Whole genome assembly of *Culex tarsalis*

Bradley J. Main[1], C. Titus Brown[2], Matteo Marcantonio[1], Christopher M. Barker[1]
[1]Department of Pathology, Microbiology and Immunology, University of California, Davis, California, USA
[2]Department of Population Health and Reproduction, University of California, Davis, CA 95616

Abstract

The mosquito, *Culex tarsalis,* is a key vector species in the western United States due to its role in transmission of zoonotic arboviruses that affect human health. Extensive research has been conducted on *Cx. tarsalis* ecology, feeding behavior, vector competence, autogeny, diapause, genetics, and insecticide resistance. However, the lack of a published reference genome has limited genomic analyses for this species, although a previous population genetic analysis of microsatellite allele frequencies across the western U.S. identified three geographic clusters. Salivary gland-specific gene expression has also revealed genes involved in blood feeding. To facilitate genomic studies in this important vector species, we have assembled and annotated a reference genome (CtarK1) based on PacBio HiFi reads from a single male and 10X linked-reads. Genome completeness was 71.5% based on the Benchmarking Universal Single-Copy Orthologs (BUSCO) tool and the N50 was 59 kb. The largest contig was 5.7Mb, including 678kb of unknown bases. CtarK1 is 37% larger than *Cx. quinquefasciatus* (800Mb vs. 579Mb); a difference that is largely transposable elements. Using the *Cx. tarsalis* transcriptome and protein sequences from *Culex quinquefasciatus*, 15,682 genes were annotated in the CtarK1 genome. Based on full mitochondrial genome alignments we present a Bayesian phylogeny and estimate the divergence time from *Culex quinquefasciatus* at 13 MYA (+/- 4).

Introduction

The mosquito, *Culex tarsalis*, is one of the most important vectors in the western United States due to its capacity to transmit several arboviruses that cause disease in humans and horses. In agricultural areas, it is the principal vector for West Nile virus (Goddard *et al.* 2002; Reisen 2013), and such areas have the highest incidence of West Nile virus disease. Extensive research has been done on *Cx. tarsalis* ecology (Reisen 2012), feeding behavior (Thiemann *et al.* 2012; Reisen *et al.* 2013), vector competence (Kramer *et al.* 1981; Reisen *et al.* 2014), autogeny (Spadoni *et al.* 1974; Reisen 1995), diapause (Reisen 1986; Buth *et al.* 1990; Reisen *et al.* 1995), and insecticide resistance (Ziegler *et al.* 1987).

The lack of a published reference genome has limited genomic analyses for *Cx. tarsalis*, but some progress has been made with the development of a genetic linkage map (Venkatesan *et al.* 2009). In addition, a population genetic analysis across the entire western U.S. identified three genetically distinct populations using 12 microsatellite

markers: the Pacific, Sonoran, and Midwest genetic clusters (Venkatesan and Rasgon 2010).

Here we describe the first publicly available genome assembly for *Cx. tarsalis.* The assembly is based on PacBio HiFi reads from a single adult male and 10X genomic data was used for assembly of the Mitochondrial genome. This resource will facilitate the characterization of sex determination in the system, future landscape genetics/genomics, and additional phylogenetic studies among mosquito species.

Methods

Mosquitoes

All *Cx. tarsalis* used for the genome assembly were sampled from the Kern National Wildlife Refuge (KNWR) colony, which was established in 2002 from mosquitoes collected at the Kern National Wildlife Refuge (35.7458°N, 118.6179°W), in Kern County, CA, USA. For PacBio sequencing, high molecular weight (HMW) DNA was extracted at the UC Berkeley DNA Sequencing Facility from a single male adult. For 10X genomics library preparation, HMW DNA was extracted by the UC Davis DNA Technologies Core from two late-eclosing, relatively large pupae (likely female). Two pupae were used for 10X because DNA yields were too low from a single individual. In addition, DNA from a single adult male was used to make a Nextera library (Illumina) for standard genome sequencing with paired-end 75bp reads.

