Journal of Heredity, 2023, **114**, 698–706
https://doi.org/10.1093/jhered/esad043
Advance access publication 10 July 2023
Genome Resources

Genome Resources

A draft reference genome assembly of the Pipevine Swallowtail butterfly, *Battus philenor hirsuta*

Samridhi Chaturvedi[1,2], Merly Escalona[3], Mohan P.A. Marimuthu[4], Oanh Nguyen[4], Noravit Chumchim[4], Colin W. Fairbairn[3], William Seligmann[3], Courtney Miller[5], H. Bradley Shaffer[5,6] and Noah K. Whiteman[1,7]

[1]Department of Integrative Biology, University of California, 142 Weill Hall #3200, Berkeley, CA 94720, United States,
[2]Department of Ecology and Evolutionary Biology, Tulane University, New Orleans, LA 70118, United States,
[3]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, United States,
[4]DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California, Davis, CA 95616, United States,
[5]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095-7239, United States,
[6]La Kretz Center for California Conservation Science, Institute of the Environment and Sustainability, University of California, Los Angeles, CA 90095-7239, United States,
[7]Department of Molecular and Cell Biology, University of California, 142 Weill Hall #3200, Berkeley, CA 94720, United States

Address correspondence to Samridhi Chaturvedi at the address above, or e-mail: schaturvedi@tulane.edu.
Address correspondence to Noah K. Whiteman at the address above, or e-mail: whiteman@berkeley.edu.

Corresponding Editor: Beth Shapiro

Abstract

The California Pipevine Swallowtail Butterfly, *Battus philenor hirsuta*, and its host plant, the California Pipevine or Dutchman's Pipe, *Aristolochia californica* Torr., are an important California endemic species pair. While this species pair is an ideal system to study co-evolution, genomic resources for both are lacking. Here, we report a new, chromosome-level assembly of *B. philenor hirsuta* as part of the California Conservation Genomics Project (CCGP). Following the sequencing and assembly strategy of the CCGP, we used Pacific Biosciences HiFi long reads and Hi-C chromatin proximity sequencing technology to produce a de novo assembled genome. Our genome assembly, the first for any species in the genus, contains 109 scaffolds spanning 443 mega base (Mb) pairs, with a contig N50 of 14.6 Mb, a scaffold N50 of 15.2 Mb, and BUSCO complete score of 98.9%. In combination with the forthcoming *A. californica* reference genome, the *B. philenor hirsuta* genome will be a powerful tool for documenting landscape genomic diversity and plant–insect co-evolution in a rapidly changing California landscape.

Key words: California Conservation Genomics Project, genomics, Lepidoptera, Papilionidae

Introduction

The California Pipevine Swallowtail Butterfly (*Battus philenor hirsuta*) is a disjunct lineage from the far more widespread Pipevine Swallowtail Butterfly (*B. philenor*), which has a broad range throughout the southern and eastern United States from Florida to Massachusetts, west to Kansas, Arizona, and southern California. *Battus philenor hirsuta* on the other hand, occurs only where its host plant occurs in north-central California, including the San Francisco Bay Area east to Sacramento, north to the edge of the Klamath ranges near Redding, and south along the western flank of the Central Sierra foothills and adjoining riparian corridors of the eastern Central Valley (Fordyce 2000; Fordyce and Agrawal 2001). There are several interesting life history traits of the California Pipevine Swallowtail that distinguish it from the Pipevine Swallowtail Butterfly, including gregarious feeding by larvae (Fordyce and Agrawal 2001). *Battus philenor hirsuta* entirely depends on its host plant, the California Pipevine or Dutchman's Pipe (*Aristolochia californica* Torr.) and anecdotal information suggests that exotic *Aristolochia* species grown in gardens pose a problem—they are not viable hosts for the larvae, yet the adult females will lay eggs on the plants, leading to larval death (Fig. 1).

Battus philenor hirsuta and its host plant *A. californica* are an iconic, California endemic species pair. The two species have likely co-evolved in an antagonistic fashion across the landscape of northern California and as a result, the butterfly may be particularly susceptible to a phenomenon known as co-extinction—where the extinction of one species results in the loss of those dependent on it. Their range overlaps with the human-dominated landscape comprising urban, suburban, and agricultural lands as well as more ecologically intact areas. Tightly coupled plant–insect systems have been identified as particularly susceptible to anthropogenic stressors, and there is a growing need to use genomic tools to predict community level responses to changing climate

