DATA NOTE

The genome sequence of the Clifden nonpareil, *Catocala fraxini* (Linnaeus, 1758) [version 1; peer review: 1 approved with reservations]

Douglas Boyes[1+], Peter W.H. Holland[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 and Hydrology, Wallingford, Oxfordshire, UK
[2]Department of Zoology, University of Oxford, Oxford, UK

[+] Deceased author

v1 **First published:** 05 Apr 2022, **7**:129
https://doi.org/10.12688/wellcomeopenres.17818.1
Latest published: 05 Apr 2022, **7**:129
https://doi.org/10.12688/wellcomeopenres.17818.1

Abstract

We present a genome assembly from an individual male *Catocala fraxini* (the Clifden nonpareil; Arthropoda; Insecta; Lepidoptera; Erebidae). The genome sequence is 781 megabases in span. The majority of the assembly (99.99%) is scaffolded into 31 chromosomal pseudomolecules, with the Z sex chromosome assembled. The mitochondrial genome was also assembled, and is 15.6 kilobases in length.

Keywords

Catocala fraxini, Clifden nonpareil, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ?

	1
version 1 05 Apr 2022	? view

1. **Surya Saha**, Boyce Thompson Institute for Plant Research, Ithaca, 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; **Holland PWH**: Supervision, Writing – Original Draft Preparation;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Wellcome Trust 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, Holland PWH, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Clifden nonpareil, *Catocala fraxini* (Linnaeus, 1758) [version 1; peer review: 1 approved with reservations]** Wellcome Open Research 2022, **7**:129 https://doi.org/10.12688/wellcomeopenres.17818.1

First published: 05 Apr 2022, **7**:129 https://doi.org/10.12688/wellcomeopenres.17818.1

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Noctuoidea; Erebidae; Erebinae; Catocala; *Catocala fraxini* (Linnaeus, 1758) (NCBI:txid423510).

Background

Catocala fraxini (Clifden nonpareil or blue underwing) is a large noctuid moth (75-100 mm wingspan) with zigzag markings on broad silver-grey forewings and striking lilac blue flashes across the hindwings. The moth has been recorded across much of the Palaearctic, from central and northern Europe to Russia, Japan and Korea, primarily in dense woodlands containing stands of its larval food plant aspen, *Populus tremula*. In the UK, *C. fraxini* was first recorded in the 1740s at 'Cleifden', Buckinghamshire (now Cliveden, famous for the Profumo scandal); the common name derives from the location plus the French 'non pareil' meaning 'without equal' (Wilkes, 1749).

Through the 1800s and early 1900s, *C. fraxini* was considered an extreme rarity in the UK with small populations in southern and eastern counties, until it became locally extinct by the 1960s. The moth achieved near mythical status amongst entomologists of the time; P.B.M. Allan was obsessed with the species and wrote that if he ever succeeded in catching the moth he would "go to Fortnum & Mason's and buy rare syrups and syllabubs for it" (Allan, 1947). Sporadic records after 1960 were attributed to occasional dispersing individuals reaching the UK from eastern Europe and Scandinavia (Waring *et al.,* 2003). Since 2010, however, there has been a dramatic increase in UK records and the species is now thought to be breeding in woodlands in the south of England (Randle *et al.,* 2019).

In northern Europe *C. fraxini* has a single generation per year, overwintering as an egg before larval development from April to July. Late instar larvae have a distinctive fringe of hairs that breaks their outline such that they are effectively camouflaged on woody twigs of the host plant. After pupation, adults emerge around September and have a short flight period; they are strictly nocturnal and can be attracted to sugary substances and to light.

The genome of *C. fraxini* 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 *C. fraxini*, based on one male specimen from Wytham Woods, Oxfordshire, UK.

Genome sequence report

The genome was sequenced from one male *C. fraxini* (Figure 1) collected from Wytham Woods, Oxfordshire (biological vice-county: Berkshire), UK (latitude 51.772, longitude -1.338). A total of 41-fold coverage in Pacific Biosciences single-molecule long reads and 107-fold coverage in 10X Genomics read clouds

Figure 1. Image of the *Catocala fraxini* (ilCatFrax1) specimen taken prior to preservation and processing.

were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 12 missing/misjoins, reducing the scaffold number by 23.26%.

