DATA NOTE

# The genome sequence of the Large Blue butterfly, *Phengaris* (*Maculinea*) *arion* (Linnaeus, 1758)

[version 1; peer review: 4 approved with reservations]

Sarah A. Meredith[1], David J. Simcox[1], Jeremy A. Thomas[1,2], Alan Sumnall[3], Peter W. H. Holland[2], Liam M. Crowley[2],
University of Oxford and Wytham Woods Genome Acquisition Lab,
Darwin Tree of Life Barcoding collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]Royal Entomological Society, St Albans, England, UK
[2]University of Oxford, Oxford, England, UK
[3]Gloucestershire Wildlife Trust, Gloucester, England, UK

**v1** **First published:** 03 Sep 2024, **9**:506
https://doi.org/10.12688/wellcomeopenres.22984.1
**Latest published:** 03 Sep 2024, **9**:506
https://doi.org/10.12688/wellcomeopenres.22984.1

## Abstract

We present a genome assembly from a female *Phengaris arion* (the Large Blue butterfly; Arthropoda; Insecta; Lepidoptera; Lycaenidae). The genome sequence is 544.50 megabases in length. Most of the assembly is scaffolded into 23 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled and is 15.7 kilobases in length.

## Keywords

Phengaris arion, Large Blue butterfly, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

**Open Peer Review**

**Approval Status** ? ? ? ?

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **version 1** 03 Sep 2024 | ? view | ? view | ? view | ? view |

1. **Jesper Boman**, Uppsala University, Uppsala, Sweden
2. **Li-Wei Wu**, Tunghai University, Taichung, Taiwan
3. **Arjen Van 't Hof**, Biology Centre of the Czech Academy of Sciences, Budějovice, Czech Republic
4. **Paula Escuer Pifarré**, University of Neuchâtel, Neuchâtel, Switzerland

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: Meredith SA**: Investigation, Resources; **Simcox DJ**: Investigation, Resources; **Thomas JA**: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Sumnall A**: Investigation, Resources; **Holland PWH**: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Crowley LM**: Investigation, Resources;

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported by Wellcome through core funding to the Wellcome Sanger Institute [206194, https://doi.org/10.35802/206194] and the Darwin Tree of Life Discretionary Award [218328, https://doi.org/10.35802/218328 ].
*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:** Meredith SA, Simcox DJ, Thomas JA *et al.* **The genome sequence of the Large Blue butterfly, *Phengaris* ( *Maculinea*) *arion* (Linnaeus, 1758) [version 1; peer review: 4 approved with reservations]** Wellcome Open Research 2024, **9**:506 https://doi.org/10.12688/wellcomeopenres.22984.1

**First published:** 03 Sep 2024, **9**:506 https://doi.org/10.12688/wellcomeopenres.22984.1

## Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Amphiesmenoptera; Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Papilionoidea; Lycaenidae; Polyommatinae; *Phengaris*; *Phengaris arion* Linnaeus, 1758 (NCBI:txid203779).

**Nomenclature:** From 1915 to 2017, all species of Large Blue butterfly were assigned to the genus *Maculinea*. The majority of publications describing their biology use this generic name, as does current legislation to protect these endangered species. In 2017 the *International Commission on Zoological Nomenclature* accepted the proposal of Fric *et al.* (2007) to use a senior synonym, *Phengaris* (Opinion 2399: Case 3508), having rejected the submission of Balletto *et al.* (2010) to conserve the genus name *Maculinea*.

## Background

The Lycaenidae is the second-largest family of butterflies with over 6,000 species worldwide. About 75% of lycaenids are known to interact with ants as larvae or pupae, living as phytophagous mutualists that provide amino-acids and sugars in exchange for protection from enemies (Pierce *et al.*, 2002). From these mutualists, about 3% of lycaenids have evolved to become social parasites of ant colonies (Fiedler, 1998). The best understood are the various species of Large Blue (Settele *et al.*, 2005; Thomas *et al.*, 2005), a Palaearctic genus of which *P. arion* is the only British representative. *P. arion* is famed for its endangered status and the extraordinary specialisations that enable the larvae to infiltrate and exploit ant societies (Thomas & Settele, 2004).

The Large Blue *P. arion* survives in isolated populations across Europe and as far east as Volograd (Russia) and Altai (Kazakhstan), although some molecular analyses suggest that the latter may represent a distinct cryptic species (Als *et al.*, 2004; Ugelvig *et al.*, 2011). Always rare, the species was declared extinct in Britain in 1979, and from 1990–2010 it experienced a catastrophic decline of over 90% across all of Europe (Thomas *et al.*, 2009; Van Swaay *et al.*, 2010). Here we describe the genome of a female Large Blue from Britain whose ancestors were translocated 30 years previously (30 generations) from the Stora Alvaret limestones of Öland, Sweden to a restored site in Somerset, 13 years after extinction of the last native British colony (Thomas *et al.*, 2009; Thomas *et al.*, 2019). The Swedish populations provided Linnaeus' type specimen and are genetically close to surviving colonies near St Petersburg and in Spain (Als *et al.*, 2004). In 1992, 281 final instar larvae, collected as eggs from 11 sub-populations on Öland, were released into host-ant territories on Green Down, Somerset, where they experienced burgeoning growth to form the largest known population in Europe (Thomas *et al.*, 2009). After establishment, 671 larvae were translocated from Green Down to restored grasslands on Daneway Banks in Gloucestershire, founding another large population (Thomas *et al.*, 2019). The specimen whose genome is reported here is from Daneway Banks.

Adult Large Blues emerge from pupae within host *Myrmica* ant nests from late May to early July. After mating, eggs are laid between the flowerbuds of thyme (*Thymus* spp) or marjoram (*Origanum vulgare*). Once hatched, the larva feeds on the flowers and seeds of its foodplants, developing rapidly but acquiring little weight (Thomas & Wardlaw, 1992). After its third moult, the larva is just 1% to 2% of its ultimate body weight, which is achieved between mid-summer and the following spring without further skin sheds by eating ant grubs inside *Myrmica* nests (Elmes *et al.*, 2001). To infiltrate a host ant society, the newly moulted larva falls from its foodplant and awaits discovery by a foraging *Myrmica* worker, which is attracted by honeydew secretions. After 30–90 minutes being 'milked', the larva mimics an escaped *Myrmica* larva by compressing its body and rearing on its prolegs (Frohawk, 1914; Thomas, 2002). This stimulates the worker to carry the *P. arion* into its underground nest and place it among the ant brood, on which it feasts (Thomas & Wardlaw, 1992). Although retrieved indiscriminately by up to five species of *Myrmica*, larval survival occurs almost exclusively with a single species, *Myrmica sabuleti* for most European *P. arion* (Tartally *et al.*, 2019; Thomas *et al.*, 1989). Acceptance is achieved in Large Blues by secretion of a cocktail of hydrocarbons that mimics pheromones of the host *Myrmica* species (Akino *et al.*, 1999; Thomas *et al.*, 2013), and survival is enhanced through mimicking the distinctive acoustics of an adult *Myrmica* queen (Barbero *et al.*, 2009). The knowledge that *P. arion* populations depend on high densities of *M. sabuleti*, combined with an understanding of this thermophilous ant's narrow niche, has enabled conservationists to restore *P. arion's* specialised habitat at multiple degraded former sites and to begin a long-term conservation programme restoring this iconic species to Britain.

