DATA NOTE

The genome sequence of the holly blue, *Celastrina argiolus* (Linnaeus, 1758) [version 1; peer review: 2 approved]

Alex Hayward¹, Charlotte Wright², 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

¹College of Life and Environmental Sciences, Department of Biosciences, University of Exeter, Penryn, UK
²Tree of Life, Wellcome Sanger Institute, Cambridge, CB10 1SA, UK

**Abstract**

We present a genome assembly from an individual male *Celastrina argiolus* (the holly blue; Arthropoda; Insecta; Lepidoptera; Lycaenidae). The genome sequence is 499 megabases in span. The majority (99.99%) of the assembly is scaffolded into 26 chromosomal pseudomolecules, with the Z sex chromosome assembled. Gene annotation of this assembly on Ensembl has identified 12,199 protein coding genes.

**Keywords**

Celastrina argiolus, holly blue, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

---

**Open Peer Review**

**Reviewer Status**

Invited Reviewers

1. Carles Lalueza-Fox, Institute of Evolutionary Biology (CSIC-UPF), Barcelona, Spain
2. Bin Liang, Inner Mongolia University, Hohhot, China

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, Visualization; Wright C: 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). AH is supported by a Biotechnology and Biological Sciences Research Council (BBSRC) David Phillips Fellowship (BB/N020146/1).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Hayward A et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Hayward A, Wright C, Darwin Tree of Life Barcoding collective et al. The genome sequence of the holly blue, Celastrina argiolus (Linnaeus, 1758) [version 1; peer review: 2 approved] Wellcome Open Research 2021, 6:340
https://doi.org/10.12688/wellcomeopenres.17478.1

First published: 14 Dec 2021, 6:340 https://doi.org/10.12688/wellcomeopenres.17478.1
**Species taxonomy**
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Papilionoidea; Lycaenidae; Polyommatinae; Celastrina; *Celastrina argiolus* (Linnaeus, 1758) (NCBI:txid203782).

**Background**
The holly blue, *Celastrina argiolus*, is a widespread butterfly, found throughout the temperate regions of Europe, Asia, North Africa and North America. It is common across the British Isles with the exception of Scotland, where it is absent. Recorded numbers of the butterfly cycles every 4-6 years due to parasitism of the larval form by the larvae of the host-specific ichneumon wasp *Listrodromus nycthemerus* (Revels, 2006; Revels, 1994). Larvae feed mainly on the flower buds, berries and terminal leaves of holly (*Ilex aquifolium*) in the spring, and ivy (*Hedera helix*) in the summer, although they can also use a wide variety of other plants. Adults are distinguished by bright blue wings with pale blue underside and small black spots. The species is typically bivoltine and overwinters as pupae. Adults are generalists, feeding on a variety of nectar sources including hawthorn, brambles and Bugle, as well as honey dew. The holly blue has an estimated genome size of 445 Mb based on flow cytometry (Mackintosh et al., 2019). The karyotype of *C. argiolus* was reported to be 25 by Federley, Lorković, Maeki, and 24 by Bigger, as described in Robinson (1971).

**Genome sequence report**
The genome was sequenced from a single male *C. argiolus* (Figure 1) collected from Oxford, England (latitude 51.74989, longitude 1.22731). A total of 32-fold coverage in Pacific Biosciences single-molecule circular consensus (HiFi) long reads (N50 13 kb) and 69-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 11 missing/misjoins and removed 3 haplotypic duplications, reducing the assembly length by 0.4%, the scaffold number by 20.6% and the scaffold N50 by 1.3%.

The final assembly has a total length of 499 Mb in 28 sequence scaffolds with a scaffold N50 of 20 Mb (Table 1). The majority, 99.99%, of assembly sequence was assigned to 26 chromosomal-level scaffolds, representing 25 autosomes (numbered by sequence length), and the Z sex chromosome (Figure 2–Figure 5; Table 2). The assembly has a BUSCO v5.1.2 (Simão et al., 2015) completeness of 97.1% 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.

**Genome annotation report**
The ilCelArgi3.1 genome has been annotated using the Ensembl rapid annotation pipeline (Table 1; GCA_905187575.1).
Table 1. Genome data for Celastrina argiolus, ilCelArgi3.1.

