The genome sequence of the bramble shoot moth, Notocelia uddmanniana (Linnaeus, 1758) [version 1; peer review: 2 approved, 1 approved with reservations]

Douglas Boyes¹⁺, Peter W.H. Holland²,
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 UK Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK
² Department of Zoology, University of Oxford, Oxford, UK

* Deceased author

Abstract
We present a genome assembly from an individual male Notocelia uddmanniana (the bramble shoot moth; Arthropoda; Insecta; Lepidoptera; Tortricidae). The genome sequence is 794 megabases in span. The majority of the assembly, 99.96%, is scaffolded into 28 chromosomal pseudomolecules, with the Z sex chromosome assembled.

Keywords
Notocelia uddmanniana, bramble shoot moth, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.
**Species taxonomy**

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Tortricoidea; Tortricidae; Olethreutinae; Eucosmini; Notocelia; *Notocelia uddmanniana* (Linnaeus, 1758) (NCBI: txid1594315).

**Background**

*Notocelia uddmanniana* (bramble shoot moth) is widely distributed across Western Europe and North Africa, with records further east from Kazakhstan to China. The larvae feed on brambles (*Rubus* sp.), occurring commonly where these species exist, and occasionally cause damage to cultivated varieties (Gordon *et al.*, 1997). Eggs are laid singly on the foodplant, where larvae feed within a folded leaf and later within the tips of growing shoots; larvae overwinter in a silken web on the foodplant stem before recommencing feeding in spring (Dicker, 1939). *Notocelia uddmanniana* also occupies woodland, and is distributed widely throughout the UK, occurring more commonly in the south. The genome of *N. uddmanniana* was sequenced as part of the Darwin Tree of Life Project, a collaborative effort to sequence all of the named eukaryotic species in the Atlantic Archipelago of Britain and Ireland. Here we present a chromosomally complete genome sequence for *N. uddmanniana*, based on one male specimen from Wytham Woods, Oxfordshire, UK.

**Genome sequence report**

The genome was sequenced from a single male *N. uddmanniana* ([Figure 1](#)) collected from Wytham Woods, Oxfordshire, UK (latitude 51.772, longitude -1.338). A total of 18-fold coverage in Pacific Biosciences single-molecule long reads (N50 16 kb) and 49-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 165 missing/misjoins and removed 72 haplotypic duplications, reducing the assembly length by 2.16% and the scaffold number by 64.71%, and increasing the scaffold N50 by 9.53%.

**Methods**

Sample acquisition, DNA extraction and sequencing

A single male *M. uddmanniana* (ilNotUddm1) was collected from Wytham Woods, Oxfordshire, UK (latitude 51.772, longitude -1.338) by Douglas Boyes, UKCEH, using a light

**Table 1. Genome data for Notocelia uddmanniana, ilNotUddm1.1.**

| Project accession data |  |
|------------------------|--|
| Assembly identifier     | ilNotUddm1 |
| Species                | *Notocelia uddmanniana* |
| Specimen               | ilNotUddm1 |
| NCBI taxonomy ID        | NCBI:txid1594315 |
| BioProject             | PRJEB42137 |
| BioSample ID           | SAMEA7519916 |
| Isolate information    | Male, whole organism |

**Raw data accessions**

| Project accession data |  |
|------------------------|--|
| PacificBiosciences SEQUEL II | ERR6590584 |
| 10X Genomics Illumina   | ERR6002710-ERR6002713 |
| Hi-C Illumina          | ERR6002707-ERR6002709 |

**Genome assembly**

| Assembly accession | GCA_905163555.1 |
|--------------------|-----------------|
| Accession of alternate haplotype | GCA_905163575.1 |
| Span (Mb)           | 794             |
| Number of contigs   | 238             |
| Contig N50 length (Mb) | 7              |
| Number of scaffolds | 49              |
| Scaffold N50 length (Mb) | 29             |
| Longest scaffold (Mb) | 51              |
| BUSCO* genome score | C:98.3%[S:97.6%,D:0.7%], F:0.5%,M:1.2%,n:5286 |

*BUSCO scores based on the lepidoptera_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/ ilNotUddm1.1/dataset/CAJHZS01/busc0.*
The specimen was identified by the same individual and preserved on dry ice.

DNA was extracted from whole organism tissue 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 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 HiSeq X instruments. Hi-C data were generated from remaining whole organism tissue using the Arima v1.0 kit and sequenced on HiSeq X.

**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), without the -e flag. 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

**Figure 3. Genome assembly of Notocelia uddmanniana, iINotUddm1.1: 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/iINotUddm1.1/dataset/CAJHZS01/blob.
was performed using gEVAL, HiGlass (Kerpedjiev et al., 2018) and Pretex. The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva et al., 2021) and annotated using MitoFinder (Allio et al., 2020). The genome was analysed and BUSCO scores generated within the Blob-ToolKit 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 agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the...
Figure 5. Genome assembly of *Notocelia uddmanniana*, ilNotUddm1.1: Hi-C contact map. Hi-C contact map of the ilNotUddm1.1 assembly, visualised in HiGlass. Chromosomes are given in order of size from left to right and top to bottom.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Notocelia uddmanniana*, ilNotUddm1.1.