The final assembly has a total length of 781 Mb in 33 sequence scaffolds with a scaffold N50 of 27.8 Mb (Table 1). The majority of the assembly sequence (99.99%) was assigned to 31 chromosomal-level scaffolds, representing 30 autosomes (numbered by sequence length), and the Z sex chromosome (Figure 2–Figure 5; Table 2). The assembly has a BUSCO v5.2.2 (Manni *et al.,* 2021) completeness of 99.0% (single 98.2%, 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.

Methods

Sample acquisition and DNA extraction

A single male *C. fraxini* (ilCatFrax1) 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 at the Tree of Life laboratory, Wellcome Sanger Institute. The ilCatFrax1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. Thorax tissue was cryogenically disrupted to a fine powder using a Covaris cryoPREP Automated Dry Pulveriser, receiving multiple impacts. 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

Table 1. Genome data for *Catocala fraxini*, ilCatFrax1.1.

Project accession data	
Assembly identifier	ilCatFrax1.1
Species	*Catocala fraxini*
Specimen	ilCatFrax1
NCBI taxonomy ID	NCBI:txid423510
BioProject	PRJEB50461
BioSample ID	SAMEA8603175
Isolate information	Male, thorax (genome assembly), abdomen (RNA-Seq), head (Hi-C)
Raw data accessions	
PacificBiosciences SEQUEL II	ERR8482049-ERR8482050
10X Genomics Illumina	ERR8373760-ERR8373763
Hi-C Illumina	ERR8373759
PolyA RNA-Seq Illumina	ERR8373764
Genome assembly	
Assembly accession	GCA_930367265.1
Accession of alternate haplotype	GCA_930367255.1
Span (Mb)	781
Number of contigs	47
Contig N50 length (Mb)	26.8
Number of scaffolds	33
Scaffold N50 length (Mb)	27.8
Longest scaffold (Mb)	31.7
BUSCO* genome score	C:99.0%[S:98.2%,D:0.8%],F:0.2%,M:0.9%,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/ilCatFrax1.1/dataset/CAKNFB01/busco.

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 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 head tissue of ilCatFrax1 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 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

