DATA NOTE

The genome sequence of the yellow loosestrife bee, Macropis europaea Warncke, 1973 [version 1; peer review: 1 approved with reservations]

Steven Falk¹, 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, Joseph Monks², Darwin Tree of Life Consortium

¹Independent Researcher, Kenilworth, Warwickshire, UK
²Department of Life Sciences, Natural History Museum, London, UK

Abstract
We present a genome assembly from an individual male Macropis europaea (the yellow loosestrife bee; Arthropoda; Insecta; Hymenoptera; Melittidae). The genome sequence is 547 megabases in span. The majority of the assembly (61.81%) is scaffolded into 11 chromosomal pseudomolecules. There is an unusually large proportion of satellite repeat, which could not be placed in the assembly. The mitochondrial genome was also assembled and is 19.2 kilobases in length.

Keywords
Macropis europaea, yellow loosestrife bee, genome sequence, chromosomal, Hymenoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status

1

version 1
23 Mar 2022

1. Christopher Ellison, Rutgers University, Piscataway, USA

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: Falk S: Investigation, Resources; Monks J: 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: © 2022 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, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective et al. The genome sequence of the yellow loosestrife bee, Macropis europaea Warncke, 1973 [version 1; peer review: 1 approved with reservations] Wellcome Open Research 2022, 7:110 https://doi.org/10.12688/wellcomeopenres.17787.1

First published: 23 Mar 2022, 7:110 https://doi.org/10.12688/wellcomeopenres.17787.1
Species taxonomy
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Hymenoptera; Apocrita; Aculeata; Apoidea; Melittinae; Macropis; *Macropis (Macropis) europaea* Warncke, 1973 (NCBI:txid253715).

Background
*Macropis (Macropis) europaea* Warncke (yellow loosestrife bee) is a medium sized oligolectic species specialised to provision its nest using floral oils and pollen collected almost exclusively from *Lysimachia vulgaris* (yellow loosestrife) (Primulaceae). The species is endemic to Europe. The genus *Macropis* Panzer represents the only oil collecting bees in the Holarctic (Pekkarinen et al., 2003). The oils are mixed with pollen to form the larval food. Additionally, the oils are used to provide a waterproof coating to the nest’s cells. This allows the species to nest in damp soils that prohibit other ground nesting bee species. Both sexes have a distinctive leg morphology. Males are easily identified by their swollen hind femora and tibiae, and yellow face. Females have specialised hairs on the forelegs and feathery hairs on the hind tibia/basitarsus to help carry oils back to the nest site (Michener, 2007). In the UK the species is restricted to wetland sites in southern England and is active from June to September.

Genome sequence report
The genome was sequenced from a single male *M. europaea* (Figure 1) collected from Wytham Woods, Oxfordshire (biological vice-county: Berkshire), UK (latitude 51.767, longitude -1.311). A total of 30-fold coverage in Pacific Biosciences single-molecule long reads and 55-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 17 misjoins, increasing the assembly size by 0.92%, the scaffold number by 4.89 and the scaffold N50 by 9.29%.

The final assembly has a total length of 547 Mb in 2273 sequence scaffolds with a scaffold N50 of 22.7 Mb (Table 1). Of the assembly sequence, 61.81% was assigned to 11 chromosomal-level scaffolds (numbered by sequence length) (Figure 2–Figure 5; Table 2). There is an unusually large proportion of satellite repeat in this assembly, which is unplaceable using the Hi-C map. The assembly has a BUSCO v5.2.2 (Manni et al., 2021) completeness of 97.3% (single 97.1%, duplicated 0.2%) using the hymenoptera_odb10 reference set (n=5991).

### Table 1. Genome data for *Macropis europaea*, iyMacEuro1.2.

| Project accession data |  |
|------------------------|--|
| Assembly identifier     | iyMacEuro1.2 |
| Species                | *Macropis europaea* |
| Specimen               | iyMacEuro1 |
| NCBI taxonomy ID       | NCBI:txid253715 |
| BioProject             | PRJEB46314 |
| BioSample ID           | SAMEA7746440 |
| Isolate information    | Thorax(genome assembly); Hi-C (head) |

| Raw data accessions |  |
|--------------------|---|
| PacificBiosciences SEQUEL II | ERR6939238 |
| 10X Genomics Illumina | ERR6688500-ERR6688503 |
| Hi-C Illumina       | ERR6688504 |

| Genome assembly |  |
|-----------------|--|
| Assembly accession | GCA_916610135.2 |
| Span (Mb)        | 547 |
| Number of contigs | 2580 |
| Contig N50 length (Mb) | 8.4 |
| Number of scaffolds | 2273 |
| Scaffold N50 length (Mb) | 22.7 |
| Longest scaffold (Mb) | 56.4 |
| BUSCO* genome score | C:97.3%;S:97.1%;D:0.2%;F:0.5%;M:2.2%;n:5991 |

