Title: Transposable element accumulation drives genome size increase in *Hylesia metabus* (Lepidoptera: Saturniidae), an urticating moth species from South America

Authors: Charles PERRIER 1*, Rémi ALLIO 1, Fabrice LEGEAI 2,3, Mathieu GAUTIER 1, Frédéric BÉNÉLUZ 4, William MARANDE 5, Anthony THERON 5, Nathalie RODDE 5, Melfran HERRERA 6, Laure SAUNE 1, Hugues PARRINELLO 7, Melanie McCLURE 8, Mónica ARIAS 9

Addresses:

1. UMR CBGP, INRAE, CIRAD, IRD, Institut Agro, Université Montpellier, Montpellier, France
2. BIPAA IGEPP, INRAE Institut Agro Univ Rennes Rennes France
3. Univ Rennes, Inria, CNRS, IRISA, 35000, Rennes, France
4. Société entomologique Antilles-Guyane, SEAG, Rémire-Montjoly, France
5. French Plant Genomic Resource Center, CNRGV, INRAE, Castanet Tolosan, France
6. Coordinación de Vigilancia Epidemiológica Ambiental, Dirección Estadal de Salud Ambiental, FUNDASALUD, Estado Sucre, Venezuela
7. MGX-Montpellier GenomiX, Université Montpellier, CNRS, INSERM, Montpellier, France
8. Laboratoire Écologie, Évolution, Interactions des Systèmes Amazoniens (LEEISA), Université de Guyane, CNRS, IFREMER, Cayenne, France
9. UMR PHIM, CIRAD, INRAE, Institut Agro, IRD, Université Montpellier, Montpellier, France

* corresponding author

Abstract:

We present the first nuclear genome assembly and a complete mitogenome for *Hylesia metabus* (Arthropoda; Insecta; Lepidoptera; Saturniidae). The assembled nuclear genome sequence is 1,271 Mb long, which is among the 10 largest lepidopteran genome assemblies published to date. It is scaffolded in 31 pseudo chromosomes, has a BUSCO score of 99.5%, and has a highly conserved synteny compared to phylogenetically close species. Repetitive elements make up 67% of the nuclear genome and are mainly located in intergenic regions, among which LINEs were predominant, with CR1-Zenon being the most abundant. Phylogenetic and comparative analyses of *H. metabus* assembly and 17 additional Saturniidae and Sphingidae assemblies suggested that an accumulation of repetitive elements likely led to the increased size of *H. metabus'* genome. Gene annotation using Helixer identified 26,122 transcripts. The Z scaffold was identified using both a synteny analysis and variations of coverage for two resequenced male and female *H. metabus*. The *H. metabus* nuclear genome and mitogenome assemblies can be found and browsed on the BIPAA website and constitute useful resources for future population and comparative genomics studies.

Keywords: ashen moth, lepidopterism, population outbreaks, repetitive elements, yellowtail moth

Introduction

The yellowtail moth *Hylesia metabus* (Saturniidae, Lepidoptera, Figure 1) known as "palometa peluda" in Venezuela and "papillon cendre" in French Guiana, is probably the most studied *Hylesia* species due to the health problems it causes. Like other species in the genus, adult females have urticating hairs that are easily released into the air, which can then come into contact with humans and cause a painful dermatitis (referred to as "Caripito itch" in Spanish or "papillonite" in French) and in extreme cases can cause respiratory problems (Rodriguez-Morales et al., 2005). Unlike other species, *H. metabus* is largely distributed in northern South America and it is responsible for epidemic outbreaks in Venezuela and French Guiana (Ciminera et al., 2019; Hernández et al., 2012; Jourdain et al., 2012). During outbreaks, hundreds to thousands of females fly simultaneously over human settlements, attracted by urban lights (Jourdain et al., 2012). The resulting abundance of urticating hairs negatively impact society by forcing citizens to shut themselves inside their houses at dusk so as to limit risks of dermatitis, and schools are forced to close to prevent children getting into contact with urticating hairs that remain on school grounds (ANSES French Agency for food environmental and occupational health & safety, 2011). *Hylesia metabus* populations are present in heterogeneous environments such as forest, savannahs and mangroves, although only populations in coastal areas are known to cause problems of epidemic dermatitis (Jourdain et al., 2012; Rodriguez-Morales et al., 2005). Using mitochondrial markers and nuclear microsatellite markers, previous studies have shown that, although *H. metabus* populations do belong to a single species, populations are genetically differentiated at relatively small spatial scale in French Guiana and Venezuela, notably between forest and mangrove habitats (Cequena et al., 2012; Ciminera et al., 2019). To better investigate the genomics of *H. metabus* populations and the potential genetic

determinants responsible for a population's propensity to produce problematic outbreaks, more in-depth genomics studies are needed. Hence, sequencing, assembling and annotating the first reference genome for *H. metabus* was essential to enable future population genomic studies, and here we achieved this using PacBio HiFi long reads scaffolded with Omni-C data.

Figure 1. Pictures of *Hylesia metabus* at different developmental stages: nest covered by urticating hairs and with first stage larvae (left); gregarious larvae at stages L2, L3, L4 (upper middle); larva stage L7 (lower middle - stage used for DNA extraction); adult female (up right); adult male (lower right). (Photos of nest and larvae ©Jean-Philippe Champenois).

Materials and methods

Sample collection

In Stoupan (4.750 N 52.331 W), French Guiana, we collected in September 2021 two *H. metabus* larvae for the genome sequencing and scaffolding, with one larva being used for the HiFi library construction and the other larva used for the Omni-C library construction. We also collected 1 adult male and 1 adult female, for whole genome individual resequencing. Each larva was flash frozen in liquid nitrogen before being stored at -80°C. Each adult was stored in ethanol 85° and stored at -20°C.

DNA extractions, libraries preparations and sequencing

For HiFi sequencing, high molecular weight (HMW) DNA was extracted from 0.3g of H. *metabus*, from the first larvae, using QIAGEN Genomic-tips 500/G kit (Qiagen, MD, USA). We followed the tissue protocol extraction, which in brief consisted of 0.3g of frozen *H. metabus* larvae abdomen ground in liquid nitrogen with a mortar and pestle. After 3h of lysis and one centrifugation step, the DNA was immobilized on the column. After several washing steps, DNA was eluted from the column, then desalted and concentrated by Isopropyl alcohol precipitation. A final wash in 70% ethanol was performed before resuspending the DNA in EB buffer. Analyses of DNA quantity and quality were performed using NanoDrop and Qubit (Thermo Fisher Scientific , MA, USA). DNA integrity was also assessed using the Agilent FP-1002 Genomic DNA 165 kb on the Femto Pulse system (Agilent, CA, USA). Hifi library was constructed using SMRTbell® Template Prep kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA) according to PacBio recommendations (SMRTbell® express template prep kit 2.0 - PN: 100-938-900). HMW DNA samples were first purified with 1X Agencourt AMPure XP beads (Beckman Coulter, Inc, CA USA), and sheared with Megaruptor 3 (Diagenode, Liège,

