



HAL
open science

Chromosome-Level Assembly and Annotation of the Pearly Heath *Coenonympha arcania* Butterfly Genome

Fabrice Legeai, Sandra Romain, Thibaut Capblancq, Paul Doniol-Valcroze, Mathieu Joron, Claire Lemaitre, Laurence Després

► **To cite this version:**

Fabrice Legeai, Sandra Romain, Thibaut Capblancq, Paul Doniol-Valcroze, Mathieu Joron, et al.. Chromosome-Level Assembly and Annotation of the Pearly Heath *Coenonympha arcania* Butterfly Genome. *Genome Biology and Evolution*, 2024, pp.1-7. 10.1093/gbe/evae055 . hal-04530573

HAL Id: hal-04530573

<https://hal.science/hal-04530573>

Submitted on 3 Apr 2024




HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Chromosome-Level Assembly and Annotation of the Pearly Heath *Coenonympha arcania* Butterfly Genome

Fabrice Legeai ^{1,2,*}, Sandra Romain¹, Thibaut Capblancq³, Paul Doniol-Valcroze⁴, Mathieu Joron⁴, Claire Lemaitre ¹, and Laurence Després ³

¹Inria, CNRS, IRISA, University of Rennes, 35000 Rennes, France

²IGEPP, INRAE, Institut Agro, University of Rennes, 35653 Le Rheu, France

³LECA, CNRS, Université Grenoble-Alpes, Université Savoie Mont Blanc, Grenoble, France

⁴CEFE, CNRS, EPHE, IRD, Université de Montpellier, Montpellier, France

*Corresponding author: E-mail: fabrice.legeai@inrae.fr.

Accepted: March 13, 2024

Abstract

We present the first chromosome-level genome assembly and annotation of the pearly heath *Coenonympha arcania*, generated with a PacBio HiFi sequencing approach and complemented with Hi-C data. We additionally compare synteny, gene, and repeat content between *C. arcania* and other Lepidopteran genomes. This reference genome will enable future population genomics studies with *Coenonympha* butterflies, a species-rich genus that encompasses some of the most highly endangered butterfly taxa in Europe.

Key words: chromosome-level assembly, *Coenonympha arcania*, butterfly, annotation, Satyrinae.

Significance

A high-quality reference genome is a prerequisite for modern population genetics and conservation genomics projects. *Coenonympha arcania* is part of a species-rich genus, including multiple taxa of high conservation concern and at least 2 intensively studied hybrid species. The newly proposed high-quality and chromosome-level assembly will be a key resource for upscaling the study of the genus to whole-genome data.

Introduction

Butterflies have been intensively studied for their ecology, biogeography, and evolutionary history and are sensitive indicators of environmental changes, such as habitat fragmentation and climate change (Kühn et al. 2005). The genus *Coenonympha* (Nymphalidae, Satyrinae) is a species-rich group comprising more than 30 species mostly distributed in Eurasia (Kodandaramaiah and Wahlberg 2009), which diverged early in the radiation of Satyrinae and lacks any close

relative genus (Kodandaramaiah et al. 2010; Wiemers et al. 2020). The phylogeny of the genus is poorly resolved with controversial conclusions depending on the molecular marker used (Peña et al. 2006; Kodandaramaiah and Wahlberg 2009). The different taxa are easily distinguished by their hindwing patterns (eyespot and color patterns) and are found in contrasted habitats. Some studies report occasional hybridization between species with overlapping distribution ranges and even the formation of stable hybrid

© The Author(s) 2024. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

species (Capblancq et al. 2015, 2019, 2020). While most *Coenonympha* species can be relatively abundant locally, some have strict habitat preferences and are at high risk of extinction. Three species *C. hero*, *C. tullia*, and *C. oedippus* are critically endangered in Europe and benefit from a high protection status (Settele et al. 2008). Yet, in *Coenonympha*, a basal genus in the Satyrinae sub-family, the lack of reference genome has limited population genetics studies to de novo reconstructed ddRAD loci (Capblancq et al. 2015, 2019, 2020; Després et al. 2019; Sherpa et al. 2022; Kebaïli et al. 2023).

