Check for updates

DATA NOTE

The genome sequence of the smoky wainscot, *Mythimna impura* (Hubner, 1808) [version 1; peer review: 1 approved]

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, Melanie Gibbs[1],
Darwin Tree of Life Consortium

[1]UK Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK

[+] Deceased author

v1 **First published:** 12 Sep 2022, **7**:226
https://doi.org/10.12688/wellcomeopenres.18104.1

Latest published: 12 Sep 2022, **7**:226
https://doi.org/10.12688/wellcomeopenres.18104.1

Abstract

We present a genome assembly from an individual female *Mythimna impura* (smoky wainscot; Arthropoda; Insecta; Lepidoptera; Noctuidae). The genome sequence is 949 megabases in span. The majority of the assembly (98.39%) is scaffolded into 32 chromosomal pseudomolecules with the W and Z sex chromosomes assembled. The complete mitochondrial genome was also assembled and is 15.3 kilobases in length. Gene annotation of this assembly on Ensembl has identified 15,441 protein coding genes.

Keywords

Mythimna impura, smoky wainscot, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓

	1
version 1 12 Sep 2022	✓ view

1. **Francois Olivier Hebert**, Université de Montréal, Montréal, Canada

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; **Gibbs M**: 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) 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: Boyes D, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective *et al.* **The genome sequence of the smoky wainscot, *Mythimna impura* (Hubner, 1808) [version 1; peer review: 1 approved]** Wellcome Open Research 2022, **7**:226 https://doi.org/10.12688/wellcomeopenres.18104.1

First published: 12 Sep 2022, **7**:226 https://doi.org/10.12688/wellcomeopenres.18104.1

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Noctuoidea; Noctuidae; Hadeninae; *Mythimna*; *Mythimna impura* (Hubner, 1808) (NCBI:txid987985).

Background

The smoky wainscot, *Mythimna impura* (Hubner, 1808), is a common, nocturnal, non-pest, macro-moth species that occurs across the Palearctic. In Great Britain, *M. impura* has been categorised using the International Union for Conservation of Nature (IUCN) Red List criteria, as a resident species of Least Concern (Fox *et al.*, 2021). Larvae feed on grasses (Gramineae; Robinson *et al.*, 2010), and overwinter as small larvae. They can have one or two broods per year. Adults fly June to October. Noctuids are relatively mobile compared with other families; *M. impura* has a 'medium' dispersal ability (Jones *et al.*, 2016).

Mythimna impura inhabit downland, sand dunes and rough grassy areas, including field margins. Moths are used as indicator species as they are sensitive to environmental change (Wagner *et al.*, 2021); *M. impura* is a good indicator species for heavily grazed plots in the steppes of Mongolia (Enkhtur *et al.*, 2017). Worldwide, the Noctuidae family contains many species considered agricultural pests. Integrated Pest Management programmes, including sex attractant trapping, have been developed for their control (Renou *et al.*, 1991). The genome of *M. impura*, along with other species from this family, will provide valuable resources for comparative studies of these economically important insects.

Genome sequence report

The genome was sequenced from a single female *M. impura* collected from Ant Hills region, Wytham, Berkshire, UK (Figure 1). A total of 38-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 54-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 47 missing/misjoins and removed 9 haplotypic duplications, reducing the assembly size by 1.27% and the scaffold number by 30.83%, and increasing the scaffold N50 by 20.56%.

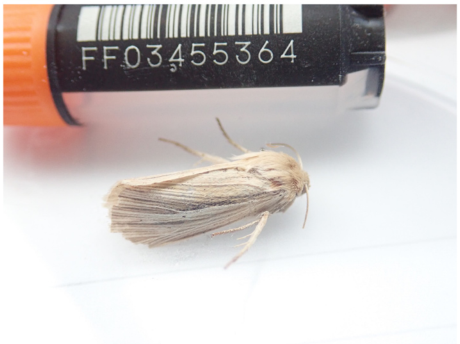

Figure 1. Image of the *Mythimna impura* specimen taken prior to preservation and processing.

