Genome Assembly and Annotation of the Dark-Branded Bushbrown Butterfly *Mycalesis mineus* (Nymphalidae: Satyrinae)

Suriya Narayanan Murugesan [1,*], Shen Tian [1], and Antónia Monteiro [1,*]

[1]Department of Biological Sciences, National University of Singapore, Singapore 117558, Singapore

*Corresponding authors: E-mails: snm_02@nus.edu.sg; antonia.monteiro@nus.edu.sg.

Accepted: March 13, 2024

Abstract

We report a high-quality genome draft assembly of the dark-branded bushbrown, *Mycalesis mineus*, a member of the Satyrinae subfamily of nymphalid butterflies. This species is emerging as a promising model organism for investigating the evolution and development of phenotypic plasticity. Using 45.99 Gb of long-read data (N50 = 11.11 kb), we assembled a genome size of 497.4 Mb for *M. mineus*. The assembly is highly contiguous and nearly complete (96.8% of Benchmarking Universal Single-Copy Orthologs lepidopteran genes were complete and single copy). The genome comprises 38.71% of repetitive elements and includes 20,967 predicted protein-coding genes. The assembled genome was super-scaffolded into 28 pseudo-chromosomes using a closely related species, *Bicyclus anynana*, with a chromosomal-level genome as a template. This valuable genomic tool will advance both ongoing and future research focused on this model organism.

Key words: *Mycalesis mineus*, genome, eyespot plasticity, Nanopore.

Significance

The dark-branded bushbrown, *Mycalesis mineus*, is a tropical butterfly that exhibits phenotypic plasticity in response to environmental cues and is also sexually dimorphic. The lack of a comprehensive genome resource has been a significant obstacle in utilizing *M. mineus* as a model system in functional genomic research focused on sexual dimorphism and phenotypic plasticity. Here, we present the first high-quality draft genome of *M. mineus*, which will serve as a valuable resource in addressing some of these questions.

Introduction

The dark-branded bushbrown butterfly *Mycalesis mineus* is a tropical satyrid that inhabits Asia, ranging from India to Indonesia and the Philippines. Its larvae primarily consume grasses, while the adult butterflies feed on decaying fruits found on the ground. The adult male and female are sexually dimorphic, and males are distinguishable by their smaller size, the presence of silver scales and hair pencils on their wings, and a darker overall coloration (Fig. 1).

Both the adults and the pupae vary in their color patterns in response to environmental cues. Adult *M. mineus* butterflies display two distinct seasonal forms, referred to as wet and dry morphs. In the dry morphs, the wings lack or have smaller eyespots compared to the wet morphs, and this plasticity is controlled, in part, by temperature (Islam et al. 2012; Van Bergen et al. 2017; Fig. 1). Moreover, pupae grown in two different relative humidity (RH) conditions in the laboratory exhibited two color morphs: brown color at low RH (60%) and green color at high RH (85%; Mayekar and Kodandaramaiah 2017). Despite this species demonstrating phenotypic plasticity in two separate traits and being sexually dimorphic, limited

© The Author(s) 2024. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-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 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.

Downloaded from https://academic.oup.com/gbe/article/16/3/evae051/7632164 by guest on 31 March 2024

