Title

Reference genome for the American rubyspot damselfly, Hetaerina americana

Journal

Journal of Heredity, 114(4)

ISSN

0022-1503

Authors

Grether, Gregory F
Beninde, Joscha
Beraut, Eric
et al.

Publication Date

2023-06-22

DOI

10.1093/jhered/esad031

Peer reviewed

Journal of Heredity, 2023, **114**, 385–394
https://doi.org/10.1093/jhered/esad031
Advance access publication 17 May 2023
Genome Resources

American Genetic Association

OXFORD

Genome Resources

Reference genome for the American rubyspot damselfly, *Hetaerina americana*

Gregory F. Grether[1,2], Joscha Beninde[2], Eric Beraut[5], Noravit Chumchim[4], Merly Escalona[3], Zachary G. MacDonald[2], Courtney Miller[1,2], Ruta Sahasrabudhe[4], Andrew M. Shedlock[6], Erin Toffelmier[1,2], H. Bradley Shaffer[1,2]

[1]Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, CA 90095-1606, United States,
[2]La Kretz Center for California Conservation Science, Institute of the Environment and Sustainability, University of California Los Angeles, Los Angeles, CA 90095-7239, United States,
[3]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, United States,
[4]DNA Technologies and Expression Analysis Core Laboratory, University of California Davis, Davis, CA 95616, United States,
[5]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, United States,
[6]Department of Biology, College of Charleston, Charleston, SC 29424, United States

Note: authors are listed in alphabetical order by last name, except for the first and last authors.
Address correspondence to Gregory F. Grether at the address above, or e-mail: ggrether@g.ucla.edu

Corresponding Editor: Arun Sethuraman

Abstract

Damselflies and dragonflies (Order: Odonata) play important roles in both aquatic and terrestrial food webs and can serve as sentinels of ecosystem health and predictors of population trends in other taxa. The habitat requirements and limited dispersal of lotic damselflies make them especially sensitive to habitat loss and fragmentation. As such, landscape genomic studies of these taxa can help focus conservation efforts on watersheds with high levels of genetic diversity, local adaptation, and even cryptic endemism. Here, as part of the California Conservation Genomics Project (CCGP), we report the first reference genome for the American rubyspot damselfly, *Hetaerina americana*, a species associated with springs, streams and rivers throughout California. Following the CCGP assembly pipeline, we produced two de novo genome assemblies. The primary assembly includes 1,630,044,487 base pairs, with a contig N50 of 5.4 Mb, a scaffold N50 of 86.2 Mb, and a BUSCO completeness score of 97.6%. This is the seventh Odonata genome to be made publicly available and the first for the subfamily Hetaerininae. This reference genome fills an important phylogenetic gap in our understanding of Odonata genome evolution, and provides a genomic resource for a host of interesting ecological, evolutionary, and conservation questions for which the rubyspot damselfly genus *Hetaerina* is an important model system.

Key words: Aquatic insect, California Conservation Genomics Project, Calopterygidae, Odonata, Zygoptera

Introduction

Insects are declining worldwide, both in diversity and aggregate biomass, with cascading effects on the ecosystems they support (Sánchez-Bayo and Wyckhuys 2019). Terrestrial insects, particularly pollinators, have received the most attention, but aquatic insects, defined as species that complete one or more life stages in freshwater, are also declining (Sánchez-Bayo and Wyckhuys 2019). Most aquatic insects that are known to be threatened with extinction are in the order Odonata (IUCN 2022). Odonata is an ancient but relatively small clade of 6,377 recognized species (Paulson et al. 2022), divided roughly equally between two major subclades: damselflies (suborder Zygoptera) and dragonflies (Anisoptera) (Bybee et al. 2021). Of the species evaluated by the IUCN with sufficient data to make a determination, 448 of 2,200 damselflies (20.4%) and 228 of 2,086 dragonflies (10.9%) are listed as Vulnerable, Endangered, or Critically Endangered (Supplementary Table S1). Thus, damselflies are nearly twice as likely to be threatened with extinction as dragonflies (IUCN 2022), perhaps because of their weaker dispersal capabilities (Corbet 1999). Habitat specialists, including lotic species that require running water to complete their life cycle, are especially sensitive to habitat loss and fragmentation (Rouquette and Thompson 2007; Ball-Damerow et al. 2014; Sánchez-Bayo and Wyckhuys 2019; Rocha-Ortega et al. 2020).

The habitat specificity of lotic damselflies makes them important indicators of riverine ecosystem health and predictors of population trends in taxa with similar habitat requirements and dispersal characteristics (Ball-Damerow et al. 2014; Kutcher and Bried 2014; Córdoba-Aguilar and Rocha-Ortega 2019; Rocha-Ortega et al. 2019). Local declines in damselfly diversity and abundance have been linked to pollution,

Received October 5, 2022; Accepted May 7, 2023

reductions in vegetation cover, and other direct and indirect effects of human activities, including climate change (Kutcher and Bried 2014; Córdoba-Aguilar and Rocha-Ortega 2019; Rocha-Ortega et al. 2019). Genomic studies of lotic damselflies should highlight the conservation value of particular watersheds and reveal hotspots of genetic diversity, local adaptation, and endemism.