Here, we present a chromosome-level genome assembly and annotation of a representative of the *Coenonympha* genus, the pearly heath *C. arcania* (Linnaeus 1761) (Lepidoptera, Nymphalidae, Satyrinae). It is a relatively common and locally abundant species widely distributed in semi-open dry grasslands in Europe from the northern Mediterranean to south-central Scandinavia and in the east from Turkey to the Urals (Besold et al. 2008). Its genome was generated with a long-read PacBio HiFi sequencing approach, complemented with Hi-C scaffolding. Additional analyses were performed on gene and repeat contents in comparison with 2 other high-quality Satyrinae genomes (*Maniola jurtina* and *Pararge aegeria*) available from the Darwin Tree of Life project (Lohse et al. 2021a, Lohse and Weir 2021b). The synteny of these species genomes was also investigated, additionally including a recently published genome assembly of another *Coenonympha* species: the chestnut heath *C. glycerion* (ENA project PRJEB71111).

Results and Discussion

Genome Sequence Statistics

HiFi long-read sequencing (PacBio Sequel II) yielded 1.4 million reads (197 Gb, N50 = 14 kb) for the pearly heath butterfly *C. arcania* butterfly, for an estimated genome coverage of 44× (for details, see [supplementary table S1, Supplementary Material](#) online). These long reads were assembled into 110 contigs with Hifiasm (Cheng et al. 2021). Next, haplotypic duplications were removed with purge_dups v1.2.5 (Guan et al. 2020), using the HiFi reads and a complementary Chromium 10× library from a *C. arcania* male for the assessment of local coverage, leading to 47 contigs (N50 = 16.5 Mb). Contigs were scaffolded with 141 million Omni-C reads using YaHS v1.2a (Zhou et al. 2023), giving an assembly of 497 Mb in 35 scaffolds (N50 = 18.8 Mb). After visual control of the Hi-C contact map, 3 scaffolds were split resulting in a final assembly of 39 scaffolds including 32 scaffolds larger than 3 Mb and considered to be of chromosome size. Of these 32 scaffolds, we identified the largest scaffold as the Z chromosome, the 2 smallest scaffolds as putatively part of the W chromosome and 29 large autosome-like scaffolds, close to the 28 autosome pairs

observed in karyotypes of *C. arcania* (de Lesse 1960). This final assembly has an N50 of 17.9 Mb and BUSCO score of 99.0%, with 98.0% of genes being single copy and complete (Fig. 1a; [supplementary fig. S1b and table S3, Supplementary Material](#) online). STAR v2.10.7b (Dobin et al. 2013) reports that 70.0% of the raw RNASeq reads uniquely mapped the genome, while 7.5% mapped at multiple locations ([supplementary table S1, Supplementary Material](#) online). When raw reads were trimmed on quality with the ncore-rnaseq workflow (Ewels et al. 2020), a better mapping rate of 86.59% was achieved.

Gene Model Predictions and Annotation

Protein-coding genes were annotated using 2 approaches: Braker v3.0.3 (Brůna et al. 2021) complemented by GUSHR (<https://github.com/Gaius-Augustus/GUSHR>) based on GeMoMa (Keilwagen et al. 2018) and Helixer v0.3.0 (Holst et al. 2023). While the first uses a de novo approach extended with evidence from protein similarities or RNASeq alignments, Helixer implements a deep learning method based only on proteome characteristics from various invertebrate genomes. The number of missing BUSCO genes was slightly higher in the Braker3 annotation, but both methods gave globally good quality results with a similar number of around 21,500 protein-coding genes and retrieved genes of interest, such as wing color pattern genes and chemosensory genes. In order to determine the number of proteins only reported by 1 of 2 methods (i.e. specific to 1 method), we ran Orthofinder (Emms and Kelly 2019) on the 2 annotations, including only the longest isoform for each gene. Around 30% of the proteins were specific to 1 method. Both programs predict untranslated regions (UTR) exons at short distance of the stop codon, in line with other detailed studies from several *Drosophila* species (Sanfilippo et al. 2017; Wang et al. 2019), but UTRs predicted by Braker3 (with RNASeq evidence) are longer and more consistent with the sizes observed in other Nymphalidae ([supplementary table S3 and fig. S3, Supplementary Material](#) online, Lo Giudice et al. 2023).

We compared the gene repertoire predicted by Helixer with other Lepidopteran species, namely 2 other Satyrinae species (*M. jurtina* and *P. aegeria*), 4 other Nymphalidae and *Bombyx mori*, using Orthofinder v2.5.5 (Emms and Kelly 2019). Among the 42,461 identified orthogroups, 12,769 include a *C. arcania* protein, and 3,802 genes (17.77%) are specific to *C. arcania* ([supplementary table S5 and fig. S4, Supplementary Material](#) online).

