DATA NOTE

The genome sequence of the bird's nest moth, *Tinea trinotella* (Thunberg, 1794) [version 1; peer review: 1 approved, 1 approved with reservations]

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, Physilia Chua[2],
Darwin Tree of Life Consortium

[1]UK Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK
[2]Wellcome Sanger Institute, Cambridge, UK

[+] Deceased author

v1 **First published:** 04 Apr 2022, **7**:124
https://doi.org/10.12688/wellcomeopenres.17826.1
Latest published: 04 Apr 2022, **7**:124
https://doi.org/10.12688/wellcomeopenres.17826.1

Abstract

We present a genome assembly from an individual male *Tinea trinotella* (the bird's nest moth; Arthropoda; Insecta; Lepidoptera; Tineidae). The genome sequence is 372 megabases in span. The majority of the assembly (99.98%) is scaffolded into 30 chromosomal pseudomolecules, with the Z sex chromosome assembled. The mitochondrial genome was also assembled, and is 16.9 kilobases in length.

Keywords

Tinea trinotella, bird's nest moth, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ? ✓

	1	2
version 1 04 Apr 2022	? view	✓ view

1. **Jacqueline Heckenhauer**, LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt, Germany

2. **Stephen Richards**, Baylor College of Medicine, Houston, 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: Boyes D: Investigation, Resources; **Chua P**: 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 bird's nest moth, *Tinea trinotella* (Thunberg, 1794) [version 1; peer review: 1 approved, 1 approved with reservations]** Wellcome Open Research 2022, **7**:124 https://doi.org/10.12688/wellcomeopenres.17826.1

First published: 04 Apr 2022, **7**:124 https://doi.org/10.12688/wellcomeopenres.17826.1

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Tineoidea; Tineidae; Tineinae; Tinea; *Tinea trinotella* (Thunberg, 1794) (NCBI:txid1594354).

Background

The bird's-nest moth (*Tinea trinotella*) belongs to the family Tineidae, known as fungus moths. *Tinea trinotella* has a wingspan of between 12 and 18 mm. The scientific name *trinotella* is derived from the three dark spots located on the bird's-nest moth's forewings, which are greyish with a golden hue. It also has a yellow-orange head tuft. These two distinctive features enable accurate identification of this small moth. T. trinotella is widespread and can be found across most of western Europe in various habitats, except Iceland. In the UK, it is described as a fairly common species.

Adults are nocturnal and are often sighted flying at night close to light sources. Their flight period is between May to August. In some parts of its range, there may be two generations occurring each year. Similar to other *Tinea* species, *T. trinotella* feeds on decayed matter.

Larvae live in portable cases which are transported by adults. They are also commonly found living in bird's nests feeding on detritus, hence its name (Boyes, 2018). To our knowledge, this is the first high-quality genome, generated for *T. trinotella* as part of the Darwin Tree of Life project. We believe that this will further our ecological understanding of this small moth by providing curated DNA reference data for molecular monitoring of this species.

Genome sequence report

The genome was sequenced from one male *T. trinotella* (Figure 1) collected from Wytham Woods, Oxfordshire (Biological vice-county: Berkshire), UK (latitude 51.772, longitude -1.338). A total of 33-fold coverage in Pacific Biosciences single-molecule long reads and 83-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 6 missing/misjoins and removed 2 haplotypic duplications, reducing the assembly size by 0.13% and the scaffold number by 6.06%.

The final assembly has a total length of 372 Mb in 31 sequence scaffolds with a scaffold N50 of 13.7 Mb (Table 1). The majority of the assembly sequence (99.98%) was assigned to 30 chromosomal-level scaffolds, representing 29 autosomes (numbered by sequence length), and the Z sex chromosome (Figure 2–Figure 5; Table 2). The single unassigned scaffold is made up of telomeric sequence. The assembly has a BUSCO v5.2.2 (Manni *et al.*, 2021) completeness of 94.6% (single 93.9%, duplicated 0.8%) using the lepidoptera_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. Image of the *Tinea trinotella* (ilTinTrin1) specimen taken prior to preservation and processing. Unfortunately, a higher-quality image of the specimen is not available.

