Rapid Differentiation of Vector Species: Development of Microsatellite Markers for Population Genetics of Biting Midges and a Potential Tool for Species Identification of Culicoides Sonorensis

Phillip Shults (ptshults@tamu.edu)  
Texas A&M University College Station  
https://orcid.org/0000-0003-2498-9637

Megan Moran  
Texas A&M University

Alexander J. Blumenfeld  
Texas A&M University

Edward L. Vargo  
Texas A&M University

Lee W. Cohnstaedt  
USDA-ARS Plains Area

Pierre-Andre Eyer  
Texas A&M University

Research

Keywords: Vector Surveillance, Biting Midges, Molecular Identification, PCR, Single-tube Assay

Posted Date: December 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1178683/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background: Proper vector surveillance relies on the ability to identify species of interest accurately and efficiently, though this can be difficult in groups containing cryptic species. *Culicoides* is a genus of small biting flies responsible for the transmission of numerous pathogens to a multitude of vertebrates. Regarding pathogen transmission, the *C. variipennis* species complex is of particular interest in North America. Of the six species within this group, only *C. sonorensis* is a proven vector of bluetongue virus and epizootic hemorrhagic disease virus. Unfortunately, subtle morphological differences, cryptic species, and mitonuclear discordance make species identification in the *C. variipennis* complex challenging. Recently, a SNP analysis enabled discrimination between the species of this group; however, this demanding approach is not practical for vector surveillance.

Methods: The aim of the current study was to develop a reliable and affordable way of differentiating the species within the *C. variipennis* complex, especially *C. sonorensis*. Twenty-five putative microsatellite markers were identified using the *C. sonorensis* genome and tested for amplification within five species of the *C. variipennis* complex. Machine learning was then used to determine which markers best explain the genetic differentiation between species. This led to the development of a subset of four and seven markers which were also tested for species differentiation.

Results: A total of 21 microsatellite markers were successfully amplified in the species tested. Clustering analyses of all of these markers recover the same species-level identification as the previous SNP data. Additionally, the subset of seven markers was equally capable of accurately differentiating the members of the *C. variipennis* complex as the 21 microsatellite markers. Finally, one microsatellite marker (*C508*) was found to be species-specific, only amplifying in the vector species *C. sonorensis* among the samples tested.

Conclusions: These microsatellites provide an affordable way in which to differentiate the species of the *C. variipennis* complex and could lead to a better understanding of the species dynamics within this group. Additionally, after further testing, marker *C508* may allow for the identification of *C. sonorensis* with a single-tube assay, potentially providing a powerful new tool for vector surveillance in North America.

Introduction

With newly diverged or cryptic species, the boundaries between taxa can be blurred and often difficult to define [1, 2]. Yet, species delimitation is vitally important as it determines the biological unit on which governmental policies, control programs, evolutionary studies, and conservation efforts rely [3, 4]. This is especially true for species that pose a risk to public or animal health, such as pathogen vectors, as misidentifications will result in unreliable transmission data. Morphological identification is commonly used in vector surveillance due to its wide accessibility and cost-effectiveness, though it can require a considerable amount of expertise if the target species closely resembles a sibling species or if it exhibits...
extensive morphological variation [5]. In many cases, sequencing a common barcoding region (i.e., COI) can be done with far less training while providing a tangible level of taxonomic identification [6]. However, barcoding is neither easily implemented nor cost-effective for use in vector surveillance programs that process hundreds, if not thousands, of specimens. For a molecular marker to be feasible in these situations, species-specific amplification is needed, denoting either the presence or absence of the vector species within pools of samples [7].

