Summer 2018

Species Distribution and Conservation Genetics of the Upland and Midland Chorus Frogs (Pseudacris) in Kentucky

Tucker Cambridge
Western Kentucky University, tuckerwc91@gmail.com

SPECIES DISTRIBUTION AND CONSERVATION GENETICS OF THE UPLAND AND MIDLAND CHORUS FROGS (*PSEUDACRIS*) IN KENTUCKY

A Thesis
Presented to
The Faculty of the Department of Biology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science

By
Tucker William Cambridge

August 2018

SPECIES DISTRIBUTION AND CONSERVATION GENETICS OF THE UPLAND AND MIDLAND CHORUS FROGS (*PSEUDACRIS*) IN KENTUCKY

Date Recommended 7/21/18

Jarrett R. Johnson, Director of Thesis

T. Keith Philips

Scott Grubbs

Cheryl A. Davis 8/13/18

Dean, Graduate School Date

ACKNOWLEDGMENTS

No part of this thesis would have been possible without the support and guidance of Dr. Jarrett Johnson, who has not only served as an excellent example of a productive academic scientist, but has also taught me everything that I know about conservation genetics. Dr. Johnson's willingness to teach and instruct those passionate about conservation biology who are otherwise inexperienced in the complicated molecular and bioinformatic methods that accompany conservation genetics is a testament to his integrity as a scientist and his fortitude as an advisor, and the only reason I will be able to pursue conservation genetics as a Ph.D. student.

I would also like to thank my committee members, Dr. Keith Philips and Dr. Scott Grubbs for their guidance on my thesis as well as other milestones of my graduate career, particularly the oral exam. I am very grateful to my cohort of graduate students for their help with fieldwork as well as the countless hours of invaluable scientific discussions. I would also like to thank Dr. Bobby Gaffney and the WKU Bioinformatics Center for providing technical assistance, and Naomi Rowland and the WKU Biotechnology Center for providing support during my molecular protocols.

Additionally, I thank the Kentucky Department of Fish and Wildlife Resources (KDFWR) for providing a Scientific/Educational Collecting Permit and John MacGregor of KDFWR for providing the putative *Pseudacris* maps, *Pseudacris* calling dates for Kentucky, contact information for other biologists, and advice on how to collect samples in the field. I also would like to thank the Texas Natural History Collection of the University of Texas Austin, and Dr. Emily Moriarty Lemmon of Florida State University

for providing tissue samples essential to my analysis. Lastly, I would like to acknowledge the various funding sources that have made this research possible: The Society for the Study of Amphibians and Reptiles, the WKU Graduate School, and the WKU Center for Biodiversity Studies.

CONTENTS

SPECIES DISTRIBUTION AND CONSERVATION GENETICS OF THE UPLAND AND MIDLAND CHORUS FROGS (*PSEUDACRIS*) IN KENTUCKY

Tucker W. Cambridge August 2018 64 Pages

Directed by: Jarrett Johnson, T. Keith Philips, Scott Grubbs

Department of Biology Western Kentucky University

The upland (*Pseudacris feriarum)* and midland (*P. triseriata)* chorus frogs are closely related cryptic species that are best distinguished genetically. The distribution of these species within the Commonwealth of Kentucky has previously been defined by only a handful of genetic samples, making delineation of range limits for each species difficult. Accurate understanding of species distributions, and the genetic structure within them, are vitally important for conservation management of amphibian species. In this study, I have collected genetic samples from across the putative ranges of *P. triseriata* and *P. feriarum* in Kentucky and used next-generation sequencing technology to generate more fine-scale estimates of species ranges. The genetic data generated in this study support the delineation of two species in Kentucky, and the species assignments of all individuals and populations are in general concordance with the previously hypothesized species distributions. However, I have identified two previously unrecognized contact zones for these species and revealed areas of hybridization. By delineating species distributions and identifying potentially important regions of genetic admixture, this study will be informative to future conservation management and conservation genetic research of chorus frogs in Kentucky.

Introduction

Understanding species distributions and the patterns of genetic structure within them is both increasingly imperative for conservation biology and accessible via modern genomic sequencing technology. Accurate species distributions are important for conservation efforts because management decisions are best implemented with precise knowledge of how populations are related, lineages are evolving, and landscape features affect gene flow. Therefore, understanding both fine-scale genetic structure among populations and broad-scale evolutionary patterns among lineages is relevant to conservation concerns. Conservation geneticists have the capacity to tackle both concerns of delimiting "species boundaries" and describing genetic structure by utilizing genomic technology.

Determining species distributions and understanding genetic structure is especially important for amphibian species, because amphibians are currently undergoing a mass extinction (Allentoft and O'Brien 2010). In fact, amphibian species are declining more rapidly than either birds or mammals (Brito 2008), with 40% of the worlds amphibian species in decline, and with a contemporary extinction rate 200 times greater than the historical background rate (Allentoft and O'Brien 2010). Modern amphibians are especially prone to severe loss in genetic diversity when compared to other vertebrate groups, in part due to contemporary anthropogenic landscape modification and habitat fragmentation, and in part due to natural life history traits such as breeding strategy, high site fidelity, and low dispersal capability (Alford and Richards 1999). Habitat fragmentation creates isolated gene pools, known as dissociated populations, which are far more prone to genetic drift and a loss of genetic diversity (Allentoft and O'Brien

2010). To further complicate the issues regarding fragmentation, the majority of amphibians exist naturally in metapopulations: populations that are spatially separated, but nevertheless are interconnected by varying amounts of gene flow between one another (Blaustein et al. 1994). The negative effects of habitat fragmentation on gene flow becomes increasingly likely, and more complicated, because metapopulation persistence depends on individual movements among subpopulations across complex landscapes.

The loss of genetic diversity resulting from reduced gene flow is a key issue in conservation biology (Lande 1988; Frankham 2005). The relationship between genetic diversity and the extinction risk of populations over time is known as a genetic-fitness correlation, whereby genetic drift causes a loss of genetic diversity that is correlated with a decrease in fitness (Reed and Frankham 2003). In conservation biology, it is understood that reduced genetic diversity results in less phenotypic variation and a decreased potential for adaptive responses to changing environments and anthropogenic threats. In the case of amphibians, these threats are complex, synergistic, and numerous, including habitat degradation and modification, acidity and toxins, infectious disease, climate change, increased UV radiation, and invasive species predation (Vitt and Cadwell 2014). Thus, amphibians face a dual threat, in that their life history attributes make gene flow difficult and increase their susceptibility to the anthropogenic stressors that cause population declines directly. Therefore, determining species distributions and understanding genetic structure is essential for developing baseline information for amphibian conservation, even prior to the observation of declining populations. We cannot hope to detect dissociated populations, complicated changes to metapopulation

dynamics, and extinction risks without detailed and comprehensive population genetic studies. The purpose of this study is to assess distribution patterns with genetic data for two cryptic species, and thereby contribute to a greater understanding of evolution and distribution of the genus *Pseudacris* in Kentucky to further conservation research and inform future management decisions.

The specific scope of this study pertains to the distribution and genetic diversity of the midland chorus frog (*P. triseriata*) and the upland chorus frog (*P. feriarum*). At its core, this project is one of cryptic species delimitation. It is apparent that upland and midland chorus frogs are diagnosable genetically (Lemmon et al. 2007), but due to sparse genetic sampling and the difficulty of distinguishing between them morphologically (Powell 2016), we do not currently understand the distributions of these two taxa in Kentucky. Therefore, this work seeks to fill this knowledge gap with sampling of individuals throughout their ranges in Kentucky and state-of-the-art genomic strategies for obtaining genetic data and assigning individuals to populations. Population genetic studies such as this one are crucially important for improving our understanding of patterns of biodiversity, with information gained being used to inform conservation and management decisions, in this case, at the state level.

The upland and midland chorus frogs have only recently been delineated as two separate species (Platz 1989; Moriarty and Cannatella 2004). *Pseudacris feriarum* and *P. triseriata* were both originally recognized as subspecies of *Pseudacris nigrita* (along with *P. n. verrucosa* and *P. n. septentrionalis*; Stejneger and Barbour 1917). However, the distributions and morphological distinctions of these four subspecies were poorly defined (Smith and Smith 1952). The first attempt to distinguish *P. feriarum* and *P. triseriata*

using morphological characters employed tibia/body size ratios and noted coloration morphology to be an unreliable identifier (Smith and Smith 1952). *Pseudacris triseriata* was then the first of the *P. nigrita* subspecies to be elevated to its own species, with *P. feriarum* at the same time being recognized as a subspecies of *P. triseriata* (Schwartz 1957). *Pseudacris feriarum* was not elevated to a species until the pulse rates of mating calls were quantified as distinct from *P. triseriata* (Platz 1989). The diagnosis of *P. feriarum* as a separate species has since been confirmed with genetic evidence (Moriarty and Cannatella 2004). Still, *P. triseriata* and *P. feriarum* remain difficult to distinguish from one another based on morphology and pulse rate is not an easily used character in the field, as it varies with temperature. The striped patterns of these two species are notoriously difficult to distinguish, with individuals occurring as striped, partially striped, spotted, or lacking patterns altogether (Fig. 1; Powell 2016). The difficulty in morphological determination of these species makes these two chorus frogs ideal candidates for a thorough study of species distributions, particularly one that employs genetic investigation. In fact, recent phylogenetic studies (Lemmon et al. 2007) have called for a major revision of range limits for both of these taxa. Within Kentucky, the putative species range boundaries were developed from only six genetic samples (Lemmon et al. 2007). With such cryptic speciation occurring in this species complex, more genetic samples are essential in order to confirm or modify the current understanding of species distributions. This work is important not simply to gain a greater understanding of chorus frog phylogeography, but for the purposes of amphibian conservation. To misunderstand the species distributions and confuse one species with

another could be detrimental to management practices attempting to maintain natural patterns of gene flow.

A recent published description of the *P. triseriata* and *P. feriarum* distributions shows *P. triseriata* distributed throughout the Western Kentucky Coal Field and portions of the Mississippian Plateau in west-central Kentucky, and the range of *P. feriarum* extending eastward from the Mississippi Embayment to the east-central section of the Mississippian Plateau (Fig. 2; Lemmon et al. 2007). The two taxa are depicted as overlapping throughout the range of *P. triseriata* in Kentucky, but the degree of potential overlap is unknown and unquantified, with no genetic evidence of sympatry or hybridization (Fig. 2; Lemmon et al. 2007). Powell et al. (2016) draw the species' ranges similarly, with the exception that no overlap is depicted, partly on the basis of unpublished maps for Kentucky based on the Lemmon et al. (2007) genetic data, pattern morphology, and the reasoning that these chorus frogs are so similar ecologically that it is unlikely there would be much overlap (John MacGregor, pers. com.).

The paucity of genetic sampling in Kentucky and subsequent uncertainty regarding the geographic distribution of these two *Pseudacris* species in Kentucky and has left many essential ecological questions unanswered. What are the existing boundaries to gene flow? Does hybridization occur? How genetically diverse are populations? Only through a clear understanding of species distributions and genetic diversity can we begin to answer these questions; questions not just scientifically interesting, but relevant to future conservation efforts.

Materials and Methods

Sample collection

Chorus frogs are far more easily heard than seen in most environments, and the presence of adults is best assessed during breeding periods when males are producing advertisement calls. Chorus frogs reproduce early in the spring season in Kentucky, usually in conjunction with cool rains, and often in shallow temporary pools (Conant and Collins 1998). Therefore, tissue samples were obtained by listening for choruses of males during the spring breeding season, beginning February 21st, 2017. Tissue from adult individuals comprised either an excised toe, in which case individuals were released at the site of capture, or liver, in which case whole animal voucher specimens were retained. Tissue samples were stored in individual tubes containing 95% ethanol and deposited in a -80C freezer to await DNA extraction. Voucher specimens were anesthetized in a 0.2% benzocaine solution buffered with sodium bicarbonate, fixed in 10% formalin, and stored in 70% ethanol at -20C. Sample location GPS data were recorded with Google Maps version 4.54.8. Sample locations were often initially wildlife management areas. However, many samples were collected along the edge of roadways, after hearing chorus frog trilling. As the breeding season waned, collection of tadpoles from roadside culverts, vernal pools, and ditches was the most reliable source of genetic material. Sampling of tadpoles continued until June 18th 2017. For all larval collections, whole tadpoles were collected, typically several from each location. Tadpoles were euthanized in a similar

manner to adult chorus frogs with the exception that the liver was not removed. Tail tips and whole legs on pre-metamorphic tadpoles were instead used for DNA extraction.