| Project accession data                  |
|-----------------------------------------|
| Assembly identifier                     | ilCelArgi3.1                          |
| Species                                 | Celastrina argiolus                   |
| Specimen                                | ilCelArgi3 (genome assembly); ilCelArgi1, ilCelArgi4 (RNA-Seq) |
| NCBI taxonomy ID                        | NCBI:txid203782                       |
| BioProject                              | PRJEB41907                            |
| BioSample ID                            | SAMEA7523268                          |
| Isolate information                     | Male, whole organisms                 |

| Raw data accessions                     |
|-----------------------------------------|
| Pacific Biosciences SEQUEL II           | ERR6558180                            |
| 10X Genomics Illumina                   | ERR6002602-ERR6002605                 |
| Hi-C Illumina                          | ERR6002606                            |
| Illumina polyA RNA-Seq                 | ERR6002607, ERR6787413                |

| Genome assembly                         |
|-----------------------------------------|
| Assembly accession                      | GCA_905187575.1                       |
| Accession of alternate haplotype        | GCA_905147145.1                       |
| Span (Mb)                               | 499                                   |
| Number of contigs                       | 137                                   |
| Contig N50 length (Mb)                  | 8                                     |
| Number of scaffolds                     | 28                                    |
| Scaffold N50 length (Mb)                | 20                                    |
| Longest scaffold (Mb)                   | 29                                    |
| BUSCO* genome score                     | C:97.1% [S:96.7%, D:0.5%], F:0.6%, M:2.3%, 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/ilCelArgi3.1/dataset/CAJJIP01/busc.

resulting annotation includes 24,102 transcribed mRNAs from 12,199 protein-coding and 1,981 non-coding genes. There are 1.98 coding transcripts per gene and 8.65 exons per transcript.

Methods
Sample acquisition and nucleic acid extraction
Three male *C. argiolus* specimens (ilCelArgi3, genome assembly; ilCelArgi1 and ilCelArgi4, RNA-Seq) were collected from Oxford, England, UK (latitude 51.74989, longitude 1.22731) using a net by Alex Hayward, University of Exeter, who also identified the sample. The samples were frozen at -80°C.

DNA was extracted from the whole organism of ilCelArgi3 at the Wellcome Sanger Institute (WSI) Scientific Operations core from the whole organism using the Qiagen MagAttract HMW DNA kit, according to the manufacturer’s instructions. RNA (from the whole organism of ilCelArgi1 and ilCelArgi4) was extracted 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 read cloud DNA sequencing libraries were constructed
according to the manufacturers’ instructions. Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit. DNA and RNA sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi), Illumina HiSeq X (10X) and Illumina HiSeq 4000 (RNA-Seq) instruments. Hi-C data were also generated from the whole organism of ilCelArgi3 using the Arima v1.0 kit and sequenced on HiSeq X.

**Genome assembly**

Assembly was carried out with Hifiasm (Cheng et al., 2021); haplotypic duplication was identified and removed with
Figure 3. Genome assembly of *Celastrina argiolus*, iICelArgi3.1: 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/iICelArgi3.1/dataset/CAJJIP01/blob.purge_dups (Guan et al., 2020). One round of polishing was performed by aligning 10X Genomics read data to the assembly with longranger align, calling variants with freebayes (Garrison & Marth, 2012). The assembly was then scaffolded with Hi-C data (Rao et al., 2014) using SALSA2 (Ghurye et al., 2019). The assembly was checked for contamination and corrected using the gEVAL system (Chow et al., 2016) as described previously (Howe et al., 2021). Manual curation (Howe et al., 2021) was performed using gEVAL, HiGlass (Kerpedjiev et al., 2018) and Pretex. The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva et al., 2021), which performed annotation using MitoFinder (Allio et al., 2020).
Figure 4. Genome assembly of *Celastrina argiolus*, ilCelArgi3.1: 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/ilCelArgi3.1/dataset/CAJJIP01/cumulative.

Gene annotation
The Ensembl gene annotation system (Aken et al., 2016) was used to generate annotation for the *Celastrina argiolus* assembly.

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.
Figure 5. Genome assembly of *Celastrina argiolus*, ilCelArgi3.1: Hi-C contact map. Hi-C contact map of the ilCelArgi3.1 assembly, visualised in HiGlass. Chromosomes are shown in size order from left to right and top to bottom.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Celastrina argiolus*, ilCelArgi3.1.

