DATA NOTE

The genome sequence of the small copper, *Lycaena phlaeas* (Linnaeus, 1760) [version 1; peer review: 1 approved with reservations]

Konrad Lohse¹, Dominik R. Laetsch¹, Roger Vila²,
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

¹Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
²Institut de Biologia Evolutiva (CSIC - Universitat Pompeu Fabra), Barcelona, Spain

Abstract

We present a genome assembly from an individual male *Lycaena phlaeas* (the small copper; Arthropoda; Insecta; Lepidoptera; Lycaenidae). The genome sequence is 420 megabases in span. The whole of the assembly is scaffolded into 24 chromosomal pseudomolecules, with the Z sex chromosome assembled. Gene annotation of this assembly on Ensembl has identified 12,147 protein coding genes.

Keywords

*Lycaena phlaeas*, small copper, genome sequence, chromosomal

This article is included in the Tree of Life gateway.
Species taxonomy
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Papilionoidea; Lycaenidae; Lycaeninae; Lycaena; Lycaena phlaeas (Linnaeus, 1761) (NCBI:txid282391).

Background
The small copper or common copper, Lycaena phlaeas, is a colourful butterfly with a vast geographical distribution throughout the Holarctic and North and East Central Africa (Tolman & Lewington, 2008). The species is listed as Least Concern in the IUCN Red List (Europe) (Van Swaay et al., 2009). On the British Isles, L. phlaeas can be found widely distributed in Southern England and Wales, and in smaller numbers in Northern England, Scotland and Ireland. It inhabits a wide range of ecosystems from chalk downlands, heathland, woodland clearings to churchyards and urban localities. Males are territorial and bask in patches of bare ground, on rocks or perching on grass stems. Females lay individual eggs on stems or leaves of host plants such as sorrel (Rumex spp.) or Buckwheats (Polygonaceae spp.) and the larvae overwinter as caterpillars. The number of annual generations are typically two or three, but in southern localities diapause seems to be skipped during warm winters. Contrary to most other Lycaenids, the myrmecoxenous L. phlaeas lacks both a dorsal nectar organ and tentacle organs, and its larvae are only weakly attended by ants in behavioural interaction experiments (Fiedler, 1991). The standard haploid karyotype of L. phlaeas consists of 24 autosomes and one sex chromosome (Lorković, 1941), the female is heterogametic (WZ), and the genome size has been estimated as 359.7 Mb (Mackintosh et al., 2019).

Genome sequence report
The genome was sequenced from a single male L. phlaeas collected from Gullane, Scotland, UK (latitude 56.04, longitude -2.82) (Figure 1). A total of 55-fold coverage in Pacific Biosciences single-molecule long reads and 89-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 4 missing/misjoins and removed 1 haplotypic duplication, reducing the assembly size by 0.02%, scaffold number by 4.00% and the scaffold N50 by 3.76%.

The final assembly has a total length of 420 Mb in 25 sequence scaffolds with a scaffold N50 of 18 Mb (Table 1). Of the assembly sequence, 100% was assigned to 24 chromosomal-level scaffolds, representing 23 autosomes (numbered by sequence length), and the Z sex chromosome (Figure 2–Figure 5; Table 2). The assembly has a BUSCO (Simão et al., 2015) v5.1.2 completeness of 98.3% (single 98.0%, duplicated 0.3%, fragmented 0.4%, missing 1.3%) 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 Lycaena phlaeas assembly (GCA_905333005.1; 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).

Figure 1. Fore and hind wings of Lycaena phlaeas specimen from which the genome was sequenced. (A) Dorsal surface view of wings from specimen SC_LP_1398 (iLycPhla1) from Gullane, Scotland, UK, used to generate Pacific Biosciences, 10X genomics, Hi-C and RNA-Seq data. (B) Ventral surface view of wings from specimen SC_LP_1398 (iLycPhla1) from Gullane, Scotland, UK, used to generate Pacific Biosciences, 10X genomics, Hi-C and RNA-Seq data.
| **Project accession data** |  |
|---------------------------|--|
| Assembly identifier       | iLycPhla1.1 |
| Species                   | *Lycaena phlaeas* |
| Specimen                  | iLycPhla1 |
| NCBI taxonomy ID          | NCBI:txid282391 |
| BioProject                | PRJEB43473 |
| BioSample ID              | SAMEA7523293 |
| Isolate information       | Male, whole organism |

| **Raw data accessions**    |  |
|---------------------------|--|
| PacificBiosciences SEQUEL II | ERR6576321 |
| 10X Genomics Illumina      | ERR6054508-ERR6054511 |
| Hi-C Illumina             | ERR6054512 |

| **Genome assembly**        |  |
|---------------------------|--|
| Assembly accession        | GCA_905333005.1 |
| Accession of alternate haplotype | GCA_905333065.1 |
| Span (Mb)                  | 420 |
| Number of contigs          | 28 |
| Contig N50 length (Mb)     | 18 |
| Number of scaffolds        | 25 |
| Scaffold N50 length (Mb)   | 18 |
| Longest scaffold (Mb)      | 27 |
| BUSCO* genome score        | C:98.3%(S:98.0%, D:0.3%), F:0.4%, M:1.3%, n:5286 |

| **Gene annotation**        |  |
|---------------------------|--|
| Number of protein coding genes | 12,147 |
| Average coding sequence length (bp) | 1,654 |
| Average number of exons per transcript | 9 |
| Average exon size (bp)     | 305 |
| Average intron size (bp)   | 2,319 |

*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/iLycPhla1.1/dataset/iLycPhla1_1/busc](https://blobtoolkit.genomehubs.org/view/iLycPhla1.1/dataset/iLycPhla1_1/busc).