Larval Identification

The process of collection itself is certainly easier for tadpoles than adult chorus frogs because the larval period is longer than the breeding season. However, the use of larval samples is difficult because clear morphological variation is often not present until later stages of ontogeny. Thus, one challenge in this study was discriminating between *P. triseriata/feriarum* and *P. crucifer*, a broadly sympatric congener. The initial strategy was to raise tadpoles to metamorphosis, at which time the species-specific patterns become apparent (Fig. 3). The patterns that distinguish *P. triseriata/feriarum* from *P. crucifer* are the typical stripes and blotches of the chorus frog species complex and a very clear "X" marking on *P. crucifer* (Powell 2016). Though this method was successful, the risk of tadpole mortality also made necessary the use of larval amphibian dichotomous key identification. While often difficult to assess, there are morphological characters useful for larval anuran identification. These characters include, spiracle morphology, body and fin shape, positioning of the eyes, and the appearance of oral apparatus (Altig and McDiarmid 2015). The oral apparatus, unique to tadpoles, is used to graze on algae and consists of beak-like keratinized jaw sheathes, various marginal and sub-marginal rows of papillae, and smaller keratinized labial tooth rows (Fig. 4; Altig and McDiarmid 2015). *Pseudacris triseriata/feriarum* and *P. crucifer* have many similar features within the oral apparatus, including a 2-3 labial tooth row configuration and overlapping sizes of the second and third lower labial tooth rows (Altig and McDiarmid 2015). The most reliable

oral apparatus structure to differentiate these species, is the mid-ventral marginal papillae, which are uniserial in *P. triseriata/feriarum*, and are biserial in *P. crucifer* (Altig and McDiarmid 2015).

From February 21st to June 18th 2017, amphibian samples were collected from 52 sites across Kentucky (Fig. 5). Of all samples collected, 97 individuals were positively identified as *P. triseriata/feriarum* from 19 sites. I selected 57 individuals from the 19 sites, with a maximum of 4 individuals per site, for sequencing. It would not be feasible to sequence all individuals, as the inclusion of too many individuals will reduce the number of sequences per sample, and thus reduce confidence in consensus sequences and depress the total number of genetic loci available for analysis. Eleven additional samples were added to the analysis from the Texas Natural History Collection of the University of Texas Austin and the research collection of Dr. Emily Moriarty Lemmon of Florida State University. These sample additions increased the final data set to 68 individuals across 30 localities (Fig. 6; Table 1).

DNA Extraction and Normalization

DNA was extracted from liver tissue, adult toe clips, metamorph legs, or tadpole tails (Table 1). DNA extraction was performed with a DNEasy Blood and Tissue Kit (Qiagen). Genomic DNA samples were checked for quality by running a total genomic gel on 2% agarose, quantified using a NanoDrop 2000 (ThermoFisher Scientific), and a Qubit fluorometer (Thermo Fisher Scientific). Final sample concentrations were normalized to $\leq$200ng/µL (Table 2)

Reduced-representation DNA library construction

The importance of understanding genetic diversity in non-model organism, such as chorus frogs, has been often eclipsed in conservation biology by the difficulty of such a task. However, comprehensive genomic studies for organisms that do not have fully sequenced genomes are increasingly becoming a reality through advances in next generation sequencing that have been developed with conservation geneticists in mind. In the past decade, next-generation sequencing methods have been some of the most significant breakthroughs in conservation biology, ecology, and evolutionary studies because of their ability to generate millions of short DNA sequences spanning genomes of many individuals in single, simple, cost effective experiments (Andrews et al. 2016). I utilized a technique known as 'restriction-site-associated DNA sequencing' (RADseq), suitable for generating a reduced genomic library for sequencing tens of thousands of distinct sections of a non-model-organism's genome for sequencing (Andrews et al. 2016). RAD sequencing is a modern technique in conservation genetics, and as such, there are many variations of this methodology. In this study, I used 3RAD (Glenn et al. 2017), which is a very recent variation of double digest restriction site associated DNA sequencing (ddRAD; Peterson et al. 2012).

Until recently, genetic studies involving the population genetics of non-model organisms relied on using small numbers of genetic markers developed specifically for the taxon of interest to score genotypes for each individual. Methods of generating genetic markers such as microsatellites or single nucleotide polymorphisms (SNPs), often required significant amounts of time and money in development before they could be utilized (Davey and Blaxter 2010). Alternatively, whole genome "shotgun" sequencing

using next-generation technology results in the generation of millions of sequences that may be useful for answering the same kinds of questions as studies using more traditional methods. However, while haphazard sequencing of whole genomes requires no prior knowledge of genomic sequences, this strategy yields much sequence data that is unwanted or unusable, and limits the number of unique individuals that can be pooled in a sequencing reaction.

RADseq is an efficient combination of the precision of specific genetic markers with the "genome-wide" power of shotgun methods. By targeting a non-random subset of the genome with restriction enzymes, RADseq provides advantage over shotgun sequencing, by creating a reduced-representation genomic library, and providing greater depth of sequence coverage per locus. Therefore generating a higher confidence in our final genotype assessments, and a greater ability to sequence many individuals simultaneously (Andrews et al. 2016). In other words, RADseq is highly beneficial for this kind of study because it does not require pre-existing genomic markers, and will allow for a more comprehensive quantitative genomic analysis.

Double-digest RADseq differs from other RADseq procedures in that it utilizes two restriction enzymes to fragment an organism's genome and generate genomic sequence fragments that are flanked on each end by a recognition site that is homologous across individuals (Peterson et al. 2012). A key advantage of this sequence-specific double fragmentation is that it allows for more precise and repeatable size selection, which greatly reduces a number of errors and biases associated with early forms of RADseq. This makes ddRAD much more forgiving to the effects of sequence "lane crowding" than earlier RADseq strategies, where the danger is pooled specimens could

be under-sampled and have suppressed read counts (Peterson et al. 2012). This is an especially important issue in conservation genetics when many individuals must be pooled on one lane of sequencing. The 3RAD method includes all of the ddRAD improvements with the extra benefit of an additional restriction enzyme. This third enzyme does not fragment the genome, but rather cuts any adapter dimers formed during the process of adapter ligation that form unwanted products (Glenn et al. 2017). This additional step, as well as several other design modifications to the methods, including quadruple-indexed libraries, greatly increases the efficacy and decreases the cost of RADseq for the purposes of conservation genetics (Glenn et al. 2017).

3RAD Protocol

Restriction Enzyme Digestion—A restriction enzyme functions by cutting the DNA only at specific sites based on each enzymes recognition sequence (Davey and Blaxter 2010). The consistent recognition sequence for each enzyme is what allows for non-random sampling across the genome, and thus enables consistent construction of genomic libraries across individuals. 3RAD depends on the sequencing of genomic fragments produced by two enzyme cut sites (generating comparatively shorter DNA fragments in the sequencing library), using two different restriction enzymes (Andrews et al. 2016).

There are many commercially available restriction enzymes, and already a number of paired enzyme designs available in the literature (Peterson et al. 2012; Glenn et al. 2017). When selecting the enzymes for use in this study, I consulted a recent successful 3RAD study on the population genomics of the foothill yellow-legged frog (*Rana boylii*; McCartney-Melstad et al. 2017). *Rana boylii* is the closest in genome size

to my study organism, as well as evolutionary relatedness among published 3RAD studies. However, I used a different rare cutter (or read 2 enzyme) in the interest of using an established protocol pair for the oligonucleotide adapter design (Glenn et al. 2017). I selected 'MspI' as the common cutter and read1 enzyme, 'BamHI' as the rare cutter and read 2 enzyme, and 'ClaI' as the adapter dimer cutter. Enzyme digestion occurs individually with each sample. At this step NEB cutsmart buffer and the custom Read 1 and Read 2 oligonucleotide adapter barcodes are also added to the genomic DNA (Glenn et al. 2017). Digestion occurs for 1 hour at 37C.

Adapter Ligation—Oligonucleotide adapters with built in barcodes (short nucleotide sequences used to identify specific samples) are ligated (or attached to the DNA via phosphodiester bonds) to the "sticky-ends" of DNA fragments produced by restriction enzyme digestion (Andrews et al. 2016). Ligation of barcoded adapters is done both to provide individual-specific identifiers at the restriction cut sites prior to pooling of samples for sequencing, and to ensure only our targeted DNA fragments are PCR amplified during library preparation (Andrews et al. 2016).

Multiplexing (or the pooling of samples) saves time and money and the 3RAD protocol allows for two different sets of unique barcode pairs (i.e., quadruple indexing, Glenn et al. 2017; Fig. 7). Quadruple indexing is very useful to save costs during sequencing because it allows multiple libraries to be pooled in one lane. The internal barcodes (at the restriction sites) function as identification for individuals within libraries and the external i5/i7 barcodes function as identification for different libraries, which could be a different organism, study system, or molecular method (Glenn et al. 2017). In

this study, there were only 68 individuals sequenced from one library, and thus quadruple indexing was unnecessary. The i5 and i7 adapters were still included, as they function as PCR primers as well as Illumina flowcell primers, but were not used in sample identification. As previously mentioned, the read 1 and read 2 adapters are added during the restriction enzyme digestion, but are not ligated until after the digestion occurs. Ligation was performed with the addition of DNA ligase, ligase buffer, and ATP, and occurred in a thermal cycler with a set incubation program (Glenn et al. 2017). After barcode ligation, all samples were pooled together.

Clean-up of small DNA fragments—Size selection converts the total genomic DNA fragment library, which consists of a wide range of fragments sizes, to a reduced-representation library of putatively homologous fragments of appropriate length for sequencing (Andrews et al. 2016). In the 3RAD protocol, two different forms of size selection were employed, and occurred at multiple steps throughout the protocol. The first type of size selection, essentially a purification process for very small, unwanted DNA fragments, utilizes paramagnetic beads, and various washes. The paramagnetic beads function by binding to larger DNA fragments, the size of fragment depending on the concentration of beads in solution. With the use of magnets, ethanol washes, and a final elution step, this protocol allows unwanted smaller DNA fragments and contamination to be removed while retaining the larger adapter-ligated products for sequencing (Rohland and Reich 2012; Fig8). For this study, I used a non-commercial Serapure bead protocol, also known as Speed Beads, which is much cheaper than name brand bead preparations (Rohland and Reich 2012). Speed Bead size selection was

performed after individual samples were pooled, and then before and after each PCR reaction.

Polymerase Chain Reaction—PCR is utilized to amplify the DNA fragments for sequencing. In the 3RAD protocol, PCR is also the step where the Illumina platform specific indices, i5 and i7 are ligated to the end of the DNA fragments. For this study, I used Adapterama dual-indices, iTru5 set 18, and iTru7 set 108, which are ligated in two separate PCR reactions (Glenn et al. 2016, 2017). Within the respective iTru5 and iTru7 adapter sets, there are many adapters to choose from, and it is vital to calculate the index diversity to ensure no nucleotide position is monomorphic (Glenn et al. 2016). PCR was performed with a Kapa HiFi PCR kit (Kapa Biosytems). A total genomic 2% agarose gel showed DNA present at the expected fragment lengths (200-1000bp; Fig. 9).

Large Fragment Size Selection and Illumina Sequencing—Large fragment size selection is performed prior to sequencing and serves to select fragments of the appropriate size for a particular sequencing protocol (in this case 150bp). This size selection was performed by Vincent J. Coates Genomics Sequencing Laboratory at University of California Berkeley on a Pippen Prep (Sage Science), which selects for a fragment range using automated gel cutting. The 3RAD protocol calls for size selection at 525 bp +/- 10. However, after this initial reaction was performed it was found this range did not generate enough product for sequencing, and the size selection range was increased to 400-650 bp, which retained sufficient DNA. Sequencing was performed on an Illumina HiSeq 4000 set for 150 bp paired end reads.

Bioinformatics

The RADseq approach results in hundreds of millions of short reads of genetic sequence data. Bioinformatics tools are then used to separate the sequence reads of pooled individuals using the barcodes, cluster homologous sequence reads into a consensus sequence within individuals, and bin the resulting homologous loci among individuals. These loci can be used in sequence form for phylogenetic applications, or distilled to SNPs (single nucleotide polymorphisms) for population genetic applications. The process of delimiting individual genomes from the pooled sequences, clustering and filtering SNPs, and assigning individuals to genetic populations involves different bioinformatics software. The protocols of each are outlined below.