103 BELGIUM) at an average size of 20 kb. After End repair, A-tailing and ligation of SMRTbell
104 adapter, the library was selected on BluePippin System (Sage Science, MA,USA) for a range
105 size of 10-50kb. The size and concentration of the library were assessed using the Agilent
106 FP-1002 Genomic DNA 165 kb on the Femto Pulse system and the Qubit dsDNA HS reagents
107 Assay kit. Sequencing primer v5 and Sequel® II DNA Polymerase 2.2 were annealed and
108 bound, respectively, to the SMRTbell library. The library was loaded on one SMRTcell 8M at
109 an on-plate concentration of 90pM. Sequencing was performed on the Sequel® II system at
110 Gentyane Genomic Platform (INRAE Clermont-Ferrand, France) with Sequel® II Sequencing
111 kit 3.0, a run movie time of 30 hours with an Adaptive Loading target (P1 + P2) at 0.75. After
112 filtering and correcting the SMRTcell output, we obtained 2,117,541 reads totalizing 42.7
113 Gb, with an N50 of 20,851 bp measured with LongQC v1.2.1 (Fukasawa et al., 2020) .

114 The Omni-C library (Dovetail Genomics®) was produced according to the
115 manufacturer instructions. In brief, this consisted of 33mg of frozen *H. metabus* abdomen of
116 the second larva, ground in liquid nitrogen and suspended in PBS. Then the DNA was fixed
117 with formaldehyde and digested using 2µl of a nuclease enzyme mix. After binding 500ng of
118 the digested DNA to chromatin capture beads, a proximity ligation was performed and the
119 crosslinks were reversed to produce the linked DNA. Finally, 107ng of the linked DNA was
120 used to produce a library then paired-end (2x150bp) sequenced on three different S4 flow-
121 cell lanes on an Illumina® NovaSeq system. The obtained raw paired-end reads were filtered
122 using fastp v0.23.2 (Chen et al., 2018) run with default options leading to a total of 171
123 millions of read pairs (51.2 Gb).

124 Two whole genome libraries of one adult female and one adult male were
125 constructed using truseq kit to produce illumina short reads. In brief, DNA from one adult
126 female and one adult male was extracted using the blood and tissue kit, column style, from

127 Qiagen (Qiagen, MD, USA). DNA concentration was estimated using both Qubit fluorometric
128 measures and nanodrop absorbance measures. Libraries were then constructed using the
129 TruSeq nano DNA kit from Illumina and paired-end (2x150bp) sequenced on a S4 flow-cell
130 lanes on an Illumina® NovaSeq system, producing 234 millions read pairs for the male and
131 325 millions read pairs for the female after filtering with fastp v0.23.2.
132
133 *Nuclear genome assembly, filtering and scaffolding*
134 We used Jellyfish (Marçais & Kingsford, 2011) and GenomeScope (Ranallo-Benavidez et al.,
135 2020; Vurture et al., 2017) to estimate the genome size from HiFi data. A whole genome
136 assembly was then built from HiFi reads HiFiasm v0.16.1 (Cheng et al., 2021) ran with
137 default options.
138 We detected and filtered out the mitogenome, potential contaminants and haplotigs
139 as follows. Using MitoFinder v1.4.2 (Allio et al., 2020) on the primary assembly, we detected
140 and annotated a contig corresponding to the complete mitochondrial genome sequence of
141 *H. metabus*. The mitochondrial genome was removed from the nuclear genome assembly
142 and we drew a circular representation using CGView (Stothard & Wishart, 2005). We used
143 blobtools (Laetsch & Blaxter, 2017) to filter out potential contaminants. We also searched
144 for potential contaminant sequences originating from the plant species' genome on which
145 larvae were found feeding on (*Tapirira guianensis*) by mapping short reads from this species
146 (https://www.ebi.ac.uk/ena/browser/view/ERR7620141) on our putative *H. metabus*
147 scaffolds using bwa-mem2 v2.2.1 (Vasimuddin et al., 2019) and samtools v1.10 (Danecek et
148 al., 2021). We used purge haplotigs v1.1.2 (Roach et al., 2018) to remove potential haplotigs
149 and small contigs exhibiting bad mapping quality or that could be considered as junk or
150 repeats.

The contig assembly was scaffolded using Omni-C sequencing data. Following Serizay et al. (2024), we separately mapped the filtered paired-end reads from the three sequencing lanes to the contig assembly using bwa-mem2 v2.2.1 (Vasimuddin et al., 2019) which was run with options *-SP5M*. The three resulting bam files generated with the samtools v. 1.14 *view* (Li et al., 2009) were then parsed, sorted and deduplicated with *pairtools v1.0.3* (Open2C et al., 2023) programs *parse* (run with options *–min-mapq 20* and *—drop-sam*), *sort,* and *dedup*, respectively. The three *pairs* files were further merged with the *pairtools merge* program and the *dump* program from the *cooler* v0.9.3 suite (Abdennur & Mirny, 2020) was used to generate a contact matrix (using 500 kb bins) that was visualized using a custom R function. The identified Omni-C pairs were finally used to scaffold the assembly using *YaHS v1.2* (Zhou et al., 2023) ran with default options but *–file-type PA5* specification to read the *pairtools* generated pair file. A contact map for the resulting scaffolded assembly was generated as described above after converting Omni-C pairs mapping coordinates from the contig assembly (based on the agp file) using a custom *awk* script.

Completeness of the contig and scaffolded assemblies was evaluated using Benchmark Universal Single Copy Orthologs (BUSCO v5.5.0, Manni et al., 2021; Simão et al., 2015) for “arthropoda_odb10” and “lepidoptera_odb10” databases. Blobtools (Laetsch & Blaxter, 2017) was used to draw a snailplot graph.

Annotation of protein coding genes and repetitive elements

Protein coding gene prediction was achieved on the non-masked genome version of *H. metabus*. Helixer v0.3.0 with the option --lineage invertebrate was used for gene prediction (Holst et al., 2023). The quality of the annotation was assessed with BUSCO v5.2.2 using lineage dataset lipidoptera_odb10, PSAURON 1.0.2 (Sommer et al., 2024) and OMArk

175 2023.10 (Nevers et al., 2024). Functional annotation of the protein sequences obtained with
176 GFFread (Pertea & Pertea, 2020) from the Helixer output were done with Diamond v2.0.13
177 (Buchfink et al., 2015) on NCBI NR 2022-12-11, Blast2GO Command Line v1.5.1 (Götz et al.,
178 2008), eggNOG v2.1.9 (Huerta-Cepas et al., 2019) with eggnog database v5.0.2 and
179 Interproscan v5.59-91.0 (Jones et al., 2014). The genome sequence and its annotations can
180 be browsed at https://bipaa.genouest.org/sp/hylesia_metabus/.

181 Repetitive elements were identified using EarlGrey (Baril et al., 2024) v4.1.0, which
182 notably uses RepeatMasker (Smit et al., 2015), RepeatModeler2 (Flynn et al., 2020), and
183 LTR_Finder (Xu & Wang, 2007). We used the Arthropoda repeat library from DFAM 3.5 as
184 the initial repeats library. We inspected the distribution of repetitive elements across the
185 genome, in intergenic regions, introns, UTR, and CDS from the Helixer GFF file.