Repeat Content

Based on transposable elements from the Insecta library of RepeatMasker (Smit et al. 2013-2015), an overall similar content in repetitive elements was found in all 3 Satyrinae genomes, respectively 12.27%, 14.87%,

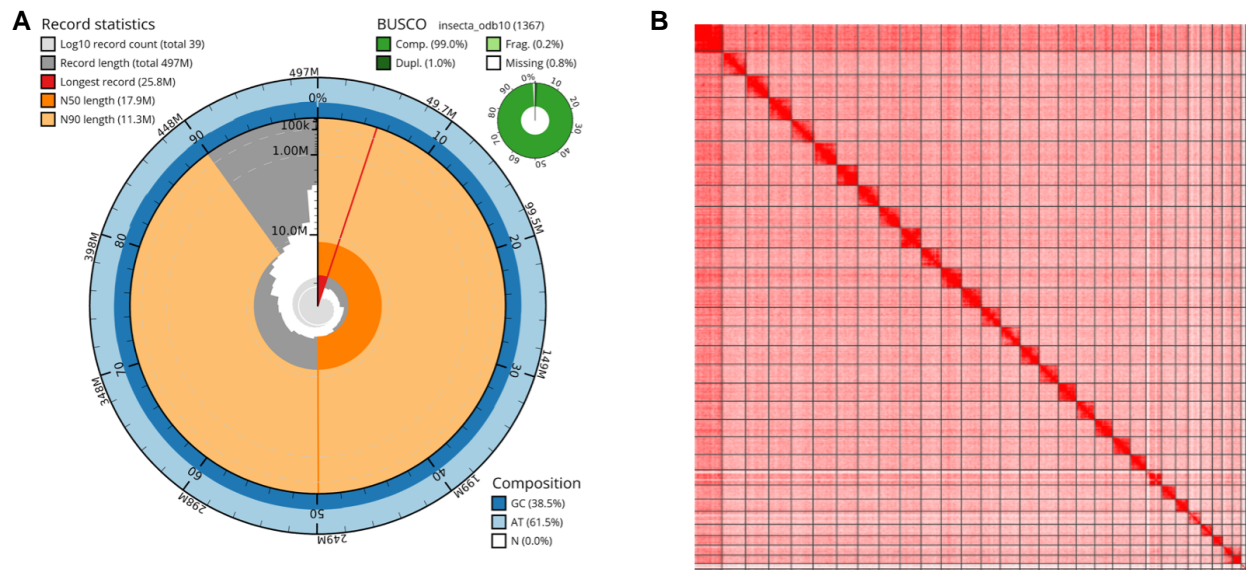


Fig. 1.—The overview of the *C. arcania* genome assembly. a) Genome assembly statistics shown as a SnailPlot from BlobToolKit, with contiguity (top-left and main circular plot), BUSCO (top right), and GC content (bottom right) metrics. b) The heatmap of chromosomal contact (Hi-C) data. Scaffolds are ordered by decreasing size from left to right, and vertical and horizontal dark lines delineate their boundaries. The color intensity of each pixel represents the frequency of interactions between genomic loci.

and 16.78% for *C. arcania*, *M. jurtina*, and *P. aegeria*. However, the genome of *C. arcania* contains more DNA transposons and LTR elements (supplementary fig. S5a and b, Supplementary Material online), and the de novo approach of RepeatModeler (Flynn et al. 2020) identified a larger proportion of interspersed repeats, namely 43.01%, 37.34%, and 39.24% for *C. arcania*, *M. jurtina*, and *P. aegeria*, respectively, yet around half of these remain unclassified.

Comparative Analysis

To assess genome synteny among Satyrinae genomes, we compared the *C. arcania* genome against *C. glycerion* (ENA project PRJEB71111), *M. jurtina* (Lohse et al. 2021a), and *P. aegeria* (Lohse and Weir 2021b). Orthologous gene order analyses with GENESPACE (Lovell et al. 2022) revealed a high degree of chromosomal synteny between these genomes, suggesting very few large-scale chromosomal rearrangements between these 4 Satyrinae taxa (Fig. 2). This strong level of synteny was also confirmed by whole genome alignments (supplementary fig. S5c, Supplementary Material online).

