Speciation in the face of long-range dispersal: population genomic structure within a species complex of biting midges

Phillip Shults[1], Matthew Hopken[2], Pierre-André Eyer[1], Alexander Blumenfeld[1], Mariana Mateos[3], Lee Cohnstaedt[4], and Edward Vargo[3]

[1]Texas A&M University College Station
[2]USDA-APHIS National Wildlife Research Center
[3]Texas A&M University
[4]USDA-ARS Arthropod-Borne Animal Diseases Research Unit

May 7, 2021

Abstract

The level of gene flow between diverging lineages ultimately determines the outcome of a speciation event. If secondary contact occurs before this process is complete, reproductive isolation barriers must exist or evolve to prevent hybridization. The selective pressures facilitating and maintaining genetic divergence do not always involve an observable phenotypic response, thus cryptic species form. The inability to distinguish between sibling species can be a particularly serious problem in groups responsible for pathogen transmission. Culicoides biting midges occur almost world-wide and vector many disease-causing pathogens that affect wildlife and livestock. In North America, the C. variipennis species complex contains three currently recognized species, only one of which is a vector, and limited molecular and morphological differences have hindered vector surveillance. Here, genomic methods were used to investigate speciation and genetic structure within this complex. Single nucleotide polymorphism (SNP) data were generated using ddRAD sequencing for 206 individuals originating from 17 locations throughout the United States and Canada. Clustering analyses consistently suggest the occurrence of five putative species with significant differentiation occurring in both sympatric and allopatric populations. Evidence of hybridization was detected in three different species pairings, indicating a lack of pre-zygotic reproductive isolation within the complex. Mitochondrial genes were used to trace the hybrid parentage of these individuals, which illuminated discordance with the SNP data. In this study, we highlight the potential role of geographic, ecological, and behavioral isolation in speciation and in maintaining species boundaries, despite hybridization and long range dispersal.

Speciation in the face of long-range dispersal: population genomic structure within a species complex of biting midges

Phillip Shults[1*], Matthew Hopken[2], Pierre-Andre Eyer[1], Alexander Blumenfeld[1], Mariana Mateos[3], Lee W. Cohnstaedt[4*], Edward L. Vargo[1]

1 - Department of Entomology, Texas A&M University, College Station, TX, 77843, USA

2 - USDA APHIS Wildlife Services National Wildlife Research Center, Fort Collins, CO, 80521, USA, and Department of Microbiology, Immunology, and Pathology, Colorado State University, Fort Collins, CO, 80523, USA

3 - Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX, 77843, USA

4 - USDA-ARS Arthropod Borne Animal Disease Research Unit, 1515 College Ave, Manhattan, KS 66502, USA.

*Corresponding authors: Phillip Shults (ptshults@tamu.edu) and Lee Cohnstaedt (lee.cohnstaedt@usda.gov).

Abstract: The level of gene flow between diverging lineages ultimately determines the outcome of a speciation event. If secondary contact occurs before this process is complete, reproductive isolation barriers must exist or evolve to prevent hybridization. The selective pressures facilitating and maintaining genetic divergence do not always involve an observable phenotypic response, thus cryptic species form. The inability to distinguish between sibling species can be a particularly serious problem in groups responsible for pathogen transmission. *Culicoides* biting midges occur almost world-wide and vector many disease-causing pathogens that affect wildlife and livestock. In North America, the *C. variipennis* species complex contains three currently recognized species, only one of which is a vector, and limited molecular and morphological differences have hindered vector surveillance. Here, genomic methods were used to investigate speciation and genetic structure within this complex. Single nucleotide polymorphism (SNP) data were generated using ddRAD sequencing for 206 individuals originating from 17 locations throughout the United States and Canada. Clustering analyses consistently suggest the occurrence of five putative species with significant differentiation occurring in both sympatric and allopatric populations. Evidence of hybridization was detected in three different species pairings, indicating a lack of pre-zygotic reproductive isolation within the complex. Mitochondrial genes were used to trace the hybrid parentage of these individuals, which illuminated discordance with the SNP data. In this study, we highlight the potential role of geographic, ecological, and behavioral isolation in speciation and in maintaining species boundaries, despite hybridization and long range dispersal.

KEYWORDS

Culicoides , species complex, *C. sonorensis* , gene flow, genetic discordance, vector

Introduction

Speciation is a dynamic evolutionary process through which populations segregate into independently evolving lineages over time (De Queiroz, 2007). When gene flow is restricted between populations, the accumulation of genetic changes, through selection or local genetic drift, may lead to genetic differentiation and potentially reproductive isolation (Coyne & Orr, 2004; Endler, 1973; Mayr, 1999; Richardson, Urban, Bolnick, & Skelly, 2014). This restriction of gene flow occurs through either geographic or ecological isolation, though these are not mutually exclusive (Nosil, 2008). Thus, the level of gene flow between divergent populations is a contributing factor to the rate of speciation, as well as to the spatial level at which it occurs (Kisel & Barraclough, 2010).

Geographic isolation reduces migration between populations, and thus, life-history traits influencing dispersal ability can drastically influence the level of gene flow among populations (*e.g.* (Bolnick & Otto, 2013; Slatkin, 1993). Species with low dispersal ability are particularly likely to exhibit highly differentiated populations resulting in the evolution of cryptic species over a limited spatial scale (Pante, Puillandre, et al., 2015). In contrast, species with high dispersal abilities are likely to maintain gene flow between populations, therefore a process outside of geographic isolation is needed to initiate the speciation process in these instances (Claramunt, Derryberry, Remsen Jr, & Brumfield, 2012; Enbody et al., 2021; Palumbi, 1994). Two such mechanisms for this are ecological and behavioral isolation whereby sympatric populations occupy distinct ecological niches (Rundle & Nosil, 2005; Wang & Bradburd, 2014). The difference between these populations, such as habitat type, differences in mating times, or sexual selection reduce the frequency of interbreeding and this can lead to genetic divergence (Coyne & Orr, 2004; Dobzhansky, 1982). Ecological and behavioral isolation can drive lineage divergence through selection, and subsequent pre-zygotic isolation can further increase divergence through reinforcement, accentuating the speciation process (Coyne & Orr, 2004). However, with an increase in specialization, fragmented distributions of either habitat or host may further reduce gene flow between populations (Thompson, 1999; Thompson & Cunningham, 2002; West & Herre, 1994).

Each reproductive isolation mechanism can lead to similar morphological adaptations and genomic signa-

tures (Keller et al., 2013; Seehausen et al., 2014), and this can make it challenging to interpret what factors contributed to the speciation event. While the most accurate assumptions about species delimitation are derived from a multifaceted approach (Carstens, Pelletier, Reid, & Satler, 2013; Schlick-Steiner et al., 2010), gene flow is directly tied to the fate of incipient species. By using a population genetic approach to study microevolution and incipient speciation, we can identify independent lineages and measure the introgression between them to better understand the evolutionary processes underlying species divergence. Species delimitation is especially important when working with organisms responsible for pathogen transmission, as misidentifications will lead to inaccurate vector competency and surveillance data. *Culicoides* Latreille (Diptera: Ceratopogonidae) biting midges are responsible for the transmission of many disease-causing agents worldwide (Borkent, 2004; Mellor, Boorman, & Baylis, 2000), including bluetongue virus (BTV) and epizootic hemorrhagic disease virus (EHDV). These viruses can cause severe symptoms and death in wild and domestic ungulates and are responsible for substantial economic losses globally (Rushton & Lyons, 2015; Tabachnick, 1996).

In North America, one of the main BTV and EHDV vectors is *Culicoides sonorensis* Wirth and Jones, which belongs to the *C. variipennis* species complex. When originally described, this group consisted of five subspecies (Wirth & Jones, 1957), but it is currently composed of three distinct species (*C. occidentalis* Wirth and Jones, *C. sonorensis* , and*C. variipennis* (Coquillet)) (Holbrook et al., 2000). Despite the current taxonomic arrangement, species identification remains difficult due to very subtle adult morphological differences and genetic similarity. Additionally, cryptic species could make vector incrimination and species distribution records potentially unreliable. Measuring genetic divergence between species and populations can be useful in vector biology as vectorial capacity and host association become increasingly variable with increased genetic distance (Byrne & Nichols, 1999; McCoy, Boulinier, Tirard, & Michalakis, 2001). Population genetic studies of *Culicoides* species in Europe, Africa, and Australia have consistently revealed frequent gene flow between populations, even at continental scales (Jacquet et al., 2015; Jacquet et al., 2016; Onyango et al., 2016; Onyango, Michuki, et al., 2015). Their high dispersal ability, likely wind-mediated (Ducheyne et al., 2007; Purse et al., 2005), decreases the likelihood of geographic isolation between populations of *Culicoides* spp. Under laboratory conditions, the species within the *C. variipennis* complex have been shown to hybridize (Velten & Mullens, 1997), and while *C. occidentalis* and *C. variipennis* are not known to be competent vectors, both species occur sympatrically with *C. sonorensis*(Wirth & Jones, 1957). This lack of post-zygotic reproductive isolation, coupled with a high dispersal ability and numerous sympatric populations, makes this species complex an intriguing system in which to study speciation and may also provide insights into the evolutionary mechanisms responsible for vector competence in this group.

