DATA NOTE

The genome sequence of the Wainscot Smudge, *Ypsolopha scabrella* (Linnaeus, 1761) [version 1; peer review: 2 approved, 1 approved with reservations]

Douglas Boyes[1+], Clare Boyes[2],
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 & Hydrology, Wallingford, England, UK
[2]Independent researcher, Welshpool, Wales, UK

[+] Deceased author

v1 **First published:** 11 Aug 2023, **8**:341
https://doi.org/10.12688/wellcomeopenres.19837.1
Latest published: 11 Aug 2023, **8**:341
https://doi.org/10.12688/wellcomeopenres.19837.1

Abstract

We present a genome assembly from an individual male *Ypsolopha scabrella* (the Wainscot Smudge; Arthropoda; Insecta; Lepidoptera; Ypsolophidae). The genome sequence is 853.6 megabases in span. Most of the assembly is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled and is 16.7 kilobases in length. Gene annotation of this assembly on Ensembl identified 20,594 protein coding genes.

Keywords

Ypsolopha scabrella, Wainscot Smudge, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ? ✓ ✓

	1	2	3
version 1 11 Aug 2023	? view	✓ view	✓ view

1. **Feng Zhang**, Nanjing Agricultural University, Nanjing, China
2. **Kai Li**, Donghua University, Shanghai, China
3. **Jerome H L Hui**, The Chinese University of Hong Kong, Hong Kong, Hong Kong

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) 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, Boyes C, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Wainscot Smudge, *Ypsolopha scabrella* (Linnaeus, 1761) [version 1; peer review: 2 approved, 1 approved with reservations]** Wellcome Open Research 2023, **8**:341 https://doi.org/10.12688/wellcomeopenres.19837.1

First published: 11 Aug 2023, **8**:341 https://doi.org/10.12688/wellcomeopenres.19837.1

DATA NOTE

The genome sequence of the Wainscot Smudge, *Ypsolopha scabrella* (Linnaeus, 1761)

Douglas Boyes[1+], Clare Boyes[2],
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 & Hydrology, Wallingford, England, UK
[2]Independent researcher, Welshpool, Wales, UK

[+] Deceased author

v1 **First published:** N/A, **N/A**: N/A N/A
Latest published: N/A, **N/A**: N/A N/A

Abstract

We present a genome assembly from an individual male *Ypsolopha scabrella* (the Wainscot Smudge; Arthropoda; Insecta; Lepidoptera; Ypsolophidae). The genome sequence is 853.6 megabases in span. Most of the assembly is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled and is 16.7 kilobases in length. Gene annotation of this assembly on Ensembl identified 20,594 protein coding genes.

Keywords

Ypsolopha scabrella, Wainscot Smudge, 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) 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, Boyes C, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Wainscot Smudge, *Ypsolopha scabrella* (Linnaeus, 1761)** Wellcome Open Research , : https://doi.org/

First published: N/A, **N/A**: N/A N/A

Species taxonomy

Eukaryota; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Amphiesmenoptera; Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Yponomeutoidea; Ypsolophidae; *Ypsolopha*; *Ypsolopha scabrella* (Linnaeus, 1761) (NCBI:txid1870435).

Background

Ypsolopha scabrella (Wainscot Smudge) is a common micro-moth in the family Ypsolophidae. The species has a southerly distribution in Britain and is found throughout Europe apart from Portugal and Greece (GBIF Secretariat, 2022).

Y. scabrella has one generation a year and flies between June and October. It readily comes to light and is found in woodland, scrub and gardens (Sterling *et al.*, 2012). The small (wingspan 15–21 mm) adult moth rests with its wings curled around its body. The forewing colour is whitish, with pale and dark brown streaks. There are three tufts of raised, darkened scales along the back (Emmet, 1996). The egg is usually laid on hawthorn or apple, but occasionally on cotoneaster where it overwinters. The larvae feed in a rather insignificant web and pupate during June and July in a boat-shaped cocoon on the ground (Langmaid *et al.*, 2018).