Pacbio genome assembly

PacBio sequencing was performed on a Sequel II SMRT cell at the UC Berkeley sequencing core. Circular consensus sequences (CCS) were then generated and filtered with high stringency to get HiFi CCS reads. We generated an initial genome assembly using Canu v1.9 with the following settings: genomeSize=0.875g, useGrid=false. The genome coverage was low, but because these were HiFi reads, we reduced the standard minimum coverage thresholds with the following options: stopOnLowCoverage=1, contigFilter="2 0 1.0 0.5 0". The Pacbio mitochondrial contig appeared to be two full mitochondrial genomes stuck end to end. As a result, we removed this contig and replaced it with the full mitochondrial genome generated from the 10X assembly. Then we performed two rounds of genome polishing using racon (v1.4.3) and 75bp Nextera reads (Illumina) from a single adult male from the KNWR colony. Annotations were performed using Maker (v.2.31.10) with the *Culex tarsalis* transcriptome (Ribeiro *et al.* 2018) and *Cx. quinquefasciatus* protein sequences (CpipJ2.4). The mitochondrial genome was annotated using DOGMA (Wyman *et al.* 2004). Repeat masking was performed in parallel using the maker annotation pipeline. As input, we used the standard RepBase database (RepBaseRepeatMaskerEdition-

20170127) and a *Cx. tarsalis*-specific repeat library that was generated using RepeatModeler (v1.0.11).

10X linked-read genome assembly

The HMW DNA extraction, 10X Genomics Chromium sequencing library, and Illumina sequencing was performed at the UC Davis DNA Technologies Core. Raw linked reads were assembled using the Supernova 2.0.1 software (10X Genomics) on an AWS instance with 480Gb on RAM and 64 logical cores. To generate a fasta formatted reference sequence, we used supernova mkoutput with --style=pseudohap. This option arbitrarily selects haplotypes across the genome resulting in one pseudo-haploid assembly composed of a mosaic of paternal and maternal haplotype stretches.

Phylogenetic analyses

The mitochondrial analysis included 15,164bp of mitochondrial sequence from *Cx. tarsalis* (CtarK1, this study), *Cx. quinquefasciatus* (GU188856.2), *Anopheles gambiae* (PEST), *Aedes aegypti* (NC_035159.1), *Aedes albopictus* (AaloF1), and *Drosophila melanogaster* (r6.29). The evolutionary analysis was performed using MEGA X (Kumar *et al.* 2018). A multiple-species alignment was generated with Multiple Sequence Comparison by Log-Expectation (MUSCLE) using default parameters. The timetree was then inferred using the Reltime method (Tamura *et al.* 2012; Kumar *et al.* 2018) and the Tamura-Nei model (Tamura and Nei 1993). The timetree was computed using published calibration constraints: 71 +/- 32 million years ago (MYA) for *Aedes* species, 179 +/- 33 MYA between *Culex* and *Aedes*, 217 +/- 37 MYA between *Culex* and *Anopheles,* and 260 +/- 30 MYA between *Drosophila* and mosquitoes (Chen *et al.* 2015). All clades with priors were considered monophyletic.

Results and Discussion

Genome assembly

The PacBio library generated from a single adult male *Cx. tarsalis* was sequenced on the Sequel II platform and yielded 5.7M raw reads and 988,512 ccs HiFi reads. The HiFi reads were used to assemble a draft assembly using Canu with reduced filtering thresholds (see methods). This resulted in 19,994 contigs and a total genome size of 800M (Quast v5.0.2). The N50 was 58,695bp, the GC content was 36% and the largest contig was 756,557bp (Table 1). To assess the quality of the CtarK1 assembly, we searched for the presence of 2799 Benchmarking Universal Single-Copy Orthologs (BUSCO) genes (Zdobnov *et al.* 2017) using BUSCO (v3) (Waterhouse *et al.* 2018). Using this approach, we detected 71.5% (2000/2799) as complete single copy genes, 8.2% (229/2799) as complete and duplicated, and 5.3% (149/2799) were fragmented. Improving the genome completeness and performing genetic scaffolding (e.g. with Hi-C) is needed to make CtarK1 comparable in quality to other mosquito genomes (Figure 1).