The American rubyspot damselfly (*Hetaerina americana;* family Calopterygidae) occurs throughout the continental United States and as far south as Nicaragua (Vega-Sánchez et al. 2020). Across this extensive geographic range, the coloration of both sexes varies considerably. Several species names were once used for regional variants (e.g. *Hetaerina californica, Hetaerina texana*), but were later synonymized under *H. americana* based on male clasper morphology (Calvert 1908; Garrison 1990). However, recent population genetic analyses have revealed deep splits and high levels of genetic differentiation between populations in different watersheds (Drury et al. 2019), and even between some sympatric populations in Mexico (Vega-Sánchez et al. 2019). Currently, *H. americana* is viewed as a complex of cryptic species at various stages of the speciation process (Vega-Sánchez et al. 2019). The most genetically distinct lineage, recently named Hetaerina *calverti* (Vega-Sánchez et al. 2020), ranges from Guatemala to Texas, and is estimated to have separated from other lineages ~ 4 Mya (Standring et al. 2022).

Rubyspot damselflies have a two-phase life cycle, with aquatic larvae, and terrestrial winged adults (Corbet 1999). Ovipositing (egg-laying) females insert eggs into submerged vegetation in flowing water. The larvae are sit-and-wait predators of smaller invertebrates and the adults feed on smaller flying insects by sallying out from perches (Corbet 1999). In temperate regions, the larvae enter a state of dormancy (diapause) during the winter months, and can take as long as two years to emerge as adults (Corbet 1999). Newly emerged adults (tenerals) are highly vulnerable to predators and adverse weather, but adults that survive the teneral period often live several weeks. Larvae cannot survive outside of water, and adults desiccate rapidly if deprived of access to a water source; adults are capable of flying greater distances but seldom disperse farther than 100 m from where they emerged, and generally stay close to water (Grether 1996). Population persistence appears to require flowing water year-round (G.F. Grether, pers. obs.).

In California, *H. americana* reproduces in springs, streams, and rivers, from sea level to over 2000 m, all across the state and in nearly every ecoregion (Fig. 1). The species has impressively wide tolerances for water chemistry, temperature, and habitat alteration. If a site has running water year-round and submerged vegetation for oviposition, it is probably suitable habitat for *H. americana*, even if the surrounding land has been fully converted for human use. However, numerous populations have gone extinct because of habitat loss related to urbanization, industrialization, dams, water-diversion projects, and prolonged drought (G.F. Grether, pers. obs.). In a re-survey of 45 sites across California and Nevada, *H. americana* was one of 25 Odonata species to show a net decrease in occurrence between 1914 to 15 and 2011 to 12 (14 species increased and 30 species showed no net change; Ball-Damerow et al. 2014).

The geographically and ecologically broad range of *H. americana*, combined with its high habitat specificity and limited dispersal ability, make this species an important candidate for conservation genomics studies. This is particularly true in California, where lotic habitats are often highly fragmented. In this paper, we report the first genome assembly for *H. americana*, produced as part of the California Conservation Genomics Project (CCGP; Shaffer et al. 2022).

Methods

Biological materials

The reference genome is based on two adult male *H. americana* (males are the heterogametic sex) collected in August 2020 from lower Cache Creek in Yolo County (38.68810° N, 121.86685° W; elevation 31 m). Damselflies were captured with an aerial net, transported alive to the UC Davis Genome Center, and flash-frozen in liquid nitrogen.

High molecular weight genomic DNA isolation

High molecular weight (HMW) genomic DNA (gDNA) was extracted from 40 mg of flash-frozen thoracic tissue (Sample ID1083.01) using the Nanobind Tissue Big DNA kit (Pacific BioSciences [PacBio], Menlo Park, CA) following manufacturer's instructions. Purity of gDNA was accessed using a NanoDrop ND-1000 spectrophotometer, which returned a 260/280 ratio of 1.9 and a 260/230 of 2.3. DNA yield was quantified using Qubit 2.0 Fluorometer (Thermo Fisher Scientific, MA). Integrity of the HMW gDNA was verified on a Femto pulse system (Agilent Technologies, Santa Clara, CA) where 80% of DNA was found in fragments above 120 kb.

HiFi library preparation and sequencing

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. The sheared gDNA was concentrated using 1.8X 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 adapter v3 at 20 °C for 60 min. The SMRTbell library was purified and concentrated with 0.45X AMpure PB beads for size selection with 40% diluted AMPure PB beads (PacBio, Cat. #100-265-900) to remove short SMRTbell templates less than 3 kb in length. The 12 to 20 kb average HiFi SMRTbell library was sequenced at UC Davis DNA Technologies Core (Davis, CA) using three 8M SMRT cells, Sequel II sequencing chemistry 2.0, and 30-h movies each on a PacBio Sequel II sequencer.

Omni-C library preparation and sequencing

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 (Sample ID: 1083.08) 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

