Check for updates

DATA NOTE

The genome sequence of the heath fritillary, *Melitaea athalia* (Rottemburg, 1775) [version 1; peer review: 2 approved]

Alex Hayward[1], Roger Vila [2], Dominik R. Laetsch [3], Konrad Lohse[3],
Tobias Baril [1], 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]University of Exeter, Penryn, UK
[2]Institut de Biologia Evolutiva (CSIC - Universitat Pompeu Fabra), Barcelona, Spain
[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK

Abstract
We present a genome assembly from an individual female *Melitaea athalia* (also known as *Mellicta athalia*; the heath fritillary; Arthropoda; Insecta; Lepidoptera; Nymphalidae). The genome sequence is 610 megabases in span. In total, 99.98% of the assembly is scaffolded into 32 chromosomal pseudomolecules, with the W and Z sex chromosome assembled. Gene annotation of this assembly on Ensembl has identified 12,824 protein coding genes.

Keywords
Melitaea athalia, Mellicta athalia, heath fritillary, genome sequence, chromosomal

This article is included in the Tree of Life
gateway.

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 1 10 Nov 2021	✓ report	✓ report

1. **Brian A. Counterman** iD, Auburn University, Auburn, USA

2. **Hikmet Budak** iD, Montana BioAgriculture Inc., Missoula, 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: Hayward A: Investigation, Resources; **Vila R**: Investigation, Resources, Writing – Review & Editing; **Laetsch DR**: Investigation, Resources; **Lohse K**: Investigation, Resources, Writing – Review & Editing; **Baril T**: Writing – Original Draft Preparation, Writing – Review & Editing;

How to cite this article: Hayward A, Vila R, Laetsch DR *et al.* **The genome sequence of the heath fritillary, *Melitaea athalia* (Rottemburg, 1775) [version 1; peer review: 2 approved]** Wellcome Open Research 2021, **6**:304 https://doi.org/10.12688/wellcomeopenres.17280.1

First published: 10 Nov 2021, **6**:304 https://doi.org/10.12688/wellcomeopenres.17280.1

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Papilionoidea; Nymphalidae; Nymphalinae; *Melitaea athalia* (also known as *Mellicta athalia*) (Rottemburg, 1775) (NCBI:txid113330).

Introduction

The heath fritillary, *Melitaea athalia* (also known as *Mellicta athalia*), is a medium-small sized butterfly found throughout the Palaearctic from western Europe to Japan. Historically, the species has been linked with the traditional practice of woodland coppicing, earning it the nickname of 'Woodman's Follower'. *M. athalia* is one of the UK's rarest butterflies and was on the brink of extinction during the 1970s, but conservation efforts have since helped to save the species (Warren, 1987). In the UK *M. athalia* is restricted to grasslands in Cornwall and Devon, heathland in Exmoor, and coppiced woodland in Kent and Essex (Tomlinson & Still, 2002) and is a species of principal importance under the Natural Environment and Rural Communities Act 2006. However, it is listed as Least Concern in the IUCN Red List (Europe) (van Swaay *et al.*, 2010). Up to eight forms and subspecies are recognized in Europe (Tolman & Lewington, 1997). The taxon *celadussa* Fruhstorfer, 1910, originally described as a subspecies of *athalia* from southwestern Europe, is now recognized by many authors as a distinct parapatric species, with a contact zone extending from France to Austria where hybrids are found (Wiemers *et al.*, 2018). Univoltine Fennoscandian

and southern European alpine subspecies fly in single broods (June-July), whilst subalpine subspecies are bivoltine and fly during May-June and late July-August (Tolman & Lewington, 1997). Females of *M. athalia* lay eggs in batches on the underside of leaves of a wide range of herbaceus food plants, with caterpillars feeding, aestivating, and hibernating together in silk nests (Wahlberg, 2000). The standard haploid karyotype of *M. athalia* consists of 30 autosomes and one sex chromsome (Bátori *et al.*, 2012), and the female is heterogametic (WZ).