Here, we evaluated the geographic connectivity within and among the species of the *C. variipennis* complex by assessing the level of gene flow within and across populations. We used a high-throughput ddRadSeq protocol to analyze 206 individuals collected from 17 sites throughout the United States and Canada. We first estimated the overall genetic similarity and population structure among these samples to determine distinct lineages within the species complex. We then estimated the level of gene flow between the inferred species, as well as uncovered hybridization between sympatric species. As previous attempts to separate these species using common barcoding genes have been inconclusive, we sequenced a region of COI to compare to the inferred SNP identifications. One species, was found to have two distinct geographic haplogroups, while three other species shared a single haplogroup. Additionally, we assessed the potential drivers of divergence in this species complex by assessing loci under selection for each species, as well as discuss the potential mechanisms controlling reproductive isolation.

Materials and Methods

Sample collection and sequencing

Culicoides midges were collected from 17 sites across the United States and Canada (Table 1). Specimens were collected either as pupae and reared to adulthood, or as adults using CDC light traps baited with CO_2 and UV light (Bioquip 2836BQ). Individuals morphologically assigned to the *C . variipennis* complex were sorted out from the by-catch and stored in 95% ethanol at -80 °C. Total DNA was extracted from

individuals (females only) using a Puregene extraction protocol (Gentra Systems, Inc., D-5500A) with the addition of glycogen (ThermoFisher, R0561) to increase yields. The DNA quality was checked using gel electrophoresis and DNA concentration was measured using a Qubit 3.0 fluorometer and a Qubit dsDNA HS assay kit (Invitrogen, Q33230). A total of 300-400 ng of DNA per sample was sent to Floragenex, Inc. for library preparation using the protocol from Truong et al. (2012). DNA was digested using the restriction enzymes*MseI* and *PstI* . After PCR amplification, the samples in each plate were pooled and sequenced on a lane of single-end 100bp sequencing on a HiSeq4000 at the University of Oregon Genomics Facility, Eugene, OR.

Raw sequence filtering and processing

Raw sequence quality was first assessed using FastQC v.0.11.9 and MultiQC v.1.7 (Andrews, 2010; Ewels, Magnusson, Lundin, & Käller, 2016), and then reads were filtered and processed using Stacks v.2.3 (Rochette, Rivera-Colon, & Catchen, 2019). Reads with a phred score below 25 were removed as well as individuals with a >75.0% missing data. Next, reads were aligned to the *C. sonorensis*genome (Morales-Hojas et al., 2018) (Accession: PRJEB19938) using the Burrows-Wheeler Aligner (BWA-mem) (Li & Durbin, 2009). Finally, aligned reads were run through the reference-based pipeline of Stacks, with filtering parameters set to keep SNPs occurring in at least half of the sampling locations and at least 50% of individuals within those sites (Pante, Abdelkrim, et al., 2015). The minimum allele frequency was set to 0.05 to protect against potential sequencing errors (Benestan et al., 2016), and only the first SNP per locus was kept to minimize linkage disequilibrium between SNPs from influencing population structure and phylogenetic analyses. All subsequent file reformatting was done with PGDSpider v.2.1.1.5 (Lischer & Excoffier, 2012).

Clustering Analysis

Population structure in the overall dataset was evaluated using fastSTRUCTURE v.1.04, with Structure_threader utilized to parallelize distinct runs of K (Pina-Martins, Silva, Fino, & Paulo, 2017; Raj, Stephens, & Pritchard, 2013). Initially, samples were grouped by location and no species data were pre-assigned to the individuals. To estimate the most likely number of genetic clusters in the dataset (K), the analysis was run for values of K ranging from 1 to 17 (*i.e.* , number of sites sampled). The best value was selected using the*chooseK.py* function from the fastSTRUCTURE package and plots were created by Distruct v.2.3 (http://distruct2.popgen.org). The clustering of individuals into the distinct genetic groups were also visualized using a principal component analysis (PCA) and a discriminant analysis of principal components (DAPC). The most likely number of genetic groups was inferred by the *find.clusters* algorithm for the PCA and the optimal number of principal components to inform the DAPC was defined using the function *optim.a.score* . Both were performed in R (Team, 2013) through the *adegenet* package (Jombart & Collins, 2015).

Any individual with more than 25% of their loci grouping with a second cluster was marked as a hybrid and removed from the phylogenetic analysis. Maximum likelihood phylogeny among individuals was run using RAxML v.8.2.12 (Stamatakis, 2014). An acquisition bias correction was applied to the likelihood calculations as alignments were solely composed of SNPs, with each invariant site removed through Phrynomics (https://github.com/bbanbury/phrynomics) (Leache, Banbury, Felsenstein, De Oca, & Stamatakis, 2015). The GTR+G nucleotide substitution model was used for each search. A rapid bootstrap analysis and search for the best-scoring maximum likelihood tree was executed using the extended majority rule-based bootstopping criterion to achieve a sufficient number of bootstrap replicates (Pattengale, Alipour, Bininda-Emonds, Moret, & Stamatakis, 2010). Additionally, to cross-validate our results, a second phylogeny was inferred in W-IQ-Tree version 1.6.12 (Trifinopoulos, Nguyen, von Haeseler, & Minh, 2016), using the TVM+F+G4 substitution model determined by ModelFinder (Kalyaanamoorthy, Minh, Wong, Von Haeseler, & Jermiin, 2017; Nguyen, Schmidt, Von Haeseler, & Minh, 2015). Branch support was calculated using 1000 ultrafast bootstraps (Hoang, Chernomor, Von Haeseler, Minh, & Vinh, 2018) and Shimodaira–Hasegawa like approximate likelihood-ratio test (SH-aRLT) (Guindon et al., 2010; Hoang et al., 2018).

As each of these clustering methods consistently supported five genetically distinct clusters, we generated

a SNP dataset with individuals assigned to both a sampling location and a cluster ("all-species" dataset) as well as four species-specific datasets. SNPs were generated from the raw reads following the processing methods above except the filtering parameters were increased to only include SNP that occurred in at least 75% of the populations and at least half of the individuals within those populations. Genetic diversity estimates (F_{IS} , H_E , andH_O), population differentiation (pairwiseF_{ST}), and isolation-by-distance (IBD) were calculated for each SNP dataset using Genepop v.4.7.0 (Rousset, 2017). Geographic distances were calculated as Euclidean distances among localities.

Mitochondrial Sequencing and haplotype network

Mitochondrial DNA haplotypes were obtained from a subset of 67 individuals from the five genetic clusters. PCR reactions were performed using a Taq-Pro COMPLETE kit (Denville Scientific, CB4065-4) targeting a partial region of the COI gene with the Lep50 primer set from Folmer et al. (1994) and the thermocycler profile from Herbert et al. (2003). PCR products were cleaned using an EXOSAP-IT kit (ThermoFisher, 78201.1.ML) and prepared for sequencing using a BigDye Terminator v.3.1 Cycle Sequencer Kit (Applied Biosystems, 4337454). Sanger sequencing was done using an Applied Biosystems 3500 Genetic Analyzer. Chromatograms were cleaned and aligned using the program Geneious v.9.1 (Kearse et al., 2012).

A haplotype network analysis was conducted using the 67 COI sequences obtained in this study combined with 218 *C. variipennis* complex sequences previously collected (M. Hopken unpublished data). Sequences were aligned in MEGA v.10.1.8 (Kumar, Stecher, Li, Knyaz, & Tamura, 2018) and trimmed to 546 bp to ensure all sequences contained identical lengths. A median-joining analysis was performed using NETWORK v.5.0.1.0 (Bandelt, Forster, & Rohl, 1999). Specimens collected in this study were assigned a color based on the results from the SNP clustering analyses while the remaining samples were left unassigned. All individuals were used to calculate the mean uncorrected$p$- divergence between and within the different groupings inferred from the haplotype network using MEGA.