A genome sequence from *Y. scabrella* will be useful for comparative studies across the Lepidoptera. The genome of *Y. scabrella* 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 *Y. scabrella* based on a male specimen from Wytham Woods, Oxfordshire, UK.

Genome sequence report

The genome was sequenced from one male *Ypsolopha scabrella* (Figure 1) collected from Wytham Woods, Oxfordshire, UK (51.77, –1.34). A total of 36-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 40-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 35 missing joins or misjoins, reducing the scaffold number by 38.46%, and increasing the scaffold N50 by 1.37%.

Figure 1. Photograph of the *Ypsolopha scabrella* (ilYpsScab1) specimen used for genome sequencing.

The final assembly has a total length of 853.6 Mb in 40 sequence scaffolds with a scaffold N50 of 29.9 Mb (Table 1). Most (99.97%) of the assembly sequence was assigned to 31 chromosomal-level scaffolds, representing 30 autosomes and the Z sex chromosome. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 2–Figure 5; Table 2). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial genome was also assembled and can be found as a contig within the multifasta file of the genome submission.

The estimated Quality Value (QV) of the final assembly is 58.8 with *k*-mer completeness of 99.99%, and the assembly has a BUSCO v5.3.2 completeness of 98.1% (single = 96.4%, duplicated = 1.6%), using the lepidoptera_odb10 reference set (n = 5,286).

Metadata for specimens, spectral estimates, sequencing runs, contaminants and pre-curation assembly statistics can be found at https://links.tol.sanger.ac.uk/species/1870435.

Genome annotation report

The *Ypsolopha scabrella* genome assembly (GCA_910592155.1) was annotated using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Ypsolopha_scabrella_GCA_910592155.1/Info/Index). The resulting annotation includes 20,761 transcribed mRNAs from 20,594 protein-coding and $NCG non-coding genes.

Methods

Sample acquisition and nucleic acid extraction

Two *Ypsolopha scabrella* specimens were collected from Wytham Woods, Oxfordshire (biological vice-county Berkshire), UK (latitude 51.77, longitude –1.34) on 2020-07-20 using a light trap. The specimens were collected and identified by Douglas Boyes (University of Oxford) and preserved on dry ice. The specimen used for genome sequencing was specimen ID Ox000642, ToLID ilYpsScab1, while the specimen used for Hi-C scaffolding was specimen ID Ox000643, ToLID ilYpsScab2.

DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute (WSI). The ilYpsScab1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. Tissue from the whole organism was disrupted using a Nippi Powermasher fitted with a BioMasher pestle. 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 the 0.8X AMpure XP purification kit prior to 10X Chromium sequencing; a minimum of 50 ng DNA was submitted for 10X

Table 1. Genome data for *Ypsolopha scabrella*, ilYpsScab1.1.

Project accession data		
Assembly identifier	ilYpsScab1.1	
Species	*Ypsolopha scabrella*	
Specimen	ilYpsScab1	
NCBI taxonomy ID	1870435	
BioProject	PRJEB45184	
BioSample ID	SAMEA7701504	
Isolate information	ilYpsScab1, male: whole organism (DNA sequencing) ilYpsScab2, female: whole organism (Hi-C scaffolding)	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	58.8	*≥ 50*
k-mer completeness	99.99%	*≥ 95%*
BUSCO**	C:98.1%[S:96.4%,D:1.6%], F:0.5%,M:1.5%,n:5,286	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	99.97%	*≥ 95%*
Sex chromosomes	Z chromosome	*localised homologous pairs*
Organelles	Mitochondrial genome assembled	*complete single alleles*
Raw data accessions		
PacificBiosciences SEQUEL II	ERR6454731, ERR6454732	
10X Genomics Illumina	ERR6054910, ERR6054912, ERR6054909, ERR6054911	
Hi-C Illumina	ERR6054913	
Genome assembly		
Assembly accession	GCA_910592155.1	
Accession of alternate haplotype	GCA_910591985.1	
Span (Mb)	853.6	
Number of contigs	91	
Contig N50 length (Mb)	26.5	
Number of scaffolds	40	
Scaffold N50 length (Mb)	29.9	
Longest scaffold (Mb)	52.4	
Genome annotation		
Number of protein-coding genes	20,594	
Number of gene transcripts	20,761	

* 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/ilYpsScab1.1/dataset/CAJUZF01.1/busco.