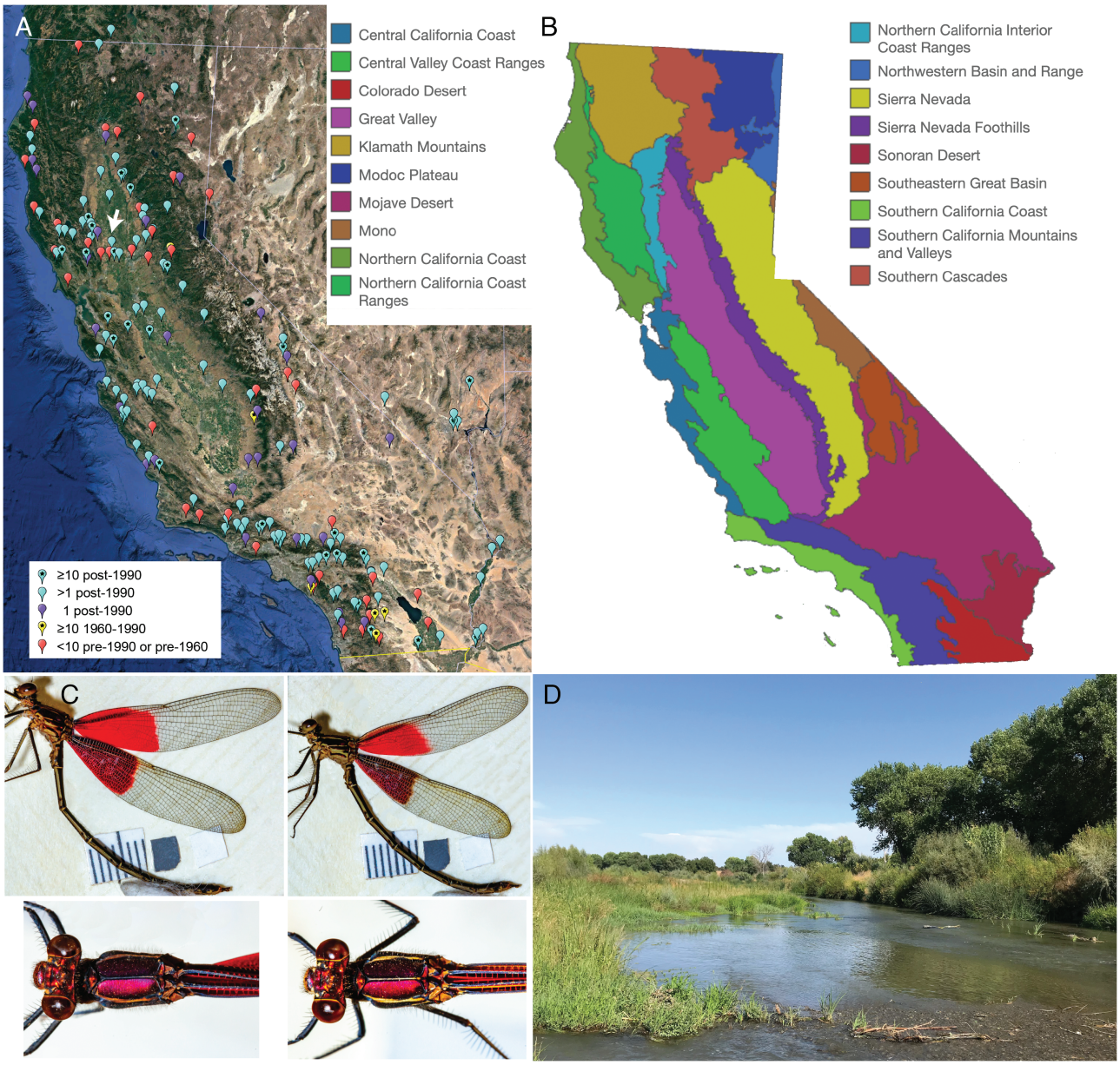

Fig. 1. The American rubyspot damselfly (*Hetaerina americana*) in California. (A) Species occurrence records from public databases, filtered to 0.1 decimal degrees, and color coded based on number and recency. The genome assembly is based on specimens collected at the site marked with a white arrow (see panel C). (B) Inset map of United States Department of Agriculture (USDA) ecoregion sections in California. (C) Photos of the specimens used for the reference genome (the scale bar is graduated in millimeters). Both specimens are males, the heterogametic sex. Based on coloration, the specimen on the left was fully mature (≥ 14 days post-emergence) and the specimen on the right was within 3–6 days of full maturity (Grether 1996). (D) Habitat at the site on lower Cache Creek in the Sacramento River drainage where specimens for this genome assembly were collected in August 2020. The species occurrence map was created in Google Earth Pro using occurrence data from GBIF (https://www.gbif.org/), Odonata Central (https://www.odonatacentral.org/) and iNaturalist (https://www.inaturalist.org/), and the USDA ecoregion map was downloaded from Conservation Biology Institute's Data Basin website (https://databasin.org/datasets/), as permitted for non-commercial use. Photos by G.F. Grether.

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. A NGS library was generated using a 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 platform (Illumina, San Diego, CA) to generate approximately 100 million 2 × 150 bp read pairs per GB of genome size.

Nuclear genome assembly

The *H. americana* genome was assembled following the CCGP assembly pipeline Version 4.0, as outlined in Table 1. As with other CCGP assemblies, our goal was to produce a high quality and highly contiguous assembly using PacBio HiFi reads and Omni-C data while minimizing manual curation.

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
Remove low-coverage, duplicated contigs	purge_dups	1.2.6
Alignment of long reads	minimap2 (-ax map-pb)	2.24-r1122
Scaffolding		
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
Organelle assembly		
Mitogenome assembly	MitoHiFi (-r, -p 50, -o 1)	2 Commit c06ed3e
Genome quality assessment		
Basic assembly metrics	QUAST (--est-ref-size)	5.0.2
Assembly completeness	BUSCO (-m geno, -l arthropoda)	5.0.0
	Merqury	2020-01-29
Genome mappability assessment	genmap index; genmap map (-K 150 -E 0)	1.3.0
Read mapping assessment	qualimap bamqc	2.2.1
Contamination screening		
Local alignment tool	BLAST+	2.1
General contamination screening	BlobToolKit	2.3.3

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

Remnant adapter sequences were removed from the PacBio HiFi dataset using HiFiAdapterFilt (Sim et al. 2022) to obtain the initial dual or partially phased diploid assembly (http://lh3.github.io/2021/10/10/introducing-dual-assembly) using HiFiasm (Cheng et al. 2021). We tagged output haplotype 1 as the primary assembly and output haplotype 2 as the alternate assembly. We scaffolded both assemblies using the Omni-C data with SALSA (Ghurye et al. 2017, 2019). Next, we identified sequences corresponding to haplotypic duplications, contig overlaps, and repeats on the primary assembly with purge_dups (Guan et al. 2020) and transferred them to the alternate assembly.