Genome sequence report

The genome was sequenced from a single female *M. athalia* collected from Lupşa, Transylvania, Romania (latitude 46.416, longitude 23.192) (Figure 1). A total of 30-fold coverage in Pacific Biosciences single-molecule long reads (N50 16 kb) and 64-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 82 missing/misjoins and removed 19 haplotypic duplications, reducing the assembly size by 1.94% and scaffold number by 45.12%, and increasing the scaffold N50 by 7.20%.

The final assembly has a total length of 610 Mb in 46 sequence scaffolds with a scaffold N50 of 20 Mb (Table 1). Of the assembly sequence, 99.98% was assigned to 32 chromosomal-level scaffolds, representing 30 autosomes (numbered by sequence length), and the W and Z sex chromosome (Figure 2–Figure 5; Table 2). The assembly has a BUSCO (Simão *et al.*, 2015)
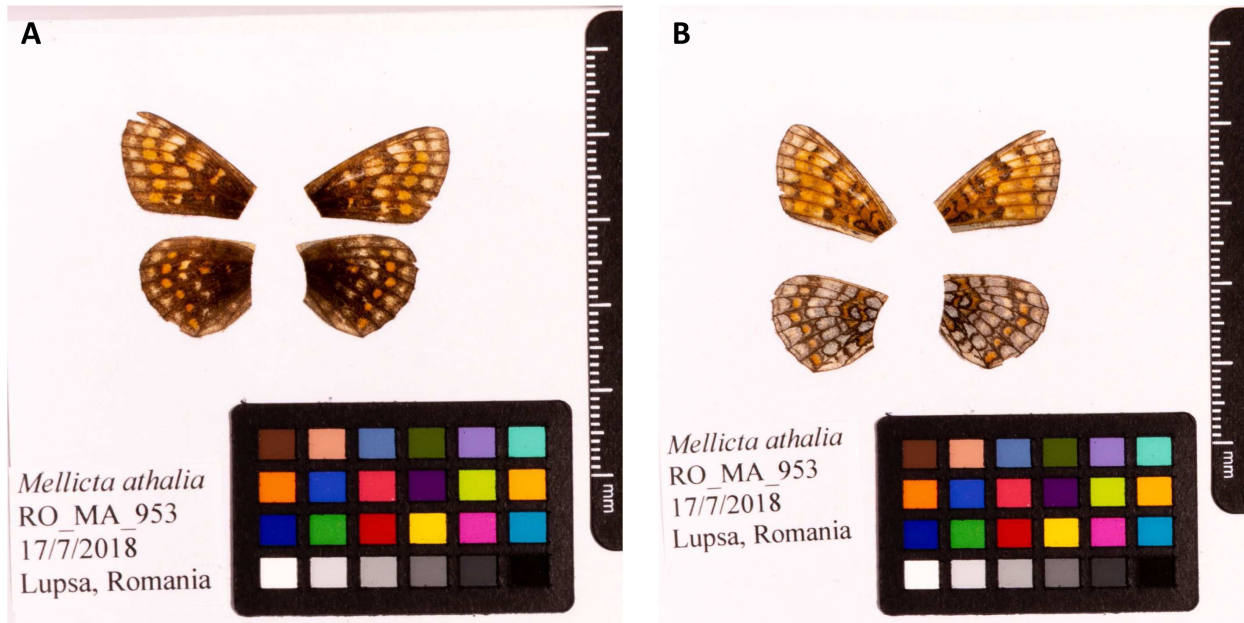
Figure 1. Fore and hind wings of *Melitaea athalia* specimen from which the genome was sequenced. (**A**) Dorsal surface view of wings from specimen RO_MA_953 (ilMelAtha1.1) from Lupşa, Transylvania, Romania, used to generate Pacific Biosciences and 10X genomics data. (**B**) Ventral surface view of wings from specimen RO_MA_953 (ilMelAtha1.1) from Lupşa, Transylvania, Romania, used to generate Pacific Biosciences and 10X genomics data.

Table 1. Genome data for *Melitaea athalia*, ilMelAtha1.1.

