The genome sequence of the Crepuscular Burnet, *Zygaena carniolica* (Scopoli, 1763)

**Abstract**

We present a genome assembly from a female specimen of *Zygaena carniolica* (the Crepuscular Burnet; Arthropoda; Insecta; Lepidoptera; Zygaenidae). The assembly includes two haplotypes with total lengths of 353.39 megabases and 323.68 megabases. Most of haplotype 1 (99.94%) is scaffolded into 31 chromosomal pseudomolecules, including the W and Z sex chromosomes. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 15.56 kilobases.

## Authors

Yannick Chittaro, Marianne Espeland, [genome curator], Hub leader, [Hub 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, Psyche Project consortium

### Author information

The following groups contributed to this project. Full author lists can be found on the respective Zenodo pages:

* [The Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team](https://doi.org/10.5281/zenodo.12162482)
* [The Wellcome Sanger Institute Scientific Operations: Sequencing Operations](https://doi.org/10.5281/zenodo.12165051)
* [The Wellcome Sanger Institute Tree of Life Core Informatics team](https://doi.org/10.5281/zenodo.12160324)
* [The Tree of Life Core Informatics collective](https://doi.org/10.5281/zenodo.12205391)
* Psyche Project consortium

## Keywords

*Zygaena carniolica*, Crepuscular Burnet, genome sequence, chromosomal, Lepidoptera

## Main body

### 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; Apoditrysia; Zygaenoidea; Zygaenidae; Zygaeninae; *Zygaena*; *Zygaena carniolica* (Scopoli, 1763) (NCBI:txid287238)

## Background

The crepuscular burnet moth (*Zygaena carniolica*) is a widespread species with extremely variable phenotype and ecology. It is distributed from eastern Spain through central Europe to southern Siberia and western Mongolia, at elevations from sea level to almost 4,000 m. As with all species in the genus *Zygaena*, *Z. carniolica* is highly aposematic due to the capability of synthesising and/or sequestering cyanogenic glucosides from its host plants. The forewings are black with six red spots of variable size, with more or less cream colouration around the spots, and hindwings are red with a back border. In extreme cases, the cream colouration completely takes over and the black colouration is hardly visible.

It occurs in a variety of habitats, mainly in calcareous regions. The species feeds on a variety of host plants mainly of the genera *Anthyllis*, *Dorycnium*, *Hippocrepis*, *Onobrychis*, *Lotus*, *Hedysarum* and *Astragalus*, all Fabaceae ([Hofmann & Tremewan, 2020](#bookmark=id.111kx3o)). It is often abundant where it occurs, and can benefit from planted *Lotus* and *Onobrychis* crops ([Rennwald *et al.*, 2012](#bookmark=id.1664s55)). Individuals are rather lethargic and many can be found roosting on a single flower head. Populations are univoltine or biovoltine depending on locality, and flight time is variable ([Hofmann & Tremewan, 2020](#bookmark=id.111kx3o)).

*Z. carniolica* is listed as Vulnerable in Switzerland ([Wermeille *et al.*, 2014](#bookmark=id.43ky6rz)) and threatened in some states in eastern Germany, where it reaches the northernmost boundary of its distribution Wachlin *et al.* ([1997](#bookmark=id.1jlao46)). It is not listed in the [IUCN Red list Europe](https://www.iucnredlist.org/regions/european-red-list). An allozyme-based population genetic study of populations in Germany and adjoining areas in Luxembourg and France found the genetic diversity of *Z. carniolica* rather low and homogeneously distributed across all populations ([Habel *et al.*, 2012](#bookmark=id.46r0co2)). Different populations are known to have between 30 and 31 chromosomes Lukhtanov & Kuznetsova ([1989](#bookmark=id.3ygebqi)).

We present a chromosome-level, haplotype-resolved genome sequence of the Crepuscular Burnet, *Zygaena carniolica*, sequenced as part of Project Psyche. The sequence data was derived from a female specimen (Figure [1](#bookmark=id.tyjcwt)), collected from Conthey, Valais.



*Figure 1: Voucher photograph of the Zygaena carniolica specimen (SAN28000139; ilZygCarn1) used for genome sequencing.*

## Methods

### Sample acquisition

An adult female *Zygaena carniolica* (specimen ID SAN28000139, ToLID ilZygCarn1) was collected from Conthey, Valais, Switzerland (latitude 46.2872, longitude 7.3116) on 2023-08-02. The specimen was collected and identified by Yannick Chittaro (Info Fauna, Avenue De Bellevaux 51, 2000 Neuchâtel, Switzerland).

### Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on [protocols.io](https://dx.doi.org/10.17504/protocols.io.8epv5xxy6g1b/v1). The ilZygCarn1 sample was prepared for DNA extraction by [weighing and dissecting](https://dx.doi.org/10.17504/protocols.io.x54v9prmqg3e/v1) it on dry ice. Tissue from the thorax was homogenised by [PowerMashing](https://dx.doi.org/10.17504/protocols.io.5qpvo3r19v4o/v1) using a PowerMasher II tissue disruptor.

HMW DNA was extracted in the WSI Scientific Operations core using the [Automated MagAttract v2 protocol](https://dx.doi.org/10.17504/protocols.io.kxygx3y4dg8j/v1). 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 following the [manual SPRI (solid-phase reversible immobilisation)](https://dx.doi.org/10.17504/protocols.io.kxygx3y1dg8j/v1) protocol, which uses AMPure PB beads to eliminate shorter fragments and concentrate the DNA. 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.