Outlier loci detection

The "all-species" and species-specific datasets analyzed in Genepop were also run through Bayescan v.2.1 to identify loci under divergent selection (Foll, 2012). Parameters of the Markov chain Monte Carlo algorithm were set to 20 pilot runs of 5000 iterations. Afterward, a burn-in of 50,000 iterations followed by 50,000 iterations were used for estimation with a thinning interval of 10. Jeffrey's scale was used to interpret selection per loci (Jeffreys, 1998). Loci with a log10 value > 0.5 are considered to have "substantial" evidence of selection and those with a log10 value > 1.0 have "strong" evidence of selection. To identify loci under selection across clusters another new SNP dataset was generated by filtering to include only those occurring in all five clusters and 75% of the individuals within each cluster. The nucleotide sequences for each locus found to be under selection were submitted for an alignment search in the InsectBase and Flybase databases (Thurmond et al., 2019; Yin et al., 2016).

Results

All of the raw sequence reads, alignment files, and genotype datasets will be deposited in the Open Science Framework database upon acceptance, https://osf.io (DOI XXX). In total, 271 individuals were subjected to the ddRADseq procedure and yielded an average of 2.08 million reads per individual. During the initial filtering, 36 individuals had a phred score of less than 25 and were removed from the dataset. Additionally, 29 individuals were found to have more than 75% missing data and were therefore removed. The final dataset included 206 individuals from 17 sites and contained 3612 SNPs. The population structure inferred by fastSTRUCTURE that best explains the data is K = 5. Structure plots showing K = 3-6 can be found in figure S1. At K = 5, most individuals (86%) were unambiguously assigned to one group (98-100% assignment score; Fig. 1). Consistent with these results, the PCA and DAPC grouped these individuals into five main clusters (Figs. 2a & S2). The main difference being that the PCA further segregated one cluster (blue, Fig. 2a) into two separate groups; east and west of the Sierra Nevada mountain range. Further support for the same five clusters was found in the maximum likelihood trees, with a high level of support from each approximation method (Figs. 2b & S3).

The geographic distributions of four of these clusters closely align with the distributions of four of the five subspecies described in Wirth & Jones (1957) (Fig. 1), suggesting these morphological descriptions accurately denoted species-level taxa within the *C. variipennis*complex. Further phylogenetic and morphological study is needed to confirm the validity of these species groupings; however, for the remainder of the manuscript we will refer to each cluster by a species name. *Culicoides occidentalis* located in Western North America,*C. sonorensis* in the Western and Southern U.S., *C. albertensis* in the Midwest U.S. and Ontario, *C. variipennis* in the Eastern U.S. and Ontario, and a fifth genetic group suggesting the occurrence of an additional, undescribed cryptic species in San Diego, CA. Notably, eight of the 17 sites had more than one species in sympatry, and one site had three species. At four sites, seven individuals were assigned to two genetic groups with an assignment score of ˜50% for three individuals (scores = 45, 47 and 41%) and of ˜25% for four individuals (scores = 34, 31, 25 and 24%), which suggests the occurrence of putative F1 or other types of hybrids (e.g. F2 or backcrosses), respectively. Interestingly, these hybrids were from three different species parings (*C. sonorensis* X *C. occidentalis* ; *C. sonorensis* X *C. variipennis* ; and*C. albertensis* X *C. variipennis*). These hybrid individuals also stood out using the PCA analysis, as they segregated between their parental clusters (Fig. 2a), as well as at the base of each parental branch in the phylogenetic tree (Fig. S3). In addition to these hybrids, 20 individuals had a secondary assignment score between 3% to 21%, signifying potential introgression between those pairings.

The samples were then rearranged by species, rather than collection site, and stricter filtering parameters were applied. This dataset contained 566 SNPs from 199 individuals (hybrids were excluded) and was used to calculate the species-level summary statistics as well as determine the loci under selection. The mean F_{ST}between the five inferred clusters was 0.7147 (0.6541-0.7470), roughly 9 times higher than the mean F_{ST} between the populations (i.e., localities) within each cluster (see below; Tables 2 & S1). The overall dataset was further split into five datasets for species-level population statistics. These datasets contained 22 individuals of *C. albertensis* from four populations (3423 SNPs), 36 individuals of *C. occidentalis* from four populations (2714 SNPs), 97 individuals of *C. sonorensis* from seven populations (2357 SNPs), and 29 individuals of *C. variipennis* from four populations (2960 SNPs). The expected and observed heterozygosity,F_{IS} , and number of private alleles for each species are reported in Table S2. No species level dataset was created for the San Diego species as only one locality was examined.

When examining each species individually, *C. albertensis,* had no evidence of population structure (K = 1), and had low genetic differentiation among populations (mean F_{ST} = 0.054) (Fig. 3a; Table 2). Although there does seem to be a pattern of isolation by distance, this was found to not be significant in this species (P = 0.238). The low number of populations sampled potentially limits our statistical power for this correlation. The results obtained for *C . occidentalis* showed much more divergence compared to the other species, with populations being strongly differentiated from each other (mean F_{ST} = 0.411) (Table 2). Additionally, fastSTRUCTURE suggests that each population of *C. occidentalis* is a distinct genetic entity (K = 4) clustering by location (Fig. 3b). While no IBD was found (P = 0.489), there seems to be a considerable amount of geographic isolation among populations of this species, with pairwise F_{ST} values ranging from 0.14 to 0.70 (Table S3). Low genetic differentiation among populations was found for *C. sonorensis* (meanF_{ST} = 0.029), despite a slight, but significant IBD in this species (P = 0.039) (Fig. 3c; Table 2). For this reason, the individuals from Colorado were combined into a single population, as were the individuals from Kansas. A fastSTRUCTURE analysis suggested the occurrence of population structure in *C. sonorensis* (K = 2), with some individuals from Kansas belonging to a distinct group. The combined Kansas populations were not divergent from any other *C. sonorensis* population (Table S3). Populations of *C .variipennis* exhibited no evidence of population structure (K = 1) or of isolation by distance (P = 0.587) (Fig. 3d). Consistently, almost no genetic differentiation was found among populations of this species (mean F_{ST} = 0.026) (Table 2).

We identified three outlier loci within the *C. variipennis*complex and an additional 23 species-specific loci: two in *C. albertensis* , seven in *C. occidentalis* , 11 in *C. sonorensis* , and two in *C. variipennis* . Each of these loci had a log10 Bayes factor value over 1 and six had values above 2, corresponding to a 95% and 99% confidence interval, respectively (Fig. 4). Searches of InsectBase were used to assign putative

functional annotations (most of which were provided by Nayduch et al. (2014), with orthologous dipteran genes subsequently found using Flybase (Table S4). Roughly 75% of the loci were matched to transcription data, and all but one associated with a dipteran orthologous gene. None of the loci found to have significant evidence of selection were shared across the different species, suggesting that each is under its own set of selective pressures.

Based on the COI gene, four distinct groupings were identified with strong genetic divergence between groups (p- distance = 2.99-3.30%) and little divergence within groups (p -distance = 0.25-0.86%; Fig. 5; Table 3). Consistent with the SNP datasets, the California population of *C. occidentalis* was separated from the rest of its range. The mean percent divergence between the two *C. occidentalis* groups (2.99%) was similar to its divergence from the other species (3.01-3.30%). The San Diego population clusters as a distinct group, with a similar level of divergence from the other species (3.01-3.03%). Interestingly, *C. albertensis* , *C. sonorensis* , and *C. variipennis* were not separated from each other, and in some cases, *C. albertensis* and *C. variipennis* shared identical haplotypes (Fig. 5). Furthermore, these three species exhibit a mean percent divergence between individuals (0.80%) similar to the divergence observed among individuals within the other clusters (Table 3). Other than the grouping of *C. occidentalis* in California, there was no other geographic clustering observed.

