DATA NOTE

The genome sequence of the white ermine, *Spilosoma lubricipeda* Linnaeus 1758 [version 1; peer review: 1 approved, 2 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 & Hydrology, Wallingford, OX10 8BB, UK
[2]Department of Zoology, University of Oxford, Oxford, OX1 3SZ, UK

v1 **First published:** 14 Oct 2021, **6**:271
https://doi.org/10.12688/wellcomeopenres.17190.1
Latest published: 14 Oct 2021, **6**:271
https://doi.org/10.12688/wellcomeopenres.17190.1

Abstract

We present a genome assembly from an individual male *Spilosoma lubricipeda* (the white ermine; Arthropoda; Insecta; Lepidoptera; Erebidae). The genome sequence is 587 megabases in span. The majority of the assembly is scaffolded into 30 chromosomal pseudomolecules, with the Z sex chromosome assembled.

Keywords

Spilosoma lubricipeda, white ermine, genome sequence, chromosomal

This article is included in the Tree of Life gateway.

Open Peer Review

Reviewer Status ? ✓ ?

	Invited Reviewers		
	1	**2**	**3**
version 1 14 Oct 2021	? report	✓ report	? report

1. **Merly Escalona**, University of California–Santa Cruz, Santa Cruz, USA
2. **Dominik R. Laetsch**, University of Edinburgh, Edinburgh, UK
3. **Petr Nguyen**, University of South Bohemia, Ceske Budejovice, Czech Republic

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: Formal Analysis, Investigation, Resources; **Holland PWH**: Investigation, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing;

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, Holland PWH, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the white ermine, *Spilosoma lubricipeda* Linnaeus 1758 [version 1; peer review: 1 approved, 2 approved with reservations]** Wellcome Open Research 2021, **6**:271 https://doi.org/10.12688/wellcomeopenres.17190.1

First published: 14 Oct 2021, **6**:271 https://doi.org/10.12688/wellcomeopenres.17190.1

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Noctuoidea; Erebidae; Arctiinae; Spilosoma; *Spilosoma lubricipeda* Linnaeus 1758 (NCBI:txid875880).

Introduction

Spilosoma lubricipeda (White ermine) is found across much of Eurasia but has decreased in abundance significantly in the UK in recent decades, with the cause(s) being unknown (Fox *et al.*, 2013) (Prescott *et al.*, 2019) The genome of *S. lubricipeda* 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 *S. lubricipeda*, based on one male specimen from Wytham Woods, Oxfordshire, UK.

Genome sequence report

The genome was sequenced from a single male *S. lubricipeda* collected from Wytham Woods, Oxfordshire, UK (latitude 51.768, longitude -1.337). A total of 39-fold coverage in Pacific Biosciences single-molecule long reads and 43-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 22 missing/misjoins and removed 2 haplotypic duplications, reducing the assembly length by 0.09% and increasing the scaffold number by 3.70%. The final assembly has a total length of 587 Mb in 37 sequence scaffolds with a scaffold N50 of 21 Mb (Table 1). Of the assembly sequence, 99.98% was assigned to 30 chromosomal-level scaffolds, representing 29 autosomes (numbered by sequence length), and the Z sex chromosome (Figure 1–Figure 4; Table 2). The assembly has a BUSCO (Simão *et al.*, 2015) v5.1.2 completeness of 98.8% 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.

Table 1. Genome data for *Spilosoma lubricipeda*, ilSpiLubr1.1.

Project accession data	
Assembly identifier	ilSpiLubr1.1
Species	*Spilosoma lubricipeda*
Specimen	ilSpiLubr1
NCBI taxonomy ID	NCBI:txid875880
BioProject	PRJEB42957
BioSample ID	SAMEA7520525
Isolate information	Male, abdomen
Raw data accessions	
PacificBiosciences SEQUEL II	ERR6406203
10X Genomics Illumina	ERR6054439-ERR6054442
Hi-C Illumina	ERR6054438
Genome assembly	
Assembly accession	GCA_905220595.1
Accession of alternate haplotype	GCA_905220605.1
Span (Mb)	587
Number of contigs	57
Contig N50 length (Mb)	21
Number of scaffolds	37
Scaffold N50 length (Mb)	21
Longest scaffold (Mb)	25
BUSCO* genome score	C:98.8%[S:98.1%,D:0.7%],F:0.3%,M:0.9%,n:5286

*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/ilSpiLubr1.1/dataset/CAJNAK01/busco.