Here we report a complete genome sequence for the Large Blue butterfly *Phengaris arion* determined as part of the Darwin Tree of Life project. The genome assembly will be useful as a reference sequence against which lower coverage data from other individuals can be compared, facilitating research into local adaptation, genetic diversity and gene flow between populations. The genome sequence will also facilitate research into the taxonomy of this endangered genus and the biochemical basis of brood parasitism in insects.

## Genome sequence report

The genome was sequenced from one female *Phengaris arion* (Figure 1) caught as a fresh adult on Daneway Banks Site of Special Scientific Interest, Gloucestershire, UK (51.43, –2.05). The natural origin of this population is Öland, Sweden (56.29, –16.28) from where it was re-introduced to Britain in 1992.

The genome of an *Phengaris arion* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of 84.99 Gb (gigabases) from 7.03 million reads, providing approximately 142-fold coverage. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data, which produced 110.65 Gbp from 732.78 million reads, yielding an approximate coverage of 203-fold. Specimen and sequencing information is summarised in Table 1.


**Figure 1. Photograph of the *Phengaris arion* specimen (ilPheArio1) used for genome sequencing.** Its wings are preserved in the Hope Entomological Collections at the University of Oxford.

**Table 1. Specimen and sequencing data for *Phengaris arion*.**

| **Project information** | | | |
|---|---|---|---|
| **Study title** | Phengaris arion (large blue) | | |
| **Umbrella BioProject** | PRJEB65386 | | |
| **Species** | *Phengaris arion* | | |
| **BioSample** | SAMEA112232493 | | |
| **NCBI taxonomy ID** | 203779 | | |
| **Specimen information** | | | |
| **Technology** | **ToLID** | **BioSample accession** | **Organism part** |
| **PacBio long read sequencing** | ilPheArio1 | SAMEA112232932 | thorax |
| **Hi-C sequencing** | ilPheArio2 | SAMEA112232934 | head |
| **Sequencing information** | | | |
| **Platform** | **Run accession** | **Read count** | **Base count (Gb)** |
| **Hi-C Illumina NovaSeq 6000** | ERR11904116 | 7.33e+08 | 110.65 |
| **PacBio Revio** | ERR11892476 | 7.03e+06 | 84.99 |

Manual assembly curation corrected 108 missing joins or mis-joins, reducing the scaffold number by 76.0%. The final assembly has a total length of 544.50 Mb in 29 sequence scaffolds with a scaffold N50 of 24.0 Mb (Table 2). The total count of gaps in the scaffolds is 165. The snail plot in Figure 2 provides a summary of the assembly statistics, while Figure 3 shows the distribution of base coverage against position per chromosome. The cumulative assembly plot in Figure 4 shows curves for subsets of scaffolds assigned to different phyla. Most (99.93%) of the assembly sequence was assigned to 23 chromosomal-level scaffolds, representing 22 autosomes and the Z sex chromosome. The specimen is a ZO female. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 3). The order and orientation of contigs on Chromosome 2 from 1 to approximately 3.63 Mb is uncertain. While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial genome was also assembled and can be found as a contig within the multifasta file of the genome submission.

The estimated Quality Value (QV) of the final assembly is 61.2 with *k*-mer completeness of 100.0%, and the assembly

**Table 2. Genome assembly data for *Phengaris arion*, ilPheArio1.1.**

| **Genome assembly** | | |
|---|---|---|
| Assembly name | ilPheArio1.1 | |
| Assembly accession | GCA_963565745.1 | |
| *Accession of alternate haplotype* | *GCA_963565765.1* | |
| Span (Mb) | 544.50 | |
| Number of contigs | 195 | |
| Contig N50 length (Mb) | 11.0 | |
| Number of scaffolds | 29 | |
| Scaffold N50 length (Mb) | 24.0 | |
| Longest scaffold (Mb) | 48.13 | |
| **Assembly metrics*** | | ***Benchmark*** |
| Consensus quality (QV) | 61.2 | *≥ 50* |
| *k*-mer completeness | 100.0% | *≥ 95%* |
| BUSCO** | C:97.4%[S:97.0%,D:0.4%], F:0.5%,M:2.1%,n:5,286 | *C ≥ 95%* |
| Percentage of assembly mapped to chromosomes | 99.93% | *≥ 95%* |
| Sex chromosomes | ZO | *localised homologous pairs* |
| Organelles | Mitochondrial genome: 15.7 kb | *complete single alleles* |

* Assembly metric benchmarks are adapted from column VGP-2020 of "Table 1: Proposed standards and metrics for defining genome assembly quality" from Rhie *et al.* (2021).

** BUSCO scores based on the lepidoptera_odb10 BUSCO set using version 5.4.3. 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/Phengaris_arion/dataset/GCA_963565745.1/busco.

has a BUSCO v5.4.3 completeness of 97.4% (single = 97.0%, duplicated = 0.4%), using the lepidoptera_odb10 reference set ($n$ = 5,286).

Metadata for specimens, BOLD barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly statistics are given at https://links.tol.sanger.ac.uk/species/203779.

## Methods

### Sample acquisition

The *Phengaris arion* specimens used for genome sequencing (specimen ID Ox002251, ToLID ilPheArio1) and Hi-C scaffolding (specimen ID Ox002252, ToLID ilPheArio2) were adult females from Daneway Banks, Gloucestershire, UK (latitude 51.43, longitude –2.05) collected on 2022-07-04 by Sarah Meredith, Peter Holland, Liam Crowley, Alan Sumnall and Jeremy Thomas. The specimens were preserved on dry ice.

The initial identification was verified by an additional DNA barcoding process according to the framework developed by Twyford *et al.* (2024). A small sample was dissected from the specimens and stored in ethanol, while the remaining parts of the specimen were shipped on dry ice to the Wellcome Sanger Institute (WSI). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding have been deposited on protocols.io (Beasley *et al.*, 2023).

### Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the WSI Tree of Life Core Laboratory includes a

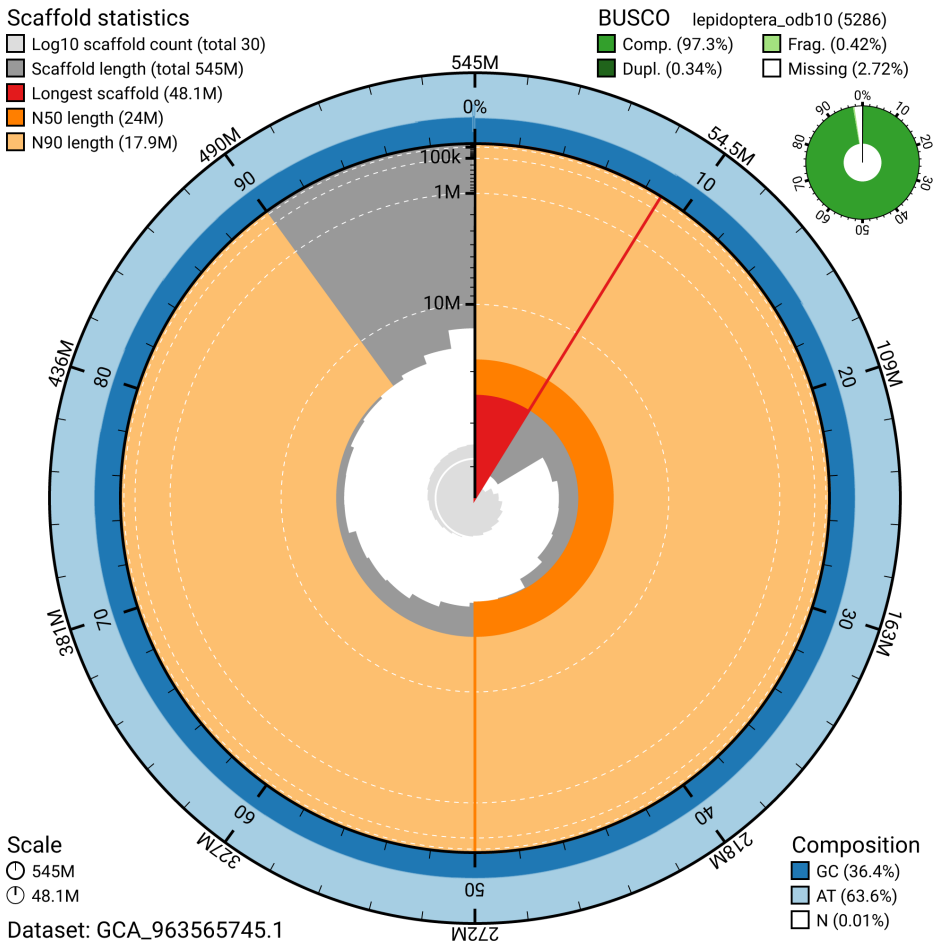


**Figure 2. Genome assembly of *Phengaris arion*, ilPheArio1.1: metrics.** The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 544,505,416 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (48,134,571 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (23,974,698 and 17,935,167 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/GCA_963565745.1/dataset/GCA_963565745.1/snail.

sequence of core procedures: sample preparation and homogenisation, DNA extraction, fragmentation, and purification. Detailed protocols are available on protocols.io (Denton *et al.*, 2023b).

In sample preparation, the ilPheArio1 sample was weighed and dissected on dry ice (Jay *et al.*, 2023). Tissue from the thorax was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a). HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023). The DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Bates *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Strickland *et al.*, 2023). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

## Sequencing

Pacific Biosciences HiFi circular consensus DNA sequencing libraries were constructed according to the manufacturers' instructions. DNA sequencing was performed by the Scientific Operations core at the WSI on a Pacific Biosciences Revio instrument. Hi-C data were also generated from head