Received April 18, 2023; Accepted July 8, 2023

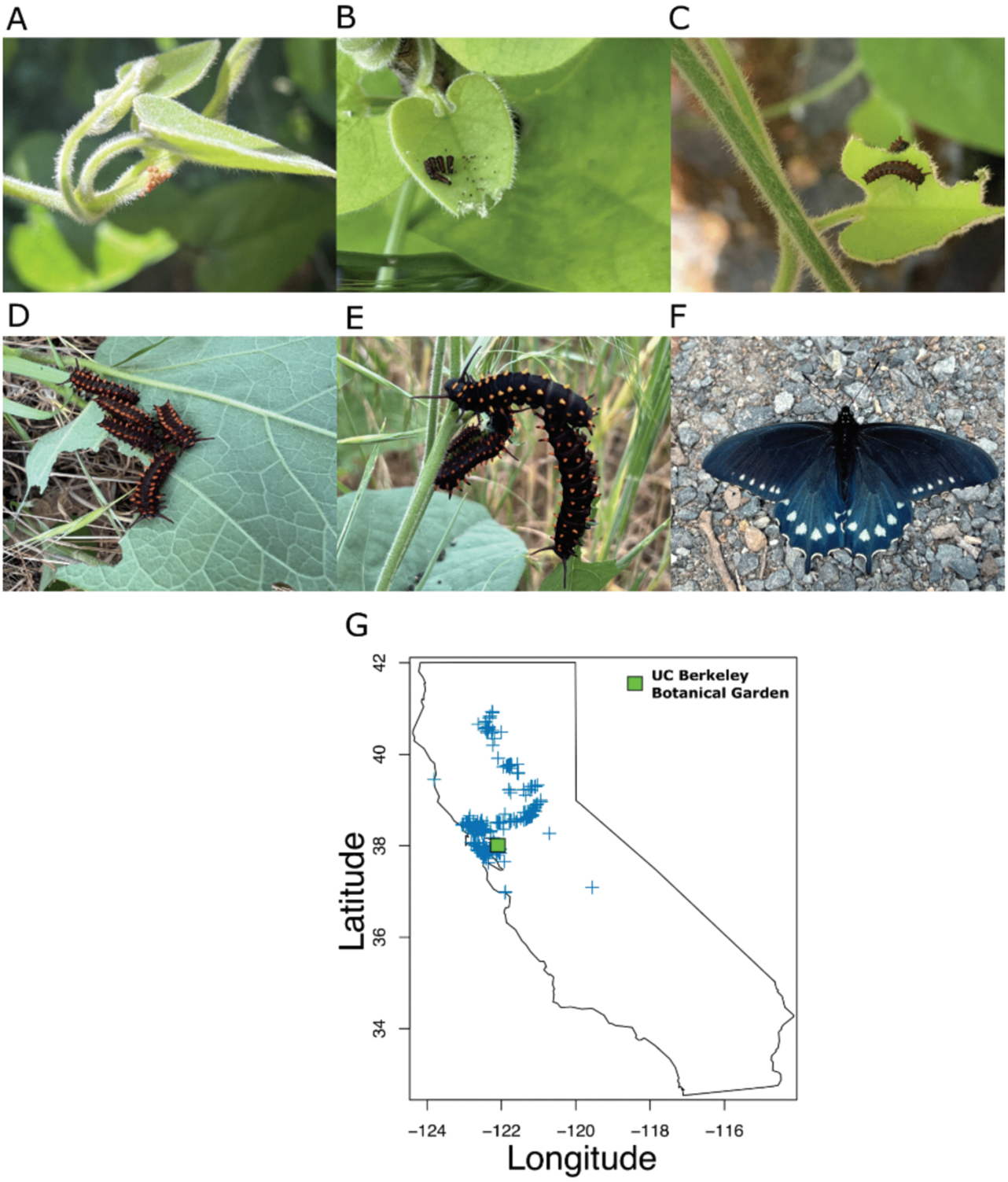

Fig. 1. *Battus philenor hirsuta* is a swallowtail butterfly species endemic to Northern California. All the life stages of the caterpillars occur on the specific host plant, *Aristolochia californica*, as shown in the images in this figure. (A) Eggs on *A. californica*, (B–E) First, second, third, and fifth instar caterpillars on *A. californica* (F) *B. philenor hirsuta* adult male, (G) A map of the species range including locations (denoted as plus signs in the map) of specimens collected based on personal observations by S. Chaturvedi. The sampling location (University of California, Berkeley Botanical Garden) of the specimen used for the genome assembly is marked in a green square. All photographs are copyright of S. Chaturvedi.

(Northfield and Ives 2013). Moreover, variations in climatic variables such as precipitation and temperature can affect plant–herbivore interactions by changing chemical make-ups, physiology, genetic effects, and range dynamics of both partners in the interaction (Leimu et al. 2012).

Here, we report the first chromosome-level genome assembly for *B. philenor hirsuta*, sequenced and assembled as part of the California Conservation Genomics Project (CCGP). This genome assembly is first of any species in the genus. Our sequencing approach helps us generate an expected

30-fold genome coverage which accounts for the 376 Mb genome size of the Tiger Swallowtail (*Papilio glaucus*) (Cong et al. 2015). The overarching goal of the CCGP is to discover patterns of genomic diversity across the state of California by sequencing the complete genomes of approximately 150 carefully selected species (Shaffer et al. 2022). The ongoing efforts of the CCGP provide an unparalleled opportunity to use the reference genome sequences of both *B. philenor hirsuta* and *A. californica* and a landscape genomics approach to gain genome-level understanding of demographic history, gene flow, and inbreeding to assess the current status of both these species, as well as potentially identify genes that may be important in their coevolutionary history. This is the first completed reference genome for the *B. philenor* butterflies and this genome assembly will provide a foundational resource for future studies on the unique ecology, biogeography, evolutionary history, behavior, and conservation of this species. It is also the first butterfly produced by the CCGP, and fills an important phylogenetic gap in the catalog of reference genomes for California (Toffelmier et al. 2022).