186

187 *Phylogenetic and comparative analyses*

188 Phylogenetic analyses have been done by comparing the BUSCO sequences from the
189 genome of *H. metabus* to the one of 4 other Saturniidae species (*Automeris io, Samia ricini,*
190 *Antheraea yamamai, Saturnia pavonia*), 12 Sphingidae species (*Cephonodes hylas, Hemaris*
191 *fuciformis, Hyles euphorbiae, Deilephila porcellus, Theretra japonica, Manduca sexta, Lapara*
192 *coniferarum, Sphinx pinastri, Clanis bilineata, Laothoe populi, Amorpha juglandis, Mimas*
193 *tiliae*), and *Bombyx mori* as outgroup. External assemblies were downloaded from NCBI
194 using their respective accession number with the command-line tool "datasets" (e.g.
195 datasets download genome accession GCF_030269925.1 --filename Bombyx_mori.zip; see
196 supplementary material 1 for accessions). For each assembly, BUSCO sequences were
197 annotated and extracted based on the predefined dataset "Lepidoptera_odb10", which
198 includes 5,286 orthologous genes for Lepidoptera. BUSCO search was performed through

199 the gVolante web server (Nishimura et al., 2017). Amino acid sequences corresponding to
200 every BUSCO marker were first individually aligned with the MAFFT (Katoh & Standley,
201 2016) algorithm FFT-NS-2. All markers were then concatenated in one supermatrix using
202 seqCat.pl and seqConverter.pl (Bininda-Emonds, 2006; Fasterius & Al-Khalili Szigyarto,
203 2019). Phylogenetic inferences were performed with Maximum-likelihood (ML) as
204 implemented in IQ-TREE V2.2.2.6 (Minh et al., 2022). One partition per BUSCO sequence
205 was defined with the option "-spp". The best evolutionary model was selected for each
206 partition using ModelFinder implemented in IQ-TREE and some partitions were merged if
207 necessary (-m MFP+MERGE). Following the recommendation of IQ-TREE developers, we also
208 set a smaller perturbation strength (-pers 0.2) and a larger number of stop iterations (-nstop
209 500) to avoid local optima. Finally, node supports were evaluated with UltraFast Bootstraps
210 (UFBS) estimated by IQ-TREE (-bb 1000). UFBS are considered robust when higher than 95%.

211 Repetitive elements content and repeat landscape in *H. metabus* was compared to
212 the one in the 17 species mentioned above, for which we ran the same EarlGrey pipeline as
213 for *H. metabus*.

214 Synteny between the genome of *H. metabus* and the genomes of *Antheraea*
215 *yamamai, Saturnia pavonia*, *Deilephila porcellus, Manduca sexta, Laothoe populi,* and
216 *Bombyx mori, extracted from NCBI (see above and supplementary table 1)* were inspected
217 with genespace (Lovell et al., 2022) using the gene positions derived from the GFF files, and
218 their corresponding orthogroups predicted with orthofinder v2.5.5 (Emms & Kelly, 2019).
219 Because protein coding genes annotation of *Antheraea yamamai, Saturnia pavonia*,
220 *Deilephila porcellus,* and *Laothoe populi* were not available at NCBI, we produced their
221 respective annotations with Helixer (v0.3.3, with the options --lineage invertebrate --
222 subsequence-length 108000 --overlap-offset 54000 --overlap-core-length 81000). Finally,

223 the sizes of intergenic and genic regions were calculated from the GFF file with a custom
224 script.
225
226 *Identification of the Z scaffold*
227 In order to identify the scaffold corresponding to the Z chromosome in *H. metabus*, we first
228 inspected synteny graphs with *Deilephila porcellus* and *Saturnia pavonia* assemblies for
229 which the Z scaffold was known. We also used Dgenies (Cabanettes & Klopp, 2018) to
230 examine more precisely the synteny between *Hylesia metabus* and *Deilephila porcellus*
231 (Boyes et al., 2022). Second, we investigated potential coverage variation among scaffolds
232 between the resequenced male and female. To do so, we first used Fastp (Chen et al., 2018)
233 to keep only good quality sequences that we then aligned to the *H. metabus* assembly using
234 bwa-mem2 v2.2.1 (Vasimuddin et al., 2019). We then used samtools to sort and index reads
235 and to estimate coverage per scaffold for each individual (Danecek et al., 2021). For each
236 scaffold we measured the female to male ratio of percentage of reads mapped to each
237 scaffold. A ratio of approximately 50% would indicate the Z scaffolds, while ratios of 100%
238 would indicate autosomal scaffolds. Finally, in order to determine the genotypic sex of the
239 individual larvae sequenced to assemble the genome, we investigated potential variations of
240 HiFi raw reads coverage between the putative Z scaffold and the other large scaffolds. A
241 coverage deficit of about 50% on the Z scaffold would illustrate that the sequenced
242 individual was a female.
243
244 **Results and discussion**
245 *Assembly of a 1.27 GB long genome scaffolded in 31 pseudo chromosomes*

246 The Jellyfish and Genomescope analysis of the HiFi reads suggested that the genome size
247 was 1.19 Gb with a heterozygosity of 2.19% (Figure 2A). The Hifiasm assembly consisted of
248 171 contigs ranging from 15kb to 57Mb and totalizing 1.41 Gb, with an N50 of 34,63 Mb, an
249 L50 of 17 and 39.85% of GC (Supplementary material 2). MitoFinder identified the smallest
250 (15393 bp) and most covered (174 X) contig, as being the complete mitochondrial genome
251 sequence and annotated it fully (Supplementary material 3). The final assembly, after
252 decontamination, purge of haplotigs, removal of the mitochondrial contig, and scaffolding
253 with Omni-C, consisted in 31 scaffolds totalizing 1.27 Gb, with an N50 of 45,18 Mb, an L50 of
254 14 and 39.58% of GC (Figure 2B, Supplementary material 4). The contact map of the primary
255 assembly showed very high contiguity even before scaffolding (Supplementary material 5),
256 enabling an efficient scaffolding toward a chromosome level (Figure 2C). The haploid read
257 depth was on average 14 X, slightly lower than the targeted read depth, as a consequence of
258 a larger genome size than expected. Arthropoda BUSCO gene representation of the final
259 assembly was 99.5% complete with 98.7% single-copy genes, and the Lepidoptera BUSCO
260 gene representation was 98.6% complete with 97.6% single-copy genes (Supplementary
261 material 4, Figure 2B), with less than 1% of duplicated genes.
262