This comparative analysis further supports our hypothesis that scaffold 29 and scaffold 30 were misassembled and belong to a single chromosome in the *C. arcania* genome. We expect only 28 autosomes from the karyotype of *C. arcania* (de Lesse 1960), and these 2 scaffolds are the most likely candidates to be fused together, as they appear linked more strongly than any other autosome pair on the Hi-C contact map (Fig. 1b) and show high synteny with a unique chromosome in *C. glycerion*, *M. jurtina*, and *P. aegeria* (Fig. 2). Unfortunately, we were not able to

confirm this fusion with the long reads and therefore kept the 2 scaffolds separated.

Chromosomes Z and W

To identify sex chromosomes, the coverage of the raw HiFi reads of the female individual (ZW) used for the assembly was compared with Chromium-10x reads obtained from a male individual (ZZ; supplementary table S4, Supplementary Material online). The scaffold 1 shows approximately half the coverage of the other scaffolds in the female only, suggesting it is the chromosome Z. This is corroborated by the strong synteny between scaffold 1 and the Z chromosomes of *C. glycerion*, *M. jurtina*, and *P. aegeria* (Fig. 2). Unfortunately, we were not able to identify the W chromosome with certainty by comparing the chromosomal median coverage of HiFi reads, probably because of its high repeat content. However, the Chromium-10x reads obtained from a male yield zero median coverage on scaffolds 31 and 32 supporting the hypothesis that these 2 small scaffolds (total size = 6.9 Mb) are part of the sex chromosome W. Furthermore, these 2 scaffolds exhibit a lower rate of polymorphism within mapped HiFi reads compared with autosomal scaffolds, as expected for sexual chromosomes in females and as observed for the Z chromosome (scaffold_1; see supplementary table S4, Supplementary Material online).

Materials and Methods

DNA Extraction and Sequencing

High molecular weight DNA was extracted using Genra Puregene kit from Qiagen, according to the manufacturer's