The final assembly has a total length of 949 Mb in 92 sequence scaffolds with a scaffold N50 of 30.6 Mb (Table 1). The majority, 98.39%, of the assembly sequence was assigned to 32 chromosomal-level scaffolds, representing 30 autosomes (numbered by sequence length) and the W and Z sex chromosomes (Figure 2–Figure 5; Table 2).

Table 1. Genome data for *Mythimna impura*, ilMytImpu1.2.

Project accession data	
Assembly identifier	ilMytImpu1.2
Species	*Mythimna impura*
Specimen	ilMytImpu1 (genome assembly, Hi-C, RNA-Seq)
NCBI taxonomy ID	987985
BioProject	PRJEB42135
BioSample ID	SAMEA7519913
Isolate information	Female; thorax/abdomen (genome assembly, RNA-Seq), head (Hi-C)
Raw data accessions	
PacificBiosciences SEQUEL II	ERR6576317;ERR6590581
10X Genomics Illumina	ERR6002686-ERR6002693
Hi-C Illumina	ERR6002694,ERR6002695
PolyA RNA-Seq Illumina	ERR6286707
Genome assembly	
Assembly accession	GCA_905147345.2
Accession of alternate haplotype	GCA_905147275.1
Span (Mb)	949
Number of contigs	139
Contig N50 length (Mb)	23.3
Number of scaffolds	92
Scaffold N50 length (Mb)	30.6
Longest scaffold (Mb)	36.2
BUSCO* genome score	C:98.9%[S:98.1%,D:0.9%],F:0.3%, M:0.8%,n:5,286
Genome annotation	
Number of protein-coding genes	15,441
Average length of coding sequence (bp)	1,387.75
Average number of exons per transcript	6.14
Average intron size (bp)	3,233.13

*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/ilMytImpu1.1/dataset/CAJHVM01.1/busco.

