A Chromosome-Level Genome Assembly and Annotation for the Clouded Apollo Butterfly (*Parnassius mnemosyne*): A Species of Global Conservation Concern

Jacob Höglund\(^1\), Guilherme Dias\(^2\), Remi-André Olsen\(^3\), André Soares\(^2\), Ignas Bunikis\(^4,5\), Venkat Talla\(^6\), and Niclas Backström\(^6\),\(^*\)

1Animal Ecology Program, Department of Ecology and Genetics (IEG), Uppsala University, Uppsala SE-752 36, Sweden
2National Bioinformatics Infrastructure Sweden (NBIS), Science for Life Laboratory, Uppsala 752 37, Sweden
3Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, Solna 17165, Sweden
4Uppsala Genome Center, Department of Immunology, Genetics and Pathology, Uppsala University, National Genomics Infrastructure hosted by SciLifeLab, Uppsala, Sweden
5Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala 752 37, Sweden
6Evolutionary Biology Program, Department of Ecology and Genetics (IEG), Uppsala University, Uppsala SE-752 36, Sweden

\(^*\)Corresponding author: E-mail: niclas.backstrom@ebc.uu.se.

Accepted: February 10, 2024

Abstract

The clouded apollo (*Parnassius mnemosyne*) is a palearctic butterfly distributed over a large part of western Eurasia, but population declines and fragmentation have been observed in many parts of the range. The development of genomic tools can help to shed light on the genetic consequences of the decline and to make informed decisions about direct conservation actions. Here, we present a high-contiguity, chromosome-level genome assembly of a female clouded apollo butterfly and provide detailed annotations of genes and transposable elements. We find that the large genome (1.5 Gb) of the clouded apollo is extraordinarily repeat rich (73%). Despite that, the combination of sequencing techniques allowed us to assemble all chromosomes (\(n_c = 29\)) to a high degree of completeness. The annotation resulted in a relatively high number of protein-coding genes (22,854) compared with other Lepidoptera, of which a large proportion (21,635) could be assigned functions based on homology with other species. A comparative analysis indicates that overall genome structure has been largely conserved, both within the genus and compared with the ancestral lepidopteran karyotype. The high-quality genome assembly and detailed annotation presented here will constitute an important tool for forthcoming efforts aimed at understanding the genetic consequences of fragmentation and decline, as well as for assessments of genetic diversity, population structure, inbreeding, and genetic load in the clouded apollo butterfly.

Key words: *Parnassius mnemosyne*, clouded apollo butterfly, genome assembly, gene annotation, repeat annotation, genome size.
Significance
The quality of the assembly of the clouded apollo (Parnassius mnemosyne) genome and the annotation of genes are well in line with the standards of the Earth BioGenome Project (https://www.earthbiogenome.org/analysis-standards-report). The analyses revealed a comparatively large genome with a high repeat content for a butterfly and considerable synteny compared with both a close and a distant lepidopteran relative. We therefore predict that the genomic resources provided here will be important tools for forthcoming investigations aimed at understanding demographic history, loss of genetic variation, population structure, and mutation load in the clouded apollo butterfly in particular and for comparative genomic approaches in Lepidoptera in general.

Introduction
Genomic data can be used to inform practical conservation efforts of threatened species, a research field generally referred to as conservation genomics (e.g. Hohenlohe et al. 2021; DeWoody et al. 2022). Genomic approaches have, for example, proven useful to infer demographic histories, to estimate genetic diversity and inbreeding, and to quantify genetic load (Supple and Shapiro 2018; Wright et al. 2020; Bortoluzzi et al. 2023; Theissinger et al. 2023), information that is crucial for informed conservation actions for species that are rare or endangered and/or difficult to monitor with traditional methods (DeWoody et al. 2022; Hogg et al. 2022). Conservation genomic efforts have traditionally been focused on flagship species that comprise a small proportion of the extant biodiversity (Bortoluzzi et al. 2023). Butterflies and moths (Lepidoptera), for example, comprise over 150,000 species (Triant et al. 2018), but compared with mammals and birds, they are extensively overlooked in assessments of species under conservation concern (Cardoso et al. 2011; Podsiadlowski et al. 2021; Duffus and Morimoto 2022; Bortoluzzi et al. 2023). Many lepidopterans have also been reported to be rapidly declining in number (e.g. Cardoso et al. 2020; Warren et al. 2021), a major concern since they often constitute key indicator species for ecosystem functioning (Schowalter et al. 2018; Warren et al. 2021).