Table 1. Genome data for *Lycaena phlaeas*, iLycPhla1.1.

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.

Methods
Sample collection, DNA extraction and sequencing
A single male *L. phlaeas* was collected from Gullane, Scotland UK (latitude 56.04, longitude -2.82) using a handnet by Konrad Lohse, University of Edinburgh, and identified by the same individual. The specimen was snap-frozen in liquid nitrogen.

DNA was extracted from the whole organism using the MagAttract HMW DNA Extraction kit, according to the manufacturer’s instructions at the Scientific Operations core at the Wellcome Sanger Institute (WSI). Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud sequencing.
Figure 2. Genome assembly of *Lycaena phlaeas*, ilLycPhla1.1: 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 420,484,508 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 (26,507,414 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (18,312,151 and 13,920,109 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.genomeshubs.org/view/ilLycPhla1.1/dataset/ilLycPhla1_1/snail.
libraries were then constructed according to the manufacturers’ instructions. Sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II and Illumina HiSeq X instruments. Hi-C data were also generated from the whole organism using the Arima v2.0 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 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 3. Genome assembly of Lycaena phlaeas, iLycPhla1.1: GC coverage.** BlobToolKit GC-coverage plot. Chromosomes are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of chromosome length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/iLycPhla1.1/dataset/iLycPhla1_1/blob.
**Figure 4. Genome assembly of *Lycaena phlaeas*, iLycPhla1.1: 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/iLycPhla1.1/dataset/iLycPhla1_1/cumulative](https://blobtoolkit.genomehubs.org/view/iLycPhla1.1/dataset/iLycPhla1_1/cumulative).

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 Pretex. The mitochondrial genome was assembled 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.

**Ethical/compliance issues**

The materials that have contributed to this genome note were supplied by a Tree of Life collaborator. The WSI employs a process whereby due diligence is carried out proportionate to
### Table 2. Chromosomal pseudomolecules in the genome assembly of *Lycaena phlaeas*, ilLycPhla1.1.

| INSDC accession | Chromosome | Size (Mb) | GC%  |
|-----------------|------------|-----------|------|
| HG995163.1      | 1          | 26.51     | 35.4 |
| HG995164.1      | 2          | 23.18     | 35.6 |
| HG995165.1      | 3          | 23.03     | 35.4 |
| HG995166.1      | 4          | 21.48     | 35.2 |
| HG995167.1      | 5          | 21.46     | 35.2 |
| HG995169.1      | 6          | 19.66     | 35.8 |
| HG995170.1      | 7          | 19.03     | 36   |
| HG995171.1      | 8          | 19.03     | 35.5 |
| HG995172.1      | 9          | 18.31     | 35.8 |
| HG995173.1      | 10         | 18.08     | 35.4 |
| HG995174.1      | 11         | 18.07     | 35.4 |
| HG995175.1      | 12         | 16.78     | 36   |
| HG995176.1      | 13         | 16.08     | 35.7 |
| HG995177.1      | 14         | 15.90     | 36.1 |
| HG995178.1      | 15         | 15.87     | 36   |
| HG995179.1      | 16         | 15.57     | 36   |
| HG995180.1      | 17         | 15.14     | 36.1 |
| HG995181.1      | 18         | 14.89     | 36.3 |
| HG995182.1      | 19         | 14.38     | 36.8 |
| HG995183.1      | 20         | 13.92     | 36.4 |
| HG995184.1      | 21         | 13.28     | 36.5 |
| HG995185.1      | 22         | 10.48     | 37.2 |
| HG995186.1      | 23         | 9.81      | 37.5 |
| HG995187.1      | Z          | 10.48     | 37.2 |
| HG995188.1      | MT         | 0.02      | 17.2 |

**Hi-C contact map of the ilLycPhla1.1 assembly, visualised in HiGlass. Chromosomes are shown in size order from left to right and top to bottom.**
the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible.

The overarching areas of consideration are:

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

Each transfer of samples is 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: Lycaena phlaeas (small copper). Accession number PRJEB43533; https://identifiers.org/ena.embl/PRJEB43533.