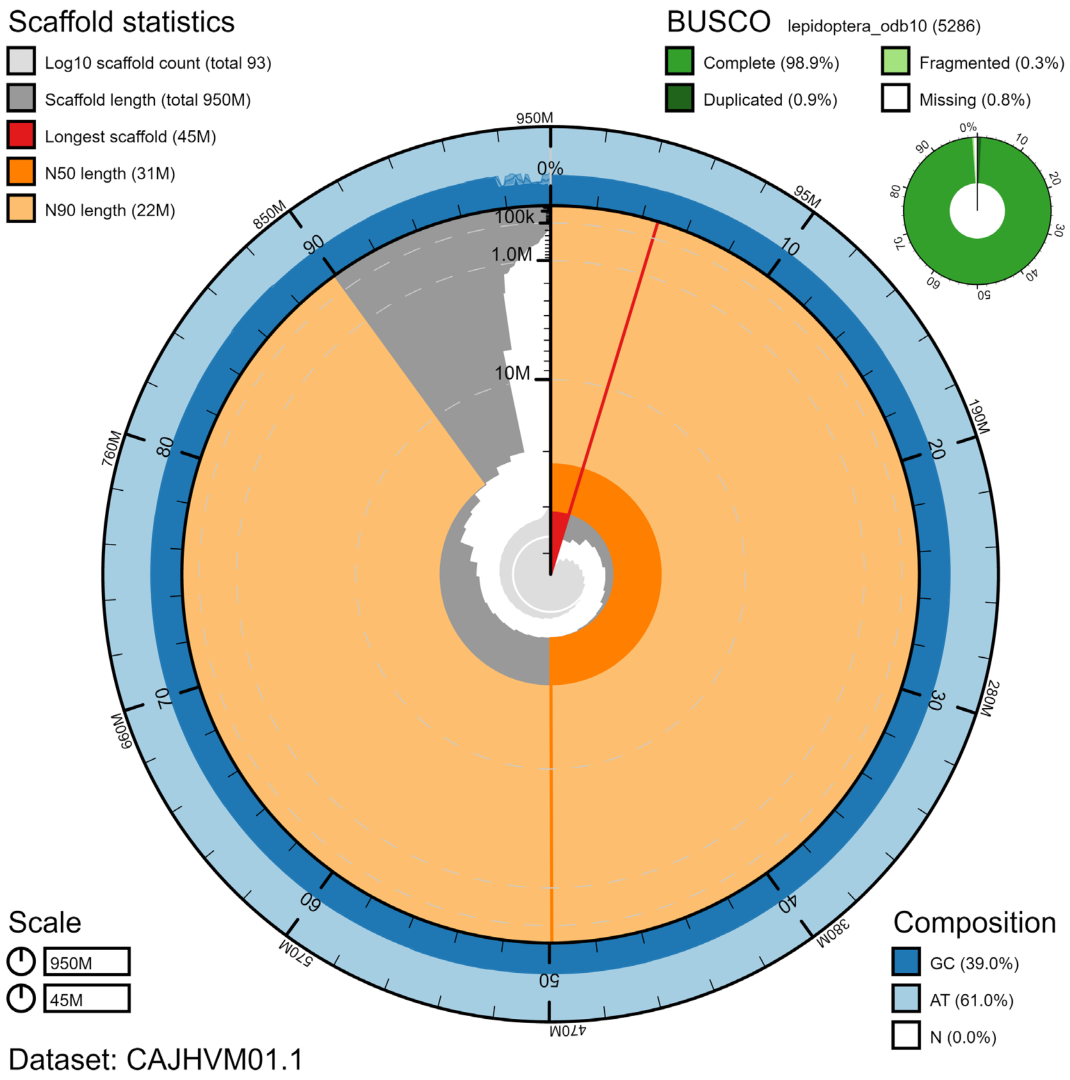

Figure 2. Genome assembly of *Mythimna impura*, ilMytImpu1.2: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 949,131,760 bp assembly. The distribution of chromosome lengths is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly (44,940,339 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (30,601,083 and 21,948,855 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/ilMytImpu1.1/dataset/CAJHVM01.1/snail.

The assembly has a BUSCO v5.2.2 (Manni *et al.*, 2021) completeness of 98.9% (single 98.1%, duplicated 0.8%) 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.

Genome annotation report

The ilMytImpu1.2 genome has been annotated using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Mythimna_impura_GCA_905147345.2/). The resulting annotation includes 28,738 transcribed mRNAs from 15,441 protein-coding and 3,690 non-coding genes. There is an average of 6.41 exons and 5.41 introns per canonical protein coding transcript, with an average intron length of 3,233.13. A total of 5,359 gene loci have more than one associated transcript.

Methods

Sample acquisition and nucleic acid extraction

A single female *M. impura* specimen (ilMytImpu1) was collected using a light trap from Ant Hills region, Wytham, Berkshire, UK (latitude 51.765, longitude -1.327) by Douglas