Table 1. Genome data for *Tinea trinotella*, ilTinTrin1.

Project accession data	
Assembly identifier	ilTinTrin1.1
Species	*Tinea trinotella*
Specimen	ilTinTrin1
NCBI taxonomy ID	NCBI:txid1594354
BioProject	PRJEB42958
BioSample ID	SAMEA7519924
Isolate information	Male, whole organism
Raw data accessions	
PacificBiosciences SEQUEL II	ERR6608655
10X Genomics Illumina	ERR6054443-ERR6054446
Hi-C Illumina	ERR6054447
Genome assembly	
Assembly accession	GCA_905220615.1
Accession of alternate haplotype	GCA_905220625.1
Span (Mb)	372
Number of contigs	34
Contig N50 length (Mb)	13.7
Number of scaffolds	31
Scaffold N50 length (Mb)	13.7
Longest scaffold (Mb)	15.9
BUSCO* genome score	C:94.6%[S:93.9%,D:0.8%],F:1.0%,M:4.4%,n:5286

*BUSCO scores based on the lepidoptera_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.genomehubs.org/view/ilTinTrin1.1/dataset/CAJNAM01/busco.

Methods

Sample acquisition, DNA extraction and sequencing

A single male *T. trinotella* (ilTinTrin1) was collected from Wytham Woods, Oxfordshire (Biological vice-county: Berkshire), UK (latitude 51.772, longitude -1.338) by Douglas Boyes, UKCEH, using a light trap in woodland. The sample was identified by the same individual, and preserved on dry ice.

DNA was extracted from whole organism tissue of ilTinTrin1 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 Chromium 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 (HiFi) and Illumina HiSeq X (10X) instruments. Hi-C data were generated from remaining whole organism tissue using the Arima Hi-C+ kit and sequenced on a HiSeq X 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 as described previously (Howe *et al.*, 2021). Manual curation

