

RESEARCH

Open Access



# A high-quality draft genome assembly of the Neotropical butterfly, *Batesia hypochlora* (Nymphalidae: Biblidinae)

Nhat Tan Pham<sup>1,2</sup> , Anne Duploux<sup>3,4</sup> , Joseph See<sup>5,6</sup> , Lucy S. Knowles<sup>7</sup>, Edgar Marquina<sup>8</sup> , Geoffrey Gallice<sup>5,9</sup> , Freerk Molleman<sup>1</sup> and Vicencio Oostra<sup>10\*</sup>

## Abstract

We report a long-read high-coverage reference genome assembly of the Neotropical butterfly, *Batesia hypochlora* (Nymphalidae: Biblidinae). This represents the first reference genome in the Biblidinae subfamily, a clade subject to ongoing studies on seasonal and climate adaptation in the Amazon. We assembled the genome from PacBio HiFi long reads (66X coverage), polished it with Illumina short reads (15X coverage), and annotated it using PacBio IsoSeq RNA data. We observed 15 chromosome-sized scaffolds, varying in length from 13.2 Mbp to 37.6 Mbp (median, 24.3 Mbp), which combined to form a total genome size of 395.788 Mbp. This assembly is highly contiguous (contig N50 of 25.14 Mbp) and complete (BUSCO completeness score of 98.6% and 0.2% duplication rate). Repeat annotation revealed that the genome comprises approximately one-third transposable elements. Gene prediction using RNA-seq evidence identified 19,395 genes, of which 17,400 were assigned to 2,883 orthogroups, including genomes of the fruit fly, silk moth, and three other Nymphalid butterfly species. The high sequencing depth also allowed us to assemble the genomes of the mitochondria and the common endosymbiotic bacterium *Wolbachia*. The mitochondrial genome was fully assembled (15,540 bp in size) with all expected genes annotated. The *Wolbachia* genome was fragmented, and we determined that it belongs to the B-super group. The high-quality assembly of *B. hypochlora* can represent the subfamily in further comparative analysis of evolution and provide a key resource for ongoing work to explore reproductive biology and adaptations to seasonality in Neotropical butterflies.

**Keywords** Reference genome, Assembly, Amazonia, Symbiont, Mitochondria, Long reads, Biblidinae

\*Correspondence:

Vicencio Oostra  
v.oostra@qmul.ac.uk

<sup>1</sup>Department of Systematic Zoology, Faculty of Biology, Institute of Environmental Biology, Adam Mickiewicz University in Poznan, Poznan, Poland

<sup>2</sup>Forest Inventory and Planning Institute, Vietnam Forest Museum, Hanoi, Vietnam

<sup>3</sup>Organismal and Evolutionary Biology Research Program, Faculty of Biological and Environmental Sciences, The University of Helsinki, PO Box 65, Viikinkaari 1, Helsinki FI00014, Finland

<sup>4</sup>Institute of Life Science, HiLIFE, The University of Helsinki, PO Box 65, Viikinkaari 3, Helsinki FI00014, Finland

<sup>5</sup>Alliance for a Sustainable Amazon, Potomac, MD 20854, USA

<sup>6</sup>Amazon Research and Conservation Collaborative, Ithaca, NY 14850, USA

<sup>7</sup>NERC Environmental Omics Facility, NEOF Visitor Facility, School of Biosciences, Alfred Denny Building, University of Sheffield, Western Bank, Sheffield S10 2TN, UK

<sup>8</sup>Crees Foundation for Manu, Fundo Mascoitania S/N, Manu, Madre de Dios, Cuzco, Peru

<sup>9</sup>Department of Engineering, Pontifical Catholic University of Peru, Lima, Peru

<sup>10</sup>School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## Background

*De novo* genome assemblies are a crucial resource for studies on genetic adaptation, evolution, and co-evolution. They provide a precise genome map [1], for which annotations identify the position of coding genes and other DNA features in the genome, and functional analysis identifies the putative function of these genes [2]. *De novo* genome assemblies also provide the backbone of population genomics studies, phylogeography, and demography [3, 4]. Through comparative genomic analyses, they provide deep insights into the macroevolution of organisms. Moreover, new research directions have emerged with the recent shift from fragmented short-read assemblies toward highly contiguous chromosome-scale assemblies based on long reads. This includes quantifying the role of recombination and structural variation for adaptation, the importance of alternative splicing in environmental responses, the identity and nature of cis-regulatory elements underlying ecological adaptations, and comparative phylogenetic analyses, including gene activities in regulatory evolution [5]. The genomes of symbionts can be sequenced, assembled, and analyzed along with their hosts. Recently, sequences from such whole genome projects have also been used for metagenomic analyses and identifying non-target organisms, such as endosymbionts. This brings new insight into the co-evolution between hosts and a wide diversity of parasitic or mutualistic microscopic organisms. In insects, for example, symbiotic bacteria like *Wolbachia* can play an important role in their hosts [6] and ecology [7]. When not discarded or ignored, these non-target assemblies can further enrich our understanding of both vertical and horizontal transmission of *Wolbachia* [6, 8, 9].

Due to their vast diversity, conservation genomics studies of insects are facing significant challenges, including the scarcity of complete reference genomes. To date, the genomes of 2,596 insect species have been assembled up to the scaffold level and are publicly available in the NCBI database (January 9, 2025). Although Lepidoptera are widely used as models in evolutionary and ecological studies or as biodiversity indicators, only 963 (0.6%) Lepidoptera species have been assembled to the scaffold level, including just 304 (1.8%) of an estimated 20,000 butterfly species globally [10]. Thus, most clades remain unrepresented or underrepresented [11]. Of the available butterfly genome assemblies, most originate from species in North America or Europe. For instance, although the Neotropics are the world's most important center of butterfly diversity, with at least 7,000 species [12], fewer than 100 butterflies from this region have had their genomes assembled. Therefore, high-quality reference genomes are missing for tropical biodiversity hotspots, which limits further studies across various fields.

One of the clades that lacks a genome assembly is the subfamily Biblidinae. This clade encompasses more than 300 species that arose about 37 Mya and are distributed in South and Central America, with some groups in the Old World tropics (Kawahara et al., 2023). These butterflies are colorful and can be of local economic importance as crop pests [13–16]. Furthermore, several species are commercially traded to be displayed in live butterfly exhibits or mounted forms [17]. They have also been the object of studies on wing pattern evolution [18], sound production [19], and life history [20]. More recently, our lab has begun using multiple species in this subfamily as part of a research program on life history and behavior, focusing on adaptation to seasonality and climate change. Therefore, a reference genome assembly for this clade is very timely.

We chose one of the most iconic Amazonian butterfly species to represent Biblidinae, *Batesia hypochlora* C. Felder & R. Felder, 1862 (NCBI ID: 127305) is distributed across the western lowland Amazonian rainforest, from central Colombia through southeastern Peru and western Brazil [21, 22]. It is associated with both intact forest interior and disturbed or edge habitats (Ramos-Artunduaga et al., 2021; GG pers. obs.). The butterfly is likely aposematic as it has a highly conspicuous color pattern and slow flight [21]. Moreover, the caterpillars feed on *Caryodendron orinocense*, an evergreen tree belonging to the milkweed (Euphorbiaceae) family (known for its toxic chemicals that, e.g., aposematic monarch butterflies sequester). Adults feed on the juices of rotting fruits [21]. However, all other aspects of the species' biology remain unexplored throughout its range, including reproductive biology, interactions with parasites, and any other interspecific interactions beyond the single known host species.

Here, we provide a reference genome for *B. hypochlora* and its *Wolbachia* endosymbiont. We combine long and short DNA reads to provide a high-quality *de novo* reference genome for *B. hypochlora* and use RNA evidence to complete gene annotations. Our assembly includes chromosome-scale nuclear scaffolds with gene and repeat element annotation, as well as functional annotation. We also provide a complete and annotated mitochondrial genome as well as a draft assembly of the associated bacterial symbiont *Wolbachia*. Finally, we compared the newly assembled genome with the high-quality reference genomes of the subfamily Nymphalinae (Nymphalidae), including *Aglaia io* and *Melitaea cinxia*.

## Methods

### Material sampling, library preparation, and sequencing

The specimens used in this study were collected at the Manu Learning Centre, a field station located along the Upper Madre de Dios River at an elevation of ca. 450 m

a.s.l., where the lower foothills of the Andes meet the lowland of the Amazon basin in southeastern Peru (Madre de Dios region; 12°47'18.9"S 71°23'29.1"W). We collected two adults of *B. hypochlora* using fruit- and fish-baited traps suspended at various heights along trails in the primary forest at the site in November 2021. We sacrificed the first specimen (PE-2021-005-R, sex unknown) and stored it in RNAlater; the second individual (PE-2021-004-E, male) was stored in 100% ethanol. We kept both samples at approximately 4 °C, then transported them at ambient temperature and stored them at -20 °C upon arrival in the laboratory.

