DATA NOTE

The genome sequence of the black clock beetle, *Pterostichus madidus* (Fabricius, 1775) [version 1; peer review: 2 approved]

Liam M. Crowley[1],
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

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

v1 **First published:** 08 Nov 2021, **6**:301
https://doi.org/10.12688/wellcomeopenres.17347.1
Latest published: 08 Nov 2021, **6**:301
https://doi.org/10.12688/wellcomeopenres.17347.1

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.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 08 Nov 2021	✓ view	✓ view

1. **James Pflug**, Oregon State University, Corvallis, USA

2. **Lukas Zangl**, University of Graz, Graz, Austria

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: Crowley LM: 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.

How to cite this article: Crowley LM, 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: 2 approved]** 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.

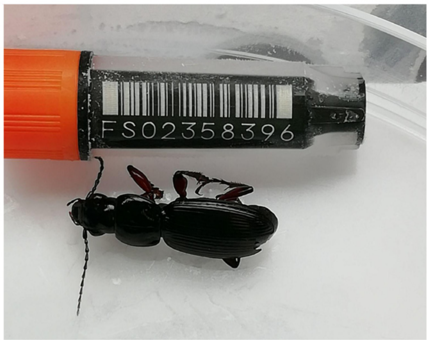

Figure 1. An image of the sequenced specimen, icPteMadi1, captured immediately prior to processing and preservation.

Table 1. Genome data for *Pterostichus madidus*, icPteMadi1.1.

Project accession data	
Assembly identifier	icPteMadi1.1
Species	*Pterostichus madidus*
Specimen	icPteMadi1
NCBI taxonomy ID	NCBI:txid767470
BioProject	PRJEB45192
BioSample ID	SAMEA7520318
Isolate information	Female, head/thorax (genome assembly), abdomen (Hi-C)
Raw data accessions	
PacificBiosciences SEQUEL II	ERR6606793
10X Genomics Illumina	ERR6054945-ERR6054948
Hi-C Illumina	ERR6054949
Genome assembly	
Assembly accession	GCA_911728475.1
Accession of alternate haplotype	GCA_911728425.1
Span (Mb)	705
Number of contigs	184
Contig N50 length (Mb)	15.8
Number of scaffolds	27
Scaffold N50 length (Mb)	37.9
Longest scaffold (Mb)	48.0
BUSCO* genome score	C:98.9%[S:98.4%,D:0.5%],F:0.6%,M:0.6%,n:2124

*BUSCO scores based on the endopterygota_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.genomehubs.org/view/icPteMadi1.1/dataset/CAJVRY01/busco.