Discussion

Our study provides valuable insights into the population genetics of the *C . variipennis* species complex and highlights the presence of potential cryptic species. For most of the species examined, minimal genetic divergence was observed across populations, suggesting the maintenance of gene flow even over large geographic distances. The only exception was *C. occidentalis* , which showed a high level of geographic isolation, as well as two distinct genetic clusters. We confirmed that mitochondrial data is not reliable to properly differentiate three out of five species, due to the lack of segregation between the mitochondrial haplotypes of *C. albertensis* , *C. sonorensis* , and *C. variipennis* . This stands in stark contrast to their clear differentiation and high level of divergence inferred from the SNP data. Though a substantial amount of divergence exists between all five species, hybridization and introgression are present at low levels in sympatry suggesting that post-zygotic isolation barriers have not evolved in this group. Thus, pre-zygotic isolation through either ecological or behavioral segregation is likely responsible for divergence within this complex. With a considerable amount of overlap between some species (Fig. 1), each sympatric population is potentially experiencing a set of unique selective pressures to maintain species boundaries.

Species delimitation and dispersal capabilities within the C. variipennis complex

The high degree of genetic differentiation between clusters inferred by the SNP data supports the current species groupings of the *C. variipennis* complex (*C. occidentalis* , *C. sonorensis* , and *C. variipennis*), as well as raising *C. albertensis* and a cryptic species in San Diego, California to species status. Little to no IBD or structure was found within populations of *C. albertensis* ,*C. sonorensis* , and *C. variipennis* (Fig. 3a,c,d). The number of populations inferred by fastSTRUCTURE for *C. sonorensis* was K=2; however, a mean pairwise F_{ST} of 0.0287 suggests that a high amount of gene flow still exists between all populations. This could also be an artifact of the propensity of delta K inferring two populations (Janes et al., 2017) or from a high level of relatedness among individuals from KS.

Interestingly, although no IBD was found in *C. occidentalis* , each location of this species clustered as a distinct population. The lack of IBD is therefore not indicative of a single, genetically homogeneous population, but rather stems from high levels of divergence between populations regardless of their geographic distances. In this species, the strong genetic divergence between the population from California and the other populations observed in the SNP data was consistently uncovered in the mtDNA (4.0% divergent, Table 3, Fig. 5). It is possible that this may represent a further cryptic species with a dispersal barrier created by the Sierra-Nevada mountain range. Patchiness of the larval habitat of *C. occidentalis* could also create isolation between populations as well as reduce the number of individuals within each population. A small population size with little to no immigration would allow for a strong effect from drift (Hare, 2001). While the

populations of *C. occidentalis* outside of California were less diverged from one other, the lowest pairwise F_{ST} values between these populations were still greater than the highest pairwise values observed for any other species, consistent with the findings of Holbrook et al. (2000) (Table 2). Interestingly, at one of the three loci found to be under selection with the complex (seipin, Table S4), all populations of *C. occidentalis* and *C. albertensis* were fixed for a single allele, whereas the other three other species were fixed for the other alternative allele. This SNP was determined to be synonymous and therefore unlikely to be the direct target of selection; however, it may be linked to a region of the genome that is.

Similar to other species of *Culicoides* (Jacquet et al., 2015; Mignotte et al., 2021; Onyango, Beebe, et al., 2015; Onyango, Michuki, et al., 2015), high values of the inbreeding coefficient were observed in all species investigated in this study (Table S2). Although these previous studies have suggested that the observed high inbreeding coefficient values are an artifact from a large number of null alleles, the consistent reporting of these findings across various species using several types of molecular markers lends support to the they hypothesis that high inbreeding has a biological origin. High levels of inbreeding and heterozygote deficiencies are common among mosquitoes (Fonseca, Smith, Kim, & Mogi, 2009; Goubert, Minard, Vieira, & Boulesteix, 2016; Lehmann et al., 1997), even when using markers with a low level of null alleles (Chapuis & Estoup, 2006; Manni et al., 2015). Goubert *et al* . (2016) considered the typical *Aedes albopictus* population as "a network of interconnected breeding sites, each with a high level of inbreeding". In this study, although we cannot rule out that the presence of null alleles and we acknowledge that a weak Wahlund effect can contribute to the level of inbreeding, our results strongly suggested that some aspects of the reproductive biology of*Culicoides* induce inbreeding within populations.

mtDNA and nuclear discordance

Culicoides albertensis , *C. sonorensis* , and *C. variipennis* have a considerable amount of genome-wide differentiation (Fig. 1); however, there was no clear differentiation of the COI gene (Fig. 5). In fact, several individuals of *C. albertensis* and*C. variipennis* shared identical haplotypes. Multiple studies have shown a high degree of genetic similarity in mtDNA between *C. sonorensis* and *C. variipennis* (Hopken, 2016; Jewiss-Gaines, Barelli, & Hunter, 2017; Shults, 2015), though it was proposed that this was due to misidentifications. As all of the individuals included in our mitochondrial haplotype analysis from the current study were identified to species using the SNP data, this lack of mitochondrial separation has an underlying biological source. Ongoing hybridization with "leaky" pre-zygotic isolation, or a semipermeable species boundaries, has been shown to produce mitochondrial introgression without detectable nuclear DNA introgression (Chan & Levin, 2005; Harrison, 1990). This is likely due to the fact the mitochondrial genome is independent of the nuclear genome and thus unlinked to the genes contributing to reproductive isolation (Harrison, 1989). This does not appear to be the case throughout the entire complex though as hybridization was also found between *C. sonorensis* and *C. occidentalis* and the mtDNA from these two species was highly divergent.

In addition to the convergence of a single haplogroup by three species,*C. occidentalis* was found to have two distinct haplogroups based on geography (Fig. 5). The mean percent divergence between *C. occidentalis* from California (CABL) and *C. occidentalis* from the other collection sites (BC-NV-UT) was equal to the divergence between the other species in the complex (Table 3). This high level of differentiation within *C. occidentalis* could be due to geographic isolation alone; however, endosymbionts have also been shown to significantly increase mitochondrial diversity in the presence of geographic structure (Ballard, Chernoff, & James, 2002; Behura, Sahu, Mohan, & Nair, 2001). Naturally occurring endosymbionts have been found in *Culicoides* midges, including *C. sonorensis* (Covey et al., 2020; Pages, Munoz-Munoz, Verdun, Pujol, & Talavera, 2017), and recently, a *Cardinium* sp. was linked to mitochondrial divergence of *C . imicola* (Pilgrim et al., 2021). Further screening is needed to determine the diversity and abundance of endosymbionts infecting *Culicoides* midges, though the possibility remains that these could be playing a role in the phylogeographical structure of*C. occidentalis* .

Hybridization and reproduction isolation

Under laboratory conditions, mating between *C. sonorensis* and *C. occidentalis* can produce viable offspring for at least six generations, though the hatch rate of the progeny is dependent on the species of the mother (Velten & Mullens, 1997). A cross of female *C. sonorensis* and male C. *occidentalis* only yields a 7% hatch rate whereas the reciprocal cross yields a 75% hatch rate. This asymmetrical hybrid viability is likely caused by cytonuclear incompatibility (Arntzen, Jehle, Bardakci, Burke, & Wallis, 2009; Gibeaux et al., 2018), though endosymbionts have also been shown to cause reproductive incompatibility (Werren, Baldo, & Clark, 2008). Upon secondary contact of closely related species, and in the absence of post-zygotic reproductive isolation, the production of unfit hybrids can induce the rapid evolution of premating barriers (Coyne & Orr, 2004; Howard, 1993; Servedio & Kirkpatrick, 1997; Yukilevich, 2012). In most populations however, *C. sonorensis* females are unlikely to come across *C. occidentalis* males due to differences in mating behavior. Conversely, *C. occidentalis* females do come into contact with *C. sonorensis* males, who do not appear to have mate discrimination (Downes, 1978), and will likely attempt to mate with these heterospecific females. As there are demographic disparities (population size and structure) between these two species, as well as viable offspring produced from this cross, rampant hybridization and asymmetric introgression would be detrimental to *C. occidentalis*(Toews & Brelsford, 2012). Strong selection against hybridization can maintain species boundaries, but as two of the ten *C. occidentalis* collected from Borax Lake in California (CABL) were F1 hybrids, another mechanism, potentially ecological or behavioral isolation, appears to be limiting directional introgression from *C. sonorensis* .