### Hi-C sample preparation

Tissue from the head of the ilZygCarn1 sample was processed at the WSI Scientific Operations core, using the Arima-HiC v2 kit. In brief, 20–50 mg of frozen tissue (stored at –80 °C) was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde. After crosslinking, the tissue was homogenised using the Diagnocine Power Masher-II and BioMasher-II tubes and pestles. Following the Arima-HiC v2 kit manufacturer's instructions, crosslinked DNA was digested using a restriction enzyme master mix. The 5’-overhangs were filled in and labelled with biotinylated nucleotides and proximally ligated. An overnight incubation was carried out for enzymes to digest remaining proteins and for crosslinks to reverse. A clean up was performed with SPRIselect beads prior to library preparation. Additionally, the biotinylation percentage was estimated using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit and Arima-HiC v2 QC beads.

### Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Pacific Biosciences HiFi circular consensus DNA sequencing libraries were prepared using the PacBio Express Template Preparation Kit v2.0 (Pacific Biosciences, California, USA) as per the manufacturer’s instructions. The kit includes the reagents required for removal of single-strand overhangs, DNA damage repair, end repair/A-tailing, adapter ligation, and nuclease treatment. Library preparation also included a library purification step using AMPure PB beads (Pacific Biosciences, California, USA) and size selection step to remove templates shorter than 3 kb using AMPure PB modified SPRI. DNA concentration was quantified using the Qubit Fluorometer v2.0 and Qubit HS Assay Kit and the final library fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument and gDNA 165 kb gDNA and 55kb BAC analysis kit. Samples were sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was between 40–135 pM. The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, as well as perform primary and secondary analysis of the data upon completion.

Pacific Biosciences SMRTbell libraries were constructed using the Revio HiFi prep kit, according to the manufacturers’ instructions. DNA sequencing was performed by the Scientific Operations core at the WSI on a Pacific Biosciences Revio instrument.

For Hi-C library preparation, DNA was fragmented to a size of 400 to 600 bp using a Covaris E220 sonicator. The DNA was then enriched, barcoded, and amplified using the NEBNext Ultra II DNA Library Prep Kit following manufacturers’ instructions. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on an Illumina NovaSeq X instrument.

### Genome assembly, curation and evaluation

#### Assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts (*k* = 31) was generated from the filtered reads using [FastK](https://github.com/thegenemyers/FASTK). GenomeScope2 ([Ranallo-Benavidez *et al.*, 2020](#bookmark=id.4bvk7pj)) was used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content. The HiFi reads were assembled using Hifiasm ([Cheng *et al.*, 2021](#bookmark=id.3fwokq0), [2022](#bookmark=id.1v1yuxt)) in Hi-C phasing mode, resulting in a pair of haplotype-resolved assemblies.

Hi-C reads were mapped to the primary contigs using bwa-mem2 ([Vasimuddin *et al.*, 2019](#bookmark=id.34g0dwd)). The contigs were further scaffolded using the provided Hi-C data ([Rao *et al.*, 2014](#bookmark=id.2r0uhxc)) in YaHS ([Zhou *et al.*, 2023](#bookmark=id.2iq8gzs)) using the --break option for handling potential mis-assemblies. The scaffolded assemblies were evaluated using Gfastats ([Formenti *et al.*, 2022](#bookmark=id.37m2jsg)), BUSCO ([Manni *et al.*, 2021](#bookmark=id.2dlolyb)) and MerquryFK ([Rhie *et al.*, 2020](#bookmark=id.25b2l0r)).

The mitochondrial genome was assembled using MitoHiFi ([Uliano-Silva *et al.*, 2023](#bookmark=id.kgcv8k)), which runs MitoFinder ([Allio *et al.*, 2020](#bookmark=id.3o7alnk)), 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](https://pipelines.tol.sanger.ac.uk/ascc) (article in preparation). Flat files and maps used in curation were generated in TreeVal ([Pointon *et al.*, 2023](#bookmark=id.1rvwp1q)). Manual curation was primarily conducted using PretextView ([Harry, 2022](#bookmark=id.2lwamvv)), with additional insights provided by JBrowse2 ([Diesh *et al.*, 2023](#bookmark=id.19c6y18)) and HiGlass ([Kerpedjiev *et al.*, 2018](#bookmark=id.206ipza)). Scaffolds were visually inspected and corrected as described by Howe *et al.* ([2021](#bookmark=id.3l18frh)). Any identified contamination, missed joins, and mis-joins were corrected, and duplicate sequences were tagged and removed. The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation>.

#### Evaluation of the final assembly

The MerquryFK tool ([Rhie *et al.*, 2020](#bookmark=id.25b2l0r)), run within a Singularity container ([Kurtzer *et al.*, 2017](#bookmark=id.4k668n3)), was used to evaluate *k*-mer completeness and assembly quality for the curated haplotypes, using the *k*-mer databases (*k* = 31) computed before genome assembly. The analysis outputs included *k*-mer plots, assembly QV scores and completeness statistics.