We used the thorax of the first individual (PE-2021-005-R) for both PacBio DNA (HiFi) and RNA (Iso-seq) sequencing (at the NERC Environmental Omics Facilities (NEOF). We extracted High Molecular Weight (HMW) DNA using the Nucleobond HMW kit (Machery-Nagel), with the following modifications: lysis buffer, proteinase K, RNase A, and doubled binding buffer volumes compared to the kit's protocol. We quantified the HMW DNA using a Qubit fluorometer and measured purity using a Nanodrop (both Thermo Fisher). To assess the integrity, we used Femto Pulse (Agilent), confirming that most of the DNA was > 50 kbp in length. RNA was extracted using the QIAGEN RNeasy Mini Kit. The tissue was initially homogenized using a TissueLyser II (QIAGEN), and the rest of the process was completed according to the kit instructions. Then, HiFi and Iso-Seq libraries were prepared and sequenced using 2 SMRT cells (HiFi library) and 1 SMRT cell (Iso-Seq library) at the NEOF Centre for Genomic Research. Combining both SMRT cell data, we obtained 2,337,639 HiFi reads with a median length of 11.4 kbp, and 78% of the reads had a Q30 or higher quality. For the Isoseq reads, we obtained 2,336,880 reads with a median length of 1,882 bp (Table S1). We used the second individual (PE-2021-004-E) for Illumina short-read DNA sequencing. Half of one thorax was removed from ethanol, dried overnight, and then homogenised using a Tissue Lyser in lysis buffer, and then DNA was isolated using a Qiagen DNeasy Blood & Tissue kit following the manufacturer's recommendations, including 3-hour Proteinase K incubation. Short insert size (ca. 350 bp) Illumina sequencing libraries were prepared (NEBNext Ultra II FS Kit with ½ volume reactions). The sample was sequenced with other samples on a Novaseq S4 lane (150 bp PE, yielding 63.5M raw reads (30.1M reads after filtering) at the NEOF Centre for Genomic Research.

#### Quality control and pre-assembly estimates

To check the quality for both HiFi long reads and Illumina short reads, we used FastQC version 0.11.9 [23]. According to the results from the given FastQC, only paired-end Illumina short reads have adapters, so we

used trimmomatic version 0.36 (LEADING: 3 TRAILING: 3 SLIDINGWINDOW: 4: 20 MINLEN: 36) to remove adapters and low-quality data [24, 25], and then the trimmed reads were rechecked. We performed a k-mer analysis on HiFi reads to estimate genome size, heterozygosity, repetitiveness, and sequencing coverage. We used Jellyfish version 2.2.10 [26] to calculate 31-mer normalized coverage, then visualized and estimated parameters using Genomescope v2 (ploidy  $p=2$  and kmer=31; at <http://qb.cshl.edu/genomescope/genomescope2.0/>) [27].

#### Nuclear genome assembly

We used HiFi long reads for initial assembly using hifiasm version 0.19.5-r587 with default parameters [28], yielding genome version 0.1. We then polished genome version 0.1 with both PacBio long reads and Illumina short reads. First, we aligned the PacBio long reads using minimap2 [29] and trimmed Illumina short reads using bwa [30] to the assembly version (0.1). We then used these alignments to polish the version 0.1 using Pilon v1.24 (--fix gaps, local, breaks) [31], resulting in version (0.2). We identified and removed xenobiotic contamination (*Wolbachia*, see Results) using Blobtoolkit version 4.2.1 [32], resulting in assembly version 0.3.

We calculated contiguity statistics and completeness for each assembly version. The module stats of bbtool v39.01 generated basic statistics, including scaffold count, N50, L50, and gap percent [33], while completeness was calculated using BUSCO version 5.5 [34] against the lepidoptera\_odb10 database.

#### Mitochondrial genome assembly

We assembled the mitochondrial genome using MitoHiFi version 3.2 from assembly version 0.3 [35]. To choose a reference mitochondrial genome for *B. hypochlora* from a close relative, we used the tool findMitoReference.py, which identified *Hamadryas epinome* (NC\_025551.1) (Nymphalidae: Biblidinae) as the optimal reference [36]. We finally ran mitohifi.py for mitochondrial annotation and reported the mitochondrial genome separately from the final nuclear genome. The remaining nuclear genome (excluding the mitochondrial genome) was named assembly version 0.4.

#### Repeat, Gene, and functional annotations

We identified repeats *de novo* in assembly version 0.4 by using the database of RepBaseRepeatMaskerEdition-20181026 with RepeatModeler version 2.0.5 [37–39]. We then split the library into known (successfully classified) elements and unknown elements (remaining unclassified or unknown) using seqkit [40]. Then, we used RepeatMasker version 4.1.5 to mask repetitive elements from the Repbase and Insecta repeat libraries

using repclassifier version 1.1 [39, 41]. Unique transposable element (TE) families were grouped into eight different TE classes, including “DNA Transposons”, “Helitrons”, “LINEs” (Long interspersed nuclear elements), “LTR Retrotransposons” (Long terminal repeats), “Low Complexity”, “SINEs” (Short interspersed nuclear elements), “Simple Repeat”, and “Unknown”. We present the soft-masked genome as the final assembly (version 1.0).

We combined the soft-masked genome version 1.0 with RNA Isoseq data to predict gene models. We followed the Iso-Seq workflow to generate both high-quality (predicted accuracy  $\geq 0.99$ ) and low-quality (predicted accuracy  $< 0.99$ ) reads. We then removed primers and barcodes from the raw Iso-Seq reads using lima and refined (by removing polyA tails) and clustered them [42]. Then, we aligned only high-quality RNA reads to genome version 1.0 using minimap2 [29]. We then used BRAKER version 3.08 to predict proteins using RNA alignment bam file, soft-masked genome assembly (version 1.0), and training with the protein sequence of Arthropoda from OrthoDB version 11 [43–45].

We used two methods to annotate the function of the protein-coding genes. The first method involved identifying *B. hypochlora* orthogroups in published reference genomes of 5 other insects (*Drosophila melanogaster* - GCA\_000001215.4, *Bombyx mori* - GCA\_014905235.2, *Danaus plexippus* - GCA\_000235995.2, *Heliconius melpomene* - GCA\_000313835.2, *Melitaeta cinxia* - GCA\_905220565.1) using Orthofinder [46]. Subsequently, we downloaded the UniProtKB database for these five insects. This yielded protein-to-GO mappings for all *B. hypochlora* proteins with an annotated ortholog in at least 1 species. The second method was to identify Gene Ontology (GO) in the genome of *Batesia* (version 1.0) by mapping the gene sequence to the precompiled database for insects. We downloaded the database for “Insecta” taxa on eggNOG DB version 5.0.2 using create\_dbs.py [47] and annotated the *Batesia* protein sequences using eggNOG-mapper v2 [48]. Then, we used an online version of GOMapper (https://go.princeton.edu/cgi-bin/GOMapper) to map the unique GO terms to GO Slim based on the Ontology aspects [49, 50]. Finally, we integrated the results of both methods to obtain the gene function.

#### Microbial symbiont detection and assembly

We used the Blobtoolkit [32] to analyze possible non-host genomic material, including sequences from common endosymbiotic bacteria, such as *Wolbachia* and *Spiroplasma* [51]. We identified 27 *Wolbachia* contigs from the host nuclear and mitochondrial assemblies (i.e. version 0.3, see Results). To further characterize these at the strain level, we isolated and screened them for the presence of the *wsp* gene and the five Multi Locus

Strain Typing (MLST) genes using the blast function in Geneious Prime® 2025.0.3 (https://www.geneious.com). These six loci are commonly used to phylogenetically assign *Wolbachia* strains to their respective taxonomic supergroup [52]. The five MLST genes (i.e. *ftsZ*, *fbpA*, *coxA*, *gatB*, and *hcpA*) were all identified and compared to orthologous genes from reference genomes of different *Wolbachia* supergroups in gene-specific phylogenetic analyses following the protocol described in [53]. In brief, a reference set for each gene was obtained from GenBank with representative strains of the *Wolbachia* A-, B-, F-, and D-supergroups [53]. Individual gene alignments were produced using the pairwise alignment with the default options in Geneious Prime. Alignments were manually screened to check and correct any errors. Phylogenetic reconstructions and tree visualizations of *Wolbachia* supergroups were carried out for each locus independently using the Tree function with default options in Geneious Prime.

