Title: *De-novo* long-read genome assembly and annotation of the luna moth (*Actias luna*) fully resolves repeat-rich silk genes

Authors and Affiliations:

*Author for Correspondence: Amanda Markee [1]McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville FL, 32611, USA. amanda.markee@ufl.edu

R. Keating Godfrey [2]Department of Biological Sciences, Florida International University, Miami FL, 33199, USA. rkeating@fiu.edu

Paul B. Frandsen [4]Department of Plant and Wildlife Sciences, Brigham Young University, Provo UT, 84602, USA. paul_frandsen@byu.edu

Yi-Ming Weng [1]McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville FL, 32611, USA. yimingweng@ufl.edu

Deborah A. Triant [3]Department of Biochemistry & Molecular Genetics, University of Virginia, Charlottesville VA 22908. dtriant@virginia.edu

Akito Y. Kawahara [1]McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville FL, 32611, USA. kawahara@flmnh.ufl.edu

Abstract

We present the first long-read *de-novo* assembly and annotation of the luna moth (*Actias luna*) and provide the full characterization of *heavy chain fibroin* (*h-fibroin*), a long and highly repetitive gene (>20 Kbp) essential in silk fiber production. There are more than 160,000 described species of moths and butterflies (Lepidoptera), but only within the last five years have we begun to recover high-quality annotated whole genomes across the order which capture *h-fibroin*. Using PacBio HiFi reads, we produce the first high-quality long-read reference genome for this species. The assembled genome has a length of 532 Mbp, a contig N50 of 16.8 Mbp, an L50 of 14 contigs, and 99.4% completeness (BUSCO). Our annotation using *Bombyx mori* protein and *A.luna* RNAseq evidence captured a total of 20,866 genes at 98.9% completeness with 10,267 functionally annotated proteins and a full-length *h-fibroin* annotation of 2,679 amino acid residues.

Significance

Silk has served an important role culturally, medically, and economically for centuries of human history. Yet, available research on the underlying genetic variation for silk production are largely incomplete in Lepidoptera due to challenges in assembling repeat-rich regions, which are often associated with silk. We provide the first highly contiguous long-read reference genome and full annotation of the repetitive silk gene (*h-fibroin*) for *A.luna,* a wild silk moth with important ecological and cultural implications. We show that long-read sequencing can be an invaluable tool for capturing historically challenging repetitive domains, such as *fibroins.*

Keywords: fibroin, genome, Lepidoptera, moth, PacBio, silk

Introduction

Silk is one of the oldest textiles manufactured by humans, with estimates of production in China starting around 3000 BC (Ball 2009). In Confucius' telling of the origin story of silk, a keenly observant empress, Leizu, accidentally discovers the unwinding fibers of a wild silk moth (*Bombyx mandrina*) cocoon that dropped into her tea from a mulberry. The silk of many moths is comprised of long, repetitive, fibroin filaments bound together by sericin polymers into threads that create a barrier to desiccation and predation for the metamorphosing moth. Thus, Leizu's discovery in her tea was important, as the processing of cocoons with hot water to unwind the long, continuous threads was a key step in the innovation of weaving silk as a textile (Babu 2018). While the human use of silks dates to approximately 3,000 BCE, silks are far more ancient. Silk proteins are produced by members of all three terrestrial arthropod subphyla (Chelicerata, Hexapoda, and Myriapoda; Sehnal and Craig 2009), with uses ranging from protection, transportation, water repellency, and prey capture by organisms both aquatic and terrestrial (Craig 1997; Collin et al. 2010; Sutherland et al. 2010). Due in part to their use in textile production (Supplemental Material 02), Lepidoptera (moths and butterflies) are one of the more commonly recognized silk producers, but surprisingly little is known about the mechanisms and variation in silk production in this group. Characterizations of lepidopteran silks are based almost exclusively on the domestic silkworm (*Bombyx mori*), for which a highly resolved genome (Mita et al. 2004) and transgenic tools are available (reviewed in Ma et al. 2018; 2024). Lepidoptera consists of over 160,000 described species (van Nieukerken et al. 2011), but almost nothing is known about the genetic variation underlying non-model "wild silks" due to gaps in genome quality and availability.