Methods

Biological materials

Two adult female *B. philenor hirsuta* were collected. The first one (BFH_F_UCBG), was collected on 8 March 2021 from University of California, Berkeley Botanical Garden in Berkeley, CA, United States (37.87118, −122.238632). The second (BFH_F_UCBO) was sampled on April 1st, 2021 from a private backyard in Berkeley, CA, United States (37.887427, −122.2712976) as a caterpillar and then reared to adulthood. The butterflies were collected near the butterfly's only host plant—California Pipevine (*Aristolochia californica*) and immediately flash frozen in liquid nitrogen. Each adult butterfly was cut in half and shipped overnight on dry ice to the UC Davis Genome Center (Davis, CA) and UC Santa Cruz (Santa Cruz, CA) sequencing cores.

Nucleic acid library preparation

High molecular weight (HMW) genomic DNA (gDNA) was extracted from 25 mg of abdominal tissue from one adult female (sample ID: BFH_F_UCBO) using the Nanobind Tissue Big DNA kit per the manufacturer's instructions (Pacific BioSciences—PacBio, Menlo Park, CA) with the following modifications. After the second resuspension step, we pelleted the tissue homogenate by centrifuging at 18,000 × *g* (4 °C for 5 min) to remove the residual wash buffer and performed the lysis step with 1.25× reaction volume. The DNA purity was estimated using absorbance ratios (260/280 = 1.91 and 260/230 = 2.12) on a NanoDrop ND-1000 spectrophotometer. The final DNA yield (3 µg) was quantified using the Quantus Fluorometer (QuantiFluor ONE dsDNA Dye assay; Promega, Madison, WI). We estimated the size distribution of the HMW DNA using the Femto Pulse system (Agilent, Santa Clara, CA) and found that 60% of the fragments were 90 Kb or longer.

The HiFi SMRTbell library was constructed using the SMRTbell Express Template Prep Kit v2.0 (PacBio; Cat. #100-938-900) according to the manufacturer's instructions. HMW gDNA was sheared to a target DNA size distribution between 12 and 20 kb using Diagenode's Megaruptor 3 system (Diagenode, Belgium; cat. B06010003). The sheared gDNA was concentrated using 1.8× of AMPure PB beads (PacBio; Cat. #100-265-900) for the removal of single-strand overhangs at 37 °C for 15 min, followed by further enzymatic steps of DNA damage repair at 37 °C for 30 min, end repair and A-tailing at 20 °C for 10 min and 65 °C for 30 min, and ligation of overhang adapters v3 at 20 °C for 60 min. The SMRTbell library was purified and concentrated with 0.45× AMPure PB beads for size selection with 2.2× of 40% v/v diluted AMPure PB beads to remove short SMRTbell templates <3 kb in length. The 12 to 20 kb average HiFi SMRTbell library was sequenced at UC Davis DNA Technologies Core (Davis, CA) using 0.5 8M SMRT cell, Sequel II sequencing chemistry 2.0, and 30-h movies each on a PacBio Sequel II sequencer.

The Omni-C library was prepared using the Dovetail Omni-C Kit (Dovetail Genomics, Scotts Valley, CA) according to the manufacturer's protocol with slight modifications. First, specimen tissue (wing and head, Sample ID: BFH_F_UCBG) was thoroughly ground with a mortar and pestle while cooled with liquid nitrogen. Subsequently, chromatin was fixed in place in the nucleus. The suspended chromatin solution was then passed through 100 µm and 40 µm cell strainers to remove large debris. Fixed chromatin was digested under various conditions of DNase I until a suitable fragment length distribution of DNA molecules was obtained. Chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter-containing ends. After proximity ligation, crosslinks were reversed and the DNA was purified from proteins. Purified DNA was treated to remove biotin that was not internal to ligated fragments. An NGS library was generated using an NEB Ultra II DNA Library Prep kit (NEB, Ipswich, MA) with an Illumina compatible y-adaptor. Biotin-containing fragments were then captured using streptavidin beads. The post-capture product was split into two replicates prior to PCR enrichment to preserve library complexity with each replicate receiving unique dual indices. The library was sequenced at Vincent J. Coates Genomics Sequencing Lab (Berkeley, CA) on an Illumina NovaSeq 6000 platform (Illumina, San Diego, CA) to generate approximately 100 million 2 × 150 bp read pairs per GB genome size, which is expected to generate a 30-fold genome coverage. The expected coverage is taking into account a genome size around 376 Mb, estimated from Tiger Swallowtail (*Papilio glaucus*) another species in the Family (Cong et al. 2015).

Nuclear genome assembly

We assembled the genome of *B. philenor hirsuta* following the CCGP assembly pipeline Version 6.0, as outlined in Table 1, which lists the tools and non-default parameters used in the assembly. The pipeline uses PacBio HiFi reads and Omni-C data to produce high quality and highly contiguous genome assemblies. First, we removed the remnant adapter sequences from the PacBio HiFi dataset using HiFiAdapterFilt (Sim et al. 2022) and generated the initial haplotype-resolved diploid genome assembly using HiFiasm (Cheng et al. 2022) on Hi-C mode, with the filtered PacBio HiFi reads and the Omni-C dataset. We then aligned the Omni-C data to both assemblies following the Arima Genomics Mapping Pipeline (https://github.com/ArimaGenomics/mapping_pipeline) and then scaffolded both assemblies with SALSA (Ghurye et al. 2017, 2019).