The final assembly was also analysed in the BlobToolKit environment. The [sanger-tol/blobtoolkit](https://zenodo.org/records/13758882) pipeline is a Nextflow ([Di Tommaso *et al.*, 2017](#bookmark=id.2u6wntf)) port of the previous Snakemake Blobtoolkit pipeline ([Challis *et al.*, 2020](#bookmark=id.vx1227)). It aligns the PacBio reads in SAMtools and minimap2 ([Li, 2018](#bookmark=id.1egqt2p)) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoaT database ([Challis *et al.*, 2023](#bookmark=id.2grqrue)) to identify all matching BUSCO lineages to run BUSCO ([Manni *et al.*, 2021](#bookmark=id.2dlolyb)). For the three domain-level BUSCO lineages, the pipeline uses DIAMOND ([Buchfink *et al.*, 2021](#bookmark=id.1hmsyys)) blastp to align the BUSCO genes to the UniProt Reference Proteomes database ([Bateman *et al.*, 2023](#bookmark=id.32hioqz)). The genome is also split into chunks according to the density of the BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Genome sequences without a hit are chunked with seqtk and aligned to the NT database with blastn ([Altschul *et al.*, 1990](#bookmark=id.ihv636)). The blobtools suite combines these outputs into a blobdir for visualisation. The blobtoolkit pipeline was developed using nf-core tooling and MultiQC ([Ewels *et al.*, 2016](#bookmark=id.3tbugp1); [Ewels *et al.*, 2020](#bookmark=id.28h4qwu)), relying on the [Conda](https://www.anaconda.com/) package manager, the Bioconda initiative ([Grüning *et al.*, 2018](#bookmark=id.1mrcu09)), the Biocontainers infrastructure ([da Veiga Leprevost *et al.*, 2017](#bookmark=id.4f1mdlm)), as well as the Docker ([Merkel, 2014](#bookmark=id.sqyw64)) and Singularity containerisation solutions.

## Genome sequence report

### Sequencing data

The genome of an adult female *Zygaena carniolica* specimen was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of 16.26 Gb (gigabases) from 1.55 million reads. Chromosome conformation Hi-C sequencing produced 103.08 Gb from 682.62 million reads. Table [1](#bookmark=id.4d34og8) summarises the raw sequence data obtained.

Based on GenomeScope *k*-mer profiling of the PacBio HiFi data (using *k* = 31), the haploid genome size was estimated to be 336.97 Mb, and the sequencing coverage of the genome was approximately 47-fold. The estimated heterozygosity was 1.27% and the repeat content was 26.65%.

*Table 1: Raw sequence information for the genome sequence of Zygaena carniolica*

|  |  |  |
| --- | --- | --- |
|  | PacBio | **Hi-C** |
| **ToLID** | ilZygCarn1 | ilZygCarn1 |
| **Biospecimen** | SAMEA11511002 | SAMEA11511002 |
| **Tissue** | thorax | head |
| **Sequencing platform and model** | PacBio – HiFi Revio | Hi-C – Arima v2 Illumina NovaSeq X |
| **BioSample** | SAMEA115110086 | SAMEA115110020 |
| **Run accessions** | ERR1348574 | ERR13494006 |
| **Read count total** | 1.55 million reads | 1.55 million reads |
| **Base count total** | 16.26 Gb | 103.08 Gb |

### Assembly statistics

Haplotype 1 and haplotype 2 were assembled using Hi-C phasing, and the two haplotypes were combined for curation. Manual assembly curation corrected 81 missing joins or mis-joins. This reduced the assembly length by 0.53% and the scaffold number by 30.36%, and increased the scaffold N50 by 1.28%. Figure [2](#bookmark=id.35nkun2) shows a Merqury assembly spectrum plot of the curated assembly.

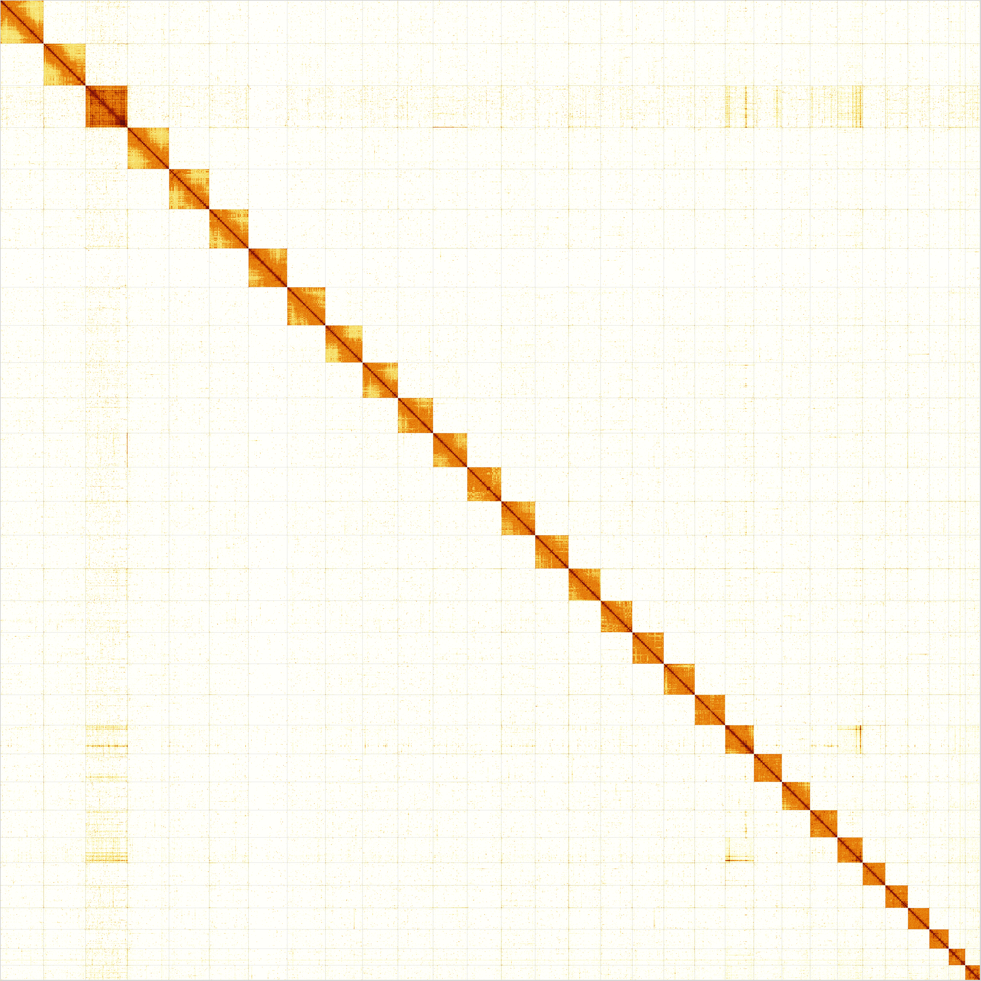
The final haplotype 1 assembly has a total length of Mb in sequence scaffolds, with 99 gaps, and a scaffold N50 of Mb (Table [2](#bookmark=id.17dp8vu)).