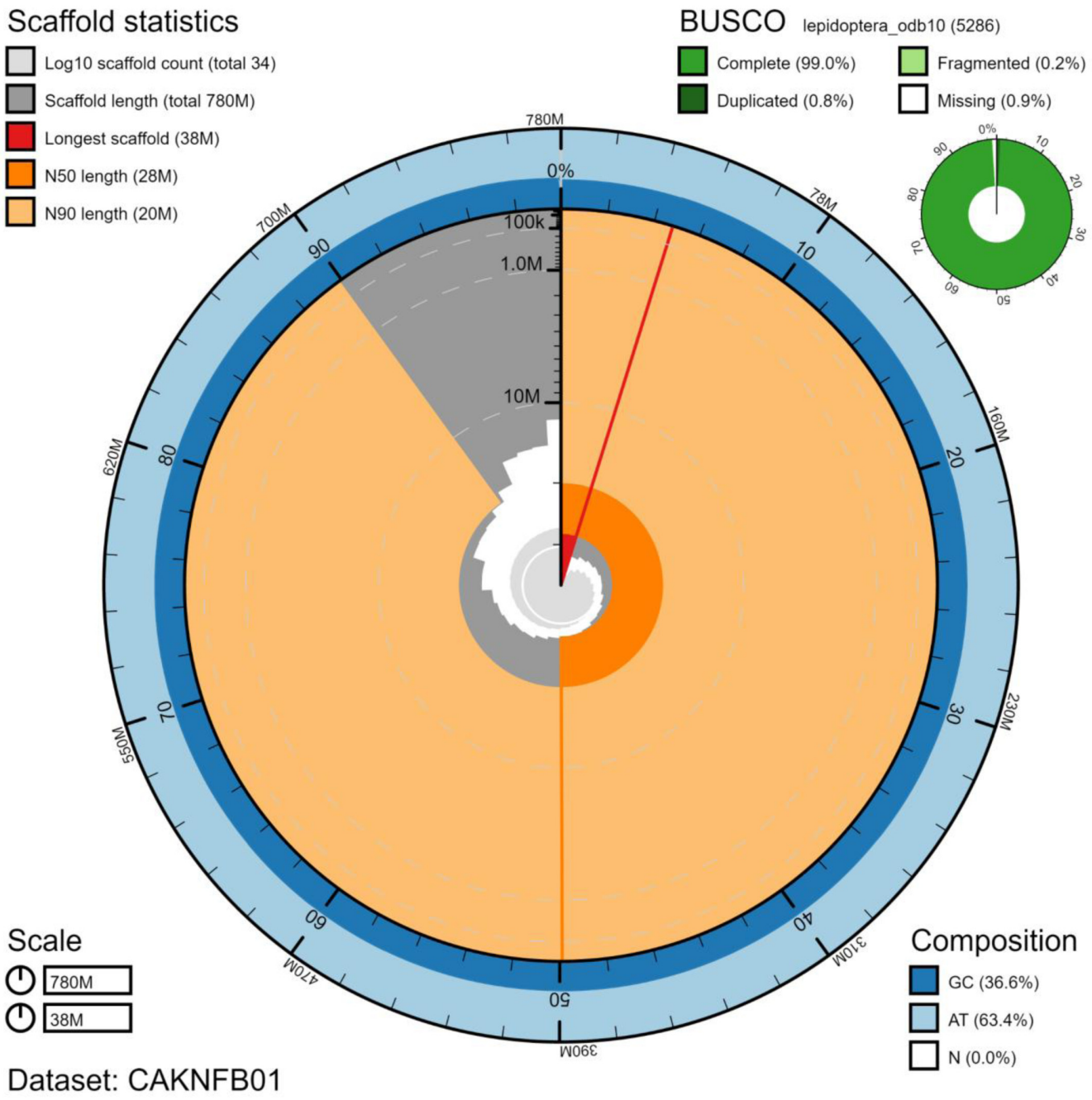

Figure 2. Genome assembly of *Catocala fraxini*, ilCatFrax1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 780,602,067 bp assembly. The distribution of chromosome lengths is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly (37,687,137 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (27,836,421 and 20,022,594 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/ilCatFrax1.1/dataset/CAKNFB01/snail.