We generated Omni-C contact maps for both assemblies by aligning the Omni-C data against the corresponding assembly 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, and checked the contact maps for major misassemblies. If, in the proximity of a join that was made by the scaffolder, we identified a strong signal off-diagonal and a lack of signal in the consecutive genome region, we marked the join. All such marked joins were dissolved by cutting the scaffolds at the coordinates of these joins. Using the PacBio HiFi reads and

YAGCloser (https://github.com/merlyescalona/yagcloser), we closed some of the remaining gaps generated during scaffolding. We then checked for contamination using the BlobToolKit Framework (Challis et al. 2020). Finally, we trimmed remnants of sequence adaptors and mitochondrial contamination identified during the contamination screening performed by NCBI.

Mitochondrial genome assembly

We assembled the mitochondrial genome of *H. americana* from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (https://github.com/marcelauliano/MitoHiFi); (Allio et al. 2020). The mitochondrial sequence of another species in the family Calopterygidae, *Mnais tenuis* (NCBI:NC_057643.1), was used as the starting sequence. After completion of the nuclear genome, we searched for matches of the resulting mitochondrial assembly sequence in the nuclear genome assembly using BLAST + (Camacho et al. 2009) and filtered out contigs and scaffolds from the nuclear genome with a percentage of sequence identity > 99% and size smaller than the mitochondrial assembly sequence.

Genome size estimation and quality assessment

We generated k-mer counts from the PacBio HiFi reads using meryl (https://github.com/marbl/meryl). The k-mer database was 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 completeness, we used BUSCO (Manni et al. 2021) with the Arthropoda ortholog database (arthropoda_odb10) containing 1,013 genes. Assessment of base level accuracy (QV) and k-mer completeness was performed using the previously generated meryl database and merqury (Rhie et al. 2020). We further estimated genome assembly accuracy via BUSCO gene set frameshift analysis using the pipeline described by Korlach et al. (2017).

Measurements of the size of the phased blocks was based on the size of the contigs generated by HiFiasm in Hi-C mode. We followed 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 known karyotype $2n = 26$ for *H. americana* (Cruden 1968; Ardila-Garcia and Gregory 2009). Quality metrics for the notation were calculated on the primary assembly.

Mappability assessment

We estimated mappability across both the primary and alternate assemblies using GenMap (Pockrandt et al. 2020). First, we generated mappability scores with 150-mers, allowing no mismatches, and obtained the number of sites with scores < 0.5, indicating hard-to-map regions. We then aligned HiFi reads to both assemblies using minimap2 (Li 2018) and sorted alignment files were generated with samtools (Li et al. 2009). Omni-C reads were similarly aligned, as previously described for the generation of the contact maps. We then used final alignments to estimate mapping statistics using Qualimap (García-Alcalde et al. 2012; Okonechnikov et al. 2016) and observed error rates (total, mismatch, insertions, and deletions) and genome coverage.

Results

The PacBio HiFi and Omni-C sequencing libraries generated 2.6 million reads and 70.9 million read pairs, respectively. The former yielded a mean coverage of 24.26X (N50 read length 14,689 bp; minimum read length 71 bp; mean read length 13,476 bp; maximum read length 52,809 bp) based on the Genomescope 2.0 genome size estimate of 1.4 Gb. Based on PacBio HiFi reads, we estimated a 0.161% sequencing error rate and 0.458% nucleotide heterozygosity rate. The k-mer spectrum based on PacBio HiFi reads showed a distribution with two peaks at ~11- and 23-fold coverage, where peaks correspond to homozygous and heterozygous states of a diploid species (Fig. 2A). The distribution presented in this k-mer spectrum supports that of a low heterozygosity profile.

The final assembly (ioHetAmer1) consists of two pseudo haplotypes, primary, and alternate, with both genome sizes close to the estimated value from Genomescope2.0 (1.4 Gb, Fig. 2A). The primary assembly consists of 1,583 scaffolds spanning 1.6 Gb with contig N50 of 4.9 Mb, scaffold N50 of 86.1 Mb, longest contig of 26.6 Mb, and longest scaffold of 123.2 Mb. The alternate assembly consists of 709 scaffolds, spanning 1.38 Gb with a contig N50 of 5.8 Mb, scaffold N50 of 73.4 Mb, longest contig 26.8 Mb, and longest scaffold of 118.5 Mb. Assembly statistics are reported in tabular form in Table 2, and graphical representation for the primary assembly in Fig. 2B (see Supplementary Fig. 1 for the alternate assembly).