Table 1. Assembly pipeline and software used.

Assembly	Software and options[§]	Version
Filtering PacBio HiFi adapters	HiFiAdapterFilt	Commit 64d1c7b
K-mer counting	Meryl (k = 21)	1
Estimation of genome size and heterozygosity	GenomeScope	2
De novo assembly (contiging)	HiFiasm (Hi-C Mode, –primary, output p_ctg.hap1, p_ctg.hap2)	0.16.1-r375
Scaffolding		
Omni-C data alignment	Arima Genomics Mapping Pipeline	Commit 2e74ea4
Omni-C Scaffolding	SALSA (-DNASE, -i 20, -p yes)	2
Gap closing	YAGCloser (-mins 2 -f 20 -mcc 2 -prt 0.25 -eft 0.2 -pld 0.2)	Commit 0e34c3b
Omni-C contact map generation		
Short-read alignment	BWA-MEM (-5SP)	0.7.17-r1188
SAM/BAM processing	samtools	1.11
SAM/BAM filtering	pairtools	0.3.0
Pairs indexing	pairix	0.3.7
Matrix generation	cooler	0.8.10
Matrix balancing	hicExplorer (hicCorrectmatrix correct --filterThreshold -2 4)	3.6
Contact map visualization	HiGlass	2.1.11
	PretextMap	0.1.4
	PretextView	0.1.5
	PretextSnapshot	0.0.3
Manual curation tools	Rapid curation pipeline (Wellcome Trust Sanger Institute, Genome Reference Informatics Team)	Commit 4ddca450
Genome quality assessment		
Basic assembly metrics	QUAST (--est-ref-size)	5.0.2
Assembly completeness	BUSCO (-m geno, -l insecta)	5.0.0
	Merqury	2020-01-29
Contamination screening		
Local alignment tool	BLAST+ (-db nt, -outfmt "6 qseqid staxids bitscore std," -max_target_seqs 1, -max_hsps 1, -evalue 1e–25)	2.1
General contamination screening	BlobToolKit	2.3.3

Software citations are listed in the text.
[§]Options detailed for non-default parameters.

Both genome assemblies were manually curated by iteratively generating and analyzing their corresponding Omni-C contact maps. To generate the contact maps we aligned the Omni-C data with BWA-MEM (Li 2013), identified ligation junctions, and generated Omni-C pairs using pairtools (Goloborodko et al. 2018). We generated a multi-resolution Omni-C matrix with cooler (Abdennur and Mirny 2020) and balanced it with hicExplorer (Ramírez et al. 2018). We used HiGlass (Kerpedjiev et al. 2018) and the PretextSuite (https://github.com/wtsi-hpag/PretextView; https://github.com/wtsi-hpag/PretextMap; https://github.com/wtsi-hpag/PretextSnapshot) to visualize the contact maps where we identified misassemblies and misjoins, and finally modified the assemblies using the Rapid Curation pipeline from the Wellcome Trust Sanger Institute, Genome Reference Informatics Team (https://gitlab.com/wtsi-grit/rapid-curation). Some of the remaining gaps (joins generated during scaffolding and curation) were closed using the PacBio HiFi reads and YAGCloser (https://github.com/merlyescalona/yagcloser). Finally, we checked for contamination using the BlobToolKit Framework (Challis et al. 2020).

Genome quality assessment

We generated k-mer counts from the PacBio HiFi reads using meryl (https://github.com/marbl/meryl). The k-mer counts were then used in GenomeScope2.0 (Ranallo-Benavidez et al. 2020) to estimate genome features including genome size, heterozygosity, and repeat content. To obtain general contiguity metrics, we ran QUAST (Gurevich et al. 2013). To evaluate genome quality and functional completeness we used BUSCO (Manni et al. 2021) with the Insecta ortholog database (insecta_odb10) which contains 1,367 genes. Assessment of base level accuracy (QV) and k-mer completeness was performed using the previously generated meryl database and merqury (Rhie et al. 2021). We further estimated genome assembly accuracy via BUSCO gene set frameshift analysis using the pipeline described in Korlach et al. (2017).

Measurements of the size of the phased blocks is based on the size of the contigs generated by HiFiasm on HiC mode. We follow the quality metric nomenclature established by (Rhie et al. 2020), with the genome quality code *x.y.P.Q.C*, where, *x* = log10[contig NG50]; *y* = log10[scaffold NG50]; *P* = log10 [phased block NG50]; *Q* = Phred base accuracy QV (quality value); *C* = % genome represented by the first "*n*" scaffolds, following a karyotype of 2*n* = 60 estimated as the median number of chromosomes from other species in the Family Papilionidae (Genome on a Tree—GoaT; tax_tree [*Battus philenor*]). Quality metrics for the notation were calculated on the assembly for haplotype 1.