We similarly screened for the presence of the *wmk* and *Oscar* genes [54, 55] and of both the *cifA* and *cifB* genes [56, 57] in our *Wolbachia* contigs, using the BLAST function with default settings in Geneious Prime. We used the WolWO-mediated killing-like protein (*wmk*) gene sequence from the *Wolbachia* strain wCauB (#MK955149.1), the amino acid sequence and domain structure of the Oscar protein given by Katsuma et al. (2020), and the cytoplasmic incompatibility factor A and B protein genes (#MG807657, #MG807658, OP947615, and #MH544806) from *Wolbachia* strain wPip. The detection of the *wmk* and *Oscar* genes, the candidate genes for the expression of the male-killing phenotype in *Wolbachia* [54, 55] could suggest that the *Wolbachia* strain, which infected *B. hypochlora* (labeled: wBhyp), induces the death of the male progeny of its host [54, 58]. Similarly, the detection of the *cifA* and *cifB* genes, which code for the expression of cytoplasmic incompatibility (CI) between individuals of incompatible infection status [56, 57], could suggest that wBhyp can induce CI in *B. hypochlora*.

#### Large-scale genome rearrangement analysis

The soft-masked genome of *B. hypochlora* (version 1.0) was compared to a published high-quality reference genome to identify any large-scale genome rearrangements. As no other genome of the Biblidinae subfamily is available, we used more distantly related species with a published chromosome-level genome assembly in the subfamily Nymphalinae. We selected *Aglais io* (Nymphalini, Nymphalinae, NCBI ID: 171585) and *Melitaeta cinxia* (Melitaeini, Nymphalinae, NCBI ID: 113334). These species have a divergence time of ca. 57 Mya, and for both, the sex chromosomes have been identified [59, 60]. We downloaded the reference genomes from



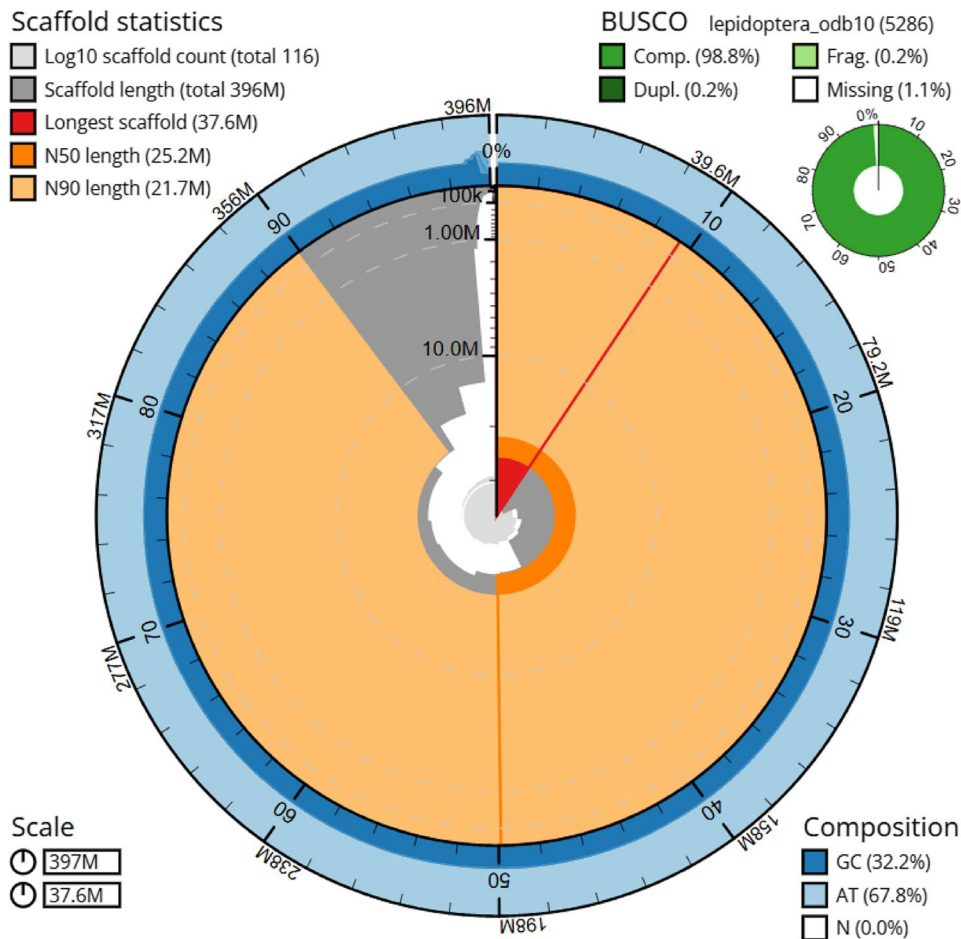
the NCBI database and only retained (for both query and references) scaffolds or chromosomes longer than 1 Mbp. We used nucmer in MUMmer version 4.0.0rc1 for alignment with a set of a minimum length of cluster of matches (c) is 100, and a minimum length of a single exact match (l) is 500bp [61]. The resulting nucmer alignment was visualized with the “circlize” or “ggplot2” packages in R 4.2.2.

Results

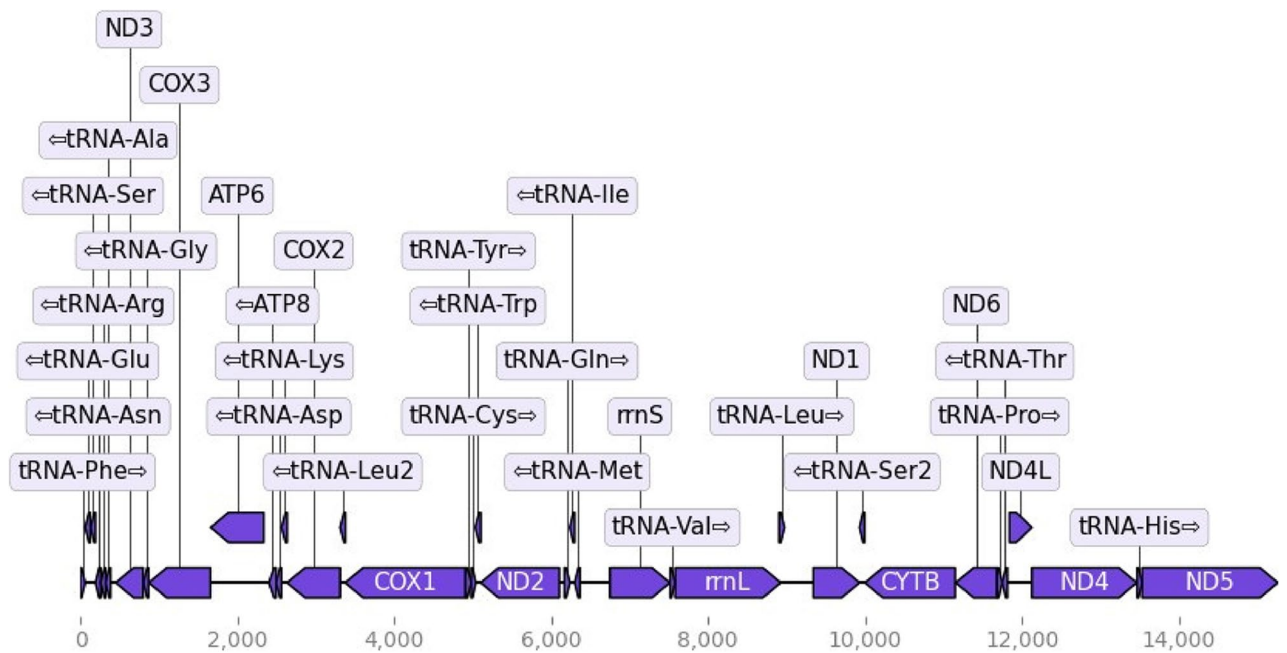
Nuclear and mitochondrial genome assemblies

Based on the k-mer analysis of the PacBio long reads with k=31 and 0.123% sequencing error, the nuclear genome of *B. hypochlora* has an estimated haploid length of 370 Mbp and a low heterozygosity of 0.43% (Fig. S2). The total genome size in assembly version 0.1 is 397.812

Mbp, comprising 143 contigs, with the longest contig being 37.56 Mbp. The contig N50 value for the genome assembly is 25.29 Mbp (Table S3). During the polishing process, we corrected the position of 863,381 bp in 321 locations, including the correction of breaks (deletion of 378,316 bp and insertion of 71,092 bp) and opening gaps (deletion of 456,931 bp and insertion of 73,739 bp). This resulted in an assembly size of 397.124 Mbp (version 0.2). The xenobiotic analysis identified 27 non-Lepidoptera scaffolds representing 1.34 Mbp of the genome. These were all version 0.2 assigned to Pseudomonadota bacteria (Fig. S3), which were later found to be *Wolbachia* (see details below). After removing the *Wolbachia* contigs, the final genome size of *B. hypochlora* is 395.788 Mbp, with 99.19% of the main genome in 16 scaffolds >= 1 Mbp (genome version 1.0, Fig. 1).