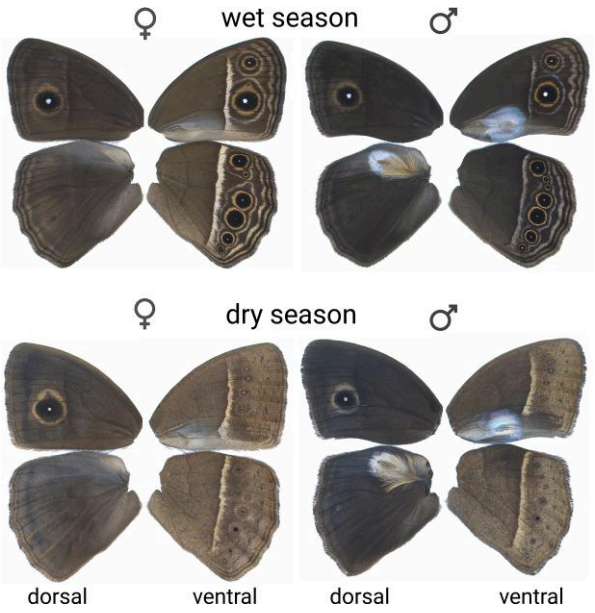

Fig. 1.—Representative images of the two seasonal forms of *M. mineus*, showing dorsal (left) and ventral sides (right) of female and male wings. The species is sexually dimorphic, and males can be distinguished by the presence of silver scales (at the base of the dorsal hindwing and ventral forewing and) and hair pencils (at the base of the dorsal hindwing) on their wings. Dry-season forms have reduced ventral eyespots in both sexes.

functional studies have been conducted to investigate the developmental basis and the evolution of these distinct traits.

Most research on butterfly wing pattern plasticity has focused on size plasticity of *Bicyclus anynana* eyespots (Brakefield and Reitsma 1991; Tian and Monteiro 2022). This species is another satyrid found in Africa, which diverged from *M. mineus* around 25 My ago (Chazot et al. 2021). Modern functional tools and a high-quality genome have been established for *B. anynana* (Monteiro et al. 2013; Banerjee and Monteiro 2018; Murugesan et al. 2022) while no tools or genome has been established for the study of *M. mineus*. Establishing a high-quality genome for *M. mineus* is an initial step for including this species in studies that aim to understand the evolution and development of eyespot size plasticity in this group of satyrid butterflies. In this study, we present a high-quality draft genome assembly and gene annotation for *M. mineus*.

Results

Genome Assembly Statistics

High molecular weight (HMW) DNA was extracted from a single female individual. A total of 7.6 µg (Qubit reading) of this HMW DNA extracted was utilized completely for Oxford Nanopore and short-read library preparation and sequencing.

Long-read sequencing using Oxford Nanopore technology produced an output of 6.06 million reads, totaling 45.5 Gb of data (~90× coverage), with an N50 of 11.1 kb. Short-read sequencing generated approximately 59 Gb of data (~118×). The long-read data were utilized for constructing the genome assembly using two different assemblers. The Flye (v2.9.2-GCC-11.3.0) assembler produced an assembly of 875 Mb with an N50 of 198 kb, and the largest contig size was 15.61 Mb. The Wtdbg2 (v.2.5) assembler generated an assembly of 505 Mb with an N50 of 4.5 Mb, and the largest contig size was 36.9 Mb. These two assemblies were merged using quickmerge (v0.3-5, Bioconda package), followed by the removal of haplotigs and heterozygosity through purge_haplotigs (v1.1.2, Bioconda package), resulting in a genome assembly with a size of 506.7 Mb and an N50 of 8.8 Mb. The purged assembly underwent further refinement through polishing, using two rounds of racon and medaka with long reads, as well as using POLishing by Calling Alternatives (POLCA; part of MaSuRCA v3.2.1 assembler) with short reads.

Assembled contigs were aligned to the genome of the closely related species *B. anynana* and showed a high degree of synteny (supplementary fig. S1A, Supplementary Material online). But we also observed that some of the contigs were likely misassembled, as a few large *M. mineus* contigs mapped to multiple smaller scaffolds/chromosomes of *B. anynana* (supplementary fig. S1A, Supplementary Material online). So, we used *ragtag_correct* to break any misassembled contigs at the places where long reads are mapped at very low coverage. After this *ragtag_correct* step, the two genomes still showed high level of synteny, but the *M. mineus* genome became fragmented into more pieces (supplementary fig. S1B, Supplementary Material online). Finally, the contigs were super-scaffolded using the *B. anynana* genome as a reference. The final assembly resulted in a genome size of 497.4 Mb with scaffolds N50 of 17.8 Mb (Fig. 2a; supplementary fig. S2A and table S1, Supplementary Material online). To assess the genome's completeness, we employed a Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis, utilizing the lepidoptera_odb10 data set. The resulting assembly demonstrated a high level of completeness, with 96.8% of the single-copy gene set from BUSCO being found in the *M. mineus* genome (Fig. 2a; supplementary fig. S2B and table S2, Supplementary Material online). In addition, 99.53% of the short reads mapped to the final assembled genome, highlighting completeness of the genome. The final *M. mineus* assembly also aligned with the *B. anynana* genome and showed high level of synteny (Fig. 2b).