The genome sequence is released openly for reuse. The *L. phlaeas* 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.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

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

Camargo AP, Sourkov V, Pereira GAG, et al.: *RNA2samba: Neural Network-Based Assessment of the Protein-Coding Potential of RNA Sequences.* NAR Genom Bioinform. 2020; 2(1): lqo024.  
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

Chow W, Brugger K, Caccamo M, et al.: *gEVAL — a Web-Based Browser for Evaluating Genome Assemblies.* Bioinformatics. 2016; 32(16): i2508-10.  
PubMed Abstract | Publisher Full Text | Free Full Text

Fiedler K: *Systematic, Evolutionary, and Ecological Implications of Myrmecophily within the Lycaenidae (Insecta: Lepidoptera: Papilionoidea).* Bonn Zool Monogr. 1991; 31: 1-210.  
Reference Source

Garrison E, Martin G: *Haplotype-Based Variant Detection from Short-Read Sequencing.* arXiv: 1207.3907. 2012.  
Reference Source

Ghurye J, Rhee A, Walenz BP, et al.: *Integrating Hi-C Links with Assembly Graphs for Chromosome-Scale Assembly.* PLoS Comput Biol. 2019; 15(6): e1007273.  
PubMed Abstract | Publisher Full Text | Free Full Text

Guan D, McCarthy SA, Wood J, et al.: *Integrating Hi-C Links with Assembly Graphs for Chromosome-Scale Assembly.* PLoS Comput Biol. 2019; 15(1): e1007273.  
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(12): 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
Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 22 November 2021

https://doi.org/10.21956/wellcomeopenres.19115.r46969

© 2021 Bombarely A. 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.

Aureliano Bombarely
Institute of Plant Molecular and Cell Biology (IBMCP), Valencia, Spain

Summary of the key results:
The manuscript titled “The genome sequence of the small copper, Lycaena phlaeas (Linnaeus, 1760)” describes the genome sequencing, assembly and annotation of the species Lycaena phlaeas , a butterfly native to the Holarctic and North and East Central Africa. The genome was sequenced with a combination of the PacBio HiFi (55X the estimated genome size) and 10X Genomics (89X) technologies, and then, scaffolded into chromosomes using Hi-C. Manual curation was used to correct 4 missing/misjoins and to remove 1 haplotypic duplication. The final assembly was 28 contigs grouped in 25 scaffolds for a total size of 420 Mb. They were assigned to 24 chromosomes. The BUSCO completeness score was 98.3% (single 98.0%, duplicated 0.3%, fragmented 0.4%, missing 1.3%). The genome annotation predicted 12,147 protein coding gene models.

Overall evaluation:
Although the manuscript is a “Data Note” there are numerous details that are missing. A more extensive evaluation of the quality of the assembly and the annotation and a whole section for the material and methods (genome annotation) is needed from my point of view. See the major concerns for more details.

Overall, the genome of the species Lycaena phlaeas will be a valuable resource for the community. The quality of the assembly looks fine with the type of evaluation described in this publication (assembly stats and BUSCO scores). Other types of parameters such as K-mer completeness and QV could complement the evaluation and give it an extra support.

Major concerns:
- The final assembly partially evaluated in terms of quality. The manuscript only presents the assembly statistics and the BUSCO score as the tools to support the quality of the assembly. Currently, there are several tools that can be applied to have a more complete perspective of the quality of the assembly, such as Merqury, which can be used to evaluate also the completeness of the assembly as well as the accuracy of the consensus. Probably it could be interesting too to estimate the heterozygosity of the individual in that regard and
see if it agrees with the assembly of separate haplotypes.

○ **There is not any information about the repeat annotation of the genome.** Although the gene annotation information is present, there is not any information about the repeat annotation process which it could be a key for the genome annotation quality.

○ **Many details are missing in the material and method section.** There are many details missing in this section, from the parameters used for the different assembly tools (if default parameters were used, it needs to be specified) to the whole annotation section that it is completely missing.

**Minor concerns:**

○ **No information about the second haplotype was supplied.** The manuscript text mentions “Contigs corresponding to the second haplotype have also been deposited” but no stats or any other information was supplied.

○ **There is not a phylogenetic analysis of the mitochondrial genome to verify the sample.** Usually, it is convenient to perform a phylogenetic analysis of the mitochondrial genome (or chloroplast for plants) to verify the identity of the sequenced taxon as quality control.

○ **Information about the sequenced libraries has not been supplied.** It is recommended also to specify in a table the amount of sequencing performed for each of the libraries as number of reads, total Gb... This type of information could include also the number of contacts detected for the Hi-C library.

○ **No BUSCO analysis for the annotation.** Although it was performed on the genome assembly, it wasn't run on the annotation. This is necessary to have an idea of how well the annotation went (in a rough way).

○ **The assembly quality could be described in the context of the EBP standards.** The Earth Biogenome Project proposed some quality standards for the genome assemblies (https://www.earthbiogenome.org/assembly-standards). It will be useful to have one sentence describing how well this assembly meets these standards.

**Discretionary comments:**

○ Sometimes the species name is not in italics.

**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:** No competing interests were disclosed.

**Reviewer Expertise:** Genomics (Genome assembly, annotation, evolution); Bioinformatics (Development of genomic tools including databases); Plant Evolution (Polyploidy and Domestication);

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.