Results

The Omni-C and PacBio HiFi sequencing libraries generated 94.08 million read pairs and 1.262 million reads respectively. The latter yielded ~33.93-fold coverage (N50 read length 13,111 bp; minimum read length 51 bp; mean read length 11,364 bp; maximum read length of 39,783 bp) based on the Genomescope 2.0 genome size estimation of 422.89 Mb. Based on PacBio HiFi reads, we estimated 0.18% sequencing error rate and 3.37% nucleotide heterozygosity rate. The k-mer spectrum based on PacBio HiFi reads show (Fig. 2A) a bimodal distribution with two major peaks at ~16- and ~33-fold coverage, where peaks correspond to homozygous and heterozygous states of a diploid species. The distribution presented in this k-mer spectrum supports that of a high heterozygosity profile.

The final assembly (ilBatPhil1) consists of two haplotypes and both assemblies are similar in size compared to the estimated value from Genomescope2.0 (Fig. 2A; Pflug et al. 2020). Haplotype 1 consists of 109 scaffolds spanning 443.21 Mb with contig N50 of 14.64 Mb, scaffold N50 of 15.24 Mb, longest contig of 18.71 Mb and largest scaffold of 19.29 Mb. Haplotype 2 was quite similar, with an assembly consisting of 76 scaffolds, spanning 422.8 Mb with contig N50 of 13.7 Mb, scaffold N50 of 15.12 Mb, largest contig 17.86 Mb, and largest scaffold of 17.88 Mb. Assembly statistics are reported in tabular form in Table 2, and graphical representation for the haplotype 1 assembly in Fig. 2B (see Fig. 1, for haplotype 2 graphical representation).

During manual curation, we generated a total of 4 breaks and 10 joins; where two breaks were made on haplotype 1 and two were made on haplotype 2. We also made five joins on each haplotype. We did not close any gaps and we did not remove any contigs due to contaminants. The Omni-C contact maps show that both assemblies are highly contiguous (Fig. 2C and D). We have deposited both assemblies on NCBI (see Table 2 and Data Availability for details).

Haplotype 1 has a BUSCO completeness score of 98.9% using the Insecta gene set, a per base quality (QV) of 66.81, a k-mer completeness of 67.41 and a frameshift indel QV of 54.47. Haplotype 2 has a BUSCO completeness score of 95.9% using the same gene set, a per base quality (QV) of 58.18, a k-mer completeness of 64.30, and a frameshift indel QV of 54.34. The Omni-C contact map shows that both assemblies are highly contiguous with some chromosome-length scaffolds (Fig. 2C and D). We have deposited scaffolds corresponding to both primary and alternate haplotype (see Table 2 and Data availability for details).

Discussion

This genome assembly of the California Pipevine Swallowtail Butterfly adds to the growing genomic resources for swallowtail butterflies in the family Papilionidae. The current genome assembly of *B. philenor hirsuta* is the fourteenth species for which a high coverage whole genome has been sequenced and assembled in the Papilionidae, the tenth in the subfamily Papilioninae, the third in the tribe Troidini and the first species in the genus *Battus* [however, see Allio et al. (2020) for low-coverage whole genome assemblies for 41 Papilionidae species based on only Illumina data]. We use the available whole genome assemblies of several species in the family Papilionidae to provide comparative metrics to assess the quality of the genome assembly presented here.

We present a summary of metrics of published genomes of species in subfamily Papilioninae in Table 3. The *B. philenor hirsuta* genome assembly presented here has the highest contig N50 at 14.6 Mb [contig N50 of the Scarce Swallowtail, *Iphiclides podalirius* genome is 5.2 Mb (Mackintosh et al. 2022), Table 3], and one of the highest scaffold N50 values at 15. 2 Mb [range of other taxa is 230 kb to 15.5 Mb (He et al. 2022), Table 3]. The current genome assembly also has the highest BUSCO completeness scores at 98.9% (range of scores for other taxa 89.9% to 98.1% (He et al. 2022; Mackintosh et al. 2022), Table 3]. Based on the available whole genome assemblies of the butterflies of subfamily Papilioninae, the genome sizes of these butterflies range between 214 Mb and 550 Mb (Table 3). The genome assembly of *B. philenor hirsuta* falls in this range with a genome size of 443 Mb. Interestingly, the size of this assembly is larger than those of other species of butterflies in the Tribe Troidine which range from 271 Mb to 336 Mb (see Table 3, He et al. 2022). These comparisons suggest that overall the genome sizes of butterflies in the subfamily Papilioninae are relatively conserved and are smaller than those of the subfamily Parnassiinae [the range of genome sizes of 5 species of butterflies in this subfamily is 449 Mb to 1115 Mb (Cong et al. 2015; He et al. 2022; Mackintosh et al. 2022)]. Furthermore, the current assembly suggests that a combined sequencing strategy of using PacBio HiFi and Omni-C Illumina reads drastically improves the assembly quality and completeness.

The current *B. philenor hirsuta* genome assembly is an excellent resource for evolutionary, ecological, and conservation biology studies. While this species has been studied from an ecological perspective, the lack of genomic resources has limited our understanding of several aspects of its biology and evolutionary history. *Battus philenor hirsuta* and its host plant *A. californica* are an iconic and California endemic species pair. This genome assembly combined with the forthcoming reference genome assembly of *A. californica* that is also being sequenced as part of the CCGP will help initiate foundational future studies focused on its ecology, biogeography, evolutionary history, behavior, and conservation. Finally, the current geographic range of this species overlaps with regions which are highly susceptible to wildfires and effects of climate change. This genome will facilitate future studies on the genomic basis of adaptation to a rapidly changing climate (Fiedler et al. 2022), predicting vulnerability of this species to extinction, and on plant–herbivore co-evolution.