Several initiatives have been started, aiming at generating genomic data for specific organism groups (e.g. 5,000 insect genomes [i5k]; Sills et al. 2011), for biodiversity in certain countries (e.g. The Darwin Tree of Life [DTol] Project Consortium et al. 2022) or larger geographical regions (e.g. The European Reference Genome Atlas [ERGA]; Mc Cartney et al. 2023). While such initiatives are promising for future access to genome information for a large proportion of the extant biodiversity, targeted efforts are needed to swiftly get data for species under immediate conservation concern. As part of the ERGA pilot effort (Mc Cartney et al. 2023), we present a high-contiguity genome assembly and detailed annotation of the clouded apollo (Parnassius mnemosyne), a butterfly of general conservation concern (Gratton et al. 2008; Bolotov et al. 2013; Talla et al. 2023). The clouded apollo is widely distributed across the western Palearctic, but population numbers have declined dramatically in many regions (Kuussaari et al. 2015; Johansson et al. 2017; Talla et al. 2023), likely a consequence of a decline in the preferred habitat—mosaics of low-intensity grazed pastures intermixed with sheltering vegetation and rich access to host plants Corydalis spp. (Konvicka and Kuras 1999; Luoto et al. 2001; Kuussaari et al. 2015; Johansson et al. 2017). Previous genetic analyses have unveiled considerable population structure, both across the distribution range in general (Gratton et al. 2008) and between populations in specific regions (e.g. Sweden; Talla et al. 2023), but currently available data are insufficient for detailed assessment of demography, genetic load, and levels of genetic diversity. The chromosome-level genome assembly we developed here will therefore be a useful tool for forthcoming investigations of inbreeding, local adaptation and loss of genetic diversity in the clouded apollo, and comparative genomic analyses in Lepidoptera in general (Bortoluzzi et al. 2023).

Results
Assembly statistics are presented in Table 1. We predicted a genome size of ~1.4 Gbp and a genome-wide heterozygosity (π) of 0.74% (supplementary fig. S1, Supplementary Material online). The mean GC content was 37.8%, and nucleotide composition was stable across all chromosomes (Fig. 1). The assembly had high contiguity, with the longest 29 scaffolds (≥25 Mbp) comprising 95.2% of the total length. Of the 29 longest scaffolds, 10 contained telomere sequences at both ends, 13 at 1 end, and only 5 lacked telomere sequences completely (Fig. 1). We estimated the quality metrics recommended by the Earth BioGenome Project (https://www.earthbiogenome.org/analysis-standards-report) and found high base pair quality scores and a low rate of false duplications (supplementary table S1, Supplementary Material online).

The results from the Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis indicate that the assembly completeness is very high, with 99.0% of the 5,286 analyzed genes completely covered (Table 1, supplementary table S2, Supplementary Material online). A BUSCO analysis using the bacteria_odb10 gene set did
not identify any genes of bacterial origin in the final assembly.

The mtDNA contig was 15,425 bp, with 18.3% GC. The gene composition was typical for an insect mtDNA, with 13 protein-coding genes, 22 transfer RNA (tRNA) genes, and 2 ribosomal RNA (rRNA) genes (supplementary fig. S2, Supplementary Material online). The gene order was similar to that of the close relative Parnassius apollo (Chen et al. 2014) and more distantly related Lepidoptera such as the Monarch (Danaus plexippus; https://www.ncbi.nlm.nih.gov/nuccore/NC_021452.1) and the silkworm (Bombyx mori; https://www.ncbi.nlm.nih.gov/nuccore/NC_002355.1).