We identified a total of four misassemblies, one on the primary assembly and three on the alternate, and broke the corresponding joins made by SALSA2 on both assemblies. We were able to close a total of 24 gaps, 19 on the primary and five on the alternate assembly. We further filtered out 73 contigs corresponding to Proteobacteria contaminants (23 contigs from the primary assembly and 51 from the alternate) and a single contig from the alternate assembly corresponding to Porifera. Finally, we filtered out 42 contigs corresponding to mitochondrial contamination (7 from the primary and 35 from the alternate assembly). No further contigs were removed.

The primary assembly has a BUSCO completeness score of 97.6% using the Arthropoda gene set, a per base quality (QV) of 59.4, k-mer completeness of 95.5, and frameshift indel QV of 51.22. The alternate assembly has a BUSCO completeness score of 92.7%, per base quality (QV) of 59.8, k-mer completeness of 88.67, and frameshift indel QV of 51.01. The Omni-C contact maps show that both assemblies are highly contiguous (Fig. 2C and D). We deposited scaffolds corresponding to both primary and alternate haplotype (see Table 2 and Data availability for details).

We assembled a mitochondrial genome with MitoHiFi, with a final mitochondrial genome size of 16,511 bp. The base composition of the final assembly version is A = 30.57%, C = 13.22%, G = 16.88%, T = 39.33%, and consists of 22 unique transfer RNAs and 13 protein coding genes.

Using a mappability threshold of 0.4, 4.07% of the primary assembly and 4.77% of the alternate assembly were identified as hard-to-map or very repetitive (see Supplementary Table S2 for percentages based on other mappability thresholds).

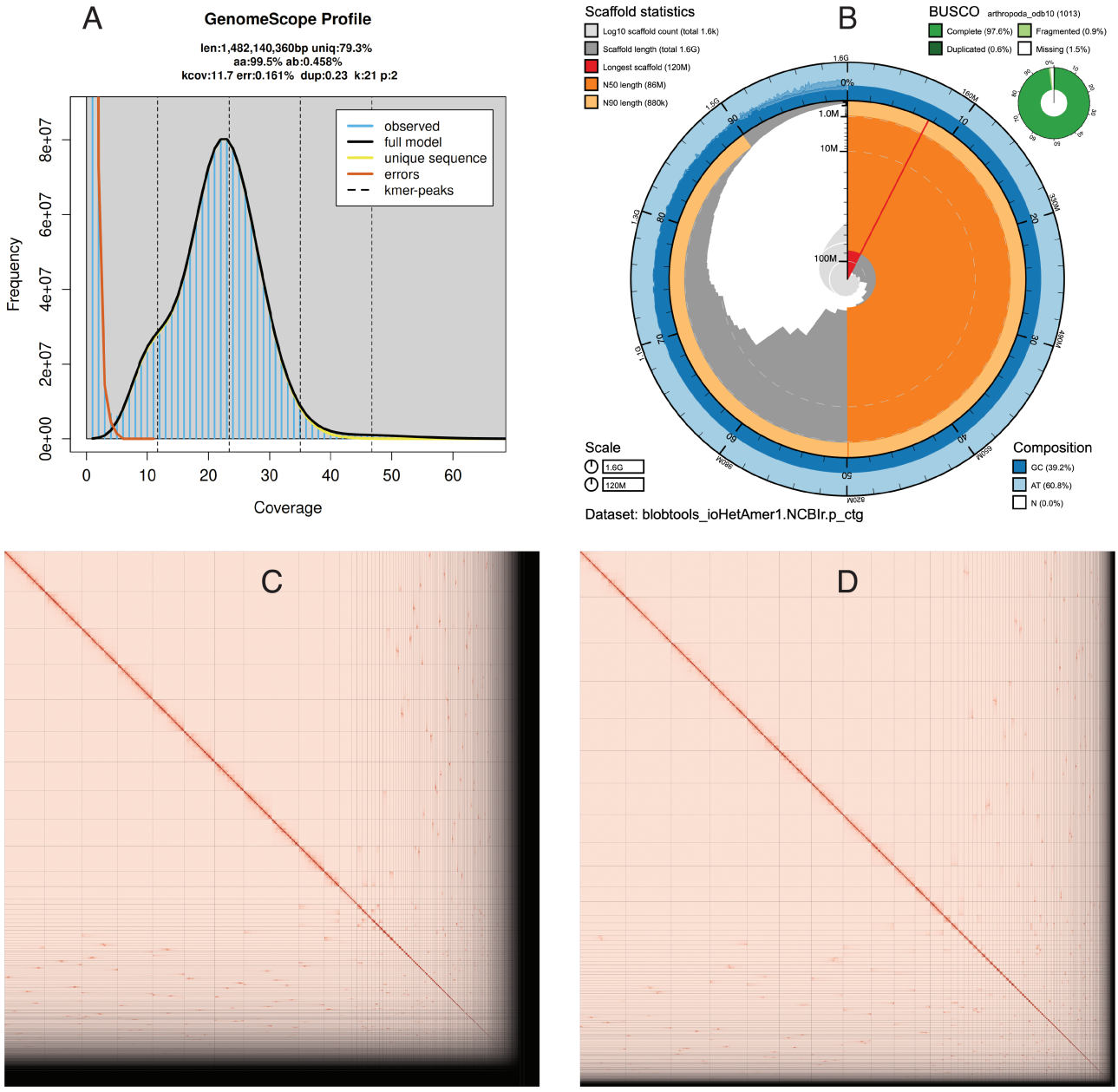

Fig. 2. Overview of genome assembly metrics for *Hetaerina americana*. (A) K-mer spectrum output from PacBio HiFi data without adapters, generated using GenomeScope2.0. K-mers at lower coverage and lower frequency correspond to differences between haplotypes, whereas the higher coverage and higher frequency k-mers correspond to the similarities between haplotypes. (B) Snail plot showing a graphical representation of the quality metrics in Table 2 for the primary assembly. The circle represents the full size of the assembly. From the inside-out, the central plot displays length-related metrics. The red line represents the longest scaffold; other scaffolds are ordered by size moving clockwise around the plot and drawn in gray starting from the outside of the central plot. Dark and light orange arcs mark the scaffold N50 and 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). (C, D) 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 3-D 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. Scaffolds are separated by black lines and higher density corresponds to higher levels of fragmentation.