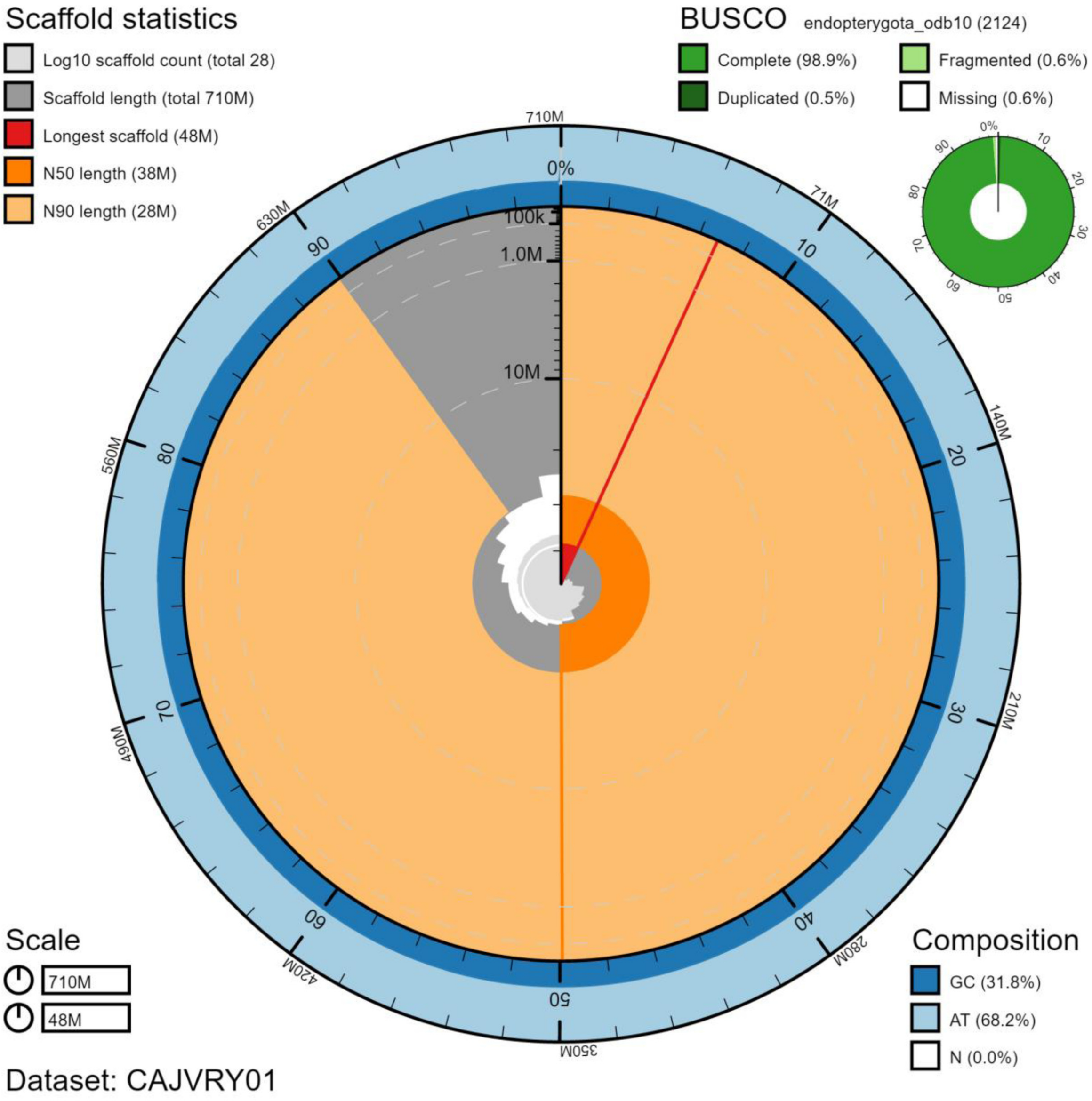

Figure 2. Genome assembly of *Pterostichus madidus*, icPteMadi1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 705,160,476 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (47,997,105 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (37,879,541 and 28,091,952 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 endopterygota_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/icPteMadi1.1/dataset/CAJVRY01/snail.