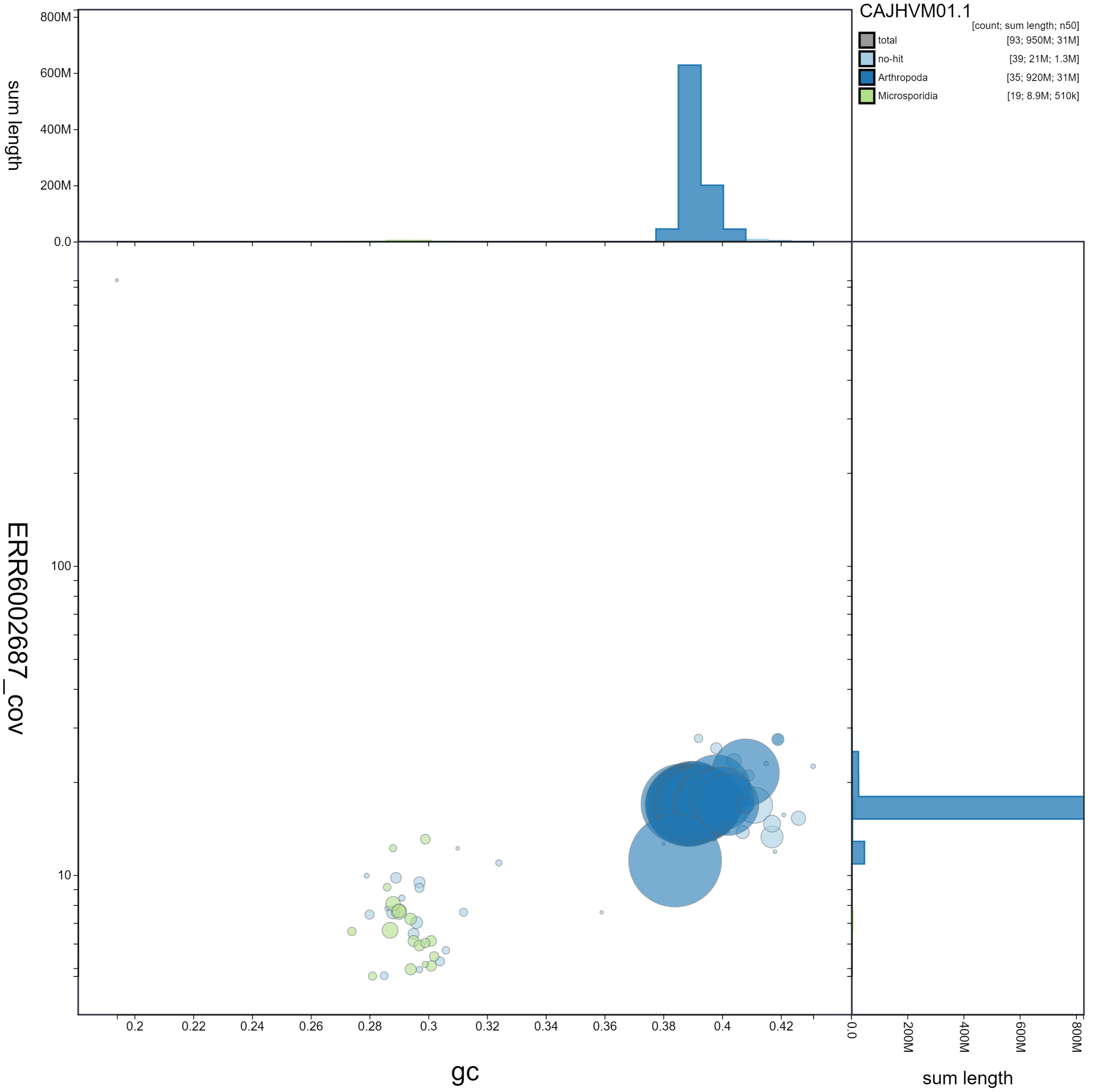

Figure 3. Genome assembly of *Mythimna impura*, ilMytImpu1.2: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilMytImpu1.1/dataset/CAJHVM01.1/blob.

Boyes (University of Oxford). The specimen was identified by Douglas Boyes and snap-frozen on dry ice.

DNA was extracted at the Scientific Operations Core, Wellcome Sanger Institute. The ilMytImpu1 sample was weighed and dissected on dry ice with head tissue set aside for Hi-C sequencing. Thorax and abdomen tissue was disrupted by manual grinding in lysis buffer with a disposable pestle. Fragment size analysis of 0.01-0.5 ng of DNA was then performed “using an Agilent FemtoPulse. High molecular weight (HMW) DNA was extracted using the Qiagen MagAttract HMW DNA extraction kit. Low molecular weight DNA was removed from a 200-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 DNA was sheared into an average fragment size between