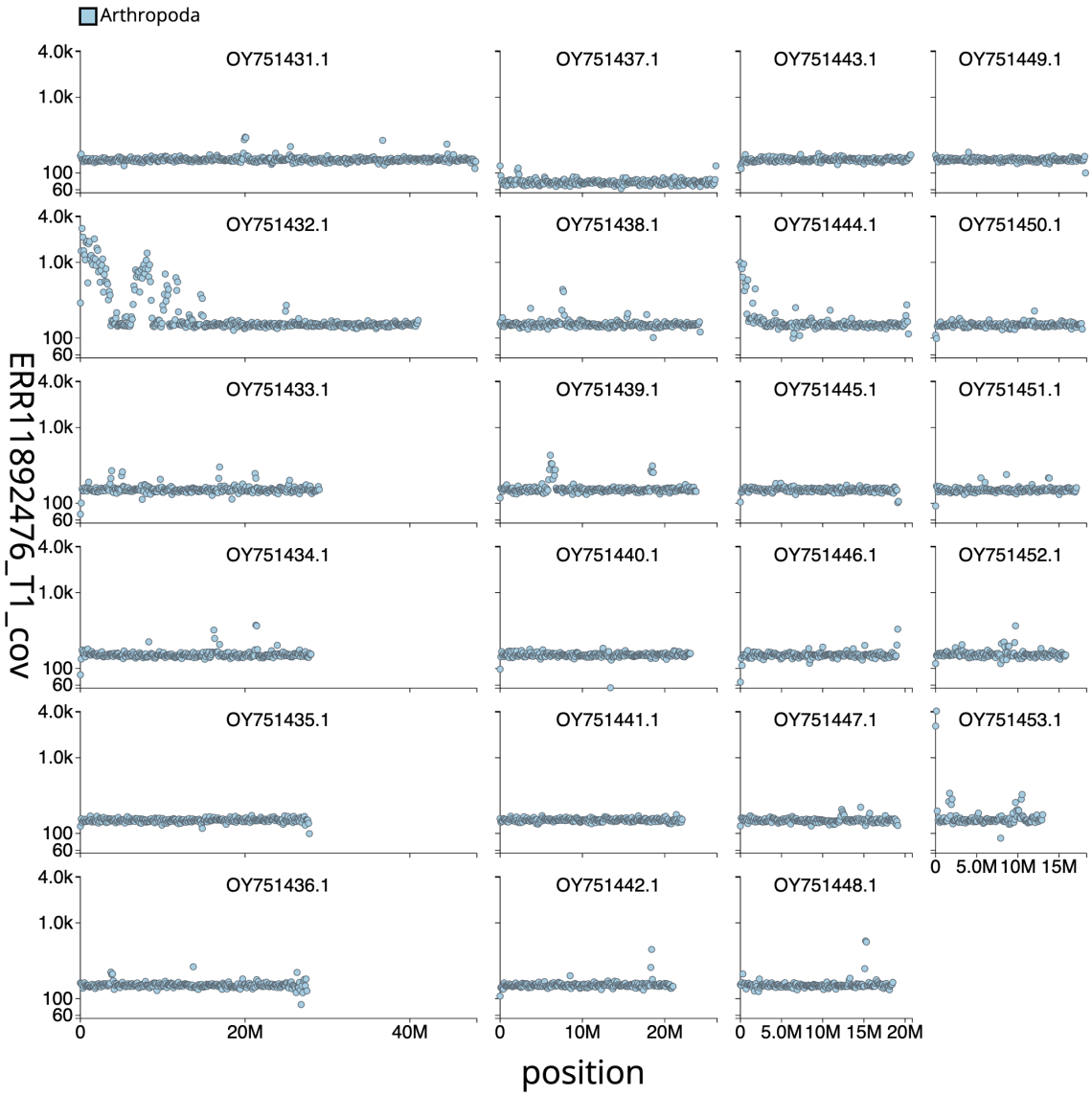


**Figure 3. Genome assembly of *Phengaris arion*, ilPheArio1.1: Distribution plot of base coverage in ERR11892476 against position for sequences in assembly GCA_963565745.1.** 100kb windows are coloured by phylum. The assembly has been filtered to exclude sequences with length < 2,550,000 An interactive version of this figure is available here.

tissue of ilPheArio2 using the Arima-HiC v2 kit. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on the Illumina NovaSeq 6000 instrument.

## Genome assembly, curation and evaluation

### *Assembly*

The HiFi reads were first assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). The Hi-C reads were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). The contigs were further scaffolded using the provided Hi-C data (Rao *et al.*, 2014) in YaHS (Zhou *et al.*, 2023) using the --break option. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

### *Assembly curation*

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline (article in preparation). Flat files and maps used in curation were

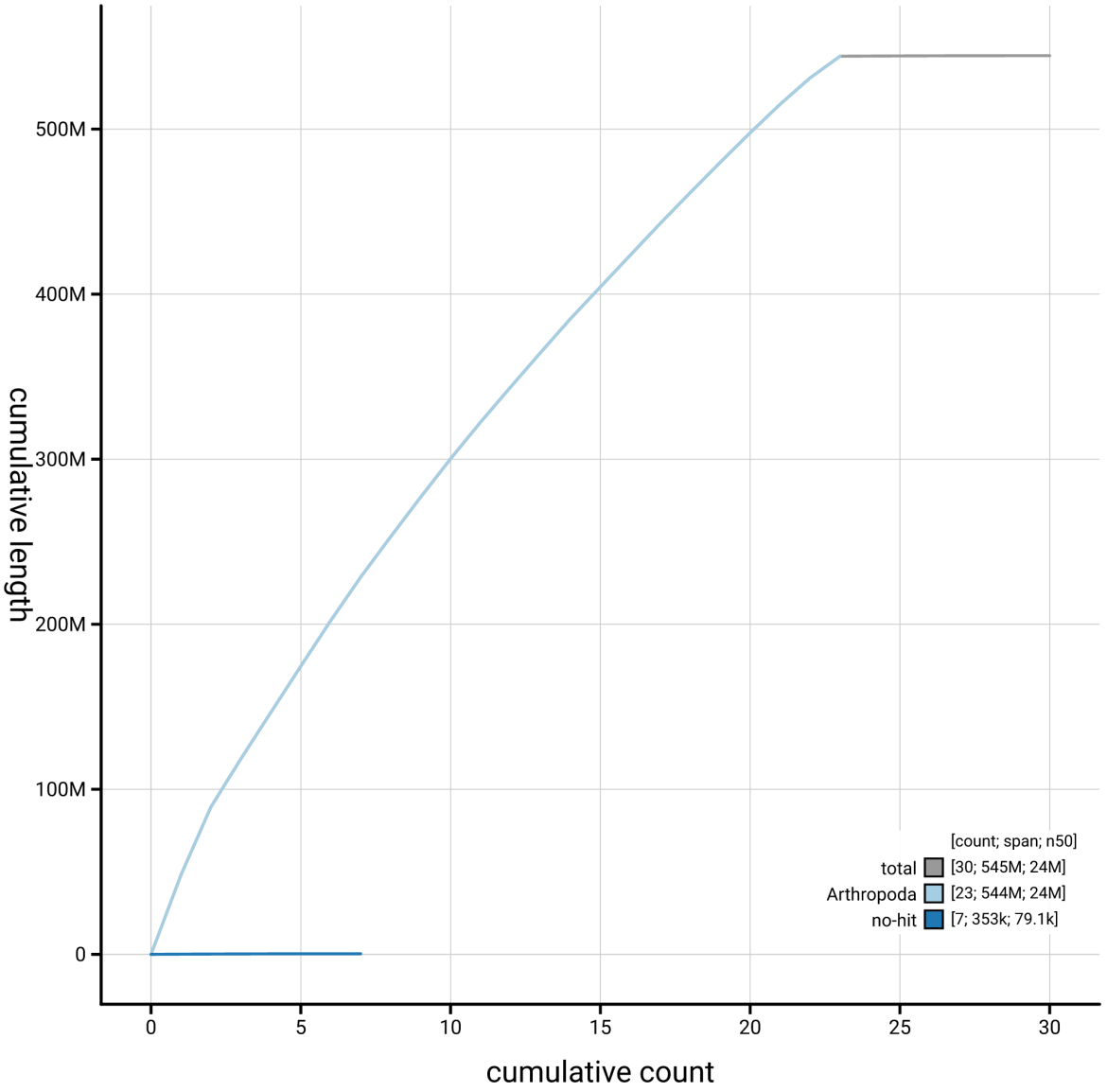


**Figure 4. Genome assembly of *Phengaris arion* ilPheArio1.1: BlobToolKit cumulative sequence plot.** The grey line shows cumulative length for all sequences. Coloured lines show cumulative lengths of sequences assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963565745.1/dataset/GCA_963565745.1/cumulative.

generated in TreeVal (Pointon *et al.*, 2023). Manual curation was primarily conducted using PretextView (Harry, 2022), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023) and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were corrected, and duplicate sequences were tagged and removed. The Z chromosome was identified based on read coverage statistics. The entire process is documented at https://gitlab.com/wtsi-grit/rapid-curation (article in preparation).