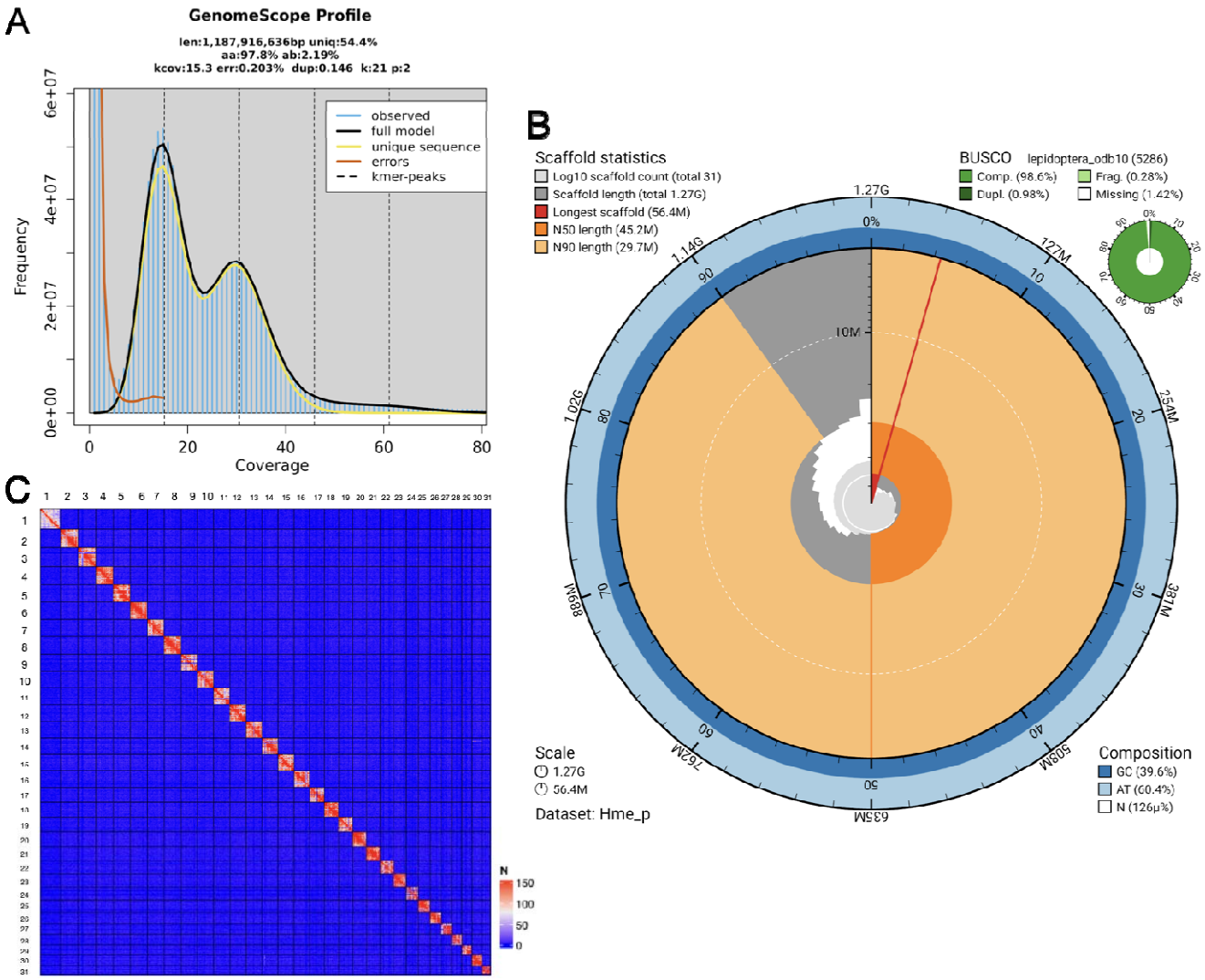

Figure 2. A) K-mer spectra output generated from corrected PacBio HiFi data using GenomeScope. The bimodal pattern observed corresponds to a diploid heterozygous genome. B) BlobToolKit Snailplot showing N50 metrics and BUSCO gene completeness. C) Omni-C contact map of the scaffolded genome sequence of *Hylesia metabus* (number of contacts per 500 kb bin).

Highly conserved synteny but size increase of intergenic regions contributing to a large genome size compared to phylogenetically close species

Synteny was highly conserved between *H. metabus* and phylogenetically close species (Figure 3A), suggesting no evidence for large chromosomal rearrangements and good quality of the assembly. In addition, the number of scaffolds for our assembly was identical to *Saturnia pavonia*, and very similar to other close species, indicating no chromosomal fusion or fission. However, the genome sequence size, 1.27 Gb, was much larger than for other available Saturniidae genome assemblies. For example, the genome assemblies for the close species *Automeris io* and *Saturnia pavonia* were both 490 Mb long (Crowley et al., 2024; Skojec et al., 2024). Yet, other lepidoptera species are known to have large genomes, notably *Euclidia mi*, 2.32 GB (Boyes & Holland, 2023), *Parnassius behrii* 1.59 GB (GCA_036936625.1), *Parnassius apollo* 1.4 GB (Podsiadlowski et al., 2021), *Tholera decimalis* 1.33 GB (Boyes et al., 2023) , *Thaumatotibia leucotreta*, 1.28 GB (Bierman et al., 2023) , *Graphium colonna* 1.27 GB (Triant & Pirro, 2023). To date, the *H. metabus* genome is amongst the top 10 largest lepidopteran genome assemblies present on NCBI database as of 24/05/2024. In line with the highly conserved synteny, the increase in total genome size, and the very low level of duplicated BUSCO genes, we found much larger intergenic regions in *H. metabus* compared to phylogenetically close species but comparable sizes of genic regions (Figure 3B), in a regular manner across chromosomes (Supplementary material 6).

