.ensembl.org/Bombylius_discolor_GCA_939192795.1). The resulting annotation includes 16,067 transcribed mRNAs from 10,411 protein-coding and 886 non-coding genes.

**Methods**

**Sample acquisition and nucleic acid extraction**

A female *B. discolor* (idBomDisc1) (Figure 1) was collected using a hand net from a garden in Tonbridge, Kent, UK (latitude 51.186304, longitude 0.286534) by Gavin Broad (Natural History
(Museum), who also identified the species. The sample was preserved by freezing at –80°C.

DNA was extracted from tissue of idBomDisc1 at the Wellcome Sanger Institute (WSI) Scientific Operations core using the Qiagen MagAttract HMW DNA kit, according to the manufacturer’s instructions. Head tissue was set aside for Hi-C sequencing.

DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute. The idBomDisc1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing.

Figure 3. Genome assembly of *B. discolor*, idBomDisc1.1: GC coverage. BlobToolKit GC-coverage plot. Chromosomes are coloured by phylum. Circles are sized in proportion to chromosome length. Histograms show the distribution of chromosome length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/CALNDU01/dataset/CALNDU01/blob.
Tissue was disrupted using a Nippi Powermasher fitted with a BioMasher pestle. Fragment size analysis of 0.01–0.5 ng of DNA was then performed using an Agilent FemtoPulse. High molecular weight (HMW) DNA was extracted using the Qiagen MagAttract HMW DNA extraction kit. Low molecular weight DNA was removed from a 20 ng aliquot of extracted DNA using 0.8X AMpure XP purification 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 fragments and concentrate the DNA sample. The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer.

Figure 4. Genome assembly of *B. discolor*, idBomDisc1.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all chromosomes. Coloured lines show cumulative lengths of chromosomes assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/CALNDU01/dataset/CALNDU01/cumulative.
Table 2. Chromosomal pseudomolecules in the genome assembly of *B. discolor*, idBomDisc1.

<table>
<thead>
<tr>
<th>INSDC accession</th>
<th>Chromosome</th>
<th>Size (Mb)</th>
<th>GC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>OWS84238.1</td>
<td>1</td>
<td>56.66</td>
<td>25.5</td>
</tr>
<tr>
<td>OWS84239.1</td>
<td>2</td>
<td>56.04</td>
<td>26.1</td>
</tr>
<tr>
<td>OWS84240.1</td>
<td>3</td>
<td>53.17</td>
<td>26.1</td>
</tr>
<tr>
<td>OWS84241.1</td>
<td>4</td>
<td>42.14</td>
<td>24.4</td>
</tr>
<tr>
<td>OWS84242.1</td>
<td>5</td>
<td>38.52</td>
<td>27.5</td>
</tr>
<tr>
<td>OWS84243.1</td>
<td>X</td>
<td>32.98</td>
<td>24.7</td>
</tr>
<tr>
<td>OWS84244.1</td>
<td>MT</td>
<td>0.02</td>
<td>25.1</td>
</tr>
<tr>
<td>-</td>
<td>unplaced</td>
<td>0.32</td>
<td>37</td>
</tr>
</tbody>
</table>

50 μl RNase-free water and its concentration was 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 using Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

Sequencing

Pacific Biosciences HiFi circular consensus libraries were constructed according to the manufacturers’ instructions. Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit. DNA and RNA sequencing were performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi) and Illumina NovaSeq 6000 (RNA-Seq) instruments. Hi-C data were generated from head tissue of idBomDisc1 using the Arima v2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

Genome assembly

The assembly of idBomDisc1.1 is based on 24x PacBio data and Arima2 Hi-C data generated by the Darwin Tree of Life Project (https://www.darwintreeoflife.org/). The assembly process included the following sequence of steps: initial PacBio assembly generation with Hi fasm (Cheng et al., 2021), retained haplotig separation with purge_dups (Guan et al., 2020), and Hi-C based scaffolding with YaHS (Zhou et al., 2022). The mitochondrial genome was assembled using MitoHiFi.
(Uliano-Silva et al., 2021). Finally, the primary assembly was analysed and manually improved using gEVAL (Chow et al., 2016). Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size.

The genome was analysed and BUSCO scores were generated within the BlobToolKit environment (Challis et al., 2020). Table 3 contains a list of all software tool versions used, where appropriate.