The genome was repeat-masked using Redmask and was annotated using Braker. For annotation, the model was trained using an arthropod protein database along with protein sequences from closely related species (see Materials and Methods for details). The annotation process yielded a total of 18,360 genes accompanied by 20,967 transcripts. The

Downloaded from https://academic.oup.com/gbe/article/16/3/evae051/7632164 by guest on 31 March 2024

A

Assembly statistics	
Size	497.4 MB
N50 scaffold	17.8 MB
Largest scaffold	24.3 MB
Gaps	296
N count	276992
BUSCO Statistics (lepidoptera_odb10)	
Complete and single-copy	5119(96.8%)
Complete and duplicated	71 (1.3%)
Fragmented	26 (0.5%)
Missing	70 (1.4%)
Total BUSO groups searched	5286

B

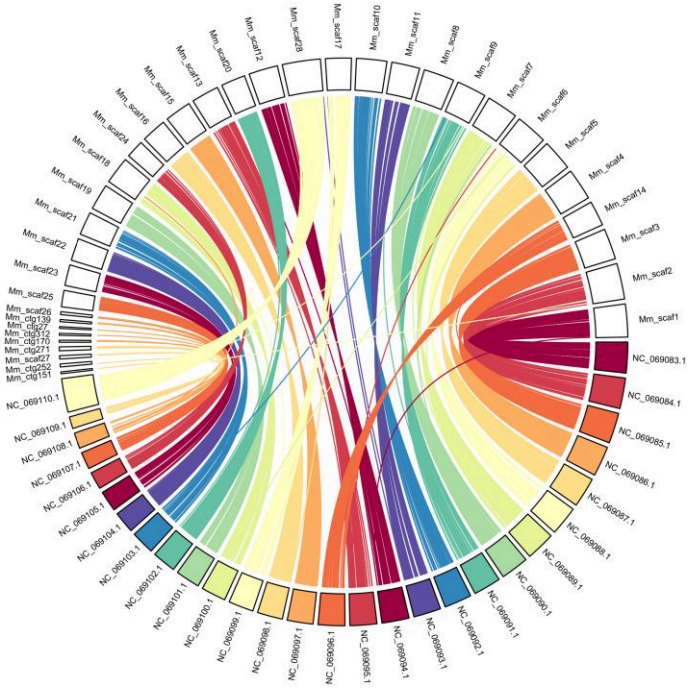

Fig. 2.—Assembly statistics and Circos plot of genome alignment between *M. mineus* and *B. anynana*. a) Genome statistics with BUSCO score for the *M. mineus_v0.6* assembly. b) Circos plot for the alignment after super-scaffolding the multiple separate *M. mineus* contigs shows a high level of synteny between the two genomes. Colored blocks are *B. anynana* scaffolds (chromosomes), and white blocks are *M. mineus* contigs and scaffolds. The ribbon lines show the orthologs of *B. anynana* sequences in the *M. mineus* genome.

protein sequence from the genome annotation shows 97% completeness with 84% of the single copy and 13% duplicated copy of BUSCO gene set (supplementary fig. S2B, Supplementary Material online).

Conclusion

Here, we present a high-quality draft genome assembly of *M. mineus*, a promising model system for studying phenotypic plasticity in response to environmental cues, as well as sexual dimorphism. Our assembly placed 93% of the 497.4 Mb genome into 28 chromosomal super-scaffolds, with an N50 of 17.8 Mb. The assembled genome is highly complete, as validated by BUSCO, and ordered similarly to other lepidopteran genomes. We successfully annotated 20,967 protein-coding transcripts. This genome assembly serves as a valuable resource for ongoing and future comparative studies on eyespot size plasticity, as well as broader studies of wing pattern evolution, utilizing *M. mineus* as a model system.