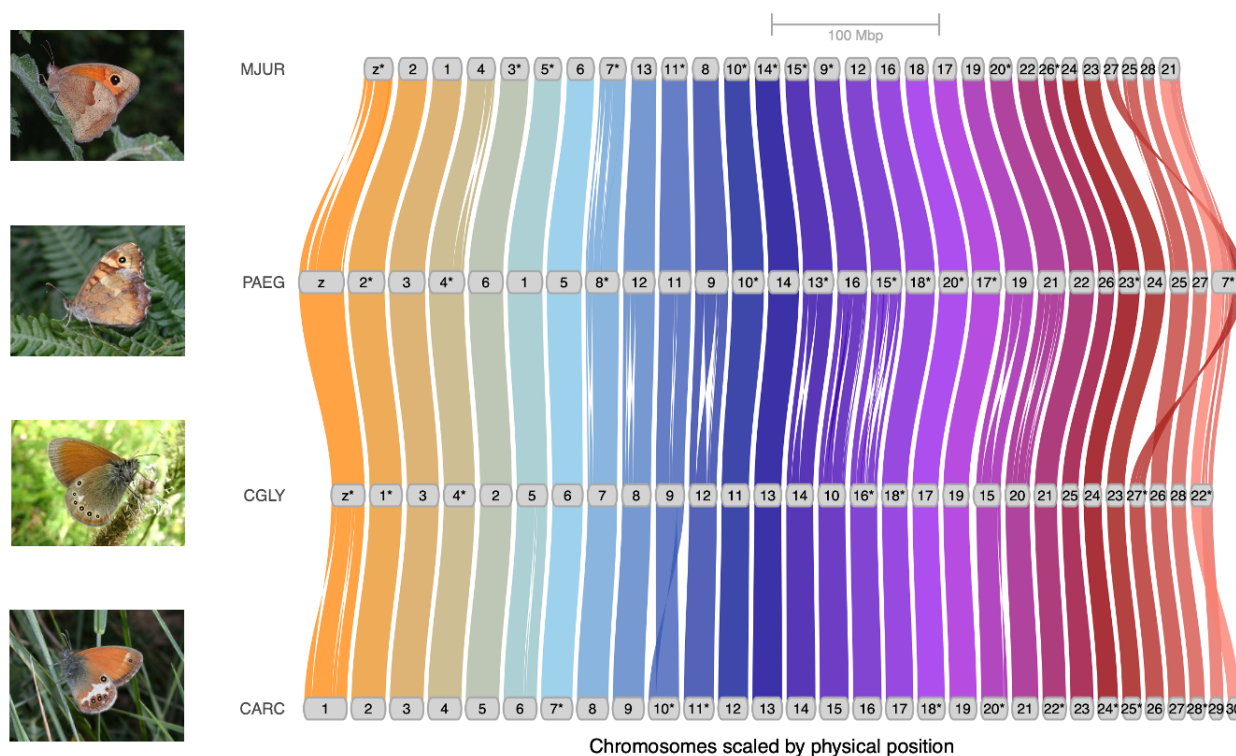


Fig. 2.—Comparative genomics between *C. arcania* and 3 other Satyrinae genomes. Syntenic blocks were obtained by comparing orthologous gene orders on *C. arcania* (CARC, bottom line) with *C. glycerion* (CGLY), *P. aegeria* (PAEG), and *M. jurtina* (MJUR). The stars (*) indicate chromosomal sequences that were reversed to ease the visualization. Photos: Philippe Mothiron (CARC, PAEG, and MJUR) and Jean-Pierre Arnaud (CGLY).

instructions. For the PacBio long-read sequencing performed at the GenoToul Platform (Toulouse, France), DNA was extracted from a freshly emerged *C. arcania* female (BST1) obtained in September 2021 from an egg of a *C. arcania* female caught in June 2021 at La Bastille (Grenoble, France; [supplementary fig. S1a, Supplementary Material](#) online). For the 10x genomics sequencing performed at the MGX platform (Montpellier, France), DNA was extracted from a wild caught male *C. arcania* (ARC2) from Saint Michel de Chaillol, France. For the Hi-C library preparation, DNA was extracted from *C. arcania* larvae reared from the eggs of wild caught females from La Mure (France). Final DNA purity and concentrations were measured by spectrometry using Nanodrop and fluorometry using Qubit (ThermoFisher).

RNA Isolation and Sequencing

Tissues from 1 lab-reared *C. arcania* adult female and 1 *C. arcania* larva were separately extracted with Trizol. mRNA purification, library preparation, and subsequent short-read sequencing (Illumina Novaseq) were performed by NOVOGENE (Cambridge, UK).

Assembly Quality Assessment

The absence of contamination in the assembly was confirmed by looking at scaffolds with unexpected coverage, GC

percent, or similarities (distinguishing putative contaminants), using Blobtoolkit v0.4.7 (Challis et al. 2020) with the HiFi sequences coverage (HiFi reads mapped with minimap2 (Li 2018) with the -ax map-pb option) or the similarities with NCBI NT database (2023-08-21 version) obtained by blastn v2.12.0 (Camacho et al. 2008), with the options -max_target_seqs 10, -max_hsps 1, and -evalue 1e-25. The level of completeness of the assembly was estimated using BUSCO v5.2.2 (Simão et al. 2015) scores against insecta_odb10 and lepidoptera_odb10 gene catalogs. We checked the level of haplotypic duplications caused by high heterozygosity level by comparing kmer abundances in the raw HiFi read set versus in the assembly using DSK v2.3.3 (Rizk et al. 2013) to count 31-mers and display the abundance profiles in a kmer-comparison plot ([supplementary fig. S2, Supplementary Material](#) online).

Identification of Sex Chromosomes

Male (ZZ) read coverage was determined using Chromium 10x reads aligned with bwa mem v 0.7.17 (Li 2013) with default parameters. Alignments with a quality score above 20 and proper pairing were retained using samtools view v1.15 (Danecek et al. 2021) with parameters -q20 -f 0x2. Coverage in 1,000-bp regions was calculated from the resulting bam file using mosdepth v0.3.4 (Pedersen and

Quinlan 2018) with options (-b 1000 -n -m). A VCF file was generated from the bam of HiFi reads alignments with bcftools v1.16 (Danecek et al. 2021) mpileup using parameters -Ou, followed by bcftools call -mv -Ob. Variants were selected if covered by at least 5 reads and called heterozygous when the alternative allele depth of coverage ranged from 20% to 80%.

Annotation of Protein-Coding Genes

Braker v3.0.3 (Brůna et al. 2021) was used with the following inputs: (i) unfiltered bams of the alignments of the raw reads of 2 RNAseq libraries obtained with STAR v2.7b (Dobin et al. 2013) using the options -outFilterMultimapNmax 5 -outFilterMismatchNmax 3 -alignIntronMin 10 -alignIntronMax 50000 -alignMatesGapMax 50000 and (ii) the fasta files of 16 butterflies proteomes downloaded from Lepbase release 4 (<http://download.lepbase.org/v4/sequence/>, Challi et al. 2016; [supplementary table S2, Supplementary Material](#) online) complemented with the proteome of *M. jurtina* (Lohse et al. 2021a). Finally, the GTF file of Braker was supplemented with the UTRs predicted with GUSHR (<https://github.com/Gaius-Augustus/GUSHR>) based on GeMoMa (Keilwagen et al. 2018) and the bam file from STAR.