Methods

A single male *S. lubricipeda* was collected from Wytham Woods, Oxfordshire, UK (latitude 51.768, longitude -1.337) by Douglas Boyes, University of Oxford using a light trap. The specimens were snap-frozen in dry ice using a CoolRack before transferring to the Wellcome Sanger Institute (WSI).

DNA was extracted at the Tree of Life laboratory, WSI. The ilSpiLubr1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. Abdomen tissue was disrupted to a fine powder using a Biomasher tissue homogeniser. 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 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.

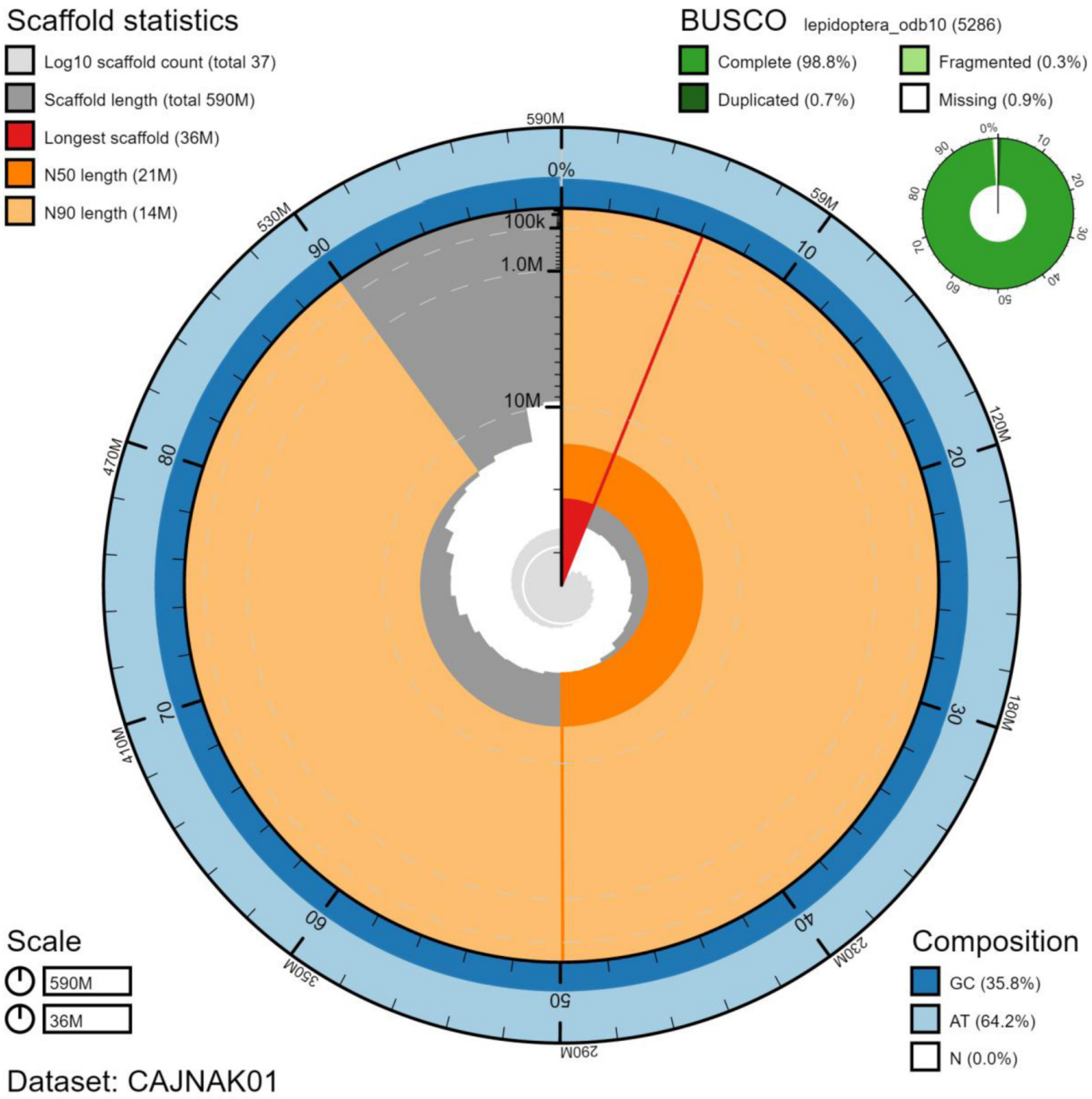

Figure 1. Genome assembly of *Spilosoma lubricipeda*, ilSpiLubr1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilSpiLubr1.1/dataset/CAJNAK01/snail.

Pacific Biosciences HiFi circular consensus and 10X Genomics 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 and Illumina HiSeq X instruments. Hi-C data were generated from abdomen tissue using the Arima v2.0 kit and sequenced on HiSeq X.