Most of the haplotype 1 assembly sequence (99.93%) was assigned to 31 chromosomal-level scaffolds, representing 29 autosomes and the Z and W sex chromosomes. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure [2](#bookmark=id.3rdcrjn); Table [3](#bookmark=id.lnxbz9)). During curation, the Z and W sex chromosomes were identified by analysing the coverage of Hi-C data and PacBio reads.

The mitochondrial genome was also assembled and is included both as a contig within the multifasta file of the genome submission and as a standalone record in GenBank.

*Table 2: Genome assembly data for the Zygaena carniolica assembly*

| **Genome assembly** | **Haplotype 1** | **Haplotype 2** |
| --- | --- | --- |
| **Assembly name** | ilZygCarn1.hap1.1 | ilZygCarn1.hap2.1 |
| **Assembly accession** | GCA\_964261665.1 | GCA\_964261595.1 |
| **Assembly level** | chromosome | scaffold |
| **Span (Mb)** | 353.39 | 323.68 |
| **Number of contigs** | 137 | 121 |
| **Contig N50** | 5.74 Mb | 5.55 Mb |
| **Number of scaffolds** | 38 | 37 |
| **Scaffold N50** | 12.29 Mb | 12.24 Mb |
| **Sex chromosomes** | Z and W | - |
| **Organelles** | Mitochondrial genome: 15.56 kb | - |

**

***Figure 2: Genome assembly of Zygaena carniolica ilZygCarn1.hap1.1:*** *Hi-C contact map of the ilZygCarn1.hap1.1 assembly, produced in PretextView. Chromosomes are shown in order of size from left to right and top to bottom.*

*Table 3: Chromosomal pseudomolecules in the haplotype 1 genome assembly of Zygaena carniolica ilZygCarn1*