Loci Filtering and Clustering

Stacks—I used Stacks (Catchen et. al 2011) to de-multiplex and filter the data. De-multiplexing to separate the data for each individual is the first bioinformatics step because I pooled samples for sequencing. The Stacks command *process_radtags* searches the raw pooled reads from the Illumina sequencing output and locates fragments belonging to particular individuals though the specified barcodes that were added during library preparation (Catchen et al. 2013). *Process_radtags* will also perform some preliminary quality filtering of the sequence reads, discarding any base pair read that is below a 90% probability of being the correct base (raw phred score of 10), as well as discarding base reads that are completely ambiguous, coded as Ns (Catchen et al. 2013). The raw sequence data were first de-multiplexed by the sequencing facility, by the

external i5/i7 indices before using *process_radtags* to locate any internal barcode adapters.

Ipyrad—Once individual genomes have been parsed, the next step is to cluster all identical sequences within individuals into consensus loci, and then to bin the homologous loci across individuals. Ipyrad can cluster loci *denovo* by aligning homologous sequence fragments across individual genomes (Eaton 2014). There are many parameters within the Ipyrad clustering analysis that may affect the outcome of locus characterization. These include loci filtering parameters, clustering thresholds, barcode mismatch error tolerance, and the number or SNPs, indels, and alleles allowed per locus (Eaton 2014). For this study, many of these parameters were kept at default levels, which are appropriate for the typical analysis of diploid organisms. The quality score threshold was left at default, at a phred of 33, which discards any base pair below approximately a 99.9% chance of being correct. Ipyrad can also adjust the amount of Ns accepted per sequence fragment, but *process_radtags* of Stacks had already removed all fragments containing Ns in this dataset.

The clustering threshold parameter defines the level of similarity at which two sequence reads are defined as homologous and are clustered together (Eaton 2014). The default value for this parameter is 85% similarity. However, I set our clustering threshold to 94% similarity, to reduce the potential for grouping of paralogous loci (genes which are not homologous but rather the product of a duplication event) into the same cluster, which will create false heterozygosity (Ilut et al. 2014). Studies have shown that higher clustering thresholds can diminish this potential bias (Ilut et al. 2014). Further, the

potential for this same bias is also increased in large genome organisms, such as many amphibians, and recent research suggests higher clustering thresholds are more appropriate for larger genomes (McCartney-Melstad et al. 2017). I determined the most optimal clustering threshold for my data by running clustering analysis at various threshold values for a reduced data set and comparing these thresholds to percent heterozygosity. The most optimal clustering threshold was determined to be 94% because it is the point where heterozygosity plateaus, meaning the clustering is splitting polymorphisms within our bins of homologous sequence fragments into separate loci (Fig. 10).

Ipyrad implements a branching workflow consisting of seven steps: 1) de-multiplexing, 2) quality filtering, 3) clustering of loci within samples, 4) estimation of heterozygosity of the loci, 5) consensus base calling, 6) clustering of loci among samples, and 7) final filtering and formatting of output files (Eaton 2014). These steps can be assimilated into a branching workflow, where multiple data sets can be assembled under different parameter settings. After the chosen threshold of 94%, the final data set was run from steps 1-7 with all samples. After it was determined 7 of the 68 samples needed to be removed due to too few homologous loci present (Table 2), workflow was branched at step 6 to cluster loci for the final data set among the 61 remaining individuals. At the end of the iPyrad analyses, output data files comprising DNA sequence data and SNPs for all retained loci were obtained.

Population Assignment Tests

Several different software programs were used to assign individuals to genetic clusters. Using assignment tests, I evaluated the hypothesis that there are two species represented by the sampled individuals. Each of the methods described below implements a different clustering algorithm to determine the number of genetic populations (K) that are present in the data. For a given study system, potential values for K might represent *a priori* hypotheses based on known factors of population ecology, or the potential values of K may be unknown. For this study, each assignment algorithm evaluated K values of 1 through 5.

Structure—STRUCTURE (Pritchard et al. 2000) implements a model-based Bayesian clustering method, which assigns individuals to populations (probabilistically) based on their genotypes, while simultaneously estimating allele frequencies assuming Hardy-Weinberg equilibrium. This is implemented by a Markov chain Monte Carlo (MCMC) method, which randomly assigns individuals to the given number of populations (K) and then assess the likelihood of those assignments given the data (Pritchard et al. 2000). For the STRUCTURE analysis, the "best" or most likely K value was determined using Structure Harvester (Earl and vonHoldt 2012) which utilizes a "Delta K" method to estimate the value of K with the greatest likelihood (Evanno et al. 2005). For this study, I ran STRUCTURE with 1,000,000 MCMC iterations and then replicated ten times for each value of K.

fastStructure—The predominant advantage of fastStructure over STRUCTURE is the alleviation of some of the computational challenges for large data sets. By implementing more efficient algorithms utilizing optimization theory, fastStructure produces the same likelihood estimates for difference values of K as STRUCTURE but more efficiently and faster (Raj et al. 2014). These algorithms approximate the log-marginal likelihood of the data by creating a family of tractable parametric posterior distributions over hidden model variables, and then finds the optimal approximation of the marginal likelihood of the data (Raj et al. 2014). Furthermore, fastStructure implements a new set of heuristic scores used to identify the potential value of K, which chooses K based on the model components that are essential to explain patterns underlying the observed data (Raj et al. 2014). For this study, I ran fastStructure with 1,000,000 MCMC iterations.

MavericK—The modeling framework utilized in MavericK is identical to the design of Structure and fastStructure (Verity and Nichols 2016). Therefore, the main benefit of implementing MavericK is its unique system for estimating K within a dataset. MavericK performs estimations of K utilizing a process known as thermodynamic integration (Verity and Nichols 2016). The key difference between this process and other processes for estimating K is that thermodynamic integration is not a heuristic, meaning it is not an approximation. Rather, thermodynamic integration uses the output of closely related MCMC chains in the data to obtain direct evidence for the estimation of K (Verity and Nichols 2016). This process has been shown to generate higher accuracy and precision when compared to heuristic estimations in simulated datasets (Verity and Nichols 2016). I ran MavericK with 5,000 MCMC iterations and 5 replicates per value of K.

ConStruct—ConStruct was designed to specifically address the problem of analyzing discrete populations which occur in the presence of continuous patterns of genetic differentiation (Bradburd et al. 2017). When sampling is incomplete or discontinuous, population assignment methods may lead us to believe our genetic populations are defined by discrete variables, such as geographic, ecological, or reproductive barriers, when in reality they are affected by the underlying continuous process of genetic isolation by distance (Bradburd et al. 2017). Construct is able to account for isolation by distance in a spatial model of population assignment by estimating a rate at which relatedness decays with distance between each population layer (Bradburd et al. 2017). The ConsStruct software also offers a non-spatial model, as well as an independent assessment of an optimal K value known as cross-validation. Cross-validation generates a random subset of 90% of the loci to estimate the posterior distribution of the model parameters, and then calculates the log-likelihood of the remaining loci averaged over the posterior (Bradburd et al. 2017). ConStruct then takes the log-likelihood and averages it over the independent data partitions of the model to generate values of predictive accuracy (Bradburd et al. 2017). I ran Construct with 10,000 MCMC iterations.

Phylogenetic tree construction and DAPC

A phylogenetic tree was constructed to assess evolutionary relatedness and genetic divergence among individuals. Randomized A(x)elerated Maximum Likelihood (RAxML) is a phylogenetic program used for the inference of large phylogenies utilizing maximum likelihood (Stamatakis 2006). The most recent version of RAxML, has been

updated to accommodate the characteristically large data sets as generated by next-generation sequencing molecular protocols (Stamatakis 2014). The RAxML phylogenetic trees for this study were built using 100 iterations of the rapid bootstrap model, and the GTRGAMMA nucleic acid model (Stamatakis et al. 2008). Phylogenetic trees were constructed using concatenated full sequence data (not SNPs).

A Discriminant Analysis of Principle Components (DAPC) was performed to asses genetic clustering of individuals, using the R package 'adegenet' (Jombart and Ahmed 2011). A DAPC differs from a principle component analysis (PCA) in that it optimizes analysis of variance between groups, instead of total variance as a traditional PCA, thereby maximizing the ability to identify genetic clusters (Jombart and Ahmed 2011).

Results

Loci Generation

A total of 945,592,522 paired end sequence reads were obtained across 68 individuals, with 582,377,514 reads (62%) retained after initial demultiplexing and quality filtering in *Process_radtags*. Individual samples retained between 54,457 and 22,216,383 reads each, with a mean of 8,564,375 and a standard deviation of 5,295,414. From the retained sequences, ipyrad recovered a total of 326,898 loci. After seven individuals were removed from the analysis for low loci coverage, and loci that were shared by fewer than 70.5% of the remaining individuals were removed, the number of loci was reduced to 1,265 in the final data set.

Number of Genetic Populations

The number of genetic populations (K), was determined separately for each implementation of population assignment tests. Structure Harvester results indicate the best support for K=2 (Fig. 11). Similarly, fastStructure heuristic estimations indicated that maximum marginal likelihood for model complexity was K=2. The thermodynamic integration estimate from MavericK also showed the highest posterior probability support for K=2 (Fig. 12). Lastly, the cross validation results of ConStruct indicate that predicative accuracy shows the highest increase in a spatial model when K=2, meaning that K is two when accounting for isolation by distance (Fig. 13).

Population Assignment tests

The results of four separate implementations of population assignment tests, STRUCTURE (Fig. 14), fastStructure (Fig. 15), MavericK (Fig. 16), and ConStruct (Fig. 17) are all displayed visually for K=2 by probability assignment bar graphs. The raw probability assignments, rounded to the nearest whole percentage, are also displayed in Table 3 for each individual by population for STRUCTURE, fastStructure and MavericK, and by population for ConStruct (Table 4). Assignment probabilities were very consistent across all analyses. No individual assignment probability differed in majority assignment over the four analyses (Tables 3, 4). Additionally, pie charts depicting the assignments probabilities averaged by location superimposed over a map of Kentucky are shown for each of the four population assignments tests (Figs. 18-21).

RAxML and DAPC

The results of the RAxML maximum likelihood rapid bootstrap phylogenetic tree are displayed in Figure 22. Each branch is labeled with its location and majority probability assignment by the population assignment tests of this study. The major split between *P.triseriata* and *P.feriarum* has 99% bootstrap support. The *P.feriarum* clade is further split into an eastern and western group with 86% and 100% bootstrap support, respectively. This split can be traced to an almost exact eastern and western grouping, with the exception of one individual from the western site 8 being assigned to the linage comprising otherwise eastern sites (Fig. 18-21; 22). The DAPC separated individuals across three clusters (Fig. 23), which match the three main RAxML lineages.

Species Range Boundaries

The population assignment results of this study suggest that the distribution of *P. feriarum* is more extensive than previously known, and many populations previously believed to be *P. triseriata* may be *P. feriarum*. In particular, the eastern most split between the two species, was previously thought to occur in Monroe County, but our data show *P. feriarum* extending as farther west into Allen County (Site 2), with some presence of admixture with *P. triseriata* in Warren County. Similarly, previous estimates of the westernmost split between the two species confined *P. feriarum* to counties west of the Cumberland River (i.e., Lake Barkley), but my data show that the range of *P. feriarum* extends eastward onto the Land Between the Lakes National Recreation Area (LBL) in Trigg and Lyon county (sites 10 and 13) and across the Cumberland River into Trigg County (site 14). Further, I have detected areas of admixture with *P. triseriata* in Livingston County (Site 20) and Caldwell County (Site 23; Fig. 18-21). This indicates a potential contact zone for the two species spanning though Livingston, Caldwell, Lyon and Trigg Counties. Other portions of the distribution of these species in Kentucky have also been confirmed, namely the previous northeastern distribution of *P. triseriata* (Sites 3, 12 15, 19, 21, 22, 27, 28, and 29) and western distribution of *P. feriarum* (Sites 4, 5, 6, 11, and 25).

Discussion

Overall the distributions of *P. triseriata,* and *P. feriarum* in Kentucky as delineated by the genetic data in this study are consistent (though not identical) with the putative distributions based on earlier work. The slight shifts in the distribution revealed by these new genetic data suggest that the geographic ranges of chorus frogs in Kentucky are not accurately delineated by major physiographic regions. Therefore, we have much to learn regarding the influence of geographic boundaries on gene flow and patterns of speciation for these species. Additionally, this study has identified areas of genomic admixture, which suggests the presence of contemporary or historical contact zones that have not been recognized previously. These hybrid zones indicate that chorus frog populations located along the distributional margins of each species in Kentucky will be difficult to assign to either chorus frog species.