**Fig. 1** Snail plot summary of assembly statistics for genome assembly version 0.3 (after removing *Wolbachia*, Table S3), generated on the Galaxy Server. The main plot is divided into 1,000 size-ordered bins around the circumference, with each bin representing 0.1% of the 397,123,821 bp assembly. The distribution of sequence lengths is shown in dark grey with the plot radius scaled to the longest sequence present in the assembly (37,552,201 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 sequence lengths (25,221,893 and 21,698,544 bp), respectively. The pale grey spiral represents the cumulative sequence count on a log scale, with white scale lines indicating successive orders of magnitude. The blue and pale blue areas around the outside of the plot show 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. The assembly has been filtered to exclude sequences with family matches Anaplasmataceae (genus matches *Wolbachia*)



**Fig. 2** Annotation of the mitochondrial genome of *Batesia hypochlora*. The horizontal axis shows the position across the mitochondrial genome (total length 15,521 bp), with the location of 13 coding genes, 22 tRNAs, and 02 rRNAs (rrnS & rrnL) indicated in purple and the transcription direction of each gene indicated with an arrow. Plot generated using MitoHiFi [35]

**Table 1** Genome assembly statistics. Evaluation of statistics based on genome version 1.0, and completeness based on genome version 0.4

Assembly statistics	
Size (Mbp)	395.788
Scaffold N50 (Mbp)	25.148
Largest scaffold (Mbp)	37.552
Number of scaffolds > 2.5 Mbp	15
BUSCOs completeness statistics	
Complete and single copy	5211 (98.6%)
Complete and duplicated	11 (0.2%)
Fragmented (F)	9 (0.2%)
Missing (M)	55 (1%)

The complete mitochondrial genome of *B. hypochlora* consists of 15,540 bp with an order of 37 genes, including 13 protein-coding genes, 22 transfer RNAs, and 2 rRNAs (Fig. 2). The majority strand (J-strand) has 23 genes (9 PCGs and 14 tRNAs), while the minority strand has 14 genes (4 PCGs, 8 tRNAs, and 2 rRNAs) (Table S5). We separated the mitochondrial genome from the nuclear genome in version 0.3, thereby maintaining a total genome size of 395.788 Mbp in genome version 0.4 (Table S3). The assessment of completeness using BUSCO, from genome version 0.1 to genome version 0.4, yielded a completeness rate of more than 98.7%, with a duplication rate of 0.2% (Table S3). After repeat masking (see details below), we got a soft-masked genome as the final genome assembly. The assembly version 1.0 size is 395.788 Mbp (Table 1), including 391.297 Mbp, which

corresponds to 15 chromosome-sized scaffolds (>2.5 Mbp). This assembly was highly contiguous, with an N90 of 21.698 Mbp (Table S3).

Microbial symbiont detection and Wolbachia assembly

Blast analysis identified 27 contigs as *Wolbachia*, with the largest ca. 100 kbp long (average 49.47 kbp), and a combined total length of 1.34 Mbp. Sequencing depth at *Wolbachia* contigs was 5.7% of host genome depth for individual PE-2021-004-E (Illumina short reads), while depth at *Wolbachia* contigs for individual PE-2021-005-R (PacBio long reads) was 0.08% of host genome depth. Through phylogenetic analyses of each MLST sequence from the *wBhyp* *Wolbachia* contigs (Fig. S4), we were able to confidently place *wBhyp* in the *Wolbachia* B-supergroup (>95% similarity with *wPip* (B-), ~ 85% similarity with *wMel* (A-), and *wBm* (D-supergroup)). We were unable to retrieve the *wsp* gene from our partial *wBhyp* assembly. Similarly, we did not find any evidence of the presence of the *cifA* and *cifB* genes [56] nor of the *Oscar* gene [55] in the *Wolbachia* contigs. However, we were able to retrieve a putative *wmk* gene from the *wBhyp* assembly, suggesting that the strain might be able to induce male-killing in its butterfly host, *B. hypochlora*.

Repeat annotation, gene model prediction, and functional annotation

We identified 1,135 different types of repeats, classified as 1,042 known and 93 unknown elements, after two rounds of running the repclassifier module, and TEs accounted

for 34% of the nuclear genome content in the final *B. hypochlora* nuclear assembly (soft-masked genome, version 1.0). The most abundant repeat class across all TE categories was LINES (47.85 Mbp,  $n = 241,677$ ), followed by Simple Repeat (9.81 Mbp,  $n = 203,837$ ) (Fig. 3). DNA transposons account for 2.8% of the total *B. hypochlora* genome size (11.16 Mbp,  $n = 34,047$ ) (Table S6).

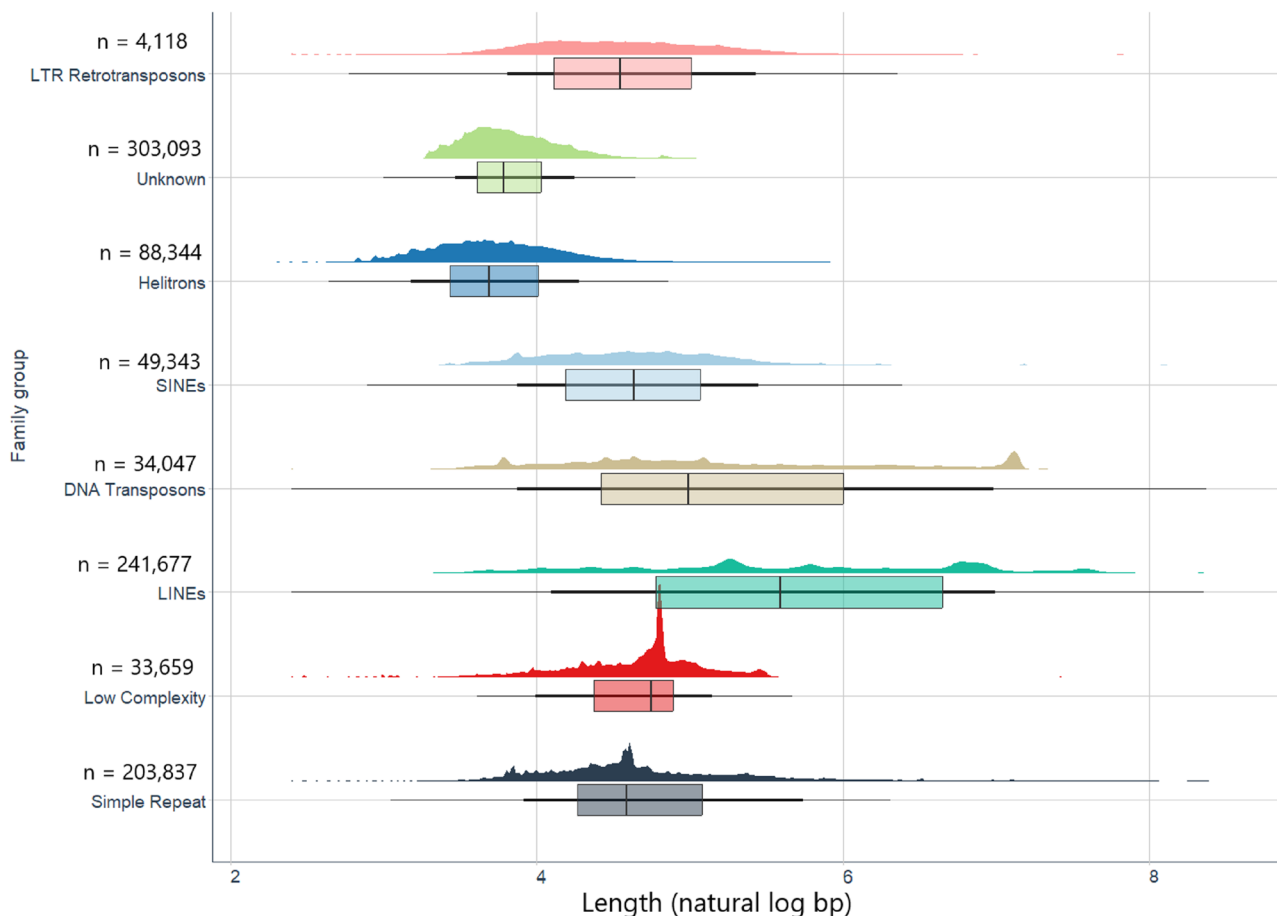
Incorporating RNAseq evidence from the thorax, we identified 21,588 mRNAs of median length 1624.5 bp (min = 6 bp, max = 354,511 bp). We constructed 19,395 gene models (mean length = 4,423.8 bp, median length = 1,461 bp) from the masked genome (version 1.0). This also consisted of 74,395 introns of median length 379 bp (min = 31 bp, max = 211,373 bp) and 95,983 exons of median length 163 bp (min = 1 bp, max = 119,64 bp). The average number of introns and exons per gene was 3.8 and 4.9, respectively.

