Genome Assembly of the Dogface Butterfly *Zerene cesonia*

Luis Rodriguez-Caro[1,3,*], Jennifer Fenner[1], Caleb Benson[1], Steven M. Van Belleghem[2], and Brian A. Counterman[1]

[1]Department of Biological Sciences, Mississippi State University

[2]Department of Biology, University of Puerto Rico—Rio Piedras

[3]Present address: Division of Biological Sciences, University of Montana, Missoula, MT

*Corresponding author: E-mail: luis.rodriguezcaro@umontana.edu.

Accepted: November 14, 2019

Data deposition: RNA-seq data used for annotation are deposited in the SRA database. Accession bioProject ID: PRJNA587792.

Abstract

Comparisons of high-quality, reference butterfly, and moth genomes have been instrumental to advancing our understanding of how hybridization, and natural selection drive genomic change during the origin of new species and novel traits. Here, we present a genome assembly of the Southern Dogface butterfly, *Zerene cesonia* (Pieridae) whose brilliant wing colorations have been implicated in developmental plasticity, hybridization, sexual selection, and speciation. We assembled 266,407,278 bp of the *Z. cesonia* genome, which accounts for 98.3% of the estimated 271 Mb genome size. Using a hybrid approach involving Chicago libraries with Hi-Rise assembly and a diploid Meraculous assembly, the final haploid genome was assembled. In the final assembly, nearly all autosomes and the Z chromosome were assembled into single scaffolds. The largest 29 scaffolds accounted for 91.4% of the genome assembly, with the remaining ~8% distributed among another 247 scaffolds and overall N50 of 9.2 Mb. Tissue-specific RNA-seq informed annotations identified 16,442 protein-coding genes, which included 93.2% of the arthropod Benchmarking Universal Single-Copy Orthologs (BUSCO). The *Z. cesonia* genome assembly had ~9% identified as repetitive elements, with a transposable element landscape rich in helitrons. Similar to other Lepidoptera genomes, *Z. cesonia* showed a high conservation of chromosomal synteny. The *Z. cesonia* assembly provides a high-quality reference for studies of chromosomal arrangements in the Pierid family, as well as for population, phylo, and functional genomic studies of adaptation and speciation.

Key words: Lepidoptera, de novo assembly, comparative genomics, Hi-Rise assembly.

Introduction

Butterflies and moths constitute a monophyletic group of insects characterized by their astonishing diversity in wing color patterns, behaviors, and ecology. Composed of more than 170,000 species, the order Lepidoptera provides a diverse array of phenotypic variation that serves as a model system for studies in genetics, development, ecology, and evolutionary biology (Mavárez et al. 2006; Fujii and Shimada 2007; Bonebrake et al. 2010; Hof et al. 2016; Van Belleghem et al. 2017).

Several aspects of lepidopteran genomes make them distinctly attractive among arthropods and eukaryotes in general: genome sizes are relatively small (~246–809 Mb), base composition is A–T rich (~68%) (Triant et al. 2018), structurally they are simple compared with other eukaryotes, and they exhibit a high degree of chromosomal synteny (Papa et al. 2008; Beldade et al. 2009; Yasukochi et al. 2009; Triant et al. 2018). A phylogenetic analysis of Lepidopteran karyotypes revealed that the ancestral number of chromosomes in Lepidoptera was most likely 31, with derived states due to chromosomal fusions documented in Nymphalids (Saura et al. 2013). However, across much of the phylogeny very few chromosomal rearrangements have been documented across the 140 Myr of divergence (Ahola et al. 2014). Comparisons between the genomes of *Bombyx* silk moths and *Heliconius* butterflies confirm that chromosomal organization is broadly conserved between the two lineages (Pringle et al. 2007; Heliconius Genome Consortium 2012), supporting high conservation of synteny across Lepidoptera.