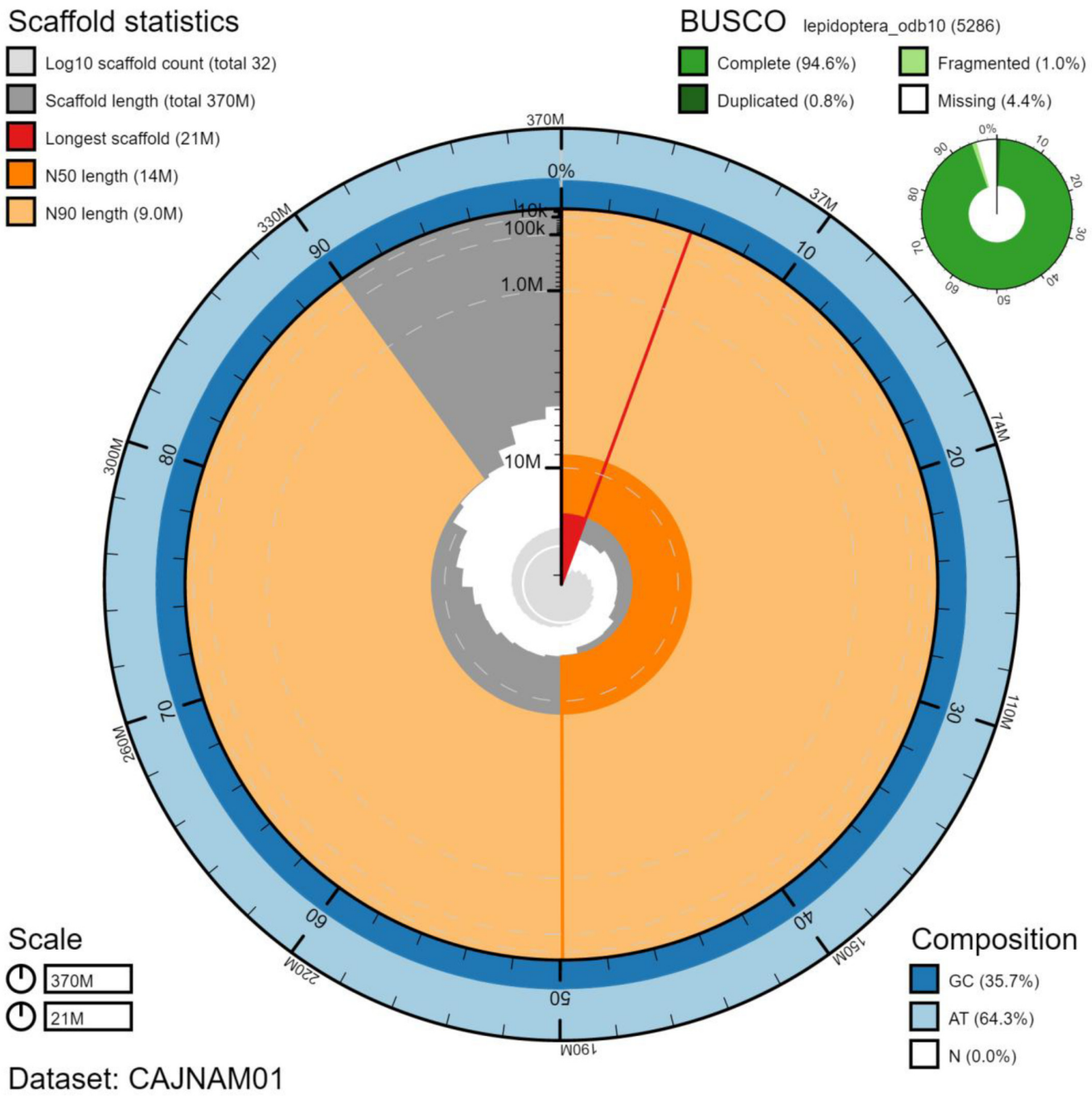

Figure 2. Genome assembly of *Tinea trinotella*, ilTinTrin1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 371,738,977 bp assembly. The distribution of chromosome lengths is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly (21,020,397 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (13,660,558 and 8,970,260 bp), respectively. The pale grey spiral shows the cumulative chromosome 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/ilTinTrin1.1/dataset/CAJNAM01/snail.