Silk genes such as *heavy chain fibroin* (*h-fibroin*) are long, repeat-rich genes averaging 20 kilobase pairs (Kbp) in length (Sezutsu and Yukuhiro 2000; Kawahara et al. 2021). The order and

number of repetitive motifs determine protein structural qualities (Sehnal and Sutherland 2008), simultaneously make it challenging to fully capture gene sequences using short-read technology. This is mainly due to average fragment lengths of 150 - 300 bp, which cannot span the length of the entire repetitive domain. High quality long-read genomes make it possible to recover challenging full-length genes, making comparative genomic, functional genomic, and phylogenomic studies of silk feasible (Triant et al. 2017; Ellis et al. 2021). Importantly, long-read sequencing can better resolve problematic repeat-rich regions of the genome, providing insights into studying silk gene structural evolution, as was recently reported for spiders, Lepidoptera, and caddisflies (order Trichoptera) (Heckenhauer et al. 2023; Frandsen et al. 2023). Studies such as these serve as an important first step in linking variation in protein sequence with mechanical properties of silk.

A.luna is a large charismatic silk moth native to the eastern United States and is distinguished by its bright green pigmentation and tail-like hindwings (Triant and Pirro 2023). It serves as a dominant food source for pollinators such as bats and birds, and feeds on a wide variety of trees including sweetgum, oak, willow, and hickory (Lindroth 1989). Silk moths are known to produce qualitatively different silks throughout larval development, in addition to cocoon silk (Chen et al. 2012), making it an ideal candidate for assessing the roles of genetic variation and phenotypic plasticity in silk production. This species has one published genome available to date (Triant and Pirro 2023; NCBI accession GCA_010014465.3) however, this genome was produced with short read DNA, and does not accurately resolve primary and haplotype assembled silk genes.