Transposable elements (TE) are abundant in lepidopteran genomes. Lepidoptera TEs have been important sources of genetic tools, such as the piggybac transposon that was initially discovered in cabbage looper moth genomes (Cary et al.

1989), and has now been engineered as a tool for gene-editing in mammalian genomes (Ding et al. 2005; Wilson et al. 2007). TEs have also been important sources of natural variation, as seen in the classic example of industrial melanism in peppered moths, where a TE insertion resulted in darker wing coloration and better camouflage, thereby saving the moth population and providing one of the best modern examples of natural selection in action (Cook and Saccheri 2013; Hof et al. 2016). A recent survey of mobile DNA in arthropods also identified the order Lepidoptera as a hotspot for potential horizontal transfer events, in association with the widespread presence of baculovirus infections, a group of viruses known for their ability to transport host TEs (Reiss et al. 2019). Collectively, the TEs that compose a large amount of lepidopteran genomes offer a remarkable array of opportunities to better understand the evolution of genome architecture and function.

Lepidopteran genomes are constantly being sequenced. However, sequencing efforts have largely concentrated on a few specific clades. As of 2018, there were 48 lepidopteran genome assemblies available, coming from only 8 of the 43 lepidopteran Superfamilies (Triant et al. 2018). Among butterflies, the best represented clade is the family Nymphalidae with more than 20 genomes available, some of which have chromosomal assemblies constructed by pedigree linkage maps, whereas groups like Papilionidae and Pieridae are represented only by a few genomes that range widely in quality. Most of these genomes are accessible through the Lepbase database (www.lepbase.org; last accessed November 30, 2019), a central repository for Lepidoptera genomes that provides an Ensemble genome browser, assembly statistics, and basic sequence analysis tools (Challi et al. 2016).

The southern dogface, *Zerene cesonia*, is a Pierid butterfly distributed across the Americas that exhibits interesting characteristics such as sexually dimorphic development, structural coloration, and developmental plasticity (Fenner 2019). The currently available Pierid genomes are mostly low-coverage draft assemblies, and only six species have genome sequences available (Cong et al. 2016; Shen et al. 2016; Talla et al. 2017). With the aim to generate high-quality genomic resources for the study of *Z. cesonia* and other Pierids, we sequenced the genome of *Z. cesonia* using the Chicago protocol (Putnam et al. 2016) with high-sequencing coverage. We provide RNA-seq based gene annotations and have compared the resulting assembly to representative genomes from other lepidopteran lineages. Our results provide high-quality genomic resources for further understanding the ecology, development, and evolution of Pierid butterflies.

Materials and Methods

DNA Sampling and Sequencing

Three female *Z. cesonia* individuals from a colony established at Mississippi State University were frozen in liquid nitrogen 24 h after pupation and sent to the Dovetail Genomics Center. DNA was extracted from two male pupae using Qiagen Genomic-tip DNA isolation protocol. Two Illumina pair-end 150-bp libraries were prepared using the TruSeq DNA PCR-free kit with insert sizes of 550 and 350 bp for DNA shotgun sequencing with HiSeq 2500 and HiSeqX technologies, respectively.

Genome Assembly

Reads were preprocessed using Trimmomatic (Bolger et al. 2014). First, ILLUMINACLIP was used to remove sequencing adapters. Next all bases with quality scores <20 were removed from the leading and trailing ends of the reads. A sliding window of 13 bp from the end of the read was then used, truncating the read when the average quality dropped <20. After this process, any read shorter than 23 bp was rejected.

Genome size was estimated from *k*-mer frequency method (Guo et al. 2015) and flow cytometry. The *k*-mer distribution with *k* equal to 79 bp best fitted the constrained heterozygous model and was therefore used to estimate the genome size. Genome size was also independently estimated using flow cytometry from DNA isolated from the heads of four *Z. cesonia* individuals and *Drosophila virilis* DNA (330 Mb genome size) as reference.