In parallel, another set of protein coding genes was annotated using Helixer v0.3.0 (Holst et al. 2023) with the option "-lineage invertebrate," which corresponds to the matrix invertebrate_v0.3_m_0100, which includes 65 training invertebrate genomes (including *Bicyclus anynana*, *B. mori*, *Papilio machaon*, *Pieris rapae*, and *Plutella xylostella*) and 136 genomes for validation.

Assignment of RNAseq fragments to genes was conducted from bam files after removing potential PCR duplicates with the nfcore rnaseq pipeline v3.10.1 (Ewels et al. 2020) with default parameters (i.e. using the STAR mapper), followed by FeatureCounts from Subread v2.0.1 (Liao et al. 2014). Eighteen wing color pattern genes were extracted from the *Danaus plexippus* V4 genome annotation (Zhan et al. 2011), and 356 chemosensory genes (43 CSP, 40 OBP, 36 IR, 167 GR, and 70 OR) were gathered from the *Ithomia salapia* genome annotation (Gauthier et al. 2023). All were aligned with BLASTP v2.12.0 (Camacho et al. 2008) to the Braker3 and Helixer protein sets with a maximal e-value of $1e-20$. *Coenonympha arcania* proteins with 60% identity with a match covering more than 80% of the subject were considered as complete.

Synteny

The identification of synteny blocks between the genomes of *C. arcania*, 2 other Satyrini, *M. jurtina* and *P. aegeria* (accessions GCA_905333055.1 and GCA_905163445.1, respectively), and *C. glycerion* (GCA_963855885.1) was performed at the gene level and by aligning whole genome sequences. Because the annotation of the *C. glycerion*

genome was not available, we first annotated it using Helixer v0.3.0 (Holst et al. 2023). At the gene level, GENESPACE (Lovell et al. 2022) was used based on the orthogroups calculated by Orthofinder (Fig. 2). Pairwise whole genome alignments were performed with SibeliaZ v1.2.5 (Minkin and Medvedev 2020) with the options -n and -a 16, and the obtained local alignments were clustered in large synteny blocks using maf2synteny v1.2 (Kolmogorov et al. 2018), with a final minimal block size of 500 bp (option -b). Each of the 500+ bp synteny blocks between *C. arcania* and *M. jurtina* was plotted as a Circos link and colored after the *M. jurtina* chromosome of the block. The order of the *C. arcania* chromosomes was rearranged to maximize the visual symmetry of the synteny blocks between the 2 species ([supplementary fig. S5, Supplementary Material](#) online).

Annotation of Transposable Elements

The genomes of *C. arcania*, *M. jurtina* and *P. aegeria* were compared with the insect library of RepeatMasker v4.1.5 (Smit et al. 2013-2015) with the options "-species insecta -xsmall -gff." Transposable elements were also identified and classified using a de novo strategy using RepeatModeler v2.0.4 (Flynn et al. 2020). Finally, the classified de novo repeats were aligned against the corresponding genomes with repeatMasker.

Supplementary Material

[Supplementary material](#) is available at *Genome Biology and Evolution* online.

Acknowledgments

We are thankful to the BIPAA bioinformatics platform for hosting the genome and to the GenOuest bioinformatics platform for supporting the calculations. We also thank Nathalie Rodde from CNRGV, Toulouse, France, for her help in the construction of the Omni-C library.

Funding

This work was supported by the French National Research Agency (ANR-20-CE02-0017).

Data Availability

The raw data (HiFi, 10X, Hi-C and RNAseq) and the genome sequence are available at NCBI under the project PRJNA1022034. The genome sequence and the 2 annotations are also available publicly at https://bipaa.genouest.org/sp/coenonympha_arcania/.