Read mapping statistics for both HiFi and Omni-C reads (Supplementary Tables S3 and S4, respectively) show that both assemblies are of high quality and accuracy.

Discussion

Currently, nine species of Odonata have publicly available reference genomes, including three damselflies (Zygoptera) and six dragonflies (Anisoptera) (Table 3). Compared to *Calopteryx splendens*, its closest relative with a reference genome [divergence time ~54 Mya; (Standring et al. 2022)], the reference genome for *H. americana* is comparable in size (1.63 Gb), more complete, based on the percentage of unfragmented, highly conserved Arthropoda genes in the assembly (97.6% versus 55%; Table 3), and has much longer scaffolds (scaffold N50 of 86.2 versus 0.42 Mb). Other

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

Bio projects and vouchers	**CCGP NCBI BioProject**		**PRJNA720569**				
	Genera NCBI BioProject		PRJNA765842				
	Species NCBI BioProject		PRJNA777181				
	NCBI BioSample		SAMN24913893, SAMN24913894				
	Specimen identification		1083.01, 1083.08				
	NCBI Genome accessions		**Primary**			**Alternate**	
	Assembly accession		JAKTNV000000000			JAKTNW000000000	
	Genome sequences		GCA_022747635.1			GCA_022747625.1	
Genome sequence	PacBio HiFi reads	Run	1 PACBIO_SMRT (Sequel II) run: 2.7M spots, 36G bases, 23.2Gb				
		Accession	SRX14688495				
	Omni-C Illumina reads	Run	2 ILLUMINA (Illumina NovaSeq 6000) runs: 70.3M spots, 21.4G bases, 7.2Gb				
		Accession	SRX14688496, SRX14688497				
Genome assembly quality metrics	Assembly identifier (Quality code*)					ioHetAmer1 (6.7.P6.Q59.C65)	
	HiFi Read coverage§					24.26	
			Primary			**Alternate**	
	Number of contigs		1,883			709	
	Contig N50 (bp)		4,979,805			5,888,495	
	Contig NG50§		5,407,786			5,406,725	
	Longest Contigs		26,688,041			26,816,108	
	Number of scaffolds		1,583			434	
	Scaffold N50		86,194,728			73,406,209	
	Scaffold NG50§		86,194,728			73,406,209	
	Largest scaffold		123,201,532			118,594,994	
	Size of final assembly (bp)		1,630,044,487			1,389,467,078	
	Phased blocks NG50		5,019,017			5,750,389	
	Gaps per Gbp (# Gaps)		184 (300)			194 (275)	
	Indel QV (Frame shift)		51.22019172			51.01334573	
	Base pair QV		59.4239			59.8344	
			Full assembly = 59.6079				
	k-mer completeness		95.5442			88.6793	
			Full assembly = 99.6573				
	BUSCO completeness (arthropoda) n = 1,013		**C**	**S**	**D**	**F**	**M**
		P‡	97.60%	97.00%	0.60%	0.90%	1.50%
		A‡	92.70%	92.60%	0.10%	1.10%	6.20%
	Organelles		1 Partial mitochondrial sequence JAKTNV010001583.1				

BUSCO scores: Complete (C); complete and Single (S); complete and Duplicated (D); Fragmented (F) and Missing (M). *n*, Number of BUSCO genes in the set/database. Bp: base pairs.
*Assembly 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 known karyotype 2*n* = 26 for *H. americana* (Cruden 1968; Ardila-Garcia and Gregory 2009).
§Read coverage and NGx statistics have been calculated based on a genome size of 1.7 Gb.
‡Primary (P) and Alternate (A) assembly values.

published Odonata reference genomes range in size from 0.66 to 1.87 Gb, with scaffold N50 values ranging from 0.4 to 206.6 Mb. Our estimate of the genome size of *H. americana* is about 50% larger than the previous estimate of ~1.1 GB, which was based on Feulgen image analysis densitometry of spermatozoa compared to that of *Drosophila melanogaster* (Ardila-Garcia and Gregory 2009).

The *H. americana* reference genome assembly will enable a wide range of investigations addressing genomic variation, evolutionary history, and biogeography of the rubyspot damselfly subfamily (Hetaerininae), and will help resolve some outstanding taxonomic issues (Standring et al. 2022). In the context of the CCGP (Shaffer et al. 2022), the *H. americana* reference genome fills an important phylogenetic gap in our understanding of California comparative genomics (Toffelmier et al. 2022), and the single nucleotide polymorphism (SNP) data from samples collected throughout the state will be used to develop a better understanding of genetic

Table 3. Comparison of Odonata genomes available on NCBI in April 2023.