The automatized repeat annotation revealed that the majority of classified repeats was retroelements but also a high proportion of unclassified repeats (data not shown). We therefore made a more detailed analysis of repeats using partly manually curated repeat libraries for other butterfly species. This analysis showed that the overall proportion of repeats was significantly higher (72.90%) than with the automatized pipeline (61.53%) (supplementary table S4, Supplementary Material online). Both the clouded apollo and the close relative, apollo P. apollo (Podsiadlowski et al. 2021), have comparatively large genome sizes and a high repeat content. We therefore compared the relative contribution of different repeat classes to the genome sizes in both species using the same pipeline. This analysis revealed that the apollo has a higher repeat content than the clouded apollo and that the difference predominantly has been driven by expansions of long interspersed nuclear elements (LINEs), long terminal repeats (LTRs), and DNA transposons in the apollo (supplementary fig. S3, Supplementary Material online).

The structural annotation of coding sequences identified 22,854 genes and 27,378 transcripts (Table 1). A completeness assessment of the predicted proteins (longest isof orm used) indicated complete coverage of 96.3% of the 5,286 lepidopteran BUSCOs (supplementary table S5, Supplementary Material online). A total of 21,635 (94.7%) of the identified genes could be assigned a functional annotation with 14,890 unique gene names (Table 1, supplementary table S6, Supplementary Material online). We also predicted 1,076 putative tRNA genes, 4,788 putative pseudogenes, and 16,610 ncRNAs (including the tRNAs; Table 1).

The alignments between larger scaffolds revealed a high degree of synteny conservation between the 2 Apollo species (supplementary fig. S4, Supplementary Material online). The comparison between P. mnemosyne and B. mori unveiled a high degree of collinearity of aligned fragments within most scaffolds (supplementary fig. S5, Supplementary Material online).

**Discussion**

The assembly of the clouded apollo genome resulted in 29 chromosome-sized scaffolds where the majority seems to have sequence information including the telomeric repeats (data not shown). This is compatible with a chromosome-scale assembly since the clouded apollo has been shown to have 29 chromosomes (Vlašánek et al. 2017). The shorter contigs, not associated with the larger scaffolds, constituted a minute fraction of the entire assembly and can be omitted from most types of downstream analyses without risks for biased results. We predicted a considerably larger set of coding genes (~22,800) compared with most other lepidopterans with annotated genome sequences (e.g. Shipilina et al. 2022; Smolander et al. 2022; Höök et al. 2023). However, the gene number was lower than the estimated number (~28,300) in the close relative P. apollo (Podsiadlowski et al. 2021), likely a consequence of the more fragmented genome assembly and/or the less strict annotation of repeats in this species (Shipilina et al. 2022). The repeat content is high in P. mnemosyne compared with other lepidopterans, and this can lead to overestimation of the number of coding genes (e.g. Baril and Hayward 2022; Shipilina et al. 2022; Höök et al. 2023). It should also be noted that, with the exception of well-conserved families, many of the predicted ncRNAs should be considered with care.
The detailed repeat annotation of both *Parnassius* species confirmed that the extreme genome sizes are a consequence of accumulation of a high number of transposable elements (TEs) (Podsiadlowski et al. 2021). LINEs in particular constitute a large portion of the genome in both species, but also LTRs, DNA transposons, and short interspersed nuclear elements (SINEs) are present in high numbers. Another *Parnassius* species (*Parnassius orleans*) has an estimated genome size of 1.2 Gb (Liu et al. 2020), which shows that large genomes are ubiquitous in the genus and that the accumulation of TEs predominantly has occurred before the radiation of currently extant *Parnassius* lineages (Liu et al. 2020; Wiemers et al. 2020; Podsiadlowski et al. 2021). However, in contrast to other butterflies with large genomes and high TE content (e.g. *Leptidea* sp.; Talla et al. 2017; Höök et al. 2023) and in agreement with previous results (Podsiadlowski et al. 2021), our results show that the overall genome structure is relatively conserved in *Parnassius*.