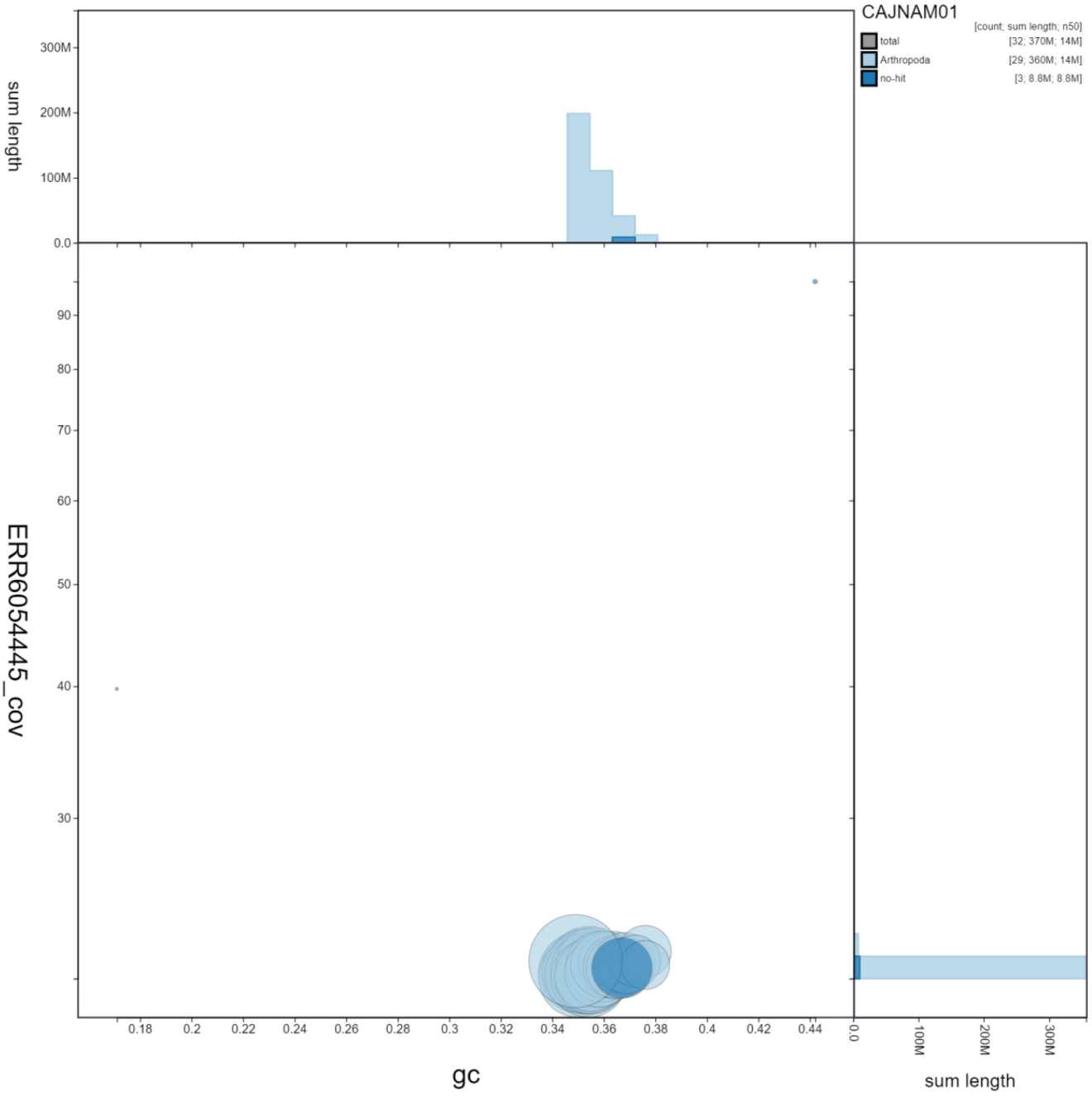

Figure 3. Genome assembly of *Tinea trinotella*, ilTinTrin1.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilTinTrin1.1/dataset/CAJNAM01/blob.

(Howe *et al.*, 2021) was performed using HiGlass (Kerpedjiev *et al.*, 2018) and Pretext. The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2021), which performs annotation using MitoFinder (Allio *et al.*, 2020). 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.