A preliminary genome assembly was generated using Meraculous (Chapman et al. 2011) for contig reconstruction, and the Chicago protocol for scaffolding. The Chicago protocol generates proximity ligation libraries using reconstituted chromatin as a substrate and then creates scaffolds using the HighRise (HiRiSE) software (Putnam et al. 2016). Both procedures were performed by Dovetail genomics. This initial preliminary assembly, named Z_cesonia_v-0.1, was constructed with a single DNA library (550 bp insert size) using Meraculous in diploid mode 1. Because this strategy failed to capture the sex (Z) chromosome, we generated a second assembly named Z_cesonia_v-0.2 with increased coverage and using Meraculous in diploid mode 2 to increase the probability of capturing all chromosomes. As Z_cesonia_v-0.2 was a diploid assembly, we used the Haplomerger pipeline (Huang et al. 2012) to assemble a single reference haplome for Z_cesonia_v-0.2. To confirm that diploid regions were successfully merged by Haplomerger, we aligned Z_cesonia_v-0.2 with Z_cesonia_v-0.1 using MUMmer (Marçais et al. 2018) and used a custom python script was designed to remove duplicate portions of scaffolds that Haplomerger failed to detect. The last step was to order and orient scaffolds from Z_cesonia_v-0.2 using the Chicago library preparation and HiRise assembly information of Z_cesonia_v-0.1. For this, we used RaGOO (Alonge et al. 2019) with zerene_cesonia_0.1 as a reference. A detailed report describing the procedures used to produce zerene_cesonia_v-1.0, including the scripts and coordinates used for duplicate removal can be found in

the GitHub repository for this project (https://github.com/LF-Rodriguez/Z_cesonia_genome_assembly_2019/tree/master/supplemental; last accessed November 30, 2019).

The mitochondrial genome was assembled using the libraries above and NOVOplasty2.7.2 (Dierckxsens et al. 2017). The Novoplasty assembler was run using recommended parameters from the documentation, with sequencing adapters trimmed from reads, and a *Z. cesonia* partial CDS of the cytochrome oxidase 1 subunit (GenBank accession no. GU164697) for the input seed sequence.

Genome Annotation

Repetitive elements (REs) were masked with RepeatMasker (www.repeatmasker.org; last accessed November 30, 2019) using a customized library containing repeats from all hexapoda including all annotated repeats from *Heliconius* butterflies updated in 2007. We used the Maker-2 pipeline (Holt and Yandell 2011) to annotate the genome, using a transcriptome of *Z. cesonia* assembled de novo from wing disc, thorax, and head tissues (SRA bioProject ID: PRJNA587792) as evidence for mRNA and exon boundaries.

To explore chromosomal synteny, we performed a MUMmer alignment of the 29 largest scaffolds of Z_cesonia_v-1.0 (91.4% of the assembly) to the 20 autosomes of *H. erato* (v.1.0), which contains chromosome information inferred from pedigree linkage maps. The scripts and references used for genome alignments and chromosome assignment are available at the GitHub repository for this project (https://github.com/LF-Rodriguez/Z_cesonia_genome_assembly_2019; last accessed November 30, 2019). Protein-coding genes in the mitochondrial genome assembly were identified and annotated using sequence similarity with *Colias erate* (GenBank accession no. NC_027253).

Results and Discussion

DNA Sampling and Sequencing

We obtained a total of 191,008,162 reads from the first DNA library (HiSeq 2500) of which 91.8% passed the trimmomatic filter with a final average length of 142.4 bp. The second DNA library (Hi-SeqX), generated 387,729,917 reads of which 97.9% passed the trimmomatic filters with final average length of 143.2 bp.

Genome Assembly

Generating an accurate estimate of the genome size of *Z. cesonia* allowed us to evaluate the completeness of the assemblies. The *k*-mer distribution analysis estimated a genome size of 271 Mb using a *k* of 79 which best fitted the distribution of the heterozygous model. Using flow cytometry and *Drosophila virilis* as a reference, the *Z. cesonia* genome was estimated to be 303 Mb, with a SD of 6 Mb. For characterizing assembly metrics, we used the 271 Mb *k*-mer genome size estimate.