*BUSCO scores based on the hymenoptera_odb10 BUSCO set using v5.2.2. 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.genomeshubs.org/view/iyMacEuro1.2/dataset/CAKAJE02/busc.

---

Figure 1. Image of the *Macropis europaea* (iyMacEuro1) specimen taken during preservation and processing.
Figure 2. Genome assembly of *Macropis europaea*, iyMacEuro1.2: 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 546,807,923 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 (56,347,481 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (22,679,440 and 72,696 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 hymenoptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/iyMacEuro1.2/dataset/CAKAJE02/snail.

**Methods**

**Sample acquisition and DNA extraction**

A male *M. europaea* specimen (iyMacEuro1) was collected from Wytham Woods, Oxfordshire (biological vice-county: Berkshire), UK (latitude 51.767, longitude -1.311) by Steven Falk, Independent Researcher, from fenland using a net. The samples were identified by the same individual and snap-frozen on dry ice.

DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute. The iyMacEuro1 sample was weighed and dissected on dry ice. Thorax 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 200-ng aliquot of extracted DNA using 0.8X AMPure XP purification kit prior to 10X Chromium sequencing; a minimum of 50 ng DNA was submitted for 10X sequencing. HMW DNA was sheared into an average fragment size between 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 and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.
Figure 3. Genome assembly of *Macropis europaea*, iyMacEuro1.2. 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/iyMacEuro1.2/dataset/CAKAJE02/blob?plotShape=circle.

Sequencing
Pacific Biosciences HiFi circular consensus and 10X Genomics 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 and Illumina NovaSeq 6000 instruments. Hi-C data were generated from head tissue of iySphMoni1 using the Arima v2.0 kit and sequenced on an Illumina NovaSeq 6000 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). Scaffolding with Hi-C data (Rao et al., 2014) was carried out with SALSA2 (Ghurye et al., 2019). The Hi-C scaffolded assembly was polished with the 10X Genomics Illumina data by aligning to the assembly with longranger align, calling variants with freebayes (Garrison & Marth, 2012). One round of the Illumina polishing was applied. The mitochondrial genome was assembled with MitoHiFi (Uliano-Silva et al., 2021), which performed annotation using MitoFinder (Allio et al., 2020). The assembly was checked for contamination as described previously (Howe et al., 2021). Manual curation (Howe et al., 2021) was performed using HiGlass (Kerpedjiev et al., 2018) and Pretext. The genome was analysed 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
**Figure 4.** Genome assembly of *Macropis europaea*, iyMacEuro1.2: 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/iyMacEuro1.2/dataset/CAKAJE02/cumulative.

**Figure 5.** Genome assembly of *Macropis europaea*, iyMacEuro1.2: Hi-C contact map. Hi-C contact map of the iyMacEuro1.2 assembly, visualised in HiGlass. Chromosomes are shown in size order from left to right and top to bottom. An interactive version of this map is available here.
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

European Nucleotide Archive: Macropis europaea (yellow loosestrife bee). Accession number PRJEB46314; https://identifiers.org/ena.embl/PRJEB46314.

The genome sequence is released openly for reuse. The *M. europaea* 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 and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Macropis europaea*, iyMacEuro1.2.

| INSDC accession | Chromosome | Size (Mb) | GC% |
|-----------------|------------|-----------|-----|
| OU744343.1      | 1          | 56.35     | 37.7|
| OU744344.1      | 2          | 46.77     | 39.0|
| OU744345.1      | 3          | 41.15     | 37.0|
| OU744346.1      | 4          | 35.81     | 39.1|
| OU744347.1      | 5          | 27.82     | 38.1|
| OU744348.1      | 6          | 27.15     | 38.1|
| OU744349.1      | 7          | 25.89     | 39.1|
| OU744350.1      | 8          | 22.68     | 39.1|
| OU744351.1      | 9          | 20.56     | 38.6|
| OU744352.1      | 10         | 17.70     | 39.0|
| OU744353.1      | 11         | 15.64     | 39.1|
| OU744354.1      | MT         | 0.02      | 14.8|
|                 | Unplaced   | 209.26    | 38.4|

