DATA NOTE

The genome sequence of the Yellow-line Quaker, *Agrochola macilenta* (Hubner, 1809) [version 1; peer review: awaiting peer review]

Douglas Boyes[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, Clare Boyes[2],
Darwin Tree of Life Consortium

[1]UK Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK
[2]Independent Researcher, Wytham, Oxfordshire, UK

[+] Deceased author

v1 **First published:** 20 Feb 2023, **8**:90
https://doi.org/10.12688/wellcomeopenres.18912.1
Latest published: 20 Feb 2023, **8**:90
https://doi.org/10.12688/wellcomeopenres.18912.1

Abstract

We present a genome assembly from an individual female *Agrochola macilenta* (the Yellow-line Quaker; Arthropoda; Insecta; Lepidoptera; Noctuidae). The genome sequence is 683 megabases in span. Most of the assembly is scaffolded into 32 chromosomal pseudomolecules, including the Z and W sex chromosomes. The mitochondrial genome has also been assembled and is 15.4 kilobases in length. Gene annotation of this assembly on Ensembl identified 18,769 protein coding genes.

Keywords

Agrochola macilenta, Yellow-line Quaker, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

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: Boyes D: Investigation, Resources; **Boyes 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, 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.

How to cite this article: Boyes D, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective *et al.* **The genome sequence of the Yellow-line Quaker, *Agrochola macilenta* (Hubner, 1809) [version 1; peer review: awaiting peer review]** Wellcome Open Research 2023, **8**:90 https://doi.org/10.12688/wellcomeopenres.18912.1

First published: 20 Feb 2023, **8**:90 https://doi.org/10.12688/wellcomeopenres.18912.1

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Noctuoidea; Noctuidae; Xyleninae; *Agrochola*; *Agrochola macilenta* (Hubner, 1809) (NCBI:txid987872).

Background

Agrochola macilenta (the Yellow-line Quaker) is a common and widespread noctuid moth found throughout Europe and east as far as Turkey. Its distribution and abundance in Great Britain and Ireland have increased significantly over the last 50 years (Randle *et al.*, 2019). *A. macilenta* was highlighted as one of the Anthropocene 'winners' as bucking the trend of general moth declines (Boyes *et al.*, 2019).

The common name of this moth is a reference to the unornamented and drab clothing traditionally worn by Quakers (Marren, 2019). The yellow line appears on its plain wings; the forewing length is 14–16 mm. In Britain, *A. macilenta* has one generation a year and the adult flies between September and December. It regularly comes to light, is attracted to sugar, and also feeds on over-ripe fruit (Waring *et al.*, 2017).

The moth is found in a variety of habitats including woodland and scrub. In the north of its range, it can be found on moorland, where the caterpillar feeds on *Calluna* spp. Elsewhere, the female lays her eggs on tree bark where they overwinter. It favours a number of species including *Quercus* spp., *Fagus sylvatica*, *Populus* spp., *Salix* spp., and *Crategus monogyna*. The eggs hatch in early spring and the larvae initially feed in spinnings on the terminal shoots of the tree, before feeding at night on the new leaves. Prior to pupation, the larvae descend from the tree to feed on forbs at ground level. By June, the larvae are fully fed, and they aestivate for a few weeks before pupating underground. It is also thought to have cannibalistic tendencies (Heath & Emmet, 1983).

The genome of *A. macilenta* 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 *A. macilenta* based on one female specimen from Wytham Woods, Oxfordshire, UK.

Genome sequence report

The genome was sequenced from one female *A. macilenta* (Figure 1) collected from Wytham Woods, UK (latitude 51.77, longitude –1.34). A total of 29-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 117-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 24 missing joins or mis-joins and removed four haplotypic duplications, reducing the scaffold number by 31.37%.

The final assembly has a total length of 683.2 Mb in 35 sequence scaffolds with a scaffold N50 of 23.0 Mb (Table 1). Most (99.98%)of the assembly sequence was assigned to 32 chromosomal-level scaffolds, representing 30 autosomes and the Z and W sex chromosomes. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 2–Figure 5; Table 2). The assembly has a BUSCO v5.3.2 (Manni *et al.*, 2021) completeness of 99.0% (single 98.2%, duplicated 0.7%) using the lepidoptera_odb10 reference set (n = 5,286). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

Figure 1. Photograph of the *Agrochola macilenta* (ilAgrMaci1) specimen used for genome sequencing.