Culicoides occidentalis females lay their eggs in highly saline environments (up to 88.0 parts per thousand (ppt)) (Smith & Mullens, 2003), and *C. sonorensis* eggs will not hatch in water with salinity over 20.0 ppt (Linley, 1986). Ecological exclusion via the larval habitat is present in this system, but alone would only limit introgression if the survival rate of the hybrids were reduced. *Culicoides occidentalis* mate at the larval habitat while *C. sonorensis* mates at or near a host (Gerry & Mullens, 1998; Holbrook et al., 2000), and this difference in mating behavior may be a more likely mechanism by which the detrimental effects of hybridization are diminished. Most *C. occidentalis* females will mate at the larval habitat, but if this does not happen, they may be mated by *C. sonorensis* while feeding at the host. As *C. occidentalis* females return to the high saline pools to lay their eggs, these hybrid individuals would have an increased chance of backcrossing with the *C. occidentalis* lineage. While only two *C. occidentalis* x *C. sonorensis* hybrids were tested in this study, both had *C. occidentalis* mothers, providing some evidence that this is the scenario taking place in nature.

Impact on vector competency and future work

The *C . variipennis* complex is one of many vector groups in which species delimitation can be challenging (Carlson, 1980; Nolan et al., 2007; Palacios et al., 2014; Rivas, Souza, & Peixoto, 2008; Sebastiani et al., 2001); however, species identification is an integral part of vector surveillance. The species status of these group members has implications for vector surveillance, as any ambiguity in identification will lead to unreliable data. For example, while *C. albertensis* and *C. sonorensis* occur in sympatry, only *C. sonorensis* is a reported as a vector species (Wilson et al., 2009). The addition of the non-competent vector species when conducting serological surveys could lead to a severe underestimation of the infection rate within the vector species. As BTV and EHDV are expanding northward into eastern Canada (Allen et al., 2019), it has been suggested that the dispersal of *C. sonorensis* to new areas could be to blame for this incursion (Jewiss-Gaines et al., 2017). Specimens assigned to *C. sonorensis* by Jewiss-Gaines et al. (2017) were included in the present study and cluster instead with *C. albertensis* ("ON", Fig. 1). Thus, there are likely alternative reasons for this expansion, including an unidentified vector species outside of the *C. variipennis* complex. Accurate species-level delimitation within this complex is sorely needed for proper vector surveillance. Additionally, elucidating the evolutionary history of these groups can lead to a better understanding of how some species become highly competent vectors while closely related taxa are not. The detection of hybridization within a vector species may be evidence of recent speciation, but it also highlights a potential path of introgression for genes controlling vector competency (Ciota, Chin, & Kramer, 2013; Mallet, 2005).

Conclusion

Our study shows that using a population genomic approach to studying sibling species can identify both species-level divergence as well as fine-scale genetic structuring. Tracing the level of gene flow within and between these species enables the detection of geographic isolation, hybridization, and cryptic species to offer a more accurate depiction of the current species dynamics. Radiation within the *C. variipennis*complex occurred despite the long-range dispersal capabilities of biting midges as well as hybridization between sympatric species. Because of this, we believe that behavioral and ecological isolation may have shaped evolution within this group or is at least maintaining the existing species boundaries. Significant geographic isolation was only found between populations of *C. occidentalis* , but more work is needed to determine if the lack of gene flow between California and the other populations represents an incipient speciation event. Delimiting the species in the *C . variipennis* complex will not only aid in vector surveillance efforts, but continued study of the speciation of closely related vector and non-vector species could produce valuable evolutionary insights into vector competency.

Acknowledgments

The authors thank Art Borkent, Adam Jewess-Gaines, Dustin Swanson, Bonnie Ryan, Nadja Mayerle, multiple vector control agencies, USDA-APHIS-Wildlife Services, and several landowners for access to property or assisting with collecting specimens used in this study. Funding was provided by a USDA Non-Assistance Cooperative Agreement: 58-3020-9-007 and the Urban Entomology Endowment fund at Texas A&M University.

Data Accessibility

The data reported in this study will be deposited in the Open Science Framework database upon acceptance, https://osf.io (DOI XXX). Mitochondrial sequences obtained in the current study have been deposited under Genbank Accession Numbers XXX-XXX.

Author Contributions

PS, MM, LC, and EV planned and designed the study, PS, MH, and LC collected samples, PS, PE, and AB carried out lab work and analyzed the data, PS and PE produced the figures, PS led writing; MH, PE, AB, MM, LC, and EV contributed to drafting and editing the manuscript, MM, LC, and EV provided supervision, PS, LC, and EV contributed to procuring funds.

ORCID

Shults 0000-0003-2498-9637, Hopken 0000-0003-3861-6153, Eyer 0000-0002-7801-0466, Blumenfeld 0000-0002-9019-1693, Mateos 0000-0001-5738-0145, Cohnstaedt 0000-0002-7885-081X, Vargo 0000-0002-8712-1248.

Disclaimer

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. The conclusions in this report are those of the authors and do not necessarily represent the views of the USDA. USDA is an equal opportunity provider and employer.

References

Allen, S. E., Rothenburger, J. L., Jardine, C. M., Ambagala, A., Hooper-McGrevy, K., Colucci, N., . . . Nemeth, N. M. (2019). Epizootic Hemorrhagic Disease in White-Tailed Deer, Canada. *Emerging infectious diseases, 25* (4), 832-834. doi:10.3201/eid2504.180743

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. In: Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Arntzen, J. W., Jehle, R., Bardakci, F., Burke, T., & Wallis, G. P. (2009). Asymmetric Viability of

Reciprocal-Cross Hybrids between Crested and Marbled Newts (*Triturus cristatus* and *T. marmoratus*).*Evolution, 63* (5), 1191-1202. doi:10.1111/j.1558-5646.2009.00611.x

Ballard, J. W. O., Chernoff, B., & James, A. C. (2002). Divergence of mitochondrial DNA is not corroborated by nuclear DNA, morphology, or behavior in *Drosophila simulans* . *Evolution, 56* (3), 527-545.

Bandelt, H. J., Forster, P., & Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular biology and evolution, 16* (1), 37-48. doi:10.1093/oxfordjournals.molbev.a026036

Behura, S., Sahu, S., Mohan, M., & Nair, S. (2001). Wolbachia in the Asian rice gall midge, *Orseolia oryzae* (Wood-Mason): correlation between host mitotypes and infection status. *Insect Molecular Biology, 10* (2), 163-171.

Benestan, L. M., Ferchaud, A. L., Hohenlohe, P. A., Garner, B. A., Naylor, G. J., Baums, I. B., . . . Luikart, G. (2016). Conservation genomics of natural and managed populations: building a conceptual and practical framework. *Molecular Ecology, 25* (13), 2967-2977.

Bolnick, D. I., & Otto, S. P. (2013). The magnitude of local adaptation under genotype-dependent dispersal. *Ecology and evolution, 3* (14), 4722-4735.

Borkent, A. (2004). *Biology of disease vectors* (W. C. Marquardt Ed. 2nd ed.). Burlington, MA: Elsevier Academic Press.

Byrne, K., & Nichols, R. A. (1999). *Culex pipiens* in London Underground tunnels: differentiation between surface and subterranean populations. *Heredity, 82* (1), 7-15.

Carlson, D. (1980). Identification of mosquitoes of *Anopheles gambiae* species complex A and B by analysis of cuticular components.*Science, 207* (4435), 1089-1091.

Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. (2013). How to fail at species delimitation. *Molecular Ecology, 22* (17), 4369-4383.

Chan, K. M., & Levin, S. A. (2005). Leaky prezygotic isolation and porous genomes: rapid introgression of maternally inherited DNA.*Evolution, 59* (4), 720-729.

Chapuis, M.-P., & Estoup, A. (2006). Microsatellite Null Alleles and Estimation of Population Differentiation. *Molecular biology and evolution, 24* (3), 621-631. doi:10.1093/molbev/msl191

Ciota, A. T., Chin, P. A., & Kramer, L. D. (2013). The effect of hybridization of *Culex pipiens* complex mosquitoes on transmission of West Nile virus. *Parasites & Vectors, 6* (1), 1-4.

Claramunt, S., Derryberry, E. P., Remsen Jr, J., & Brumfield, R. T. (2012). High dispersal ability inhibits speciation in a continental radiation of passerine birds. *Proceedings of the Royal Society B: Biological Sciences, 279* (1733), 1567-1574.

Covey, H., Hall, R. H., Krafsur, A., Matthews, M. L., Shults, P. T., & Brelsfoard, C. L. (2020). Cryptic *Wolbachia* (Rickettsiales: Rickettsiaceae) Detection and Prevalence in *Culicoides* (Diptera: Ceratopogonidae) Midge Populations in the United States. *Journal of Medical Entomology, 57* (4), 1262-1269. doi:10.1093/jme/tjaa003

Coyne, J. A., & Orr, H. A. (2004). Speciation. In. Sunderland, MA: Sinauer Associates, Inc.

