



**HAL**  
open science

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

Charles Perrier, Rémi Allio, Fabrice Legeai, Mathieu Gautier, Frédéric Bénéluz, William Marande, Anthony Théron, Nathalie Rodde, Melfran Herrera, Laure Saune, et al.

## ► To cite this version:

Charles Perrier, Rémi Allio, Fabrice Legeai, Mathieu Gautier, Frédéric Bénéluz, et al.. Transposable element accumulation drives genome size increase in *Hylesia metabus* (Lepidoptera: Saturniidae), an urticating moth species from South America. *Journal of Heredity*, In press, pp.esae069. 10.1093/jhered/esae069 . hal-04792049v3

HAL Id: hal-04792049

<https://cnrs.hal.science/hal-04792049v3>

Submitted on 7 Jan 2025

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

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



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License



## Genome Resources

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

Charles Perrier<sup>1,\*</sup>, Rémi Allio<sup>1</sup>, Fabrice Legeai<sup>2,3</sup>, Mathieu Gautier<sup>1</sup>, Frédéric Bénélu<sup>4</sup>, William Marande<sup>5</sup>, Anthony Theron<sup>5</sup>, Nathalie Rodde<sup>5</sup>, Melfran Herrera<sup>6</sup>, Laure Sauné<sup>1</sup>, Hugues Parrinello<sup>7</sup>, Melanie McClure<sup>8</sup> and Mónica Arias<sup>9,10</sup>

<sup>1</sup>UMR CBGP, INRAE, CIRAD, IRD, Institut Agro, Université Montpellier, Montpellier, France,

<sup>2</sup>BIPAA IGEPP, INRAE Institut Agro Univ Rennes Rennes France,

<sup>3</sup>Univ Rennes, Inria, CNRS, IRISA, 35000 Rennes, France,

<sup>4</sup>Société entomologique Antilles-Guyane, SEAG, Rémire-Montjoly, France,

<sup>5</sup>French Plant Genomic Resource Center, CNRGV, INRAE, Castanet Tolosan, France,

<sup>6</sup>Coordinación de Vigilancia Epidemiológica Ambiental, Dirección Estatal de Salud Ambiental, FUNDASALUD, Estado Sucre, Venezuela,

<sup>7</sup>MGX-Montpellier GenomiX, Université Montpellier, CNRS, INSERM, Montpellier, France,

<sup>8</sup>Laboratoire Écologie, Évolution, Interactions des Systèmes Amazoniens (LEEISA), Université de Guyane, CNRS, IFREMER, Cayenne, France,

<sup>9</sup>CIRAD, UMR PHIM, F-34398, Montpellier, France

<sup>10</sup>UMR PHIM, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France

\*Corresponding author: Charles Perrier, UMR CBGP, INRAE, CIRAD, IRD, Institut Agro, Université Montpellier, Montpellier, France. Email: [charles.perrier@inrae.fr](mailto:charles.perrier@inrae.fr)

Corresponding Editor: Warren Booth

## 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 with 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.

**Key words:** ashen moth, lepidopterism, population outbreaks, repetitive elements, yellowtail moth

## Introduction

The yellowtail moth *Hylesia metabus* (Saturniidae, Lepidoptera, Fig. 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 (Rodríguez-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 (Hernández et al. 2012; Jourdain et al. 2012; Ciminera

et al. 2019). 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 (Rodríguez-Morales et al. 2005; Jourdain et al. 2012). Using mitochondrial markers and

Received July 19, 2024; Accepted November 14, 2024

© The American Genetic Association. 2024.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [reprints@oup.com](mailto:reprints@oup.com) for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com).



**Fig. 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 males (lower right). (Photos of nest and larvae ©Jean-Philippe Champenois).

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.

## 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  $-80^{\circ}\text{C}$ . Each adult was stored in ethanol  $85^{\circ}$  and stored at  $-20^{\circ}\text{C}$ .

### 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, BELGIUM) at an average size of 20 kb. After End repair, A-tailing and ligation of SMRTbell adapter, the library was selected on BluePippin System (Sage Science, MA, USA) for a range size of 10–50 kb. The size and concentration of the library were assessed using the Agilent FP-1002 Genomic DNA 165 kb on the Femto Pulse system and the Qubit dsDNA HS reagents Assay kit. Sequencing primer v5 and Sequel® II DNA Polymerase 2.2 were annealed and bound, respectively, to the SMRTbell library. The library was loaded on one SMRTcell 8M at an on-plate concentration of 90pM. Sequencing was performed on the Sequel® II system at Gentyane Genomic Platform (INRAE

Clermont-Ferrand, France) with Sequel® II Sequencing kit 3.0, a run movie time of 30 hours with an Adaptive Loading target (P1 + P2) at 0.75. After filtering and correcting the SMRTcell output, we obtained 2,117,541 reads totaling 42.7 Gb, with an N50 of 20,851 bp measured with LongQC v1.2.1 (Fukasawa et al. 2020).

The Omni-C library (Dovetail Genomics®) was produced according to the manufacturer instructions. In brief, this consisted of 33mg of frozen *H. metabus* abdomen of the second larva, ground in liquid nitrogen and suspended in PBS. Then the DNA was fixed with formaldehyde and digested using 2µl of a nuclease enzyme mix. After binding 500 ng of the digested DNA to chromatin capture beads, a proximity ligation was performed and the crosslinks were reversed to produce the linked DNA. Finally, 107 ng of the linked DNA was used to produce a library then paired-end (2 × 150 bp) sequenced on three different S4 flow-cell lanes on an Illumina® NovaSeq system. The obtained raw paired-end reads were filtered using fastp v0.23.2 (Chen et al. 2018) run with default options leading to a total of 171 millions of read pairs (51.2 Gb).