The 10X chromium library was sequenced on Illumina's Novaseq platform, yielding approximately 507 million clusters passing filter. Based on several trial assemblies, we downsampled the total read input to 350 million paired-end reads to yield approximately 56x coverage. A contig containing the complete mitochondrial genome was then identified and redundant sequence (due to circular genome) was trimmed at each end of the contig.

Genome annotation

We identified 15,682 genes in this assembly version using the maker annotation pipeline (see methods). This is approximately 4k fewer genes (79%) than *Cx. quinquefasciatus* (Table 1). As the BUSCO score was 71.5%, the genome size may increase significantly in later versions. Repetitive elements were annotated in parallel using RepeatMasker and a custom repeat library. In total, approximately 60% of CtarK1 was annotated as a repeat feature. This is double the estimate from *Cx. quinquefasciatus* (Arensburger *et al.* 2010), indicating that the *Cx. tarsalis*-specific genome expansion (800Mb vs 579Mb) is mostly composed of transposable elements.

Table 1. Assembly statistics

Assembly	Genes	Contig #	Median contig (N50)	Total length
Cx. quinquefasciatus (CpipJ2.4)	19,793	3,172	486,756 bp	~579Mb
Cx. tarsalis (CtarK1)	15,682	20,008	58,695 bp	~800Mb

Phylogenetic analysis

Based on multiple-species alignments of complete mitochondrial genomes (~15kb), we estimated the divergence time between *Cx. tarsalis* and *Cx. quinquefasciatus* to be approximately 13MYA (95% Credible Interval = 9-16; Figure 2). The relatively recent divergence estimate compared to the the *Aedes* species is notable and thus it will be important to compare this estimate with whole genome comparisons. The closer phylogenetic relationship between *Culex tarsalis* and *Aedes mosquitoes* compared to *Anopheles* is consistent with previous genome-wide estimates (Severson and Behura 2012).

Conclusions

Here, we present the first genome assembly and genomic analysis for *Cx. tarsalis*. Based on full mitochondrial genome alignments, we estimate that *Cx. tarsalis* diverged from *Cx. quinquefasciatus* approximately 13 MYA, well after the last common ancestor between *Aedes albopictus* and *Aedes aegypti*. Phylogenetic analyses based on whole

genome alignments are still needed to improve these estimates. The *Cx. tarsalis* genome (793Mb) is 37% larger than the *Cx. quinquefasciatus* genome (578Mb), and this difference appears to be mostly composed of transposable elements. Future work will include comparative genomics within *Culex* versus among *Anopheles* species (Neafsey *et al.* 2015), with a particular focus on genes involved in diapause. *Cx. tarsalis* is a well-studied species that is now wide open for basic genetic studies and landscape genomics.

Acknowledgements:
We thank the Alameda County Mosquito Abatement District for sponsoring a portion of the genome assembly work, and especially Eric Haas-Stapleton and Ryan Clausnitzer for their enthusiastic support of our research on *Culex tarsalis* genomics. We acknowledge funding from an Innovative Development Award from the UC Davis Academic Senate and the Vector-Borne Disease Pilot Grant program of the UC Davis School of Veterinary Medicine. The sequencing was carried out by the DNA Technologies Core and Expression Analysis Core at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. MM and CMB acknowledge funding support from the Pacific Southwest Center of Excellence in Vector-Borne Diseases funded by the U.S. Centers for Disease Control and Prevention (Cooperative Agreement 1U01CK000516). CTB was supported by the Gordon and Betty Moore Foundation under grant GBMF4551.

Data availability
The reference genome and annotation file were deposited at the open science framework DOI: 10.17605/OSF.IO/MDWQX

References Cited

Arensburger, P., K. Megy, R. M. Waterhouse, J. Abrudan, P. Amedeo *et al.*, 2010 Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. Science 330: 86–88.

Buth, J. L., R. A. Brust, and R. A. Ellis, 1990 Development time, oviposition activity and onset of diapause in Culex tarsalis, Culex restuans and Culiseta inornata in southern Manitoba. J. Am. Mosq. Control Assoc. 6: 55–63.