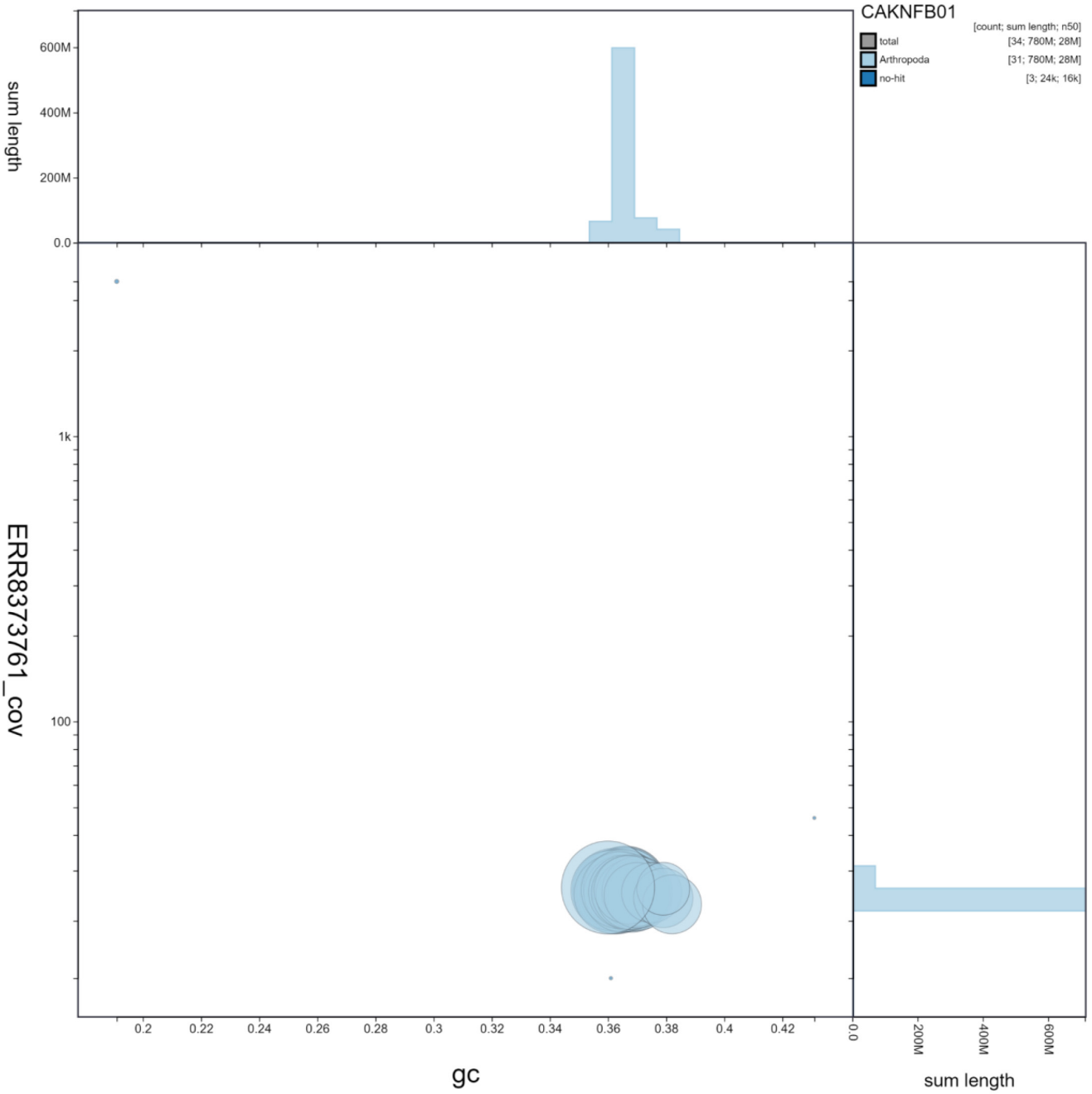

Figure 3. Genome assembly of *Catocala fraxini*, ilCatFrax1.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilCatFrax1.1/dataset/CAKNFB01/blob.

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 NovaSeq 6000 (10X) and Illumina HiSeq 4000 (RNA-Seq) instruments. Hi-C data were generated from head tissue using the Arima Hi-C+ kit and sequenced on NovaSeq 6000.