*Culicoides* is a genus of small, biting midges that are responsible for the transmission of many pathogens affecting both wildlife and livestock worldwide [8, 9]. Viruses, such as bluetongue virus (BTV) and epizootic hemorrhagic disease virus (EHDV), are of particular interest as these can cause a high rate of mortality in infected animals [10, 11]. In the past two decades, *Culicoides* spp. have contributed to notable disease outbreaks in Australia [12], Europe [13, 14], and North America [15, 16], leading to significant morbidity, mortality, and economic loss in these regions [11, 17]. These outbreaks highlight the need for *Culicoides* vector surveillance and population management programs; however, these are complicated by the fact that several of the vector species belong to complexes of closely related species that are not easily distinguishable [18]. The inclusion of a non-vector cryptic species into vector surveillance data can artificially lower seroprevalence rates, overestimate species distributions, or even interfere with the detection of other vector species. The *C. imicola, C. obsoletus, C. pulcairis, and C. variipennis* complexes all play a key role in the transmission of BTV and EHDV [9, 18]; however, proper species-level identification in these groups remains challenging. Molecular tools have been developed to aid in species identification in certain groups of *Culicoides* [19, 20, 21, 22], though cryptic diversity is often noted in *Culicoides* taxa regarded as a single species [23, 24, 25]. Additionally, there have been no molecular markers developed for the identification of *C. sonorensis*, the North American vector of BTV and EHDV.

The *C. variipennis* species complex is found throughout much of North America and is comprised of at least six species (*C. albertensis, C. australis, C. occidentalis, C. sonorensis, C. variipennis*, and *C. mullensi*) [26], only one of which (*C. sonorensis*) is a proven vector [27, 28, 29, 30]. Species delimitation within the *C. variipennis* complex is particularly challenging due to subtle morphological differences between these species [31]. Species identification is further hampered by a lack of segregation between mitochondrial haplotypes of three of these species, including the vector species *C. sonorensis* (plus *C. albertensis* and *C. variipennis*) [32]. The absence of mitochondrial discrimination prevents genetic identification using the traditional COI barcode [33, 34, 35], and to further complicate the situation, *C. sonorensis* occurs in sympathy with each of the other members of this species complex [36]. Overall, the lack of clear morphological differences, the unavailability of readily applied genetic identification, and the occurrence of several species within a single location have introduced ambiguity to vector surveillance in this group. Recently, genomic analyses using a SNP dataset shed light on species delimitation in the *C. variipennis* complex and served as a useful tool for population genomic analyses [32]. However, this method is expensive and requires bioinformatic analyses, rendering it unsuitable for the rapid and affordable species identification necessary for effective vector surveillance.
Here, we first aimed to provide an easy and cost-effective way to identify species within the *C. variipennis* complex, especially the vector species *C. sonorensis*. We developed microsatellite markers from the available genome of *C. sonorensis* and tested these markers for species differentiation within the *C. variipennis* complex. These results were compared to the species delimitation obtained through SNP analyses from Shults *et al.* (2021) [32]. We then used machine-learning analyses to estimate the influence of each microsatellite marker in discriminating between the different species in the *C. variipennis* complex. This analysis was used to determine the minimum number of markers required for identification while still maintaining a high level of confidence in species differentiation. Finally, a single marker was found to uniquely amplify in the *C. sonorensis* samples tested and may prove to be a fast and inexpensive tool for discriminating between this species and the non-vector species of the *C. variipennis* complex.

**Methods**

**Microsatellite Marker Selection**

The reference genome of *C. sonorensis* (RefSeq GCA_900258525.2) [37] was assessed with the software QDD v. 3.1 [38] to determine suitable microsatellite repeat motifs. Microsatellite repeats containing less than five repetitions, as well as mononucleotide repeats, were discarded. For each microsatellite repeat, 200bp flanking regions on either side of the repeat were extracted. Overall, microsatellite repeat motifs were identified in 60,026 reads. To maximize polymorphism, loci with the highest number of repeats were selected, all of which had dinucleotide repeats. Twenty-five loci were selected, and their corresponding primers were generated using the online Primer-BLAST software through NCBI (https://www.ncbi.nlm.nih.gov/tools/primer-blast). A broad range of PCR products (110-490) were targeted to ease the development of multiplex arrangements. Primer sequences, microsatellite repeat information, and product size are displayed for each of the microsatellite markers in Table 1.