Acknowledgments

We thank the staff at the University of California, Berkeley Botanical Garden who assisted in sample collection for this

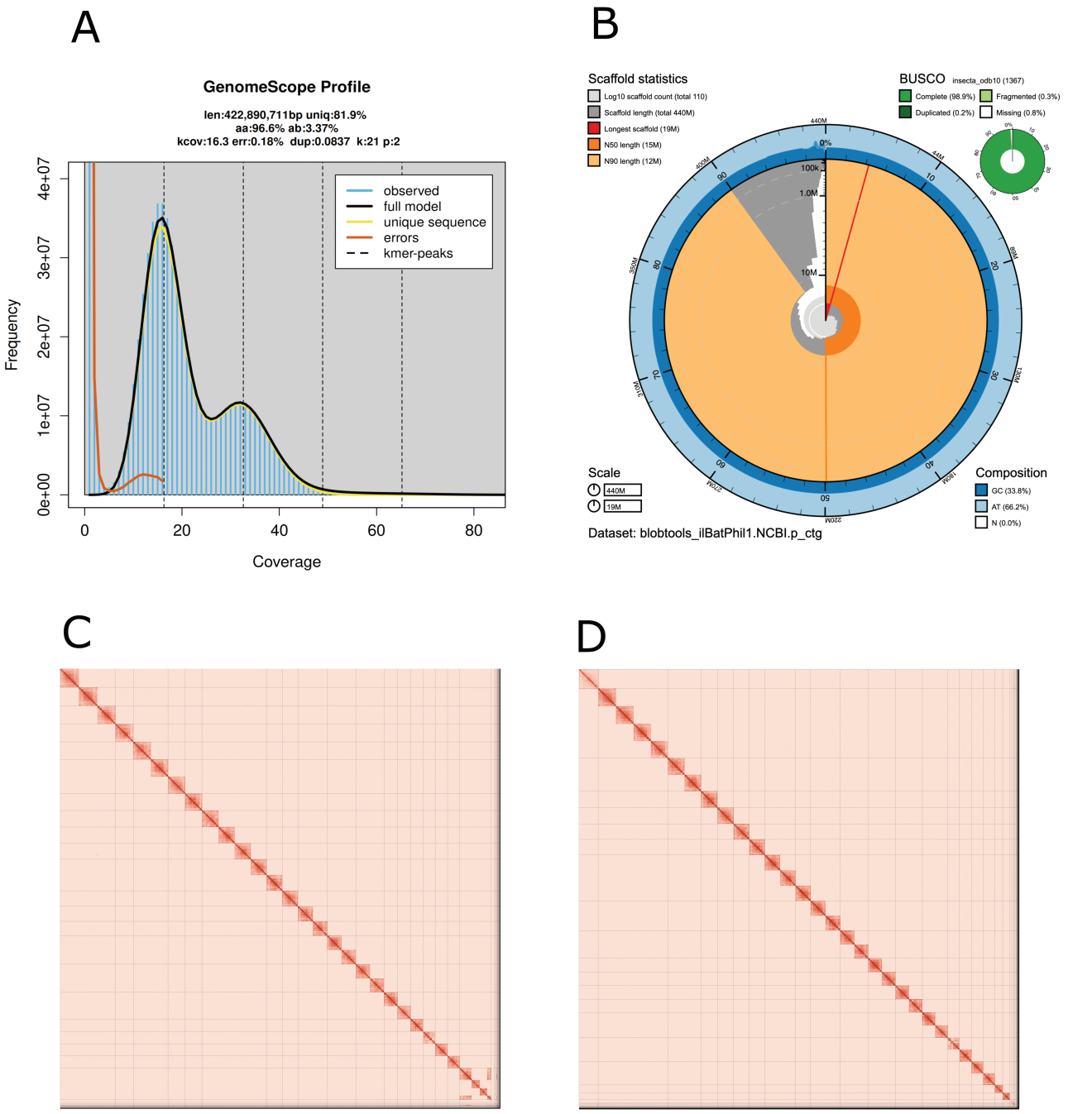

Fig. 2. Visual overview of genome assembly metrics. (A) K-mer spectra output generated from PacBio HiFi data without adapters using GenomeScope2.0. The bimodal pattern observed corresponds to a diploid genome and the k-mer profile matches that of high (>1%) heterozygosity. K-mers covered at lower coverage and high frequency correspond to differences between haplotypes, whereas the higher coverage and lower frequency k-mers correspond to the similarities between haplotypes. (B) BlobToolKit Snail plot showing a graphical representation of the quality metrics presented in Table 2 for the *B. philenor hirsuta* primary assembly (iBatPhil1.NCBI.p_ctg). The plot circle represents the full size of the assembly. From the inside-out, the central plot covers length-related metrics. The red line represents the size of the longest scaffold; all other scaffolds are arranged in size-order moving clockwise around the plot and drawn in gray starting from the outside of the central plot. Dark and light orange arcs show the scaffold N50 and scaffold N90 values. The central light gray spiral shows the cumulative scaffold count with a white line at each order of magnitude. White regions in this area reflect the proportion of Ns in the assembly; the dark versus light blue area around it shows mean, maximum, and minimum GC versus AT content at 0.1% intervals (Challis et al. 2020). Hi-C Contact maps for the primary (C) and alternate (D) genome assembly generated with PretextSnapshot. Hi-C contact maps translate proximity of genomic regions in 3D space to contiguous linear organization. Each cell in the contact map corresponds to sequencing data supporting the linkage (or join) between two of such regions.