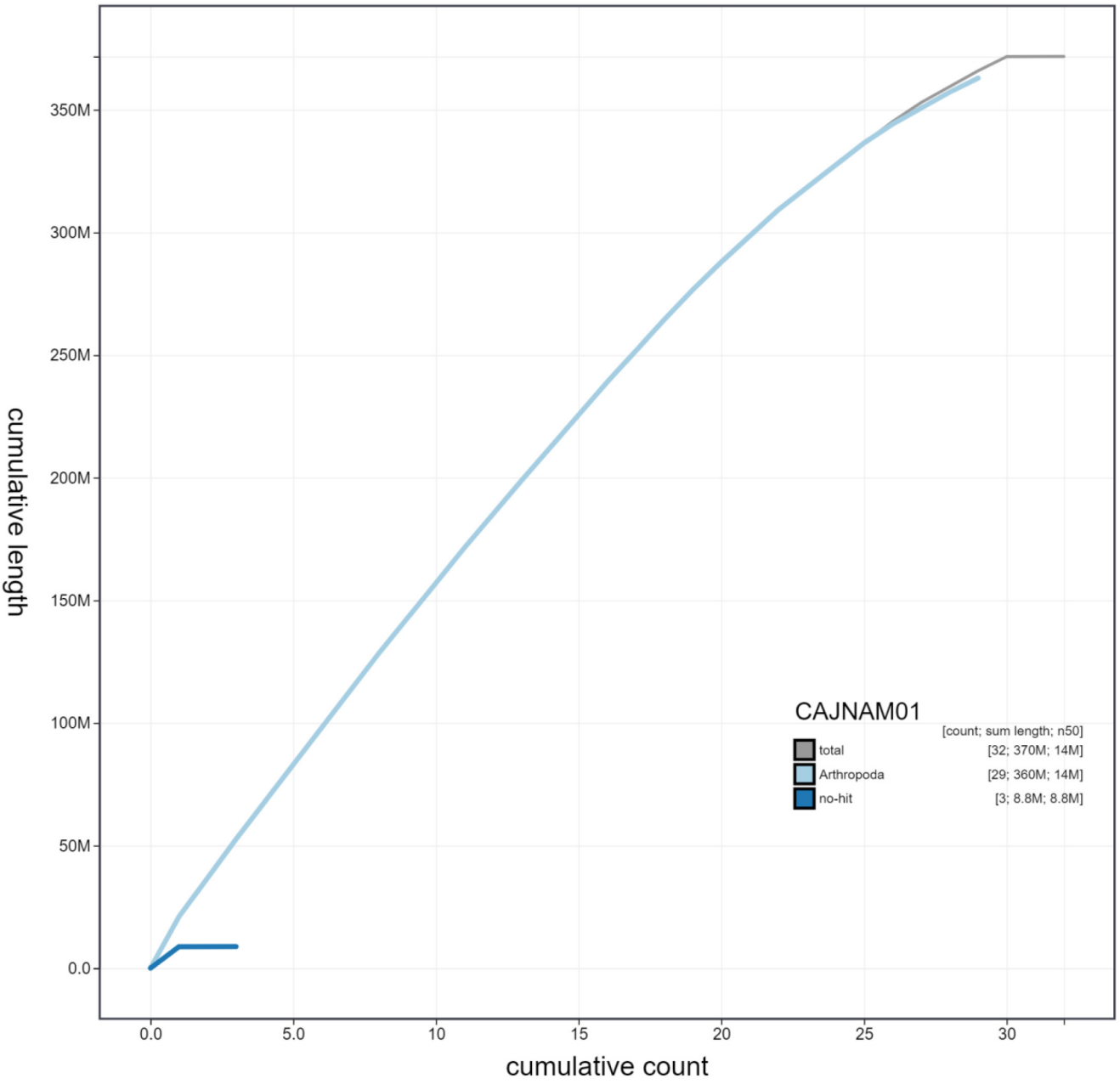

Figure 4. Genome assembly of *Tinea trinotella*, ilTinTrin1.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilTinTrin1.1/dataset/CAJNAM01/cumulative.