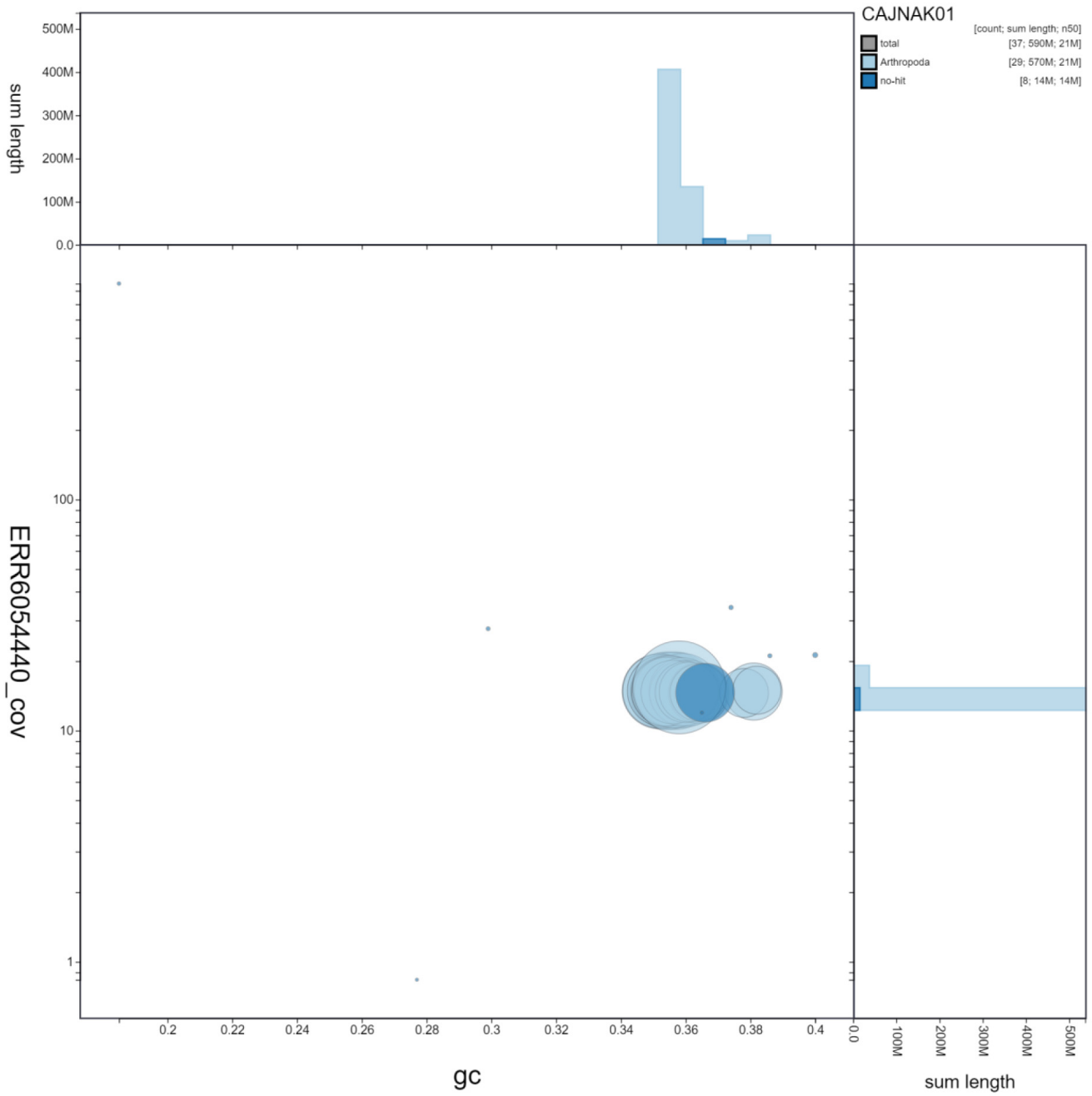

Figure 2. Genome assembly of *Spilosoma lubricipeda*, ilSpiLubr1.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilSpiLubr1.1/dataset/CAJNAK01/blob.

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021); haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2020). The assembly was polished with the 10X Genomics Illumina data by aligning to the assembly with longranger align, calling variants with freebayes (Garrison & Marth, 2012). One round of the Illumina polishing was applied. Scaffolding with Hi-C data (Rao *et al.*, 2014) was carried out with 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 and Pretext. The mitochondrial genome was assembled