***Evaluation of the final assembly***

The final assembly was post-processed and evaluated with the three Nextflow (Di Tommaso *et al.*, 2017) DSL2 pipelines "sanger-tol/readmapping" (Surana *et al.*, 2023a), "sanger-tol/genomenote" (Surana *et al.*, 2023b), and "sanger-tol/blobtoolkit" (Muffato *et al.*, 2024). The pipeline sanger-tol/readmapping aligns the Hi-C reads with bwa-mem2 (Vasimuddin *et al.*, 2019) and combines the alignment files with SAMtools (Danecek *et al.*, 2021). The sanger-tol/genomenote pipeline transforms the Hi-C alignments into a contact map with BEDTools

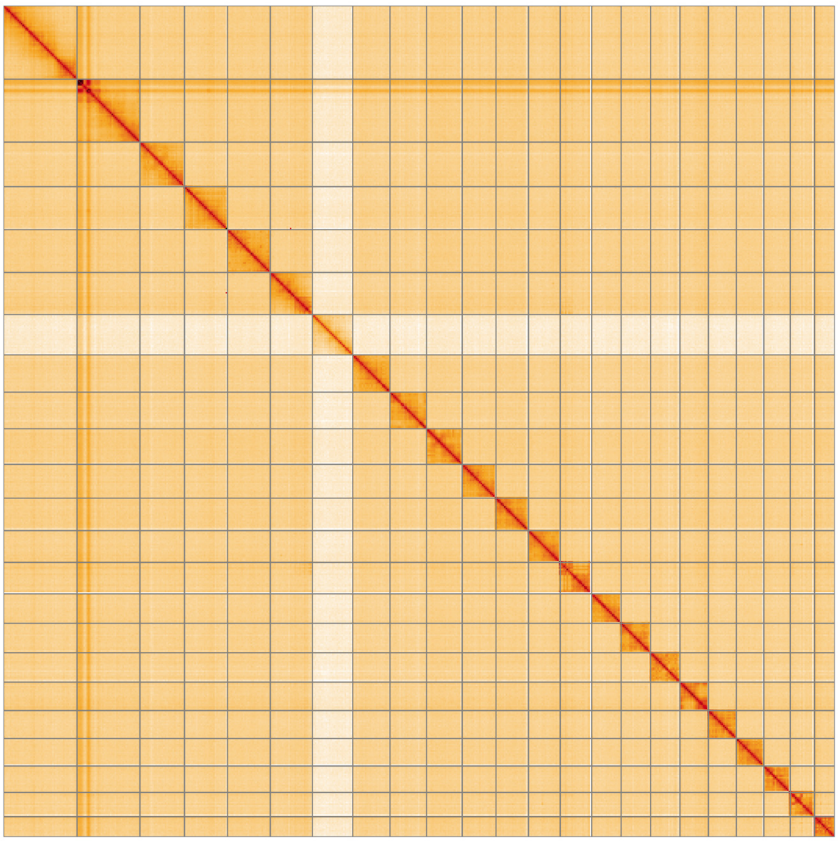

**Figure 5. Genome assembly of *Phengaris arion* ilPheArio1.1: Hi-C contact map of the ilPheArio1.1 assembly, visualised using HiGlass.** Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=S-3x7SQNTnqFEOXQdWo26A.

**Table 3. Chromosomal pseudomolecules in the genome assembly of *Phengaris arion*, ilPheArio1.**

| INSDC accession | Name | Length (Mb) | GC% |
|---|---|---|---|
| OY751431.1 | 1 | 48.13 | 36.0 |
| OY751432.1 | 2 | 41.19 | 37.5 |
| OY751433.1 | 3 | 29.16 | 36.0 |
| OY751434.1 | 4 | 28.11 | 36.0 |
| OY751435.1 | 5 | 27.99 | 36.5 |
| OY751436.1 | 6 | 27.65 | 36.0 |
| OY751438.1 | 7 | 24.42 | 36.0 |
| OY751439.1 | 8 | 23.97 | 36.0 |
| OY751440.1 | 9 | 23.25 | 36.5 |
| OY751441.1 | 10 | 22.2 | 36.0 |
| OY751442.1 | 11 | 21.19 | 36.5 |
| OY751443.1 | 12 | 20.88 | 36.5 |
| OY751444.1 | 13 | 20.53 | 36.5 |
| OY751445.1 | 14 | 19.31 | 36.5 |
| OY751446.1 | 15 | 19.28 | 36.5 |
| OY751447.1 | 16 | 19.22 | 36.0 |
| OY751448.1 | 17 | 18.63 | 37.0 |
| OY751449.1 | 18 | 18.34 | 36.5 |
| OY751450.1 | 19 | 17.94 | 36.5 |
| OY751451.1 | 20 | 17.3 | 36.5 |
| OY751452.1 | 21 | 15.93 | 36.0 |
| OY751453.1 | 22 | 13.17 | 37.5 |
| OY751437.1 | Z | 26.36 | 35.5 |
| OY751454.1 | MT | 0.02 | 17.0 |

(Quinlan & Hall, 2010) and the Cooler tool suite (Abdennur & Mirny, 2020), which is then visualised with HiGlass (Kerpedjiev *et al.*, 2018). It also provides statistics about the assembly with the NCBI datasets (Sayers *et al.*, 2024) report, computes *k*-mer completeness and QV consensus quality values with FastK and MERQURY.FK, and a completeness assessment with BUSCO (Manni *et al.*, 2021).

The sanger-tol/blobtoolkit pipeline is a Nextflow port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads with SAMtools and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoaT database (Challis *et al.*, 2023) to identify all matching BUSCO lineages to run BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineage, the pipeline aligns the BUSCO genes to the Uniprot Reference Proteomes database (Bateman *et al.*, 2023) with DIAMOND (Buchfink *et al.*, 2021) blastp. The genome is also split into chunks according to the density of the BUSCO genes from the closest taxonomically lineage, and each chunk is aligned to the Uniprot Reference Proteomes database with DIAMOND blastx. Genome sequences that have no hit are then chunked with seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). All those outputs are combined with the blobtools suite into a blobdir for visualisation.

The genome assembly and evaluation pipelines were developed using the nf-core tooling (Ewels *et al.*, 2020), use MultiQC (Ewels *et al.*, 2016), and make extensive use of the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), and the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

Table 4 contains a list of relevant software tool versions and sources.