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

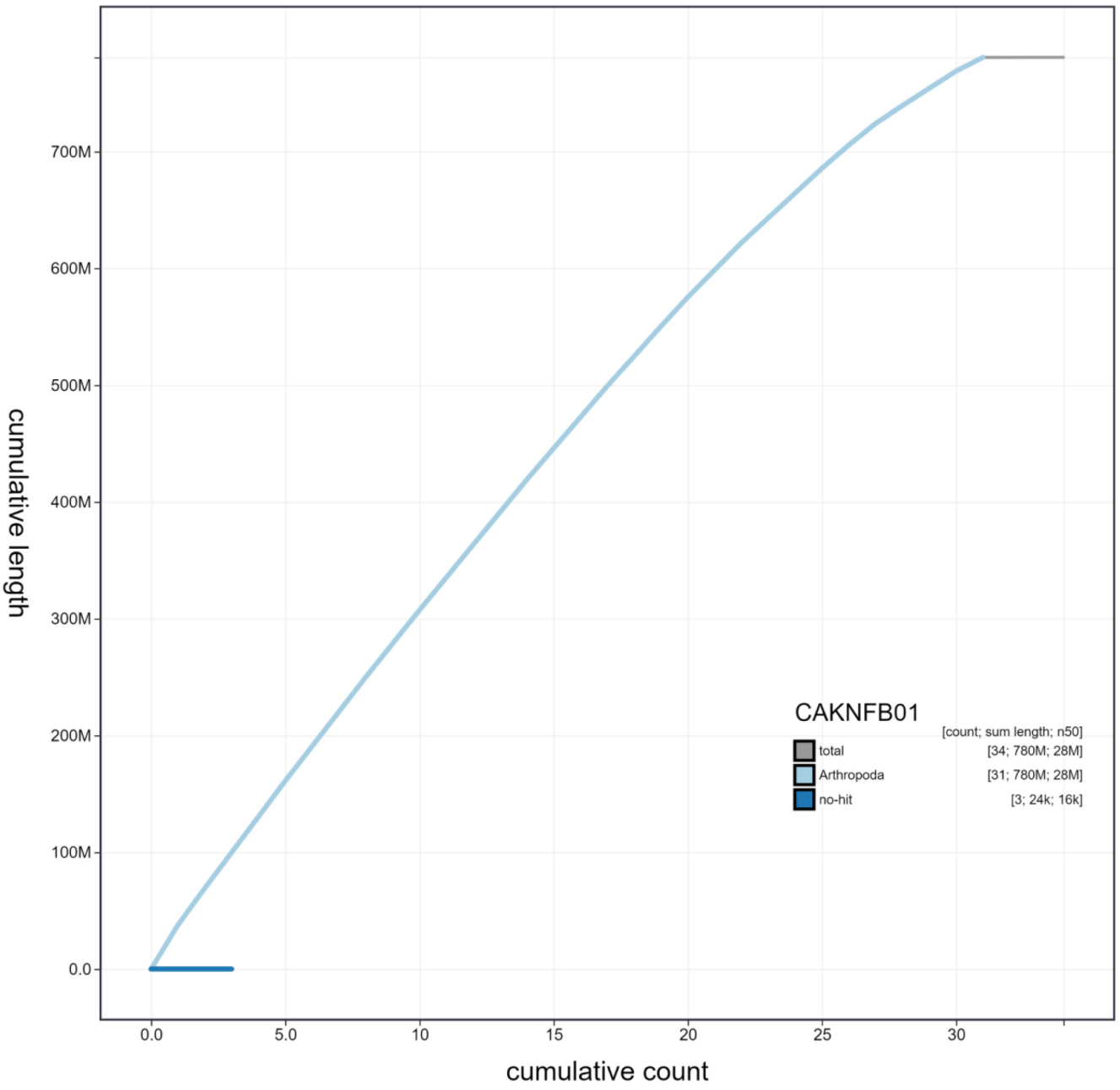

Figure 4. Genome assembly of *Catocala fraxini*, ilCatFrax1.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilCatFrax1.1/dataset/CAKNFB01/cumulative.

(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 (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.

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

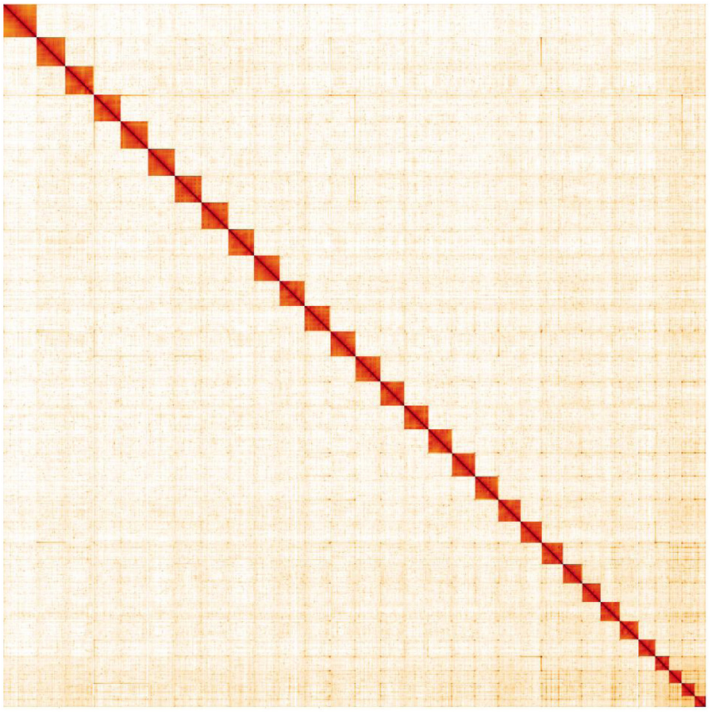

Figure 5. Genome assembly of *Catocala fraxini*, ilCatFrax1.1: Hi-C contact map. Hi-C contact map of the ilCatFrax1.1 assembly, visualised in HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this map is available here.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Catocala fraxini*, ilCatFrax1.1.

INSDC accession	Chromosome	Size (Mb)	GC%
OV884025.1	1	31.69	36.5
OV884026.1	2	31.05	36.6
OV884027.1	3	30.58	36.4
OV884028.1	4	30.14	36.2
OV884029.1	5	29.81	36.6
OV884030.1	6	29.74	36.4
OV884031.1	7	29.27	36.3
OV884032.1	8	28.82	36.6
OV884033.1	9	28.63	36.3
OV884034.1	10	28.17	36.2
OV884035.1	11	27.97	36.3
OV884036.1	12	27.84	36.4
OV884037.1	13	27.64	36.1
OV884038.1	14	27.17	36.2
OV884039.1	15	26.59	36.5
OV884040.1	16	26.40	36.8
OV884041.1	17	25.72	36.5
OV884042.1	18	25.44	36.4
OV884043.1	19	24.95	36.8
OV884044.1	20	23.33	36.6
OV884045.1	21	23.28	36.7
OV884046.1	22	21.38	37.0
OV884047.1	23	21.28	36.8
OV884048.1	24	21.04	37.0
OV884049.1	25	20.02	36.7
OV884050.1	26	18.44	37.0
OV884051.1	27	15.26	37.5
OV884052.1	28	14.85	37.9
OV884053.1	29	14.70	38.2
OV884054.1	30	11.69	37.9
OV884024.1	Z	37.69	36.0
OV884055.1	MT	0.02	19.2
-	Unplaced	0.01	39.1