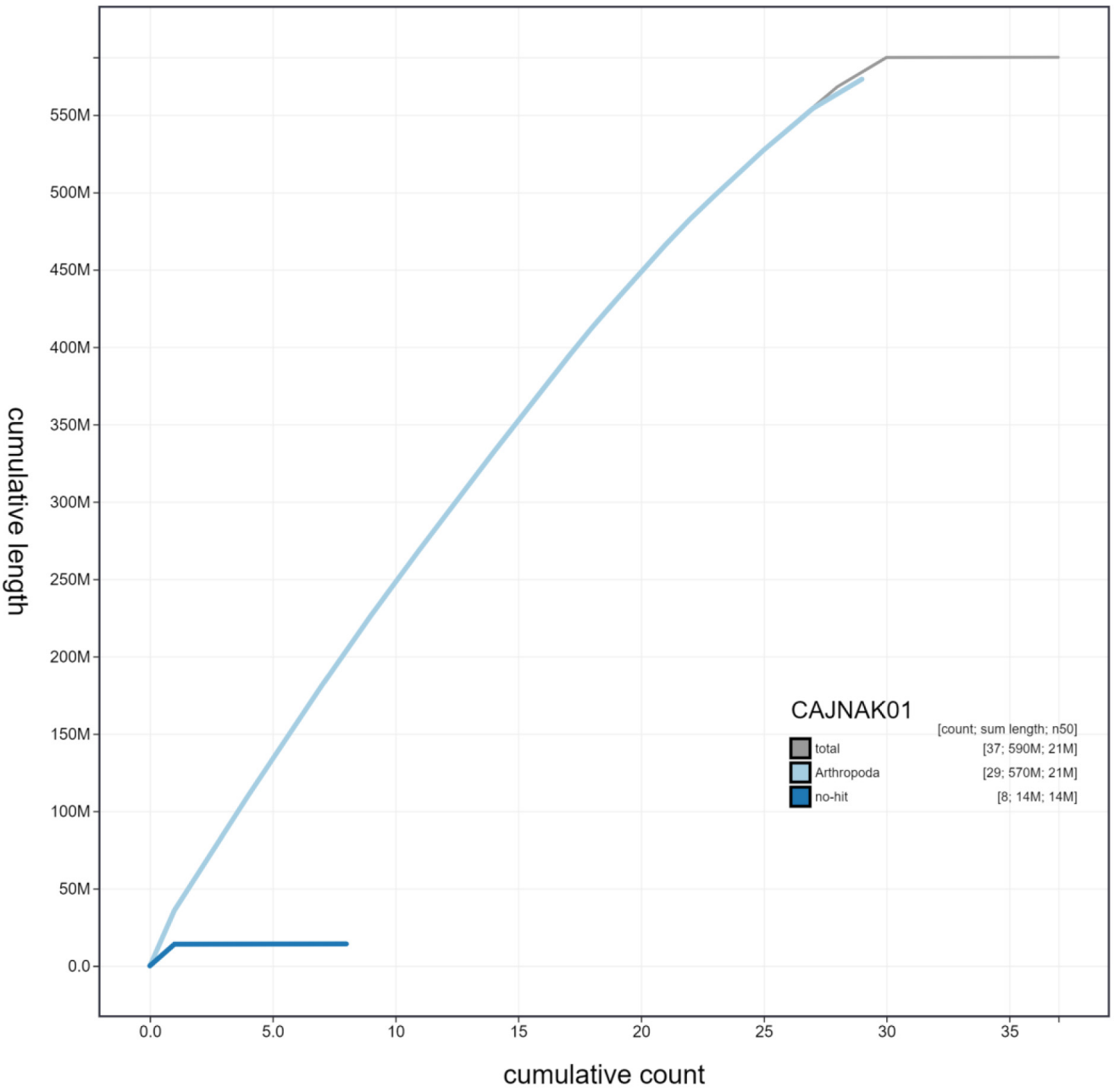

Figure 3. Genome assembly of *Spilosoma lubricipeda*, ilSpiLubr1.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all chromosomes. Coloured lines show cumulative lengths of chromosomes assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilSpiLubr1.1/dataset/CAJNAK01/cumulative.

using MitoHiFi (Uliano-Silva *et al.*, 2021). 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.