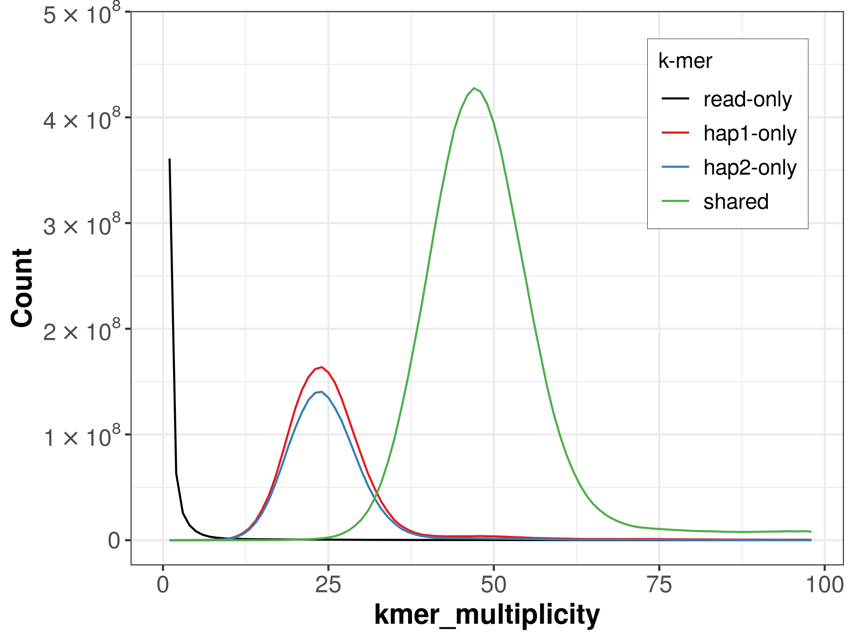
| **INSDC accession** | **Molecule** | **Length (Mb)** | **GC%** |
| --- | --- | --- | --- |
| OZ179962.1 | 1 | 15.56 | 37 |
| OZ179963.1 | 2 | 15.15 | 37 |
| OZ179965.1 | 3 | 15.0 | 36.5 |
| OZ179966.1 | 4 | 14.47 | 36.5 |
| OZ179967.1 | 5 | 14.19 | 36.5 |
| OZ179969.1 | 6 | 13.79 | 36 |
| OZ179970.1 | 7 | 13.38 | 36.5 |
| OZ179971.1 | 8 | 12.72 | 36.5 |
| OZ179972.1 | 9 | 12.69 | 36 |
| OZ179973.1 | 10 | 12.34 | 37 |
| OZ179974.1 | 11 | 12.29 | 36 |
| OZ179975.1 | 12 | 12.16 | 36.5 |
| OZ179976.1 | 13 | 12.02 | 36.5 |
| OZ179977.1 | 14 | 11.64 | 37 |
| OZ179978.1 | 15 | 11.39 | 37.5 |
| OZ179979.1 | 16 | 11.35 | 36 |
| OZ179980.1 | 17 | 11.15 | 37 |
| OZ179981.1 | 18 | 10.97 | 36.5 |
| OZ179982.1 | 19 | 10.38 | 39.5 |
| OZ179983.1 | 20 | 10.13 | 36.5 |
| OZ179984.1 | 21 | 10.13 | 37.5 |
| OZ179985.1 | 22 | 9.82 | 37 |
| OZ179986.1 | 23 | 9.14 | 38 |
| OZ179987.1 | 24 | 8.18 | 37 |
| OZ179988.1 | 25 | 8.12 | 37.5 |
| OZ179989.1 | 26 | 7.72 | 36.5 |
| OZ179990.1 | 27 | 6.99 | 37 |
| OZ179991.1 | 28 | 5.95 | 38 |
| OZ179992.1 | 29 | 5.35 | 38.5 |
| OZ179964.1 | W | 15.11 | 39 |
| OZ179968.1 | Z | 13.93 | 36 |
| OZ179993.1 | MT | 0.02 | 18.5 |

### Assembly quality metrics

The estimated Quality Value (QV) and *k*-mer completeness metrics, along with BUSCO completeness scores, were calculated for each haplotype and the combined assembly. The QV indicates the base-level accuracy of the assembly, while *k*-mer completeness indicates the proportion of expected *k*-mers identified in the assembly. BUSCO scores provide a measure of completeness based on benchmarking universal single-copy orthologues.

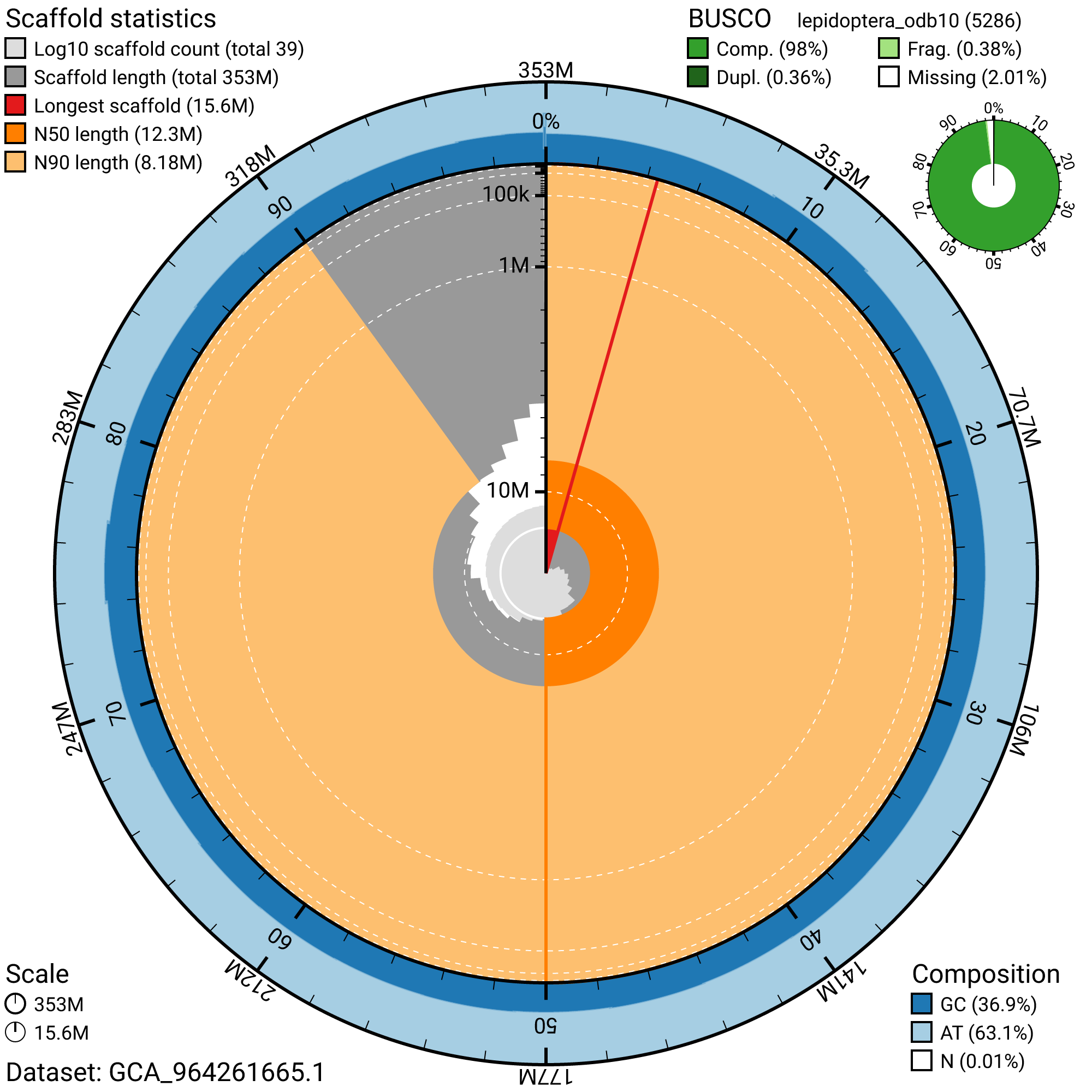
For haplotype 1, the QV is 64.3; for haplotype 2, 64.5. Combining both haplotypes yields a QV of 64.4 (less than 1 error per million bases). The *k*-mer completeness is 77.49% for haplotype 1 and 73.39% for haplotype 2; for the combined assemblies the *k*-mer completeness is 99.65%. Figure 3 shows the frequency distribution of *k-*mers that are unique to each haplotype (red and blue curves) and shared between the haplotypes (green curve).