Species	Size (Gb)	No. of scaffolds	Scaffold N50	Scaffold L50	No. of contigs	Contig N50	Contig L50	GC%	BUSCO (arthropoda) (%)					
									S	D	F	C	M	Ref.
Calopteryx splendens	1.6	8,896	0.42	1,013	645,677	0.0045	83,269	39	54	1.7	32	55	13	1
Hetaerina americana	1.6	1,582	86.2	8	1,882	5.0	84	39	97	0.6	0.9	98	1.5	2
Ischnura elegans	1.7	110	123.6	7	359	13.1	36	38	98	1.1	–	99	0.8	3
Ladona fulva	1.2	9,411	1.2	297	46,619	0.0673	4,253	36	80	2	13	82	5.3	4
Pantala flavacens	0.7	43	56	6	100	16.2	13	34	94	3.3	–	97	3.1	4
Platycnemis pennipes	1.8	87	144.8	6	252	18.2	32	39	–	–	–	–	–	–
Rhinocypha anisoptera	1.9	754,445	0.4	12	770,162	0.0249	5,505	40	–	–	–	–	–	–
Sympetrum striolatum	1.3	550	103.2	6	900	6.1	49	35	–	–	–	–	–	–
Tanypteryx hageni	1.7	1,032	206.6	4	1,919	4.3	109	38	96	0.6	1.0	97	2.2	5

N50 in Mb. BUSCO scores: Single (S), Duplicated (D), Fragmented (F), Complete (C = S + D), Missing (M). BUSCO references: 1, (Ioannidis et al. 2017); 2, this paper; 3, NCBI; 4, (Liu et al. 2022); 5, (Tolman et al. 2023).

connectivity within and among California's watersheds, identify specific barriers to gene flow, and map broad patterns of genomic diversity. These damselflies have great potential to serve as sentinels for the quality and connectivity of their lotic habitats, given that they require running water to complete their life cycle and are capable of only weak overland dispersal (Ball-Damerow et al. 2014; Kutcher and Bried 2014; Córdoba-Aguilar and Rocha-Ortega 2019; Rocha-Ortega et al. 2019). *Hetaerina americana* occurs in perennial streams and rivers throughout the state, including vulnerable aquatic ecosystems such as desert springs, and as such this genome and the subsequent population genomic analyses currently underway will be of considerable value to ongoing freshwater protection efforts in the state (Fiedler et al. 2022). Isolation of some of these springs may be sufficient to have facilitated the evolution of cryptic endemic species, which may only be identified through analyses of genomic data, and integrative species delimitation (Dayrat 2005; Campbell et al. 2022). Thus, what we learn from the genomics of this widespread species could be of substantial value for identifying and conserving threatened species and subspecies, as well as the unique ecosystems they inhabit.

Supplementary Material

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

Acknowledgments

PacBio Sequel II library prep and sequencing was carried out at the DNA Technologies and Expression Analysis Cores 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 Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. We thank the staff at the UC Davis DNA Technologies and Expression Analysis Cores and the UC Santa Cruz Paleogenomics Laboratory for their diligence and dedication to generating high quality sequence data.

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]. G.F.G. was partially supported by National Science Foundation grant DEB-2040883.

Data Availability

Data generated for this study are available under NCBI BioProject PRJNA777181. Raw sequencing data for samples 1083.01 and 1083.08 (NCBI BioSample SAMN24913893, SAMN24913894) are deposited in the NCBI Short Read Archive (SRA) under SRX14688495 for PacBio HiFi sequencing data, and SRX14688496 and SRX14688497 for the Omni-C Illumina sequencing data. GenBank accessions for both primary and alternate assemblies are GCA_022747635.1 and GCA_022747625.1; and for genome sequences JAKTNV000000000 and JAKTNW000000000. The GenBank organelle genome assembly for the mitochondrial genome is JAKTNV010001583.1. 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(1):311–316.

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol Ecol Resour*. 2020;20(4):892–905.

Ardila-Garcia AM, Gregory TR. An exploration of genome size diversity in dragonflies and damselflies (Insecta: Odonata). *J Zool*. 2009;278(3):163–173.

Ball-Damerow JE, M'Gonigle LK, Resh VH. Changes in occurrence, richness, and biological traits of dragonflies and damselflies (Odonata) in California and Nevada over the past century. *Biodivers Conserv*. 2014;23:2107–2126.

Bybee SM, Kalkman VJ, Erickson RJ, Frandsen PB, Breinholt JW, Suvorov A, Dijkstra KDB, Cordero-Rivera A, Skevington JH, Abbott JC, et al. Phylogeny and classification of Odonata using targeted genomics. *Mol Phylogenet Evol*. 2021;160:107115.

Calvert PP. Odonata. In: Eaton AE, Calvert PP, editors. *Biologia Centrali Americana: Insecta Neuroptera*. London: RH Porter and Dulau Co.; 1908. p. 17–342.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinf*. 2009;10:421.

Campbell EO, MacDonald ZG, Gage EV, Gage RV, Sperling FAH. Genomics and ecological modelling clarify species integrity in a confusing group of butterflies. *Mol Ecol*. 2022;31(8):2400–2417.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit—interactive quality assessment of genome assemblies. *G3 Genes Genomes Genet*. 2020;10(4):1361–1374.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with Hifiasm. *Nat Methods*. 2021;18:170–175.

Corbet, P. S. 1999. *Dragonflies: behavior and ecology of Odonata*. Ithaca, N.Y: Cornell University Press.

Córdoba-Aguilar A, Rocha-Ortega M. Damselfly (Odonata: Calopterygidae) population decline in an urbanizing watershed. *J Insect Sci*. 2019;19(3):1–6.