The eggNOG-mapper detected 18,460 genes with an average of 63.31 GO terms per gene, including 8,560 genes containing at least one GO term (mean = 136.5 GO terms, median = 101 GO terms). We mapped 5,646 unique GO terms to GO Slim in three ontology aspects

(Fig. S5). Finally, we also used the protein sequence for orthology analysis with five different species (*Drosophila melanogaster*, *Bombyx mori*, *Danaus plexippus*, *Heliconius melpomene*, *Melitaea cinxia*), and OrthoFinder assigned a total of 19,393 genes (81.8% of all genes in 6 species) to 2,957 orthogroups, with a mean orthogroup size of 6.6. In *Batesia hypochlora* protein sequences, 17,400 genes (80.6%) were assigned to 2,883 orthogroups. However, our reference assembly contained no match for the *Heliconius charithonia* W-linked gene UVRh1 [62], further confirming that the first sample PE-2021-005-R was a male specimen.

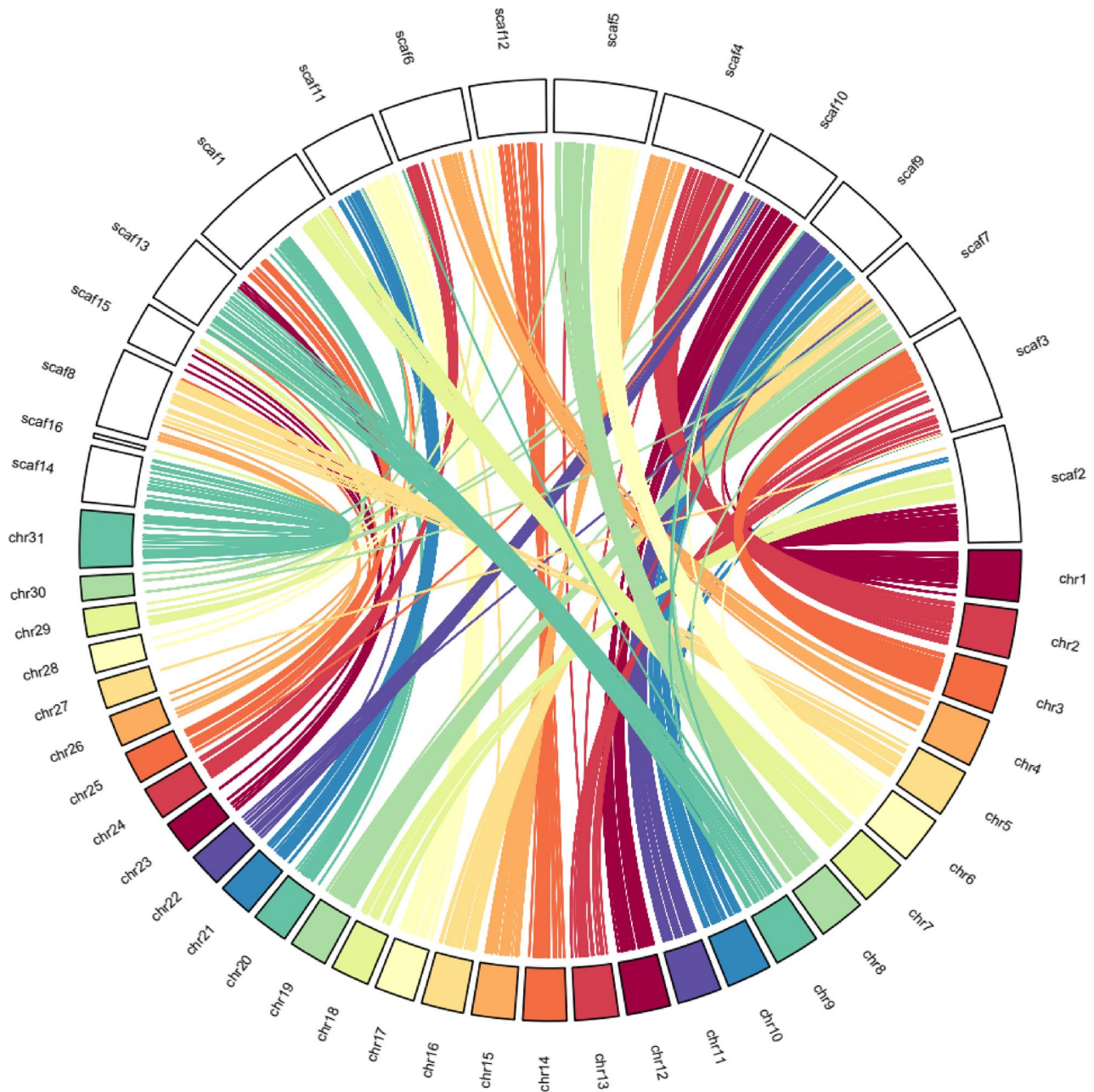
### Large-scale genome rearrangements

Examining chromosomal rearrangements between our *B. hypochlora* genome assembly and (*A. io* genome, which shares a last common ancestor ca. 57 Mya [59], shows strong syntenic relationships across scaffolds (Fig. 4). We identified 17,496 alignments between the (*B. hypochlora* on 17 scaffolds and all chromosomes of *A. io*. We detected large-scale rearrangements in all 15 chromosome-sized scaffolds (mean length = 610 bp, ca. 2.63% genome size)



**Fig. 3** Length distributions of transposable elements in eight TE superfamilies. The horizontal axis is the length (natural logarithm of bp), with the vertical axis showing the frequency. Eight superfamilies are shown from top to bottom in different colors, annotated with their number of elements





**Fig. 4** Large-scale rearrangement and genome evolution compared with the genome of *Aglais io*. Circos plot shows results from the whole-genome alignment using nucmer for contigs > 1 Mbp, with the minimum length of a cluster of matches (c) being 100 bp, and the percentage of identity higher than 80%. Chromosome 31 is the sex chromosome in the *A.io* genome

of the *B. hypochlora* genome, except for chromosome 27 (8.7 Mbp) in the *A.io* genome, which showed no alignment. In general, many scaffolds in the *B. hypochlora* genome appear as fusions of chromosomes compared to *(A) io*. For instance, scaffold 1 in the *(B) hypochlora* genome appears to have resulted from a fusion of chromosome 7 with an inversion of major portions of chromosomes 20 and 26 of the *(A) io* genome. Similarly, scaffold 3 in *(B) hypochlora* appears to have resulted from a fusion of chromosome 3 with an inversion of the entire

chromosome 13 from the *(A) io* genome (Fig. S6). We found that scaffold 14 of the *(B) hypochlora* genome was mapped to chromosome Z from the *(A) io* genome. We also observed a similar pattern of chromosome fusion when comparing *(B) hypochlora* and *M. cinxia*, with 14,894 alignments across 15 chromosome-sized regions (mean length = 605 bp, ca. 2.12% of the *B. hypochlora* genome size). The supposed sex chromosome was also mapped to the sex chromosome in *M. cinxia* (Fig. S6 & S7). Moreover, when we mapped the Illumina short reads



of our male specimen (PE-2021-004-E) to the reference genome, the coverage against scaffold 14 (the putative *B. hypochlora* sex chromosome) was identical to that of the autosomes (Table S7 & S8), consistent with the reference assembly being from a ZZ individual. Therefore, we tentatively conclude that scaffold 14 represents chromosome Z of *B. hypochlora*.