Two whole genome libraries of one adult female and one adult male were constructed using truseq kit to produce illumina short reads. In brief, DNA from one adult female and one adult male was extracted using the blood and tissue kit, column style, from Qiagen (Qiagen, MD, USA). DNA concentration was estimated using both Qubit fluorometric measures and nanodrop absorbance measures. Libraries were then constructed using the TruSeq nano DNA kit from Illumina and paired-end (2 × 150 bp) sequenced on a S4 flow-cell lanes on an Illumina® NovaSeq system, producing 234 millions read pairs for the male and 325 million read pairs for the female after filtering with fastp v0.23.2.

### Nuclear genome assembly, filtering and scaffolding

We used Jellyfish (Marçais and Kingsford 2011) and GenomeScope (Vurture et al. 2017; Ranallo-Benavidez et al. 2020) to estimate the genome size from HiFi data. A whole genome assembly was then built from HiFi reads HiFiasm v0.16.1 (Cheng et al. 2021) ran with default options.

We detected and filtered out the mitogenome, potential contaminants and haplotigs as follows. Using MitoFinder v1.4.2 (Allio et al. 2020) on the primary assembly, we detected and annotated a contig corresponding to the complete mitochondrial genome sequence of *H. metabus*. The mitochondrial genome was removed from the nuclear genome assembly and we drew a circular representation using CGView (Stothard and Wishart 2005). We used blobtools (Laetsch and Blaxter 2017) to filter out potential contaminants. We also searched for potential contaminant sequences originating from the plant species' genome on which larvae were found feeding on (*Tapirira guianensis*) by mapping short reads from this species (https://www.ebi.ac.uk/ena/browser/view/ERR7620141) on our putative *H. metabus* scaffolds using bwa-mem2 v2.2.1 (Vasimuddin et al. 2019) and samtools v1.10 (Danecek et al. 2021). We used purge haplotigs v1.1.2 (Roach et al. 2018) to remove potential haplotigs and small contigs exhibiting bad mapping quality or that could be considered as junk or 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 and 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 *agg* 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, Simão et al. 2015; Manni et al. 2021) for “arthropoda\_odb10” and “lepidoptera\_odb10” databases. Blobtools (Laetsch and 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 lepidoptera\_odb10, PSAURON 1.0.2 (Sommer et al. 2024) and OMArk 2023.10 (Nevers et al. 2024). Functional annotation of the protein sequences obtained with GFFread (Perteau and Perteau 2020) from the Helixer output were done with Diamond v2.0.13 (Buchfink et al. 2015) on NCBI NR 2022-12-11, Blast2GO Command Line v1.5.1 (Götz et al. 2008), eggNOG v2.1.9 (Huerta-Cepas et al. 2019) with eggnog database v5.0.2 and Interproscan v5.59-91.0 (Jones et al. 2014). The genome sequence and its annotations can be browsed at https://bipaa.genouest.org/sp/hylesia\_metabus/.

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

### Phylogenetic and comparative analyses

Phylogenetic analyses have been done by comparing the BUSCO sequences from the genome of *H. metabus* to the one of 4 other Saturniidae species (*Automeris io*, *Samia ricini*, *Antheraea yamamai*, *Saturnia pavonia*), 12 Sphingidae species (*Cephonodes hylas*, *Hemaris fuciformis*, *Hyles euphorbiae*, *Deilephila porcellus*, *Theretra japonica*, *Manduca sexta*, *Lapara coniferarum*, *Sphinx pinastri*, *Clanis bilineata*, *Laothoe populi*, *Amorpha juglandis*, *Mimas tiliae*), and *Bombyx mori* as outgroup. External assemblies were downloaded from NCBI using their respective accession number with the

command-line tool “datasets” (e.g. datasets download genome accession GCF\_030269925.1—filename *Bombyx\_mori*.zip; see [Supplementary Material S1](#) for accessions). For each assembly, BUSCO sequences were annotated and extracted based on the predefined dataset “Lepidoptera\_odb10,” which includes 5,286 orthologous genes for Lepidoptera. BUSCO search was performed through the gVolante web server ([Nishimura et al. 2017](#)). Amino acid sequences corresponding to every BUSCO marker were first individually aligned with the MAFFT ([Katoh and Standley 2016](#)) algorithm FFT-NS-2. All markers were then concatenated in one supermatrix using seqCat.pl and seqConverter.pl ([Bininda-Emonds 2006](#); [Fasterius and Al-Khalili Szigyarto 2019](#)). Phylogenetic inferences were performed with Maximum-likelihood (ML) as implemented in IQ-TREE V2.2.2.6 ([Minh et al. 2022](#)). One partition per BUSCO sequence was defined with the option “-spp.” The best evolutionary model was selected for each partition using ModelFinder implemented in IQ-TREE and some partitions were merged if necessary (-m MFP + MERGE). Following the recommendation of IQ-TREE developers, we also set a smaller perturbation strength (-pers 0.2) and a larger number of stop iterations (-nstop 500) to avoid local optima. Finally, node supports were evaluated with UltraFast Bootstraps (UFBS) estimated by IQ-TREE (-bb 1000). UFBS are considered robust when higher than 95%.

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

Synteny between the genome of *H. metabus* and the genomes of *Antheraea yamamai*, *Saturnia pavonia*, *Deilephila porcellus*, *Manduca sexta*, *Laothoe populi*, and *Bombyx mori*, extracted from NCBI (see above and [Supplementary Table S1](#)) were inspected with genespace ([Lovell et al. 2022](#)) using the gene positions derived from the GFF files, and their corresponding orthogroups predicted with orthofinder v2.5.5 ([Emms and Kelly 2019](#)). Because protein-coding genes annotation of *Antheraea yamamai*, *Saturnia pavonia*, *Deilephila porcellus*, and *Laothoe populi* were not available at NCBI, we produced their respective annotations with Helixer (v0.3.3, with the options—lineage invertebrate—subsequence-length 108,000—overlap-offset 54,000—overlap-core-length 81,000). Finally, the sizes of intergenic and genic regions were calculated from the GFF file with a custom script.