Project accession data	
Assembly identifier	ilMelAtha1.1
Species	*Melitaea athalia* (also known as *Mellicta athalia*)
Specimen	ilMelAtha1, RO_MA_953
NCBI taxonomy ID	NCBI:txid113330
BioProject	PRJEB42954
BioSample ID	SAMEA7523312
Isolate information	Female, whole organism
Raw data accessions	
PacificBiosciences SEQUEL II	ERR6576319
10X Genomics Illumina	ERR6054423-ERR6054426
Hi-C Illumina	ERR6054427
Genome assembly	
Assembly accession	GCA_905163435.1
Accession of alternate haplotype	GCA_905163405.1
Span (Mb)	576
Number of contigs	70
Contig N50 length (Mb)	18
Number of scaffolds	43
Scaffold N50 length (Mb)	19
Longest scaffold (Mb)	23
BUSCO* genome score	C:98.6%[S:97.9%,D:0.7%],F:0.4%,M:1.0%,n:5286
Gene annotation	
Number of protein coding genes	12,824
Average coding sequence length (bp)	1,492
Average number of exons per transcript	8
Average exon size (bp)	264
Average intron size (bp)	2,892

*BUSCO scores based on the lepidoptera_odb10 BUSCO set using v5.1.2. C= complete [S= single copy, D=duplicated], F=fragmented, M=missing, n=number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/ilMelAtha1.1/dataset/CAJNAG01/busco.