## Discussion

We present a high-quality genome assembly for the Neotropical butterfly, *Batesia hypochlora*, the first reference genome in the subfamily Biblidinae. We identified 15 chromosome-sized scaffolds, which is less than half of the reconstructed ancestral chromosome number of 31 for the Biblidinae subfamily and the Nymphalidae as a whole based on cytological counts [63–65]. We initially considered a technical error, suspecting that our genome assembly may have accidentally fused (fragments from) different chromosomes onto the same scaffold. Our analysis of coverage variation of PacBio HiFi and Illumina short-fragment paired-end reads failed to identify putative incorrect fusion points. Therefore, we conclude that the putative chromosome number of 15 in *Batesia hypochlora* is not a methodological artifact, and instead reflects the remarkable biological variation in chromosome number observed across the subfamily Biblidinae, as previously described using karyotype analysis [63]. For example, within the same tribe (Ageroniini), representatives of other genera, such as *Panacea* (the most closely related genus) and *Hamadryas*, have the ancestral chromosome number ( $n = 31$ ), but *Ectima* also has about half ( $n = 16$ ). Still, several other tribes show considerable variation, including Epiphilini, which include species with  $n = 7, 10–14, 27–34$ , and 54, Biblidini ( $n = 15$  and 28–33), and Epicaliini ( $n = 7–8, 11, 14–16, 21–31$ ). There is even substantial variation within genera and even within species (*Eunica malvina*,  $n = 14$  and 31). In contrast, two tribes (Callicorini and Eubagini) exclusively have genera with the ancestral or close to ancestral chromosome number ( $n = 28–31$ ). Thus, overall chromosomal fusions (and to a lesser extent, fissions) are common in Biblidinae, consistent with our finding of  $n = 15$  for *B. hypochlora*. The driving forces of chromosome number diversity in Biblidinae are unknown. In other Neotropical Nymphalidae taxa, mimicry, host plant diversification, hybridization, and geographic isolation are key evolutionary forces that may play important roles in chromosome number evolution [66–68]. For instance, in *Agrodiaetus* (Lycaenidae), chromosomal rearrangements have been implicated as key drivers of speciation [67]. Biblidinae wing patterns are colourful and highly diverse, and likely include aposematic, cryptic, and mimetic colouration patterns [18, 63]. Biblidinae, therefore, provides an excellent taxon to test the link between chromosome number variation and

wing pattern diversity. However, we lacked Hi-C data to verify the number of chromosomes in the final assembly version, so our estimated chromosome number should be interpreted cautiously.

The rate of chromosomal rearrangements over evolutionary time varies among different groups of organisms [69], with butterflies exhibiting a particularly high rate [70]. While we have information about chromosome numbers in some Biblidinae, we lack another genome in this subfamily to gain insight into which parts were rearranged in what fashion. Therefore, we compared the *B. hypochlora* genome with those of two species from Nymphalinae (divergence time 57 Mya), (*A. io* and *M. cinxia*). Over such evolutionary timescales, chromosomal rearrangements can accumulate, leading to reduced synteny or more fragmented collinearity signals. Interestingly, we observed many large-scale rearrangements in the (*B. hypochlora* genome, with the joining and inversion of multiple chromosomes. In Lepidoptera, the frequency of chromosomal rearrangements is often higher between closely related species, suggesting that such rearrangements may play a role in the early stages of speciation. However, chromosomal rearrangement could also occur after speciation [65, 71]. Some rearrangements can contribute to reproductive isolation [67] by causing barriers to gene flow [72]. Chromosomal rearrangements with fusion and fission are common in Nymphalidae, including both cladogenesis and anagenetic events. The anagenetic event rates are higher than the cladogenesis event rates in Nymphalidae, except in Ithomiine butterflies [71]. Hence, further tests are needed to evaluate the chromosomal rearrangement rate in relation to fusion or fission events during *B. hypochlora* speciation.

With 34%, the observed TE content in *B. hypochlora* fell within the range of variation in Nymphalid butterflies (7% in *Danaus plexippus* to 55% in *Melanargia galathea* [65]) and is similar to *M. cinxia* with 42% [73]. There is strong evidence that larger genomes tend to have a higher TE content [74]. Moreover, the percentage of TE in the *Batesia* genome was close to the mean TE content across published Nymphalinae genomes (34.88%), which diverged 57 Mya, and slightly lower than the average TE content (41.63%) across Satyrinae, Heliconiinae, and Limenitidinae, which diverged 63 Mya [65]. Previous studies have shown that certain types of TEs, such as LINEs and LTRs, are associated with fusion-prone regions [75]. Additionally, shorter chromosomes, which are more frequently involved in fusion events, tend to have a high frequency of TEs [64]. Further studies are needed in taxa that have a wide range of chromosome numbers, such as Biblidinae, to demonstrate the relationship between TE content and speciation rate or divergence time.

We obtained 1.34 Mbp of the *Wolbachia* genome during *B. hypochlora* genome assembly using male specimens. This is similar to the median of *Wolbachia* genome size (ca.1.3 Mbp) in Lepidoptera [55]. The low depth of read mapping observed suggests a low *Wolbachia* titer, which might be expected from thorax tissue and male hosts. As both sequencing approaches used here produced reads for the *Wolbachia* assembly, we can confidently say that the symbiont infected both *B. hypochlora* specimens we sequenced. Butterflies are well-known hosts of *Wolbachia* (Duploux and Hornett 2018), providing many textbook examples of the role of *Wolbachia* in the ecology and evolutionary histories of these insects [76–78]. Our partial assembly of the wBhyp strain provides evidence that this particular *Wolbachia* strain sits in the *Wolbachia* B-super group, which has also been often found in Lepidoptera [6, 79]. Finally, we successfully isolated the *wmk* gene, a putative gene for the expression of male killing in *Wolbachia* [54]. Although we were unable to retrieve the *wsp* gene, CI-coding (*cifA* and *cifB*), and the *Oscar* gene in our partial assembly of wBhyp, this could be attributed to the low coverage when mapping Illumina short reads to assembly version 0.2. To our knowledge, there is no record of female-only broods in this species, and sex-ratio distortion was not mentioned when the species was reared under laboratory conditions [21]. Here, the specimens used for sequencing were both infected with the same strains and were both males. This could suggest that the *Wolbachia*-induced male-killing phenotype is not expressed or potentially repressed in *B. hypochlora* [80], as was previously described in *Hypolimnas bolina* [81, 82]. The effect of this *Wolbachia* strain on the reproductive system of the butterfly host, and its ecology and evolutionary biology, thus, deserves to be further experimentally tested [80, 83–85].

## Conclusion

The assembly of the *Batesia hypochlora* genome assigned 391.297 Mbp (98.87% of the genome) into 15 chromosome-size scaffolds, with an N50 of 25.15 Mbp. The genome is highly complete, with 98.8% of BUSCO represented. We also present 15.54 kbp of mitochondrial assembly and 1.34 Mbp of *Wolbachia* assembly. With the analysis of TE content, gene annotation, and symbiont infection, this genome assembly is a valuable resource for Lepidoptera genomics, ecology, and evolution.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-025-12394-z>.

Supplementary Material 1.

Supplementary Material 2.

## Acknowledgements

We thank Jared Shorma, other field station staff, and volunteers for their help with butterfly collections at Manu Learning Centre, and Carl Yung for DNA isolation advice. We acknowledge NEOF staff and facilities at the University of Sheffield and the University of Liverpool (Centre for Genomic Research). We acknowledge the support of the Freiburg Galaxy Team for the Galaxy server (<https://usegalaxy.eu>). This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT. <https://doi.org/10.5281/zenodo.438045>. We acknowledge the assistance of the ITS Research team at Queen Mary University of London. The authors thank Peru's Servicio Nacional Forestal y de Fauna Silvestre (SERFOR) for permission to conduct field and laboratory research (permit no. D000443-2021-MIDAGRI-SERFOR-DGGSPFFS).

## Authors' contributions

J.S., E.M., G.G., and V.O. contributed to sample collection. L.S.K. and V.O. performed DNA and RNA extractions. T.P.N. conducted all main analyses supervised by V.O. and A.D. for symbiont analyses. T.P.N. wrote the manuscript, with guidance and contributions by V.O., E.M., and G.G. All authors read, revised, and approved the final version of the manuscript.

## Funding

This work was supported by the UK Natural Environment Research Council (NERC) Environmental Omics Facility, grant NEOF1388, to VO (HMW DNA and RNA isolation, PacBio library preparation and sequencing). VO was additionally supported by a UKRI Future Leaders Fellowship (MR/V024744/2), Queen Mary University of London, the University of Liverpool, and the British Ecological Society (grant SR20-1273, funding the fieldwork). NTP and FM were supported by Poland NCN 2021/43/B/NZ8/00966. AD was funded by the Research Council of Finland (grant #321543). GG acknowledges the support of Wild Green Future.

## Data availability

All scripts with the used commands are available on GitHub at [github.com/tanpham15/B\\_hypochlora](https://github.com/tanpham15/B_hypochlora). Raw HiFi, Isoseq, and Illumina reads were deposited at ENA (PRJEB87368). The final nuclear assembly (version 1.0), mitochondrial genome, and annotation are available at NCBI (PRJNA1240833). The genome assembly dataset, including the final genome, protein sequences, protein annotations, \*Wolbachia\* sequences, and scripts, is also available separately at Zenodo (<https://doi.org/10.5281/zenodo.17576225>).