The preliminary assembly, Z_cesonia_v-0.1, was constructed with ~195× coverage into large scaffolds, most of which were near the expected chromosomal sizes, and covered a total of 229,153,833 bases (84.5% of the estimated genome size). This assembly conducted by Dovetail benefitted from the Chicago library prep and Hi-Rise assembly, but had insufficient coverage to assemble the Z chromosome, and a relatively large number of ambiguous bases (i.e., *N*) inserted in the genome (12.1%). The second assembly, Z_cesonia_v-0.2, was constructed with increased coverage (~322×) using a second male individual, and the Meraculous assembler in diploid mode 2. This resulted in a diploid genome assembly of ~516.4 Mb, with most autosome sequences present twice as expected from the Meraculous diploid mode 2 configuration. This approach allowed us to assemble the full Z chromosome in a single scaffold (12.4 Mb). Merging the diploid Z_cesonia_v-0.2 assembly into a haplome resulted in a final haploid assembly (Z_cesonia_v-1.0) of size 266,407,278 bp, which is 98.2% of the expected genome size (271 Mb), represented by a total of 276 scaffolds with an N50 of 9.2 Mb (fig. 1). Notably, 91.4% of the genome assembly is contained in the 29 largest scaffolds.

The assembly of *H. erato's* genome is a valuable reference for butterfly genomics because of its chromosomal assemblies that were constructed empirically using pedigree-based linkage maps (Van Belleghem et al. 2017). Assuming high-synteny conservation between Pierid and Nymphalid genomes, we mapped the genome of *Z. cesonia* to the genome of *H. erato* to determine the identity of the scaffolds. According to previous studies on lepidopteran genomes (Maeki 1960; Saura et al. 2013), the genome of *Z. cesonia* has 31 chromosomes (29 autosomes). We found that the 29 largest scaffolds of the assembly largely show homology to only one or two chromosomes in *H. erato* (fig. 1*A*). This suggests that these 29 scaffolds likely reflect near full assemblies of 28 autosomes and the Z chromosome. This also suggests that, similar to other major clades of Lepidoptera, Pierid genomes also exhibit high conservation of chromosomal synteny.

The *Z. cesonia* mitochondrial genome assembled into a single contiguous sequence of 15,138 bp with a GC-content of ~19.2%, and the positions of 13 protein-coding genes identified.

Genome Annotation

Extensive research in butterfly genomics has generated a thorough repertoire of annotated genomic features including coding sequences and a curated library of REs for butterflies (Challi et al. 2016). Taking advantage of those resources, we used Maker (Holt and Yandell 2011) to annotate the genome and RepeatMasker (Tarailo-Graovac and Chen 2009) to identify and mask REs.