De Queiroz, K. (2007). Species Concepts and Species Delimitation.*Systematic biology, 56* (6), 879-886. doi:10.1080/10635150701701083

Dobzhansky, T. (1982). *Genetics and the Origin of Species* : Columbia university press.

Downes, J. A. (1978). The *Culicoides variipennis* complex: a necessary re-alignment of nomenclature (Diptera: Ceratopogonidae).*The Canadian Entomologist, 110* (1), 63-69.

Ducheyne, E., De Deken, R., Becu, S., Codina, B., Nomikou, K., Mangana-Vougiaki, O., . . . Hendrickx, G. (2007). Quantifying the wind dispersal of *Culicoides* species in Greece and Bulgaria.*Geospatial Health* , 177-189.

Enbody, E. D., Pettersson, M. E., Sprehn, C. G., Palm, S., Wickstrom, H., & Andersson, L. (2021). Ecological adaptation in European eels is based on phenotypic plasticity. *Proceedings of the National Academy of Sciences, 118* (4).

Endler, J. A. (1973). Gene flow and population differentiation: studies of clines suggest that differentiation along environmental gradients may be independent of gene flow. *Science, 179* (4070), 243-250.

Ewels, P., Magnusson, M., Lundin, S., & Kaller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics, 32* (19), 3047-3048. doi:10.1093/bioinformatics/btw354

Foll, M. (2012). BayeScan v2. 1 user manual. *Ecology, 20* , 1450-1462.

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology, 3* (5), 294-299.

Fonseca, D. M., Smith, J. L., Kim, H.-C., & Mogi, M. (2009). Population genetics of the mosquito *Culex pipiens* pallens reveals sex-linked asymmetric introgression by *Culex quinquefasciatus* .*Infection, Genetics and Evolution, 9* (6), 1197-1203.

Gerry, A. C., & Mullens, B. A. (1998). Response of Male*Culicoides variipennis sonorensis* (Diptera: Ceratopogonidae) to Carbon Dioxide and Observations of Mating Behavior on and Near Cattle.*Journal of Medical Entomology, 35* (3), 239-244. doi:10.1093/jmedent/35.3.239

Gibeaux, R., Acker, R., Kitaoka, M., Georgiou, G., van Kruijsbergen, I., Ford, B., . . . Heald, R. (2018). Paternal chromosome loss and metabolic crisis contribute to hybrid inviability in *Xenopus* . *Nature, 553* , 337. doi:10.1038/nature25188

Goubert, C., Minard, G., Vieira, C., & Boulesteix, M. (2016). Population genetics of the Asian tiger mosquito *Aedes albopictus* , an invasive vector of human diseases. *Heredity, 117* (3), 125-134.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0.*Systematic biology, 59* (3), 307-321. doi:10.1093/sysbio/syq010

Hare, M. P. (2001). Prospects for nuclear gene phylogeography.*Trends in Ecology & Evolution, 16* (12), 700-706.

Harrison, R. G. (1989). Animal mitochondrial DNA as a genetic marker in population and evolutionary biology. *Trends in Ecology & Evolution, 4* (1), 6-11.

Harrison, R. G. (1990). Hybrid zones: windows on evolutionary process.*Oxford surveys in evolutionary biology, 7* , 69-128.

Hebert, P. D., Cywinska, A., Ball, S. L., & Dewaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 270* (1512), 313-321.

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation.*Molecular biology and evolution, 35* (2), 518-522.

Holbrook, F. R., Tabachnick, W. J., Schmidtmann, E. T., McKinnon, C. N., Bobian, R. J., & Grogan, W. L. (2000). Sympatry in the *Culicoides variipennis* Complex (Diptera: Ceratopogonidae): a Taxonomic Reassessment. *Journal of Medical Entomology, 37* (1), 65-76. doi:10.1603/0022-2585-37.1.65

Hopken, M. W. (2016). *Pathogen vectors at the wildlife-livestock interface: molecular approaches to elucidating Culicoides (Diptera: Ceratopogonidae) biology.* (PhD dissertation), University of Colorado, Fort Collins, CO.

Howard, D. J. (1993). Reinforcement: origin, dynamics, and fate of an evolutionary hypothesis. *Hybrid zones and the evolutionary process* , 46-69.

Jacquet, S., Garros, C., Lombaert, E., Walton, C., Restrepo, J., Allene, X., . . . Huber, K. (2015). Colonization of the Mediterranean basin by the vector biting midge species *Culicoides imicola* : an old story.*Molecular Ecology, 24* (22), 5707-5725. doi:10.1111/mec.13422

Jacquet, S., Huber, K., Pages, N., Talavera, S., Burgin, L. E., Carpenter, S., . . . Garros, C. (2016). Range expansion of the Bluetongue vector, *Culicoides imicola* , in continental France likely due to rare wind-transport events. *Scientific Reports, 6* . doi:10.1038/srep27247

Janes, J. K., Miller, J. M., Dupuis, J. R., Malenfant, R. M., Gorrell, J. C., Cullingham, C. I., & Andrew, R. L. (2017). The K= 2 conundrum.*Molecular Ecology, 26* (14), 3594-3602.

Jeffreys, H. (1998). *The theory of probability* : OUP Oxford.

Jewiss-Gaines, A., Barelli, L., & Hunter, F. F. (2017). First Records of *Culicoides sonorensis* (Diptera: Ceratopogonidae), a Known Vector of Bluetongue Virus, in Southern Ontario. *Journal of Medical Entomology, 54* (3), 757-762. doi:10.1093/jme/tjw215

Jombart, T., & Collins, C. (2015). Analysing genome-wide SNP data using adegenet 2.0. 0. In.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods, 14* (6), 587-589.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., . . . Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics, 28* (12), 1647-1649. doi:10.1093/bioinformatics/bts199

Keller, I., Wagner, C., Greuter, L., Mwaiko, S., Selz, O., Sivasundar, A., . . . Seehausen, O. (2013). Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Molecular Ecology, 22* (11), 2848-2863.

Kisel, Y., & Barraclough, T. G. (2010). Speciation has a spatial scale that depends on levels of gene flow. *The American Naturalist, 175* (3), 316-334.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms.*Molecular biology and evolution, 35* (6), 1547-1549.

Leache, A. D., Banbury, B. L., Felsenstein, J., De Oca, A. N.-M., & Stamatakis, A. (2015). Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies.*Systematic biology, 64* (6), 1032-1047.

Lehmann, T., Besansky, N., Hawley, W., Fahey, T., Kamau, L., & Collins, F. (1997). Microgeographic structure of *Anopheles gambiae* in western Kenya based on mtDNA and microsatellite loci. *Molecular Ecology, 6* (3), 243-253.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics, 25* (14), 1754-1760.

Linley, J. (1986). The effect of salinity on oviposition and egg hatching in *Culicoides variipennis sonorensis* (Diptera: Ceratopogonidae). *Journal of the American Mosquito Control Association, 2* (1), 79-82.

Lischer, H. E., & Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics, 28* (2), 298-299.

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution, 20* (5), 229-237.

Manni, M., Gomulski, L. M., Aketarawong, N., Tait, G., Scolari, F., Somboon, P., . . . Gasperi, G. (2015). Molecular markers for analyses of intraspecific genetic diversity in the Asian Tiger mosquito, *Aedes albopictus* . *Parasites & Vectors, 8* (1), 1-11.

Mayr, E. (1999). *Systematics and the origin of species, from the viewpoint of a zoologist* : Harvard University Press.

McCoy, K., Boulinier, T., Tirard, C., & Michalakis, Y. (2001). Host specificity of a generalist parasite: genetic evidence of sympatric host races in the seabird tick Ixodes uriae. *Journal of Evolutionary Biology, 14* (3), 395-405.

Mellor, P., Boorman, J., & Baylis, M. (2000). *Culicoides* biting midges: their role as arbovirus vectors. *Annual Review of Entomology, 45* (1), 307-340.

Mignotte, A., Garros, C., Dellicour, S., Jacquot, M., Gilbert, M., Gardes, L., . . . De Wavrechin, M. (2021). High dispersal capacity of*Culicoides obsoletus* (Diptera: Ceratopogonidae), vector of bluetongue and Schmallenberg viruses, revealed by landscape genetic analyses. *Parasites & Vectors, 14* (1), 1-14.