Cruden RW. Chromosome numbers of some North American dragonflies (Odonata). *Can J Genet Cytol*. 1968;10:200–214.

Dayrat B. Towards integrative taxonomy. *Biol J Linn Soc*. 2005;85(3):407–415.

Drury JP, Anderson CN, Castillo MBC, Fisher J, McEachin S, Grether GF. A general explanation for the persistence of reproductive interference. *Am Nat*. 2019;194(2):268–275.

Fiedler PL, Erickson B, Esgro M, Gold M, Hull JM, Norris J, Shapiro B, Westphal M, Toffelmier E, Shaffer HB. Seizing the moment: the opportunity and relevance of the California Conservation Genomics Project to state and federal conservation policy. *J Hered*. 2022;113(6):589–596.

García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*. 2012;28(20):2678–2679.

Garrison RW. A synopsis of the genus *Hetaerina* with descriptions of four new species (Odonata: Calopterygidae). *Trans Am Entomol Soc*. 1990;116(1):175–259.

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;8(8):e1007273.

Goloborodko A, Abdennur N, S. Venev, hbbrandao, and gfudenberg. 2018. mirnylab/pairtools: v0.2.0. (v0.2.0). Zenodo. https://doi.org/10.5281/zenodo.1490831.

Grether GF. Sexual selection and survival selection on wing coloration and body size in the rubyspot damselfly *Hetaerina americana*. *Evolution*. 1996;50(5):1939–1948.

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 2020;36(9):2896–2898.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072–1075.

Ioannidis P, Simao FA, Waterhouse RM, Manni M, Seppey M, Robertson HM, Misof B, Niehuis O, Zdobnov EM. Genomic features of the damselfly *Calopteryx splendens* representing a sister clade to most insect orders. *Genome Biol Evol*. 2017;9(2):415–430.

IUCN. 2022. The IUCN Red List of Threatened Species. https://www.iucnredlist.org. [accessed 2022 Sep 14].

Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir N, Kumar A, 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(10):gix085.

Kutcher TE, Bried JT. Adult Odonata conservatism as an indicator of freshwater wetland condition. *Ecol Indic*. 2014;38:31–39.

Li, H. 2013. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. arXiv [q-bio.GN], preprint: not peer reviewed.

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094–3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079.

Liu H, Jiang F, Wang S, Wang H, Wang A, Zhao H, Xu D, Yang B, Fan W. Chromosome-level genome of the globe skimmer dragonfly (*Pantala flavescens*). *GigaScience*. 2022;11:1–8.

Manni M, Berkeley MR, Seppey M, Zdobnov EM. BUSCO: assessing genomic data quality and beyond. *Curr Protoc*. 2021;1:1–41.

Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*. 2016;32(2):292–294.

Paulson D, Schorr M, Deliry C. *World Odonata List*; 2022. https://www.pugetsound.edu/slater-museum-natural-history-0/biodiversity-resources/insects/dragonflies/world-odonata-list

Pockrandt C, Alzamel M, Iliopoulos CS, Reinert K. GenMap: ultra-fast computation of genome mappability. *Bioinformatics*. 2020;36(12):3687–3692.

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

Rocha-Ortega M, Rodríguez P, Bried J, Abbott J, Córdoba-Aguilar A. Why do bugs perish? Range size and local vulnerability traits as surrogates of Odonata extinction risk. *Proc Biol Sci*. 2020;287(1924):20192645.

Rocha-Ortega M, Rodríguez P, Córdoba-Aguilar A. Can dragonfly and damselfly communities be used as bioindicators of land use intensification? *Ecol Indic*. 2019;107:105553.

Rouquette JR, Thompson DJ. Patterns of movement and dispersal in an endangered damselfly and the consequences for its management. *J Appl Ecol*. 2007;44:692–701.

Sánchez-Bayo F, Wyckhuys KAG. Worldwide decline of the entomofauna: a review of its drivers. *Biol Conserv*. 2019;232:8–27.

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:(6):1–12.

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(1):157.

Standring S, Sánchez-Herrera M, Guillermo-Ferreira R, Ware JL, Vega-Sánchez YM, Clement R, Drury JP, Grether GF, González-Rodríguez A, Mendoza-Cuenca L, et al. Evolution and biogeographic history of rubyspot damselflies (Hetaerininae: Calopterygidae: Odonata). *Diversity*. 2022;14(9):757.

Toffelmier E, Beninde J, Shaffer HB. The phylogeny of California, and how it informs setting multi-species conservation priorities. *J Hered*. 2022:113(6):esac045.

Tolman ER, Beatty CD, Bush J, Kohli M, Moreno CM, Ware JL, Weber KS, Khan R, Maheshwari C, Weisz D, et al. A chromosome-length assembly of the Black Petaltail (*Tanypteryx hageni*) Dragonfly. *Genome Biol Evol*. 2023;15(3):1–8.

Vega-Sánchez YM, Mendoza-Cuenca LF, Gonzálezrodríguez A. *Hetaerina calverti* (Odonata: Zygoptera: Calopterygidae) sp. nov., a new cryptic species of the American Rubyspot complex. *Zootaxa*. 2020;4766(3):485–497.

Vega-Sánchez YM, Mendoza-Cuenca LF, González-Rodríguez A. Complex evolutionary history of the American Rubyspot damselfly, *Hetaerina americana* (Odonata): evidence of cryptic speciation. *Mol Phylogenet Evol*. 2019;139:106536.