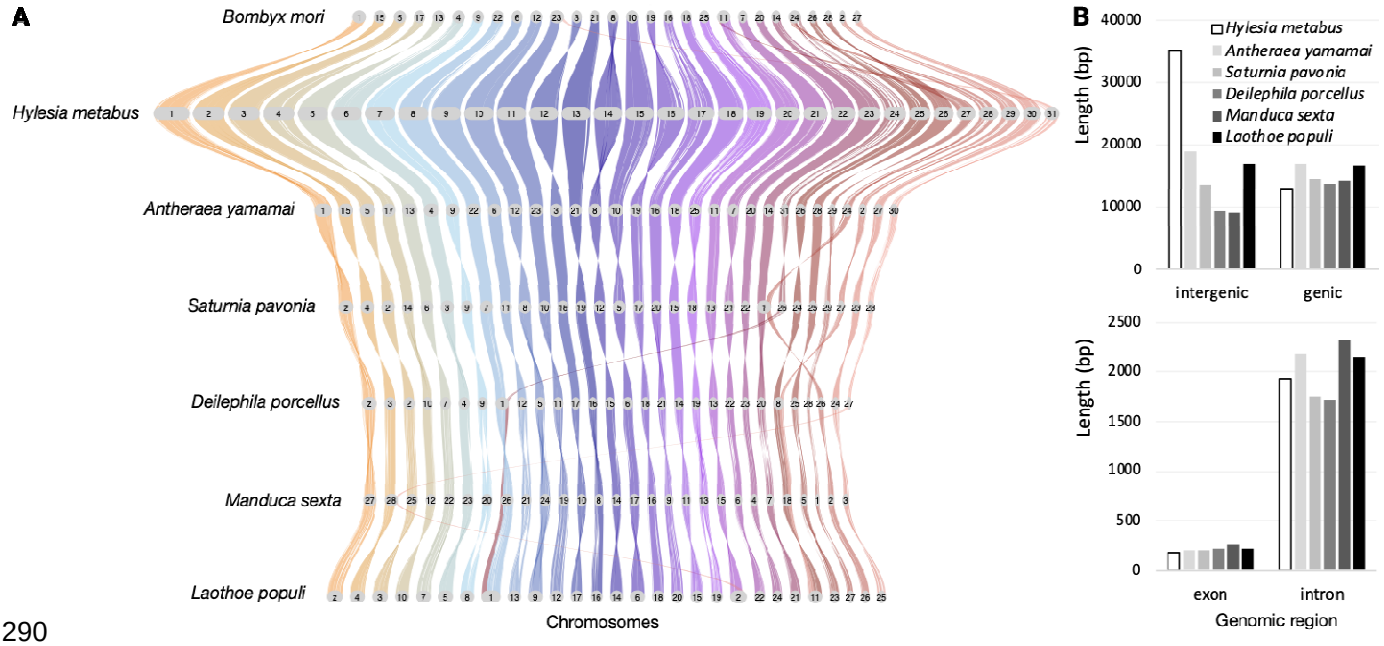

Figure 3. A) Synteny for *Hylesia metabus* and 2 other Saturniidae species, 3 Sphingidae species, and *Bombyx mori*. B) Length of intergenic and genic regions, and of exons and introns, for *H. metabus* and 2 other Saturniidae species, 3 Sphingidae species, and *Bombyx mori*.

Invasion of repetitive elements

Analysis of repetitive elements was achieved with the fully automated EarlGrey pipeline and determined that 67% of the *H. metabus* genome sequence was made-up of repeats (Figure 4A & 4B, Supplementary material 1 & 4). Compared to other species of Saturniidae and species of Sphingidae, *H. metabus* had a higher proportion of repetitive elements (Figure 4A, Supplementary material 1 & 4), although this was similar to the 65% reported for the aforementioned large genome of *Parnassius apollo* (Podsiadlowski et al., 2021). In general, TE invasion in lepidoptera, and more broadly in arthropods, is associated with increased genome size (Gilbert et al., 2021; Muller et al., 2021; Petersen et al., 2019).

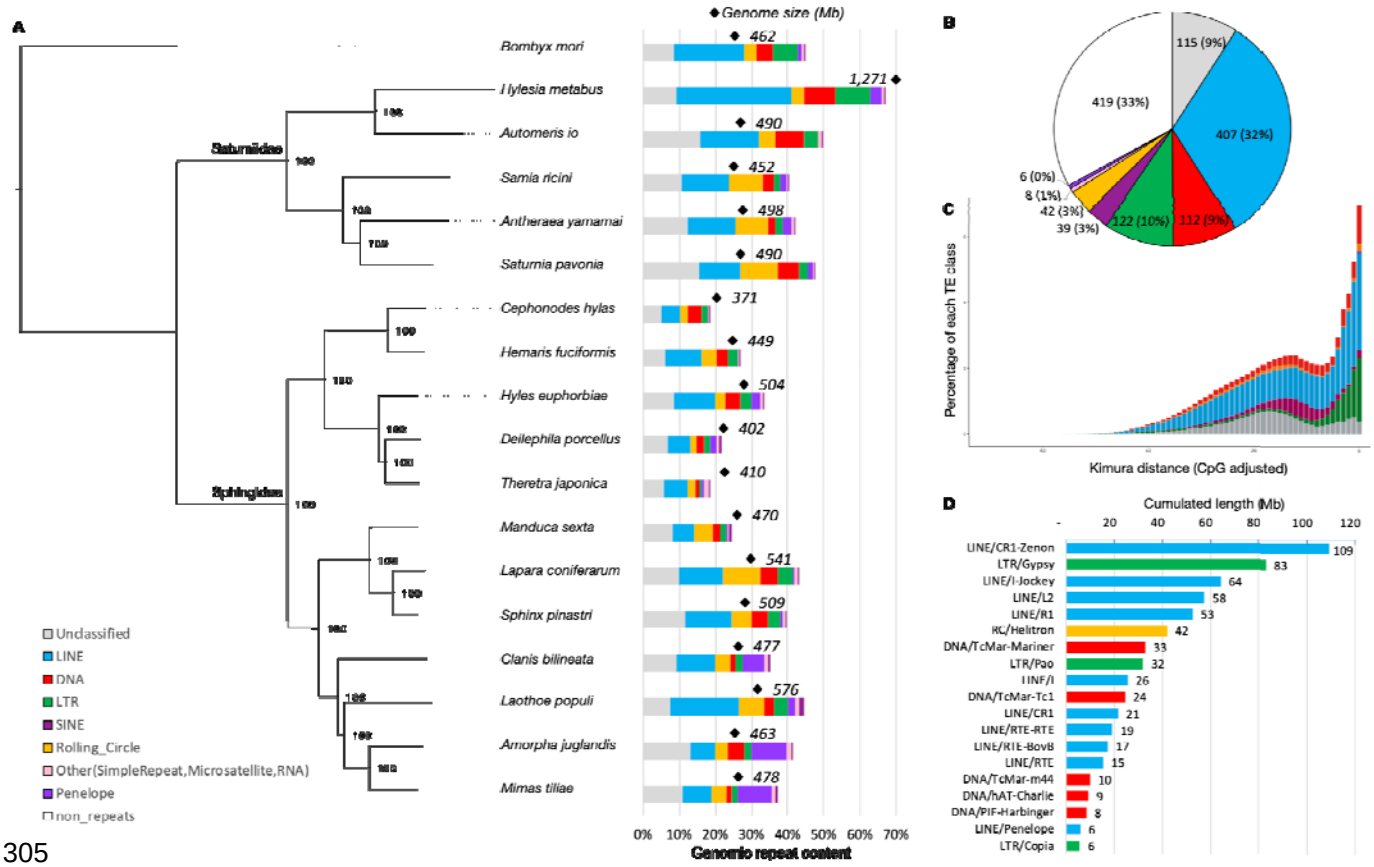

Figure 4. A) Phylogeny and comparative analysis of repeat content for *Hylesia metabus* and 4 other Saturniidae species, 12 Sphingidae species, and *Bombyx mori* as an outgroup. Phylogenetic tree (left panel), genome sequence length and repeat content (right panel). Repeat content (B), repeat landscape (C) and cumulated length for each of the most abundant TE families (D) in the genome assembly of *H. metabus*.

Repetitive elements were found more often in intergenic regions (77.1%) than in introns (21.5%), UTR (0.3%) and CDS (1.1%) (Supplementary material 7A). Moreover, intergenic regions consisted of repeats (68.1%) more than introns (60.4%), UTR (25.8%) and CDS (28.0%; Supplementary material 7B). This reinforces the hypothesis that the invasion of repetitive elements explains the large size of intergenic regions and of the entire genome in *H. metabus*.