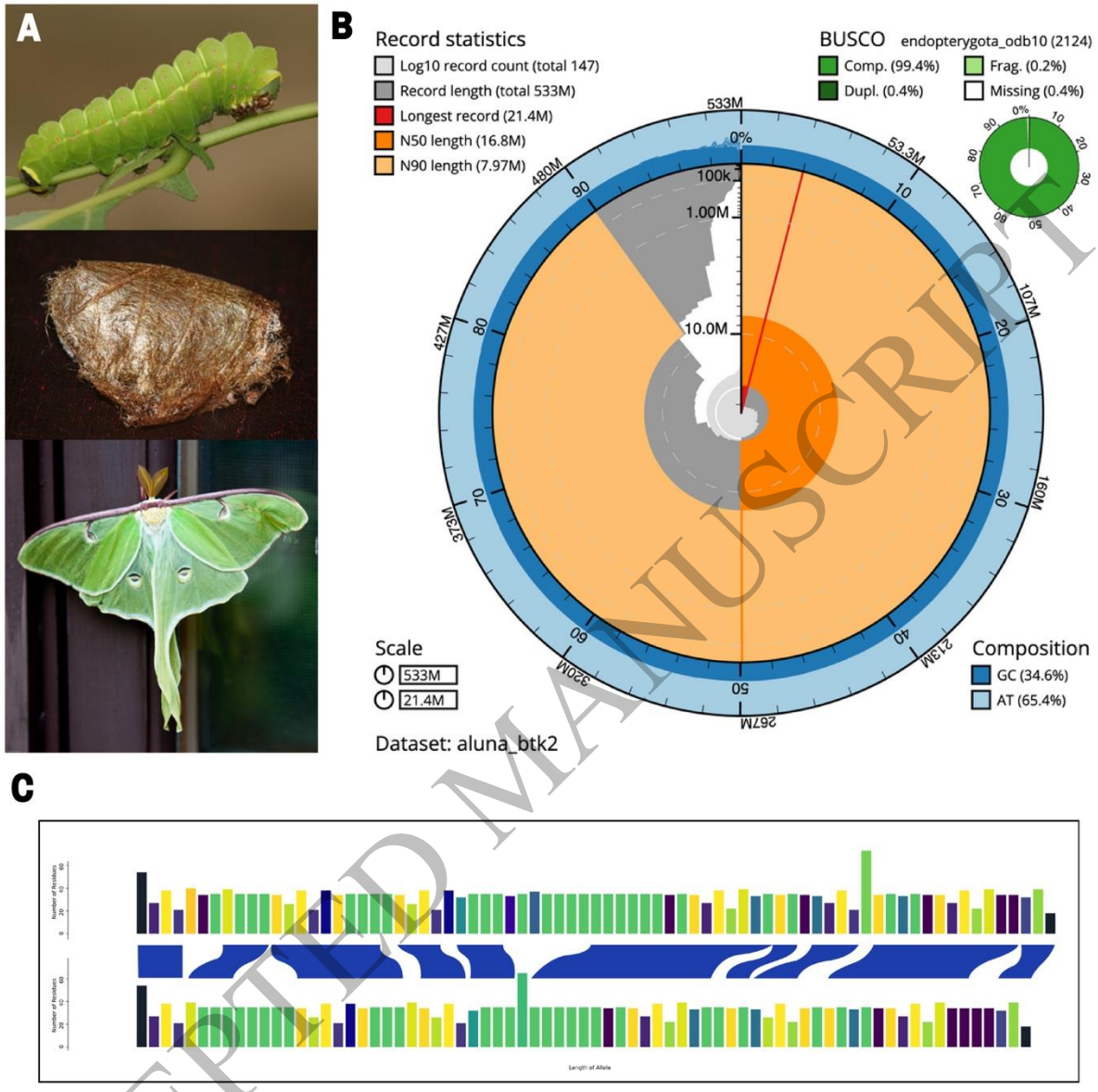

Figure 1. (A) Three of the five life stages of *A.luna* including larva (Credit: Dean Morley, some rights reserved (CC BY-ND), pupa encased in a cocoon (Credit: Dean Morley, some rights reserved (CC BY-ND), and adult (Credit: Patrick, some rights reserved (CC BY)). (B) Snail plot generated in blobtools2 representing general genome statistics, and assembly completeness with BUSCO. (C) Annotation for *h-fibroin* in *A. luna* produced from this study, with both haplotypes. Each bar represents a corresponding repetitive motif (see Supplemental Material 8) with variation in the number of repetitive motifs present, and motif identity. Blue ribbons highlight alignment of

identical amino acid repeats, with genetic variation arising in unaligned motifs. Figure is oriented starting with the N-terminal, stretched to the C-terminal.

New high-quality genomic resources are necessary to understanding how functional genes can give rise to diverse silk phenotypes, and to be able to study these genes in a comparative approach. Here, we present the first comprehensive long read *de-novo* genome assembly and annotation of *A. luna* using PacBio HiFi sequencing, with protein and RNAseq evidence to fully annotate the genome and characterize the entirety of the *h-fibroin* gene.

Results and Discussion

Genome Assembly and QC

From our HiFi DNA library, we recovered 5.3 million polymerase reads containing 2.7 million HiFi reads, with a raw-read N50 of 184 Kbp and an average raw-read length of approximately 90 Kbp. The mean HiFi read length, however, was 7.6 Kbp. Genome size estimations and heterozygosity were predicted using kmer count (KMC) (Kokot et al. 2017; RRID:SCR_001245) and GenomeScope 2.0 (Ranallo-Benavidez et al. 2020) and the predicted genome length was approximately 420 Mbp, with 1.2% heterozygosity (Supplementary Material 03) and 40x coverage.

Whole genome assembly yielded a primary assembly length of approximately 532 Mbp, with quality statistics of N50 = 16.8 Mbp, L50 = 14 contigs, and 99.4% completeness based on the universal, single-copy ortholog gene set for Endopterygota (BUSCO with Endopterygota_odb10 database; Manni et al. 2021; RRID:SCR_015008; Table 1). The shortest contig length in the assembly was 6.1 Kbp, the median contig length was 78.3 Kbp, the mean contig length was 3.4 Mbp, and the longest contig length was 21.4 Mbp. Following non-target

filtering with BlobTools (Laetsch and Blaxter 2017; RRID:SCR_017618), N=11 contigs were predicted to encode primarily Microsporidia (fungal) proteins, and Streptophyta (plant) proteins (Supplemental Material 04). The presence of microsporidian pathogens has been observed in other wild silk moths, especially in the sericulture industry (Subrahmanyam 2019). Thus, these putative non-target sequences are worth assessing further in the future but is beyond the scope of this study. Non-target sequences were removed after assembly, and the resulting cleaned primary assembly was used for all downstream analyses.

Genome Annotation

For the structural annotation, we employed BRAKER3 using two forms of evidence: protein sequences from the closely related *Bombyx mori* (Supplemental Material 06) and RNAseq from 16 individuals of varying tissue types and developmental stages of *A.luna*, totaling 38 independent tissues sequenced (Supplemental Material 07). Each individual represented a particular life stage (egg, 1st-4th instar, pre-pupal, pupal, or adult), and represented one of three tissue types, including head, thorax, abdomen or whole body (e.g., luna_01_head contained head tissue from a 4th instar caterpillar). This annotation recovered a total of 20,866 genes with 98.9% BUSCO completeness (Single-copy:98.1%,Duplicated:0.8%). 17,537 proteins had significant blast hits in the NCBI non-redundant protein database. Of these, blast2go assigned functional annotations to 10,133 proteins, and mapped GO terms to 11,840 proteins. The two major biological processes identified in the *A. luna* genome were cellular and metabolic processes, where binding and catalytic activity were the largest subcategories for molecular function (Supplemental Material 05). When annotating cellular components, most genes were assigned to

cell structure and function, including to membrane, intracellular anatomical structure, and organelle function.