Table 3. Software tools used.

| Software tool   | Version | Source                          |
|-----------------|---------|---------------------------------|
| Hifiasm         | 0.15.1  | Cheng et al., 2021             |
| purge_dups      | 1.2.3   | Guan et al., 2020              |
| SALSA2          | 2.2     | Ghurye et al., 2019            |
| lonranger align | 2.2.2   | https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines |
| freebayes       | v1.3.1-17-gaa2ace8 | Garrison & Marth, 2012 |
| MitoHiFi        | 2       | Uliano-Silva et al., 2021      |
| HiGlass         | 1.11.6  | Kerpedjiev et al., 2018        |
| PretextView     | 0.2.x   | https://github.com/wtsi-hpag/PretextView |
| BlobToolKit     | 3.0.5   | Challis et al., 2020           |

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.

References

Allio R, Schomaker-Bastos A, Romiguier J, et al.: MitoFinder: Efficient Automated Large-Scale Extraction of Mitogenomic Data in Target Enrichment Phylogenomics. Mol Ecol Resour. 2020; 20(4): 892-905. PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Richards E, Rajan J, et al.: BlobToolKit - Interactive Quality Assessment of Genome Assemblies. G3 (Bethesda). 2020; 10(4): 1361-74. 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): 170-75. PubMed Abstract | Publisher Full Text | Free Full Text

Garrison E, Marth G: Haplotype-Based Variant Detection from Short-Read Sequencing. arXiv. 2012. Reference Source
Ghurye J, Rhee A, Walenz BP, et al.: Integrating Hi-C Links with Assembly Graphs for Chromosome-Scale Assembly. PLoS Comput Biol. 2019; 15(8): e1007273. 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): 2896–98. PubMed Abstract | Publisher Full Text | Free Full Text

Howe K, Chow W, Collins J, et al.: Significantly Improving the Quality of Genome Assemblies through Curation. GigaScience. 2021; 10(1): gisaa153. PubMed Abstract | Publisher Full Text | Free 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

Manni M, Berkeley MR, Seppey 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): 4647–54. PubMed Abstract | Publisher Full Text | Free Full Text

Michener CD: The Bees of the World. Johns Hopkins University Press. 2007. Reference Source

Pekkarinen A, Berg Ö, Calabuig J, et al.: Distribution and Co-Existence of the Macropis Species and Their Cleptoparasite Epeoloides Coecutiens (Fabr.) in NW Europe (Hymenoptera: Apoidea, Melittidae and Apidae). Entomol Fenn. 2003; 14(1): 53–59. Publisher Full Text

Rao SS, Huntley MH, Durand NC, et al.: A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014; 158(7): 1665–80. PubMed Abstract | Publisher Full Text | Free Full Text

Uliano-Silva M, Nunes JGF, Krasheninnikova K, et al.: marcelauliano/MitoHiFi: mitohifi_v2.0. 2021. Publisher Full Text
Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 05 April 2022

https://doi.org/10.21956/wellcomeopenres.19687.r49424

© 2022 Ellison 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.

Christopher Ellison
Department of Genetics, Human Genetics Institute of New Jersey, Rutgers University, Piscataway, NJ, USA

Falk et al present a high-quality genome assembly for *M. europaea* using state-of-the-art sequencing technologies and computational approaches. The report is detailed and easy to follow. Including the information below would be useful to other readers:

1. There is no clear rationale for generating this dataset other than the mention of the Tree of Life Consortium. Do other genome assemblies exist for this species? Does this species have unique qualities or phylogenetic placement that led to its selection by the consortium?

2. Is the number of chromosomes known from previous work? Or from other close relatives? If so, please include this information.

3. It is not clear why the 10x genomics read clouds were generated. Were these used for genome assembly/scaffolding or only polishing?

4. What is the abundant satellite repeat? Is it a single repeat or many different satellites?

5. I assume the male is haploid? If so, why would there be haplotypic duplication?

6. The Hi-C contact map looks unusual due to the pattern that looks like a large plus sign "+". Could this be due to shorter contigs containing large amounts of the satellite repeat? If the authors are aware of what is creating the "+" pattern, it would be useful to include it here.

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?  
Yes

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

**Reviewer Expertise:** Evolutionary 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, however I have significant reservations, as outlined above.