**Methods**

**DNA and RNA Extractions and Library Preparations**

Two adult female clouded apollos from a captive population, recently established (2019) from a natural population (Blekinge; Talla et al. 2023) to aid the conservation actions in Sweden (https://en.nordensark.se/conservation/), were used for sequencing (sampled 2021 May 11). One individual (PM1) was used for PacBio HiFi sequencing and the other individual (PM2) for Illumina Hi-C and RNA-sequencing. For the PacBio HiFi and Illumina Hi-C libraries, frozen muscle tissue from the thorax was ground with pestle and mortar. Muscle tissue was also used for RNA-sequencing with Illumina short inserts and PacBio Iso-Seq. High molecular weight DNA was extracted from the muscle tissue using SDS lysis, followed by phenol:chloroform:isoamyl alcohol (25:24:1) treatment. DNA was precipitated with a high salt/low ethanol solution, washed twice with 70%
ethanol, and eluted in TE buffer. RNA was also extracted from muscle tissue using a standard TRIzol protocol including DNase treatment (RNeasy, Invitrogen). Eluted RNA was stored at −70 °C. The PacBio HiFi library preparation and sequencing was performed at the National Genomics Infrastructure (NGI) node in Uppsala, Sweden. All DNA and RNA extractions were done at Uppsala Genome Center (https://www.ucg.igp.uu.se/). Both RNA-Seq libraries and the Illumina Hi-C library were prepared at the ERGA node in Antwerp, Belgium, and sequenced at the node in Florence, Italy. Both individuals were kept as voucher specimens (PM1; ERGA-ID: ilParMnem1 as UPSZTY 184741 and PM2 as UPSZTY 184901) and stored at the Evolution Museum at Uppsala University (supplementary fig. S6, Supplementary Material online).

Genome Assembly

The genome was assembled using a combination of PacBio (HiFi) and Illumina Hi-C data (supplementary table S7, Supplementary Material online). The HiFi reads had a coverage of ~30x and were assembled using HiLiasm v0.16.0 (Cheng et al. 2021). Purge_Dups v1.2.5 (Guan et al. 2020) was used to remove putative duplications. The Hi-C sequence reads were aligned to the purged assembly and processed with pairtools v0.3.0 (Abdennur et al. 2023), and contigs were scaffolded with YaHS v1.1a (Zhou et al. 2022). Hi-C scaffolds were manually edited with JBAT (https://github.com/tokit/telomeric-identifier) to produce the final assembly.

Potential contamination was assessed using Mash v2.3 (Ondov et al. 2019) and the National Center for Biotechnology Information (NCBI) RefSeq database (https://gembox.cbcb.umd.edu/mash/refsseq.genomes.k21s10000.msh), and no substantial contamination was found. Genome properties were estimated from both the HiFi reads and the final assembly using the k-mer counter FastK (https://github.com/thegenemers/FastK) with 31 mers and GenomeScopeFK (https://github.com/theGenemers/GENESCOPE.FK), a modified version of GenomeScope v2.0 (Ranallo-Benavidez et al. 2020). Additionally, MERURY.FK (https://github.com/theGenemers/MERURY.FK) was used to estimate the k-mer completeness and the false duplication rate. Assembly quality was assessed with general statistics (Table 1), and completeness and duplication rate were evaluated using BUSCO v5.4.6 (Manni et al. 2021), with the lepidoptera_o_dbb10 gene set from OrthoDB v10 (Kriventseva et al. 2018). The mitochondrial genome was recovered from the primary assembly after haplotig removal using MitophiFi v2.2 (Uliano-Silva et al. 2023) and annotated using Mitofinder v1.4.1 (Allio et al. 2020). The mtDNA assembly was reoriented to match the P. apollo mtDNA assembly (NCBI accession NC_024727.1; Chen et al. 2014).

Gene Annotation

A custom repeat library was created with RepeatModeler2 v2.0.2a (Flynn et al. 2020), and the genome was soft-masked with RepeatMasker (https://www.repeatmasker.org/) v4.1.5 prior to gene annotation. The candidate repeats obtained by RepeatModeler were vetted against UniProt/SwissProt to exclude nucleotide motifs stemming from protein-coding sequences. For comparative purposes, we also generated a detailed annotation of the different classes of repeats present in the genomes of both the apollo (Podsiadlowski et al. 2021) and the clouded apollo butterfly and estimated the proportions of the genomes covered by each class of repeat, again using RepeatModeler2 v2.0.2a (Flynn et al. 2020), but including manually curated repeat libraries from other butterfly species (Shipilina et al. 2022; Höök et al. 2023).