The annotated h-fibroin protein sequence was 2,679 amino acids in length in the primary sequence. The gene is structured by a short initial exon (641bp) followed by an intron (749bp), and a terminal exon with the entire repetitive region (7,997bp). As reported previously in other species (Frandsen et al. 2023) there was a substantial amount of allelic variation in h-fibroin, with the alternate copy spanning 2,597 amino acids. The *h-fibroin* allele for the primary haplotype included a total of 73 internal repetitive motifs broken up by poly-alanine chains $(AAA)_3$, and 2 conserved termini regions, while the alternate haplotype included 71 repetitive motifs, and 2 conserved termini regions. The majority of allelic differences stem from insertions and deletions (indels) of repetitive amino acid motifs, such as the long alanine chains and glycine-rich repeats (Figure 1), known to be responsible for the physical properties of silk fibers (Yonemura & Sehnal, 2006).

Prior to 2019, only seven full-length *h-fibroin* gene sequences were available across Lepidoptera (Kono et al., 2019). Of these sequences, *h-fibroin* was available for two *Antherea* species, and a single *Samia* species, which all share similarity in protein length (~2,500 amino acid residues), gene architecture and repetitive motif composition to that of *A.luna* (Family: Saturniidae). Other annotated *fibroins* from more distantly related lepidopterans such as *Bombyx mori* (Family: Bombycidae) and *Plodia interpunctella* (Family: Pyralidae) are much longer (5,263 amino acid residues and 4,714 amino acid residues, respectively; Zhou et al. 2001, Kawahara et al. 2022) with distinct repetitive motifs compared to saturniid species. *A. luna* and *A. yamamai* have similar silk gene architecture, with long alanine chains and a double cystine complex in the conserved N terminal, which have been identified by previous studies to being

homologous across several different families of lepidopterans (Yonemura et al. 2009). However, Saturniidae are unique in their lack of a conserved cystine residue at the tail end of the h-fibroin protein. Importantly, the conserved cystines and other amino acid residues are indicators of which regions of these molecules (h-fibroin, l-fibroin and P25) are involved with their interactions (Yonemura & Sehnal 2006).

Materials and Methods

Sample Preparation, DNA/RNA Extraction, and Sequencing

All specimens used in this study were reared in the lab from eggs collected from gravid adult females wild-caught in Gainesville, Florida. All samples were flash frozen in liquid nitrogen and stored in -80°C until extraction. High molecular weight DNA was isolated from a single, fifth-instar caterpillar by the University of Florida's Interdisciplinary Center for Biotechnology Research (ICBR) (RRID:SCR_019152) using the Qiagen DNeasy Blood and Tissue kit following manufacturer protocol. RNA was isolated using a TRIzol-chloroform extraction protocol (Rio et al. 2010) from whole eggs and first instar caterpillars, and separately from the head, thoracic and abdominal tissue of larval instars 4 and 5, pre-pupa, pupa, and adults of both sexes.

High molecular weight DNA was sheared to 15 Kbp using mechanical shearing (Covaris). HiFi SMRT bell libraries were prepared following manufacturer protocol for PacBio HiFi (P/N 101-763-800) with a few modifications (Supplemental Material 01). DNA preparations were cleaned using the MoBio PowerClean DNA Cleanup Kit, were cleaned and concentrated using the ZYMO Research RNA Clean and Concentrator kit. DNA preparations

were evaluated for quality on the Agilent TapeStation. Libraries were sequenced on two PacBio Sequel IIe 30-hour SMRT cells at UF ICBR's Next Generation Sequencing Core.

All RNAseq samples (N=38) were pooled and prepared for sequencing using the TruSeq RNA Library Prep Kit v2, and sequenced on an Illumina HiSeq3000 at 2x100 cycles, and with one lane per barcoded and pooled samples (with a total of 8 lanes).