The currently understood distribution of chorus frogs in Kentucky includes broad western and east-central zones of parapatry. The eastern parapatric distributions extend from Allen County to Hart County, with assumed parapatry of species distributions in Marion, Nelson, and Washington counties. Unfortunately sampling of this study was not able to include samples in these areas in order to verify the presence or absence of any contact zone in the northern area of this central species gap. Further sampling is required in this area, perhaps consisting of more than a single breeding season. I have found evidence of a previously undescribed contact zone in the southern portion of this parapatric distribution in Allen and Warren Counties. All four specimens from Site 2 in Allen County were found to be admixed (Table 3), as were specimens from one site in

Warren County (Site 18). These results indicate either that hybridization is contemporary or that historic admixture is persisting despite the lack of ongoing gene flow. However, further study, with greater individual genetic sampling in the area, is needed to confirm any contemporary hybridization. Additionally, greater sampling along transects perpendicular to the contact zone would benefit the understanding of the extent of hybridization.

In the western region of parapatry, the presence of admixed populations on the east side of the Cumberland River (i.e., Lake Barkley), as well as on Land Between the Lakes (LBL), is an important finding of this study. This contact zone was previously not known to exist, as the species were thought to be separated by the Cumberland River. This study has shown gene flow occurs between *P. feriarum* and *P. triseriata* to the east of the Cumberland River in Kentucky. This is the first study to evaluate population distributions for chorus frogs at this scale in Kentucky, but other studies have evaluated the geographic features that define the range-wide distributional patterns of chorus frogs at larger scales. Lemmon et al. 2007 have suggested that four major river systems significantly reduce gene flow in *P. feriarum* by showing divergent genetic distance in populations opposite each side of rivers. I have shown gene flow across river boundaries as evidenced by admixed populations on either side across Lake Barkley (Table 3).

It is possible these results do not imply current gene flow, but rather historic admixture of these species that has not yet been diminished over evolutionary time. Furthermore, as LBL is not a true island, gene flow from the southwestern portion of Kentucky may permeate through northern Tennessee and up into LBL as well as the counties east of the Cumberland River. Nevertheless, these results, when compared to the

continental–scale studies of *Pseudacris* gene flow, highlight the importance of fine-scale studies in evaluating gene flow and population structure. Whether the genetic admixture in the western zone of parapatry is the result of ongoing or historical hybridization, it appears that the Tennessee River (i.e., Kentucky Lake) and the Cumberland River (i.e., Lake Barkley) are not barriers to gene flow.

This is the first study to identify admixture and potential hybridization in *P. feriarum* and *P. triseriata*. However, hybridization within the genus *Pseudacris* has been documented previously (Lemmon et al. 2007). Interestingly, *P. feriarum* and *P. triseriata*, have both been shown to hybridize with other chorus frogs which display larger differences in reproductive behaviors (Lemmon et al. 2007). This suggests an increase likelihood for contemporary hybridization in *P. feriarum* and *P. triseriata*, as their respective mating calls differ only in pulse rate, not in frequency (Platz 1989). Future studies of the contact zones revealed by this study are needed to answer important questions: What is the nature of the contact zones, for example, does interspecific competition affect distributions? What is the nature of hybridization, if it does occur? Previous studies in *Pseudacris* have found *P. feriarum* experiences a 44% reduction in fitness when hybridized with *P. nigrita* (Lemmon and Lemmon 2010). If hybridization does occur in the study species, further fitness estimates relating specifically to the effects of hybridization would be interesting.

Through the use of four independent population assignments tests and their accompanying assessments of the most optimal number of populations, I have concluded this dataset shows clearly two genetic populations, or species. This is not an unexpected result, as the premise of this study was not to identify alternate subspecies or new species

altogether, but rather to make clearer the distributions of the two species in Kentucky. However, the maximum-likelihood tree (Fig. 22) and the DAPC (Fig. 23) indicate three lineages or three genetic clusters, respectively. Although the population assignment tests did reveal some minimal support for three populations, all four tests showed greater support for two populations.

The third population comprises the eastern *P. feriarum* group, genetically distinct from *P. triseriata* as well as the western *P. feriarum* group. However, the results of the phylogenetic tree and DAPC must be put into the correct context. As this study was conducted with Kentucky management agencies in mind, the sampling of the true species ranges, which are ecologically uninfluenced by state boundaries, is certainly incomplete. Though conservation management occurs at the state level, population boundaries do not. Specifically, there is likely abundant gene flow through Tennessee connecting the eastern and western regions of *P. feriarum* in Kentucky.

The bias in our sampling is accounted for by ConStruct, which implements a correction for genetic isolation by distance. The ConStruct results show that when isolation by distance is accounted for during population assignment, the support for K=2 increases greatly, where the support for K=3 changes minimally (Fig. 13). Therefore, from the ConStruct results I conclude that the western and eastern *P.feriarum* groups are not more genetically different than one would expect them to be given the large geographic distance between them. The two *P. feriarum* groups in Kentucky are certainly distinct. However, it is very likely that the sampling scheme of this study, which did not include samples from Tennessee has exaggerated the extent of the differentiation. More sampling of *P. feriarum* populations in the areas between the eastern and western regions

of Kentucky is needed to better evaluate the evidence for a third genetic population of chorus frogs in Kentucky.

Though the eastern and western regions of the *P. feriarum* distribution in Kentucky perhaps do not represent isolated genetic populations, they do demonstrate extensive genetic differentiation. Therefore, conservation management initiatives would be ill advised to mix populations of *P. feriarum* from great distances across the state. In fact, it may be best to treat populations of chorus frogs in the state of Kentucky as being comprised of three separate management units, *P. triseriata* of central and northeast Kentucky, and a separate eastern and western group of *P. feriarum*. Future results with much denser sampling would be needed to fully justify these designations, and it is clear by the results of this study that at the species level, there are at present two genetic populations.

This study will be primarily useful to future research in three regards. First of all, researchers seeking to begin conservation genetics studies on the genus *Pseudacris* and closely related genera can be confident in the use of loci generated *de novo* for this genus with the 3RAD method, and may find useful insights in the methods of this study, particularly choice of restriction enzyme pairs, and locus clustering thresholds. Secondly, researchers wishing to make clear the distributions of closely related cryptic species may see the results of this study as exemplifying the potential implications of unknown distributions and how geographic assumptions can be wrong. In other words, this study has shown that in a cryptic system such as this one, only genetic evidence is sufficient to accurately define populations and our resulting understanding of their gene flow. Finally, this study will most be useful to future conservation effort of chorus frogs within the state

of Kentucky. Future research investigating genetic diversity of these species could be well improved by more specific and concentrated sampling with more individual samples per population. Furthermore, now that contact zones have been identified, future research can prioritize more efficient sampling by focusing on these areas.

Future conservation genetic goals for chorus frogs in Kentucky could be the identification of evolutionary significant units (ESUs), genetic groupings that represent important units of diversity for the management and preservation of a species (Moritz 1994). As ESUs are not discrete species but rather groupings within them, a more fine scale comprehensive study within, and perhaps outside of Kentucky would be necessary for the chorus frogs of Kentucky. For example, to make clear the potential genetic distance this study has indicated between eastern and western populations of *P. feriarum*, sampling though Tennessee would be highly desirable to accurately quantify gene flow. The failure of this study to sample locations outside of Kentucky is an important lesson for conservation management. Though management often occurs at the state level, collaboration between state management agencies can be important.

Though the chorus frog is not presently in great ecological peril, the need for conservation investigation is no less important. *Pseudacris* species are certainly not invulnerable to population declines. Adult recruitment studies have suggested naturally low levels of adult survivorship (Smith 1987). *Pseudacris* may also be vulnerable to abiotic effects. While this has not been determined for our study species, studies have shown *Pseudacris crucifer* had decreased survival rates when exposed to increased UV-B radiation, and increased exposure to copper (Baud and Beck 2005). These abiotic effects are certainly possible concerns for the *Pseudacris* populations of Kentucky, with

increased UV-B radiation from habitat modification, and potential aquatic contamination of copper or other heavy metals from atmospheric deposition from coal fired power plants (Baud and Beck 2005). Lastly, climate change is an enduring issue and cause for conservation investigation for all amphibian populations. Amphibians are ectothermic, and as such have a more narrow range of temperature fluctuations they can tolerate compared to other vertebrates (Vitt and Cadwell 2014). While some vertebrate species may be able to mitigate the effects of climate change with migration, amphibian species are at a disadvantage due low dispersal capability (Vitt and Cadwell 2014). Due to these potential threats and numerous others, effective conservation must always be concerned with distribution and diversity estimates, regardless of the current status. Loss of amphibian diversity is a persistent and pervasive global issue, and as such we must be as equally vigilant in our monitoring of imperiled species, as well as the species in our own ecological backyard.

Conclusions

The results of this study have shown the utility of loci generated *de novo* to delineate genetic structure at a species level across a large geographic distance. This simultaneously exemplifies the effectiveness of recent methods in conservation genetics, as well as the insights that can be gained when genetic evidence is applied to an otherwise cryptic species system. The distribution of chorus frog species of Kentucky has now been defined more accurately, with several key differences between previous assumed distributions. These new distribution data offer insights into potential boundaries and corridors of gene flow, contact zones between the two species, as well as setting important groundwork into future conservation genetics studies of these species within the state, such as proper designation of future management units. Lastly, future studies of the *Pseudacris* genus may view this study as potentially useful as a test of conservation genetic molecular methods, and similarly cryptic species systems may find this study a useful justification for the clairvoyance of genetic evidence in cases of cryptic speciation.

References

Agencourt AMPure XP (2016) Agencourt AMPure Xp Information For Use Guide PCR Purification

Alford RA, Richards and SJ (1999) Global amphibian declines: A problem in applied ecology. Annu Rev Ecol Syst 30:133–165

Allentoft M, O'Brien J (2010) Global amphibian declines, loss of genetic diversity and fitness: A Review. Diversity 2:47–71

Altig R, McDiarmid RW (2015) Handbook of larval amphibians of the United States and Canada. Cornell University Press, Ithaca, NY

Andrews KR, Good JM, Miller MR, et al (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet 17:81–92

Baud DR, Beck ML (2005) Interactive effects of UV-B and copper on spring peeper tadpoles (*Pseudacris crucifer*). Southeast Nat 4:15–22

Blaustein AR, Wake DB, Sousa WP (1994) Amphibian declines: Judging stability, persistence, and susceptibility of populations to local and global extinctions. Conserv Biol 8:60–71

Bradburd G, Coop G, Ralph P (2017) Inferring continuous and discrete population genetic structure across space. bioRxiv preprint doi: http://dx.doi.org/10.1101/189688

Brito D (2008) Amphibian conservation: Are we on the right track? Biol Conserv 141:2912–2917

Catchen J, Hohenlohe PA, Bassham S, et al (2013) Stacks: An analysis tool set for population genomics. Mol Ecol 22:3124–3140

Conant R, Collins JT (1998) A field guide to reptiles and amphibians: Eastern and central North America. Houghton Mifflin, Boston

Davey JW, Blaxter ML (2010) RADSeq: Next-generation population genetics. Brief Funct Genomics 9:416–423

Devi LM, Thabah CM, Hooroo RNK, Dey S (2016) Morphological and microstructural changes of the oral apparatus in two Anuran tadpoles, in regard to pH. Micron 82:41–51

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359–361

Eaton DAR (2014) PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30:1844–1849

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

Frankham R (2005) Genetics and extinction. Biol Conserv 126:131–140

Glenn TC, Bayona-Vasquez NJ, Kieran TJ, et al (2017) Adapterama III: Quadruple-indexed, triple-enzyme RADseq libraries for about $1USD per Sample (3RAD). bioRxiv preprint doi: https://doi.org/10.1101/205799

Glenn TC, Nilsen R, Kieran TJ, et al (2016) Adapterama I: Universal stubs and primers for thousands of dual-indexed Illumina libraries (iTru & iNext). bioRxiv preprint doi: https://doi.org/10.1101/049114

Ilut DC, Nydam ML, Hare MP (2014) Defining Loci in Restriction-Based Reduced Representation Genomic Data from Nonmodel Species: Sources of Bias and Diagnostics for Optimal Clustering. BioMed Res Int 2014:1–9

Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27:3070–3071

Lande R (1988) Genetics and demography in biological conservation. Science 241:1455–1460