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

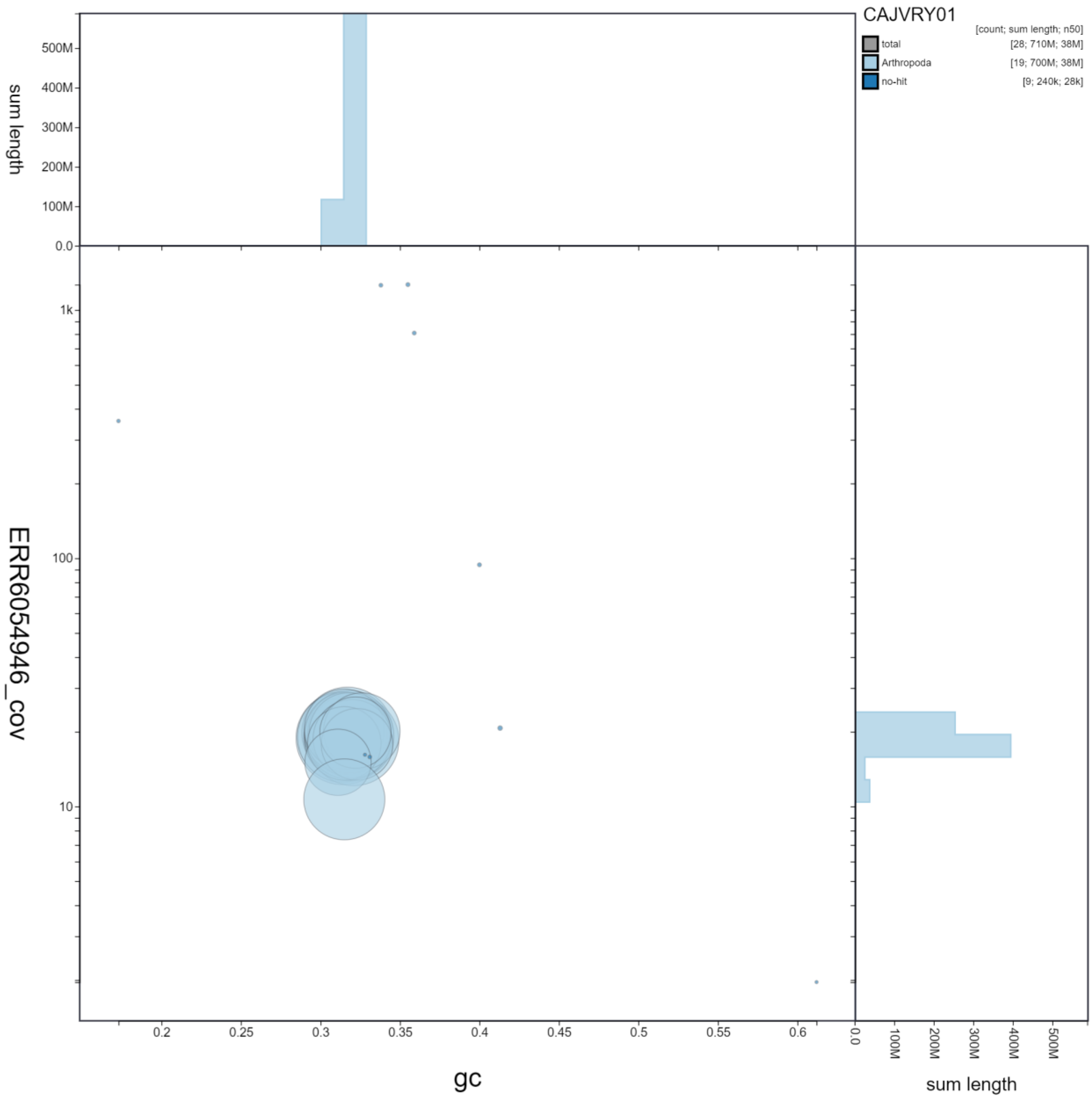

Figure 3. Genome assembly of *Pterostichus madidus*, icPteMadi1.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/icPteMadi1.1/dataset/CAJVRY01/blob.