318 The majority of repetitive elements identified in *H. metabus* were of type LINE (32%
319 of the genome, hence 55% of the classified repeats; Figure 4B), followed by LTR, DNA and
320 SINE elements (respectively 10%, 9% and 3% of the genome, hence 17%, 16% and 5% of the
321 classified repeats). Only 9% of the genome was made of unclassified repetitive elements.
322 These LINEs and LTRs proportions are higher than for other species of Saturniidae and
323 species of Sphingidae (Figure 4A, Supplementary material 1). This higher proportion of LINEs
324 and LTRs is comparable to *Parnassius apollo*.

325 The repeat landscape of *H. metabus* suggests that repetitive elements likely invaded
326 the genome relatively regularly over time, with a particularly recent burst of invasion of LTRs
327 (Figure 4C). The repeat landscape of *H. metabus* was similar to the one found in the closest
328 species, *Automeris io*, but slightly contrasted with other Saturniidae and Sphingidae, with
329 increasing phylogenetic distance (Supplementary material 8).

330 The five most abundant repetitive elements were LINE/CR1-Zenon (109M bp -
331 319,358 copies, Figure 4D, Supplementary material 9), LTR/Gypsy (83M bp), LINE/I-Jockey
332 (64M bp), LINE/L2 (58M bp) and LINE/R1 (42M bp). For the four other Saturniidae species
333 considered in this study (Supplementary material 10), LINE/R1 was the most abundant
334 family in *Automeris io,* followed by RC/Helitron and LINE/CR1-Zenon, and RC/Helitron was
335 the most abundant family in the three other species, followed by LINE/I-Jockey in two
336 species and by LINE/L2 and LINE/R1 in the last one. This more detailed analysis therefore
337 shows that the TE invasion of *H. metabus* genome is primarily driven by a few TE families
338 that are also among the most abundant ones in the closest species genome sequences. In
339 particular, the LINE CR1 *Zenon* has been shown to successfully invade other lepidopteran
340 species' genomes (Wang et al., 2019).

341

342 *Gene prediction, Z scaffold and data accessibility*

343 Gene prediction with Helixer identified 26,122 transcripts on the non-masked version of the
344 genome (Supplementary material 4). The BUSCO score for the annotated genes was 88.9%
345 complete (Supplementary material 4). Comparatively OMArk identified 6361 complete
346 proteins among the 6779 Obtectomera HOG, including 1466 duplicated, in the range of
347 other closely related species. OMArk also determined that 77.25% of the protein sequences
348 were placed at a consistent lineage, while 16.31% were unknown and reported no
349 contamination. The psauron score, reflecting the likelihood of being a genuine protein
350 coding sequence, was 96.9.

351 The synteny inspection showed that the largest *H. metabus* scaffold (scaffold_1)
352 corresponded to the Z scaffold in *Deilephila porcellus* and *Saturnia pavonia* assemblies
353 (Figure 3A, Supplementary material 11). This was confirmed by the comparisons of read
354 mapping between one male and one female (that had average coverage of 16X and 25X,
355 respectively), revealing a 0.54 female to male relative ratio of coverage on scaffold 1
356 (Supplementary material 12). Finally, mapping HiFi reads of the assembled individual on the
357 final assembly, we estimated that Scaffold_1 had on average 57% lower read depth than the
358 other 30 largest scaffolds, furthermore confirming that scaffold_1 indeed corresponds to
359 the Z chromosome and suggesting that the individual used to assemble the genome was a
360 female.

361

362 **Conclusion**

363 Here we present a high-quality genome sequence assembly for *H. metabus*, a Saturniidae
364 moth species known for causing painful human dermatitis especially during recurrent
365 demographic outbreaks. This genome sequence is among the 10 largest lepidopteran
366 genome sequences published to date. The genome expansion could be explained by an
367 invasion of repetitive elements, especially of LINEs and LTRs, in intergenic regions, as
368 observed in a few other lepidopteran species. Both genome size, intergenic regions length
369 and repeat content contrast with the closest species of Saturniidae and Sphingidae
370 sequenced so far. It will be interesting to use several of the numerous *Hylesia* species (more
371 than 110 species (Lemaire, 2002)) and additional Saturniidae species (Hamilton et al., 2019;
372 Rougerie et al., 2022) as models to study repetitive element dynamics, notably by
373 comparing their repeat contents, genomes sizes, genetic diversity and effective population
374 sizes. Studying other *Hylesia* species genomes is also important as several of these species
375 also result in health problems (Glasser et al., 1993; Iserhard et al., 2007; Molina, 2019;
376 Salomón et al., 2005) and/or agricultural damage (Carrillo-Sánchez et al., 1998; Fronza et al.,
377 2011). The genomic resource presented here will also be useful for future comparative
378 studies of urticating insects' genomes (Battisti et al., 2011), and for future population
379 genomic studies of *H. metabus* aiming to better understand differences in population
380 genetics and demography of this species in South America (Ciminera et al., 2019).

381

Data availability:

The nuclear and mitochondrial assemblies, and the HiFi, Omni-C and resequencing data will soon be available on NCBI under the BioProject ID PRJNA1132489. The nuclear and mitochondrial assemblies will also soon be available on the Bioinformatics BIPAA Platform, together with annotations of genes and repetitive elements, at the following address: https://bipaa.genouest.org/sp/hylesia_metabus/. Supplementary material are available at: 10.6084/m9.figshare.26197073.

Authors contributions:

Conceptualization: C. Perrier, M. Arias; Funding acquisition: C. Perrier, M. McClure, M. Arias; Biological samples: F. Bénéluz; Pictures: M. Herrera; Wet lab and sequencing: C. Perrier, W. Marrande, A. Theron, N. Rodde, L. Sauné, H. Parrinello, M. Arias; Statistical analysis and visualization: C. Perrier, R. Allio, F. Legeai, M. Gautier, W. Marrande, M. Arias; Writing of the original draft: C. Perrier; Review & editing: All the authors.

Acknowledgment:

C. Perrier acknowledges INRAE CBGP for funding Omni-C analyses. M. McClure and M. Arias acknowledge CNRS-MITI (Mission pour les Initiatives Transverses) for funding HiFi analyses. We thank Genobioinfo and GenOuest INRAE platforms for giving access to bioinformatic computing facilities, Gentyane INRAE Genomic Platform for Pacific Biosciences sequencing, Pierre Nouhaud for advice on transposable element annotation, ARS Guyane in Cayenne for discussions regarding health issues caused by *H. metabus*, and Jean-Philippe Champenois for sharing pictures.

Bibliography

Abdennur, N., & Mirny, L. A. (2020). Cooler: Scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*, *36*(1), 311–316.

Allio, R., Schomaker□Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2020). MitoFinder: Efficient automated large□scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources*, *20*(4), 892–905.