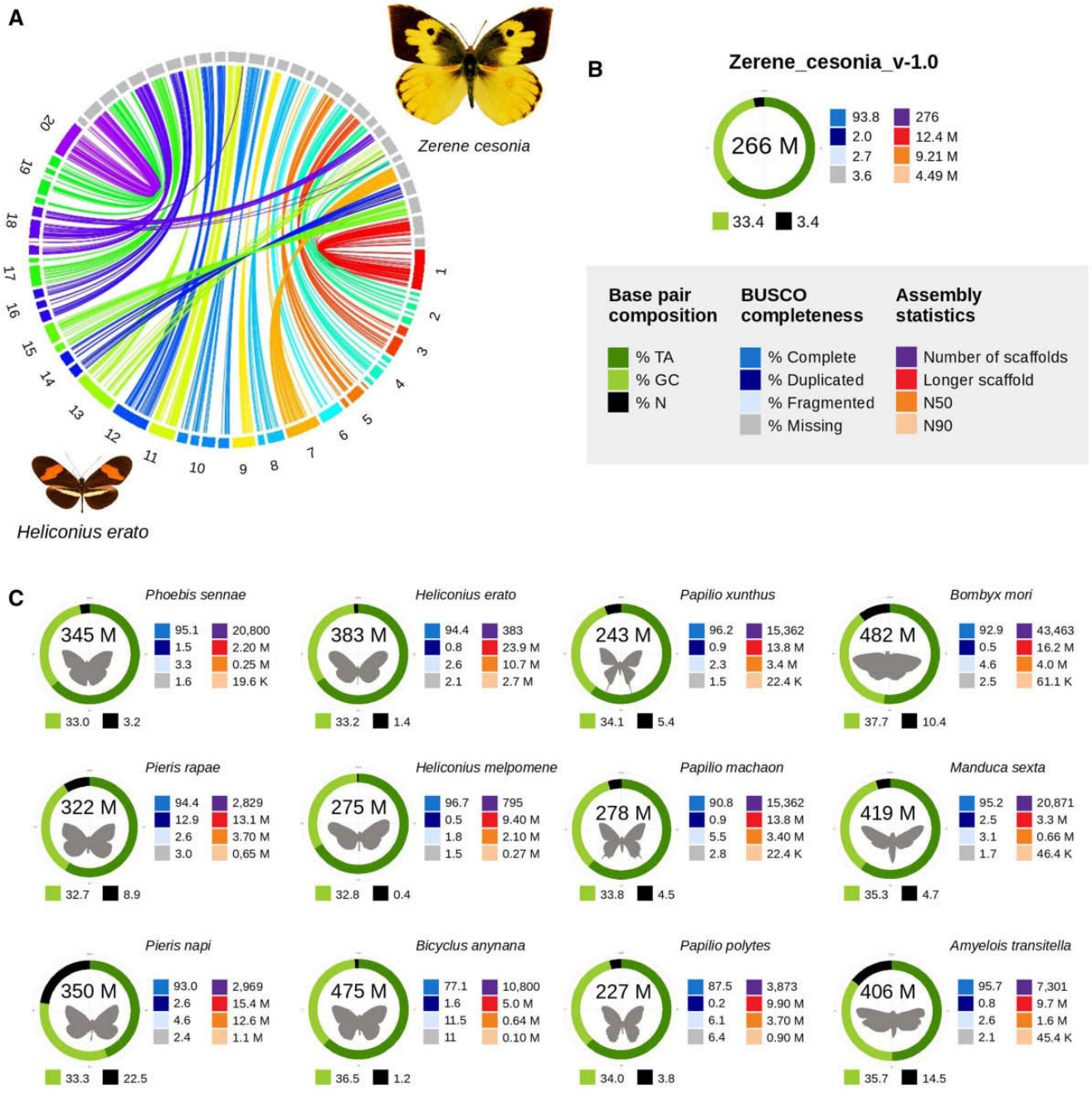

Fig. 1.—*Zerene cesonia* genome assembly in comparison to other lepidopterans. (*A*) Synteny conservation between *Z. cesonia* and *Heliconius erato* genomes. (*B*) Genome assembly and composition for *Z. cesonia* assemblies. (*C*) Graphical comparison of genome assemblies across butterflies and moths.

We identified 16,442 genes with an average gene span of 5,757 bp and found that the 9.01% of the genome is composed of RE, most of which are helitrons. This is a small portion of RE compared with butterflies like *H. erato* (27.95%; Van Belleghem et al. 2017), *Heliconius melpomene* (25.36%; Ray et al. 2019), the silkworm moth, *Bombyx mori* (35.4%; Osanai-Futahashi et al. 2008), and other pierid butterflies (*Phoebis sennae*—22.7%, and *Pieris rapae*—17.2%; Shen et al. 2016; Talla et al. 2017) (fig. 2*B*). Divergence plots of TE families shows clear differences in TE content among the different Lepidoptera lineages. Divergence of TE families was measured as the percentage of divergent bases for each genomic copy compared with the consensus sequence generated from RepeatMasker.