## Wellcome Sanger Institute – Legal and Governance

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'**, which can be found in full on the Darwin Tree of Life website here. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they

**Table 4. Software tools: versions and sources.**

| Software tool | Version | Source |
|---|---|---|
| BEDTools | 2.30.0 | https://github.com/arq5x/bedtools2 |
| BLAST | 2.14.0 | ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/ |
| BlobToolKit | 4.3.7 | https://github.com/blobtoolkit/blobtoolkit |
| BUSCO | 5.4.3 and 5.5.0 | https://gitlab.com/ezlab/busco |
| bwa-mem2 | 2.2.1 | https://github.com/bwa-mem2/bwa-mem2 |
| Cooler | 0.8.11 | https://github.com/open2c/cooler |
| DIAMOND | 2.1.8 | https://github.com/bbuchfink/diamond |
| fasta_windows | 0.2.4 | https://github.com/tolkit/fasta_windows |
| FastK | 427104ea91c78c3b8b8b49f1a7d6bbeaa869ba1c | https://github.com/thegenemyers/FASTK |
| Gfastats | 1.3.6 | https://github.com/vgl-hub/gfastats |
| GoaT CLI | 0.2.5 | https://github.com/genomehubs/goat-cli |
| Hifiasm | 0.19.5-r587 | https://github.com/chhylp123/hifiasm |
| HiGlass | 44086069ee7d4d3f6f3f0012569789ec138f42b84a a44357826c0b6753eb28de | https://github.com/higlass/higlass |
| Merqury.FK | d00d98157618f4e8d1a9190026b19b471055b22e | https://github.com/thegenemyers/MERQURY.FK |
| MitoHiFi | 3 | https://github.com/marcelauliano/MitoHiFi |
| MultiQC | 1.14, 1.17, and 1.18 | https://github.com/MultiQC/MultiQC |
| NCBI Datasets | 15.12.0 | https://github.com/ncbi/datasets |
| Nextflow | 23.04.0-5857 | https://github.com/nextflow-io/nextflow |

| Software tool | Version | Source |
|---|---|---|
| PretextView | 0.2 | https://github.com/sanger-tol/PretextView |
| purge_dups | 1.2.5 | https://github.com/dfguan/purge_dups |
| samtools | 1.16.1, 1.17, and 1.18 | https://github.com/samtools/samtools |
| sanger-tol/ascc | - | https://github.com/sanger-tol/ascc |
| sanger-tol/genomenote | 1.1.1 | https://github.com/sanger-tol/genomenote |
| sanger-tol/readmapping | 1.2.1 | https://github.com/sanger-tol/readmapping |
| Seqtk | 1.3 | https://github.com/lh3/seqtk |
| Singularity | 3.9.0 | https://github.com/sylabs/singularity |
| TreeVal | 1.0.0 | https://github.com/sanger-tol/treeval |
| YaHS | 1.2a.2 | https://github.com/c-zhou/yahs |

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.

Further, 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)

**The Large Blue is listed as an Endangered Species in England and as a species of the Habitats Directive (Annex IV) in EU nations, which applied to the UK in 2022. The statutory licensing authority, Natural England, issued licence number 2022-61093-SCI-SC to authorise David Simcox, Jeremy Thomas, Sarah Meredith and Peter Holland to collect up to five specimens for genome sequencing from Daneway Banks in accordance with The Conservation of Habitats and Species Regulations 2017 (as amended) and Wildlife and Countryside Act 1981 (as amended); only two specimens were collected. The co-owners of Daneway Banks, Gloucestershire Wildlife Trust and the Royal Entomological Society, gave separate written permission to take specimens for genome sequencing from the site.** Each transfer of samples is further undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Darwin Tree of Life Partner, Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances other Darwin Tree of Life collaborators.

## Data availability

European Nucleotide Archive: Phengaris arion (large blue). Accession number PRJEB65386; https://identifiers.org/ena.embl/PRJEB65386 (Wellcome Sanger Institute, 2023). The genome sequence is released openly for reuse. The *Phengaris arion* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project. All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

### Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.12157525.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.12158331

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: https://doi.org/10.5281/zenodo.12162482.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: https://doi.org/10.5281/zenodo.12165051.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: https://doi.org/10.5281/zenodo.12160324.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.12205391.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783558.

## References


Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316.
PubMed Abstract | Publisher Full Text | Free Full Text

Akino T, Knapp JJ, Thomas JA, *et al.*: **Chemical mimicry and host specificity in the butterfly *Maculinea rebeli*, a social parasite of *Myrmica* ant colonies.** *Proc Biol Sci.* 1999; **266**(1427): 1419–1426.
Publisher Full Text | Free Full Text

Allio R, Schomaker-Bastos A, Romiguier J, *et al.*: **MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics.** *Mol Ecol Resour.* 2020; **20**(4): 892–905.
PubMed Abstract | Publisher Full Text | Free Full Text

Als TD, Vila R, Kandul NP, *et al.*: **The evolution of alternative parasitic life histories in large blue butterflies.** *Nature.* 2004; **432**(7015): 386–90.
PubMed Abstract | Publisher Full Text

Altschul SF, Gish W, Miller W, *et al.*: **Basic local alignment search tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text

Balletto E, Bonelli S, Settele J, *et al.*: ***Maculinea* Van Eecke, 1915 (Lepidoptera: Lycaenidae): proposed precedence over *Phengaris* Doherty, 1891.** *Bull Zool Nom.* 2010; **67**(2): 129–132.
Publisher Full Text

Barbero F, Thomas JA, Bonelli S, *et al.*: **Queen ants make distinctive sounds that are mimicked by a butterfly social parasite.** *Science.* 2009; **323**(5915): 782–5.
PubMed Abstract | Publisher Full Text

Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the universal protein knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531.
PubMed Abstract | Publisher Full Text | Free Full Text

Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor®3 for LI PacBio.** *protocols.io.* 2023.
Publisher Full Text

Beasley J, Uhl R, Forrest LL, *et al.*: **DNA barcoding SOPs for the Darwin Tree of Life project.** *protocols.io.* 2023; [Accessed 25 June 2024].
Publisher Full Text

Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368.
PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24.
PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374.
PubMed Abstract | Publisher Full Text | Free Full Text

Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175.
PubMed Abstract | Publisher Full Text | Free Full Text

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial arthropod fauna [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 123.
PubMed Abstract | Publisher Full Text | Free Full Text

da Veiga Leprevost F, Grüning BA, Alves Aflitos S, *et al.*: **BioContainers: an open-source and community-driven framework for software standardization.** *Bioinformatics.* 2017; **33**(16): 2580–2582.
PubMed Abstract | Publisher Full Text | Free Full Text

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008.
PubMed Abstract | Publisher Full Text | Free Full Text

Denton A, Oatley G, Cornwell C, *et al.*: **Sanger Tree of Life sample homogenisation: PowerMash.** *protocols.io.* 2023a.
Publisher Full Text

Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life wet laboratory protocol collection V.1.** *protocols.io.* 2023b.
Publisher Full Text

Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.
PubMed Abstract | Publisher Full Text