Chen, X.-G., X. Jiang, J. Gu, M. Xu, Y. Wu *et al.*, 2015 Genome sequence of the Asian Tiger mosquito, Aedes albopictus, reveals insights into its biology, genetics, and evolution. Proc. Natl. Acad. Sci. U. S. A. 112: E5907–15.

Goddard, L. B., A. E. Roth, W. K. Reisen, and T. W. Scott, 2002 Vector competence of California mosquitoes for West Nile virus. Emerg. Infect. Dis. 8: 1385–1391.

Kramer, L. D., J. L. Hardy, S. B. Presser, and E. J. Houk, 1981 Dissemination barriers for western equine encephalomyelitis virus in Culex tarsalis infected after ingestion of low viral doses. Am. J. Trop. Med. Hyg. 30: 190–197.

Kumar, S., G. Stecher, M. Li, C. Knyaz, and K. Tamura, 2018 MEGA X: Molecular

Evolutionary Genetics Analysis across Computing Platforms. Mol. Biol. Evol. 35: 1547–1549.

Neafsey, D. E., R. M. Waterhouse, M. R. Abai, S. S. Aganezov, M. A. Alekseyev *et al.*, 2015 Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. Science 347: 1258522.

Reisen, W. K., 2013 Ecology of West Nile virus in North America. Viruses 5: 2079–2105.

Reisen, W. K., 1995 Effect of temperature on Culex tarsalis (Diptera: Culicidae) from the Coachella and San Joaquin Valleys of California. J. Med. Entomol. 32: 636–645.

Reisen, W. K., 1986 Overwintering Studies on Culex tarsalis (Diptera: Culicidae) in Kern County, California: Life Stages Sensitive to Diapause Induction Cues. Ann. Entomol. Soc. Am. 79: 674–676.

Reisen, W. K., 2012 The Contrasting Bionomics of Culex Mosquitoes in Western North America. J. Am. Mosq. Control Assoc. 28: 82–91.

Reisen, W. K., Y. Fang, and V. M. Martinez, 2014 Effects of temperature on the transmission of West Nile virus by Culex tarsalis (Diptera: Culicidae). J. Med. Entomol. 43: 309–317.

Reisen, W. K., H. D. Lothrop, and T. Thiemann, 2013 Host Selection Patterns of Culex tarsalis (Diptera: Culicidae) at Wetlands Near the Salton Sea, Coachella Valley, California, 1998–2002. J. Med. Entomol. 50: 1071–1076.

Reisen, W. K., P. T. Smith, and H. D. Lothrop, 1995 Short-term reproductive diapause by Culex tarsalis (Diptera: Culicidae) in the Coachella Valley of California. J. Med. Entomol. 32: 654–662.

Ribeiro, J. M. C., I. Martin-Martin, F. R. Moreira, K. A. Bernard, and E. Calvo, 2018 A deep insight into the male and female sialotranscriptome of adult Culex tarsalis mosquitoes. Insect Biochem. Mol. Biol. 95: 1–9.

Severson, D. W., and S. K. Behura, 2012 Mosquito genomics: progress and challenges. Annu. Rev. Entomol. 57: 143–166.

Spadoni, R. D., R. L. Nelson, and W. C. Reeves, 1974 Seasonal Occurrence, Egg Production, and Blood-feeding Activity of Autogenous Culex tarsalis,. Ann. Entomol. Soc. Am. 67: 895–902.

Tamura, K., F. U. Battistuzzi, P. Billing-Ross, O. Murillo, A. Filipski *et al.*, 2012 Estimating divergence times in large molecular phylogenies. Proc. Natl. Acad. Sci. U. S. A. 109: 19333–19338.

Tamura, K., and M. Nei, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10: 512–526.

Thiemann, T. C., D. A. Lemenager, S. Kluh, B. D. Carroll, H. D. Lothrop *et al.*, 2012 Spatial variation in host feeding patterns of Culex tarsalis and the Culex pipiens complex (Diptera: Culicidae) in California. J. Med. Entomol. 49: 903–916.