**Molecular Techniques**

A total of 79 individuals from five selected species of the *C. variipennis* complex (14 for *C. albertensis*, 15 for *C. mullensi*, 16 for *C. occidentalis*, 19 for *C. sonorensis*, and 15 for *C. variipennis*; Supplementary Information Table S1) were used for testing primer amplification. These individuals were previously assigned to species based on genomic analyses of 3609 SNP loci from Shults *et al.* (2021) and were selected to cover most of the geographic distributions of the different species (Figure 1a & 1b). A modified Gentra Puregene extraction method (Gentra Systems, Inc. Minneapolis, MN, USA) was used to extract the genomic DNA of each individual. Each of the 25 primers was amplified in standard simplex PCR conditions using a Bio-Rad thermocycler T100 (Bio-Rad, Pleasanton, CA, USA). Each PCR reaction contained 2.0 µl of DNA, 0.75 µM of a primer pair, 5.0 µl of 5x reaction buffer, 0.15 µl of Taq, and 16.35 µl of deionized water. The cycling conditions used for the amplification of microsatellite markers consisted of 95°C for 3 min, followed by 35 cycles of 95°C for 1 min, 57°C for 1.5 min, and 72°C for 2 min, with a final extension step at 72°C for 5 min. All microsatellite markers were tested at 57°C, regardless of their
species of origin. The M13-tailed primer method was used to label amplicons to facilitate multiplexing after PCR amplification. Each forward primer had an M13 tail attached, which was 5’-fluorescently labeled with 6-NED, VIC, PET, or FAM. An ABI 3500 capillary sequencer with a LIZ500 internal standard (Applied Biosystems, Foster City, CA, USA) was used to visualize PCR products. Alleles were scored using Geneious v.9.1 (Biomatters, Auckland, New Zealand) [39]. Four primers were discarded due to inconsistent or nonexistent amplification. The final primer set includes 21 microsatellite markers grouped in five different multiplexes (Table 1).

### Allelic Diversity and Summary Statistic for Each Species

For each marker, the number of alleles and allelic frequency was calculated for each species as well as the entire dataset using GENEPOP on the web [40]. This software was also used to calculate linkage disequilibrium on the entire dataset between each pair of microsatellite markers. GENEPOP on the Web was also used separately on each species dataset to calculate the expected ($H_e$) and observed ($H_o$) heterozygosity, the occurrence of a significant deviation from Hardy-Weinberg equilibrium (HWE), and the inbreeding coefficient ($F_{IS}$). $H_e$, $H_o$, deviation from HWE and $F_{IS}$ were calculated separately for each marker.

### Assessing Species Differentiation of the Microsatellite Markers

Species differentiation using all 21 microsatellite markers was first visualized by plotting individuals on a principal component analysis (PCA) using the adegenet R package [41]. A clear delimitation between species is denoted by a non-overlap of individuals from different species on the PCA. Bayesian clustering implemented in STRUCTURE v.2.3.4 [42] was used to estimate the number of genetic clusters (K) and determine whether individuals from distinct species cluster together. Simulations were run with values of K from 1 to 20 and repeated 10 times for each K-value. Each run consisted of a $5 \times 10^4$ burn-in period followed by $1 \times 10^5$ iterations of the MCMC. The most likely number of genetic clusters (K) was inferred using the method in Puechmaille (2016) [43] implemented in the web-based software StructureSelector [44]. The outputs were visualized using CLUMPAK [45]. The clustering of individuals from the different species using the microsatellite datasets was compared to the clustering of the same individuals using the SNP dataset from Shults et al. (2021) [32] (Figures 1c, Supplementary Methods, SNP dataset availability at DOI 10.17605/OSF.IO/E3Z72).