Morales-Hojas, R., Hinsley, M., Armean, I. M., Silk, R., Harrup, L. E., Gonzalez-Uriarte, A., . . . Fife, M. (2018). The genome of the biting midge *Culicoides sonorensis* and gene expression analyses of vector competence for bluetongue virus. *BMC Genomics, 19* (1), 624. doi:10.1186/s12864-018-5014-1

Nayduch, D., Lee, M. B., & Saski, C. A. (2014). The reference transcriptome of the adult female biting midge (*Culicoides sonorensis*) and differential gene expression profiling during teneral, blood, and sucrose feeding conditions. *PLoS One, 9* (5).

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution, 32* (1), 268-274.

Nolan, D. V., Carpenter, S., Barber, J., Mellor, P. S., Dallas, J. F., Mordue, A. J., & Piertney, S. B. (2007). Rapid diagnostic PCR assays for members of the *Culicoides obsoletus* and *Culicoides pulicaris* species complexes, implicated vectors of bluetongue virus in Europe. *Veterinary Microbiology, 124* (1-2), 82-94.

Nosil, P. (2008). Ernst Mayr and the integration of geographic and ecological factors in speciation. *Biological Journal of the Linnean Society, 95* (1), 26-46.

Onyango, M. G., Aitken, N. C., Jack, C., Chuah, A., Oguya, J., Djikeng, A., . . . Duchemin, J.-B. (2016). Genotyping of whole genome amplified reduced representation libraries reveals a cryptic population of*Culicoides brevitarsis* in the Northern Territory, Australia.*BMC Genomics, 17* (1), 769. doi:10.1186/s12864-016-3124-1

Onyango, M. G., Beebe, N. W., Gopurenko, D., Bellis, G., Nicholas, A., Ogugo, M., . . . Duchemin, J.-B. (2015). Assessment of population genetic structure in the arbovirus vector midge, *Culicoides brevitarsis* (Diptera: Ceratopogonidae), using multi-locus DNA microsatellites. *Veterinary Research, 46* (1), 108. doi:10.1186/s13567-015-0250-8

Onyango, M. G., Michuki, G. N., Ogugo, M., Venter, G. J., Miranda, M. A., Elissa, N., . . . Duchemin, J.-B. (2015). Delineation of the population genetic structure of *Culicoides imicola* in East and South Africa. *Parasites & Vectors, 8* (1), 660. doi:10.1186/s13071-015-1277-4

Pages, N., Munoz-Munoz, F., Verdun, M., Pujol, N., & Talavera, S. (2017). First detection of Wolbachia-infected *Culicoides*(Diptera: Ceratopogonidae) in Europe: Wolbachia and *Cardinium*infection across *Culicoides* communities revealed in Spain.*Parasites & Vectors, 10* (1), 582. doi:10.1186/s13071-017-2486-9

Palacios, G., Tesh, R. B., Savji, N., da Rosa, A. P. T., Guzman, H., Bussetti, A. V., . . . Lipkin, W. I. (2014). Characterization of the Sandfly fever Naples species complex and description of a new Karimabad species complex (genus *Phlebovirus* , family Bunyaviridae). *The Journal of general virology, 95* (Pt 2), 292.

Palumbi, S. R. (1994). Genetic divergence, reproductive isolation, and marine speciation. *Annual review of ecology and systematics, 25* (1), 547-572.

Pante, E., Abdelkrim, J., Viricel, A., Gey, D., France, S., Boisselier, M.-C., & Samadi, S. (2015). Use of RAD sequencing for delimiting species. *Heredity, 114* (5), 450-459.

Pante, E., Puillandre, N., Viricel, A., Arnaud-Haond, S., Aurelle, D., Castelin, M., . . . Valero, M. (2015). Species are hypotheses: avoid connectivity assessments based on pillars of sand. *Molecular Ecology, 24* (3), 525-544.

Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R., Moret, B. M., & Stamatakis, A. (2010). How many bootstrap replicates are necessary? *Journal of computational biology, 17* (3), 337-354.

Pilgrim, J., Siozios, S., Baylis, M., Venter, G., Garros, C., & Hurst, G. D. D. (2021). *Cardinium* symbiosis as a potential confounder of mtDNA based phylogeographic inference in *Culicoides imicola*(Diptera: Ceratopogonidae), a vector of veterinary viruses. *Parasites & Vectors, 14* (1), 100. doi:10.1186/s13071-020-04568-3

Pina-Martins, F., Silva, D. N., Fino, J., & Paulo, O. S. (2017). Structure_threader: An improved method for automation and parallelization of programs structure, fastStructure and MavericK on multicore CPU systems. *Molecular ecology resources, 17* (6), e268-e274.

Purse, B. V., Mellor, P. S., Rogers, D. J., Samuel, A. R., Mertens, P. P., & Baylis, M. (2005). Climate change and the recent emergence of bluetongue in Europe. *Nature Reviews Microbiology, 3* (2), 171-181.

Raj, A., Stephens, M., & Pritchard, J. K. (2013). Variational Inference of Population Structure in Large SNP Datasets. *bioRxiv* , 001073.

Richardson, J. L., Urban, M. C., Bolnick, D. I., & Skelly, D. K. (2014). Microgeographic adaptation and the spatial scale of evolution. *Trends in Ecology & Evolution, 29* (3), 165-176.

Rivas, G., Souza, N., & Peixoto, A. A. (2008). Analysis of the activity patterns of two sympatric sandfly siblings of the *Lutzomyia longipalpis* species complex from Brazil. *Medical and Veterinary Entomology, 22* (3), 288-290.

Rochette, N. C., Rivera-Colon, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology, 28* (21), 4737-4754.

Rousset, F. (2017). Genepop Version 4.7. 0. In.

Rundle, H. D., & Nosil, P. (2005). Ecological speciation. *Ecology Letters, 8* (3), 336-352.

Rushton, J., & Lyons, N. (2015). Economic impact of Bluetongue: a review of the effects on production. *Veterinaria italiana, 51* (4), 401-406.

Schlick-Steiner, B. C., Steiner, F. M., Seifert, B., Stauffer, C., Christian, E., & Crozier, R. H. (2010). Integrative taxonomy: a multisource approach to exploring biodiversity. *Annual Review of Entomology, 55* , 421-438.

Sebastiani, F., Meiswinkel, R., Gomulski, L., Guglielmino, C., Mellor, P., Malacrida, A., & Gasperi, G. (2001). Molecular differentiation of the Old World *Culicoides imicola* species complex (Diptera, Ceratopogonidae), inferred using random amplified polymorphic DNA markers. *Molecular Ecology, 10* (7), 1773-1786.

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., . . . Brannstrom, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics, 15* (3), 176-192.

Servedio, M. R., & Kirkpatrick, M. (1997). The effects of gene flow on reinforcement. *Evolution, 51* (6), 1764-1772. doi:doi:10.1111/j.1558-5646.1997.tb05100.x

Shults, P. (2015). *A study of the taxonomy, ecology, and systematics of Culicoides species (Diptera: Ceratopogonidae) including those associated with deer breeding facilities in southeast Texas.*Texas A&M University,

Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium populations. *Evolution, 47* (1), 264-279.

Smith, H., & Mullens, B. A. (2003). Seasonal activity, size, and parity of *Culicoides occidentalis* (Diptera: Ceratopogonidae) in a coastal southern California salt marsh. *Journal of Medical Entomology, 40* (3), 352-355. doi:10.1603/0022-2585-40.3.352

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30* (9), 1312-1313.

Tabachnick, W. J. (1996). *Culicoides vriipennis* and Bluetongue-Virus eidemiology in the United States. *Annual Review of Entomology, 41* (1), 23-43. doi:10.1146/annurev.en.41.010196.000323

Team, R. C. (2013). R: A language and environment for statistical computing.

Thompson, J. N. (1999). The evolution of species interactions.*Science, 284* (5423), 2116-2118.

Thompson, J. N., & Cunningham, B. M. (2002). Geographic structure and dynamics of coevolutionary selection. *Nature, 417* (6890), 735-738.

Thurmond, J., Goodman, J. L., Strelets, V. B., Attrill, H., Gramates, L. S., Marygold, S. J., . . . Trovisco, V. (2019). FlyBase 2.0: the next generation. *Nucleic acids research, 47* (D1), D759-D765.

Toews, D. P., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology, 21* (16), 3907-3930.

Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A., & Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic acids research, 44* (W1), W232-W235.