Materials and Methods

DNA Extraction and Sequencing

A female *M. mineus* specimen was collected from Clementi Forest, Singapore, under the permission from the National Parks Board (permit ID: NP/RP14-063-6a). Species identification was based on *The Butterflies of the Malay Peninsula*, 4th edition (Corbet and Pendlebury 1992). The thorax (including legs) and head regions of the specimen were used to extract HMW DNA. The extraction was performed using the Monarch HMW DNA Extraction Kit (T3060S), following the manufacturer's user manual. The quantity and quality of the extracted HMW DNA were assessed using a NanoDrop and Qubit.

Library preparation and sequencing were conducted by Novogene in Singapore. Nanopore library preparation was performed using the Ligation Sequencing Kit for gDNA (SQK-LSK114) following the user's protocol. Sequencing was carried out on a FLO-PRO002 cell in the PromethION machine, and base calling was performed using Guppy 6.5.7. Illumina short-read (2 × 150 bp) paired-end library preparation was done by Novogene, Singapore, and sequencing was carried out using Illumina NovaSeq 6000 sequencer.

De Novo Genome Assembly

Initial genome assembly was performed using the Flye (v2.9.2-GCC-11.3.0) assembler (Kolmogorov et al. 2019) and the wtdbg2 assembler (Ruan and Li 2020) separately. Flye was run with the default settings, with polisher iteration set to 1, while wtdbg2 (v2.5) was run with its default settings. For both assemblers, the genome size was set to 500 Mb, based on the genome size of the closely related

Downloaded from https://academic.oup.com/gbe/article/16/3/evae051/7632164 by guest on 31 March 2024

species *B. anynana*. The two genome assemblies were merged using quickmerge (Chakraborty et al. 2016), with wtdbg2 as the reference assembly and the Flye assembly as the query. Heterozygosity and duplication in the assembly were removed using purge_haplotigs (Roach et al. 2018), keeping only the haploid contigs. The output assembly from purge_haplotigs was polished with two rounds of racon (v1.5; Vaser et al. 2017) and medaka (https://github.com/nanoporetech/medaka) using the long reads and further polished with short reads using POLCA (Zimin and Salzberg 2020). Lepidoptera genomes are highly syntenic and ordered (D'Alençon et al. 2010; Traut et al. 2023). The draft assembly was checked for putative misassembly using ragtag_correct (Alonge et al. 2022) using *B. anynana* genome as reference (Saccheri et al. 2023) and the long reads. Ragtag_correct using the long reads mapped to the genome validates the misassemblies identified between reference genome and query genome and breaks the assembly when at breakpoints with very low or very high coverage. The corrected genome was scaffolded in two steps by first using Scaffolding Assemblies with Multiple Big Alignments (SAMBA), which uses long read to improve the contiguity of the contigs. This was followed by scaffolding using chromosome_scaffolder tools in MaSuRCA assembler with *B. anynana* genome as a reference (Zimin et al. 2013; Zimin and Salzberg 2022). The completeness of the genome at each step was assessed using BUSCO scores (Simão et al. 2015; supplementary table S1 and fig. S2B, Supplementary Material online).

Repeat Masking and Genome Annotations

The genome was repeat-masked using Redmask (https://github.com/nextgenusfs/redmask). The repeat-masked genome was annotated using the Braker2 automated pipeline (Stanke et al. 2006, 2008; Hoff et al. 2016, 2019; Brůna et al. 2021). We utilized the Arthropoda protein data set from OrthoDB v.10 (Kriventseva et al. 2019), along with protein sequences from *B. anynana*, *Junonia orithya*, and *Bombyx mori*, to train the model and predict gene structures within the genome. We used the protein sequences from the three above species for this annotation because their genomes are nearly complete and have high gene completeness. Functional annotation of the predicted genes was performed using eggNOG-mapper v.2.1.12 (Cantalapiedra et al. 2021) against the eggNOG arthropod database v.5.0.1 (Huerta-Cepas et al. 2019) via the Diamond BLAST tool (Buchfink et al. 2014).