Although a more accurate RE annotation and standardized divergence metrics are required to make inferences about the history and evolution of REs in the genome (Platt et al. 2016), raw percent divergence is informative to identify overall

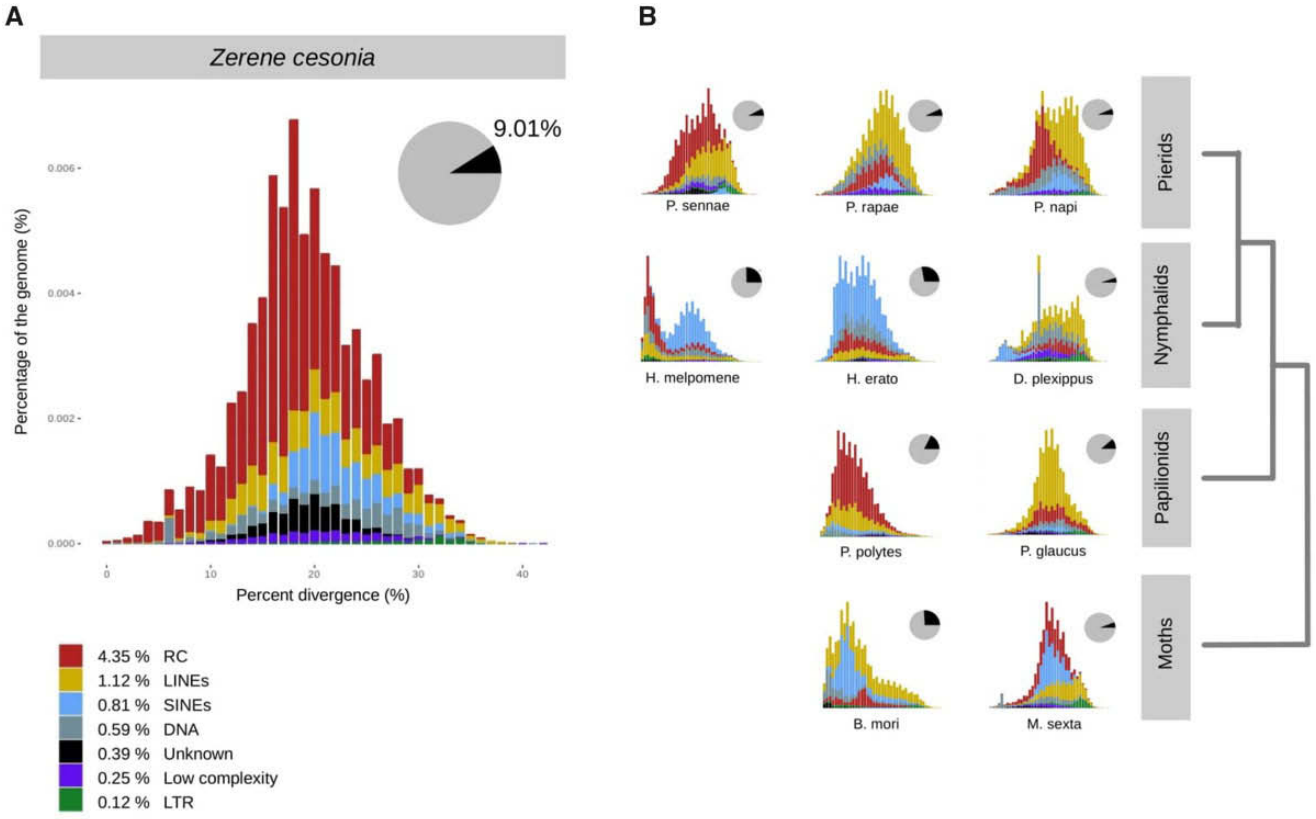

Fig. 2.—Comparison of repetitive element (RE) abundance in *Zerene cesonia* and other lepidopterans. (*A*) Bar plot of RE types abundance versus their divergence. Divergence is estimated as the percentage of divergent sites on the RepeatMasker hits. The pie chart shows the portion of the genome containing REs in black. (*B*) Bar plots of abundance versus divergence of REs in other lepidopterans. All bar plots range from 0% to 40% in the *x* axis.