Lemmon EM, Lemmon AR (2010) Reinforcement in chorus frogs: Lifetime fitness estimates including intrinsic natural selection and sexual selection against hybrids. Evolution 64:1748–1761

Lemmon EM, Lemmon AR, Collins JT, et al (2007) Phylogeny-based delimitation of species boundaries and contact zones in the trilling chorus frogs (*Pseudacris*). Mol Phylogenet Evol 44:1068–1082

McCartney-Melstad E, Gidiş M, Shaffer HB (2017) Population Genomics of the Foothill Yellow-Legged Frog (*Rana boylii*) and RADseq Parameter Choice for Large-Genome Organisms. bioRxiv preprint doi: http://dx.doi.org/10.1101/186635

Moriarty EC, Cannatella DC (2004) Phylogenetic relationships of the North American chorus frogs (*Pseudacris*: Hylidae). Mol Phylogenet Evol 30:409–420

Moritz C (1994) Defining 'Evolutionarily Significant Units' for conservation. Trends Ecol Evol 9:373–375

Peterson BK, Weber JN, Kay EH, et al (2012) Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PLOS ONE 7:e37135

Platz JE (1989) Speciation within the chorus frog *Pseudacris triseriata*: Morphometric and mating call analyses of the boreal and western subspecies. Copeia 1989:704

Powell R, Conant R, Collins JT (2016) Peterson field guide to reptiles and amphibians of eastern and central North America, 4th ed. Houghton Mifflin Harcourt

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: Variational inference of population structure in large snp data sets. Genetics 197:573–589

Reed DH, Frankham R (2003) Correlation between fitness and genetic diversity. Conserv Biol 17:230–237

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res 22:939–946

Schwartz A (1957) Chorus frogs (*Pseudacris nigrita* LeConte) in South Carolina

Smith DC (1987) Adult recruitment in chorus frogs: effects of size and date at metamorphosis. Ecology 68:344–350

Smith PW, Smith DM (1952) The relationship of the chorus frogs, *Pseudacris nigrita feriarum* and *Pseudacris n. triseriata*. Am Midl Nat 48:165–180

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinforma Oxf Engl 22:2688–2690

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

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol 57:758–771

Stejneger L, Barbour T (1917) A check list of North American amphibians and reptiles. Camb Harv Univ Press 126

Verity R, Nichols RA (2016) Estimating the number of subpopulations (k) in structured populations. Genetics 203:1827–1839

Vitt LJ, Cadwell JP (2014) Herpetology: An introductory biology of amphibians and reptiles, 4th edn. Academic Press

APPENDIX A: TABLES

Table 1. Sample site localities and the type of sample collected. 'X' denotes sites not included in final analyses. 'JRJ' are samples collected for this study, 'ECM' are samples donated from the laboratory of Emily Lemmon at Florida State University, and 'TNHC' are samples loaned from the Texas Natural History Collection at the University of Texas, Austin.

Site	County	Latitude	Longitude	Sample	Sample ID
1	Butler	37.1954	-86.6494	Toe Clip	JRJ9233, 9234, 9237, 9238
2	Allen	36.8934	-86.1345	Liver	JRJ9239, 9240, 9241, 9242
3	Ohio	37.5502	-86.7126	Liver	JRJ9249, 9250
4	Adair	37.2399	-85.1896	Liver	JRJ9251
5	Whitley	36.7027	-83.9546	Liver	JRJ9252
6	Whitley	36.7454	-84.0204	Liver	JRJ9253
7	Graves	36.6611	-88.4890	Legs	JRJ9254, 9255, 9256, 9257
8	Graves	36.5054	-88.4917	Legs	JRJ9259, 9260, 9261, 9262
9	Carlisle	36.9097	-89.0336	Legs	JRJ9264, 9265, 9266, 9267
10	Trigg	36.8922	-88.0390	Legs	JRJ9268, 9269
11	Lincoln	37.4864	-84.5132	Legs	JRJ9270, 9271, 9272, 9273
12	Henderson	37.8877	-87.4665	Legs	JRJ9284, 9285, 9286, 9287
13	Lyon	37.0136	-88.0923	Legs	JRJ9290, 9291
14	Lyon	36.9729	-87.9094	Legs	JRJ9313, 9314, 9315, 9316
15	Webster	37.5058	-87.6893	Tadpole	JRJ9292, 9293, 9294, 9295
16	Warren	36.8712	-86.5342	Liver	JRJ9188, 9198
17	Warren	37.0330	-86.4898	Liver	JRJ9199, 9200, 9201
18	Warren	36.9538	-86.3864	Liver	JRJ9223, 9224, 9225
19	Owen	38.5236	-85.0186	Liver	ECM16636
20	Livingston	37.2689	-88.2520	Liver	ECM16650
21	Christian	37.1196	-87.6536	Liver	ECM16652
22	Henderson	37.8365	-87.3763	Liver	ECM16659
23	Caldwell	37.1650	-87.8655	Liver	ECM4170
24	Graves	36.8290	-88.5381	Liver	ECM4472
25	Whitley	36.6010	-84.0476	Liver	ECM5240
26	McCracken	37.1340	-88.8143	Liver	TNHC63393
27	Breckinridge	38.0193	-86.5089	Liver	TNHC65812
28	Daviess	37.8661	-87.0285	Liver	TNHC63392
29	Jefferson	38.0947	-85.8454	Liver	TNHC63394
x	Christian	37.1041	-87.6514	Tadpole	JRJ9299, 9300, 9302, 9303

Table 2. Individual sample data. Sample ID (denoting species, site# and individual), county locality, total genomic mass sequenced per sample, total base pairs recovered per sample post sequencing, and total loci recovered post bioinformatics filtering and clustering. Asterisks indicate samples that were removed from final filtering for too few loci.

Sample ID	County	Input DNA (ng/ul)	Raw sequence reads (bp)	Loci recovered	Loci retained after filtering
T1.1	Butler	2.9	5658296	30104	1089
T1.2	Butler	114.0	8411308	26765	980
T1.3	Butler	34.5	1221200	21867	963
T1.4	Butler	35.8	2187167	29060	1099
F2.1	Allen	71.2	3783885	28235	1076
F2.2	Allen	113.0	7519136	25617	921
F2.3	Allen	75.4	3199078	28273	1082
F2.4	Allen	89.3	1955961	24736	985
T3.1	Ohio	123.0	2126078	21992	1073
*	Ohio	46.9	26962	8	-
F4.1	Adair	86.2	3528050	24036	1004
F5.1	Whitley	113.0	2174947	23849	974
F6.1	Whitley	105.0	3728321	21372	920
F7.1	Graves	47.6	2758297	25626	1094
F7.2	Graves	47.0	3529236	28280	1125
F7.3	Graves	78.4	4911012	24238	1034
F7.4	Graves	104.0	6349043	26010	1074
F8.1	Graves	69.9	2214163	25166	1061
F8.2	Graves	35.3	3134752	27665	1112
F8.3	Graves	38.6	7942661	27281	988
F8.4	Graves	136.0	7918562	23029	996
F9.1	Carlisle	72.9	6456472	26786	1099
F9.2	Carlisle	38.7	3611114	26276	1113
F9.3	Carlisle	32.4	3718458	28625	1123
F9.4	Carlisle	216.0	7882675	22195	959
F10.1	Trigg	93.0	7963682	22386	965
F10.2	Trigg	197.0	7499911	21756	961
F11.1	Lincoln	88.4	3274314	25475	1001
F11.2	Lincoln	50.8	2668936	26652	1046
F11.3	Lincoln	61.1	8524710	27364	1007
F11.4	Lincoln	97.8	8953751	26802	996
T12.1	Henderson	45.0	5385100	30494	1145
T12.2	Henderson	98.6	5233313	27255	1093

T12.3	Henderson	104.0	6639889	24662	1060
T12.4	Henderson	123.0	7119756	24294	1045
F13.1	Lyon	79.0	3987074	30074	1123
F13.2	Lyon	201.0	3670574	22138	1004
F14.1	Lyon	63.2	5380896	29255	1099
F14.2	Lyon	121.0	4024155	28440	1128
F14.3	Lyon	85.6	3825537	29693	1156
F14.4	Lyon	123.0	3581262	28532	1119
T15.1	Webster	143.0	2941329	21600	1012
T15.2	Webster	112.0	1821893	21148	998
T15.3	Webster	112.0	1557472	20096	940
*	Webster	75.7	189761	379	-
*	Christian	49.8	538182	1197	-
*	Christian	39.3	926583	1398	-
*	Christian	57.3	644755	1393	-
*	Christian	154.0	1196967	1373	-
*	Christian	0.6	183098	223	-
T16.1	Warren	37.9	11095314	25149	892
T17.1	Warren	97.0	3941808	28401	1110
T17.2	Warren	110.0	5687000	24578	1029
T17.3	Warren	126.0	6132737	27342	994
T18.1	Warren	127.0	6158540	25307	980
T18.2	Warren	183.0	9756097	28360	1007
T18.3	Warren	124.0	6925249	28046	1063
T19.1	Owen	121.0	5890513	21845	991
F20.1	Livingston	188.0	805197	9248	616
T21.1	Christian	102.0	3691577	23507	939
T22.1	Henderson	206.0	3515352	21723	959
T23.1	Caldwell	138.0	6932815	25705	1019
F24.1	Graves	138.0	6506070	26616	953
F25.1	Whitley	91.4	1122825	13274	610
F26.1	McCracken	70.6	3476207	24861	991
T27.1	Breckinridge	87.3	1017270	11856	503
T28.1	Daviess	82.4	2685991	26388	1057
T29.1	Jefferson	40.8	1811810	20813	947

Table 3. Individual population assignment probabilities by site and population assignment software. 'T' indicates percent assignment to *P. triseriata,* and 'F' indicates percent assignment to *P. feriarum.*

Site	Structure		FastStructure		MavericK	
	T	F	T	F	T	F
1	100	0	100	0	98	2
	100	0	100	0	98	2
	100	0	100	0	100	0
	100	0	100	0	100	0
2	25	75	20	80	26	74
	37	63	36	64	45	55
	20	80	18	82	20	80
	27	73	27	73	28	72
3	100	0	100	0	99	1
4	1	99	0	100	4	96
5	0	100	0	100	0	100
6	0	100	0	100	2	98
7	0	100	0	100	0	100
	0	100	0	100	3	97
	0	100	0	100	8	92
	0	100	0	100	3	97
8	0	100	0	100	0	100
	0	100	0	100	0	100
	28	72	24	76	29	72
	0	100	0	100	4	96
9	1	99	0	100	0	100
	0	100	0	100	0	100
	0	100	0	100	4	96
	0	100	0	100	0	100
10	15	85	12	88	24	76
	12	88	6	94	17	83
11	1	100	0	100	4	96
	0	100	0	100	0	100
	0	100	0	100	3	97
	0	100	0	100	2	98
12	100	0	100	0	100	0
	99	1	100	0	100	0
	100	0	100	0	97	3
	100	0	100	0	99	1
13	8	92	3	97	6	94
	7	94	2	98	3	97
14	28	72	23	77	18	82
	34	66	30	70	33	67

	26	74	20	80	24	76
	28	72	24	76	27	73
15	100	0	100	0	99	1
	100	0	100	0	100	0
	100	0	100	0	100	0
16	93	7	100	0	94	7
17	94	6	100	0	97	3
	99	1	100	0	99	1
	100	0	100	0	97	3
18	70	30	77	23	86	14
	74	26	80	20	81	19
	79	21	85	15	94	6
19	100	0	100	0	100	0
20	42	58	42	58	43	57
21	100	0	100	0	100	0
22	98	2	100	0	100	0
23	90	10	95	5	90	10
24	0	100	0	100	3	97
25	0	100	0	100	0	100
26	1	99	0	100	0	100
27	100	0	100	0	100	0
28	100	0	100	0	100	0
29	100	0	100	0	100	0

Table 4. Population assignment probabilities by site from ConStruct. 'T' indicates assignment to *P. triseriata,* and 'F' indicates assignment to *P. feriarum.*

Site	ConStruct	
	T	F
1	100	0
2	40	60
3	99	1
4	1	99
5	4	96
6	0	100
7	12	88
8	14	86
9	12	88
10	28	72
11	3	97
12	100	0
13	27	73
14	42	58
15	100	0
16	100	0
17	100	0
18	65	35
19	99	1
20	50	50
21	100	0
22	100	0
23	92	8
24	12	88
25	0	100
26	18	82
27	80	20
28	100	0
29	100	0