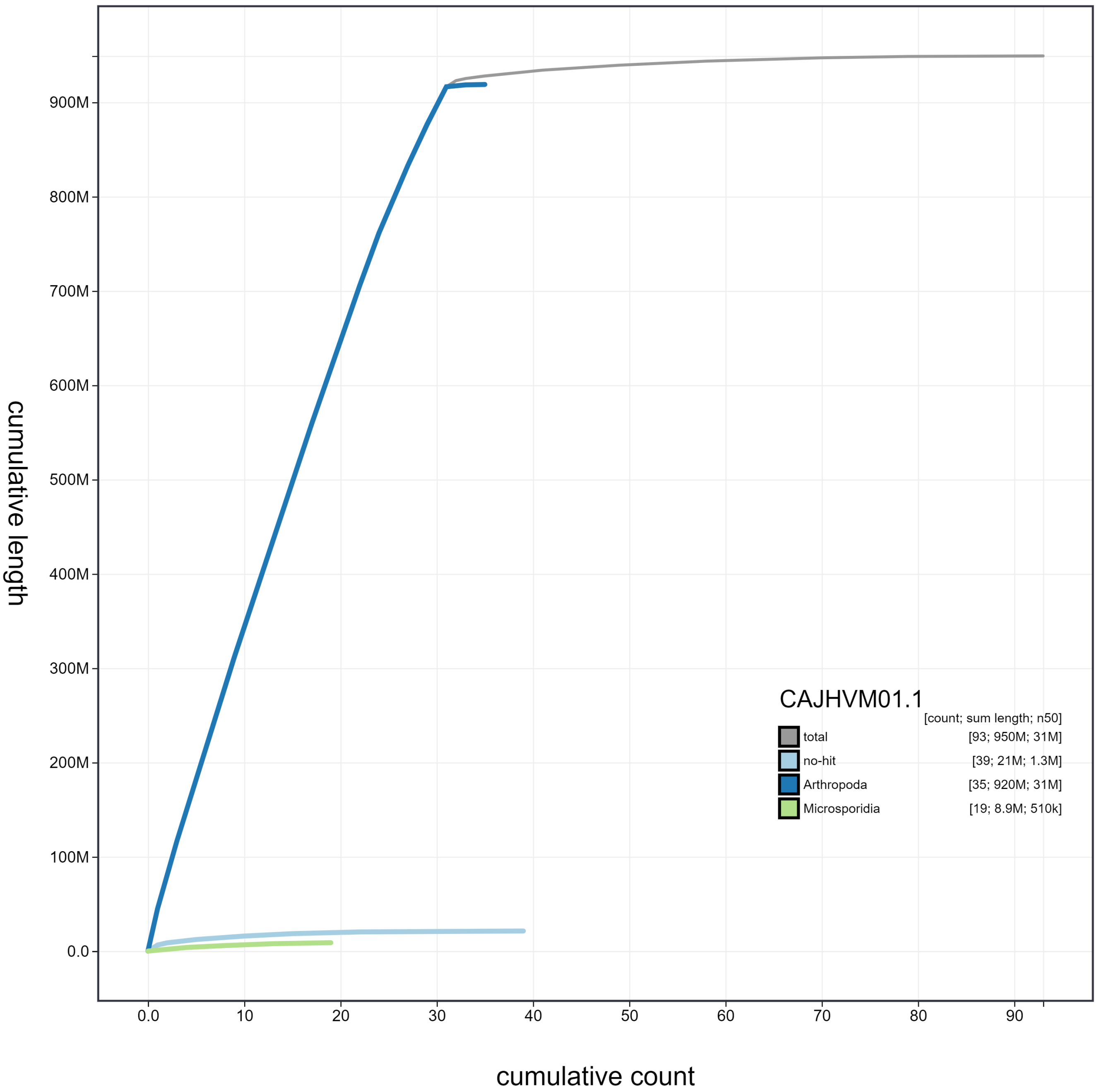

Figure 4. Genome assembly of *Mythimna impura*, ilMytImpu1.2: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilMytImpu1.1/dataset/CAJHVM01.1/cumulative.

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 remaining thorax and abdomen tissue of ilMytImpu1 in the Tree of Life Laboratory at the WSI using TRIzol, according to the manufacturer's instructions. RNA was then eluted in 50 μl RNAse-free water and