Quality Assessment and Whole Genome Assembly

Raw reads were assessed using the FastQC quality assessment tool (Andrews 2010) for general sequence quality statistics including Phred score, GC content, over-representation, and sequence length distribution. We estimated genome characteristics such as size, heterozygosity and repetitiveness using k-mer counter (KMC) v.3.2.1 (Kokot et al. 2017; RRID:SCR_001245) and GenomeScope 2.0 (Ranallo-Benavidez et al. 2020; RRID:SCR_017014) with a default k-mer length set to 21, and the ploidy set to 2.

We assembled the genome using the assembler hifiasm v0.13-r307 (RRID:SCR_021069) with standard duplicate purging enabled (option –l 2). The resulting primary contig assembly (*.p_ctg.gfa) was used for all downstream analyses. We measured genome contiguity using the assembly_stats.py script (Trizna 2020) and genome completeness using BUSCO v.5.2.2 (Manni et al. 2021; RRID:SCR_015008) with the obd10 reference for Endopterygota.

We identified non-target DNA and potential contaminants using BlobTools v1.0 (Laetsch and Blaxter 2017; RRID:SCR_017618), which indexes the assembly using samtools (Danecek et al. 2021; RRID:SCR_002105), and maps HiFi reads back to the indexed assembly with minimap2 (Li 2018; RRID:SCR_018550). Non-target DNA plots were visualized using BlobPlot, where megablast (Morgulis et al. 2008) searches from BLASTN v 2.10.1

(RRID:SCR_001598) were used to assign taxonomic ID to each contig with an e-value cutoff of 1e-25. Putative non-target sequences, based on non-Arthropoda assignment and clear deviations of GC content and mean sequence depth, were checked against prior BUSCO results. All putative non-target sequences were confirmed to be absent from prior BUSCO results before manually removing them from the assembly, to ensure that genome completeness would remain the same after purging non-target hits.

Structural Annotation and Gene Predictions

To identify repeat elements in the genome, we used RepeatModeler2 (Flynn et al. 2020; RRID:SCR_015027) and masked repetitive elements using RepeatMasker (Smit and Hubley 2013; RRID:SCR 012954). To be the most inclusive due to uncertainty of true repeat elements, we only use soft masking for the structural annotation moving forward.

For gene modeling, we use GeneMark-ETP mode in the BRAKER3 annotation pipeline (Bruna et al. 2021; Hoff et al. 2016; Hoff et al. 2021; Stanke et al. 2006; Stanke et al. 2008; Buchfink et al. 2015; Li et al. 2009; Barnett et al. 2011; Kovaka et al. 2019; Pertea and Pertea 2020; Quinlan 2014; RRID:SCR_018964) to predict the protein coding genes from the soft-masked genome. For the focus of this study, a subset of the Illumina RNAseq reads were used for whole-genome structural annotation. We combined all reads originating from the same individual (including multiple tissue types) and used the concatenated sequences as a representative for each life stage (i.e., luna_01_head, luna_01_thorax, and luna_01_abdomen were combined into one luna_01_instar4 to represent RNAseq from that developmental stage). We mapped reads to the genome using default parameters in hisat2 (Kim et al. 2019; RRID:SCR_015530), and the resulting mapped reads were used to jointly perform GeneMark training and prediction with hints from all available protein sequences of the closely related

Bombyx mori on NCBI (accessed January 16, 2023; Supplemental Material 06) using ProtHint (Bruna et al. 2020; RRID:SCR_021167). The two resulting gene models (Braker and Augustus) were evaluated using BUSCO v.4.1.4 (Manni et al. 2021) with the Endopterygota odb10 core ortholog sets. We selected the model predicted by Augustus (augustus.hints) to represent the gene model of this assembly based on the higher quality BUSCO results, with more complete single copy recovery and less duplication.

To assign function to the genes identified with BRAKER3, we used BLASTP (Sayers et al. 2022) to search the predicted proteins against the NCBI non-redundant protein database with an e-value cutoff of 1e-4 for the top hits up to ten hits. In addition, we used BLAST2GO (Sayers et al. 2022) to assign gene ontology terms and functional annotations to the predicted proteins.

For functional annotation of the *h-fibroin* silk gene, we used the primary and haplotype assemblies from the hifiasm output from the previous assembly methods. N- and C- termini for *h-fibroin* are fairly conserved across Lepidoptera, thus we used tBLASTn with termini sequences from *Antherea mylitta*, a species from the same family (Hwang et al. 2001), to search the primary and alternate haplotype *A. luna* assemblies. We then extracted the target sequences and flanking regions were extracted from the primary and haplotype assemblies and annotated the region in Augustus v.3.3.2.