Synteny Analysis

We used nucmer from mummer4 (Marçais et al. 2018) to align our *M. mineus* genome to 28 chromosomes of *B. anynana* genome and visualized the synteny in R using scripts from Tunstrom et al. (2022).

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Author Contributions

S.N.M. extracted HMW DNA, analyzed the data, and wrote the first draft. S.T. collected the sample from the field and imaged the specimens. A.M. supervised the project. S.N.M., S.T., and A.M. edited the draft. All authors approve of the manuscript.

Funding

This work was supported by the National Research Foundation Singapore, under its Investigatorship program (award NRF-NRFI05-2019-0006). S.T. was supported by a PhD scholarship from the NRF-CRP program (award NRF-CRP20-2017-0001). We thank NUS HPC for access to their server, which was used to analyze the data.

Data Availability

The raw nanopore reads and short reads used for the genome assembly are available in NCBI under BioProject PRJNA1037280, and the genome assembly was submitted to NCBI (GCA_034621355.1). The assembly and annotation are also available in dryad (https://doi.org/10.5061/dryad.47d7wm3mr). Codes used for the construction of the assembly are in the supplementary files.

Literature Cited

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. 2022:23(1):258. https://doi.org/10.1186/s13059-022-02823-7.

Banerjee TD, Monteiro A. Crispr-Cas9 mediated genome editing in *Bicyclus anynana* butterflies. Methods Protoc. 2018:1(2):16. https://doi.org/10.3390/mps1020016.

Brakefield PM, Reitsma N. Phenotypic plasticity, seasonal climate and the population biology of *Bicyclus* butterflies (Satyridae) in Malawi. Ecol Entomol. 1991:16(3):291–303. https://doi.org/10.1111/j.1365-2311.1991.tb00220.x.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 2021:3(1):lqaa108. https://doi.org/10.1093/nargab/lqaa108.

Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2014:12(1):59–60. https://doi.org/10.1038/nmeth.3176.

Cantalapiedra CP, Herņandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021:38(12):5825–5829. https://doi.org/10.1093/molbev/msab293.

Downloaded from https://academic.oup.com/gbe/article/16/3/evae051/7632164 by guest on 31 March 2024

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 2016:44(1):e147. https://doi.org/10.1093/nar/gkw654.

Chazot N, Condamine FL, Dudas G, Peña C, Kodandaramaiah U, Matos-Maraví P, Aduse-Poku K, Elias M, Warren AD, Lohman DJ, et al. Conserved ancestral tropical niche but different continental histories explain the latitudinal diversity gradient in brush-footed butterflies. Nat Commun. 2021:12(1):5717. https://doi.org/10.1038/s41467-021-25906-8.

Corbet SA, Pendlebury HM. 1992. In: Eliot J, editors. The butterflies of the Malay Peninsula. 4th ed. Kuala Lumpur, Malaysia: Malayan Nature Society. p. 595.

D'Alençon E, Sezutsu H, Legeai F, Permal E, Bernard-Samain S, Gimenez S, Gagneur C, Cousserans F, Shimomura M, Brun-Barale A, et al. Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements. Proc Natl Acad Sci U S A. 2010:107(17):7680–7685. https://doi.org/10.1073/pnas.0910413107.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 2016:32(5):767–769. https://doi.org/10.1093/bioinformatics/btv661.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. Methods Mol Biol. 2019:176(12):65–95. https://doi.org/10.1007/978-1-4939-9173-0_5.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei 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(D1):D309–D314. https://doi.org/10.1093/nar/gky1085.