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

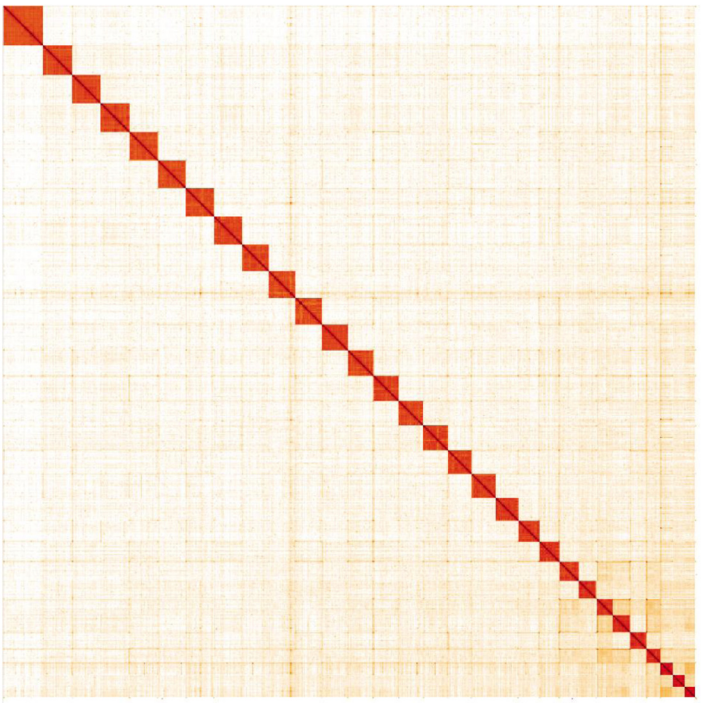

Figure 5. Genome assembly of *Tinea trinotella*, ilTinTrin1: Hi-C contact map. Hi-C contact map of the ilTinTrin1.1 assembly, visualised in HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this map is available here.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Tinea trinotella*, ilTinTrin1.1.

INSDC accession	Chromosome	Size (Mb)	GC%
HG992306.1	1	15.89	35.1
HG992307.1	2	15.75	35.4
HG992308.1	3	15.25	35.3
HG992309.1	4	15.20	35.0
HG992310.1	5	15.16	35.0
HG992311.1	6	15.11	35.4
HG992312.1	7	14.99	35.3
HG992313.1	8	14.40	35.1
HG992314.1	9	14.33	35.4
HG992315.1	10	14.31	35.2
HG992316.1	11	13.78	35.4
HG992317.1	12	13.66	35.8
HG992318.1	13	13.47	35.6
HG992319.1	14	13.43	35.4
HG992320.1	15	13.22	35.7
HG992321.1	16	12.77	35.5
HG992322.1	17	12.70	35.5
HG992323.1	18	12.19	35.9
HG992324.1	19	11.31	35.8
HG992325.1	20	10.71	36.3
HG992326.1	21	10.58	36.0
HG992327.1	22	9.31	36.4
HG992328.1	23	8.98	36.5
HG992329.1	24	8.97	36.8
HG992330.1	25	8.75	36.7
HG992331.1	26	7.69	36.9
HG992332.1	27	6.52	37.6
HG992333.1	28	6.48	37.2
HG992334.1	29	5.75	37.6
HG992305.1	Z	21.02	34.9
HG992335.1	MT	0.02	17.1
-	Unplaced	0.04	44.2

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
SALSA	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	1.0	Uliano-Silva *et al.*, 2021
HiGlass	1.11.6	Kerpedjiev *et al.*, 2018
PretextView	0.1.x	https://github.com/wtsi-hpag/PretextView
BlobToolKit	3.0.5	Challis *et al.*, 2020

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: Tinea trinotella (bird's-nest moth). Accession number PRJEB42958; https://identifiers.org/ena.embl/PRJEB42958.

The genome sequence is released openly for reuse. The *T. trinotella* 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.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

Boyes DH: **Natural History of Lepidoptera Associated with Bird Nests in Mid-Wales.** *The Entomologist's Record and Journal of Variation.* [manuscript in Review]. 2018.
Reference Source

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: 1207.3907. 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**

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 12 April 2022

https://doi.org/10.21956/wellcomeopenres.19737.r49717

✓ **Stephen Richards**
Human Genome Sequencing Center, Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, TX, USA

I agree very much with the comments of the firsts reviewer Dr Heckenhauer, but perhaps I am a bit less strict on a data note. I have minor issues:

1. For the rational of why this dataset was created, I believe you should also point out that the Darwin Tree of Life aims genetically describe all species in the UK, and as a species in the UK this is one of the targets of this impressive project.