BUSCO 5.5.0 analysis using the lepidoptera\_odb10 reference set (*n* = 5,286) identified 98.0% of the expected gene set (single 97.6%, duplicated 0.4%) for haplotype 1, and 93.5% (single 93.3%, duplicated 0.3%) for haplotype 2.



***Figure 3: Evaluation of k-mer completeness using MerquryFK****.* *This plot illustrates the recovery of k‐mers from the original read data in the final assemblies. The horizontal axis shows k‐mer multiplicity, and the vertical axis shows the number of k‐mers. The grey curve represents k‐mers that appear in the reads but are not assembled. The green curve corresponds to k‐mers shared by both assemblies (the homozygous peak), and the red and blue curves show k‐mers found only in haplotype 1 or haplotype 2, respectively (the heterozygous peaks).*

The snail plot in Figure [3](#bookmark=id.44sinio) provides a scaffold-level overview of the assembly. Scaffolds are arranged by length around the circumference, with arcs highlighting the largest scaffold and the N50 and N90 lengths. The plot also displays GC, AT, and N content across scaffold bins, as well as a cumulative count of scaffolds in the centre. The blob plot in Figure [4](#bookmark=id.2jxsxqh) visualises GC content on the horizontal axis and coverage on the vertical axis, with each point representing a scaffold sized by length.



***Figure 3: Genome assembly of Zygaena carniolica, ilZygCarn1.hap1.1: metrics.*** *The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the set is presented at the top right. An interactive version of this figure can be accessed on the* [*BlobToolKit viewer*](https://blobtoolkit.genomehubs.org/view/GCA_964261665.1/snail)*.*



***Figure 4: Genome assembly of Zygaena carniolica, ilZygCarn1.hap1.1****: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the* [*BlobToolKit viewer*](https://blobtoolkit.genomehubs.org/view/GCA_964261665.1/blob)*.*

### 

Table [4](#bookmark=id.z337ya) shows the assembly metric benchmarks adapted from Rhie *et al.* ([2021](#bookmark=id.3q5sasy)) and the Earth BioGenome Project Report on Assembly Standards [September 2024](https://www.earthbiogenome.org/report-on-assembly-standards) for the chromosome-level assembly haplotype 1.

In Table 4, the notation **6.C.Q64** indicates a contig N50 of around 6 Mb, chromosome-level assembly (“C”), and a QV of approximately 64. These values exceed the minimum reference standard of 6.C.Q40 recommended by the [Earth BioGenome Project](https://www.earthbiogenome.org/report-on-assembly-standards).

*Table 4: EBP summary metrics for the Zygaena carniolica assembly (haplotype 1)*

| **Measure** (Benchmark) | **Value** |
| --- | --- |
| EBP summary (haplotype 1) | 6.C.Q64 |
| Contig N50 length (≥ 1 Mb) | 5.74 Mb |
| Scaffold N50 length (= chromosome N50) | 12.29 Mb |
| Consensus quality (QV) (≥ 40) | Haplotype 1: 64.3; haplotype 2: 64.5; combined: 64.4 |
| *k*-mer completeness (≥ 95%) | Haplotype 1: 77.49%; haplotype 2: 73.39%; combined: 99.65% |
| BUSCO \*(S > 90%; D < 5%) | Haplotype 1: C:98.0%[S:97.6%,D:0.4%],F:0.4%,M:1.6%,n:5286; Haplotype 2: C:93.5%[S:93.3%,D:0.3%],F:0.4%,M:6.1%,n:5286 |
| Percentage of assembly assigned to chromosomes (≥ 90%) | 99.93% |

\* BUSCO scores based on the lepidoptera\_odb10 BUSCO set using version 5.5.0. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison.