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

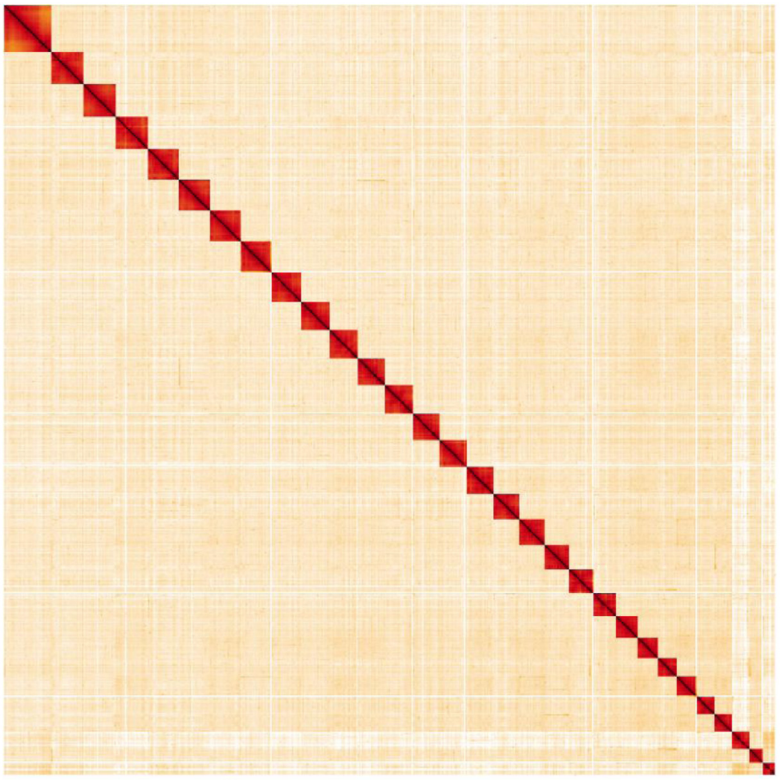

Figure 4. Genome assembly of *Spilosoma lubricipeda*, ilSpiLubr1.1: Hi-C contact map. Hi-C contact map of the ilSpiLubr1.1 assembly, visualised in HiGlass.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Spilosoma lubricipeda*, ilSpiLubr1.1.