APPENDIX B: FIGURES

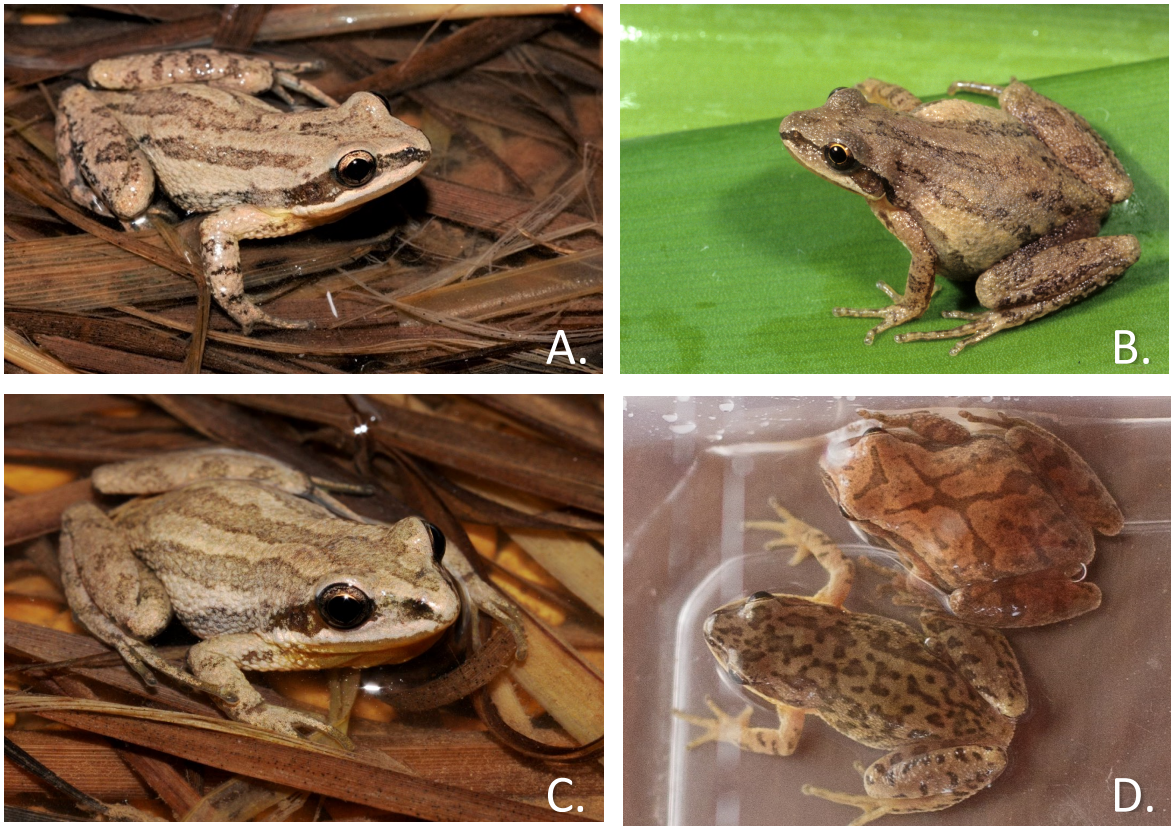

Figure 1. Morphological variation in adult *Pseudacris*. (A) *P. triseriata* from Breckenridge Co., KY, (B) *P. feriarum* from Lincoln Co., KY, (C) *P. triseriata* from Owen Co., KY, and (D) *P. crucifer* above and unusual mottled *P. triseriata* below from Butler Co., KY. Photo Credits: John MacGregor (A-C), and Jarrett Johnson (D)

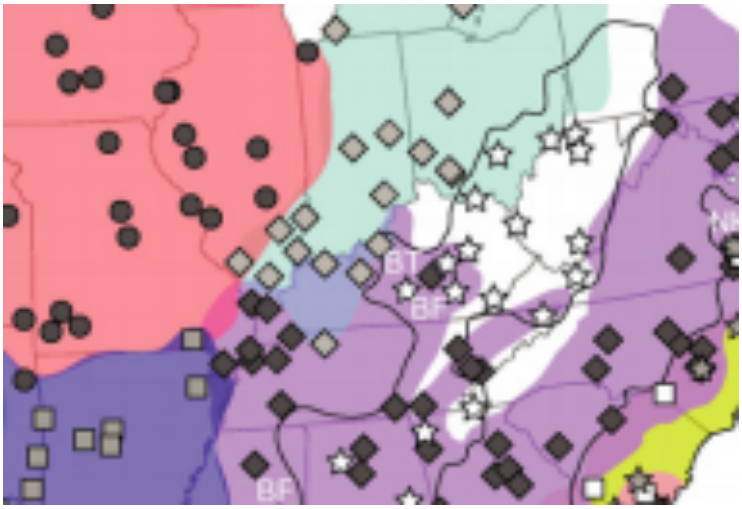

Figure 2. Current range limits of *Pseudacris feriarum* (black diamonds) and *P. triseriata* (gray diamonds) in Kentucky (modified from Lemmon et al. 2007).

Figure 3. Chorus frog tadpole in the process of metamorphosis at a point of development where definitive positive assignment to the genus *P. triseriata/feriarum* is possible, as indicated by striped pattern on dorsum.

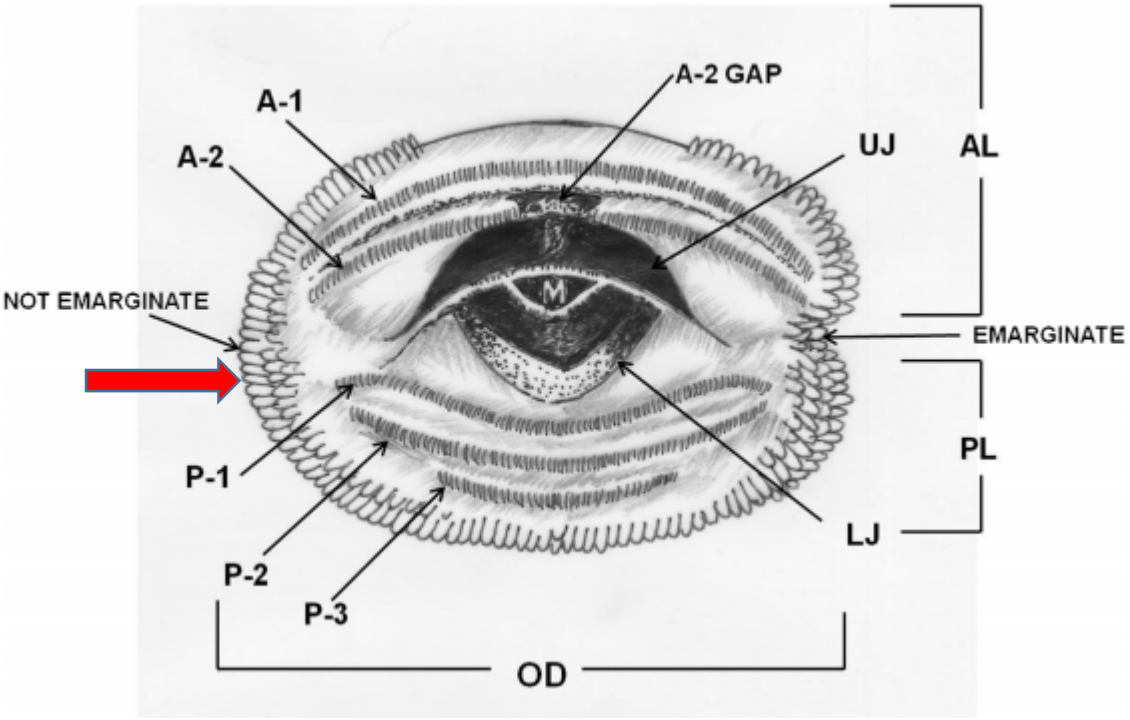

Figure 4. Larval anuran oral apparatus structures. P-1, P-2, P-3 refers to the lower labial tooth rows. The red arrow refers to biserial mid-ventral marginal papillae. Image depicted is in the 2/3 labial tooth row formation (modified from Devi et al. 2016).

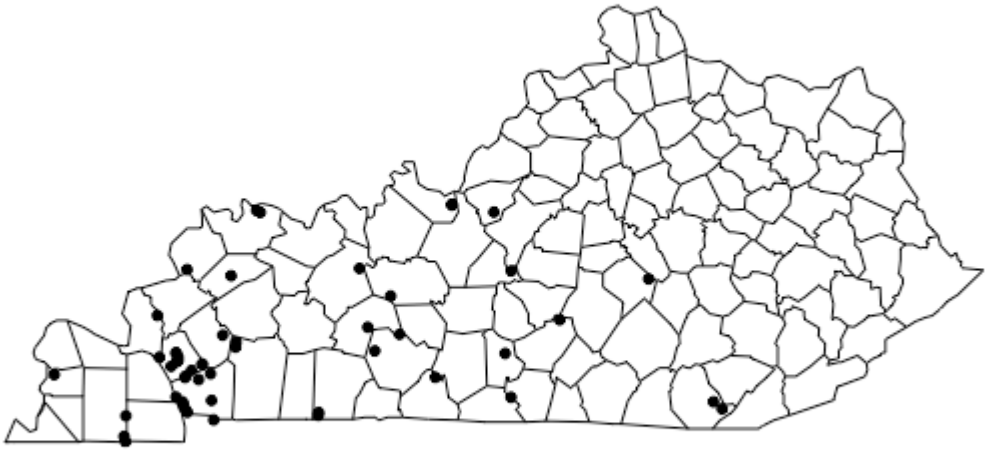

Figure 5. Total sites sampled from 21 February to 18 June 2017 for this study. Not shown: sites where samples were collected by other researchers.

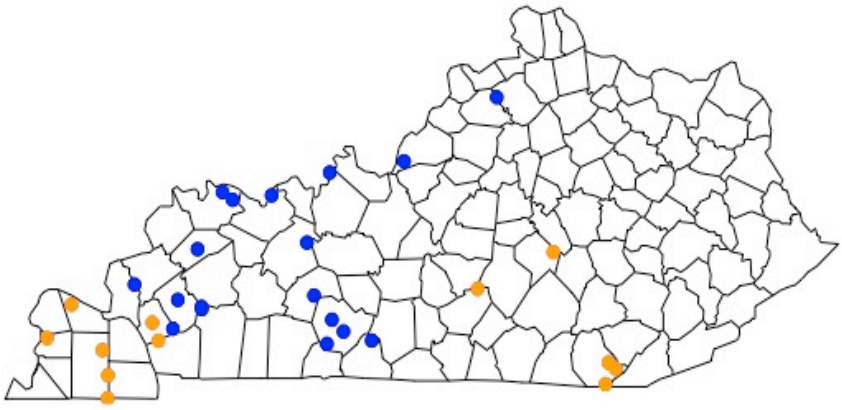

Figure 6. Thirty sites from which chorus frog specimens were included in the genomic libraries for sequencing. The yellow sites indicate upland chorus frog, *Pseudacris feriarum.* The blue sites indicates midland chorus frog, *P. triseriata.* Maximum individual samples per site=4. This map includes samples loaned from other sources.

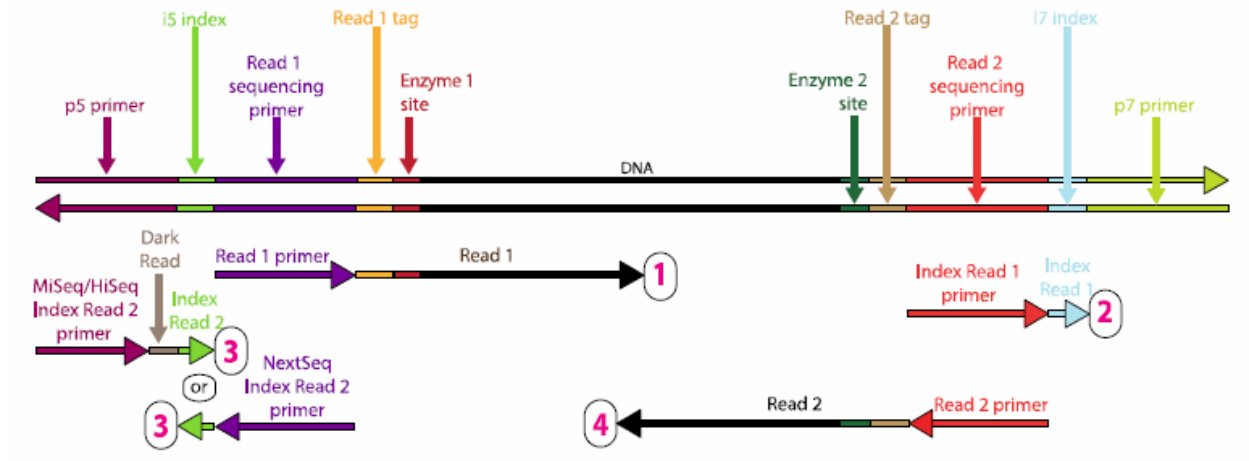

Figure 7. Composition of genomic DNA fragments resulting from 3RAD library preparation (from Glenn et al. 2017). Each genomic DNA fragment (upper) is flanked by two different enzyme recognition sites (cut sites). The quadruple-index design comprises the internal adapters (read 1 and read 2 tags), which were ligated to the cut sites along with the sequencing primer binding sites (read 1 and read 2 sequencing primer) and the outer barcode sequences (i5 and i7 indices). The p7 and p5 primer sites are added via small cycle PCR reactions.