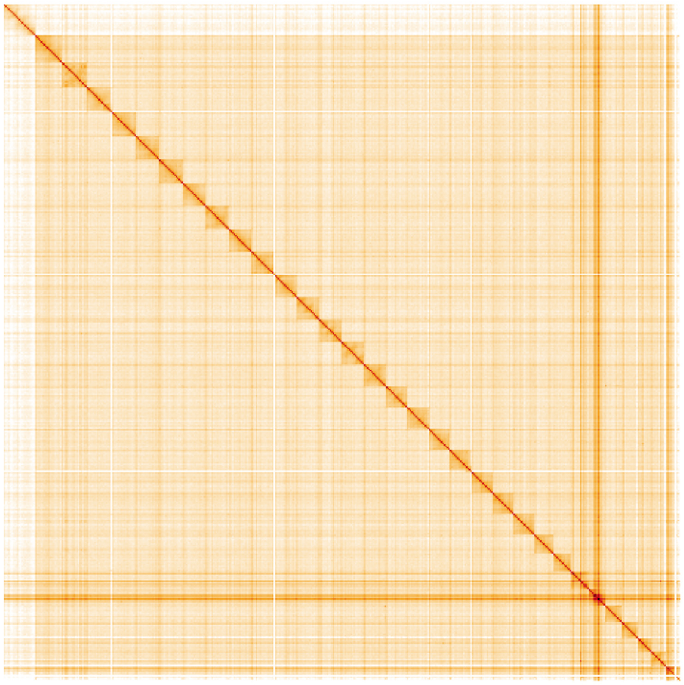

Figure 5. Genome assembly of *Mythimna impura*, ilMytImpu1.2: Hi-C contact map. Hi-C contact map of the ilMytImpu1.2 assembly, visualised in HiGlass. Chromosomes are arranged in size order from left to right and top to bottom. The interactive Hi-C map can be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=GU924wLKRTmysOazbqg-jw.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Mythimna impura*, ilMytImpu1.2.

INSDC accession	Chromosome	Size (Mb)	GC%
LR990340.1	1	36.23	39
LR990341.1	2	35.32	39
LR990342.1	3	33.47	38.9
LR990343.1	4	33	38.6
LR990344.1	5	32.88	38.8
LR990345.1	6	32.37	38.8
LR990346.1	7	32.32	39
LR990347.1	8	31.94	38.9
LR990348.1	9	31.3	38.9
LR990349.1	10	31.13	38.9
LR990350.1	11	31.02	38.9
LR990351.1	12	30.8	38.7
LR990352.1	13	30.72	39
LR990353.1	14	30.6	38.9
LR990354.1	15	30.44	39.3
LR990355.1	16	30.28	38.7
LR990356.1	17	29.88	39
LR990357.1	18	29.73	39
LR990358.1	19	29.32	39.3
LR990359.1	20	29.05	39.2
LR990360.1	21	29.04	39.2
LR990361.1	22	28.49	39.3
LR990362.1	23	27.04	39
LR990363.1	24	24.43	39.5
LR990364.1	25	23.89	39.8
LR990365.1	26	23.42	40.8
LR990366.1	27	21.95	39.5
LR990367.1	28	21.59	39.6
LR990368.1	29	20.36	40
LR990369.1	30	19.51	40.2
LR990370.1	W	6.45	41.1
LR990339.1	Z	44.94	38.4
LR990371.2	MT	0.02	20.1
-	Unplaced	26.21	34.8

its concentration RNA 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 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), Illumina HiSeq (10X) and Illumina HiSeq 4000 (RNA-Seq) instruments. Hi-C data were generated in the Tree of Life laboratory from head tissue of ilMytImpu1 using the Qiagen kit and sequenced on an Illumina HiSeq (10X) 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), 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.

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	1.0	Uliano-Silva *et al.*, 2021
HiGlass	1.11.6	Kerpedjiev *et al.*, 2018
PretextView	0.2.x	https://github.com/wtsi-hpag/PretextView
BlobToolKit	3.0.5	Challis *et al.*, 2020

Genome annotation

The Ensembl gene annotation system (Aken *et al.*, 2016) was used to generate annotation for the *Mythimna impura* assembly (GCA_905147345.2). Annotation was created primarily through alignment of transcriptomic data to the genome, with gap filling via protein-to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 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: Mythimna impura (smoky wainscot). Accession number PRJEB42135; https://identifiers.org/ena.embl/PRJEB42135.

The genome sequence is released openly for reuse. The *M. impura* 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.6418202.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.6418156.

Members of the Wellcome Sanger Institute Tree of Life programme are listed here: https://doi.org/10.5281/zenodo.6866293.

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.6418363.