Truong, H. T., Ramos, A. M., Yalcin, F., de Ruiter, M., van der Poel, H. J. A., Huvenaars, K. H. J., . . . van Eijk, M. J. T. (2012). Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS One, 7* (5), e37565-e37565. doi:10.1371/journal.pone.0037565

Velten, R. K., & Mullens, B. A. (1997). Field morphological variation and laboratory hybridization of *Culicoides variipennis sonorensis*and *C . v . occidentalis* (Diptera:Ceratopogonidae) in southern California. *Journal of Medical Entomology, 34* (3), 277-284.

Wang, I. J., & Bradburd, G. S. (2014). Isolation by environment.*Molecular Ecology, 23* (23), 5649-5662.

Werren, J. H., Baldo, L., & Clark, M. E. (2008). *Wolbachia* : master manipulators of invertebrate biology. *Nature Reviews Microbiology, 6* (10), 741.

West, S. A., & Herre, E. A. (1994). The ecology of the New World fig-parasitizing wasps *Idarnes* and implications for the evolution of the fig–pollinator mutualism. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 258* (1351), 67-72.

Wilson, W. C., Mecham, J. O., Schmidtmann, E., Sanchez, C., Herrero, M., & Lager, I. (2009). Current status of bluetongue virus in the Americas.*Bluetongue* , 197-220.

Wirth, W. W., & Jones, R. H. (1957). The North American Subspecies of*Culicoides variipennis* (Diptera, Heleidae). *U. S. Dep. Agric. Tech. Bull, 1170* , 1-35.

Yin, C., Shen, G., Guo, D., Wang, S., Ma, X., Xiao, H., . . . Zhang, Y. (2016). InsectBase: a resource for insect genomes and transcriptomes.*Nucleic acids research, 44* (D1), D801-D807.

Yukilevich, R. (2012). Asymmetrical patterns of speciation uniquely support reinforcement in *Drosophila* . *Evolution, 66* (5), 1430-1446. doi:doi:10.1111/j.1558-5646.2011.01534.x

Tables

Table 1. Collection site information and numbers of individuals retained for the SNP analyses.

Country	State/Provence	Lat	Long	Collection date	Collection method	N	Abbreviation
Canada	British Columbia	49.3065	-119.6323	5/7/2019	Pupal rearing	5	BC
USA	California	39.0245	-122.8515	8/14/2018	Pupal rearing	12	CACL
USA	California	38.9811	-122.6731	8/14/2018	Pupal rearing	9	CABL
USA	California	32.5522	-117.0628	11/7/2014	Light trap	15	CASD
USA	Idaho	43.7065	-116.4236	8/19/2014	Light trap	14	ID
USA	Nevada	40.0521	-118.4681	7/29/2013	Light trap	17	NV
USA	Arizona	34.5792	-112.4258	7/21/2010	Light trap	17	AZ
USA	Utah	40.7844	-112.1090	9/10/2018	Light trap	16	UT
USA	South Dakota	43.7438	-101.9509	8/6/2018	Light trap	10	SD
USA	Colorado	40.6560	-104.9878	8/8/2019	Light trap	15	COFC
USA	Colorado	39.0546	-108.5170	7/16/2013	Light trap	7	COME
USA	Kansas	38.8793	-98.4481	9/25/2018	Pupal rearing	16	KSLI
USA	Kansas	39.2234	-96.5906	7/17/2018	Light trap	18	KSMA
USA	Texas	29.9515	-99.6010	7/29/2017	Light trap	8	TX
Canada	Ontario	43.2167	-79.9500	7/5/2013	Light trap	8	ON
USA	South Carolina	34.3080	-81.7550	7/23/2014	Light trap	16	SC
USA	Florida	30.4782	-84.6401	8/27/2018	Light trap	3	FL

Table 2. Mean pairwise F_{ST} within and between species. The between species F_{ST} values (below diagonal) were calculated using 566 SNPs and the within-species values (on diagonal) is the mean F_{ST} calculated from individual species-specific datasets (see Table S3).

Species	*C. albertensis*	*C. occidentalis*	*C. sonorensis*	*C. variipennis*
C. albertensis	0.055 *(-0.009–0.116)*	-	-	-
C. occidentalis	0.707	0.411 *(0.143 – 0.704)*	-	-
C. sonorensis	0.709	0.730	0.029 *(0.006 – 0.069)*	-
C. variipennis	0.654	0.747	0.730	0.026 *(-0.006 – 0.045)*
San Diego pop.	0.714	0.719	0.706	0.734

Table 3. Mean percent divergence (p-distance) within and between species clusters based on the COI gene (ranges listed in parentheses). Based on overall similarity, *C. occidentalis* was split into two groups (CABL; and BC-NV-UT) and *C. albertensis* ,*C. sonorensis* , and *C. variipennis* were grouped into a single clade (alb-son-var).

Clade	occ (CABL)	occ (BC-NV-UT)	San Diego pop.	alb-son-var
occ (CABL)	0.48 (0.00 – 0.73)	-	-	-
occ (BC-NV-UT)	3.99 (3.20 – 5.49)	0.86 (0.00 – 1.65)	-	-

Clade	occ (CABL)	occ (BC-NV-UT)	San Diego pop.	alb-son-var
San Diego pop.	3.01 (2.38 – 4.21)	3.66 (2.75 – 4.76)	0.25 (0.00 – 0.66)	-
alb-son-var	3.30 (2.75 – 5.12)	3.76 (3.30 – 6.04)	3.03 (2.38 – 4.21)	0.80 (0.00 – 2.74)

Figures

Figure 1. Geographic distribution and structure plots for each collection site (black squares) overlaid on the historical distribution of the species described in Wirth and Jones 1957. The fastSTRUCTURE results are for 206 individuals inferred by 3612 SNPs and assuming five populations (K=5). The vertical bars within each collection site represents an individual, with each color representing a cluster. The putative species identity of each clusters are as follows: *Culicoides occidentalis* (blue),*C. sonorensis* (teal), *C. albertensis* (yellow), *C. variipennis* (red), and an unidentified population in San Diego, CA (CASD) (green). The black bars above structure plot indicates an individual for which the COI gene was also sequenced. The individuals inferred to be hybrids are labeled h1-7.

Figure 2. (a) A 3D representation of the principal Component Analysis (PCA) of all individuals included in the study. Each color represents the cluster inferred from the structure analysis; *C. albertensis* (yellow), *C. occidentalis* (blue), *C. sonorensis* (teal), *C. variipennis* (red), and the unidentified San Diego population (green). Hybrids (h1–h7) are designated with a black circle and their inferred parental ancestry is depicted with pie graphs. The geographic locations of the two *C. occidentalis*clusters are labeled next to each grouping (see table 1 for abbreviation). **(b)**Unrooted maximum likelihood phylogenetic tree based on 199 individuals inferred from 3612 SNPs (the hybrids were removed here but are included in Fig. S3.). Clade colors represent the clusters inferred from the structure analysis; *C. albertensis* (yellow), *C. occidentalis* (blue), *C. sonorensis* (teal), *C. variipennis*(red), and the unidentified San Diego population (green). Support values written on the branches: rapid bootstrap (%) / SH-aLRT support (%) / ultrafast bootstrap support (%). For clarity, the values within each cluster are not shown.

Figure 3. For each species, an independent SNP dataset was used to calculate the best K using fastSTRUCTURE v.1.04 with the inferred clusters denoted by varying shades. The IBD (shown as pairwiseF_{ST} by log geographic distance) for each species were calculated in Genepop v.4.7.0. The individuals from San Diego, CA are not included here as they were only found in a single population.

Figure 4. Loci under selection. Individual loci from the "all-species" dataset (566 SNPs) and the species-specific datasets are plotted against their corresponding log10 values. A log10 over 1.0 is considered to have high support (95% CI) for being under selection with a log10 value over 2.0 corresponding 99% CI for being under selection. The individuals from San Diego, CA do not have a species-specific dataset as they were only found in a single population, however, they were still included in the "all species" analysis.

Figure 5. A haplotype network inferred by a median-joining method, using 285 mitochondrial (mt) DNA sequences of the *C. variipennis* complex from 27 states in the U.S. as well as British Columbia and Ontario, Canada. The size of each circle represents the frequencies of the haplotype. The 67 sequences obtained in the present study, see figure 1, are colored according the clusters assigned from the structure analysis. The four main groups of haplotypes are demarcated by ellipses (see main text).