## Identification of the Z scaffold

In order to identify the scaffold corresponding to the Z chromosome in *H. metabus*, we first inspected synteny graphs with *Deilephila porcellus* and *Saturnia pavonia* assemblies for which the Z scaffold was known. We also used Dgenies ([Cabanettes and Klopp 2018](#)) to examine more precisely the synteny between *Hylesia metabus* and *Deilephila porcellus* ([Boyes et al. 2022](#)). Second, we investigated potential coverage variation among scaffolds between the resequenced male and female. To do so, we first used Fastp ([Chen et al. 2018](#)) to keep only good quality sequences that we then aligned to the *H. metabus* assembly using bwa-mem2 v2.2.1 ([Vasimuddin et al. 2019](#)). We then used samtools to sort and index reads and to estimate coverage per scaffold for each individual ([Danecek et al. 2021](#)). For each scaffold we measured the female to male ratio of percentage of reads mapped to each scaffold. A ratio of approximately 50% would indicate the Z

scaffolds, whereas ratios of 100% would indicate autosomal scaffolds. Finally, in order to determine the genotypic sex of the individual larvae sequenced to assemble the genome, we investigated potential variations of HiFi raw reads coverage between the putative Z scaffold and the other large scaffolds. A coverage deficit of about 50% on the Z scaffold would illustrate that the sequenced individual was a female.