Venkatesan, M., K. W. Broman, M. Sellers, and J. L. Rasgon, 2009 An initial linkage map of the West Nile Virus vector Culex tarsalis. Insect Mol. Biol. 18: 453–463.

Venkatesan, M., and J. L. Rasgon, 2010 Population genetic data suggest a role for mosquito-mediated dispersal of West Nile virus across the western United States. Mol. Ecol. 19: 1573–1584.

Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis *et al.*, 2018 BUSCO

Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. 35: 543–548.

Wyman, S. K., R. K. Jansen, and J. L. Boore, 2004 Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252–3255.

Zdobnov, E. M., F. Tegenfeldt, D. Kuznetsov, R. M. Waterhouse, F. A. Simão *et al.*, 2017 OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res. 45: D744–D749.

Ziegler, R., S. Whyard, A. E. R. Downe, G. R. Wyatt, and V. K. Walker, 1987 General esterase, malathion carboxylesterase, and malathion resistance in Culex tarsalis. Pestic. Biochem. Physiol. 28: 279–285.

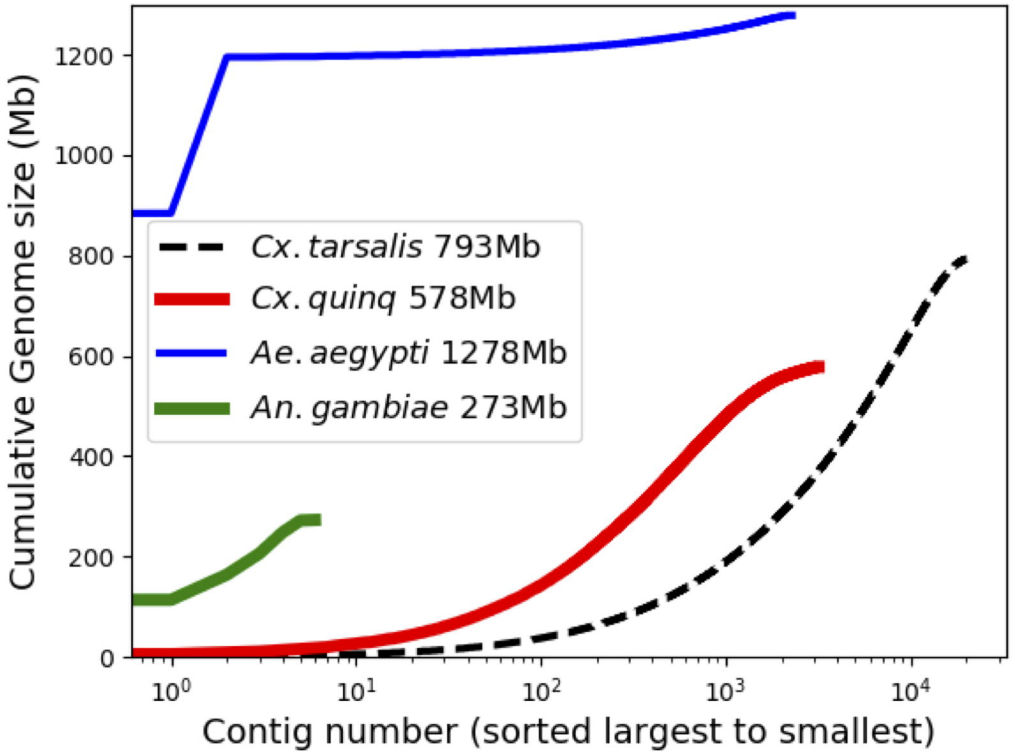

Figure 1. Cumulative genome size (minus N's) comparison between the *Cx. tarsalis* CtarK1 assembly (dashed black) and the complete reference genomes of *Cx. quinquefasciatus* (CpipJ2, solid red), *Ae. aegypti* (LVP_AGWG, solid blue), and *An. gambiae* (AgamP4, solid green). Higher quality, chromosome-scale genomes have steeper slopes and fewer contigs. X-axis is on a log scale.