project. PacBio Sequel II library prep and sequencing was carried out at the DNA Technologies and Expression Analysis Core at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. Deep sequencing of Omni-C libraries used the Novaseq S4 sequencing platforms at the Vincent J. Coates Genomics Sequencing Laboratory at UC

Table 2. Sequencing and assembly statistics, and accession numbers.

Bio projects and vouchers	**CCGP NCBI BioProject**		**PRJNA720569**			
	Genera NCBI BioProject		PRJNA766277			
	Species NCBI BioProject		PRJNA904298			
	NCBI BioSample		SAMN31838219, SAMN35734393			
	Specimen identification		BPH_F_UCBO, BPH_F_UCBG			
	NCBI Genome accessions		**Haplotype 1**		**Haplotype 2**	
	Assembly accession		JAQMYT000000000		JAQMYU000000000	
	Genome sequences		GCA_028537555.1		GCA_028537355.1	
Genome sequence	PacBio HiFi reads	Run	1 PACBIO_SMRT (Sequel II) run: 1.3M spots, 14.4G bases, 7.3Gb			
		Accession	SRR23572040			
	Omni-C Illumina reads	Run	2 ILLUMINA (Illumina NovaSeq 6000) runs: 94.1 M spots, 28.4 bases, 8.8 Gb			
		Accession	SRR23572038, SRR23572039			
Genome assembly quality metrics	Assembly identifier (Quality code*)		ilBatPhil1(7.7.P7.Q66.C97)			
	HiFi Read coverage§		33.93×			
			Haplotype 1		**Haplotype 2**	
	Number of contigs		119		86	
	Contig N50 (bp)		14,644,048		13,729,011	
	Contig NG50§		15,242,119		13,729,011	
	Longest Contigs		18,712,882		17,865,774	
	Number of scaffolds		109		76	
	Scaffold N50		15,242,119		15,128,050	
	Scaffold NG50§		15,494,611		15,128,050	
	Largest scaffold		19,297,608		17,889,339	
	Size of final assembly		443,215,375		422,808,241	
	Phased block NG50§		15,242,119		13,729,011	
	Gaps per Gbp (# Gaps)		23(10)		24(10)	
	Indel QV (Frame shift)		54.4781931		54.34909394	
	Base pair QV		66.81		66.84	
			Full assembly = 66.83			
	k-mer completeness		67.41		64.3	
			Full assembly = 99.85			
	BUSCO completeness (arthropoda_odb10) *n* = 1,013		**C**	**S**	**D**	**F**	**M**
		H1‡	98.90%	98.80%	0.10%	0.60%	0.50%
		H2‡	96.20%	96.20%	0.00%	0.90%	2.90%

*Assembly quality code *x.y.P.Q.C* derived notation, from (Rhie et al. 2021). *x* = log10[contig NG50]; *y* = log10[scaffold NG50]; *P* = log10 [phased block NG50]; *Q* = Phred base accuracy QV (quality value); C = % genome represented by the first "*n*" scaffolds, following a karyotype for of 2*n* = 60 estimated from other species in the family. Quality code for all the assembly denoted by primary assembly (ilBatPhil1.0.hap1).
§Read coverage and NGx statistics have been calculated based on the estimated genome size of 422 Mb.
‡(H1) Haplotype 1 and (H2) Haplotype 2 assembly values.

Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. We thank the staff at the UC Davis DNA Technologies and Expression Analysis Core and the UC Santa Cruz Paleogenomics Laboratory for their diligence and dedication to generating high quality sequence data. Partial support was provided by Illumina for Omni-C sequencing.

Funding

This work was supported by the California Conservation Genomics Project, with funding provided to the University of California by the State of California, State Budget Act of 2019 [UC Award ID RSI-19-690224]. N.K.W. was also supported

Table 3. A summary of currently available whole genome assemblies of butterflies in the family Papilionidae, subfamily Papilioninae.

Species	Genome size (Mb)	Contig N50 (Mb)	Scaffold N50 (Mb)	BUSCO %	Citation
Battus philenor hirsuta **(Tribe Troidini)**	**443**	**14.6**	**15.2**	**98.9**	**Current study**
Troides helena (Tribe Troidini)	336	NA	10.5	96.6	He et al. (2022)
Byasa hedistus (Tribe Troidini)	271	NA	9.2	98.1	He et al. (2022)
Lamproptera curius	550	NA	2.5	89.9	He et al. (2022)
Teinopalpus imperialis	535	NA	12.5	95.8	He et al. (2022)
Meandrusa payeni	406	NA	12.5	96.3	He et al. (2022)
Papilio demoleus	240	NA	9.1	97.5	He et al. (2022)
Papilio protenor	214	NA	5.4	97.4	He et al. (2022)
Iphiclides podalirius	430.6	5.2	15.5	96.5	Makintosh et al. (2022)
Papilio glaucus	376	NA	0.23	NA	Cong et al. (2015)

The three species from Tribe Troidini, including *Battus philenor hirsuta* are highlighted in bold.

by a grant from the National Institute of General Medical Sciences of the NIH (award no. R35GM119816).