completeness of 98.6% (single 97.9%, duplicated 0.7%, fragmented 0.4%, missing 1.0%) 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.

Gene annotation
The Ensembl gene annotation system (Aken *et al.*, 2016) was used to generate annotation for the *Melitaea athalia* assembly (GCA_905220545.1, see https://rapid.ensembl.org/Mellicta_athalia_GCA_905220545.1/; Table 1). The 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)) and OrthoDB (Kriventseva *et al.*, 2008). Prediction tools, CPC2 (Kang *et al.*, 2017) and RNAsamba (Camargo *et al.*, 2020), were used to aid determination of protein coding genes.
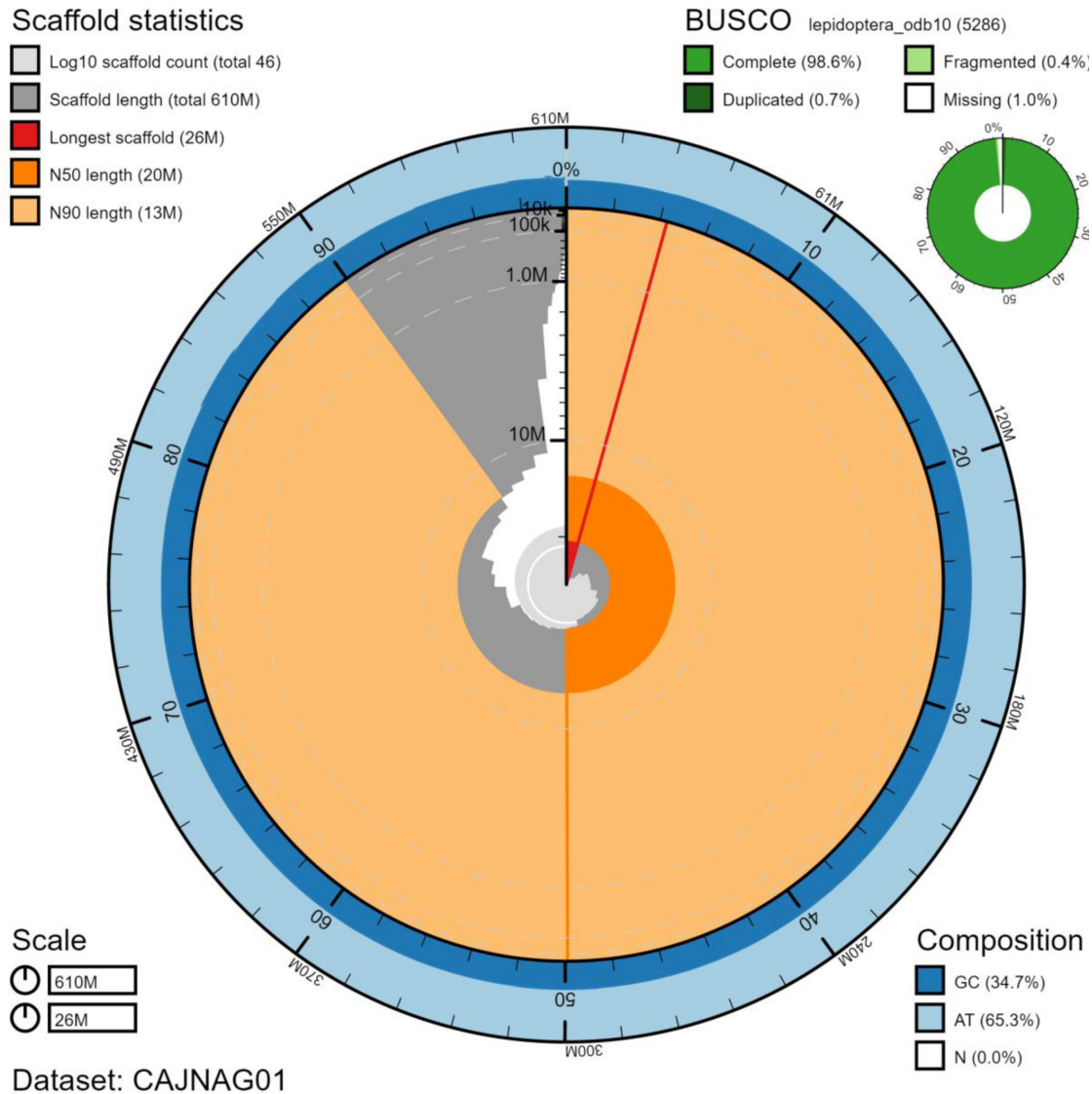
Figure 2. Genome assembly of *Mellitaea athalia*, ilMelAtha1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 609,564,789 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly (26,233,870 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (20,295,254 and 13,271,753 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/ilMelAtha1.1/dataset/CAJNAG01/snail.