### Selecting a Subset of Markers for Optimized Species Differentiation

A random forest (RF) classification analysis was carried out on an 18 microsatellite dataset using the R package randomForest [46]. Three markers ($C424$, $C995$, and $C508$) were discarded due to the presence of missing data (e.g., non-amplifying marker in some species), which cannot be handled by an RF analysis. This analysis aims at estimating the confidence rate in determining an individual’s species of
origin for each of the microsatellite markers developed. The RF analysis was performed using 1,000 trees. The default values were used for the number of input variables randomly selected to build each node of the tree, and for the number of observations not used for building the tree (i.e., the out of bag sample - OOB). The OOB samples were used to build the confusion matrix and to estimate the OOB error rate. Low OOB error rates indicate a high ability of the variables in predicting the species of origin of the individual. In addition, RF analysis on the 18-marker dataset was used to determine the importance of each microsatellite marker in classifying the individuals in the five species. This analysis enables the selection of a subset of microsatellite markers most capable of differentiating species.

The markers determined to have the highest influence in differentiating species were grouped into two subsets (a four- and seven-marker dataset), from which a PCA and STRUCTURE analysis were subsequently applied. STRUCTURE assignments using the four- and seven-marker datasets were compared to the assignments from the entire 21 microsatellite marker dataset as well as the SNP dataset. In addition, RF analyses were re-applied on these datasets to estimate the confidence (i.e., OOB error rate) in estimating an individual’s species of origin using only four or seven microsatellite markers.

Results

The 21 selected microsatellite markers amplified in most of the species of the C. variipennis complex, with a few exceptions (Table 3.) All 21 markers were found to be polymorphic, with the number of alleles ranging from 11 to 37 (Mean ± SD = 26.4 ± 7.4; Table 2). More specifically, allelic diversity ranged from 3 to 15 (Mean ± SD = 8.6 ± 3.4) alleles per marker for C. albertensis, 4 to 12 (8.4 ± 3.2) for C. mullensi, 4 to 16 (8.0 ± 3.9) for C. occidentalis, 4 to 20 (13.0 ± 3.8) for C. sonorensis, and 4 to 14 (8.6 ± 3.7) for C. variipennis. Deviation from HWE was observed for most markers per species. This result originated from significantly positive $F_{IS}$ inbreeding coefficients observed for the majority of the markers and most species, with levels of observed heterozygosity lowered than expected (Table 3). It is important to note that the positive $F_{IS}$ values can be overestimated due to the sampling of a few individuals per species over an expansive range (i.e., the Wahlund effect). Results from the linkage disequilibrium analysis suggest that most genotypes at one locus were independent from genotypes at another locus. An exception was that markers C45, C728, and C995 appeared to be linked ($P$ = 0.004, 0.03 and 0.058), as were markers C94 and C2085 ($P$ = 0.05) (Supplementary Table S2). Note that only a single marker from each of these two groups was later used in the four- and seven-marker datasets.

The overall dataset of 21 markers was successful in the species-level differentiation of all specimens, though the clustering of individuals using a PCA revealed that two species, C. albertensis and C. variipennis, slightly overlap (Figure 2a). The clustering of individuals using a STRUCTURE analysis suggested the presence of five distinct clusters in the dataset (best K = 5; Figure 1d; individual assignments for other values of K are provided in Supplementary figure S1). This clustering using microsatellite markers corresponds to five different species, as it closely mirrors the results of the SNP dataset with the same samples from Shults et al. (2021) (Figure 1c). Importantly, individuals mostly belonged (>85% [mean = 98%] assignment probability) to a single genetic cluster when using the overall
dataset of 21 microsatellite markers (i.e., unambiguous assignment to the correct species). RF analysis on the overall dataset also suggests that markers C226, C728, C838, and C1450 had the highest influence in differentiating species, followed by markers C589, C2085, and C1241 (Figure S2). When using most of the microsatellite markers (18-marker dataset), the OOB error rate was 1.3%. The confusion matrix found that a potential low rate of misidentifications might occur with C. sonorensis and C. variipennis samples, while no misidentification occurs among samples from the three other species (Figure S3).