Literature Cited

Besold J, Schmitt T, Tammaru T, Cassel-Lundhagen A. Strong genetic impoverishment from the centre of distribution in Southern Europe

- to peripheral Baltic and isolated Scandinavian populations of the pearly heath butterfly. *J Biogeogr.* 2008;35(11):2090–2101. <https://doi.org/10.1111/j.1365-2699.2008.01939.x>.
- Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. bioinform.* 2021;3(1):lqaa108. <https://doi.org/10.1093/nargab/lqaa108>.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinform.* 2008;10(1):421. <https://doi.org/10.1186/1471-2105-10-421>.
- Capblancq T, Després L, Mavárez J. Genetic, morphological and ecological variation across a sharp hybrid zone between two alpine butterfly species. *Evol Appl.* 2020;13(6):1435–1450. <https://doi.org/10.1111/eva.12925>.
- Capblancq T, Després L, Rioux D, Mavárez J. Hybridization promotes speciation in *Coenonympha* butterflies. *Mol Ecol.* 2015;24(24):6209–6222. <https://doi.org/10.1111/mec.13479>.
- Capblancq T, Mavárez J, Rioux D, Després L. Speciation with gene flow: evidence from a complex of alpine butterflies (*Coenonympha*, Satyridae). *Ecol Evol.* 2019;9(11):6444–6457. <https://doi.org/10.1002/ece3.5220>.
- Challis RJ, Kumar S, Dasmahapatra KK, Jiggins CD, Blaxter M. 2016. Lepbase: the Lepidopteran genome database. *bioRxiv* 056994. <https://doi.org/10.1101/056994>.
- Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit—interactive quality assessment of genome assemblies. *G3 (Bethesda)*. 2020;10(4):1361–1374. <https://doi.org/10.1534/g3.119.400908>.
- Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods.* 2021;18(2):170–175. <https://doi.org/10.1038/s41592-020-01056-5>.
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. *GigaScience.* 2021;10(2):giab008. <https://doi.org/10.1093/gigascience/giab008>.
- de Lesse H. Spéciation et variation chromosomique chez les Lépidoptères Rhopalocères. *Ann Soc Nat Zool.* 1960;12:1–223.
- Després L, Henniaux C, Rioux D, Capblancq T, Zupan S, Čelik T, Ficetola GF. Inferring the biogeography and demographic history of an endangered butterfly in Europe from multilocus markers. *Biol J Linn Soc.* 2019;126(1):95–113. <https://doi.org/10.1093/biolinnean/bly160>.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
- Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20(1):238. <https://doi.org/10.1186/s13059-019-1832-y>.
- Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol.* 2020;38(3):276–278. <https://doi.org/10.1038/s41587-020-0439-x>.
- Flynn JM, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A.* 2020;117(17):9451–9457. <https://doi.org/10.1073/pnas.1921046117>.
- Gauthier J, Meier J, Legeai F, McClure M, Whibley A, Bretaudeau A, Boulain H, Parrinello H, Mugford ST, Durbin R, et al. First chromosome scale genomes of ithomiine butterflies (Nymphalidae: Ithomiini): comparative models for mimicry genetic studies. *Mol Ecol Resour.* 2023;23(4):872–885. <https://doi.org/10.1111/1755-0998.13749>.
- Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics.* 2020;36(9):2896–2898. <https://doi.org/10.1093/bioinformatics/btaa025>.
- Holst F, Bolger A, Günther C, Maß J, Triesch S, Kindel F, Kiel N, Saadat N, Ebenhöh O, Usadel B, et al. 2023. Helixer—de novo prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. *bioRxiv* 527280. <https://doi.org/10.1101/2023.02.06.527280>.
- Kebaili C, Sherpa S, Guéguen M, Renaud J, Rioux D, Després L. Comparative genetic and demographic responses to climate change in three peatland butterflies in the Jura massif. *Biol Conserv.* 2023;287:110332. <https://doi.org/10.1016/j.biocon.2023.110332>.
- Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics.* 2018;19(1):189. <https://doi.org/10.1186/s12859-018-2203-5>.
- Kodandaramaiah U, Peña C, Braby MF, Grund R, Müller CJ, Nylin S, Wahlberg N. Phylogenetics of Coenonymphina (Nymphalidae: Satyrinae) and the problem of rooting rapid radiations. *Mol Phylogenet Evol.* 2010;54(2):386–394. <https://doi.org/10.1016/j.ympev.2009.08.012>.
- Kodandaramaiah U, Wahlberg N. Phylogeny and biogeography of Coenonympha butterflies (Nymphalidae: Satyrinae)—patterns of colonization in the Holarctic. *Syst Entomol.* 2009;34(2):315–323. <https://doi.org/10.1111/j.1365-3113.2008.00453.x>.
- Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, Yang F, Odom D, Flicek P, Keane TM, Thybert D, et al. Chromosome assembly of large and complex genomes using multiple references. *Genome Res.* 2018;28:1720–1732. <https://doi.org/10.1101/gr.236273.118>.
- Kühn E, Settele J, Thomas JA. Studies on the ecology and conservation of butterflies in Europe. Pensoft; 2005.
- Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997v2 [q-bio.GN]*.
- Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>.
- Liao Y, Smyth GK, Shi W. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics.* 2014;30(7):923–930. <https://doi.org/10.1093/bioinformatics/btt656>.
- Linnaeus C. *Philosophia Botanica*. Impensis Direct. Laurentii Salvii; 1761.
- Lo Giudice C, Zambelli F, Chiara M, Pavesi G, Tangaro MA, Picardi E, Pesole G. UTRdb 2.0: a comprehensive, expert curated catalog of eukaryotic mRNAs untranslated regions. *Nucleic Acids Res.* 2023;51(D1):D337–D344. <https://doi.org/10.1093/nar/gkac1016>.
- Lohse K, Taylor-Cox E. The genome sequence of the speckled wood butterfly, *Pararge aegeria* (Linnaeus, 1758). *Wellcome Open Res.* 2021a;6:287. <https://doi.org/10.12688/wellcomeopenres.17278.1>.
- Lohse K, Weir J. The genome sequence of the meadow brown, *Maniola jurtina* (Linnaeus, 1758). *Wellcome Open Res.* 2021b;6:296. <https://doi.org/10.12688/wellcomeopenres.17304.1>.
- Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, Emms D, Goodstein DM, Schmutz J. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife.* 2022;11:e78526. <https://doi.org/10.7554/eLife.78526>.
- Minkin I, Medvedev P. Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. *Nat Commun.* 2020;11:6327. <https://doi.org/10.1038/s41467-020-19777-8>.
- Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics.* 2018;34(5):867–868. <https://doi.org/10.1093/bioinformatics/btx699>.