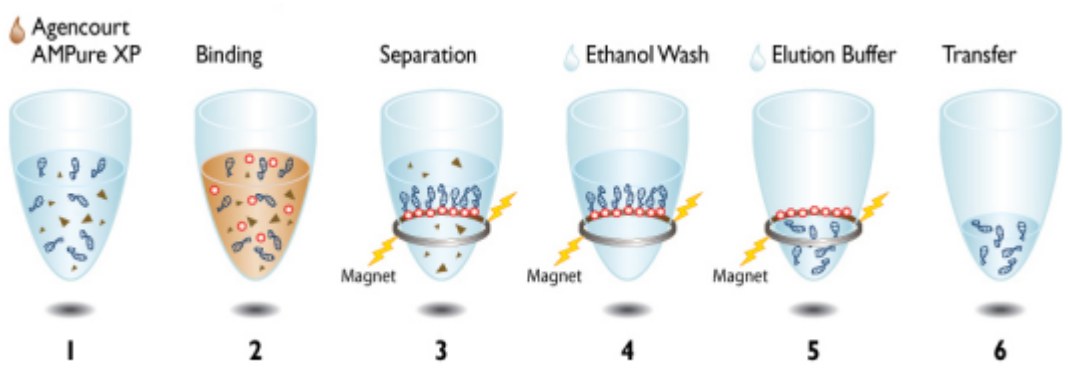

Figure 8. Steps in paramagnetic speed bead small fragment size selection and purification process (modified from Agencourt AMPure XP 2016).

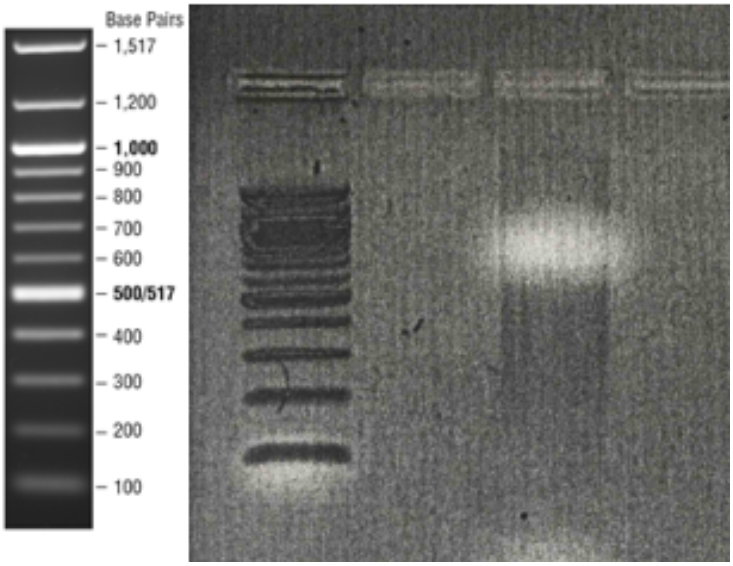

Figure 9. Total genomic gel after sucessful 3RAD library preparation before sequencing and large fragment size selection. Right-most smear depicts the genomic library, left bands indicate DNA ladder.

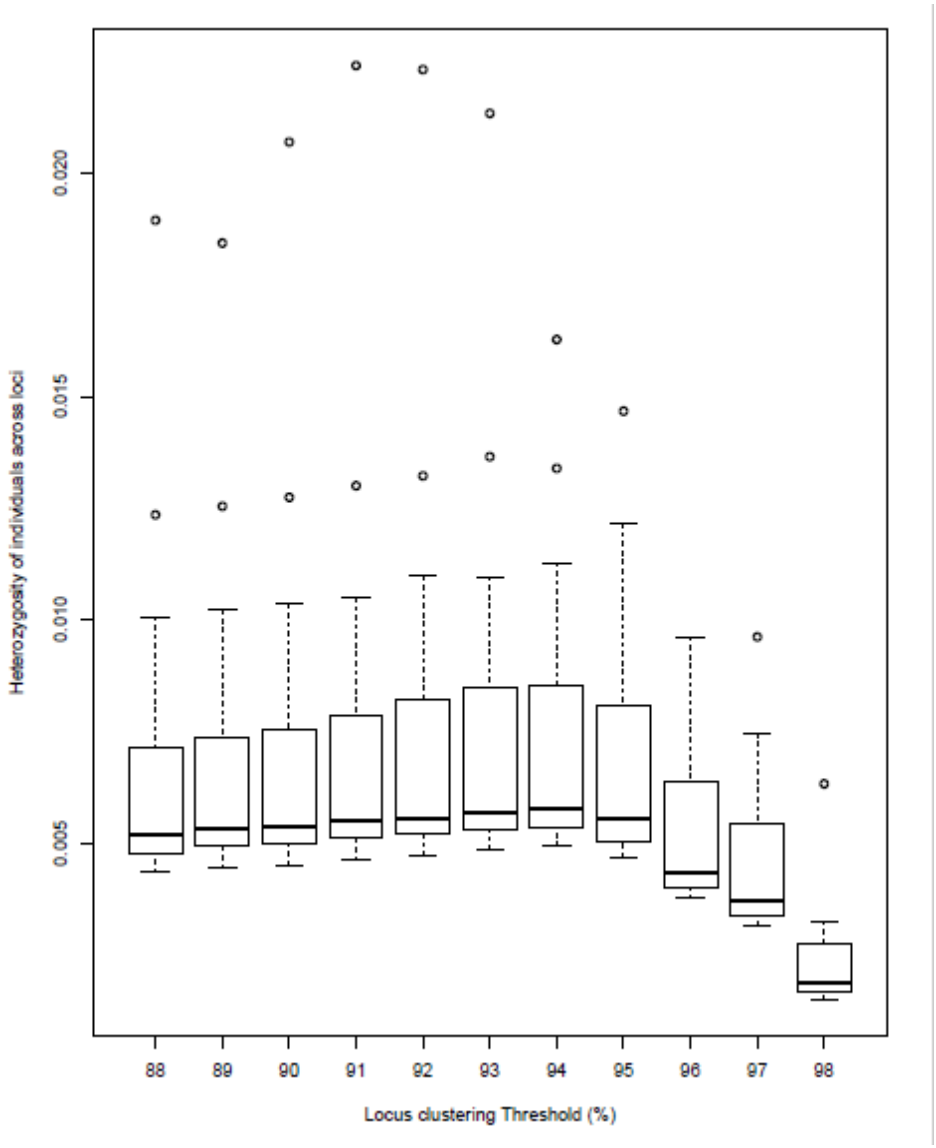

Figure 10. Heterozygosity of individuals across loci by locus clustering threshold.

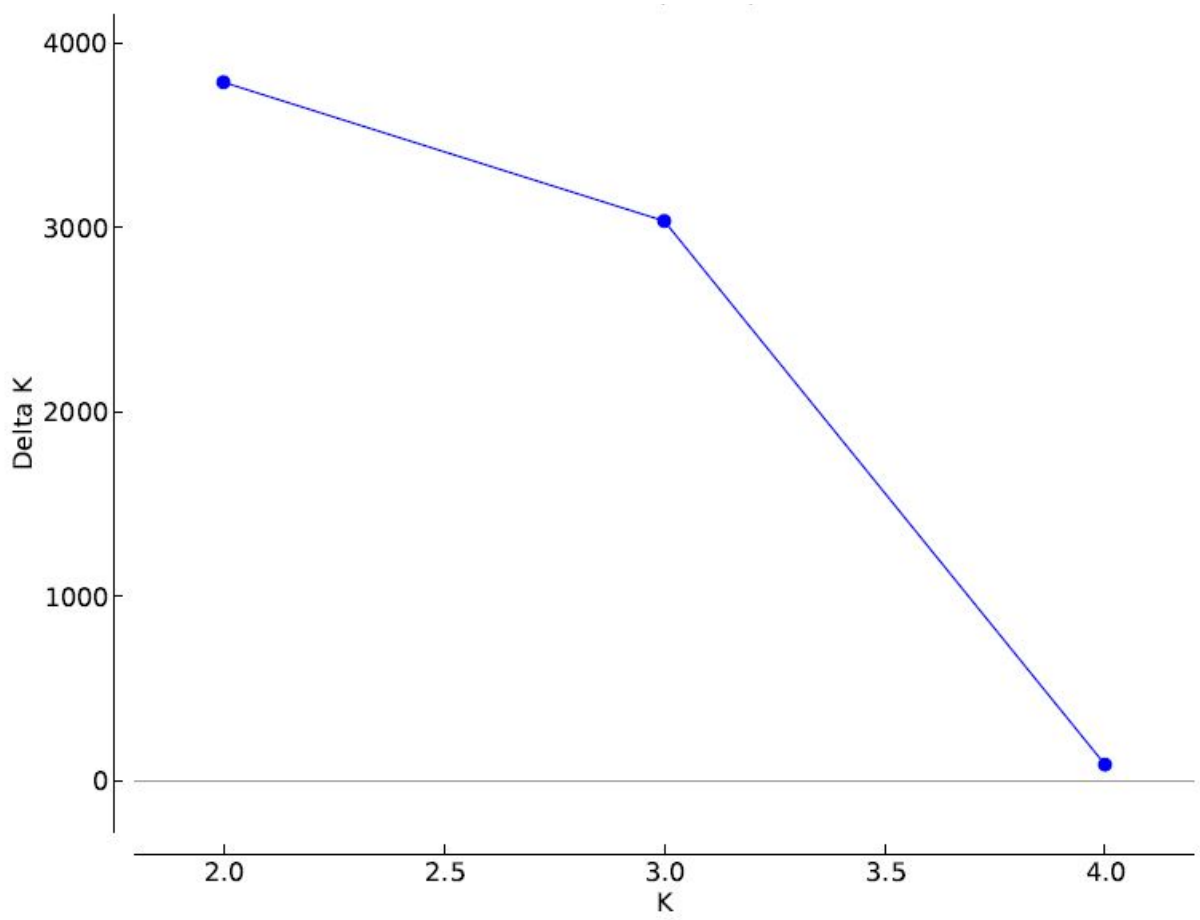

Figure 11. Delta K values across values of K as generated by the Structure Harvester analysis.