| INSDC accession | Chromosome | Size (Mb) | GC% |
|-----------------|------------|-----------|-----|
| LR994577.1      | 1          | 29.05     | 36.2|
| LR994578.1      | 2          | 24.85     | 36  |
| LR994579.1      | 3          | 24.55     | 35.8|
| LR994580.1      | 4          | 24.51     | 36.2|
| LR994581.1      | 5          | 24.42     | 35.8|
| LR994582.1      | 6          | 24.04     | 36.1|
| LR994584.1      | 7          | 21.46     | 36.1|
| LR994585.1      | 8          | 21.39     | 35.9|
| LR994586.1      | 9          | 20.68     | 35.7|
| LR994587.1      | 10         | 20.43     | 36.3|
| LR994588.1      | 11         | 19.02     | 36.1|
| LR994589.1      | 12         | 18.92     | 35.9|
| LR994590.1      | 13         | 18.52     | 36.4|
| LR994591.1      | 14         | 18.52     | 36.2|
| LR994592.1      | 15         | 18.42     | 35.9|
| LR994593.1      | 16         | 18.32     | 36.1|
| LR994594.1      | 17         | 17.10     | 36.4|
| LR994595.1      | 18         | 16.97     | 36  |
| LR994596.1      | 19         | 16.88     | 36.2|
| LR994597.1      | 20         | 16.52     | 36.5|
| LR994598.1      | 21         | 16.32     | 36.3|
| LR994599.1      | 22         | 15.74     | 36.2|
| LR994600.1      | 23         | 11.45     | 36.9|
| LR994601.1      | 24         | 10.30     | 37  |
| LR994602.1      | 25         | 6.95      | 36.4|
| LR994583.1      | Z          | 23.77     | 35  |
| LR994603.1      | MT         | 18.00     | 18  |
| -               | Unplaced   | 0.02      | 49  |
Table 3. Software tools used.

| Software tool       | Version | Source                              |
|---------------------|---------|-------------------------------------|
| Hifiasm             | 0.7     | Cheng et al., 2021                  |
| purge_dups          | 1.2.3   | Guan et al., 2020                   |
| SALSA2              | 2.2     | Ghurye et al., 2019                 |
| longranger align    | 2.2.2   | https://support.10xgenomics.com/     |
|                     |         | genome-exome/software/pipelines/latest/advanced/other-pipelines |
| freebayes           | 1.3.1-17-gaa2ace8 | Garrison & Marth, 2012             |
| 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               |

(GCA_905187575.1; 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.

Data availability
European Nucleotide Archive: Celastrina argiolus (holly blue) genome assembly, iiCelArgi3. Accession number PRJEB41907: https://www.ebi.ac.uk/ena/browser/view/PRJEB41907

The genome sequence is released openly for reuse. The C. argiolus 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 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.4783586.

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

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

Allio R, Schomaker-Bastos A, Romeiguier 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

Bigger T: Chromosome Numbers of Lepidoptera. Part II. Entomologist's Gazette. 1961; 12: 85–89.

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

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
Open Peer Review

Current Peer Review Status: ✔️ ✔️

Version 1

Reviewer Report 04 January 2022

https://doi.org/10.21956/wellcomeopenres.19326.r47621

© 2022 Liang B. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Bin Liang

College of Life Sciences, Inner Mongolia University, Hohhot, China

The authors submitted a chromosome-level genome assembly of the holly blue, Celastrina argiolus. In the method part, the authors scientifically described how they finished sampling, genome DNA and RNA sequencing, assembly and annotation. In this good work, the authors combined several popular sequencing techniques to guarantee the assembly quality, including a long-read sequencing, Pacbio HiFi, 10X Genomics sequencing, RNA-seq and Hi-C mapping. All results are perfect. In order to verify the identity of the sample, I suggest author had better conduct nucleotide blast in ncbi database using their assembly of the mitochondrial genome. I believe they already did it and can show sequence identity comparing with published mitochondrial DNA of Celastrina argiolus. All protocols and software are described clearly to ensure the reader can follow the authors’ performance.

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: Phylogeny, Comparative genomics, Mitochondrial genome

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 16 December 2021

https://doi.org/10.21956/wellcomeopenres.19326.r47619

© 2021 Lalueza-Fox C. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Carles Lalueza-Fox
Institute of Evolutionary Biology (CSIC-UPF), Barcelona, Spain

The paper presents the assembly and annotation of a common Lycaenidae butterfly, the holly blue (Celastrina argiolus). The methods are clearly written and well described, and the standard parameters of quality (such as N50 of 20Mb or BUSCO over 97%) indicates this a sound annotated genome that can be of utility not only to explore the holly blue's own evolutionary traits, but also, by means of its widespread distribution, to explore adaptive strategies to different environmental conditions. The authors cite the relevant literature for this common and yet understudied butterfly; precisely in Mackintosh et al. (2019)¹, the authors compare the genetic diversity across a large dataset of butterflies, finding that the holly blue, despite being a generalist, has a relatively low genetic diversity (as compared to other butterflies, some of them more geographically restricted). I think this could also be mentioned in the Introduction because, obviously, a reference genome for a species with low diversity is scientifically much more valuable than one for a species that can be highly structured and displays high diversity parameters.

References
1. Mackintosh A, Laetsch D, Hayward A, Charlesworth B, et al.: The determinants of genetic diversity in butterflies. Nature Communications. 2019; 10 (1). Publisher Full Text

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

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

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

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolutionary 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.