Genome annotation report

The *A. macilenta* genome assembly GCA_916701695.1 was annotated using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Agrochola_macilenta_GCA_916701695.1/). The resulting annotation includes 18,959 transcribed mRNAs from 18,769 protein-coding genes.

Methods

Sample acquisition and nucleic acid extraction

Two *A. macilenta* specimens (ilAgrMaci1 and ilAgrMaci2) were collected in Wytham Woods, Oxfordshire (biological vice-county: Berkshire), UK (latitude 51.77, longitude –1.34) on 8 October 2020 using a light trap. The specimens were collected and identified by Douglas Boyes (University of Oxford) and snap-frozen on dry ice.

DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute (WSI). The ilAgrMaci1 sample was weighed and dissected on dry ice with head 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. Low molecular weight DNA was removed from a 20 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

Table 1. Genome data for *Agrochola macilenta*, ilAgrMaci1.1.

Project accession data		
Assembly identifier	ilAgrMaci1.1	
Species	*Agrochola macilenta*	
Specimen	ilAgrMaci1	
NCBI taxonomy ID	987872	
BioProject	PRJEB46326	
BioSample ID	SAMEA8603207	
Isolate information	ilAgrMaci1; head (Hi-C) ilAgrMaci2: abdomen (RNA-Seq)	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	58.4	*≥ 50*
k-mer completeness	100%	*≥ 95%*
BUSCO**	C:99.0%[S:98.2%,D:0.7%], F:0.3%,M:0.8%,n:5,286	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	99.98%	*≥ 95%*
Sex chromosomes	W and Z chromosomes	*localised homologous pairs*
Organelles	Mitochondrial genome assembled	*complete single alleles*
Raw data accessions		
PacificBiosciences SEQUEL II	ERR6808008	
10X Genomics Illumina	ERR6688561–ERR6688564	
Hi-C Illumina	ERR6688560	
PolyA RNA-Seq Illumina	ERR9435010	
Genome assembly		
Assembly accession	GCA_916701695.1	
Accession of alternate haplotype	GCA_916701705.1	
Span (Mb)	683.2	
Number of contigs	65	
Contig N50 length (Mb)	22.3	
Number of scaffolds	35	
Scaffold N50 length (Mb)	23.0	
Longest scaffold (Mb)	29.4	
Genome annotation		
Number of protein-coding genes	18,769	
Number of gene transcripts	18,959	

* 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 lepidoptera_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/ilAgrMaci1.1/dataset/CAKAOO01.1/busco.