Table 1a. A comparison of genome statistics for the genome produced in this study, to the existing wild silk moth genome assemblies (Family: Saturniidae) available on NCBI.

Table 1b. Comparison of genome completeness between the only other existing *A.luna* genome assembly available to date.

	Assembly type	Size (Mb)	N50 (Mb)	L50	Number of contigs	Genome coverage	Ref
*Actias luna**	contig	532	16.8	14	155	40x	Authors
*Samia ricini**	chromosome	452.5	33.1	8	1489	100x	GCA_014132275.2
Saturnia japonica	chromosome	581.8	20.2	13	959	100x	GCA_033032175.1
*Saturnia pavonia**	chromosome	489.9	17.7	13	71	50x	GCA_947532125.1
Antheraea pernyi	scaffold	787	0.338	4256	103,035	100x	GCA_015888305.1
Actias luna	scaffold	559.1	0.002	63,632	529,540	140x	GCA_010014465.3
*Antheraea mylitta**	scaffold	698.4	5.3	2894	37,190	166.6x	GCA_014332785.1
*Saturnia pavonia**	scaffold	459.9	2.4	62	418	50x	GCA_947532135.1

Note: *genomes produced with long-read sequencing platforms e.g. PacBio or Oxford Nanopore

	BUSCO complete (%)	Single Copy (%)	Duplicated (%)	Fragmented (%)	Missing (%)	Ref
*Actias luna**	99.4	99.0	0.4	0.2	0.4	Authors
Actias luna	71.4	63.9	7.5	15.4	13.2	GCA_010014465.3

Note: *genomes produced with long-read sequencing platforms e.g. PacBio or Oxford Nanopore

Supplementary Material

See supplemental materials document, and the DOI for genome assembly materials (10.6084/m9.figshare.25483282; https://figshare.com/s/e0381962bda32f013804) genome structural annotation materials (10.6084/m9.figshare.25483330; https://figshare.com/s/9793ed051033ea39f6b1) and genome functional annotation (10.6084/m9.figshare.25483372; https://figshare.com/s/b45bdd01f1d071d3ac7e).

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under NSF Award No. 1938103. This research was additionally supported by the McGuire Center for Lepidoptera and Biodiversity (MGCL), the Florida Museum of Natural History (FLMNH), The University of Florida Research Opportunity Fund (UF-ROF), University of Florida's School of Natural Resources and Environment (SNRE), and the National Science Foundation grants NSF MCB-2217159 (AYK) and NSF MCB-2217155 (PBF). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation. We thank Ashlyn Powell for her help generating allelic diversity elements in our figure, and we thank JT Tubera for his help with figure editing.

Data Availability

PacBio HiFi reads and raw RNAseq reads, and the draft genome assembly are publicly available on NCBI under the following BioProject number: PRJNA1072661. The final assembly, BRAKER3 annotation output, heavy-fibroin protein sequence, *Bombyx mori* protein evidence, and all bioinformatic scripts used in this study are all available on FigShare (DOI 10.6084/m9.figshare.25483375)

Literature Cited

Andrews, S. (2010.). *FastQC: A Quality Control tool for High Throughput Sequence Data*. Retrieved December 18, 2023, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Babu, K. M. (2019). *Silk: Processing, properties and applications* (Second edition). The Textile Institute : Woodhead Publishing, an imprint of Elsevier.

Ball, P. (2009). Rethinking silk's origins. *Nature*. https://doi.org/10.1038/457945a

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., & Marth, G. T. (2011). BamTools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics (Oxford, England)*, *27*(12), 1691–1692. https://doi.org/10.1093/bioinformatics/btr174

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, *3*(1), lqaa108. https://doi.org/10.1093/nargab/lqaa108

Brůna, T., Lomsadze, A., & Borodovsky, M. (2020). GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics*, *2*(2), lqaa026. https://doi.org/10.1093/nargab/lqaa026

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

Chen, F., Porter, D., & Vollrath, F. (2012). Morphology and structure of silkworm cocoons. *Materials Science and Engineering: C*, *32*(4), 772–778. https://doi.org/10.1016/j.msec.2012.01.023

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. https://doi.org/10.1038/s41592-020-01056-5

Collin, M. A., Mita, K., Sehnal, F., & Hayashi, C. Y. (2010). Molecular Evolution of Lepidopteran Silk Proteins: Insights from the Ghost Moth, *Hepialus californicus*. *Journal of Molecular Evolution*, *70*(5), 519–529. https://doi.org/10.1007/s00239-010-9349-8