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

Received: 16 May 2025 / Accepted: 28 November 2025

Published online: 06 December 2025

## References

1. Sohn J-i, Nam J-W. The present and future of *de Novo* whole-genome assembly. *Brief Bioinform*. 2016;19(1):23–40.
2. Ranz JM, González PM, Clifton BD, Nazario-Yepiz NO, Hernández-Cervantes PL, Palma-Martínez MJ, Valdivia DI, Jiménez-Kaufman A, Lu MM, Markow TA, et al. A *de Novo* transcriptional atlas in *Danaus plexippus* reveals variability in dosage compensation across tissues. *Commun Biol*. 2021;4(1):791.
3. Yan SM, Sherman RM, Taylor DJ, Nair DR, Bortvin AN, Schatz MC, McCoy RC. Local adaptation and archaic introgression shape global diversity at human structural variant loci. *eLife*. 2021;10:e67615.
4. Mérot C, Oomen RA, Tigano A, Wellenreuther M. A roadmap for Understanding the evolutionary significance of structural genomic variation. *Trends Ecol Evol*. 2020;35(7):561–72.

5. van Dijk EL, Naquin D, Gorrichon K, Jaszczyszyn Y, Ouazahrour R, Thermes C, Hernandez C. Genomics in the long-read sequencing era. *Trends Genet.* 2023;39(9):649–71.
6. Ahmed MZ, Breinholt JW, Kawahara AY. Evidence for common horizontal transmission of *Wolbachia* among butterflies and moths. *BMC Evol Biol.* 2016;16(1):118.
7. Russell JA, Goldman-Huertas B, Moreau CS, Baldo L, Stahlhut JK, Werren JH, Pierce NE. Specialization and geographic isolation among *Wolbachia* symbionts from ants and Lycaenid butterflies. *Evolution.* 2009;63(3):624–40.
8. Miyata M, Nomura M, Kageyama D. Rapid spread of a vertically transmitted symbiont induces drastic shifts in butterfly sex ratio. *Curr Biol.* 2024;34(10):R490–2.
9. Lohman DJ, Peggie D, Pierce NE, Meier R. Phylogeography and genetic diversity of a widespread old world butterfly, *Lampides boeticus* (Lepidoptera: Lycaenidae). *BMC Evol Biol.* 2008;8(1):301.
10. Shirey V, Larsen E, Doherty A, Kim CA, Al-Sulaiman FT, Hinolan JD, Itliong MGA, Naive MAK, Ku M, Beltz M, et al. LepTraits 1.0 A globally comprehensive dataset of butterfly traits. *Sci Data.* 2022;9(1):382.
11. Bortoluzzi C, Wright CJ, Lee S, Cousins T, Genez TAL, Thybert D, Martin FJ, Haggerty L, Blaxter M, Durbin R. Lepidoptera genomics based on 88 chromosomal reference sequences informs population genetic parameters for conservation. *bioRxiv.* 2023:2023.2004.2014.536868.
12. Garwood K, Jaramillo JG. Mariposas Del Neotrópico lista de chequeo - Suramérica/ Neotropical butterflies Checklist - South America (Lepidoptera: Papilionoidea). *BioButterfly Database.* 2021;V1.
13. Dias FMS, Casagrande MM, Mielke OHH. Biology and external morphology of the immature stages of the butterfly *Callicore Pygas eucale*, with comments on the taxonomy of the genus *Callicore* (Nymphalidae: Biblidinae). *J Insect Sci.* 2014;14:91.
14. Dias FMS, Carneiro E, Casagrande MM, Mielke OHH. Biology and external morphology of immature stages of the butterfly, *Diaethria Candrena Candrena*. *J Insect Sci.* 2012;12(1). <https://doi.org/10.1673/031.012.0901>
15. Francesconi W, Nair PKR, Levey DJ, Daniels J, Cullen L. Butterfly distribution in fragmented landscapes containing agroforestry practices in southeastern Brazil. *Agrofor Syst.* 2013;87(6):1321–38.
16. Pacheco Gómez T, Córdova Horna S, Del Castillo Mozombite DM, Rios Zumaeta R, Cabanillas Oliva EG, Pinedo Jiménez J, Angulo Ruiz P, Tello Espinoza R, Vásquez Bardales J. Supervivencia e insectos Plaga de especies forestales y frutales En Puerto Almendra, región Loreto, Perú. *Folia Amazónica.* 2022;31:209–26.
17. Wang Z, Chan W-P, Pham NT, Zeng J, Pierce NE, Lohman DJ, Meng W. One in five butterfly species sold online across borders. *Biol Conserv.* 2023;283:110092.
18. Garzón-Orduña IJ, Silva-Brandão KL, Willmott K, Freitas AVL, Wahlberg N, Brower AVZ. Wing pattern diversity in *Eunica* butterflies (Nymphalidae: Biblidinae): phylogenetic analysis implies decoupled adaptive trends in dorsal sexual dimorphism and ventral eyespot evolution. *Cladistics.* 2024;40(1):1–20.
19. Yack JE, Otero LD, Dawson JW, Surlykke A, Fullard JH. Sound production and hearing in the blue cracker butterfly *Hamadryas Feronia* (Lepidoptera, Nymphalidae) from Venezuela. *J Exp Biol.* 2000;203(24):3689–702.
20. Vasquez J, Lamas G, Couturier Guy KM. Aspectos biológicos de *Panacea prola Amazonica* (Fruhstorfer) (Lepidoptera: Nymphalidae). *En La Amazonia Peruana. Folia Amazónica.* 2012;21(1–2):71–6.
21. DeVries PJ, Penz CM, Walla TR. The biology of *Batesia hypochlora* in an Ecuadorian rainforest (Lepidoptera: Nymphalidae). *Trop Lepid Res.* 1999:43–6.
22. Schoch CL, Ciufo S, Domrachev M, Hottot CL, Kannan S, Khovanskaya R, Leipe D, McVeigh R, O'Neill K, Robertse B, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database.* 2020;2020.
23. Andrews S. FastQC: a quality control tool for high throughput sequence data. In: Cambridge, United Kingdom: Babraham Bioinformatics, Babraham Institute; 2010.
24. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics.* 2014;30(15):2114–20.
25. Tilak M-K, Botero-Castro F, Galtier N, Nabholz B. Illumina library Preparation for sequencing the GC-rich fraction of heterogeneous genomic DNA. *Genome Biol Evol.* 2018;10(2):616–22.
26. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* 2011;27(6):764–70.
27. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics.* 2017;33(14):2202–4.
28. 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):170–5.
29. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094–100.
30. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997.* 2013.
31. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE.* 2014;9(11):e112963.
32. Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit – Interactive quality assessment of genome assemblies. *G3 (Bethesda).* 2020;10(4):1361–74.
33. Bushnell B. BBMap short-read aligner, and other bioinformatics tools. Berkeley: University of California; 2015.
34. Manni M, Berkeley MR, Seppely 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. *Mol Biol Evol.* 2021;38(10):4647–54.
35. Uliano-Silva M, Ferreira JGRN, Krashenninnikova K, Blaxter M, Mieszkowska N, Hall N, Holland P, Durbin R, Richards T, Kersey P, et al. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics.* 2023;24(1):288.
36. Cally S, Lhuillier E, Iribar A, Garzón-Orduña I, Coissac E, Muriene J. Shotgun assembly of the complete mitochondrial genome of the Neotropical cracker butterfly *Hamadryas epinome*. *Mitochondrial DNA DNA Mapp Seq Anal.* 2016;27(3):1864–6.
37. Flynn JM, Hubley R, 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):9451–7.
38. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA.* 2015;6(1):11.
39. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. In. 2024. <http://www.repeatmasker.org>.
40. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE.* 2016;11(10):e0163962.
41. Card D. GenomeAnnotation. In. 2022.
42. Pacific Biosciences. IsoSeq - scalable de novo isoform discovery from single-molecule PacBio reads. In: Bioconda: Pacific Biosciences of California, Inc. 2019.
43. Gabriel L, Brūna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, Stanke M. BRAKER3: fully automated genome annotation using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *bioRxiv.* 2023.
44. Brūna T, Gabriel L, Hoff KJ. Navigating eukaryotic genome annotation pipelines: a route map to Braker, Galba, and Tsebra. *arXiv.* 2024.
45. Kuznetsov D, Tegenfeldt F, Manni M, Seppely M, Berkeley M, Kriventseva Evgenia V, Zdobnov EM. OrthoDB v1.1: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* 2022;51(D1):D445–51.
46. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20(1):238.
47. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol Biol Evol.* 2017;34(8):2115–22.
48. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol.* 2021;38(12):5825–9.
49. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. GO::TermFinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics.* 2004;20(18):3710–5.
50. Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.* 2004;32(suppl1):D258–61.
51. Duploup A, Hornett EA. Uncovering the hidden players in lepidoptera biology: the heritable microbial endosymbionts. *PeerJ.* 2018;6:e4629.
52. Baldo L, Hotopp JCD, Jolley KA, Bordenstein SR, Biber SA, Choudhury RR, Hayashi C, Maiden MCJ, Tettelin H, Werren JH. Multilocus sequence typing system for the endosymbiont *Wolbachia pipiensis*. *Appl Environ Microbiol.* 2006;72(11):7098–110.
53. Twort VG, Blande D, Duploup A. One's trash is someone else's treasure: sequence read archives from lepidoptera genomes provide material