INSDC accession	Chromosome	Size (Mb)	GC%
HG992275.1	1	24.84	35.7
HG992276.1	2	24.64	35.5
HG992277.1	3	24.48	35.4
HG992278.1	4	23.97	35.4
HG992279.1	5	23.77	35.6
HG992280.1	6	23.46	35.2
HG992281.1	7	22.82	35.9
HG992282.1	8	22.61	35.2
HG992283.1	9	21.71	35.4
HG992284.1	10	21.21	35.5
HG992285.1	11	21.05	35.3
HG992286.1	12	21.03	35.5
HG992287.1	13	20.64	35.7
HG992288.1	14	20.45	35.6
HG992289.1	15	20.26	35.4
HG992290.1	16	20.07	35.5
HG992291.1	17	19.45	35.4
HG992292.1	18	18.38	35.9
HG992293.1	19	17.77	36.1
HG992294.1	20	17.67	35.9
HG992295.1	21	16.57	35.6
HG992296.1	22	15.29	36
HG992297.1	23	14.91	36.4
HG992298.1	24	14.47	36.1
HG992299.1	25	14.02	36.6
HG992300.1	26	13.41	36.4
HG992301.1	27	13.33	38.1
HG992302.1	28	9.66	37.8
HG992303.1	29	9.33	38.2
HG992274.1	Z	35.93	35.8
HG992304.1	MT	0.02	18.6
-	Unplaced	0.15	36.7

Table 3. Software tools used.

Software tool	Version	Source
Hifiasm	0.1.2	Cheng *et al.*, 2021
purge_dups	1.2.3	Guan *et al.*, 2020
longranger	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
SALSA2	2.2	Ghurye *et al.*, 2019
gEVAL	N/A	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

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: Spilosoma lubricipeda (white ermine). Accession number PRJEB42957: https://identifiers.org/ena.embl:PRJEB42957

The genome sequence is released openly for reuse. The *S. lubricipeda* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project. All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1.

Acknowledgements

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.4789929.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.4893704.

Members of the Wellcome Sanger Institute Tree of Life programme collective are listed here: https://doi.org/10.5281/zenodo.5377053.

Members of Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.4790456.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.5013542.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783559.

References

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

Fox R, Parsons MS, Chapman JW, *et al.*: **The State of Britain's Larger Moths 2013.** 2013; 29.
Reference Source

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**

Prescott T, Tordoff GM, Fox R: **Atlas of Britain and Ireland's Larger Moths.** Pisces Publications, Newbury. 2019.
Reference Source

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**

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

Open Peer Review

Current Peer Review Status: ? ✓ ?

Version 1

Reviewer Report 03 November 2021

https://doi.org/10.21956/wellcomeopenres.18993.r46441

?

Petr Nguyen
University of South Bohemia, Ceske Budejovice, Czech Republic

The authors report a chromosome-level assembly of the white ermine, *Spilosoma lubricipeda*, sequenced by PacBio and 10x Genomics technologies. Decrease of *S. lubricipeda* abundance in the UK is mentioned in the introduction but it is not clear how it is relevant for genome sequencing. It should be elaborated.

The *S. lubricipeda* assembly is 587 Mb long. Although the *S. lubricipeda* genome size is not known, some comparison with a genome size inferred from k-mers would be nice. Anyway, the assembly length seems to be close to a genome size of its congener, *S. virginica* (626 Mb according to www.genomesize.com). 99.98% of sequence was assigned to 30 chromosome-level scaffolds, which is interesting as Robinson (1971, Lepidoptera genetics)[1] provides chromosome number n=31 for this species citing two independent references. The Z chromosome is surprisingly large in *S. lubricipeda.* The karyotype n=31 is ancestral in Lepidoptera and comparison with genome of some other species with n=31 such as *Plutella xylostella*, *Spodoptera litura*, or *Melitaea cinxia* could reveal fusion between the Z chromosome and an autosome which occurred in the sampled *S. lubricipeda* population. The authors should address the above mentioned in their genome sequence report.