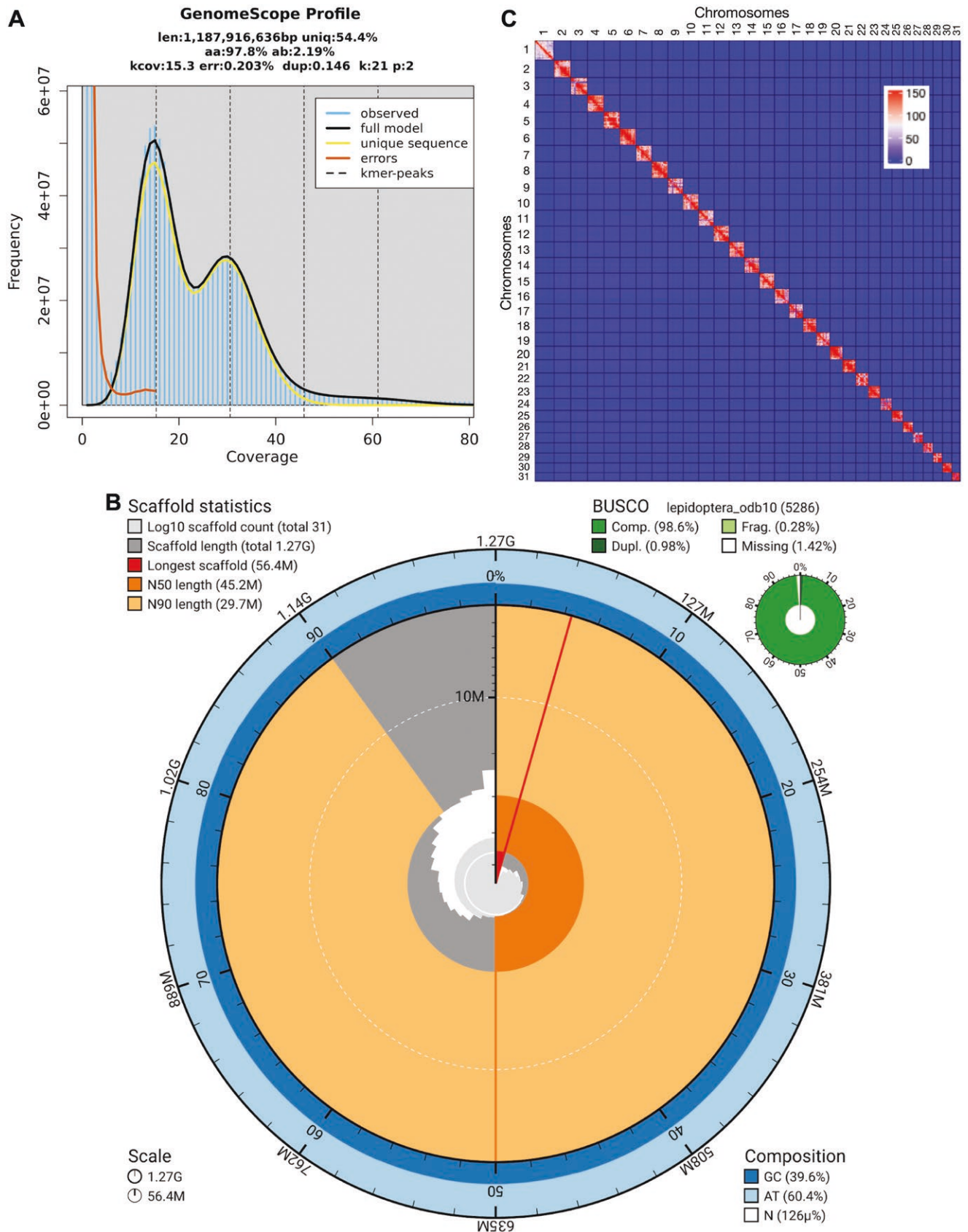
## Results and discussion

### Assembly of a 1.27 GB long genome scaffolded in 31 pseudo chromosomes

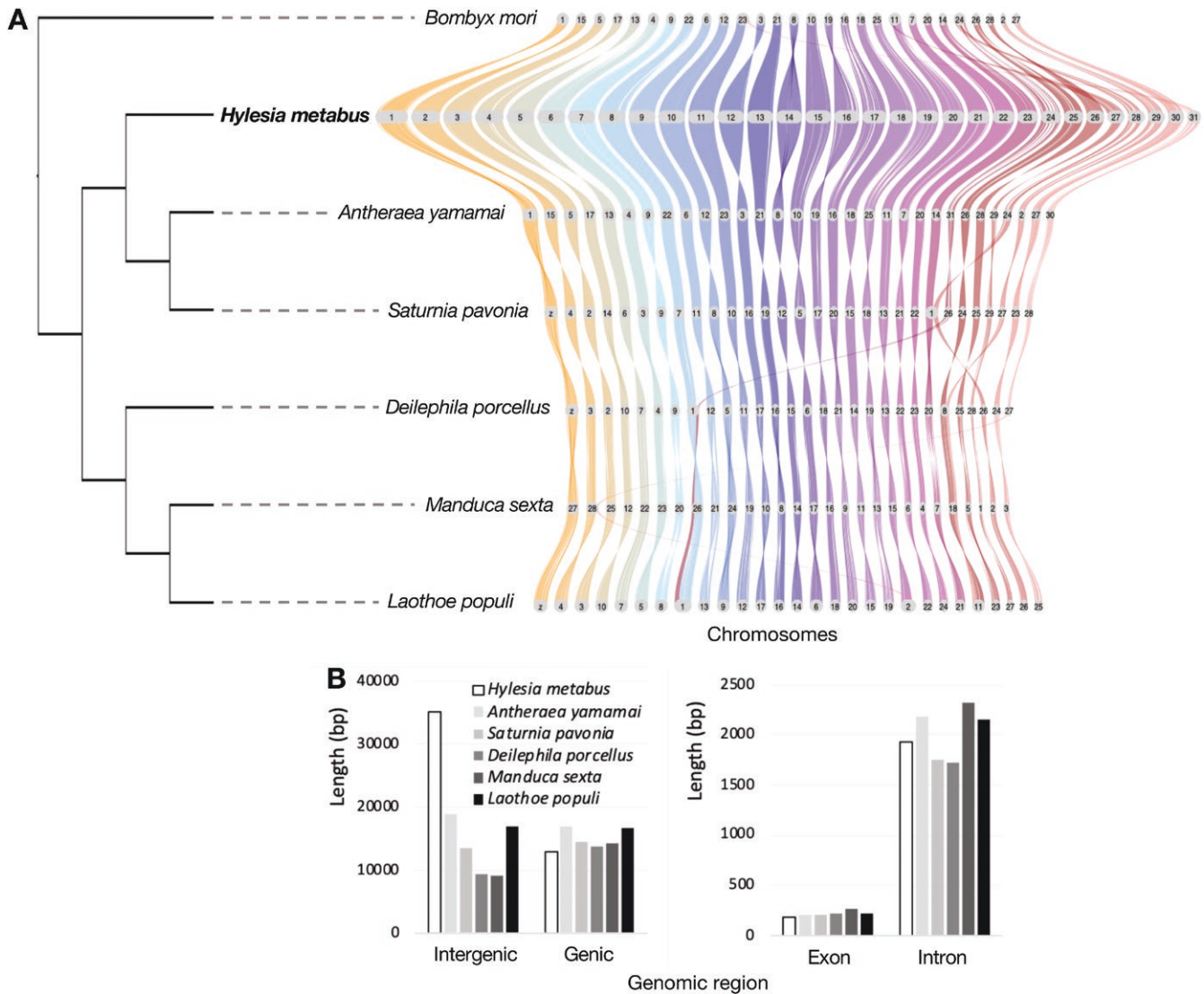
The Jellyfish and Genomescope analysis of the HiFi reads suggested that the genome size was 1.19 Gb with a heterozygosity of 2.19% ([Fig. 2A](#)). The Hifiasm assembly consisted of 171 contigs ranging from 15 kb to 57 Mb and totalizing 1.41 Gb, with an N50 of 34,63 Mb, an L50 of 17 and 39.85% of GC ([Supplementary Material S2](#)). MitoFinder identified the smallest (15,393 bp) and most covered (174 X) contig, as being the complete mitochondrial genome sequence and annotated it fully ([Supplementary Material S3](#)). The final assembly, after decontamination, purge of haplotigs, removal of the mitochondrial contig, and scaffolding with Omni-C, consisted in 31 scaffolds totalizing 1.27 Gb, with an N50 of 45,18 Mb, an L50 of 14 and 39.58% of GC ([Fig. 2B](#), [Supplementary Material S4](#)). The contact map of the primary assembly showed very high contiguity even before scaffolding ([Supplementary Material S5](#)), enabling an efficient scaffolding toward a chromosome level ([Fig. 2C](#)). The haploid read depth was on average 14 X, slightly lower than the targeted read depth, as a consequence of a larger genome size than expected. Arthropoda BUSCO gene representation of the final assembly was 99.5% complete with 98.7% single-copy genes, and the Lepidoptera BUSCO gene representation was 98.6% complete with 97.6% single-copy genes ([Supplementary Material S4](#), [Fig. 2B](#)), with less than 1% of duplicated genes.

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

Synteny was highly conserved between *H. metabus* and phylogenetically close species ([Fig. 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. 2024b](#)). Yet, other lepidoptera species are known to have large genomes, notably *Euclidia mi* 2.32 GB ([Boyes et al. 2023a](#)), *Parnassius behrii* 1.59 GB (GCA\_036936625.1), *Parnassius apollo* 1.4 GB ([Podsiadlowski et al. 2021](#)), *Tholera decimialis* 1.33 GB ([Boyes et al. 2023b](#)), *Thaumatotibia leucotreta*, 1.28 GB ([Bierman et al. 2023](#)), *Graphium colonna* 1.27 GB ([Triant and 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



**Fig. 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).



**Fig. 3.** A) Phylogeny and synteny for *Hylesia metabus* and two other Saturniidae species, three Sphingidae species, and *Bombyx mori*. B) Length of intergenic and genic regions, and of exons and introns, for *H. metabus* and two other Saturniidae species, three Sphingidae species, and *Bombyx mori*.

larger intergenic regions in *H. metabus* compared with phylogenetically close species but comparable sizes of genic regions (Fig. 3B), both for introns and exons, and in a regular manner across chromosomes (Supplementary Material S6).