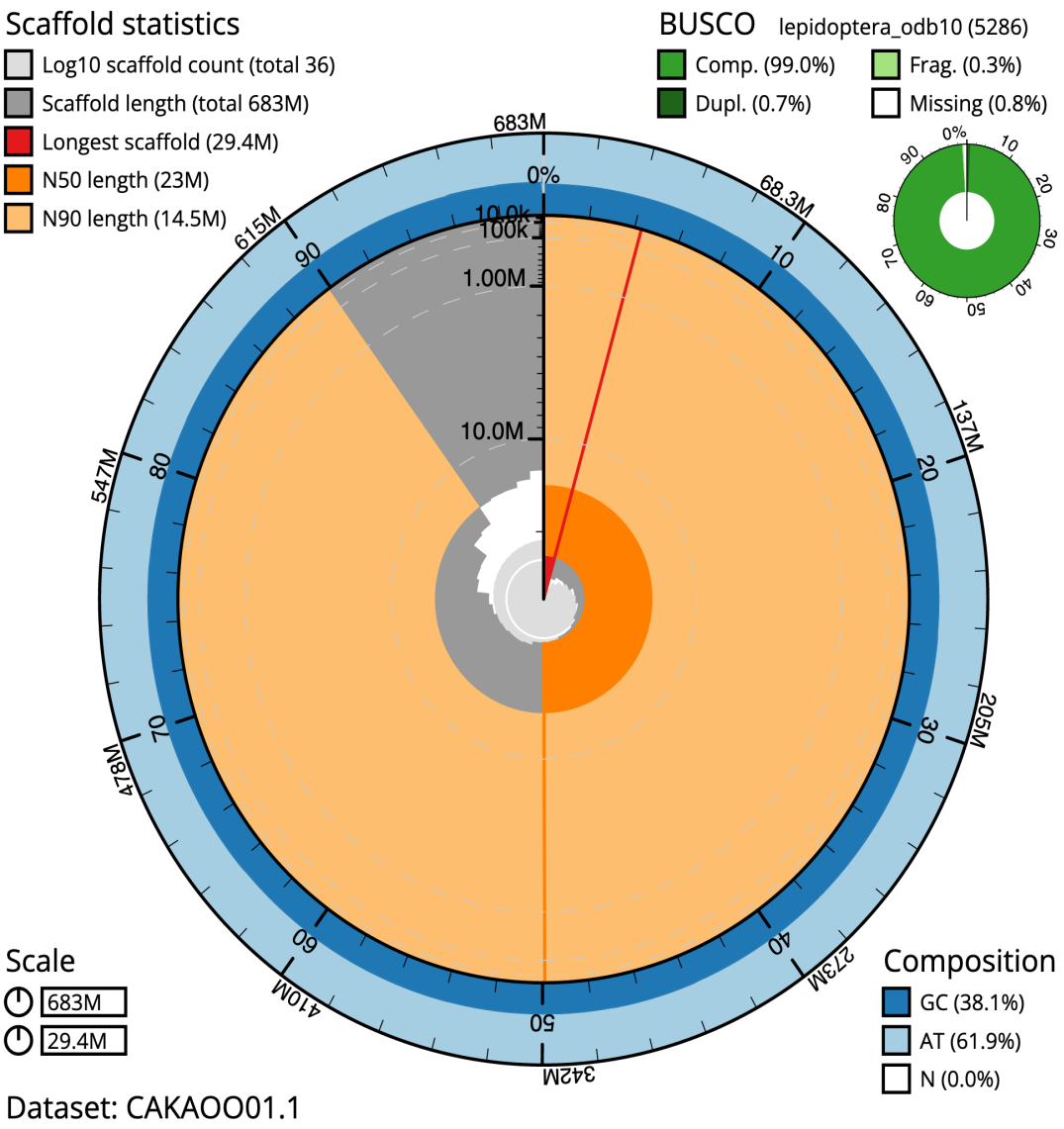

Figure 2. Genome assembly of *Agrochola macilenta,* ilAgrMaci1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 683,223,563 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (29,426,042 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (22,977,954 and 14,528,028 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 lepidoptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilAgrMaci1.1/dataset/CAKAOO01.1/snail.

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 and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

RNA was extracted from abdomen tissue of ilAgrMaci2 in the Tree of Life Laboratory at the WSI using TRIzol, according to the manufacturer's instructions. RNA was then eluted in