Methods

Sample acquisition, nucleic acid extraction and sequencing

A single female *M. athalia* was collected from Lupşa, Transylvania, Romania (latitude 46.416, longitude 23.192) by Alex Hayward (University of Exeter), Roger Vila (Universitat Pompeu Fabra), Dominik Laetsch and Konrad Lohse (both University of Edinburgh), using a net. The specimen was identified by Roger Vila and was snap-frozen in liquid nitrogen.

DNA was extracted from the whole organism of ilMelAtha1 using the Qiagen MagAttract HMW DNA kit, according to
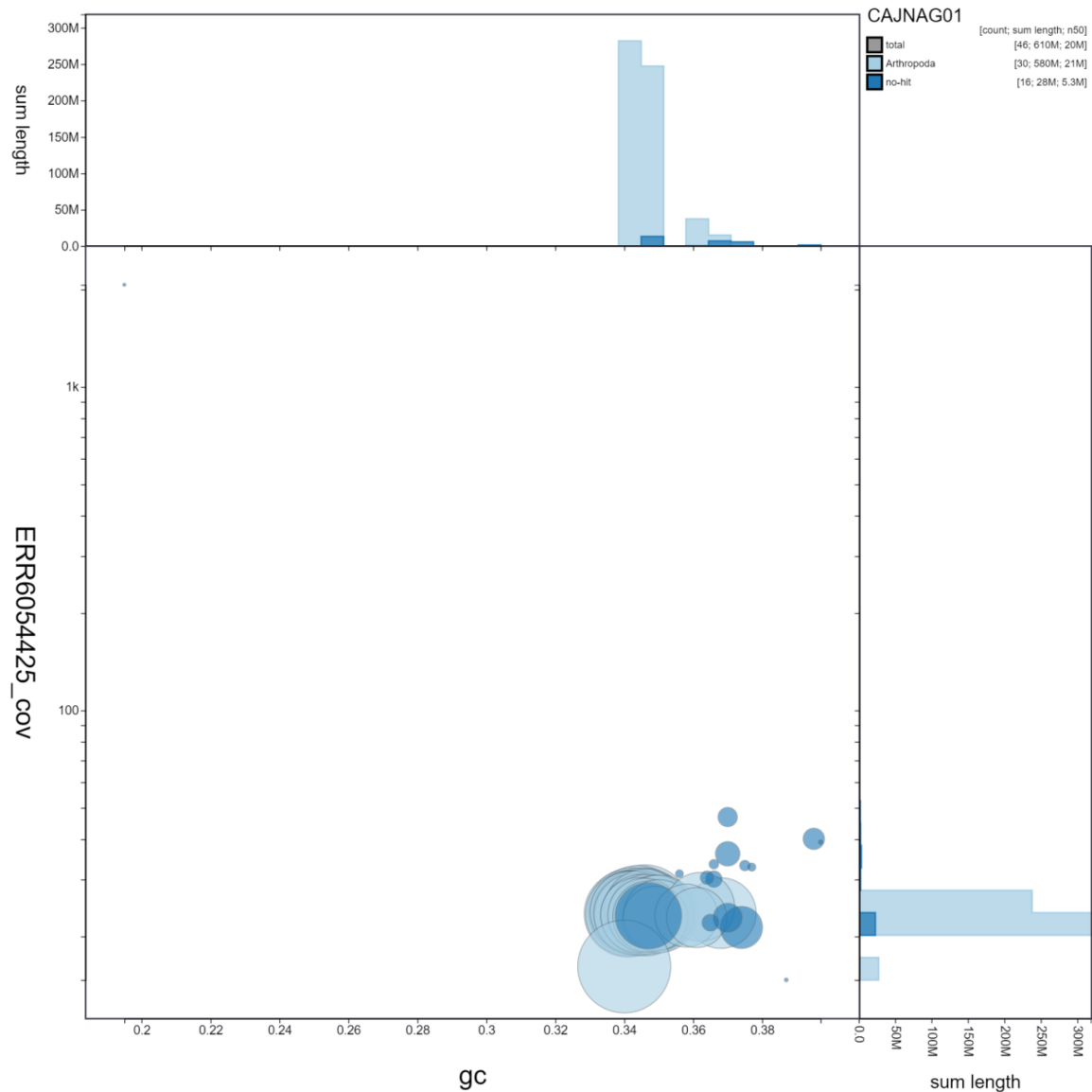
Figure 3. Genome assembly of *Mellitaea athalia* , ilMelAtha1.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilMelAtha1.1/dataset/CAJNAG01/blob.

the manufacturer's instructions. Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud sequencing libraries were then constructed according to the manufacturers' instructions. Sequencing was performed by the Scientific Operations core at the Wellcome Sanger Institute on Pacific Biosciences SEQUEL II and Illumina HiSeq X instruments. Hi-C data were generated using the Arima v1.0 kit and sequenced on HiSeq X.