patterns of RE activity across lineages. The distributions of RE divergence in *Z. cesonia* reflects high divergence of all RE types (peak ~20%) which suggests inactivity of REs in the recent past. This contrasts with the patterns observed in the genus *H. melpomene* (fig. 2B), which is known to have experienced recent transposon activity (Lavoie et al. 2013) and substantial genome diversification due to TE activity (Ray et al. 2019). When comparing the TE landscape among lepidopteran genomes, two patterns can be identified. First, LTR divergence peak is close to 20% in most groups included, suggesting that their activity ceased before the split between moths and butterflies, which is consistent with the findings of previous studies of RE activity in Lepidoptera (Lavoie et al. 2013; Talla et al. 2017; Reiss et al. 2019). Secondly, helitrons and LINEs are the most abundant type of REs among Pierid butterflies, but overall the species included here show low abundance of REs (range: 6.17–22.7%) and no signs of RE activity in their genomes in the recent past. Additionally, *P. sennae* and *Z. cesonia* show a relatively large abundance of undescribed RE's compared with other lepidopterans, which reflects an incomplete characterization of Pierid-specific RE's.

Together, these results provide a comprehensive summary of the composition and architecture of the genome of *Z. cesonia*. The assembly here presented covered 98.2% of the genome with chromosome sized scaffolds and provide an initial characterization of the TE landscape of *Z. cesonia* and other lepidopterans.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the National Science Foundation (NSF) (1736026 and 1755329 to B.A.C.). For computational resources, we thank the University of Puerto Rico, the Puerto Rico INBRE Grant P20 GM103475 from the National Institute for General Medical Sciences (NIGMS), a component of the National Institutes of Health (NIH) and Matt Brown's Lab, and awards 1010094 and 1002410 from the EPSCoR program of

the NSF. We thank Brice Noonan for sequencing resources for the *Z. cesonia* transcriptome.

Literature Cited

Ahola V, et al. 2014. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nat Commun. 5(1):4737.

Alonge M, et al. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 20(1):224.

Beldade P, Saenko SV, Pul N, Long AD. 2009. A gene-based linkage map for *Bicyclus anynana* butterflies allows for a comprehensive analysis of synteny with the Lepidopteran reference genome. PLoS Genet. 5(2):e1000366.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Bonebrake TC, Ponisio LC, Boggs CL, Ehrlich PR. 2010. More than just indicators: a review of tropical butterfly ecology and conservation. Biol Conserv. 143(8):1831–1841.

Cary LC, et al. 1989. Transposon mutagenesis of Baculoviruses: analysis of Trichoplusia Ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. Virology 172(1):156–169.

Challi RJ, Kumar S, Dasmahapatra KK, Jiggins CD, Blaxter M. 2016. Lepbase: the Lepidopteran genome database. *BioRxiv*, June, 056994.

Chapman JA, et al. 2011. Meraculous: De Novo Genome Assembly with Short Paired-End Reads. PLOS ONE 6(8):e23501.

Cong Q, et al. 2016. Speciation in cloudless sulphurs gleaned from complete genomes. Genome Biol Evol. 8(3):915–931.

Cook LM, Saccheri IJ. 2013. The peppered moth and industrial melanism: evolution of a natural selection case study. Heredity 110(3):207–212.

Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487(7405):94–98.

Dierckxsens N, Mardulyn P, Smits G. 2017. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 45(4):e18–e18.