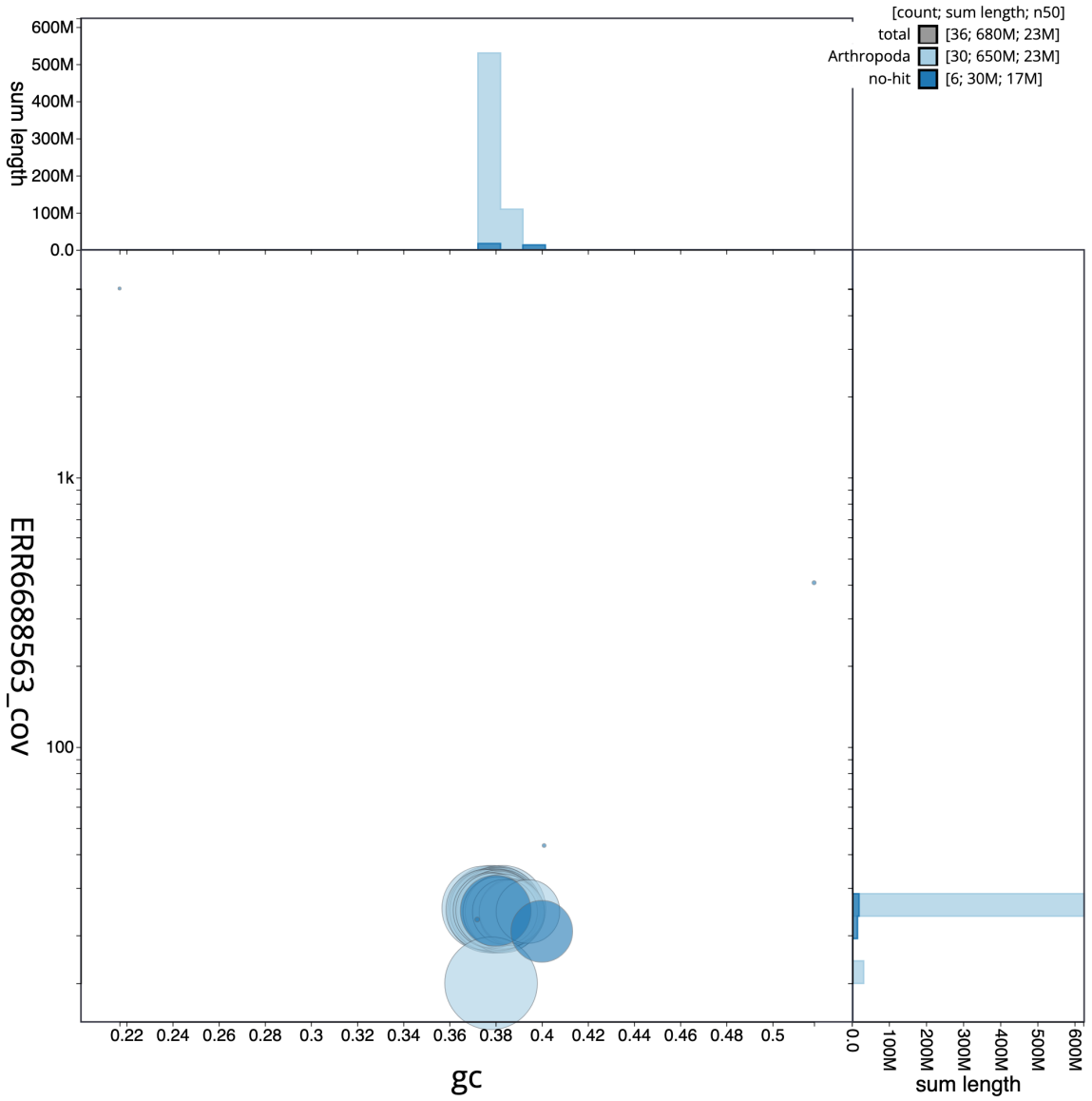

Figure 3. Genome assembly of *Agrochola macilenta*, ilAgrMaci1.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilAgrMaci1.1/dataset/CAKAOO01.1/blob.

50 µl RNAse-free water and its concentration 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 and 10X Genomics read cloud DNA sequencing 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 was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi), Illumina HiSeq 4000 (RNA-Seq) and Illumina NovaSeq 6000 (10X) instruments. Hi-C data were also generated from head tissue of ilAgrMaci1D using the Arima v2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

Genome assembly

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021) and haplotypic duplication was identified and removed