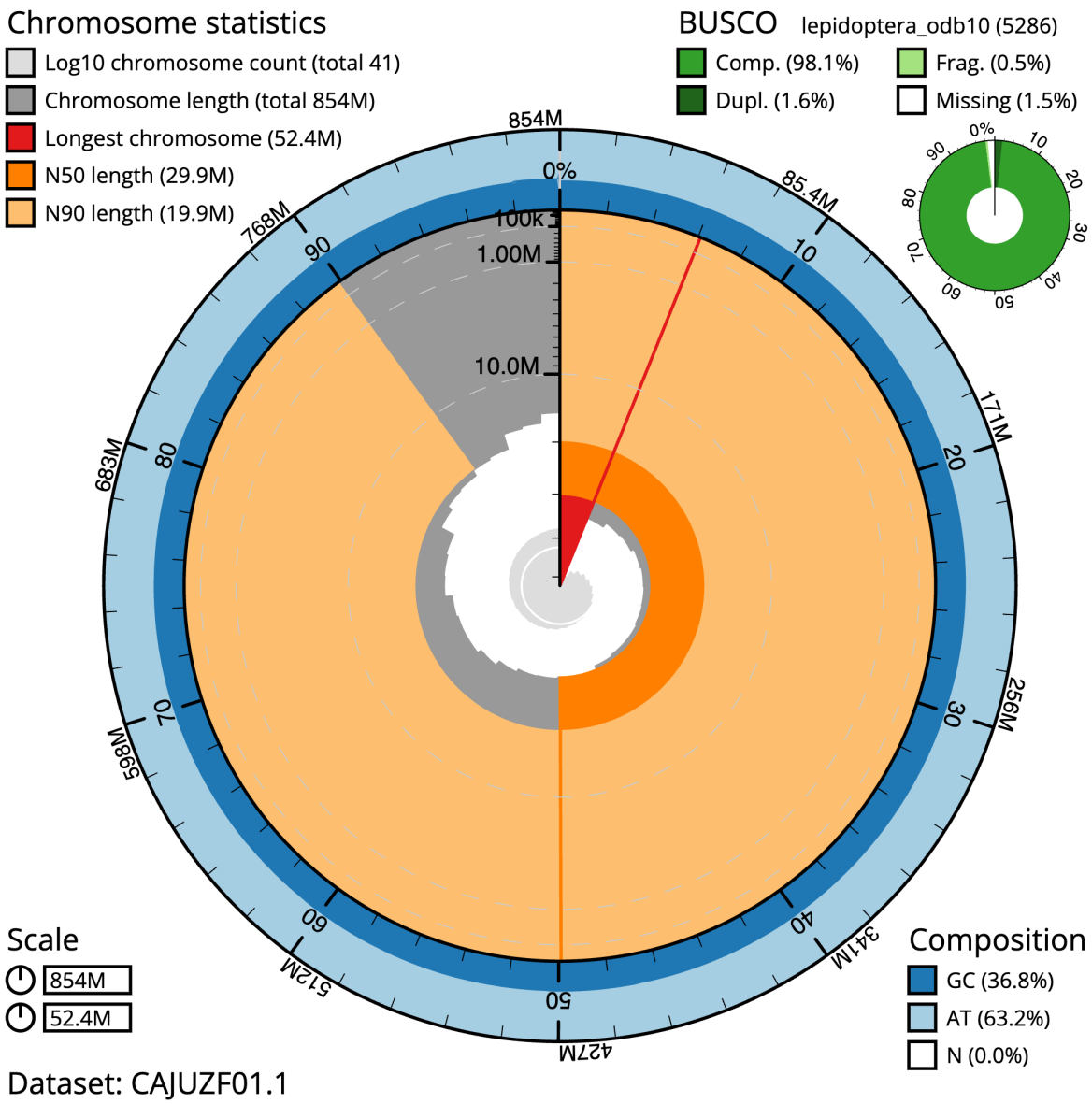

Figure 2. Genome assembly of *Ypsolopha scabrella*, ilYpsScab1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 853,595,150 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (52,420,632 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (29,858,840 and 19,866,168 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/ilYpsScab1.1/dataset/CAJUZF01.1/snail.

sequencing. HMW 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.