Genome assembly
Assembly was carried out with HiCanu (Nurk *et al.*, 2020). 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 was performed using gEVAL, HiGlass (Kerpedjiev *et al.*, 2018) and Pretext. The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2021). The genome was analysed and BUSCO scores generated within the BlobToolKit
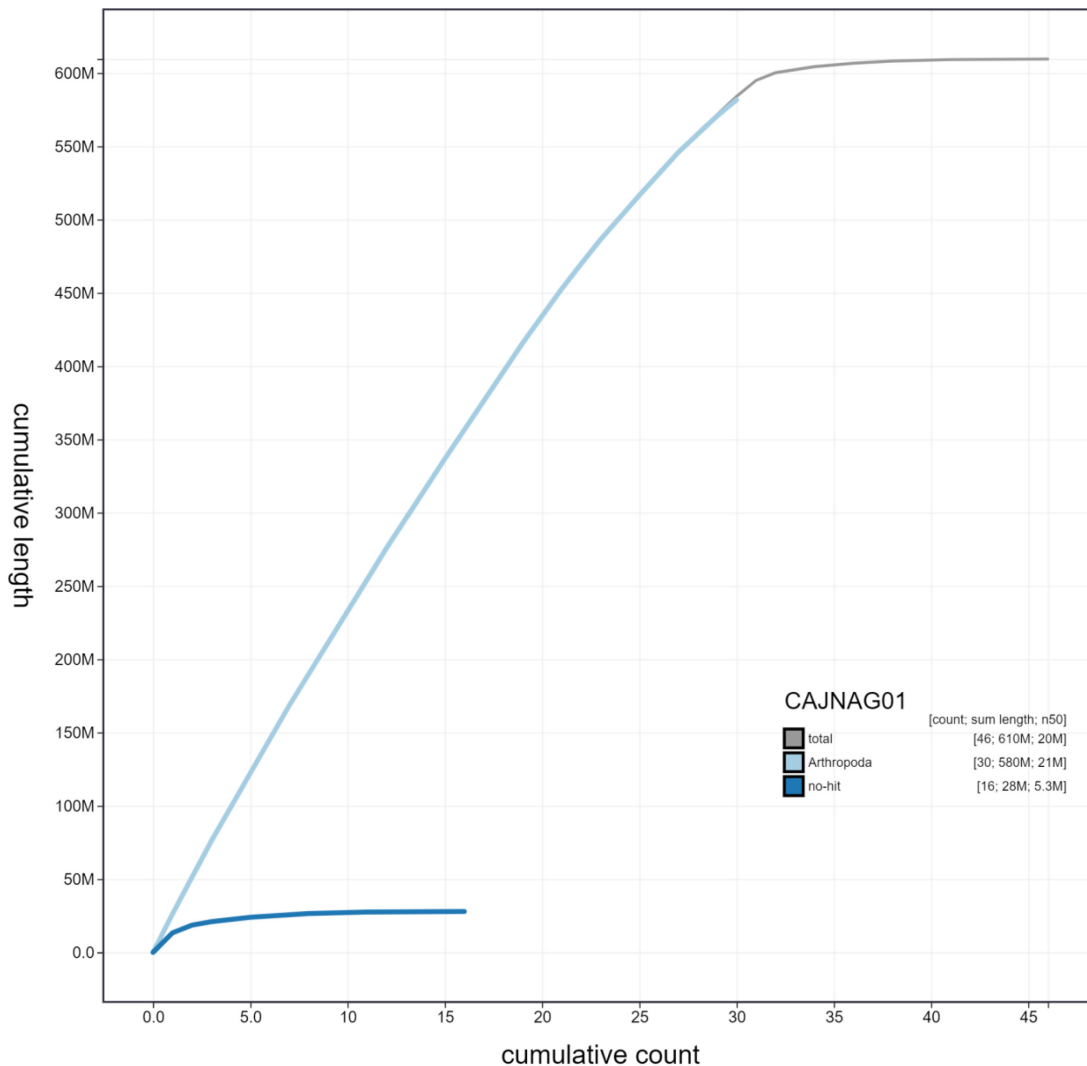
Figure 4. Genome assembly of *Mellitaea athalia*, ilMelAtha1.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilMelAtha1.1/dataset/CAJNAG01/cumulative.

environment (Challis *et al.*, 2020). Table 3 contains a list of all software tool versions used, where appropriate.

Ethical/compliance issues

The materials that have contributed to this genome note were supplied by a Tree of Life collaborator. 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;

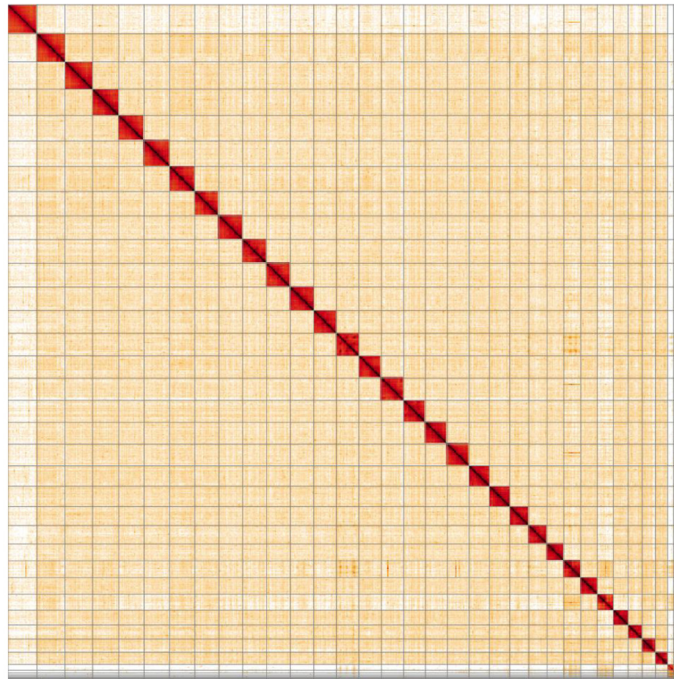- Legality of collection, transfer and use (national and international).