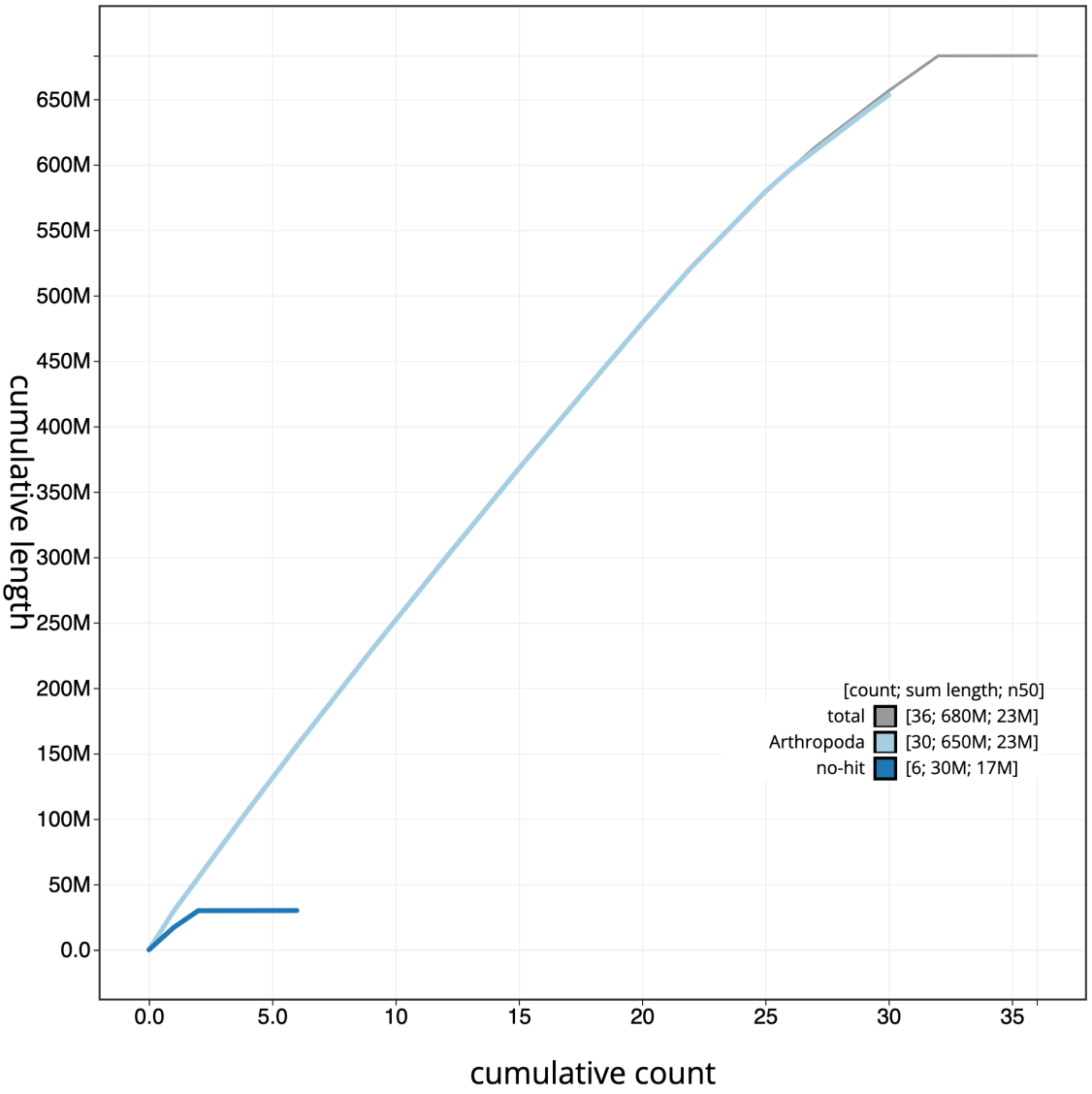

Figure 4. Genome assembly of *Agrochola macilenta*, ilAgrMaci1.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilAgrMaci1.1/dataset/CAKAOO01.1/cumulative.

with purge_dups (Guan *et al.*, 2020). One round of polishing was performed by aligning 10X Genomics read data to the assembly with Long Ranger 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 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 performed annotation using MitoFinder (Allio *et al.*, 2020). 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 BRAKER2 pipeline (Brůna *et al.*, 2021) was used in the default protein mode to generate annotation for the *Agrochola macilenta* assembly (GCA_916701695.1) in Ensembl Rapid Release.

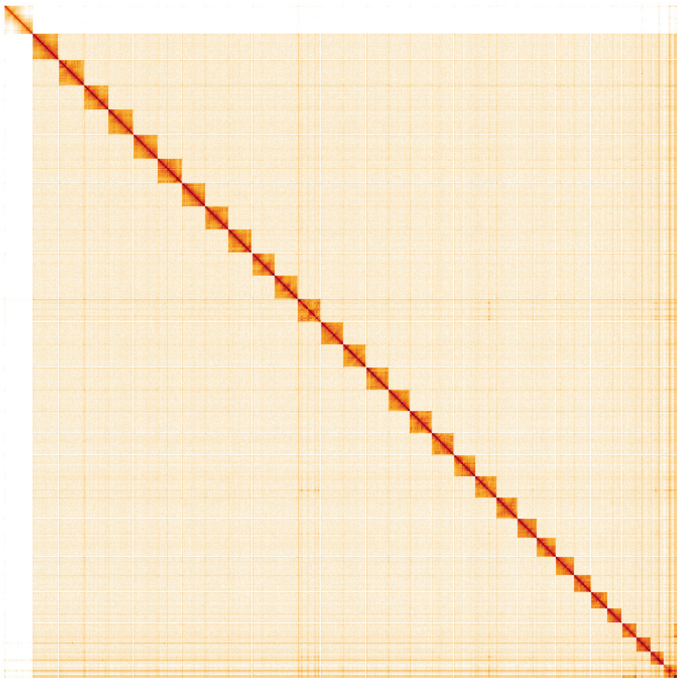

Figure 5. Genome assembly of *Agrochola macilenta*, ilAgrMaci1.1: Hi-C contact map. Hi-C contact map of the ilAgrMaci1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=I8ttdgVjRtazOuazFgsjgg.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Agrochola macilenta*, ilAgrMaci1.