Sequencing

Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud DNA sequencing libraries were constructed according to the manufacturers' instructions. DNA sequencing was performed by the Scientific Operations core at the WSI on

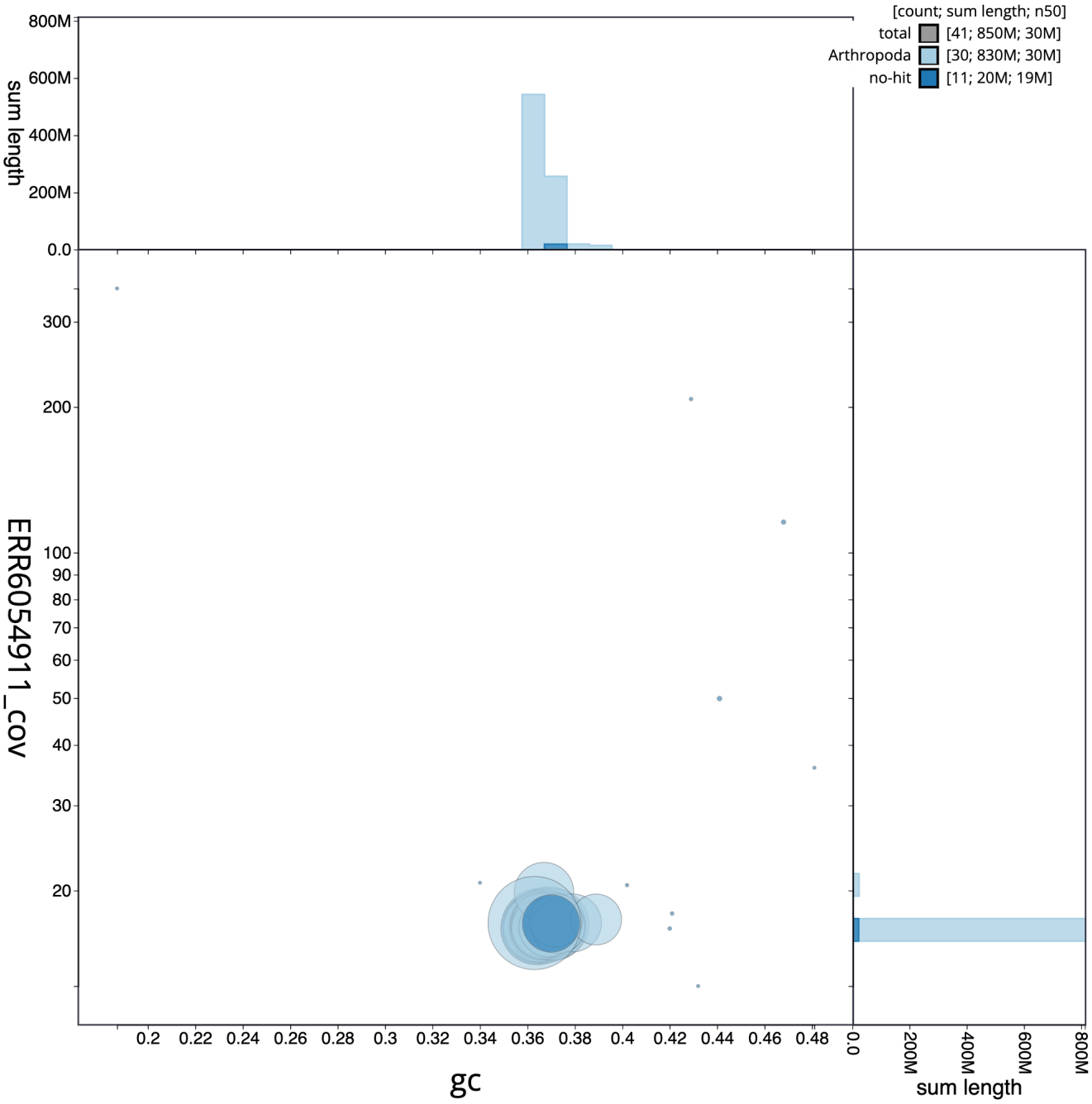

Figure 3. Genome assembly of *Ypsolopha scabrella*, ilYpsScab1.1: BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilYpsScab1.1/dataset/CAJUZF01.1/blob.

Pacific Biosciences SEQUEL II (HiFi) and Illumina NovaSeq 6000 (10X) instruments. Hi-C data were also generated from whole organism tissue of ilYpsScab2 using the Arima2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021) and 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 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 contamination and corrected using the gEVAL system (Chow *et al.*, 2016) as described previously (Howe *et al.*, 2021).