When the seven-marker dataset was analyzed (i.e., C226, C728, C838, C1450, C589, C2085, and C1241), almost no overlap was found between individuals from distinct species on the PCA (Figure 2b). Similarly, STRUCTURE analysis revealed confident segregation of the individuals into the different species (Figure 1e), as most individuals (N = 75) were unambiguously assigned to the correct species (> 85% [mean = 95%] assignment probability). Only four samples had assignment probabilities lower than 85% to the correct species cluster, with one sample of C. occidentalis (63%), one sample of C. albertensis (71%), and two samples of C. variipennis (80 and 83%). Additionally, the clustering closely mirrored the results from both the entire 21 microsatellite marker dataset and the SNP dataset. This finding suggests robust segregation of samples into the different species using seven microsatellite markers. The RF analysis provides further support for species delineation using these markers, revealing an OOB error rate of 1.9% (Figure S3). The confusion matrix found a misclassified sample of C. mullensi, C. sonorensis, and C. variipennis.

When plotting individuals on a PCA using the four-marker dataset (i.e., C226, C728, C838, and C1450), individuals within the same species mostly clustered together, despite small overlap (Figure 2c). Similarly, the STRUCTURE analysis revealed that individuals mostly cluster into their respective species (Figure 1f), with most individuals (N = 65) being correctly assigned (> 85% [mean = 87%] assignment probability). However, 14 individuals had a mixed assignment (< 85% assignment probability), with four of them having less than 50% assignment to their correct species, hampering full confidence in identifying species using only four markers. This finding was confirmed by an RF analysis that revealed a small, but non-negligible OOB error rate of 6.3% (Figure S3). The confusion matrix revealed multiple misclassified samples belonging to C. mullensi, C. sonorensis, and C. variipennis.

Lastly, the microsatellite locus C508 was found to amplify only in C. sonorensis (Figure 3), the only proven vector species within the C. variipennis complex. In total, 79 individuals spanning 14 geographic locations were tested at this marker: C. albertensis from four populations, C. mullensi from one population, C. occidentalis from three populations, C. sonorensis from nine populations, and C. variipennis from four populations (Figure 1a and Table S1). Many more samples and populations need to be tested to confirm this species-specific amplification; however, in the samples tested here, there does not appear to be any geographical bias in amplification. Individuals of C. albertensis, C. occidentalis, and C. variipennis collected from the same location as individuals of C. sonorensis showed no amplification at this marker. It is also important to note that this marker was not included in the RF analyses above due to the substantial amount of missing data (i.e., non-amplification in four of the sibling species) (Table 3).
Discussion

As only about 2% of the known species of Culicoides are vectors [8], differentiating these from non-vector species remains vital to surveillance efforts. Since this can be complicated by morphologically similar cryptic taxa, molecular species delimitation tools are needed. This study identified a relatively simple, reproducible, and economical tool for the molecular differentiation of the species within the C. variipennis species complex. We generated a set of 21 microsatellite markers that can assign species-level identities to the members of this complex. These markers also exhibit consistent polymorphism for each species and should lead to a better understanding of the population structure and species dynamics within this group. Machine learning was utilized to detect a set of seven microsatellite markers optimal for species differentiation, further reducing costs. Finally, the locus C508 was found to only amplify in C. sonorensis and appears to be a promising marker to improve vector surveillance for this species, though additional testing is needed.

In populations with closely related or cryptic species, approaching species delimitation at a population level can help identify independent, or mostly independent, gene pools. Shults et al. (2021) [32] provided insight into the number of biological species within the C. variipennis complex; however, SNP data is expensive to produce and cannot be easily combined with new datasets. Conversely, the microsatellite data produced here was far less expensive while achieving the same level of species delimitation as the SNPs. This will allow these new markers to be integrated into most studies of this species complex to improve accurate species identification. It is highly likely that the species distribution records and serological data within this group need to be revisited. Additionally, as morphological identification of the larvae within the C. variipennis complex is not possible, these markers will help to decipher the immature habitat of each species. Hybridization is possible between C. sonorensis and several other members of the C. variipennis complex [32, 47], but it remains to be seen if microsatellites can be used to identify hybrid individuals. Unfortunately, this study was unable to obtain specimens of the newly elevated C. australis [26]; however, if this is truly a valid species, these markers should differentiate it as well.