Gene prediction was performed in 3 steps that were later combined, incorporating standard RNA-seq, protein sequences from multiple organisms, and PacBio Iso-Seq as evidence. (i) The RNA-seq reads for P. mnemosyne (Table 1) were aligned to the assembly with HiSat2 v2.1.0 (Kim et al. 2019), and BRAKER v3.0.3 (Gabriel et al. 2023) was used to extract splicing signals and to train and predict genes using Augustus. (ii) Arthropod proteins from OrthoDB v11 (https://bioinf.uni-greifswald.de/bioinf/partitioned_odb11; Kriventseva et al. 2018) were aligned to the assembly with miniprot v0.10-r226-dirty (Li, 2023) and used to train and predict genes with Augustus within the GALBA pipeline v1.0.6 (Brüna et al. 2023). (iii) High-quality transcripts from PacBio Iso-Seq were aligned to the genome using mimimap2 v2.26 (https://github.com/lh3/minimap2/releases) and used to obtain gene predictions with GeneMarkS-T v5.1 (Tang et al. 2015), following the long-read protocol from BRAKER (https://github.com/Gaius-Augustus/BRAKER/blob/master/docs/long_reads/long_read_protocol.md). Finally, the gene models from BRAKER, GALBA, and GeneMarkS-T were combined and filtered using TSEBRA (long_reads branch, commit 1f2614c; Gabriel et al. 2021). The combined gene models were processed with AGAT v1.2.0 (https://zenodo.org/record/8178877) to remove overlapping genes, and functional annotation was done using the NCBI functional_annotation nextflow pipeline v2.0.0 (https://github.com/NCBISweden/pipelines-nextflow), which uses BLAST for similarity searches between the annotated proteins and the UniProtKB/SwissProt database (Magrane and UniProt Consortium 2011) (downloaded 2022 to 2012, 568,363 proteins), InterProScan (Jones et al. 2014) to query the proteins against InterPro v59-91 (Paysan-Lafosse et al. 2022), and merge results using AGAT (https://zenodo.org/record/8178877). To reduce potential false-positives, single-exon genes without any InterPro annotation were removed. Each predicted protein sequence was blasted against the
UniProt/SwissProt reference data set (downloaded 2022 to 2012) in order to infer, when available, the gene and protein name. The inference was made using the best hit with a maximum e-value cutoff of $1 \times 10^{-6}$. Gene names that were encountered several times were suffixed to keep each gene name unique (complete list of gene names is available upon request). tRNAs were predicted using tRNAscan-SE v2.0.12 (https://github.com/UCSC-LoweLab/tRNAscan-SE) with default parameters for eukaryotes. Noncoding RNAs (ncRNAs) were predicted with Infernal v1.1.4 (Nawrocki and Eddy 2013) and the Rfam v14.1 covariance models (Nawrocki et al. 2014).

To assess synteny conservation, whole genome alignments between *P. mnemosyne* and *P. apollo* and *B. mori*, respectively, were done using D-GENIES (Cabanettes and Klopp 2018) with default settings. The scaffold plot in Fig. 1 was generated with the R package Rideogram (Hao et al. 2020).

**Supplementary Material**

Supplementary material is available at Genome Biology and Evolution online.

**Acknowledgments**

The authors would also like to acknowledge the support from the Science for Life Laboratory (SciLifeLab), NGI, and Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) for providing assistance and Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under the Project SNIC 2020-5-20. We thank Veronika Mrazek for the help with the dissections.

**Author Contributions**


**Funding**

This work was supported by the Swedish Research Council (VR research grant #019-04791 to N.B.), NBIS/SciLifeLab long-term bioinformatics support (WABI to N.B.), and Swedish Rescue Program for *P. mnemosyne* through the local administrative board (Länstyrelsen) of Blekinge (to J.H. and N.B.).

**Data Availability**

All data have been deposited at the European Nucleotide Archive under accession PRJEB67749. Scripts used for the TE analysis are available at GitHub (https://github.com/EBC-butterfly-genomics-team/clouded_apollo_genomics). Assembly and annotation files are also available via the ERGA portal (https://portal.erga-biodiversity.eu/data_portal/Parnassius%20mnemosyne).

**Literature Cited**