Diesh C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of synteny and structural variation.** *Genome Biol.* 2023; **24**(1): 74.
PubMed Abstract | Publisher Full Text | Free Full Text

Elmes GW, Thomas JA, Munguira ML, *et al.*: **Larvae of lycaenid butterflies that parasitize ant colonies provide exceptions to normal insect growth rules.** *Biol J Linn Soc.* 2001; **73**(3): 259–278.
Publisher Full Text

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
PubMed Abstract | Publisher Full Text

Fiedler K: **REVIEW: lycaenid-ant interactions of the *Maculinea* type: tracing their historical roots in a comparative framework.** *J Insect Conserv.* 1998; **2**: 3–14.
Publisher Full Text

Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
PubMed Abstract | Publisher Full Text | Free Full Text

Fric Z, Wahlberg N, Pech P, *et al.*: **Phylogeny and classification of the *Phengaris-Maculinea* clade (Lepidoptera: Lycaenidae): total evidence and phylogenetic species concepts.** *Syst Entomol.* 2007; **32**(3): 558–567.
Publisher Full Text

Frohawk FW: **Natural history of British Butterflies**. London: Hutchinson, 1914.
Reference Source

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
PubMed Abstract | Publisher Full Text | Free Full Text

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898.
PubMed Abstract | Publisher Full Text | Free Full Text

Harry E: **PretextView (Paired Read Texture Viewer): a desktop application for viewing pretext contact maps.** 2022; [Accessed 19 October 2022].
Reference Source

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

Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life sample preparation: triage and dissection.** *protocols.io.* 2023.
Publisher 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

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
PubMed Abstract | Publisher Full Text | Free Full Text

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text | Free Full Text

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | Publisher Full Text | Free Full Text

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2.
Reference Source

Muffato M, Butt Z, Challis R, *et al.*: **Sanger-tol/blobtoolkit: v0.3.0 – poliwag**. 2024.
Publisher Full Text

Oatley G, Denton A, Howard C: **Sanger Tree of Life HMW DNA extraction: automated MagAttract v.2.** *protocols.io.* 2023.
Publisher Full Text

Pierce NE, Braby MF, Heath A, *et al.*: **The ecology and evolution of ant association in the Lycaenidae (Lepidoptera).** *Annu Rev Entomol.* 2002; **47**: 733–71.
PubMed Abstract | Publisher Full Text

Pointon DL, Eagles W, Sims Y, *et al.*: **Sanger-tol/treeval v1.0.0 – Ancient Atlantis**. 2023.
Publisher Full Text

Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–842.
PubMed Abstract | Publisher Full Text | Free Full Text

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

Sayers EW, Cavanaugh M, Clark K, *et al.*: **GenBank 2024 update.** *Nucleic Acids Res.* 2024; **52**(D1): D134–D137.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

Settele J, Kuehn R, Thomas JA: **Studies in the ecology & conservation of Butterflies in Europe 2. Species ecology along a European gradient: *Maculinea* butterflies as a model**. Sofia: Pensoft, 2005; **2**.
**Reference Source**

Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *protocols.io.* 2023.
**Publisher Full Text**

Surana P, Muffato M, Qi G: **Sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo.* 2023a.
**Publisher Full Text**

Surana P, Muffato M, Sadasivan Baby C: **Sanger-tol/genomenote (v1.0.dev).** *Zenodo.* 2023b.
**Publisher Full Text**

Tartally A, Thomas JA, Anton C, *et al.*: **Patterns of host use by brood parasitic *Maculinea* butterflies across Europe.** *P Philos Trans R Soc Lond B Biol Sci.* 2019; **374**(1769): 20180202.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

Thomas JA: **Larval niche selection and evening exposure enhance adoption of a predacious social parasite, *Maculinea arion* (large blue butterfly), by *Myrmica* ants.** *Oecologia.* 2002; **132**(4): 531–537.
**PubMed Abstract** | **Publisher Full Text**

Thomas JA, Elmes GW, Sielezniew M, *et al.*: **Mimetic host shifts in an endangered social parasite of ants.** *Proc Biol Sci.* 2013; **280**(1751): 20122336.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

Thomas JA, Elmes GW, Wardlaw JC, *et al.*: **Host specificity among *Maculinea* butterflies in *Myrmica* ant nests.** *Oecologia.* 1989; **79**(4): 452–457.
**PubMed Abstract** | **Publisher Full Text**

Thomas JA, Schönrogge K, Elmes GW: **Specializations and host associations of social parasites of ants**. In: *Insect Evolutionary Ecology: Proceedings of the Royal Entomological Society's 22nd Symposium, Reading, UK 2003*. UK: CABI Publishing, 2005; 479–518.
**Publisher Full Text**

Thomas JA, Settele J: **Evolutionary biology: butterfly mimics of ants.** *Nature.* 2004; **432**(7015): 283–284.
**PubMed Abstract** | **Publisher Full Text**

Thomas JA, Simcox DJ, Clarke RT: **Successful conservation of a threatened *Maculinea* butterfly.** *Science.* 2009; **325**(5936): 80–83.
**PubMed Abstract** | **Publisher Full Text**

Thomas JA, Simcox DJ, Meredith SA: **Re-establishing the large blue butterfly in Britain.** *Br Wildl.* 2019; **31**: 7–14.
**Reference Source**

Thomas JA, Wardlaw JC: **The capacity of a *Myrmica* ant nest to support a predacious species of *Maculinea* butterfly.** *Oecologia.* 1992; **91**(1): 101–109.
**PubMed Abstract** | **Publisher Full Text**

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer review: awaiting peer review].** *Wellcome Open Res.* 2024; **9**: 339.
**Publisher Full Text**

Ugelvig LV, Vila R, Pierce NE, *et al.*: **A phylogenetic revision of the *Glaucopsyche* section (Lepidoptera: Lycaenidae), with special focus on the *Phengaris-Maculinea* clade.** *Mol Phylogenet Evol.* 2011; **61**(1): 237–243.
**PubMed Abstract** | **Publisher Full Text**

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

Van Swaay C, Cuttelod A, Collins S, *et al.*: **European red list of Butterflies**. IUCN (International Union for Conservation of Nature) and Butterfly Conservation Europe in collaboration with the European Union, 2010.
**Reference Source**

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems**. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2019; 314–324.
**Publisher Full Text**

Wellcome Sanger Institute: **The genome sequence of the Large Blue butterfly, *Phengaris (Maculinea) arion* (Linnaeus, 1758)**. European Nucleotide Archive. [dataset], accession number PRJEB65386, 2023.