Genome annotation
The Ensembl gene annotation system (Aken et al., 2016) was used to generate annotation for the B. discolor assembly (GCA_939192795.1). Annotation was created primarily through alignment of transcriptomic data to the genome, with gap filling via protein-to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 2019).

<table>
<thead>
<tr>
<th>Table 3. Software tools used.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Software tool</strong></td>
</tr>
<tr>
<td>BlobToolKit</td>
</tr>
<tr>
<td>gEVAL</td>
</tr>
<tr>
<td>hifiasm</td>
</tr>
<tr>
<td>MitoHiFi</td>
</tr>
<tr>
<td>purge_dups</td>
</tr>
<tr>
<td>YaHS</td>
</tr>
</tbody>
</table>

Data availability
European Nucleotide Archive: *Bombylus discolor* (dotted bee fly) Accession number PRJEB50790; [https://identifiers.org/ena.embl/PRJEB50790](https://identifiers.org/ena.embl/PRJEB50790). The genome sequence is released openly for reuse. The *Bombylus discolor* 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 Natural History Museum Genome Acquisition Lab are listed here: [https://doi.org/10.5281/zenodo.4790042](https://doi.org/10.5281/zenodo.4790042).

Members of the Darwin Tree of Life Barcoding collective are listed here: [https://doi.org/10.5281/zenodo.4893703](https://doi.org/10.5281/zenodo.4893703).

Members of the Darwin Tree of Life Barcoding collective are listed here: [https://doi.org/10.5281/zenodo.4783585](https://doi.org/10.5281/zenodo.4783585).


Members of Wellcome Sanger Institute Tree of Life programme are listed here: [https://doi.org/10.5281/zenodo.5013541](https://doi.org/10.5281/zenodo.5013541).

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

Members of the Tree of Life Core Informatics collective are listed here: [https://doi.org/10.5281/zenodo.5013541](https://doi.org/10.5281/zenodo.5013541).


References


Open Peer Review

Current Peer Review Status: ✔️ ✔️

Review Version 1

Reviewer Report 26 April 2023

https://doi.org/10.21956/wellcomeopenres.20641.r56075

© 2023 Navarro-Escalante L. 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.

Lucio Navarro-Escalante
Federacion Nacional de Cafeteros de Colombia, Bogotá, Bogota, Colombia

In general it is a well written report with high quality results for the genome assembly.

Some comments and suggestions:
1. Authors should include any scientific reference (e.g. a paper reference) for the statement: "it appears to be increasing its range". Otherwise this should be removed.

2. Authors should also include a reference for the statement: "Bee-flies are obligate flower visitors as the females require pollen to mature their eggs".

3. I suggest to cite Table 1 at the end of statement "Contigs corresponding to the second haplotype have also been deposited", so the readers can find the accession number of it.

4. I suggest to clarify that gEVAL tool is a genome browser in the Methods. Maybe: "manually improved using gEVAL genome browser".

5. If there was something special in the "select set of proteins from UniProt" in the Methods, then it should be better described here.

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:** Insect molecular genetics and genomics, plant-insect molecular interactions, insect microbiome analysis

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 07 February 2023**

https://doi.org/10.21956/wellcomeopenres.20641.r54560

© 2023 Li X. 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.

✅ Xuankun Li
Department of Biological Sciences, University of Memphis, Memphis, TN, USA

The present paper is one of the series of Darwin Tree of Life genome project, with standardized and clear methods and high-quality results.

Some modifications are needed in the Background portion:

1. 'it appears to be increasing its range' is there any reference to cite about this claim? One possibility on the increased known distribution is the increasing use of social media. If there isn't any scientific study on the expanding range of this species, please remove this claim.

2. "Excellent resources exist for the identification of B. discolor and other bombyliids" change to "Excellent resources exist for the identification of B. discolor and other British bombyliids", remember there are ~5000 species of Bombyliidae found worldwide. Cited references did not cover the fauna out side Britain.

3. "Bee-flies are obligate flower visitors as the females require pollen to mature their eggs." change to "Bee-flies are obligate flower visitors, females require pollen to mature their eggs, males visiting flowers as a resource based encounter site but also feed on nectar and pollen".

**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: bee fly systematics, HiFi sequencing and genome analysis

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.