Table 3. Software tools used.

Software tool	Version	Source
Hifiasm	0.15.3	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	2.0	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

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: Catocala fraxini (Cliften non-pareil). Accession number PRJEB50461; https://identifiers.org/ena.embl/PRJEB50461.

The genome sequence is released openly for reuse. The *C. fraxini* 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 using the RNA-Seq data 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

Allan PBM: **A Moth Hunter's Gossip (Watkins & Doncaster).** *Moths and Memories"(Watkins & Doncaster).* 1947.

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): 1361–1374.
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): 170–75.
PubMed Abstract | Publisher Full Text | Free Full Text

Garrison E, Marth G: **Haplotype-Based Variant Detection from Short-Read Sequencing, July**, 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): 2896–2898.
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): 4647–4654.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

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

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.** *Cell.* 2014; **159**(7): 1665–80.
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

Waring P, Townsend M, Lewington R: **Field Guide to the Moths of Great Britain and Ireland**. British Wildlife Publishing, Hampshire. 2003.

Wilkes B: **The English Moths and Butterflies: Together with the Plants, Flowers, and Fruits Whereon They Feed, and Are Usually Found. All Drawn and Coloured in Such a Manner, as to Represent Their Several Beautiful Appearances ... Together with an Attempt Towards a Natural History of the Said Moths and Butterflies**. Printed for and sold by B. Wilkes. 1749.

Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 13 January 2023

https://doi.org/10.21956/wellcomeopenres.19727.r53699

?

Surya Saha
Boyce Thompson Institute for Plant Research, Ithaca, NY, USA

The manuscript is well-written and present a high quality and contiguous *Catocala fraxini* (Cliften non-pareil) genome assembly with DNA sourced from a single individual which is the gold standard for genome assembly. The NCBI Bioproject is well structured and although this is not a completely phased assembly, provides access to both the primary and alternate haplotype assemblies.

The rapid release of data for use by the community is commendable and the Wellcome Open Research Data Notes are an excellent practice adopted by the DToL project.

The Bioproject mentioned in the paper is well documented
https://www.ncbi.nlm.nih.gov/bioproject/PRJEB50461

But I was confused by the Bioprojects linked from the primary and alternate assembly pages

- https://www.ncbi.nlm.nih.gov/assembly/GCA_930367265.1/ ->
  https://www.ncbi.nlm.nih.gov/bioproject/PRJEB50839/

- https://www.ncbi.nlm.nih.gov/assembly/GCA_930367255.1 ->
  https://www.ncbi.nlm.nih.gov/bioproject/PRJEB50840/

It is possible that this is an artifact of the submission system. If not, these should both link back to https://www.ncbi.nlm.nih.gov/bioproject/PRJEB50461.

I only have minor comments:

Although the versions and tools are mentioned, the methods lack enough details for reasonable reproducibility. Parameters used for execution (even if they were the defaults) need to be reported for readers. Sharing a script with commands to reproduce the major steps in the assembly (using a github repository release with a zenodo DOI) or adding to an additional column in Table 3 might be a solution worth considering.

I could not find the HiC read coverage in the Data Note.

Figures 3 and 4 don't provide a lot of insights for genomes of this quality. Using a kmer mining toolkit like KAT (https://github.com/TGAC/KAT) can help uncover the heterozygosity and repetitiveness in the assembly considering this is not a completely phased assembly.

More details of the polishing process will be helpful. I assume you applied bcftools consensus on the filtered Freebayes VCF files. What was the improvement in quality after the polishing? Was a metric like QV-score used to evaluate this?

Was RepeatMasker run on the assembly? A summary table of repeat elements can be helpful for understanding the architecture of this genome.

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?
Partly

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: assembly, annotation, hemipteran 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.