INSDC accession	Chromosome	Size (Mb)	GC%
OU753553.1	1	26.03	37.9
OU753554.1	2	25.53	38.3
OU753555.1	3	25.39	37.8
OU753556.1	4	25.15	38
OU753557.1	5	24.51	38.1
OU753558.1	6	24.37	37.5
OU753559.1	7	23.96	38.1
OU753560.1	8	23.96	37.8
OU753561.1	9	23.61	38
OU753562.1	10	23.4	38
OU753563.1	11	23.35	37.7
OU753564.1	12	23.11	38.2
OU753565.1	13	22.98	37.9
OU753566.1	14	22.97	37.8
OU753567.1	15	22.49	38
OU753568.1	16	22.35	38.1
OU753569.1	17	22.32	37.6
OU753570.1	18	22.1	38.2
OU753571.1	19	21.96	38.1
OU753572.1	20	21.59	38.3
OU753573.1	21	21.14	38.2
OU753574.1	22	19.51	37.8
OU753575.1	23	19.16	37.9
OU753576.1	24	19.08	38.3
OU753577.1	25	17.03	38
OU753578.1	26	16.74	38.1
OU753579.1	27	14.53	38.7
OU753580.1	28	14.37	38.7
OU753581.1	29	14.34	38.4
OU753582.1	30	13.73	39.4
OU753583.1	W	12.9	40
OU753552.1	Z	29.43	37.8
OU753584.1	MT	0.02	21.8
-	unplaced	0.11	42.3

Table 3. Software tools and versions used.

Software tool	Version	Source
BlobToolKit	3.4.0	Challis *et al.*, 2020
freebayes	1.3.1-17-gaa2ace8	Garrison & Marth, 2012
Hifiasm	0.15.3	Cheng *et al.*, 2021
HiGlass	1.11.6	Kerpedjiev *et al.*, 2018
Long Ranger ALIGN	2.2.2	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines
MitoHiFi	2	Uliano-Silva *et al.*, 2022
PretextView	0.2	Harry, 2022
purge_dups	1.2.3	Guan *et al.*, 2020
SALSA	2.2	Ghurye *et al.*, 2019

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: *Agrochola macilenta* (yellow-line quaker). Accession number PRJEB46326; https://identifiers.org/ena.embl/PRJEB46326. (Wellcome Sanger Institute, 2021)

The genome sequence is released openly for reuse. The *Agrochola macilenta* 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 listed here: https://doi.org/10.5281/zenodo.4893703.

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 listed here: https://doi.org/10.5281/zenodo.5013541.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783558.

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

Boyes DH, Fox R, Shortall CR, *et al.*: **Bucking the trend: the diversity of Anthropocene 'winners' among British moths.** *Front Biogeogr.* 2019; **11**(3): e43862.
Publisher Full Text

Brůna T, Hoff KJ, Lomsadze A, *et al.*: **BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database.** *NAR Genom Bioinform.* 2021; **3**(1): lqaa108.
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): 1361–1374.
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): 170–175.
PubMed Abstract | Publisher Full Text | Free Full Text

Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing**. 2012.
Publisher Full Text

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): 2896–2898.
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

Heath J, Emmet AM: **The Moths and Butterflies of Great Britain and Ireland Noctuidae (Part II) and Agaristidae**. Colchester: Harley Books, 1983.
Reference Source

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* Oxford University Press,

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**

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): 4647–4654.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Marren P: **Emperors, Admirals, and Chimney Sweepers.** Dorset: Little Toller, 2019.
Reference Source

Randle Z, Evans-Hill LJ, Parsons MS, *et al.*: **Atlas of Britain & Ireland's Larger Moths**. Newbury: NatureBureau, 2019.
Reference Source

Rao SSP, 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): 1665–1680.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio High Fidelity reads.** *bioRxiv.* [Preprint]. 2022.
Publisher Full Text

Waring P, Townsend M, Lewington R: **Field Guide to the Moths of Great Britain and Ireland: Third Edition**. Bloomsbury Wildlife Guides. 2017.
Reference Source

Wellcome Sanger Institute: T**he genome sequence of the Yellow-line Quaker, *Agrochola macilenta* (Hubner, 1809).** European Nucleotide Archive. [dataset], accession number PRJEB46326. 2021.