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

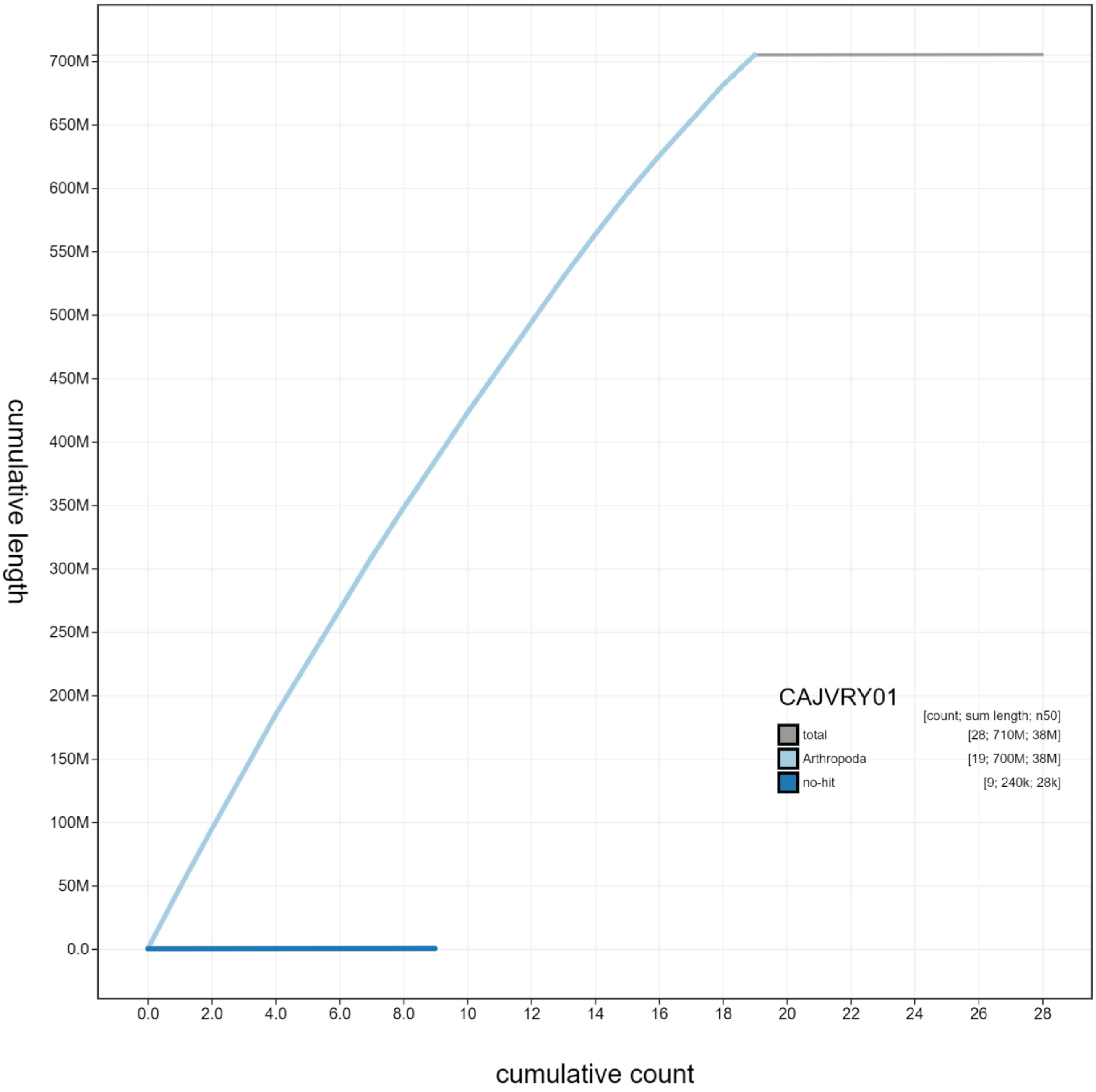

Figure 4. Genome assembly of *Pterostichus madidus*, icPteMadi1.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/icPteMadi1.1/dataset/CAJVRY01/cumulative.

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.