Zhou C, McCarthy SA, Durbin R: **YaHS: yet another Hi-C scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

# Open Peer Review

## Current Peer Review Status: ? ? ? ?

**Version 1**

Reviewer Report 25 October 2024

https://doi.org/10.21956/wellcomeopenres.25309.r103759



?

**Paula Escuer Pifarré**
University of Neuchâtel, Neuchâtel, Switzerland

In this paper the authors present the genome assembly of the butterfly *Phengaris arion*. The genome presents 544.50 Mb of length and 22 scaffolds corresponding to the autosomes plus the Z sex chromosome and the mitochondrial genome. This resource represents an important material to shed light about lepidoptera genomics. The procedure is well detailed; however, I would like to mention some improvements for this genome note:

- First the authors say that they used 1 female individual for genome sequencing (ilPheArio1), but then in the methods they mention that another individual was used for Hi-C libraries (ilPheArio2). It is a bit confusing because when you start reading it seems only one individual has been used and they should clarify that earlier, even if the genetic information is mainly from the first individual one. On the other hand, they also only explain how they weighed and dissected for DNA extraction the first individual but then they just say that the head tissue of iPheArio2 was used for Hi-C (I assume they did the same for that individual?).

- They affirm *P. arion* is Z0 but they don't include any citation about this. They mention "The Z chromosome was identified based on read coverage statistics." I guess they mapped the reads against the genome, but it would be better if they were more concrete about the procedure. Also, I consider it is better to redact it like "the sex determination system is Z0 based on the coverage statistics" because just with the genome assembly of one individual is not enough to confirm.

- Looking at the Hi-C plot, the second larger scaffold looks a bit weird, there is a small region with higher contacts, and they don't explain why that might be. It would need more manual curation because it could be an artifact.

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

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

Partly

**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:*** Genomics, Epigenomics, Speciation

**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 22 October 2024

https://doi.org/10.21956/wellcomeopenres.25309.r101276



?

**Arjen Van 't Hof**
Biology Centre of the Czech Academy of Sciences, Budějovice, Czech Republic

The Darwin Tree of Life Consortium has added the genome sequence of the Large Blue butterfly (Phengaris arion) to their ever-expanding list of assemblies.

Since I have no hands-on experience with most of the bioinformatics tools used, I have to rely on the assembly statistics, which are very impressive as we are used to with dtol genomes.

The Large blue butterfly is a very interesting species because of its conservation status and its exploitation of ant colonies. The Background section explains very nicely how fascinating this species is.

When it comes to the genome assembly, I miss an interpretation in a wider biological context. In 1971, M.J.D. White wrote in his Preface to the third edition of Animal Cytology and Evolution: “many of the younger workers, rightly anxious to be fully informed on all the latest discoveries, seem to have lost historical perspective.”

More than 50 years later, this seems to be still the case. The authors claim that the species has a ZO sex chromosome composition and this is mentioned in a six-word sentence. Those interested in chromosome evolution would argue that this deserves more attention, and since it is so relevant, evidence that the W chromosome is truly absent would also be welcome.

The female Luperina nickerlii assembly also lacks a W-scaffold, but for this species, the wording is chosen more carefully “No W chromosome could be identified, and the species appears to be ZO.”

I would prefer such a reserved statement about the sex chromosome composition in the Large Blue as well until it is thoroughly investigated.

**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:*** Lepidoptera genetics

**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 21 October 2024

https://doi.org/10.21956/wellcomeopenres.25309.r102525



?

**Li-Wei Wu**
Tunghai University, Taichung, Taichung City, Taiwan

This work presents a high-quality genome of the Large Blue, *Phengaris arion*. This data is highly valuable and is expected to provide significant scientific support for future conservation studies. Meredith *et al.* provided clear background information on this species, which allowed me to easily understand the context of this organism. However, I have the following suggestions regarding the presented data:

1. Page 4, Figure 1: The samples should include images of both dorsal and ventral wing patterns to provide clear morphological information for species identification.

2. Page 4, last line of the left paragraph: "The specimen is a ZO female." This contradicts our

current understanding of Lepidoptera being a ZW system. Is there a citation that explains *Phengaris arion* as a ZO system? Or was the presence of the W chromosome not confirmed in this data analysis?

3. In the "Genome sequence report" section on page 3, it is mentioned that genome analysis was conducted using a female,
but on page 5 it states that two samples were used. I suggest providing a consistent explanation in the text regarding the use of different samples.

4. Page 6, DNA concentration was measured using both a Nanodrop spectrophotometer and a Qubit fluorometer. I would like to know which concentration was used for the subsequent sample analysis.

5. Page 10, "Table 4 contains a list of relevant software tool versions and sources" is an odd paragraph and should be revised.

**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:*** butterfly conservation, genomic analysis

**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 09 October 2024

https://doi.org/10.21956/wellcomeopenres.25309.r103710



?

**Jesper Boman**
Uppsala University, Uppsala, Sweden

Meredith *et al* presents a high-quality genome-assembly of the large blue (*Phengaris arion*). The DToL is a commendable project that provides freely available genome assemblies to the research community. I would like to extend my appreciation to the authors on a well-written Background-section. However, I think the rest of the report needs some work.

A few points to consider:

1) You write that the specimen is Z0. Are there any support for this karyotype from the literature or do you claim it as a discovery in this study?

2) You write: *"The order and orientation of contigs on Chromosome 2 from 1 to approximately 3.63 Mb is uncertain."* Is it really wise then to actually include that sequence in the same scaffold? This would surely be missed and create confusion in comparative analyses. An alternative approach would be to call these smaller scaffolds:  >2_unplaced_1, >2_unplaced_2 etc.

3) In the Results section you write as if only a single specimen was used while in the Sample acquisition section you use "specimens". This is confusing for the reader. It is not clear for a reader like me if the **ToLID** is an individual identification or not. It could also be read as if it is an identification for a sample, e.g. the head and thorax of the same individual.

4) Please clearly indicate the Z chromosome in Figure 5.

5)  It would be better to use user-friendly scaffold names in Figure 3. E.g. "Chromosome_2" or "Scaffold_1".

I also want to reiterate point 1), 2) and 4) from a previous peer review:

Boman J. Peer Review Report For: The genome sequence of the Bright-line Brown-eye moth, Lacanobia oleracea Linnaeus, 1758 [version 1; peer review: 1 approved with reservations]. Wellcome Open Res 2024, 9:515 (https://doi.org/10.21956/wellcomeopenres.25228.r98875)

They are valid for this report as well.

Best regards,
Jesper Boman

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

**Are the protocols appropriate and is the work technically sound?**
Partly

**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?**

Partly

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Speciation, Genomics, Population Genetics, Transposons, Regulatory Evolution, Recombination

**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.**