- Peña C, Wahlberg N, Weingartner E, Kodandaramaiah U, Nylin S, Freitas AVL, Brower AVZ. Higher level phylogeny of Satyrinae butterflies (Lepidoptera: Nymphalidae) based on DNA sequence data. *Mol Phylogenet Evol.* 2006;40(1):29–49. <https://doi.org/10.1016/j.ympev.2006.02.007>.
- Rizk G, Lavenier D, Chikhi R. DSK: k-mer counting with very low memory usage. *Bioinformatics.* 2013;29(5):652–653. <https://doi.org/10.1093/bioinformatics/btt020>.
- Sanfilippo P, Wen J, Lai EC. Landscape and evolution of tissue-specific alternative polyadenylation across *Drosophila* species. *Genome Biol.* 2017;18(1):229. <https://doi.org/10.1186/s13059-017-1358-0>.
- Settele J, Kudrna O, Harpke A, Kühn I, Van Swaay C, Verovnik R, Warren M, Wiemers M, Hanspach J, Hickler T, et al. Climatic risk atlas of European butterflies. Sofia: Pensoft; 2008. p. 392–393.
- Sherpa S, Kebaili C, Rioux D, Guéguen M, Renaud J, Després L. Population decline at distribution margins: assessing extinction risk in the last glacial relictual but still functional metapopulation of a European butterfly. *Divers Distrib.* 2022;28(2):271–290. <https://doi.org/10.1111/ddi.13460>.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 2015;31(19):3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>.
- Smit AFA, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. <http://www.repeatmasker.org>.
- Wang W, Fang D-H, Gan J, Shi Y, Tang H, Wang H, Fu M-Z, Yi J. Evolutionary and functional implications of 3′ untranslated region length of mRNAs by comprehensive investigation among four taxonomically diverse metazoan species. *Genes Genomics.* 2019;41(7):747–755. <https://doi.org/10.1007/s13258-019-00808-8>.
- Wiemers M, Chazot N, Wheat CW, Schweiger O, Wahlberg N. A complete time-calibrated multi-gene phylogeny of the European butterflies. *Zookeys.* 2020;938:97–124. <https://doi.org/10.3897/zookeys.938.50878>.
- Zhan S, Merlin C, Boore JL, Reppert SM. The monarch butterfly genome yields insights into long-distance migration. *Cell.* 2011;147(5):1171–1185. <https://doi.org/10.1016/j.cell.2011.09.052>.
- Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics.* 2023;39(1):btac808. <https://doi.org/10.1093/bioinformatics/btac808>.

Associate editor: Christopher Wheat