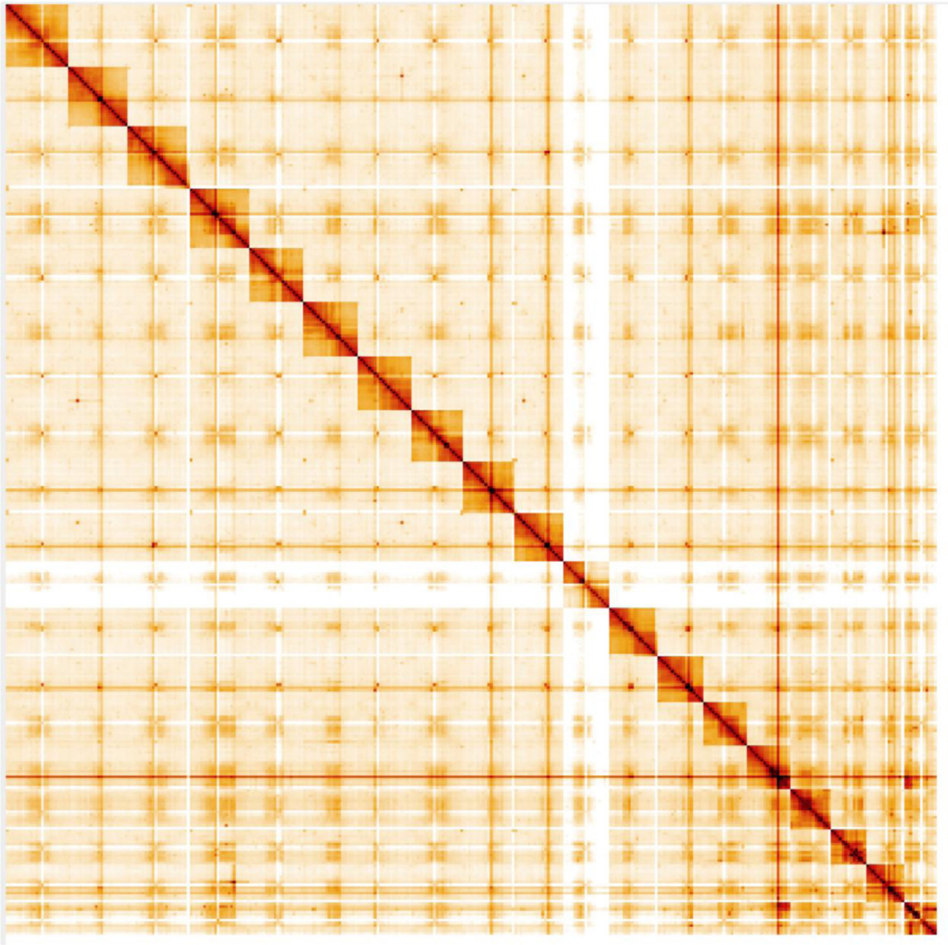

Figure 5. Genome assembly of *Pterostichus madidus*, icPteMadi1.1: Hi-C contact map. Hi-C contact map of the icPteMadi1.1 assembly, visualised in HiGlass.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Pterostichus madidus*, icPteMadi1.1.

INSDC accession	Chromosome	Size (Mb)	GC%
OU452301.1	1	48.00	31.4
OU452302.1	2	46.20	31.7
OU452303.1	3	45.37	31.4
OU452304.1	4	44.84	32.1
OU452305.1	5	42.01	31.7
OU452306.1	6	41.31	31.9
OU452307.1	7	40.83	31.7
OU452308.1	8	38.74	31.8
OU452309.1	9	37.88	31.8
OU452310.1	10	37.21	31.6
OU452312.1	11	35.34	31.8
OU452313.1	12	35.33	31.5
OU452314.1	13	33.82	32
OU452315.1	14	32.21	32.6
OU452316.1	15	29.68	31.5
OU452317.1	16	28.09	32.3
OU452318.1	17	27.89	32.2
OU452319.1	18	23.99	31.1
OU452311.1	X	36.19	31.5
OU452320.1	MT	0.02	17.3
-	Unplaced	0.22	38.4

Table 3. **Software tools used.**

Software tool	Version	Source
Hifiasm	0.14-r312	Cheng *et al.*, 2021
purge_dups	1.2.3	Guan *et al.*, 2020
SALSA2	2.2	Ghurye *et al.*, 2019
longranger align	2.2.2	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines
freebayes	1.3.1-17-gaa2ace8	Garrison & Marth, 2012
MitoHiFi	2.11.3	Uliano-Silva *et al.*, 2021
gEVAL	N/A	Chow *et al.*, 2016
HiGlass	1.11.6	Kerpedjiev *et al.*, 2018
PretextView	0.2.x	https://github.com/wtsi-hpag/PretextView
BlobToolKit	2.6.2	Challis *et al.*, 2020

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

European Nucleotide Archive: Pterostichus madidus (black clock beetle). Accession number PRJEB45192; https://identifiers.org/ena.embl/PRJEB45192.