| INSDC accession | Chromosome | Size (Mb) | GC%  |
|-----------------|------------|-----------|------|
| LR991053.1      | 1          | 51.12     | 38.3 |
| LR991054.1      | 2          | 44.98     | 38.3 |
| LR991055.1      | 3          | 34.90     | 38.6 |
| LR991056.1      | 4          | 33.91     | 38.6 |
| LR991057.1      | 5          | 33.55     | 38.5 |
| LR991058.1      | 6          | 31.63     | 38.3 |
| LR991059.1      | 7          | 31.24     | 38.4 |
| LR991060.1      | 8          | 30.81     | 38.7 |
| LR991061.1      | 9          | 29.12     | 38.8 |
| LR991062.1      | 10         | 28.99     | 38.6 |
| LR991063.1      | 11         | 28.96     | 38.8 |
| LR991064.1      | 12         | 28.81     | 38.6 |
| LR991065.1      | 13         | 27.46     | 38.5 |
| LR991066.1      | 14         | 25.91     | 38.7 |
| LR991067.1      | 15         | 25.88     | 38.7 |
| LR991068.1      | 16         | 25.71     | 38.5 |
| LR991069.1      | 17         | 25.63     | 38.7 |
| LR991070.1      | 18         | 24.47     | 39   |
| LR991071.1      | 19         | 21.77     | 38.9 |
| LR991072.1      | 20         | 19.36     | 38.8 |
| LR991073.1      | 21         | 18.18     | 39.2 |
| LR991074.1      | 22         | 17.67     | 38.9 |
| LR991075.1      | 23         | 17.30     | 39   |
| LR991077.1      | 24         | 15.87     | 38.9 |
| LR991076.1      | 25         | 16.51     | 39.8 |
| LR991078.1      | 26         | 15.07     | 39.4 |
| LR991079.1      | 27         | 12.65     | 39.3 |
| LR991052.1      | Z          | 75.62     | 38.4 |
| LR991080.1      | MT         | 0.02      | 18.8 |
| -               | Unplaced   | 1.01      | 40.8 |
Table 3. Software tools used.

| Software tool     | Version | Source                                                                 |
|-------------------|---------|------------------------------------------------------------------------|
| Hifiasm           | 0.12    | 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          | 1       | https://github.com/marcelauliano/MitoHiFi                              |
| gEVAL             | N/A     | Chow et al., 2016                                                     |
| HiGlass           | 1.11.6  | Kerpedjie et al., 2018                                                |
| PretextView       | 0.1.x   | https://github.com/wtsi-hpag/PretextView                              |
| BlobToolKit       | 2.6.2   | Challis et al., 2020                                                  |

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: Notocelia uddmanniana (bramble shoot). Accession number PRJEB42137: https://www.ebi.ac.uk/ena/browser/view/PRJEB42037.

The genome sequence is released openly for reuse. The N. uddmanniana 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.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.5744840.

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.5743293.

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: 1207.3967. 2012. Reference Source

Ghurye J, Rhie A, Walenz BP, et al.: Integrating Hi-C Links with Assembly Graphs
Open Peer Review

Current Peer Review Status: ? ✔ ✔

Version 1

Reviewer Report 21 August 2024

https://doi.org/10.21956/wellcomeopenres.19338.r54332

© 2024 Bono H. 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.

✔ Hidemasa Bono
Hiroshima University, Kagamiyam, Japan

Authors described genome sequencing of the bramble shoot moth, *Notocelia uddmanniana*. There are no other assemblies in the INSDC records besides the one submitted by the author. There are no problem in sample acquisition and DNA extraction. Furthermore, the data analysis method described in the manuscript is modern and appropriate for the genome assembly of insect species.

There is a typo in the Data availability section; PRJEB42137 should be PRJEB42037. It should be fixed before the final indexing because this information is very critical for 'Data Note' manuscript.

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:** Genome biology

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.
Violaine Llaurens

CNRS, Paris, France

The article accurately describes the chromosome-level assembly of the genome of a male *Notocelia uddmanniana* sampled in Oxfordshire. These new data open new research avenue in different scientific fields from ecology to evolution in Lepidoptera. A single male was sequenced, and the assembly show that the genome size is 794 mb long, and is arranged in 28 different chromosomes. At the moment, this genome remains to be annotated.

I have only a minor comment on this genome release: It is unclear to me whether the chromosome 1, which is the largest chromosome, corresponds to the Z-chromosome?

**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:** Evolutionary biology

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.
In this manuscript, the authors describe the sequencing and assembly of the *Notocelia uddmanniana* genome using DNA from a single male specimen collected in the UK. The primary genome sequence assembly includes proposed chromosomal pseudomolecule sequences for 27 autosomes, the Z sex chromosome, and a complete mitochondrial genome. No females were collected, so the W sex chromosome was not sequenced. Overall, this is a solid contribution to the scientific literature.

Some suggestions to the authors:

1. Methods: The date of collection for the sequenced *Notocelia uddmanniana* specimen is not included in the manuscript. It is best practice to include collection dates for all wild-caught insect specimens used for genome sequencing because it is not unusual for cryptic species pairs to be distinguishable by differences in phenology of adult emergence. If it should come to pass that genus *Notocelia* includes cryptic species, knowing the specimen collection date can help future researchers determine to which taxon this sequenced genome should be assigned, especially since it appears that the entire specimen was used for DNA extraction and is no longer available for morphological examination.

2. For future work, I suggest that the researchers preferentially sequence the heterogametic sex when assembling genomes for previously unstudied species (in the case of Lepidoptera, the heterogametic sex is female), so that draft assembles can be prepared for both sex chromosomes.

**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?**
No

**Are the datasets clearly presented in a useable and accessible format?**
Yes

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

**Reviewer Expertise:** Evolutionary biology of insects, phylogenomics

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.