- for genome reconstruction of their endosymbionts. *BMC Microbiol.* 2022;22(1):209.
54. Perlmutter JI, Bordenstein SR, Unckless RL, LePage DP, Metcalf JA, Hill T, Martinez J, Jiggins FM, Bordenstein SR. The phage gene *Wmk* is a candidate for male killing by a bacterial endosymbiont. *PLoS Pathog.* 2019;15(9):e1007936.
55. Katsuma S, Hirota K, Matsuda-Imai N, Fukui T, Muro T, Nishino K, Kosako H, Shoji K, Takanashi H, Fujii T, et al. A *Wolbachia* factor for male killing in lepidopteran insects. *Nat Commun.* 2022;13(1):6764.
56. LePage DP, Metcalf JA, Bordenstein SR, On J, Perlmutter JI, Shropshire JD, Layton EM, Funkhouser-Jones LJ, Beckmann JF, Bordenstein SR. Prophage WO genes recapitulate and enhance *Wolbachia*-induced cytoplasmic incompatibility. *Nature.* 2017;543(7644):243–7.
57. Beckmann JF, Ronau JA, Hochstrasser M. A *Wolbachia* deubiquitylating enzyme induces cytoplasmic incompatibility. *Nat Microbiol.* 2017;2(5):17007.
58. Duploup A, Iturbe-Ormaetxe I, Beatson SA, Szubert JM, Brownlie JC, McMeniman CJ, McGraw EA, Hurst GDD, Charlat S, O'Neill SL, et al. Draft genome sequence of the male-killing *Wolbachia* strain wBol1 reveals recent horizontal gene transfers from diverse sources. *BMC Genomics.* 2013;14(1):20.
59. Kawahara AY, Storer C, Carvalho APS, Plotkin DM, Condamine FL, Braga MP, Ellis EA, St Laurent RA, Li X, Barve V, et al. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. *Nat Ecol Evol.* 2023;7(6):903–13.
60. 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 [version 1; peer review: 2 approved]. *Wellcome Open Res.* 2023;8(24). [10.12688/wellcomeopenres.18658.1](https://doi.org/10.12688/wellcomeopenres.18658.1)
61. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol.* 2018;14(1):e1005944.
62. Chakraborty M, Lara AG, Dang A, McCulloch KJ, Rainbow D, Carter D, Ngo LT, Solares E, Said I, Corbett-Detig RB, et al. Sex-linked gene traffic underlies the acquisition of sexually dimorphic UV color vision in *Heliconius* butterflies. *Proc Natl Acad Sci U S A.* 2023;120(33):e230141120.
63. Brown S, Keith J, Lucci Freitas V, Wahlberg A, Von Schoutz N, Saura BO, Saura A. Chromosomal evolution in the South American nymphalidae. *Hereditas.* 2007;144(4):137–48.
64. Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P, Välimäki N, Paulin L, Kvist J, Wahlberg N, et al. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in lepidoptera. *Nat Commun.* 2014;5(1):4737.
65. Wright CJ, Stevens L, Mackintosh A, Lawniczak M, Blaxter M. Comparative genomics reveals the dynamics of chromosome evolution in lepidoptera. *Nat Ecol Evol.* 2024;8(4):777–90.
66. Saura A, Schoutz BV, Saura AO, Brown Jr. KS: chromosome evolution in Neotropical butterflies. *Hereditas.* 2013;150(2–3):26–37.
67. Kandul NP, Lukhtanov VA, Pierce NE. Karyotypic diversity and speciation in *Agrodiaetus* butterflies. *Evolution.* 2007;61(3):546–59.
68. Janz N, Nylin S, Wahlberg N. Diversity begets diversity: host expansions and the diversification of plant-feeding insects. *BMC Evol Biol.* 2006;6(1):4.
69. Augustijn H, Batscher L, Cesanek M, Chkhartishvili T, Dincă V, Iankoshvili G, Ogawa K, Vila R, Klopstein S, de Vos JM, et al. A macroevolutionary role for chromosomal fusion and fission in *Erebia* butterflies. *Sci Adv.* 2024;10(16):eadl0989.
70. Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L. Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends Genet.* 2005;21(12):673–82.
71. de Vos JM, Augustijn H, Batscher L, Lucek K. Speciation through chromosomal fusion and fission in lepidoptera. *Philos Trans R Soc Lond B Biol Sci.* 2020;375(1806):20190539.
72. Mackintosh A, Vila R, Laetsch DR, Hayward A, Martin SH, Lohse K. Chromosome fissions and fusions act as barriers to gene flow between *Brenthis* fritillary butterflies. *Mol Biol Evol.* 2023;40(3). [10.1093/molbev/msad043](https://doi.org/10.1093/molbev/msad043)
73. Smolander O-P, Blande D, Ahola V, Rastas P, Tanskanen J, Kammonen JJ, Oostra V, Pellegrini L, Ikonen S, Dallas T, et al. Improved chromosome-level genome assembly of the Glanville fritillary butterfly (*Melitaea cinxia*) integrating Pacific biosciences long reads and a high-density linkage map. *Gigascience.* 2022;11.
74. Talla V, Suh A, Kalsoom F, Dincă V, Vila R, Friberg M, Wiklund C, Backström N. Rapid increase in genome size as a consequence of transposable element hyperactivity in Wood-White (*Leptidea*) butterflies. *Genome Biol Evol.* 2017;9(10):2491–505.
75. Höök L, Näsvall K, Vila R, Wiklund C, Backström N. High-density linkage maps and chromosome level genome assemblies unveil direction and frequency of extensive structural rearrangements in wood white butterflies (*Leptidea* spp.). *Chromosome Res.* 2023;31(1):2.
76. Hornett EA, Duploup AMR, Davies N, Roderick GK, Wedell N, Hurst GDD, Charlat S. You can't keep a good parasite down: evolution of a male-killer suppressor uncovers cytoplasmic incompatibility. *Evolution.* 2008;62(5):1258–63.
77. Charlat S, Duploup A, Hornett EA, Dyson EA, Davies N, Roderick GK, Wedell N, Hurst GDD. The joint evolutionary histories of *Wolbachia* and mitochondria in *Hypolimnas Bolina*. *BMC Evol Biol.* 2009;9(1):64.
78. Salunkhe RC, Narkhede KP, Shouche YS. Distribution and evolutionary impact of *Wolbachia* on butterfly hosts. *Indian J Microbiol.* 2014;54(3):249–54.
79. Duploup A, Nair A, Nyman T, van Nouhuys S. Long-term spatiotemporal genetic structure of an accidental parasitoid introduction, and local changes in prevalence of its associated *Wolbachia* symbiont. *Mol Ecol.* 2021;30(18):4368–80.
80. Charlat S, Reuter M, Dyson EA, Hornett EA, Duploup A, Davies N, Roderick GK, Wedell N, Hurst Gregory DD: Male-killing bacteria trigger a cycle of increasing male fatigue and female promiscuity. *Curr Biol.* 2007;17(3):273–7.
81. Charlat S, Hornett EA, Dyson EA, Ho PP, Loc NT, Schilthuizen M, Davies N, Roderick GK, Hurst GD. Prevalence and penetrance variation of male-killing *Wolbachia* across Indo-Pacific populations of the butterfly *Hypolimnas Bolina*. *Mol Ecol.* 2005;14(11):3525–30.
82. Charlat S, Hornett EA, Fullard JH, Davies N, Roderick GK, Wedell N, Hurst GDD. Extraordinary flux in sex ratio. *Science.* 2007;317(5835):214–214.
83. Hornett EA, Charlat S, Wedell N, Jiggins CD, Hurst GDD. Rapidly shifting sex ratio across a species range. *Curr Biol.* 2009;19(19):1628–31.
84. Hiroki M, Kato Y, Kamito T, Miura K. Feminization of genetic males by a symbiotic bacterium in a butterfly, *Eurema hecabe* (Lepidoptera: Pieridae). *Naturwissenschaften.* 2002;89(4):167–170.
85. Jiggins H, Dolman M. High-prevalence male-killing *Wolbachia* in the butterfly *Acraea Encedana*. *J Evol Biol.* 2000;13(3):495–501.

## Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.