Ding S, et al. 2005. Efficient transposition of the PiggyBac (PB) transposon in mammalian cells and mice. Cell 122(3):473–483.

Fujii T, Shimada T. 2007. Sex determination in the silkworm, *Bombyx mori*: a female determinant on the W chromosome and the sex-determining gene cascade. Semin Cell Dev Biol. 18(3):379–388.

Guo LT, et al. 2015. Flow cytometry and K-mer analysis estimates of the genome sizes of *Bemisia tabaci* B and Q (Hemiptera: Aleyrodidae). Front Physiol. 6:144.

Hof A. E V, et al. 2016. The industrial melanism mutation in British peppered moths is a transposable element. Nature 534(7605):102–105.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12(1):491.

Huang S, et al. 2012. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Res. 22(8):1581–1588.

Lavoie CA, Platt RN, Novick PA, Counterman BA, Ray DA. 2013. Transposable element evolution in *Heliconius* suggests genome diversity within Lepidoptera. Mob DNA. 4(1):21.

Maeki K. 1960. Studies of the chromosomes of North American Rhopalocera 2. Hesperildle, Megathymidle, and Pieridle. J Lepid Soc. 14(1):37–57.

Marçais G, et al. 2018. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol. 14(1):e1005944.

Mavárez J, et al. 2006. Speciation by hybridization in *Heliconius* butterflies. Nature 441(7095):868–871.

Osanai-Futahashi M, Suetsugu Y, Mita K, Fujiwara H. 2008. Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, *Bombyx mori*. Insect Biochem Mol Biol. 38(12):1046–1057.

Papa R, et al. 2008. Highly conserved gene order and numerous novel repetitive elements in genomic regions linked to wing pattern variation in *Heliconius* butterflies. BMC Genomics 9(1):345.

Platt RN, Blanco-Berdugo L, Ray DA. 2016. Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biol Evol. 8(2):403–410.

Pringle EG, et al. 2007. Synteny and chromosome evolution in the Lepidoptera: evidence from mapping in *Heliconius melpomene*. Genetics 177(1):417–426.

Putnam NH, et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26(3):342–350.

Ray DA, et al. 2019. Simultaneous TE analysis of 19 Heliconiine butterflies yields novel insights into rapid TE-based genome diversification and multiple SINE births and deaths. Genome Biol Evol. 11(8):2162–2177.

Reiss D, et al. 2019. Global survey of mobile DNA horizontal transfer in arthropods reveals Lepidoptera as a prime hotspot. PLoS Genet. 15(2):e1007965.

Saura A, Schoultz BV, Saura AO, Brown KS. 2013. Chromosome evolution in neotropical butterflies: chromosome evolution in neotropical butterflies. Hereditas 150(2–3):26–37.

Shen J, et al. 2016. Complete genome of *Pieris rapae*, a resilient alien, a cabbage pest, and a source of anti-cancer proteins. F1000Res. 5(November):2631.

Talla V, et al. 2017. Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (Leptidea) butterflies. Genome Biol Evol. 9(10):2491–2505.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. Chapter. 4(March):Unit 4.10.

Triant DA, Cinel SD, Kawahara AY. 2018. Lepidoptera genomes: current knowledge, gaps and future directions. Curr Opin Insect Sci. 25(February):99–105.

Van Belleghem SM, et al. 2017. Complex modular architecture around a simple toolkit of wing pattern genes. Nat Ecol Evol. 1(3):0052.

Wilson MH, Coates CJ, George AL. 2007. PiggyBac transposon-mediated gene transfer in human cells. Mol Ther. 15(1):139–145.

Yasukochi Y, et al. 2009. Extensive Conserved Synteny of Genes between the Karyotypes of Manduca Sexta and Bombyx Mori Revealed by BAC-FISH Mapping. PLos One 4(10): e7465.

Associate editor: Adam Eyre-Walker