References

Aken BL, Ayling S, Barrell D, *et al.*: **The Ensembl Gene Annotation System.** *Database (Oxford).* 2016; **2016**: baw093.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

Enkhtur K, Pfeiffer M, Lkhagva A, *et al.*: **Response of Moths (Lepidoptera: Heterocera) to Livestock Grazing in Mongolian Rangelands.** *Ecological Indicators.* 2017; **72**: 667–74.
Publisher Full Text

Fox R, Dennis EB, Harrower CA, *et al.*: **The State of Britain's Larger Moths 2021.** Wareham: Butterfly Conservation, Rothamsted Research and UK Centre for Ecology & Hydrology. 2021.
Reference Source

Garrison E, Marth G: **Haplotype-Based Variant Detection from Short-Read Sequencing**. arXiv: 1207.3907, 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): 2896–98.
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

Jones HBC, Lim KS, Bell JR, *et al.*: **Quantifying Interspecific Variation in Dispersal Ability of Noctuid Moths Using an Advanced Tethered Flight Technique.** *Ecol Evol.* 2016; **6**(1): 181–90.
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

Renou M, Lucas P, Dore JC, *et al.*: **A Comparative Study of Sex Pheromone Reception in the Hadeninae (Lepidoptera: Noctuidae).** *Physiol Entomol.* 1991; **16**(1): 87–97.
PubMed Abstract | Publisher Full Text | Free Full Text

Robinson GS, Ackery PR, Kitching IJ, *et al.*: **HOSTS - a Database of the World's Lepidopteran Hostplants.** Natural History Museum. 2010.
Reference Source

Uliano-Silva M, Nunes JGF, Krasheninnikova K, *et al.*: **marcelauliano/MitoHiFi: mitohifi_v2.0.** 2021.
Publisher Full Text

UniProt Consortium: **UniProt: A Worldwide Hub of Protein Knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–15.
PubMed Abstract | Publisher Full Text | Free Full Text

Wagner DL, Fox R, Salcido DM, *et al.*: **A Window to the World of Global Insect Declines: Moth Biodiversity Trends Are Complex and Heterogeneous.** *Proc Natl Acad Sci U S A.* 2021; **118**(2): e2002549117.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✓

Version 1

Reviewer Report 27 October 2022

https://doi.org/10.21956/wellcomeopenres.20075.r52729

✓ **Francois Olivier Hebert**

Centre de Recherche du CHUM, axe neurosciences, Université de Montréal, Montréal, QC, Canada

This paper presents the genome assembly and brief protein-coding gene annotation of a female smoky wainscot specimen collected from Ant Hills region (UK). The 949Mb-long genome containing 15,441 protein-coding genes was generated using the 'hifiasm' program and resulted in a total of 132 contigs and 92 scaffolds, representing 30 autosomes and 2 sex chromosomes (W and Z chromosomes). The genome was thoroughly cleaned and well-assembled using a combination of HiFi and 10X reads, along with transcriptomic data. With a high level of completeness (assessed with BUSCO scores), this genome is of very high quality and its assembly is well described in the manuscript. The manuscript is written in clear words and sentences, straight to the point, efficient and precise. Methods are sufficiently described, with references to the exact programs used, including version numbers. Good work!

In the background section, I would maybe add another "category" of justification for the genome sequencing of this insect, along with the economic impacts. Really, it is just a detail, but surely if these insects are economically important, they are also "environmentally" or "ecologically" important species. I felt like it was important to mention both aspects: economy and biology.

Please, could the authors provide, somewhere in a table for instance (table 1?) the number of raw reads for each set of raw data (HiFi, 10X, HiSeq)?

Are the scripts available somewhere? I recognize that not everybody is comfortable with the concept of giving away scripts because they can be specifically tailored to a "bioinformatic context" and hardly reusable in other contexts (e.g., different computational resources, different platforms), but I think they are potentially useful to understand exactly how the programs were used to obtain the assembly. It's a good opportunity for other research teams to understand which parameters were used in the programs and what values they were assigned. Basically, it tells other teams how to use the programs, which sometimes can be a little bit tricky. I would suggest including either the main options used in the programs cited, or simply make the scripts available in open access (github repo?).

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/ecological 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.