### Invasion of repetitive elements

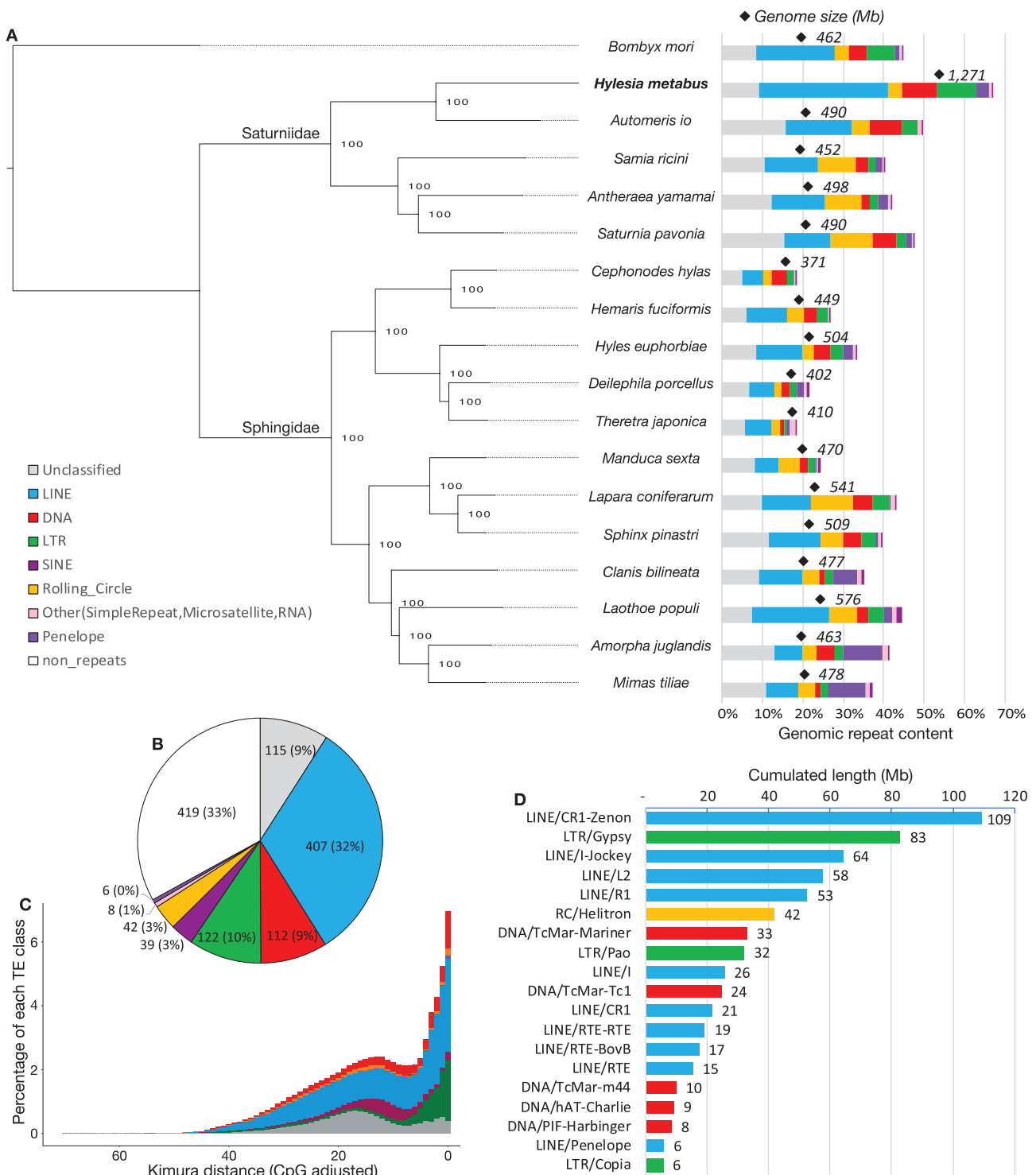
Analysis of repetitive elements was achieved with the fully automated EarlGrey pipeline and determined that 67% of the *H. metabus* genome sequence was made-up of repeats (Fig. 4A and 4B, Supplementary Materials S1 and S4). Compared with other species of Saturniidae and species of Sphingidae, *H. metabus* had a higher proportion of repetitive elements (Fig. 4A, Supplementary Materials S1 and S4), 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 (Petersen et al. 2019; Gilbert et al. 2021; Muller et al. 2021).

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 S7A). Moreover, intergenic regions consisted of repeats (68.1%) more than introns (60.4%), UTR (25.8%), and CDS (28.0%; Supplementary Material S7B). 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*.

The majority of repetitive elements identified in *H. metabus* were of type LINE (32% of the genome, hence 55% of the classified repeats; Fig. 4B), followed by LTR, DNA and SINE elements (respectively, 10%, 9%, and 3% of the genome, hence 17%, 16%, and 5% of the classified repeats). Only 9% of the genome was made of unclassified repetitive elements. These LINES and LTRs proportions are higher than for other species of Saturniidae and species of Sphingidae (Fig. 4A, Supplementary Material S1). This higher proportion of LINES and LTRs is comparable to *Parnassius apollo*.