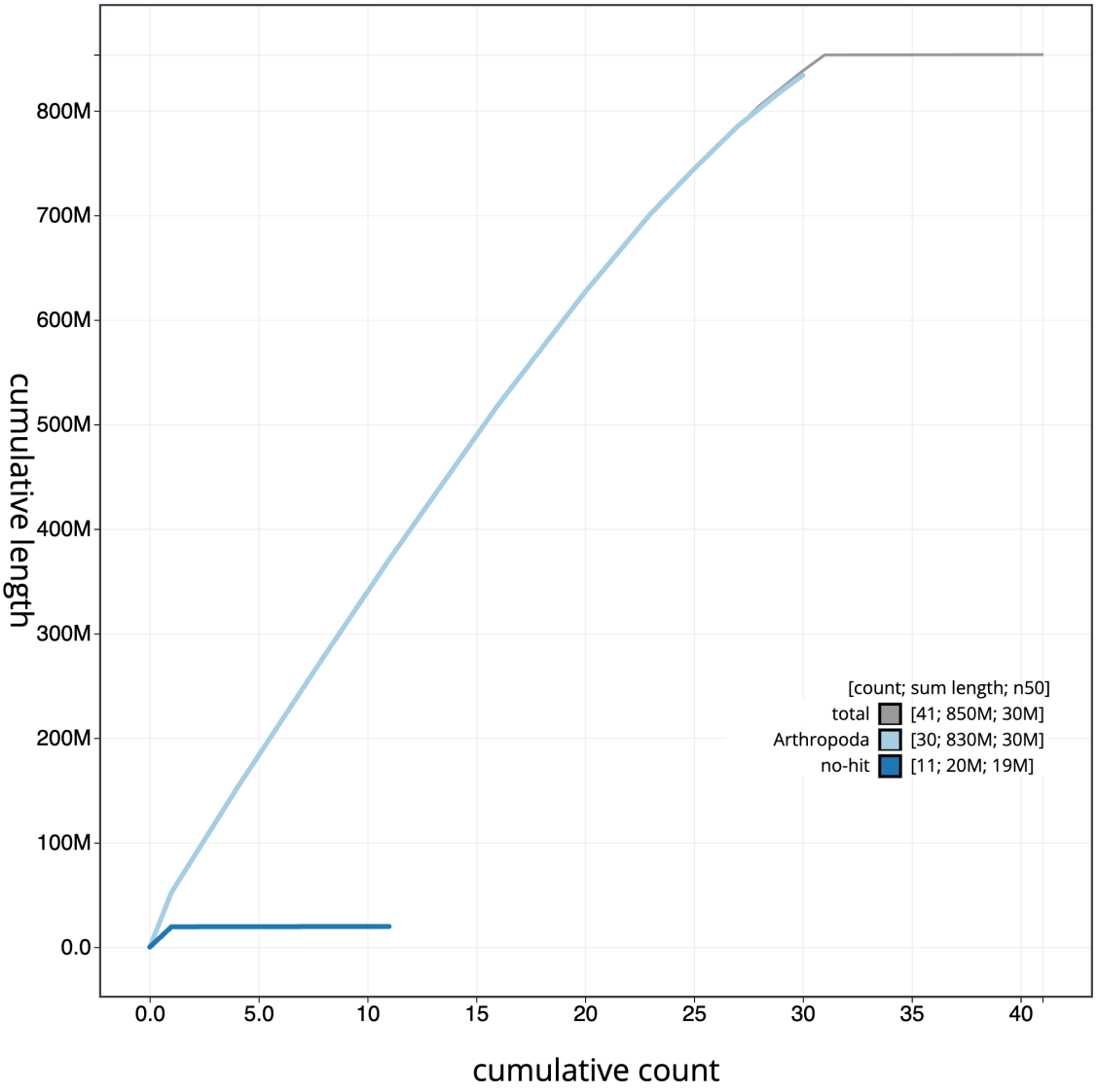

Figure 4. Genome assembly of *Ypsolopha scabrella*, ilYpsScab1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilYpsScab1.1/dataset/CAJUZF01.1/cumulative.

Manual curation was performed using gEVAL, HiGlass (Kerpedjiev *et al.*, 2018) and Pretext (Harry, 2022). The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) or MITOS (Bernt *et al.*, 2013) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). To assess the assembly metrics, the *k*-mer completeness and QV consensus quality values were calculated in Merqury (Rhie *et al.*, 2020). This work was done using Nextflow (Di Tommaso *et al.*, 2017) DSL2 pipelines "sanger-tol/readmapping" (Surana *et al.*, 2023a) and "sanger-tol/genomenote" (Surana *et al.