Islam A, Razzak M, Islam M, Saifullah A, Endo K, Yamanaka A. Dry and wet season polymorphism in the butterflies, *Melanitis leda* and *Mycalesis mineus* (Satyridae: Lepidoptera). Univ J Zool Rajshahi Univ. 2012:29:33–39. https://doi.org/10.3329/ujzru.v29i1.9463.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019:37(5):540–546. https://doi.org/10.1038/s41587-019-0072-8.

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2019:47(D1):D807–D811. https://doi.org/10.1093/nar/gky1053.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol. 2018:14(1):e1005944. https://doi.org/10.1371/journal.pcbi.1005944.

Mayekar HV, Kodandaramaiah U. Pupal colour plasticity in a tropical butterfly, *Mycalesis mineus* (Nymphalidae: Satyrinae). PLoS One 2017:12(2):e0171482. https://doi.org/10.1371/journal.pone.0171482.

Monteiro A, Chen B, Ramos DM, Oliver JC, Tong X, Guo M, Wang WK, Fazzino L, Kamal F. Distal-less regulates eyespot patterns and melanization in *Bicyclus* butterflies. J Exp Zool B Mol Dev Evol. 2013:320(5):321–331. https://doi.org/10.1002/jez.b.22503.

Murugesan SN, Connahs H, Matsuoka Y, Das Gupta M, Tiong GJL, Huq M, Gowri V, Monroe S, Deem KD, Werner T, et al. Butterfly eyespots evolved via cooption of an ancestral gene-regulatory network that also patterns antennae, legs, and wings. Proc Natl Acad Sci U S A. 2022:119(8):e2108661119. https://doi.org/10.1073/pnas.2108661119.

Roach MJ, Schmidt SA, Borneman AR. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 2018:19(1):460. https://doi.org/10.1186/s12859-018-2485-7.

Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020:17(2):155–158. https://doi.org/10.1038/s41592-019-0669-3.

Saccheri IJ; 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. The genome sequence of the squinting bush brown, *Bicyclus anynana* (Butler, 1879). Wellcome Open Res. 2023:8:280. https://doi.org/10.12688/wellcomeopenres.19432.1.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015:31(19):3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

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

Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 2006:7(1):62. https://doi.org/10.1186/1471-2105-7-62.

Tian S, Monteiro A. A transcriptomic atlas underlying developmental plasticity of seasonal forms of *Bicyclus anynana* butterflies. Mol Biol Evol. 2022:39(6):msac126. https://doi.org/10.1093/molbev/msac126.

Traut W, Sahara K, ffrench-Constant RH. Lepidopteran Synteny Units reveal deep chromosomal conservation in butterflies and moths. G3 (Bethesda). 2023:13(8):jkad134. https://doi.org/10.1093/g3journal/jkad134.

Tunstrom K, Wheat CW, Parmesan C, Singer MC, Mikheyev AS. A genome for Edith's checkerspot butterfly: an insect with complex host-adaptive suites and rapid evolutionary responses to environmental changes. Genome Biol Evol. 2022:14(8):evac113. https://doi.org/10.1093/gbe/evac113.

Van Bergen E, Osbaldeston D, Kodandaramaiah U, Brattström O, Aduse-Poku K, Brakefield PM. Conserved patterns of integrated developmental plasticity in a group of polyphenic tropical butterflies. BMC Evol Biol. 2017:17(1):59. https://doi.org/10.1186/s12862-017-0907-1.

Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017:27(5):737–746. https://doi.org/10.1101/gr.214270.116.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics 2013:29(21):2669–2677. https://doi.org/10.1093/bioinformatics/btt476.

Zimin AV, Salzberg SL. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS Comput Biol. 2020:16(6):e1007981. https://doi.org/10.1371/journal.pcbi.1007981.

Zimin AV, Salzberg SL. The SAMBA tool uses long reads to improve the contiguity of genome assemblies. PLoS Comput Biol. 2022:18(2):e1009860. https://doi.org/10.1371/journal.pcbi.1009860.

Associate editor: Christopher Wheat

Downloaded from https://academic.oup.com/gbe/article/16/3/evae051/7632164 by guest on 31 March 2024