*Table 5: 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 | 4.3.9 | <https://github.com/blobtoolkit/blobtoolkit> |
| BUSCO | 5.5.0 | <https://gitlab.com/ezlab/busco> |
| bwa-mem2 | 2.2.1 | <https://github.com/bwa-mem2/bwa-mem2> |
| 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 | 0.19.8-r603 | <https://github.com/chhylp123/hifiasm> |
| Merqury.FK | d00d98157618f4e8d1a9190026b19b471055b22e | <https://github.com/thegenemyers/MERQURY.FK> |
| Minimap2 | 2.24-r1122 | <https://github.com/lh3/minimap2> |
| MitoHiFi | 3 | <https://github.com/marcelauliano/MitoHiFi> |
| MultiQC | 1.14, 1.17 and 1.18 | <https://github.com/MultiQC/MultiQC> |
| Nextflow | 23.04.0-5857 | <https://github.com/nextflow-io/nextflow> |
| PretextView | 0.2 | <https://github.com/sanger-tol/PretextView> |
| samtools | 1.19.2 | <https://github.com/samtools/samtools> |
| sanger-tol/blobtoolkit | 0.5.1 | <https://github.com/sanger-tol/blobtoolkit> |
| Seqtk | 1.3 | <https://github.com/lh3/seqtk> |
| Singularity | 3.9.0 | <https://github.com/sylabs/singularity> |
| TreeVal | 1.1.0 | <https://github.com/sanger-tol/treeval> |
| YaHS | 1.2a.2 | <https://github.com/c-zhou/yahs> |

### Data availability

European Nucleotide Archive: Zygaena carniolica (crepuscular burnet). Accession number [PRJEB78806](https://identifiers.org/ena.embl/PRJEB78806). The genome sequence is released openly for reuse. The *Zygaena carniolica* genome sequencing initiative is part of Project Psyche, BioProject [PRJEB71705](https://www.ebi.ac.uk/ena/browser/view/PRJEB71705). 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 Tables [1](#bookmark=id.xvir7l) and [2](#bookmark=id.4d34og8).

### Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Tree of Life collaborator. The Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so, we align with best practice wherever possible. The overarching areas of consideration are:

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

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.

## References

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F: [MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics](https://doi.org/10.1111/1755-0998.13160). *Molecular Ecology Resources*. 2020; **20**(4) 892–905.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: [Basic local alignment search tool](https://doi.org/10.1016/S0022-2836(05)80360-2). *Journal of Molecular Biology*. 1990; **215**(3) 403–10.

Bateman A, Martin M-J, Orchard S, Magrane M, Ahmad S, et al.: [UniProt: The universal protein knowledgebase in 2023](https://doi.org/10.1093/nar/gkac1052). *Nucleic Acids Research*. 2023; **51**(D1) D523–31.

Buchfink B, Reuter K, Drost H-G: [Sensitive protein alignments at tree-of-life scale using DIAMOND](https://doi.org/10.1038/s41592-021-01101-x). *Nature Methods*. 2021; **18**(4) 366–68.

Burgeff H, Haupt G: Chromosomzahlen bei der gattung *zygaena* (lep.). *Nachrichten der Akademie der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*. 1967; **9** 89–94.

Challis R, Kumar S, Sotero-Caio C, Brown M, Blaxter M: [Genomes on a Tree (GoaT): A versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life](https://doi.org/10.12688/wellcomeopenres.18658.1). *Wellcome Open Research*. 2023; **8** 24.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M: [BlobToolKit – interactive quality assessment of genome assemblies](https://doi.org/10.1534/g3.119.400908). *G3: Genes, Genomes, Genetics*. 2020; **10**(4) 1361–74.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H: [Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm](https://doi.org/10.1038/s41592-020-01056-5). *Nature Methods*. 2021; **18**(2) 170–75.

Cheng H, Jarvis ED, Fedrigo O, Koepfli KP, Urban L, et al.: [Haplotype-resolved assembly of diploid genomes without parental data](https://doi.org/10.1038/s41587-022-01261-x). *Nature Biotechnology*. 2022; **40**(9) 1332–35.

da Veiga Leprevost F, Grüning BA, Alves Aflitos S, Röst HL, Uszkoreit J, et al.: [BioContainers: An open-source and community-driven framework for software standardization](https://doi.org/10.1093/bioinformatics/btx192). *Bioinformatics*. 2017; **33**(16) 2580–82.

Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C: [Nextflow enables reproducible computational workflows](https://doi.org/10.1038/nbt.3820). *Nature Biotechnology*. 2017; **35**(4) 316–19.

Diesh C, Stevens GJ, Xie P, De Jesus Martinez T, Hershberg EA, et al.: [JBrowse 2: A modular genome browser with views of synteny and structural variation](https://doi.org/10.1186/s13059-023-02914-z). *Genome Biology*. 2023; **24**(1) 74.

Ewels P, Magnusson M, Lundin S, Käller M: [MultiQC: Summarize analysis results for multiple tools and samples in a single report](https://doi.org/10.1093/bioinformatics/btw354). *Bioinformatics*. 2016; **32**(19) 3047–48.

Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, et al.: [The nf-core framework for community-curated bioinformatics pipelines](https://doi.org/10.1038/s41587-020-0439-x). *Nature Biotechnology*. 2020; **38**(3) 276–78.

Fischer U, Sobczyk T: Rote liste der schwärmer und spinnerartigen schmetterlinge. Sächsisches Landesamt für Umwelt und Geologie, Dresden.

Formenti G, Abueg L, Brajuka A, Brajuka N, Gallardo-Alba C, et al.: [Gfastats: Conversion, evaluation and manipulation of genome sequences using assembly graphs](https://doi.org/10.1093/bioinformatics/btac460). *Bioinformatics*. 2022; **38**(17) 4214–16.

Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, et al.: [Bioconda: Sustainable and comprehensive software distribution for the life sciences](https://doi.org/10.1038/s41592-018-0046-7). *Nature Methods*. 2018; **15**(7) 475–76.

Habel JC, Engler JO, Rödder D, Schmitt T: [Contrasting genetic and morphologic responses on recent population decline in two burnet moths (*lepidoptera*, *zygaenidae*)](https://doi.org/10.1007/s10592-012-0372-8). *Conservation Genetics*. 2012; **13** 1293–1304.

Harry E: [PretextView (Paired REad TEXTure Viewer): A desktop application for viewing pretext contact maps](https://github.com/sanger-tol/PretextView).

Hofmann A, Tremewan WG: *The Natural History of Burnet Moths. Part III.2*. Munich: Museum Witt.

Howe K, Chow W, Collins J, Pelan S, Pointon DL, et al.: [Significantly improving the quality of genome assemblies through curation](https://doi.org/10.1093/gigascience/giaa153). *GigaScience*. 2021; **10**(1).

Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, et al.: [HiGlass: Web-based visual exploration and analysis of genome interaction maps](https://doi.org/10.1186/s13059-018-1486-1). *Genome Biology*. 2018; **19**(1) 125.

Kurtzer GM, Sochat V, Bauer MW: [Singularity: Scientific containers for mobility of compute](https://doi.org/10.1371/journal.pone.0177459). *PLOS ONE*. 2017; **12**(5) e0177459.

Larsen TB: Chromosome numbers in genus *zygaena* f. (*Lepidoptera*, *zygaenidae*). *Entomological Gazette*. 1976; **27** 157–60.

Li H: [Minimap2: Pairwise alignment for nucleotide sequences](https://doi.org/10.1093/bioinformatics/bty191). *Bioinformatics*. 2018; **34**(18) 3094–3100.

Lukhtanov VA, Kuznetsova VC: Karyotype structure in higher lepidoptera (*papilionomorpha*). *Entomological Review*. 1989; **68** 12–31.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM: [BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes](https://doi.org/10.1093/molbev/msab199). *Molecular Biology and Evolution*. 2021; **38**(10) 4647–54.

Merkel D: [Docker: Lightweight Linux containers for consistent development and deployment](https://dl.acm.org/doi/10.5555/2600239.2600241). *Linux J.* 2014; **2014**(239).

Oatley G, Sampaio F, Howard C: [Sanger Tree of Life fragmented DNA clean up: Automated SPRI](https://doi.org/10.17504/protocols.io.q26g7p1wkgwz/v1).

Pointon D-L, Eagles W, Sims Y, Muffato M, Surana P: [Sanger-tol/treeval v1.0.0 – Ancient Atlantis](https://doi.org/10.5281/zenodo.10047653).

Ranallo-Benavidez TR, Jaron KS, Schatz MC: [GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes](https://doi.org/10.1038/s41467-020-14998-3). *Nature Communications*. 2020; **11** 1432.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, et al.: [A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping](https://doi.org/10.1016/j.cell.2014.11.021). *Cell*. 2014; **159**(7) 1665–80.

Rennwald E, Sobczyk T, Hofmann A: Rote liste und gesamtartenliste der spinnerartigen falter (*lepidoptera*: Bombyces, sphinges s.l.) deutschlands. In *Rote Liste Gefährdeter Tiere, Pflanzen Und Pilze Deutschlands. Band 3: Wirbellose Tiere (Teil 1)*, eds M Binot-Hafke, S Balzer, N Becker, H Gruttke, H Haupt, et al., pp. 243–83. Bonn: Bundesamt für Naturschutz.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, et al.: [Towards complete and error-free genome assemblies of all vertebrate species](https://doi.org/10.1038/s41586-021-03451-0). *Nature*. 2021; **592**(7856) 737–46.

Rhie A, Walenz BP, Koren S, Phillippy AM: [Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies](https://doi.org/10.1186/s13059-020-02134-9). *Genome Biology*. 2020; **21**(1).

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, Blaxter M, Mieszkowska N, et al.: [MitoHiFi: A python pipeline for mitochondrial genome assembly from PacBio high fidelity reads](https://doi.org/10.1186/s12859-023-05385-y). *BMC Bioinformatics*. 2023; **24**(1) 288.

Vasimuddin Md, Misra S, Li H, Aluru S: [Efficient architecture-aware acceleration of BWA-MEM for multicore systems](https://doi.org/10.1109/IPDPS.2019.00041). *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 2019; 314–24.

Wachlin V, Kallies A, Hoppe H: ROTE LISTE der gefährdeten großschmetterlinge mecklenburg-vorpommerns. Der Minister für Landwirtschaft und Naturschutz des Landes Mecklenburg-Vorpommern, Schwerin.

Wermeille E, Chittaro Y, Gonseth Y: Rote liste tagfalter und widderchen. Gefährdete arten der schweiz, stand 2012. Bundesamt für Umwelt, Bern, und Schweizer Zentrum für die Kartografie der Fauna, Neuenburg.

Zhou C, McCarthy SA, Durbin R: [YaHS: Yet another Hi-C scaffolding tool](https://doi.org/10.1093/bioinformatics/btac808). *Bioinformatics*. 2023; **39**(1).