*, 2023b). The genome was analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021; Simão *et al.*, 2015) were calculated.

Table 3 contains a list of relevant software tool versions and sources.

Genome annotation

The BRAKER2 pipeline (Brůna *et al.*, 2021) was used in the default protein mode to generate annotation for the *Ypsolopha scabrella* assembly (GCA_910592155.1) in Ensembl Rapid Release.

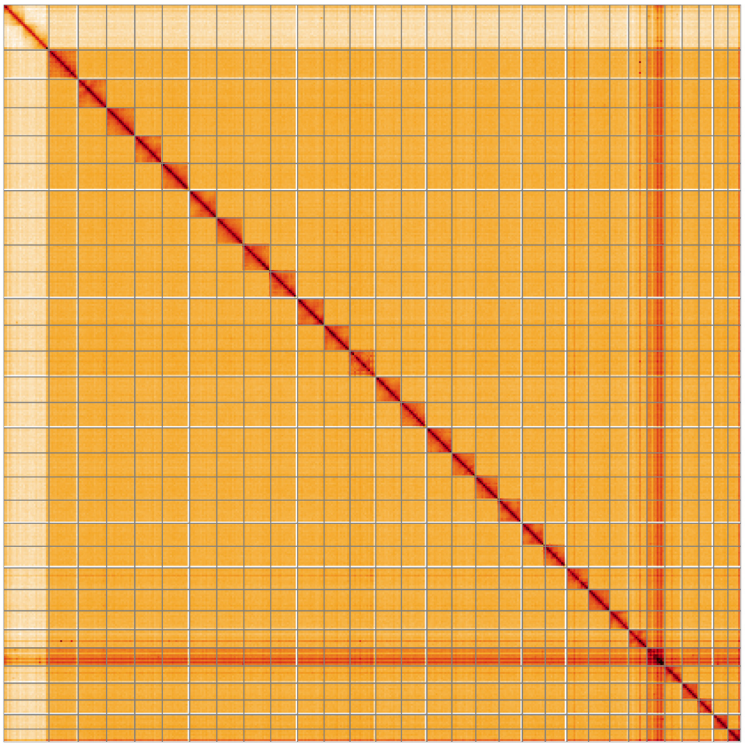

Figure 5. Genome assembly of *Ypsolopha scabrella*, ilYpsScab1.1: Hi-C contact map of the ilYpsScab1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=MEkHt8scQR6xmc-6DNy3Gg.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Ypsolopha scabrella*, ilYpsScab1.