The repeat landscape of *H. metabus* suggests that repetitive elements, especially of type LINE, invaded the genome relatively regularly over time, but that LTRs had a recent burst of invasion (Fig. 4C). While comparing the repeat landscape of *H. metabus* to the 4 other species of Saturniidae and 12



**Fig. 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*.

species of Sphingidae (Supplementary Material S8), we found that repeat landscapes were in general relatively similar between close species and dissimilar between more distant ones. For example, while the repeat landscape of *H. metabus* was very similar to the one of its closest species, *Automeris io*, it was very dissimilar to the repeats landscapes of the three other Saturniidae species *Samia ricini*, *Antheraea yamamai*

and *Saturnia pavonia* which all showed a more ancient burst of invasion and a higher accumulation of Rolling circles. Regarding the timing of accumulation of these transposable elements, under uncorrelated relaxed clock model, Rougerie et al. (2022) estimated the origin of the crown group *Hylesia* at 10 to 13 MY (unfortunately *H. metabus* was not included in their study), the divergence time between the genera



*Hylesia* and *Automeris* at about 25 MY and about 46 MY between *Hylesia* and the genera *Samia*, *Antheraea* and *Saturnia*. Skojec et al. (2024a) study found similar divergence times (although showing a 18MY divergence between *Automeris* and *io* clades). This suggests that the similar repeat landscapes of *H. metabus* and *Automeris io* might have evolved between 46 and 25 MY, and that *H. metabus* accumulation of TE might have occurred during the last 25 MY, and perhaps more likely during the *Hylesia* genus diversification between 10 and 13 MY or during the more recent evolution of *H. metabus*. Sequencing genomes for more Saturniidae species, especially from the genus *Hylesia*, would enable more detailed comparative analyses of the accumulation of transposable elements in this family.

The five most abundant repetitive elements were LINE/CR1-Zenon (109 M bp—319,358 copies, Fig. 4D, Supplementary Material S9), LTR/Gypsy (83M bp), LINE/I-Jockey (64M bp), LINE/L2 (58M bp) and LINE/R1 (42M bp). For the four other Saturniidae species considered in this study (Supplementary Material S10), LINE/R1 was the most abundant family in *Automeris io*, followed by RC/Helitron and LINE/CR1-Zenon, and RC/Helitron was the most abundant family in the three other species, followed by LINE/I-Jockey in two species and by LINE/L2 and LINE/R1 in the last one. This more detailed analysis therefore shows that the TE invasion of *H. metabus* genome is primarily driven by a few TE families that are also among the most abundant ones in the closest species genome sequences. In particular, the LINE CR1 *Zenon* has been shown to successfully invade other lepidopteran species' genomes (Wang et al. 2019).

### Gene prediction, Z scaffold and data accessibility

Gene prediction with Helixer identified 26,122 transcripts on the non-masked version of the genome (Supplementary Material S4). The BUSCO score for the annotated genes was 88.9% complete (Supplementary Material S4). Comparatively OMArk identified 6,361 complete proteins among the 6,779 Obctomera HOG, including 1,466 duplicated, in the range of other closely related species. OMArk also determined that 77.25% of the protein sequences were placed at a consistent lineage, whereas 16.31% were unknown and reported no contamination. The psauron score, reflecting the likelihood of being a genuine protein-coding sequence, was 96.9. Generating RNAseq data in order to further annotate this genome would probably greatly improve the precision of the annotation, by keeping only transcripts with evidence of expression.

The synteny inspection showed that the largest *H. metabus* scaffold (scaffold\_1) corresponded to the Z scaffold in *Deilephila porcellus* and *Saturnia pavonia* assemblies (Fig. 3A, Supplementary Material S11). This was confirmed by the comparisons of read mapping between one male and one female (that had average coverage of 16X and 25X, respectively), revealing a 0.54 female to male relative ratio of coverage on scaffold 1 (Supplementary Material S12). Finally, mapping HiFi reads of the assembled individual on the final assembly, we estimated that Scaffold\_1 had on average 57% lower read depth than the other 30 largest scaffolds, further confirming that scaffold\_1 indeed corresponds to the Z chromosome and suggesting that the individual used to assemble the genome was a female.

## Conclusion

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

## Supplementary Material

Supplementary material can be found at <http://www.jhered.oxfordjournals.org/>.

## Funding

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 Nouhau 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.

## 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.

## 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/](https://bipaa.genouest.org/sp/hylesia_metabus/). Supplementary material are available at: 10.6084/m9.figshare.26197073.