Figure 5. Genome assembly of _Mellitaea athalia_, ilMelAtha1.1: Hi-C contact map. Hi-C contact map of the ilMelAtha1.1 assembly, visualised in HiGlass. Chromosomes are arranged in size order from left to right and top to bottom.

Table 2. Chromosomal pseudomolecules in the genome assembly of _Melitaea athalia_, ilMelAtha1.1.

INSDC accession	Chromosome	Size (Mb)	GC%
HG992177.1	1	25.13	34.6
HG992178.1	2	24.88	34.4
HG992179.1	3	23.65	34.6
HG992180.1	4	22.93	34.2
HG992181.1	5	22.90	34.1
HG992182.1	6	22.79	34.5
HG992183.1	7	21.87	34.4
HG992184.1	8	21.42	34.3
HG992185.1	9	21.39	34.2
HG992186.1	10	21.38	34.1
HG992187.1	11	21.23	34.4
HG992188.1	12	20.51	34.3
HG992189.1	13	20.30	34.8
HG992190.1	14	20.21	34.2
HG992191.1	15	19.99	34.3
HG992192.1	16	19.82	34.6

INSDC accession	Chromosome	Size (Mb)	GC%
HG992193.1	17	19.64	34.5
HG992194.1	18	19.55	34.7
HG992195.1	19	18.44	34.9
HG992196.1	20	18.37	35
HG992197.1	21	17.06	34.4
HG992198.1	22	16.62	34.6
HG992199.1	23	15.22	34.7
HG992200.1	24	15.15	36.8
HG992201.1	25	14.97	35
HG992202.1	26	14.40	36.2
HG992203.1	27	13.27	34.7
HG992204.1	28	12.76	34.9
HG992205.1	29	11.97	35.8
HG992206.1	30	10.90	36.1
HG992207.1	W	5.27	37.4
HG992176.1	Z	26.23	34
HG992208.1	MT	0.02	19.7
-	Unplaced	9.34	37.2

Table 3. Software tools used.

Software tool	Version	Source
HiCanu	2.1	Nurk *et al.*, 2020
purge_dups	1.2.3	Guan *et al.*, 2020
SALSA2	2.2	Ghurye *et al.*, 2019
longranger align	2.2.2	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines
freebayes	1.3.1-17-gaa2ace8	Garrison & Marth, 2012
MitoHiFi	1	https://github.com/marcelauliano/MitoHiFi
gEVAL	2016	Chow *et al.*, 2016
HiGlass	1.11.6	Kerpedjiev *et al.*, 2018
PretextView	0.1.x	https://github.com/wtsi-hpag/PretextView
BlobToolKit	2.6.2	Challis *et al.*, 2020

Each transfer of samples is undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Tree of Life collaborator, Genome Research Limited (operating as the Wellcome Sanger Institute) and in some circumstances other Tree of Life collaborators.

Data availability

European Nucleotide Archive: Mellicta athalia (heath fritillary). Accession number PRJEB42954; https://identifiers.org/ena.embl/PRJEB42954.

The genome sequence is released openly for reuse. The *M. athalia* 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.