INSDC accession	Chromosome name	Length (MB)	GC Percent
OU342961.1	1	33.96	36.5
OU342962.1	2	32.86	36.5
OU342963.1	3	32.56	36.5
OU342964.1	4	31.86	36
OU342965.1	5	31.59	36.5
OU342966.1	6	31.45	36.5
OU342967.1	7	31.33	36.5
OU342968.1	8	31.04	36
OU342969.1	9	31.02	36
OU342970.1	10	30.74	36.5
OU342971.1	11	29.93	36.5
OU342972.1	12	29.86	37
OU342973.1	13	29.63	36.5
OU342974.1	14	29.51	36.5
OU342975.1	15	28.82	36.5
OU342976.1	16	27.76	36
OU342977.1	17	26.93	36.5
OU342978.1	18	26.87	36.5
OU342979.1	19	26.54	36.5
OU342980.1	20	25.07	36.5
OU342981.1	21	24.93	37
OU342982.1	22	24.6	36.5
OU342983.1	23	21.95	36.5
OU342984.1	24	21.05	36.5
OU342985.1	25	20.85	36.5
OU342986.1	26	19.87	37.5
OU342987.1	27	19.5	37
OU342988.1	28	16.98	37
OU342989.1	29	16.64	37
OU342990.1	30	15.21	38.5
OU342960.1	Z	52.42	36
OU342991.1	MT	0.2	

Table 3. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	4.0.7	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.3.2	https://gitlab.com/ezlab/busco
FreeBayes	1.3.1-17-gaa2ace8	https://github.com/freebayes/freebayes
gEVAL	N/A	https://geval.org.uk/
Hifiasm	0.15.1-r328	https://github.com/chhylp123/hifiasm
HiGlass	1.11.6	https://github.com/higlass/higlass
Long Ranger ALIGN	2.2.2	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines
Merqury	MerquryFK	https://github.com/thegenemyers/MERQURY.FK
MitoHiFi	2	https://github.com/marcelauliano/MitoHiFi
PretextView	0.2	https://github.com/wtsi-hpag/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
SALSA	2.2	https://github.com/salsa-rs/salsa
sanger-tol/genomenote	v1.0	https://github.com/sanger-tol/genomenote
sanger-tol/readmapping	1.1.0	https://github.com/sanger-tol/readmapping/tree/1.1.0

Wellcome Sanger Institute – Legal and Governance

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'**, which can be found in full on the Darwin Tree of Life website here. 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.

Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible. The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international)

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: *Ypsolopha scabrella* (wainscot smudge). Accession number PRJEB45184; https://identifiers.org/ena.embl/PRJEB45184. (Wellcome Sanger Institute, 2021)

The genome sequence is released openly for reuse. The *Ypsolopha scabrella* 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

Abdennur N, Mirny LA: **Cooler: Scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316.
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

Bernt M, Donath A, Jühling F, *et al.*: **MITOS: Improved *de novo* metazoan mitochondrial genome annotation.** *Mol Phylogenet Evol.* 2013; **69**(2): 313–319.
PubMed Abstract | 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

Chow W, Brugger K, Caccamo M, *et al.*: **gEVAL — a web-based browser for evaluating genome assemblies.** *Bioinformatics.* 2016; **32**(16): 2508–2510.
PubMed Abstract | Publisher Full Text | Free Full Text

Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.
PubMed Abstract | Publisher Full Text

Emmet AM: **The Moths and Butterflies of Great Britain and Ireland - Yponomeutidae - Elachistidae.** Colchester: Harley Books, 1996.

Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing.** arXiv: 1207.3907v2, 2012; [Accessed 26 July 2023].
Publisher Full Text

GBIF Secretariat: ***Ypsolopha scabrella* (Linnaeus, 1761).** *GBIF Backbone Taxonomy.* 2022; [Accessed 12 July 2023].
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): 2896–2898.
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

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

Langmaid JR, Palmer S, Young MR: **A Field Guide to the Smaller Moths of Great Britain and Ireland.** 3rd ed. British Entomological and Natural History Society, 2018.
Reference Source

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

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

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–3212.
PubMed Abstract | Publisher Full Text

Sterling P, Parsons M, Lewington R: **Field Guide to the Micro Moths of Great Britain and Ireland**. Gillingham, Dorset: British Wildlife Publishing, 2012.
Reference Source

Surana P, Muffato M, Qi G: **sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0)**. Zenodo. 2023a.
Publisher Full Text

Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev)**. Zenodo. 2023b.
Publisher Full Text

Uliano-Silva M, Ferreira GJRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio High Fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.
PubMed Abstract | Publisher Full Text | Free Full Text

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems**. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2019; 314–324.
Publisher Full Text

Wellcome Sanger Institute: **The genome sequence of the Wainscot Smudge, *Ypsolopha scabrella* (Linnaeus, 1761).** European Nucleotide Archive, [dataset], accession number PRJEB45184, 2021.

Open Peer Review

Current Peer Review Status: ? ✓ ✓

Version 1

Reviewer Report 15 July 2024

https://doi.org/10.21956/wellcomeopenres.21969.r86877

✓ **Jerome H L Hui**
The Chinese University of Hong Kong, Hong Kong, Hong Kong

In this Data Note, Boyes and colleagues sequenced and assembled the genome of moth *Ypsolopha scabrella* (Linnaeus, 1761), commonly known as wainscot hooktip or wainscot smudge. According to the UKmoths and iNaturalist, this species can be found in both England and Wales. Molecular data of this species are scarce prior to this report, and are mainly mitochondrial cytochrome oxidase subunit 1 (COI) gene sequences deposited to the NCBI database. Therefore, this new genome resource will be useful for further studies, ranging from understanding the effect and impact of climate change on them, revealing their population structures, to understanding their evolution with other insects.

This genome resource is excellent from the summary statistics, with high BUSCO number scores, high sequence continuity, and majority of sequences contained on the 30 pseudochromosomes (plus sex chromosome and mitochondrion). All in all, this is another valuable contribution.

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: I have published with Peter Holland more than three years ago, and confirm that this potential conflict of interest did not affect my ability to write an objective and unbiased

review of the article.

Reviewer Expertise: Genomics, evolution, invertebrates

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 15 July 2024

https://doi.org/10.21956/wellcomeopenres.21969.r74112

Kai Li
Donghua University, Shanghai, China

The genome sequence obtained from *Y. scabrella* will serve as a valuable resource for conducting comparative studies within the Lepidoptera order. The author provides a relatively clear description of sample acquisition, DNA library preparation, high-throughput sequencing, bioinformatics, and other related aspects. Therefore, judging from the key points of the journal review, it is a qualified study, and the relevant sequences are helpful for other researchers.

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: Bioinformatics; Insect Molecular Ecology.

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 15 July 2024

https://doi.org/10.21956/wellcomeopenres.21969.r86883

?

Feng Zhang
Nanjing Agricultural University, Nanjing, China

This manuscript presents a high-quality genome assembly but a low-quality annotation. See detailed comments below:

- "The genome was sequenced from one male *Ypsolopha scabrella* (Figure 1) collected from Wytham Woods, Oxford-shire, UK (51.77, –1.34)."
- Add the 'E, N' and measurement unit for the longitude and latitude.
- "The resulting annotation includes 20,761 transcribed mRNAs from 20,594 protein-coding and $NCG non-coding genes." Explain "$NCG" in detail.
- I don't think that it is a good idea to annotate the insect genome using BRAKER2, which usually generate more gene models than other pipelines. Those models are often of shorter gene length (mean < 500 aa) and lower BUSCO completeness. In this case, the gene number 20,594 is much higher than those of most lepidopteran published genomes, including world-wide pests. In addition, it doesn't make much sense to annotate the genome while the RNA-seq data is lacking.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Partly

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: Insect genomics and 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.