While these microsatellite markers will be helpful to future studies of the C. variipennis complex, their practicality in vector surveillance may be limited. However, locus C508 could be incredibly useful for the rapid identification of the vector species, C. sonorensis. If amplification of this marker is specific to this species, screening individuals or pools of individuals can be completed with a single PCR and agarose gel. Of the samples tested here, amplification was 100% for C. sonorensis, which included individuals from nine populations across its known range (Table S1). Conversely, no amplification was seen in the other members of the C. variipennis complex (Figure 3). More samples need to be tested to be confident in locus C508's ability to identify C. sonorensis in all populations. Additionally, this marker has yet to be tested on Culicoides species outside of this complex. Should this marker be cross-reactive with another species, it would produce a false positive if used as the sole method for identifying C. sonorensis. Fortunately, the C. variipennis complex is morphologically distinguishable from other species of Culicoides [26, 48], thus if this were the case, a combination of the two methods would still allow for rapid identification of the vector species. Single-tube molecular identification assays already exist for vector
species in other *Culicoides* species complexes; however, most of these are based on mitochondrial data. If similar patterns of mitonuclear discordance (as seen in the *C. variipennis* complex) exist in these groups, these assays have the potential to miss cryptic species. As microsatellite markers have already been developed for several of these groups [49, 50], it would be interesting to compare the number of species recovered between these mitochondrial and nuclear markers.

**Abbreviations**

BTV  
bluetongue virus

COI  
cytochrome oxidase subunit 1

EHDV  
epizootic hemorrhagic disease

$F_{IS}$  
inbreeding coefficient

$H_e$  
expected heterozygosity

$H_o$  
observed heterozygosity

HWE  
Hardy-Weinberg equilibrium

OOB  
out of bag

PCA  
principal component analysis

PCR  
polymerase chain reaction

RF  
random forest

USA  
United States of America.

**Declarations**

**Acknowledgements**

We would like to thank Matthew Hopken and JR Ewing for aiding in collecting specimens and shipping material.

**Funding**
Funding for this project was provided by the Texas A&M University Urban Entomology Endowment Fund and the United States Department of Agriculture, Agricultural Research Service.

**Availability of data and materials**

Not Applicable

**Authors’ contributions**

PS and PAE designed the study; ELV, LWC, and PAE provided supervision of the project; PS and LWC obtained the samples; MM and PAE performed the molecular work; PS, AJB, and PAE analyzed the results; PS, ELV, and PAE wrote the original draft of the manuscript. All authors revised the manuscript and read and approved the final draft.

**Ethics approval and consent to participate**

Not Applicable

**Consent for publication**

Not Applicable

**Competing interests**

The authors declare that they have no competing interests.

**References**


Tables

Due to technical limitations, tables 1 to 3 are only available as a download in the Supplemental Files section.

Figures

Figure 1

(a.) The sites used to collect specimens of the *C. variipennis* complex. Each location is colored (corresponding to the phylogenetic tree) to represent which species were tested from these collection sites. Corresponding to the phylogenetic tree. (b.) A phylogenetic tree and (c.) STRUCTURE plot based on SNP data from the subset of individuals sequenced in Shults et al. (2021) (best K = 5). Structure plots inferred from (d.) the 21-marker dataset (best K = 5), the seven-marker dataset (best K = 5) (e.), and the four-marker dataset (best K = 5) (f.).

Figure 2

Principal component analysis (PCA) of individuals of the *C. variipennis* complex using the (a.) 21-marker, (b.) seven-marker, and (c.) four-marker datasets. Each dot represents an individual and each color corresponds to the species assignment generated from the STRUCTURE analyses. The shaded ovals denote the confidence ellipse for the mean of each species.

Figure 3

The PCR product of marker *C508* imaged after gel electrophoresis for individuals of the *C. variipennis* complex. This marker is roughly 350 bp in length and, of the samples tested, shows amplification in only *C. sonorensis*. 
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables.xlsx
- SuppFigureS1.pdf
- SuppFigureS2.pdf
- SuppFigureS3.pdf
- graphicalabstract.pdf
- SupplementalInformation.docx