ANSES French Agency for food environmental and occupational health & safety. (2011). *Opinion of the French Agency for Food, Environmental and Occupational Health & Safety on the analysis of the risks to health and the environment related to strategies in French Guiana to combat the Hylesia metabus moth (Lepidoptera: Saturniidae), the agent responsible for "Caripito itch" dermatitis*.

Baril, T., Galbraith, J., & Hayward, A. (2024). Earl Grey: A fully automated user-friendly transposable element annotation and analysis pipeline. *Molecular Biology and Evolution*, *41*(4), msae068.

Battisti, A., Holm, G., Fagrell, B., & Larsson, S. (2011). Urticating Hairs in Arthropods: Their Nature and Medical Significance. *Annual Review of Entomology*, *56*(Volume 56, 2011), 203–220. https://doi.org/10.1146/annurev-ento-120709-144844

Bierman, A., Karsten, M., & Terblanche, J. S. (2023). Genome assembly of Thaumatotibia leucotreta, a major polyphagous pest of agriculture in sub-Saharan Africa. *G3 Genes|Genomes|Genetics*, *13*(3), jkac328. https://doi.org/10.1093/g3journal/jkac328

Bininda-Emonds, O. (2006). seqConverter. Pl, version 1. *Institut Fur Spezeille Zoologie Und Evolutionsbiologie Mit Phyletischem Museum, Friedrich-Schiller-Universitat Jena*.

Boyes, D., & Holland, P. W. H. (2023). The genome sequence of the Mother Shipton moth , Euclidia mi (Clerck, 1759). *Wellcome Open Research*, *8*, 108. https://doi.org/10.12688/wellcomeopenres.19098.1

Boyes, D., Holland, P. W. H., University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life

programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, & Darwin Tree of Life Consortium. (2023). The genome sequence of the Feathered Gothic, Tholera decimalis (Poda, 1761). *Wellcome Open Research*, *8*, 200. https://doi.org/10.12688/wellcomeopenres.19395.1

Boyes, D., of Oxford, U., Lab, W. W. G. A., of Life, W. S. I. T., Sivess, L., & Darwin Tree of Life Consortium. (2022). The genome sequence of the small elephant hawk moth, Deilephila porcellus (Linnaeus, 1758). *Wellcome Open Research*, *7*.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Cabanettes, F., & Klopp, C. (2018). D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *6*, e4958. https://doi.org/10.7717/peerj.4958

Carrillo-Sánchez, J., Equihua-Martínez, A., Sosa-Torres, C., & Fernández-Sosa, R. (1998). *The defoliator of black cherry and maize, Hylesia iola Dyar (Lepidoptera: Saturniidae), a pest of increasing importance in Tlaxcala, Mexico.*

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890.

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175.

Ciminera, M., Auger-Rozenberg, M.-A., Caron, H., Herrera, M., Scotti-Saintagne, C., Scotti, I., Tysklind, N., & Roques, A. (2019). Genetic variation and differentiation of Hylesia metabus (Lepidoptera: Saturniidae): Moths of public health importance in French Guiana and in Venezuela. *Journal of Medical Entomology*, *56*(1), 137–148.

Crowley, L. M., Baker, E., Holland, P. W., University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life Management, S. and L. team, Wellcome Sanger Institute Scientific Operations: Sequencing Operations, Wellcome Sanger Institute Tree of Life Core

Informatics team, Tree of Life Core Informatics collective, & Darwin Tree of Life Consortium. (2024). The genome sequence of the Emperor moth, Saturnia pavonia (Linnaeus, 1758). *Wellcome Open Research*, *9*, 48.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*(2), giab008.

Cequena, H., Arrivillaga, J., Sainz-Borgo, C., & Hernandez, J. (2012). Variabilidad, estructura genetica y filogenia de Hylesia metabus. In *Estudio multidisciplinario de la palometa peluda Hylesia metabus* (IVIC, Instituto Venezolano de Investigaciones Científicas, pp. 113–128).

Ciminera, M., Auger-Rozenberg, M.-A., Caron, H., Herrera, M., Scotti-Saintagne, C., Scotti, I., Tysklind, N., & Roques, A. (2019). Genetic variation and differentiation of Hylesia metabus (Lepidoptera: Saturniidae): Moths of public health importance in French Guiana and in Venezuela. *Journal of Medical Entomology*, *56*(1), 137–148.

Fasterius, E., & Al-Khalili Szigyarto, C. (2019). seqCAT: a bioconductor R-package for variant analysis of high throughput sequencing data. *F1000Research*, *7*, 1466.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457.

Fronza, E., Specht, A., & Corseuil, E. (2011). Butterflies and moths (Insecta: Lepidoptera) associated with erva-mate, the South American Holly (Ilex paraguariensis St. Hil.), in Rio Grande do Sul, Brazil. *Check List*, *7*(4), 496–504.

Fukasawa, Y., Ermini, L., Wang, H., Carty, K., & Cheung, M.-S. (2020). LongQC: A Quality Control Tool for Third Generation Sequencing Long Read Data. *G3 (Bethesda, Md.)*, *10*(4), 1193–1196. https://doi.org/10.1534/g3.119.400864

Gilbert, C., Peccoud, J., & Cordaux, R. (2021). Transposable elements and the evolution of insects. *Annual Review of Entomology*, *66*, 355–372.

Glasser, C. M., Cardoso, J. L., Carréri-Bruno, G. C., Domingos, M. de F., Moraes, R. H. P.,

& Ciaravolo, R. M. de C. (1993). Surtos epidêmicos de dermatite causada por mariposas do gênero Hylesia (Lepidóptera: Hemileucidae) no Estado de São Paulo, Brasil. *Revista de Saúde Pública*, *27*, 217–220.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talón, M., Dopazo, J., & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, *36*(10), 3420–3435. https://doi.org/10.1093/nar/gkn176

Hamilton, C. A., St Laurent, R. A., Dexter, K., Kitching, I. J., Breinholt, J. W., Zwick, A., Timmermans, M. J. T. N., Barber, J. R., & Kawahara, A. Y. (2019). Phylogenomics resolves major relationships and reveals significant diversification rate shifts in the evolution of silk moths and relatives. *BMC Evolutionary Biology*, *19*(1), 182. https://doi.org/10.1186/s12862-019-1505-1

Hernández, J. V., Osborn, F., & Conde, J. E. (2012). Estudio multidisciplinario de la palometa peluda Hylesia metabus. In *Instituto Venezolano de Investigaciones Científicas*. Caracas.

Holst, F., Bolger, A., Günther, C., Maß, J., Triesch, S., Kindel, F., Kiel, N., Saadat, N., Ebenhöh, O., & Usadel, B. (2023). Helixer–de novo prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. *BioRxiv*, 2023–02.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Iserhard, C. A., Kaminski, L. A., Marchiori, M. O., Teixeira, E. C., & Romanowski, H. P. (2007). Occurrence of lepidopterism caused by the moth Hylesia nigricans (Berg)(Lepidoptera: Saturniidae) in Rio Grande do Sul state, Brazil. *Neotropical Entomology*, *36*, 612–615.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics (Oxford, England)*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Jourdain, F., Girod, R., Vassal, J.-M., Chandre, F., Lagneau, C., Fouque, F., Guiral, D., Raude, J., & Robert, V. (2012). The moth Hylesia metabus and French Guiana lepidopterism: Centenary of a public health concern. *Parasite: Journal de La Société Française de Parasitologie*, *19*(2), 117.