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 Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.4790456.

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.

References

Butterfield J: **Carabid Life-Cycle Strategies and Climate Change: A Study on an Altitude Transect.** *Ecol Entomol.* 1996; **21**(1): 9–16.
Publisher 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

Chow W, Brugger K, Caccamo M, *et al.*: **gEVAL - a web-based browser for evaluating genome assemblies.** *Bioinformatics.* 2016; **32**(16): 2508–10.
PubMed Abstract | Publisher Full Text | Free Full Text

Garrison E, Marth G: **Haplotype-Based Variant Detection from Short-Read Sequencing.** arXiv: 1207.3907. 2012.
Reference Source

Ghurye J, Rhie 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): giaa153.
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

Luff ML: **Adult and Larval Feeding Habits of *Pterostichus Madidus* (F.) (Coleoptera: Carabidae).** *J Nat Hist.* 1974; **8**(4): 403–9.
Publisher Full Text

Luff ML, Others: **The Annual Activity Pattern and Life Cycle of Pterostichus

Madidus (F.) (Col. Carabidae). *Entomologica Scandinavica. Supplement.* 1973; **4**(4): 259–73.

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**

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; **159**(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 25 July 2022

https://doi.org/10.21956/wellcomeopenres.19177.r51468

✓ **Lukas Zangl**
Institute of Biology, University of Graz, Graz, Austria

In this study, the authors provide a full length chromosome-level quality genome of the ground beetle Pterostichus madidus. The genome was composed of PacBio long-reads as well as 10X Genomics Illumina short-read data. High BUSCO scores suggest high levels of coding gene completeness and homogenous GC values as well as low numbers of non-traget scaffolds indicate low levels of contamination.

All raw data as well as assembly and annotation are publicly available and the methods and tools described will allow for good reproducibility. My only remark in this context regards the manual edits, which are mentioned, but not described in further detail. Maybe there are some comments on rationale and procedure that would be appropriate/needed to insure full reproducibility.

Furthermore, information about the two subspecies is given, however the authors missed to provide a comment on which subspecies they have actually sequenced.

Finally, despite the apparent good quality of the assembled and annotated genome, I had a hard time interpreting the Snailplot (specifically the grey and orange parts), however, I have to admit that I am not too familiar with his kind of representation. Therefore, I would suggest adding a few comments in the main text about what can be seen in order to increase clarification for all readers in general.

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: zoology, biodiversity research, genetics (inlcuding some 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.

Reviewer Report 20 December 2021

https://doi.org/10.21956/wellcomeopenres.19177.r46962

✓

James Pflug
Department of Integrative Biology, Oregon State University, Corvallis, OR, USA

Summary:
This paper reports on the successful sequencing of the genome of the ground beetle *Pterostichus madidus*. The authors employed PacBio and 10X Genomics sequencing to produce a 705 Mb assembly. The authors also performed Hi-C experiments, with the data suggesting that the genome is divided into 19 chromosomal pseudomolecules. The assembly contained matches to the vast majority of BUSCO endopterygota_odb10 reference genes, suggesting that the coding gene content of the assembly is relatively complete.

The paper provides a clear justification for the purpose of the study, as well as the methods and means used to obtain the data. The resulting genome assembly appears to be relatively high quality based on the analyses performed. All relevant data have been deposited in the appropriate locations and are readily accessible. Overall, the experiments conducted are sound, and the manuscript is well written and appropriate for the Data Note format.

General Comments:

- Has *P. madidus*'s genome size (or that of a closely related species) been established via flow cytography or other fluorometric methods? If so, how does it compare to the 705 Mb assembly?

- Figure 2 (i.e., the BlobToolKit Snailplot) is somewhat difficult for me to understand at first look. In particular, the orange and grey histogram section in the center of the large circle. I'm aware this style of plot is fairly common, and others who are more familiar with snailplots will not have trouble interpreting them, but readers like me would appreciate some additional clarification of what's going on.

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: Genomics and Systematics of the beetle family Carabidae

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.