Conflict of Interest

The authors declare no conflict of interests.

Data Availability

Data generated for this study are available under NCBI BioProject PRJNA904298. Raw sequencing data for sample BPH_F_UCBG (NCBI BioSample SAMN31838219) are deposited in the NCBI Short Read Archive (SRA) under SRR23572040 for PacBio HiFi sequencing data, and SRR23572038 and SRR23572039 for the Omni-C Illumina sequencing data. GenBank accessions for both primary and alternate assemblies are GCA_028537555.1 and GCA_028537355.1; and for genome sequences JAQMYT000000000 and JAQMYU000000000. Assembly scripts and other data for the analyses presented can be found at the following GitHub repository: www.github.com/ccgproject/ccgp_assembly.

References

Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*. 2020;36:311–316.

Allio R, Scornavacca C, Nabholz B, Clamens A-L, Sperling FA, Condamine FL. Whole genome shotgun phylogenomics resolves the pattern and timing of Swallowtail Butterfly evolution. *Syst Biol*. 2020;69:38–60.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit—interactive quality assessment of genome assemblies. *G3 Genes Genomes Genet*. 2020;10:1361–1374.

Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, Li H. Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol*. 2022;40:1332–1335.

Cong Q, Borek D, Otwinowski Z, Grishin NV. Tiger Swallowtail genome reveals mechanisms for speciation and caterpillar chemical defense. *Cell Rep*. 2015;10:910–919.

Fiedler PL, Erickson B, Esgro M, Gold M, Hull JM, Norris JM, Shapiro B, Westphal M, Toffelmier E, Shaffer EB. Seizing the moment: the opportunity and relevance of the California Conservation Genomics Project to State and Federal Conservation Policy. *J Hered*. 2022;113:589–596.

Fordyce JA. A model without a mimic: aristolochic acids from the California Pipevine Swallowtail, *Battus philenor hirsuta*, and its host plant, *Aristolochia californica*. *J Chem Ecol*. 2000;26:2567–2578.

Fordyce JA, Agrawal AA. The role of plant trichomes and caterpillar group size on growth and defence of the Pipevine Swallowtail *Battus Philenor*. *J Anim Ecol*. 2001;70:997–1005.

Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. Scaffolding of long read assemblies using long range contact information. *BMC Genom*. 2017;18:527.

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol*. 2019;15:e1007273.

Goloborodko A, Abdennur N, Venev S, Hbbrandao G. *Mirnylab/Pairtools: V0. 2.0*. 2018.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29:1072–1075.

He J-W, Zhang R, Yang J, Chang Z, Zhu L-X, Lu S-H, Xie F-A, Mao J-L, Dong Z-W, Liu G-C, et al. High-quality reference genomes of Swallowtail Butterflies provide insights into their coloration evolution. *Zool Res*. 2022;43:367–379.

Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A, Kumar N, et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol*. 2018;19:125.

Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, Cantin L, Jarvis ED. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience*. 2017;6:1–16.

Leimu R, Muola A, Laukkanen L, Kalske A, Prill N, Mutikainen P. Plant–herbivore coevolution in a changing world. *Entomol Exp Appl*. 2012;144:3–13.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, ArXiv [q-Bio.GN], arXiv:http://arxiv.org/abs/1303.3997, preprint: not peer reviewed.

Mackintosh A, Laetsch DR, Baril T, Ebdon S, Jay P, Vila R, Hayward A, Lohse K. The genome sequence of the Scarce Swallowtail, *Iphiclides podalirius*. *G3*. 2022;12:jkac193. doi: 10.1093/g3journal/jkac193.

Manni M, Berkeley MR, Seppey 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.

Northfield TD, Ives AR. Coevolution and the effects of climate change on interacting species. *PLoS Biol*. 2013;11:e1001685.

Pflug JM, Holmes VR, Burrus C, Spencer Johnston J, Maddison DR. Measuring genome sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles (Coleoptera). *G3*. 2020;10:3047–3060.

Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. High-resolution TADs

reveal DNA sequences underlying genome organization in flies. *Nat Commun*. 2018;9:189.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 2020;11:1432.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. 2021;592:737–746.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;21:245.

Shaffer HB, Toffelmier E, Corbett-Detig RB, Escalona M, Erickson B, Fiedler P, Gold M, Harrigan RJ, Hodges S, Luckau TK, et al. Landscape genomics to enable conservation actions: the California Conservation Genomics Project. *J Hered*. 2022;113:577–588.

Sim SB, Corpuz RL, Simmonds TJ, Geib SM. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genom*. 2022;23:157.

Toffelmier E, Joscha B, Bradley Shaffer H. The phylogeny of California, and how it informs setting multispecies conservation priorities. *J Hered*. 2022;113:597–603.