Katoh, K., & Standley, D. M. (2016). A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics*, *32*(13), 1933–1942.

Laetsch, D. R., & Blaxter, M. L. (2017). BlobTools: Interrogation of genome assemblies. *F1000Research*, *6*(1287), 1287.

Lemaire, C. (2002). *Saturniidae of America: Hemileucinae*. Antiquariat Geock & Evers.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lovell, J. T., Sreedasyam, A., Schranz, M. E., Wilson, M., Carlson, J. W., Harkess, A., Emms, D., Goodstein, D. M., & Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife*, *11*, e78526. https://doi.org/10.7554/eLife.78526

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654.

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770.

Minh, B. Q., Trifinopoulos, J., Schrempf, D., Schmidt, H., & Lanfear, R. (2022). IQ-TREE

version 2.0: Tutorials and Manual Phylogenomic software by maximum likelihood. *Nucleic Acids Research*, *44*(W1), W232–W235.

Molina, U. (2019, September 27). *Se elevan los casos de alergia causados por la polilla 'Hylesia.'* https://www.prensa.com/impresa/panorama/elevan-alergia-causados-polilla-Hylesia_0_5405459491.html

Muller, H., Ogereau, D., Da Lage, J.-L., Capdevielle, C., Pollet, N., Fortuna, T., Jeannette, R., Kaiser, L., & Gilbert, C. (2021). Draft nuclear genome and complete mitogenome of the Mediterranean corn borer, Sesamia nonagrioides, a major pest of maize. *G3 Genes|Genomes|Genetics*, *11*(7), jkab155. https://doi.org/10.1093/g3journal/jkab155

Nevers, Y., Warwick Vesztrocy, A., Rossier, V., Train, C.-M., Altenhoff, A., Dessimoz, C., & Glover, N. M. (2024). Quality assessment of gene repertoire annotations with OMArk. *Nature Biotechnology*, 1–10. https://doi.org/10.1038/s41587-024-02147-w

Nishimura, O., Hara, Y., & Kuraku, S. (2017). gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics*, *33*(22), 3635–3637.

Open2C, Abdennur, N., Fudenberg, G., Flyamer, I. M., Galitsyna, A. A., Goloborodko, A., Imakaev, M., & Venev, S. V. (2023). *Pairtools: From sequencing data to chromosome contacts* (p. 2023.02.13.528389). bioRxiv. https://doi.org/10.1101/2023.02.13.528389

Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research*, *9*, ISCB Comm J-304. https://doi.org/10.12688/f1000research.23297.2

Petersen, M., Armisén, D., Gibbs, R. A., Hering, L., Khila, A., Mayer, G., Richards, S., Niehuis, O., & Misof, B. (2019). Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Ecology and Evolution*, *19*, 1–15.

Podsiadlowski, L., Tunström, K., Espeland, M., & Wheat, C. W. (2021). The Genome Assembly and Annotation of the Apollo Butterfly Parnassius apollo, a Flagship Species for Conservation Biology. *Genome Biology and Evolution*, *13*(8), evab122.

https://doi.org/10.1093/gbe/evab122

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), 1432. https://doi.org/10.1038/s41467-020-14998-3

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*(1), 460. https://doi.org/10.1186/s12859-018-2485-7

Rodriguez-Morales, A. J., Arria, M., Rojas-Mirabal, J., Borges, E., Benitez, J. A., Herrera, M., Villalobos, C., Maldonado, A., Rubio, N., & Franco-Paredes, C. (2005). Lepidopterism due to exposure to the moth Hylesia metabus in northeastern Venezuela. *American Journal of Tropical Medicine and Hygiene*, *73*(5), 991.

Rougerie, R., Cruaud, A., Arnal, P., Ballesteros-Mejia, L., Condamine, F. L., Decaëns, T., Elias, M., Gey, D., Hebert, P. D. N., Kitching, I. J., Lavergne, S., Lopez-Vaamonde, C., Murienne, J., Cuenot, Y., Nidelet, S., & Rasplus, J.-Y. (2022). *Phylogenomics Illuminates the Evolutionary History of Wild Silkmoths in Space and Time (Lepidoptera: Saturniidae)* (p. 2022.03.29.486224). bioRxiv. https://doi.org/10.1101/2022.03.29.486224

Salomón, A. D., Simón, D., Rimoldi, J. C., Villaruel, M., Pérez, O., Pérez, R., & Marchán, H. (2005). Lepidopterismo por Hylesia nigricans (mariposa negra): Investigación y acción preventiva en Buenos Aires. *Medicina (Buenos Aires)*, *65*(3), 241–246.

Serizay, J., Matthey-Doret, C., Bignaud, A., Baudry, L., & Koszul, R. (2024). Orchestrating chromosome conformation capture analysis with Bioconductor. *Nature Communications*, *15*(1), 1072. https://doi.org/10.1038/s41467-024-44761-x

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Skojec, C., Godfrey, R. K., & Kawahara, A. Y. (2024). Long read genome assembly of

Automeris io (Lepidoptera: Saturniidae) an emerging model for the evolution of deimatic displays. *G3: Genes, Genomes, Genetics*, jkad292.

Smit, A., Hubley, R., & Green, P. (2015). *RepeatMasker Open-4.0. 2013–2015*.

Sommer, M. J., Zimin, A. V., & Salzberg, S. L. (2024). PSAURON: A tool for assessing protein annotation across a broad range of species. *bioRxiv: The Preprint Server for Biology*, 2024.05.15.594385. https://doi.org/10.1101/2024.05.15.594385

Stothard, P., & Wishart, D. S. (2005). Circular genome visualization and exploration using CGView. *Bioinformatics (Oxford, England)*, *21*(4), 537–539. https://doi.org/10.1093/bioinformatics/bti054

Triant, D., & Pirro, S. (2023). The Complete Genome Sequences of 9 Species of Swallowtail Butterflies (Papilionidae, Lepidoptera). *Biodiversity Genomes*. https://doi.org/10.56179/001c.73927

Vasimuddin, Md., Misra, S., Li, H., & Aluru, S. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 314–324. https://doi.org/10.1109/IPDPS.2019.00041

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202–2204.

Wang, P.-L., Luchetti, A., Alberto Ruggieri, A., Xiong, X.-M., Xu, M.-R.-X., Zhang, X.-G., & Zhang, H.-H. (2019). Successful invasions of short internally deleted elements (SIDEs) and its partner CR1 in Lepidoptera insects. *Genome Biology and Evolution*, *11*(9), 2505–2516.

Xu, Z., & Wang, H. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, *35*(Web Server issue), W265–W268. https://doi.org/10.1093/nar/gkm286

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