The method section describes the pipeline but not the parameters used. For better reproducibility, parameters used for individual processing and assembly steps should be specified.

References

1. Robinson R: Lepidoptera Genetics: International Series of Monographs in Pure and Applied Biology: Zoology. *Pergamon Press*. 1971; **46**. Publisher Full Text

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

Partly

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genetics of Lepidoptera

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.

Reviewer Report 02 November 2021

https://doi.org/10.21956/wellcomeopenres.18993.r46443

Dominik R. Laetsch
Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK

This article describes the sampling, sequencing, assembly, scaffolding, curation, and quality control that led to the *Spilosoma lubricipeda* genome assembly. The steps are described in sufficient detail to allow replication and are consistent with current best practices. Relevant data has been deposited on INSDC databases and is accessible.

Minor comments:

- Richard Fox in the comments points out that the reference Prescott *et al.*, 2019 should be edited. Since he is one of the authors of that book he is probably right.
- The same sentence (first sentence in introduction) is also missing a full stop after the citations.
- The introduction could be expanded a bit to include more information about why the genome is of interest to the research community.

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 work with the Darwin Tree of Life initiative before, but I don't think this affects my ability to review impartially.

Reviewer Expertise: Evolutionary Biology, Bioinformatics, Genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 28 October 2021

https://doi.org/10.21956/wellcomeopenres.18993.r46442

?

Merly Escalona
Department of Biomolecular Engineering, University of California–Santa Cruz, Santa Cruz, CA, USA

The paper reports a chromosome level assembly of the White ermine moth. The quality of the genome assembly seems impressive at first sight and corresponds to an excellent resource for the scientific community. However, there are some things missing from the report that are important for a type of data note/genome announcement paper that is presented here.

- First, seems necessary and in line with published genome notes from the same journal to have a picture of the species.

- There are metrics other than BUSCO scores that complement the quality assessment of a genome assembly, i.e. k-mer completeness and per base quality (QV).

- While there's information related to the amount of read coverage generated for Pacific Biosciences (PacBio) HiFi reads and 10X Chromium data, there's no information regarding the coverage for the Arima HiC libraries. It might not be common to describe 'coverage' for HiC data but knowing the number of reads generated makes the description of the sequencing data generated consistent.

- Related to the mitochondrial genome assembly: MitoHiFi looks for a closely related species and uses it to guide the assembly for the species at hand. It is then relevant to know which species was used.

- Description of the software version used is not consistent throughout the paper. At the Genome sequence report section, you identify the version of the BUSCO score/program with v5.1.2 but then the rest of the software tools do not have the version next to them but on Table 3.

- There are missing citations in the text. Methods section, 4th paragraph. HiGlass and Pretext.

- Also, in the Methods section, 4th paragraph, on the sentence before last:
"The genome was analyzed and BUSCO scores generated within the BlobToolKit environment (Challis et al., 2020)." - Although I understand the context of the sentence, the flow of the paragraphs seems to indicate that this sentence refers to the mitochondrial assembly which is incorrect.

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?
No

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

Competing Interests: I am part of the California Conservation Genomics Project, which along with the Darwin Tree of Life Project, is affiliated with the EarthBiogenome Project.

Reviewer Expertise: Genome assembly

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.

Comments on this article

Version 1

Reader Comment 18 Oct 2021

Richard Fox, Butterfly Conservation, Wareham, UK

Excellent work sequencing genomes of UK moths. Just a very minor point that the citation Prescott *et al.* 2019 is incorrect. This full reference for this publication is:

Randle, Z., Evans-Hill, L.J., Parsons, M.S., Tyner, A., Bourn, N.A.D., Davis, A.M., Dennis, E.B., O'Donnell, M., Prescott, T., Tordoff, G.M. and Fox, R. (2019). *Atlas of Britain & Ireland's Larger Moths*. Pisces Publications, Newbury.

I expect that this reference will be cited on many moth genome data papers based on UK samples, so thought you would want to be aware of the error.

Keep up the great work!

Competing Interests: No competing interests were disclosed.