## Bibliography

- Open2C, Abdennur N, Fudenberg G, Flyamer IM, Galitsyna AA, Goloborodko, A, Imakaev M, Venev SV. 2023. *Pairtools: From sequencing data to chromosome contacts* (p. 2023.02.13.528389). bioRxiv. <https://doi.org/10.1101/2023.02.13.528389>
- Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*. 2020;36:311–316.
- Allio R, Schomaker-Bastos A, Romiguier J, Prodocimi F, Nabholz B, Delsuc F. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol Ecol Resour*. 2020;20:892–905.
- 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. Maison-Alfort, Île-de-France, France.
- Baril T, Galbraith J, Hayward A. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. *Mol Biol Evol*. 2024;41:msae068.
- Battisti A, Holm G, Fagrell B, Larsson S. Urticating hairs in arthropods: their nature and medical significance. *Annu Rev Entomol*. 2011;56:203–220. doi:10.1146/annurev-ento-120709-144844
- Bierman A, Karsten M, Terblanche JS. Genome assembly of *Thaumotobia leucotreta*, a major polyphagous pest of agriculture in sub-Saharan Africa. *G3 (Bethesda, Md.)* 2023;13:jkac328. doi:10.1093/g3journal/jkac328
- Bininda-Emonds O. seqConverter. Pl, version 1. Friedrich-Schiller-Universität Jena; 2006.
- Boyes D, Holland PWH; University of Oxford and Wytham Woods Genome Acquisition Lab. The genome sequence of the Mother Shipton moth, *Euclidia mi* (Clerck, 1759). *Wellcome Open Res*. 2023a;8:108. doi:10.12688/wellcomeopenres.19098.1
- Boyes D, Holland PWH; University of Oxford and Wytham Woods Genome Acquisition Lab. The genome sequence of the Feathered Gothic, *Tholera decimalis* (Poda, 1761). *Wellcome Open Res*. 2023b;8:200. doi:10.12688/wellcomeopenres.19395.1
- Boyes D, Lewin T; University of Oxford and Wytham Woods Genome Acquisition Lab. The genome sequence of the small elephant hawk moth, *Deilephila porcellus* (Linnaeus, 1758). *Wellcome Open Res*. 2022;7:258.
- Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;12:59–60. doi:10.1038/nmeth.3176
- Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*. 2018;6:e4958. doi:10.7717/peerj.4958
- Carrillo-Sánchez J, Equihua-Martínez A, Sosa-Torres C, Fernández-Sosa R. The defoliator of black cherry and maize, *Hylesia iola* Dyar (Lepidoptera: Saturniidae), a pest of increasing importance in Tlaxcala, Mexico. *Folia Entomológica Mexicana*, 1998;103:99–100.
- Cequena H, Arrivillaga J, Sainz-Borgo C, Hernandez J. Variabilidad, estructura genética y filogenia de *Hylesia metabus*. In: Estudio multidisciplinario de la palometa peluda *Hylesia metabus*. Caracas, Venezuela: IVIC, Instituto Venezolano de Investigaciones Científicas; 2012. p. 113–128.
- Chacón J. 2019. *Alerta: aumentan los casos de alergia causados por pelusa de la polilla ‘Hylesia.’ Panamá América*. [accessed 2019 Oct 15] <https://www.panamaamerica.com.pa/sociedad/alerta-aumentan-los-casos-de-alergia-causada-por-pelusa-de-la-polilla-hylesia-1145024>
- Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34:i884–i890.
- Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;18:170–175.
- Ciminera M, Auger-Rozenberg M-A, Caron H, Herrera M, Scotti-Saintagne C, Scotti I, Tysklind N, Roques A. Genetic variation and differentiation of *Hylesia metabus* (Lepidoptera: Saturniidae): moths of public health importance in French Guiana and in Venezuela. *J Med Entomol*. 2019;56:137–148.
- Crowley LM, Baker E, Holland PW; 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. The genome sequence of the Emperor moth, *Saturnia pavonia* (Linnaeus, 1758). *Wellcome Open Res*. 2024;9:48.
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10:giab008.
- Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* 2019;20(1):238. doi:10.1186/s13059-019-1832-y.
- Fasterius E, Al-Khalili Szigyarto C. seqCAT: a bioconductor R-package for variant analysis of high throughput sequencing data. *F1000Research* 2019;7:1466.
- 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 USA*. 2020;117:9451–9457.
- Fronza E, Specht A, Corseuil E. 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* 2011;7:496–504.
- Fukasawa Y, Ermini L, Wang H, Carty K, Cheung M-S. LongQC: a quality control tool for third generation sequencing long read data. *G3 (Bethesda, Md.)* 2020;10:1193–1196. doi:10.1534/g3.119.400864
- Gilbert C, Peccoud J, Cordaux R. Transposable elements and the evolution of insects. *Annu Rev Entomol*. 2021;66:355–372.
- Glasser CM, Cardoso JL, Carréri-Bruno GC, Domingos M. de F, Moraes RHP, Ciaravolo R. M. de C. Surtos epidêmicos de dermatite causada por mariposas do gênero *Hylesia* (Lepidoptera: Hemileucidae) no Estado de São Paulo, Brasil. *Revista de Saúde Pública* 1993;27:217–220.
- Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*. 2008;36:3420–3435. doi:10.1093/nar/gkn176
- Hamilton CA, St Laurent RA, Dexter K, Kitching IJ, Breinholt JW, Zwick A, Timmermans MJTN, Barber JR, Kawahara AY. Phylogenomics resolves major relationships and reveals significant diversification rate shifts in the evolution of silk moths and relatives. *BMC Evol Biol*. 2019;19:182. doi:10.1186/s12862-019-1505-1
- Hernández JV, Osborn F, Conde JE (Eds.). Estudio multidisciplinario de la palometa peluda *Hylesia metabus*. Caracas, Venezuela: Edition IVIC, Instituto Venezolano de Investigaciones Científicas (IVIC); 2012.
- Holst F, Bolger A, Günther C, Maß J, Triesch S, Kindel F, Kiel N, Saadat N, Ebenhöf O, Usadel B. Helixer—de novo prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. *BioRxiv* 2023:2023–2002.
- Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattai T, Jensen LJ, et al.

- eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019;47:D309–D314. doi: [10.1093/nar/gky1085](https://doi.org/10.1093/nar/gky1085)
- Iserhard CA, Kaminski LA, Marchiori MO, Teixeira EC, Romanowski HP. Occurrence of lepidopterism caused by the moth *Hylesia nigricans* (Berg)(Lepidoptera: Saturniidae) in Rio Grande do Sul state, Brazil. *Neotrop Entomol.* 2007;36:612–615.
- Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* (Oxford, England) 2014;30:1236–1240. doi: [10.1093/bioinformatics/btu031](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. The moth *Hylesia metabus* and French Guiana lepidopterism: centenary of a public health concern. *Parasite* (Paris, France) 2012;19:117–128.
- Katoh K, Standley DM. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics.* 2016;32:1933–1942.
- Laetsch DR, Blaxter ML. BlobTools: interrogation of genome assemblies. *F1000Research* 2017;6:1287.
- Lemaire C. The Saturniidae of America–Hemileucinae. Keltern, Germany: Antiquariat Geock & Evers; 2002.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25:2078–2079. doi: [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352)
- Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, Emms D, Goodstein DM, Schmutz J. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife* 2022;11:e78526. doi: [10.7554/eLife.78526](https://doi.org/10.7554/eLife.78526)
- 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;38:4647–4654.
- Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* 2011;27:764–770.
- Minh BQ, Trifinopoulos J, Schrempf D, Schmidt H, Lanfear R. IQ-TREE version 2.0: tutorials and manual phylogenomic software by maximum likelihood. *Nucleic Acids Res.* 2022;44:W232–W235.
- Muller H, Ogereau D, Da Lage J-L, Capdevielle C, Pollet N, Fortuna T, Jeannette R, Kaiser L, Gilbert C. Draft nuclear genome and complete mitogenome of the Mediterranean corn borer, *Sesamia nonagrioides*, a major pest of maize. *G3* (Bethesda, MD) 2021;11:jkab155. doi: [10.1093/g3journal/jkab155](https://doi.org/10.1093/g3journal/jkab155)
- Nevers Y, Warwick Vesztrocy A, Rossier V, Train C-M, Altenhoff A, Dessimoz C, Glover NM. Quality assessment of gene repertoire annotations with OMArk. *Nat Biotechnol.* 2024;1:1–10. doi: [10.1038/s41587-024-02147-w](https://doi.org/10.1038/s41587-024-02147-w)
- Nishimura O, Hara Y, Kuraku S. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics.* 2017;33:3635–3637.
- Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. *F1000Research* 2020;9:ISCB Comm J–ISCB Com304. doi: [10.12688/f1000research.23297.2](https://doi.org/10.12688/f1000research.23297.2)
- Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, Misof B. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Ecol Evol.* 2019;19:1–15.
- Podsiadlowski L, Tunström K, Espeland M, Wheat CW. The genome assembly and annotation of the Apollo butterfly *Parnassius apollo*, a flagship species for conservation biology. *Genome Biol Evol.* 2021;13:evab122. doi: [10.1093/gbe/evab122](https://doi.org/10.1093/gbe/evab122)
- Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun.* 2020;11:1432. doi: [10.1038/s41467-020-14998-3](https://doi.org/10.1038/s41467-020-14998-3)
- Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinf.* 2018;19:460. doi: [10.1186/s12859-018-2485-7](https://doi.org/10.1186/s12859-018-2485-7)
- Rodriguez-Morales AJ, Arria M, Rojas-Mirabal J, Borges E, Benitez JA, Herrera M, Villalobos C, Maldonado A, Rubio N, Franco-Paredes C. Lepidopterism due to exposure to the moth *Hylesia metabus* in northeastern Venezuela. *Am J Trop Med Hyg.* 2005;73:991–993.
- Rougerie R, Cruaud A, Arnal P, Ballesteros-Mejia L, Condamine FL, Decaëns T, Elias M, Gey D, Hebert PDN, Kitching IJ, et al. *Phylogenomics Illuminates the Evolutionary History of Wild Silkmoths in Space and Time (Lepidoptera: Saturniidae)* (p. 2022.03.29.486224). 2022. bioRxiv. <https://doi.org/10.1101/2022.03.29.486224>
- Salomón AD, Simón D, Rimoldi JC, Villaruel M, Pérez O, Pérez R, Marchán H. Lepidopterismo por *Hylesia nigricans* (mariposa negra): Investigación y acción preventiva en Buenos Aires. *Medicina* (Buenos Aires) 2005;65:241–246.
- Serizay J, Matthey-Doret C, Bignaud A, Baudry L, Koszul R. Orchestrating chromosome conformation capture analysis with bioconductor. *Nat Commun.* 2024;15:1072. doi: [10.1038/s41467-024-44761-x](https://doi.org/10.1038/s41467-024-44761-x)
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* (Oxford, England) 2015;31:3210–3212. doi: [10.1093/bioinformatics/btv351](https://doi.org/10.1093/bioinformatics/btv351)
- Skojec C, Earl C, Couch CD, Masonick P, Kawahara AY. Phylogeny and divergence time estimation of Io moths and relatives (Lepidoptera: Saturniidae: Automeris). *PeerJ.* 2024a;12:e17365. doi: [10.7717/peerj.17365](https://doi.org/10.7717/peerj.17365)
- Skojec C, Godfrey RK, Kawahara AY. Long read genome assembly of *Automeris io* (Lepidoptera: Saturniidae) an emerging model for the evolution of deimatic displays. *G3 Genes Genomes Genet.* 2024b;14:jkad292.
- Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. 2015. <http://www.repeatmasker.org>.
- Sommer MJ, Zimin AV, Salzberg SL. PSAURON: a tool for assessing protein annotation across a broad range of species. *bioRxiv: The Preprint Server for Biology* 2024. doi: [10.1101/2024.05.15.594385](https://doi.org/10.1101/2024.05.15.594385)
- Stothard P, Wishart DS. Circular genome visualization and exploration using CGView. *Bioinformatics* (Oxford, England) 2005;21:537–539. doi: [10.1093/bioinformatics/bti054](https://doi.org/10.1093/bioinformatics/bti054)
- Triant D, Pirro S. The Complete Genome Sequences of 9 Species of Swallowtail Butterflies (Papilionidae, Lepidoptera). *Biodivers Genomes.* 2023; doi: [10.56179/001c.73927](https://doi.org/10.56179/001c.73927)
- Vasimuddin Md, Misra S, Li H, Aluru S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. *Rio de Janeiro, Brazil: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS);* 2019. p. 314–324. doi: [10.1109/IPDPS.2019.00041](https://doi.org/10.1109/IPDPS.2019.00041)
- Vurtture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics.* 2017;33:2202–2204.
- Wang P-L, Luchetti A, Alberto Ruggieri A, Xiong X-M, Xu M-R-X, Zhang X-G, Zhang H-H. Successful invasions of short internally deleted elements (SIDES) and its partner CR1 in Lepidoptera insects. *Genome Biol Evol.* 2019;11:2505–2516.
- Xu Z, Wang H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 2007;35:W265–W268. doi: [10.1093/nar/gkm286](https://doi.org/10.1093/nar/gkm286)
- Zhou C, McCarthy SA, Durbin R. YaHS: Yet another Hi-C scaffolding tool. *Bioinformatics.* 2023;39:btac808. doi: [10.1093/bioinformatics/btac808](https://doi.org/10.1093/bioinformatics/btac808)