2. On sequencing a male. I'm sure you would have sequenced a female if you could have, but perhaps a sentence describing why not (something like - "we were unable to collect a female at this time, and so sequence of the W is not represented here" - or whatever the truth is).

3. On the quality - The scaffolding onto chromosome sized pieces looks really good and as always I can find the data in the databases - Bravo.

4. On methods in the long term for all these DToL data note papers, it would be nice to see the computational methods presented as actual commands in a script - perhaps in one kind of container for may installation - so the other may run and install easier, but for now the methods meet the usual standards.

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?
Yes

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Arthropod genomics, genome assembly, human genomics, transcriptomics, comparative 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 11 April 2022

https://doi.org/10.21956/wellcomeopenres.19737.r49719

?

Jacqueline Heckenhauer
LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt, Germany

In their manuscript, authors present the genome sequence of the bird's nest moth, *Tinea trinotella* using a *de novo* assembly method following the approach used by the Darwin Tree of Life Project. The presented genome is of high quality and released openly for reuse. This is the first time this species has been assembled, therefore this high-quality reference genome is beneficial to the field and I approve the article with reservations and recommend a very minor revision based on the following suggestions:

The rationale for sequencing the genome stated by the authors is a better ecological understanding and monitoring of this species. In my opinion, these topics might be difficult to address with just the genome sequence without genome annotation. Authors state that the genome will be annotated and presented through the pipeline at the European Bioinformatics Institute. I wonder if it might be easier for the reader if authors would present the genome sequence together with its annotation in one data note. However, I highly appreciate that this genome assembly as well as the raw sequence data already have been deposited in public databases and are easy to access. This is really important and useful for the scientific community.

Abstract:
- Authors state that the mitogenome was assembled. Would it be possible to upload the mitogenome separately? Where can I find just the mitogenome sequence?

Introduction:
- I would appreciate if authors could add references in the first two paragraphs of the introduction.

- The species name should be written in italics throughout the text.

- The beginning of the third paragraph is not clear to me. “Larvae live in portable cases which are transported by adults.” What does this mean? Do larvae pupate inside these portable cases? May be some further information about the life-cycle could be added here. I think it is also important to add how the cases build, i.e. that they made out of debris which is held together by larval silk.

- Regarding the rational, authors could state why is it important to monitor this specific species.

- I suggest to add information on the karyotype and the expectation of the number of chromosomes and genome size.

Genome sequence report:

- Why was a male individual used for sequencing (W chromosome)?

- It might be interesting to put this species in context with other members of this genus with respect to availability of genome data. If I am correct, this is the first genome published of this genus and authors should make this clear. Comparison of this assembly with previously published Lepidoptera genomes would have helped in appreciating the high quality achieved (see also Hotaling *et al.*, 2021[1]).

- How did authors calculate sequencing coverage? Is this based on assembly length?

- Regarding genome completeness, besides looking at the BUSCOs it would be interesting to compare the assembly length with the genome size, preferably estimated by flow cytometry if available. If genome size estimates are not available for this individual / species, authors could estimate genome size with Genomescope2 or similar using their PacBio reads.

- What is meant by “the sample has been identified by the same individual”? Morphologically? Via barcoding? I would also reformulate this sentence “this specimen has been identified by …”

- I am not sure if the description of genome assembly does not allow replication by others. Parameters used for each program are not given. Did authors use the default parameters? If yes, please state.

Figure 1:

- Can authors present a higher-quality image of another specimen of the same species in addition to the photo in figure 1?

References

1. Hotaling S, Sproul J, Heckenhauer J, Powell A, et al.: Long Reads Are Revolutionizing 20 Years of Insect Genome Sequencing. *Genome Biology and Evolution*. 2021; **13** (8). Publisher Full Text

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?
No

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: biodiversity genomics, aquatic insects, phylogenetics, comparative 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.