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**

Bátori E, Pecsenye K, Pál Tóth J, *et al.*: **Patterns of Genetic and Morphometric Differentiation in *Melitaea* (*Mellicta*) *athalia* (Lepidoptera: Nymphalidae).** *Biol J Linn Soc.* Linnean Society of London. 2012; **107**(2): 398–413.
Publisher Full Text

Camargo AP, Sourkov V, Pereira GAG, *et al.*: **RNAsamba: Neural Network-Based Assessment of the Protein-Coding Potential of RNA Sequences.** *NAR Genom Bioinform.* 2020; **2**(1): lqz024.
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**

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**

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**

Kang YJ, Yang DC, Kong L, *et al.*: **CPC2: A Fast and Accurate Coding Potential Calculator Based on Sequence Intrinsic Features.** *Nucleic Acids Res.* 2017; **45**(W1): W12–16.
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**

Kriventseva EV, Rahman N, Espinosa O, *et al.*: **OrthoDB: The Hierarchical Catalog of Eukaryotic Orthologs.** *Nucleic Acids Res.* 2008; **36**(Database issue): D271–75.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Nurk S, Walenz BP, Rhie A, *et al.*: **HiCanu: Accurate Assembly of Segmental Duplications, Satellites, and Allelic Variants from High-Fidelity Long Reads.** *Genome Res.* 2020; **30**(9): 1291–1305.
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**

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–12.
PubMed Abstract | **Publisher Full Text**

Tolman T, Lewington R: **Butterflies of Britain & Europe. Collins Field Guide.** Harper Collins, London. 1997.
Reference Source

Tomlinson D, Still R: **Britain's Butterflies.** WILDGuides. 2002.
Reference Source

Uliano-Silva M, Gabriel Ferreira Nunes J, 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**

van Swaay C, Wynhoff I, Verovnik R, *et al.*: **IUCN Red List of Threatened Species: Melitaea Athalia.** *IUCN Red List of Threatened Species*, 2010; e.T174355A7056436.
Publisher Full Text

Wahlberg N: **Comparative Descriptions of the Immature Stages and Ecology of Five Finnish Melitaeine Butterfly Species (Lepidoptera: Nymphalidae).** *Entomologica Fennica.* 2000; **11**(3): 167–74.
Publisher Full Text

Warren MS: **The Ecology and Conservation of the Heath Fritillary Butterfly, Mellicta athalia. III. Population Dynamics and the Effect of Habitat Management.** *J Appl Ecol.* 1987; **24**(2): 499–513.
Publisher Full Text

Wiemers M, Balletto E, Dincă V, *et al.*: **An updated checklist of the European Butterflies (*Lepidoptera, Papilionoidea*).** *Zookeys.* 2018; (811): 9–45.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 1

Reviewer Report 27 January 2022

✓ **Hikmet Budak** iD

Department of Genomics and Genome Editing, Montana BioAgriculture Inc., Missoula, MT, USA

I believe that there is data which contributes to the community. Both the genome sequencing and assembly methods and applications looks good and presented well although it would be great to go into details a little more so that other scientists can use/compare/benefit form their experiences. I am unsure if Figure 4 is necessary to show cumulative seq. The authors were even able to assemble the W chromosome. This would definitely help sequencing and chromosome level assembly of wheat stem sawfly and orange wheat blossom midge, dangerous insects in the USA.

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: Plant genomics and biology, Next generation sequencing and annotations, smallRNAs, microRNAs, LncRNAs.

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 25 November 2021

✅ **Brian A. Counterman** (iD)

Department of Biological Sciences, Auburn University, Auburn, AL, USA

This manuscript presents a concise report of the genome sequencing and assembly of *Melitaea athalia*, one of the rarest butterflies in Europe. The methods used are appropriate and appear to have been conducted appropriately. The result is a highly contiguous, chromosome-level assembly of the genome. It is noteworthy that the authors were able to assemble the W chromosome and release the unphased assemblies of both diploid copies. This genome is certain to be a valuable resource for comparative genomics of Lepidoptera and future conservation efforts of the species.

I found no major errors or concerns with the manuscript, and felt the tables and figures included sufficiently and succinctly presented relevant information to assess the methods used and quality of the genome assembly. I was unclear which version of BUSCO was used and if the "lepidoptera_odb10" database was the 2019 version. These details could be useful for future users of the data, and perhaps could be simply added to Table 3. Overall, I commend the authors for their clear and concise presentation of the work.

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: Lepidoptera, Evolution, 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.