DATA NOTE

The genome sequence of the black clock beetle, *Pterostichus madidus* (Fabricius, 1775) [version 1; peer review: awaiting peer review]

Liam Crowley1, 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

1Department of Zoology, University of Oxford, Oxford, UK

---

**Abstract**

We present a genome assembly from an individual female *Pterostichus madidus* (the black clock beetle; Arthropoda; Insecta; Coleoptera; Carabidae). The genome sequence is 705 megabases in span. The majority (99.96%) of the assembly is scaffolded into 19 chromosomal pseudomolecules, with the X sex chromosome assembled.

**Keywords**

Pterostichus madidus, black clock beetle, genome sequence, chromosomal

---

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

Author roles: Crowley L: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Wellcome Trust 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: © 2021 Crowley 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: Crowley L, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective et al. The genome sequence of the black clock beetle, *Pterostichus madidus* (Fabricius, 1775) [version 1; peer review: awaiting peer review] Wellcome Open Research 2021, 6:301 https://doi.org/10.12688/wellcomeopenres.17347.1

First published: 08 Nov 2021, 6:301 https://doi.org/10.12688/wellcomeopenres.17347.1
Species taxonomy
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Coleoptera; Adephaga; Caraboidea; Carabidae; Harpalinae; Pterostichini; Pterostichus; Steropus; *Pterostichus madidus* (Fabricius, 1775) (NCBI:txid767470).

Background
The black clock beetle, *Pterostichus madidus*, is a large, common species of ground beetle. It occurs across western and northern Europe and in the UK it is the most frequently recorded beetle in the family Carabidae. It can be found throughout a wide range of habitats where it is active during both the night and day. It is a relatively large (13-18 mm), black carabid with smoothly rounded pronotal hind angles. There are two subspecies, *Pterostichus madidus validus* Dejean, 1828, which has black femora, and *Pterostichus madidus concinnus* (Sturm, 1818), which has distinctive ‘wine red’ femora. *Pterostichus madidus* is omnivorous, being a predator and scavenger, but also feeding on plant material (Luff, 1974). It is predominantly an annual species, laying eggs in late summer/autumn and larvae developing over the winter (Luff & Others, 1973). Overwintered adults are active from spring/early summer and some adults, particularly at higher altitudes, are biennial (Butterfield, 1996).

Genome sequence report
The genome was sequenced from one female *P. madidus* collected from Wytham Woods, Oxfordshire (biological vice-county: Berkshire), UK (latitude 51.775, longitude -1.326) (Figure 1). A total of 34-fold coverage in Pacific Biosciences single-molecule long reads and 53-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 142 missing/misjoins and removed 6 haplotypic duplications, reducing the assembly length by 0.18% and the scaffold number by 80.00%, and increasing the scaffold N50 by 58.29%.

The final assembly has a total length of 705 Mb in 27 sequence scaffolds with a scaffold N50 of 37.9 Mb (Table 1). The majority, 99.96%, of the assembly sequence was assigned to 19 chromosomal-level scaffolds, representing 18 autosomes (numbered by sequence length), and the X sex chromosome (Figure 2–Figure 5; Table 2). Some regions of the genome have large repeats with less certain structure than the rest of the assembly, most notably chromosomes 14, 15 and 18. Chromosome 14 from 23.8 Mb onwards has strong Hi-C association with chromosome 18. The assembly has a BUSCO v5.1.2 (Manni *et al*., 2021) completeness of 98.9% (single 98.4%, duplicated 0.5%) using the endopterygota_odb10 reference set. While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

![Figure 1. An image of the sequenced specimen, icPteMadi1, captured immediately prior to processing and preservation.](image)
Methods
Sample acquisition, DNA extraction and sequencing
A single female *P. madidus* was collected from Wytham Woods, Oxfordshire (biological vice-county: Berkshire), UK (latitude 51.775, longitude -1.326) by Liam Crowley, University of Oxford, using a pooter. The sample was identified by the same individual, snap-frozen on dry ice and stored using a CoolRack.

DNA was extracted from the head/thorax tissue of *P. madidus* (icPteMadi1) at the Wellcome Sanger Institute (WSI) Scientific Operations core from the whole organism using the Qiagen...
MagAttract HMW DNA kit, according to the manufacturer’s instructions. Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud DNA sequencing libraries were constructed according to the manufacturers’ instructions. Sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II and Illumina HiSeq X instruments. Hi-C data were generated from abdomen tissue of icPteMadi1 using the Arima v2 Hi-C 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). 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 the gEVAL system (Chow et al., 2016) as described previously (Howe et al., 2021). Manual curation (Howe et al., 2021) was performed using gEVAL, HiGlass (Kerpedjiev et al., 2018) and Pretext. The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva et al., 2021). 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.

**Figure 4.** Genome assembly of *Pterostichus madidus*, icPteMadi1.1: 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/icPteMadi1.1/dataset/CAJY1R01/cumulative](https://blobtoolkit.genomehubs.org/view/icPteMadi1.1/dataset/CAJY1R01/cumulative).
**Table 2.** Chromosomal pseudomolecules in the genome assembly of *Pterostichus madidus*, icPteMadi1.1.

<table>
<thead>
<tr>
<th>INSDC accession</th>
<th>Chromosome</th>
<th>Size (Mb)</th>
<th>GC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>OU452301.1</td>
<td>1</td>
<td>48.00</td>
<td>31.4</td>
</tr>
<tr>
<td>OU452302.1</td>
<td>2</td>
<td>46.20</td>
<td>31.7</td>
</tr>
<tr>
<td>OU452303.1</td>
<td>3</td>
<td>45.37</td>
<td>31.4</td>
</tr>
<tr>
<td>OU452304.1</td>
<td>4</td>
<td>44.84</td>
<td>32.1</td>
</tr>
<tr>
<td>OU452305.1</td>
<td>5</td>
<td>42.01</td>
<td>31.7</td>
</tr>
<tr>
<td>OU452306.1</td>
<td>6</td>
<td>41.31</td>
<td>31.9</td>
</tr>
<tr>
<td>OU452307.1</td>
<td>7</td>
<td>40.83</td>
<td>31.7</td>
</tr>
<tr>
<td>OU452308.1</td>
<td>8</td>
<td>38.74</td>
<td>31.8</td>
</tr>
<tr>
<td>OU452309.1</td>
<td>9</td>
<td>37.88</td>
<td>31.8</td>
</tr>
</tbody>
</table>

**Figure 5.** Genome assembly of *Pterostichus madidus*, icPteMadi1.1: Hi-C contact map. Hi-C contact map of the icPteMadi1.1 assembly, visualised in HiGlass.
Table 3. Software tools used.

<table>
<thead>
<tr>
<th>Software tool</th>
<th>Version</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hifiasm</td>
<td>0.14-r312</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>2.11.3</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>0.2.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>2.6.2</td>
<td>Challis et al., 2020</td>
</tr>
</tbody>
</table>

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 P. madidus 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.

Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.4789929.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.4893704.

Members of the Wellcome Sanger Institute Tree of Life programme collective are listed here: https://doi.org/10.5281/zenodo.5377053.


Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.5013542.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783559.