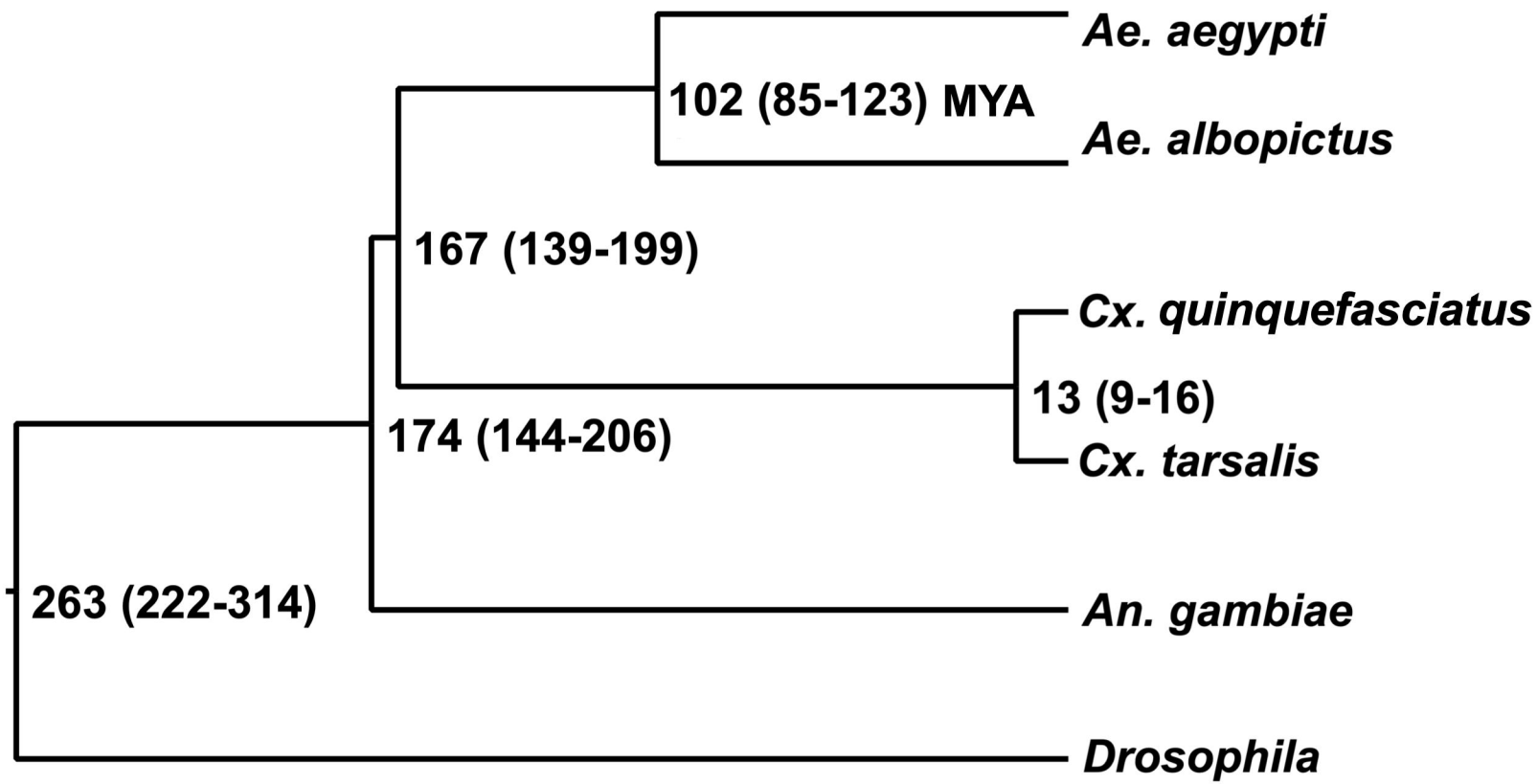

Figure 2. Phylogeny based on a Bayesian analysis of Mitochondrial genomes. Divergence times are in million years ago (with 95% credible intervals). Divergence time priors were selected from Chen et al. 2015.