Craig, C. L. (1997). Evolution of Arthropod Silks. *Annual Review of Entomology*, *42*(1), 231–267. https://doi.org/10.1146/annurev.ento.42.1.231

Danecek, P. et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Ellis, E. A., Storer, C. G., & Kawahara, A. Y. (2021). *De novo* genome assemblies of butterflies. *GigaScience*, *10*(6), giab041. https://doi.org/10.1093/gigascience/giab041

Flynn, J. M. et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M., & Stanke, M. (2021). TSEBRA: Transcript selector for BRAKER. *BMC Bioinformatics*, *22*(1), 566. https://doi.org/10.1186/s12859-021-04482-0

Gotoh, O. (2008). Direct mapping and alignment of protein sequences onto genomic sequence. *Bioinformatics*, *24*(21), 2438–2444. https://doi.org/10.1093/bioinformatics/btn460

Gupta K, A., Mita, K., Arunkumar, K. P., & Nagaraju, J. (2015). Molecular architecture of silk fibroin of Indian golden silkmoth, Antheraea assama. *Scientific Reports*, *5*(1), 12706. https://doi.org/10.1038/srep12706

Heckenhauer, J. et al. (2023). Characterization of the primary structure of the major silk gene, *h-fibroin*, across caddisfly (Trichoptera) suborders. *iScience*, *26*(8), 107253. https://doi.org/10.1016/j.isci.2023.107253

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics (Oxford, England)*, *32*(5), 767–769. https://doi.org/10.1093/bioinformatics/btv661

Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-Genome Annotation with BRAKER. In M. Kollmar (Ed.), *Gene Prediction* (Vol. 1962, pp. 65–95). Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_5

Hwang, J.-S., et al. (2001). Cloning of the fibroin gene from the oak silkworm, *Antheraea yamamai* and its complete sequence. *Biotechnology Letters*, *23*(16), 1321–1326. https://doi.org/10.1023/A:1010542011150

Kawahara, A. Y., et al. (2022). Long-read HiFi sequencing correctly assembles repetitive *heavy fibroin* silk genes in new moth and caddisfly genomes. *Gigabyte*, *2022*, 1–14. https://doi.org/10.46471/gigabyte.64

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kokot, M., Długosz, M., & Deorowicz, S. (2017). KMC 3: Counting and manipulating k-mer statistics. *Bioinformatics*, *33*(17), 2759–2761. https://doi.org/10.1093/bioinformatics/btx304

Kono, N., et al. (2019). The bagworm genome reveals a unique fibroin gene that provides high tensile strength. *Communications Biology*, *2*(1), 148. https://doi.org/10.1038/s42003-019-0412-8

Kovaka, S. et al. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*(1), 278. https://doi.org/10.1186/s13059-019-1910-1

Laetsch, D. R., & Blaxter, M. L. (2017). BlobTools: Interrogation of genome assemblies. *F1000Research*, *6*, 1287. https://doi.org/10.12688/f1000research.12232.1

Lee, J et al. (2021). The genome sequence of *Samia ricini*, a new model species of lepidopteran insect. *Molecular Ecology Resources*, *21*(1), 327–339. https://doi.org/10.1111/1755-0998.13259

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lindroth, R. L. (1989). Chemical ecology of the luna moth: Effects of host plant on detoxification enzyme activity. *Journal of Chemical Ecology*, *15*(7), 2019–2029. https://doi.org/10.1007/BF01207434

Ma, S. et al. (2024). High-throughput and genome-scale targeted mutagenesis using CRISPR in a nonmodel multicellular organism, *Bombyx mori*. *Genome Research*. https://doi.org/10.1101/gr.278297.123

Ma, S.-Y., Smagghe, G., & Xia, Q.-Y. (2019). Genome editing in *Bombyx mori*: New opportunities for silkworm functional genomics and the sericulture industry. *Insect Science*, *26*(6), 964–972. https://doi.org/10.1111/1744-7917.12609

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. https://doi.org/10.1093/molbev/msab199

Mita, K. (2004). The Genome Sequence of Silkworm, *Bombyx mori*. *DNA Research*, *11*(1), 27–35. https://doi.org/10.1093/dnares/11.1.27

Morgulis, A. et al. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics*, *24*(16), 1757–1764. https://doi.org/10.1093/bioinformatics/btn322

Peigler, R. S. (1993). Wild Silks of the World. *American Entomologist*, *39*(3), 151–162. https://doi.org/10.1093/ae/39.3.151

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

Quinlan, A. R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current Protocols in Bioinformatics*, *47*, 11.12.1-34. https://doi.org/10.1002/0471250953.bi1112s47

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

Reddy, N., & Yang, Y. (2012). Investigation of the Structure and Properties of Silk Fibers Produced by *Actias luna*. *Journal of Polymers and the Environment*, *20*(3), 659–664. https://doi.org/10.1007/s10924-012-0482-x

Rio, D. C., Ares, M., Hannon, G. J., & Nilsen, T. W. (2010). Purification of RNA Using TRIzol (TRI Reagent). *Cold Spring Harbor Protocols*, *2010*(6), pdb.prot5439. https://doi.org/10.1101/pdb.prot5439

Sehnal, F., & Craig, C. (2009). Chapter 235—Silk Production. In V. H. Resh & R. T. Cardé (Eds.), *Encyclopedia of Insects (Second Edition)* (pp. 921–924). Academic Press. https://doi.org/10.1016/B978-0-12-374144-8.00244-7

Sehnal, F., & Sutherland, T. (2008). Silks produced by insect labial glands. *Prion*, *2*(4), 145–153. https://doi.org/10.4161/pri.2.4.7489

Sezutsu, H., & Yukuhiro, K. (2000). Dynamic Rearrangement Within the *Antheraea pernyi* Silk Fibroin Gene Is Associated with Four Types of Repetitive Units. *Journal of Molecular Evolution*, *51*(4), 329–338. https://doi.org/10.1007/s002390010095

Smit, A. F., & Hubley, R. (n.d.). *RepeatMasker Open-4.0*. Retrieved December 18, 2023, from https://www.repeatmasker.org/

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*(5), 637–644. https://doi.org/10.1093/bioinformatics/btn013

Stanke, M. et al. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(Web Server), W435–W439. https://doi.org/10.1093/nar/gkl200

Subrahmanyam, G. et al. (2019). Isolation and Molecular Identification of Microsporidian Pathogen Causing Nosemosis in Muga Silkworm, *Antheraea assamensis* Helfer (Lepidoptera: Saturniidae). *Indian Journal of Microbiology*, *59*(4), 525–529. https://doi.org/10.1007/s12088-019-00822-0

Sutherland, T. D., Young, J. H., Weisman, S., Hayashi, C. Y., & Merritt, D. J. (2010). Insect Silk: One Name, Many Materials. *Annual Review of Entomology*, *55*(1), 171–188. https://doi.org/10.1146/annurev-ento-112408-085401

Triant, D. A., Cinel, S. D., & Kawahara, A. Y. (2018). Lepidoptera genomes: Current knowledge, gaps and future directions. *Current Opinion in Insect Science*, *25*, 99–105. https://doi.org/10.1016/j.cois.2017.12.004

Triant, D. A., & Pirro, S. (2023). The Complete Genome Sequence of *Actias luna* (Saturniidae, Lepidoptera), the luna moth. *Biodiversity Genomes*, *2023*. https://doi.org/10.56179/001c.75356

Trizna, M. (2020). *Assembly_stats 0.1.4* (0.1.4) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.3968775

Tuskes, P. M., Tuttle, J. P., & Collins, M. M. (1996). *The wild silk moths of North America: A natural history of the Saturniidae of the United States and Canada*. Comstock Pub. Associates.

Van Nieukerken, E. J. et al. (2011). Order Lepidoptera Linnaeus, 1758. In: Zhang, Z.-Q. (Ed.) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. *Zootaxa*, *3148*(1). https://doi.org/10.11646/zootaxa.3148.1.41

Yonemura, N., Mita, K., Tamura, T., & Sehnal, F. (2009). Conservation of Silk Genes in Trichoptera and Lepidoptera. *Journal of Molecular Evolution*, *68*(6), 641–653. https://doi.org/10.1007/s00239-009-9234-5

Yonemura, N., & Sehnal, F. (2006). The design of silk fiber composition in moths has been conserved for more than 150 million years. *Journal of Molecular Evolution*, *63*(1), 42–53. https://doi.org/10.1007/s00239-005-0119-y

Zhou, C. et al. (2001). Silk fibroin: Structural implications of a remarkable amino acid sequence. *Proteins: Structure, Function, and Bioinformatics*, *44*(2), 119–122. https://doi.org/10.1002/prot.1078