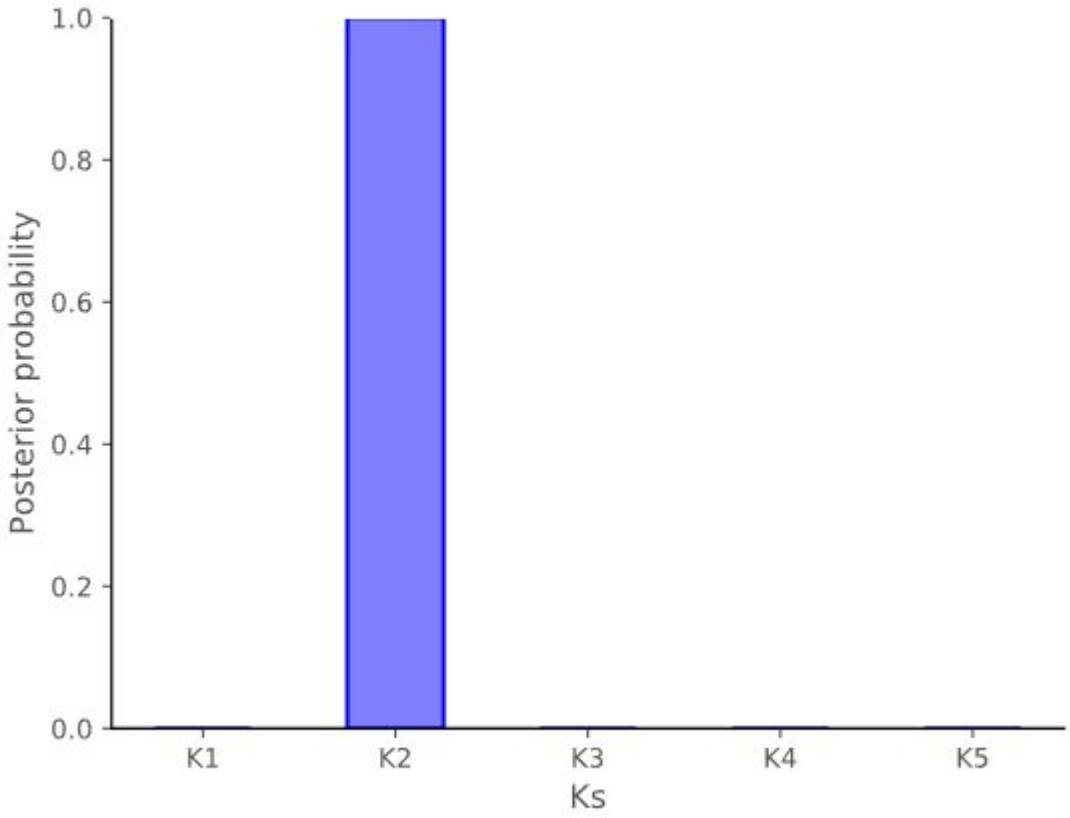

Figure 12. Posterior Probability values across values of K from the Thermodynamic Integration analysis of MavericK

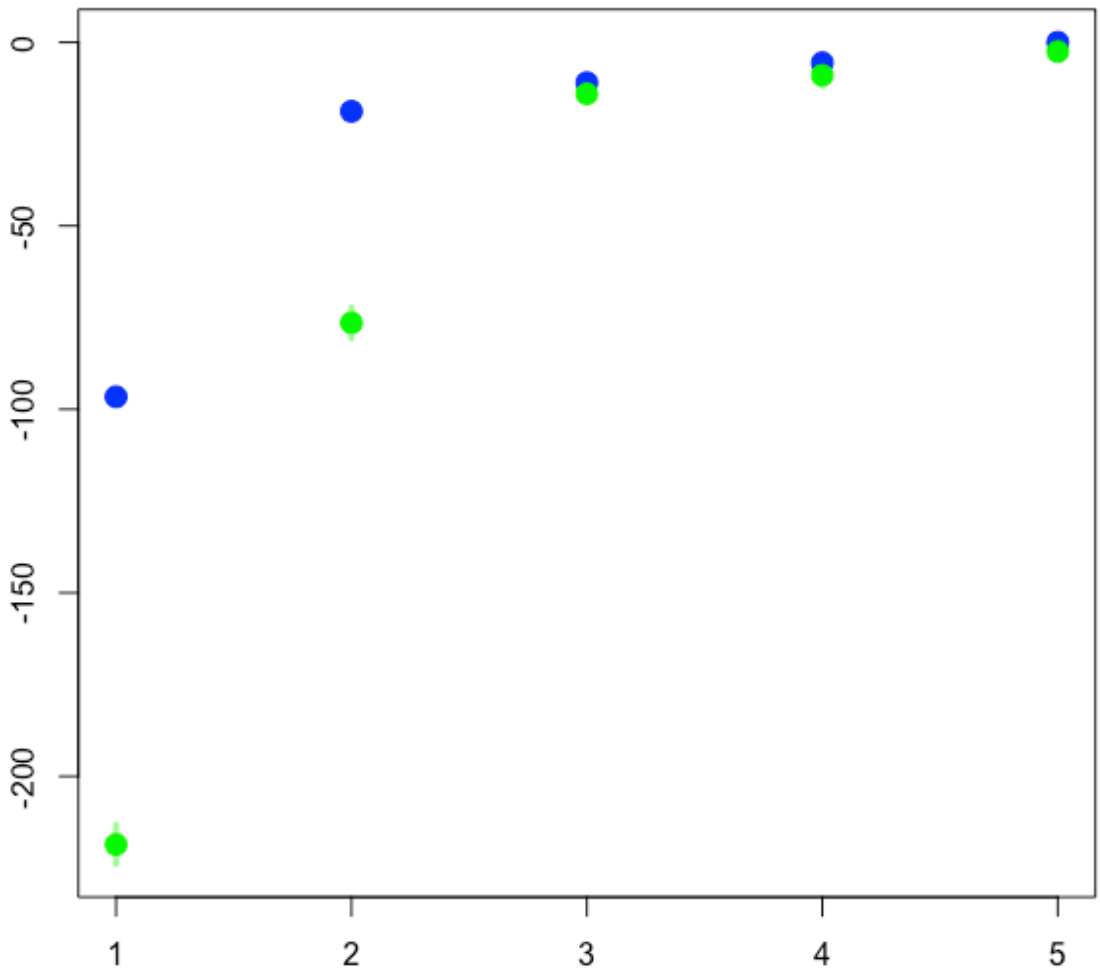

Figure 13. Cross validation results of the Construct analysis. The y-axis indicates predictive accuracy and the x-axis indicates values of K. Blue points indicate a spatial model of the data and green points indicates a non-spatial model of the data. For K=2, the spatial model (in blue) shows nearly equivalent support to higher values of K. Under the non-spatial model (in green), K=3 is preferred.

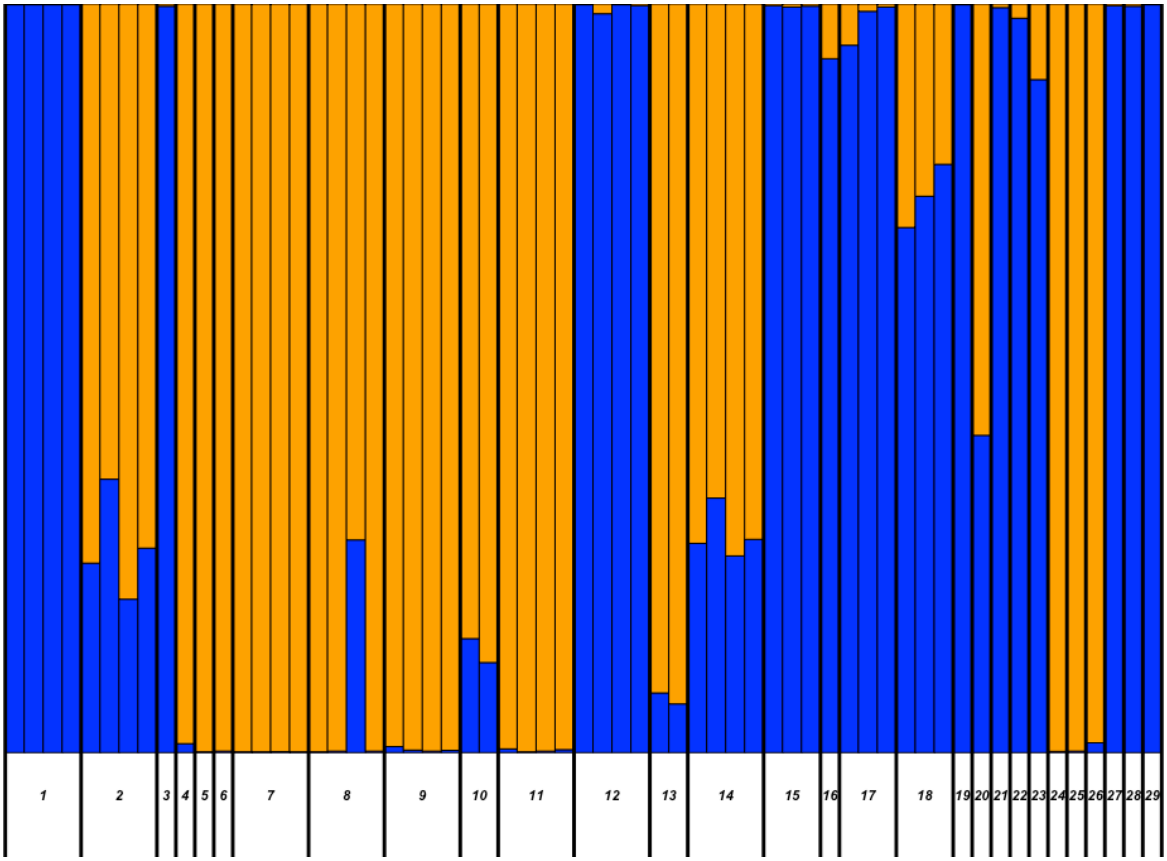

Figure 14 STRUCTURE admixture assignment plot for K=2. Individuals are represented by each column and the bold vertical lines represent locations as indicated by the numbers. Blue indicates assignment to *Pseudacris triseriata,* and orange indicates assignment to *P. feriarum.*

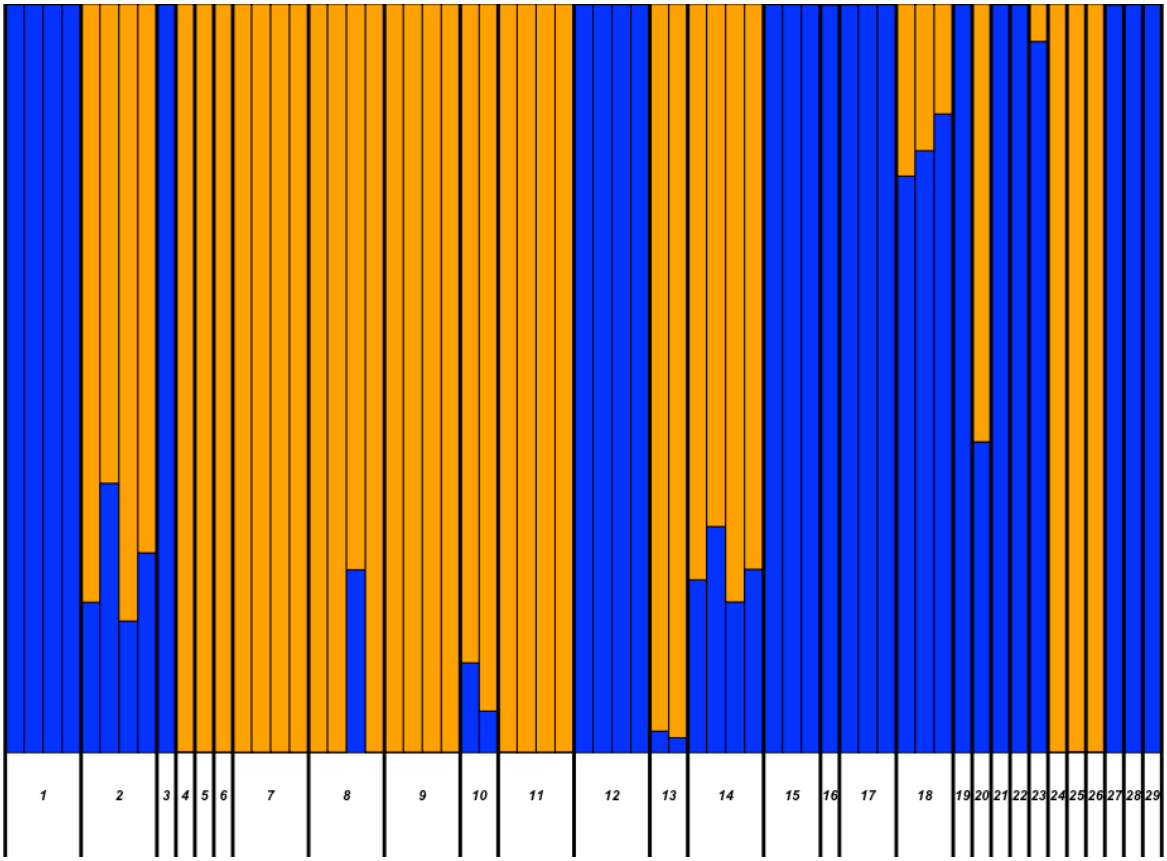

Figure 15 FastStructure admixture assignment plot for K=2. Individuals are represented by each column and the bold vertical lines represent locations as indicated by the numbers. Blue indicates assignment to *Pseudacris triseriata,* and orange indicates assignment to *P. feriarum.*

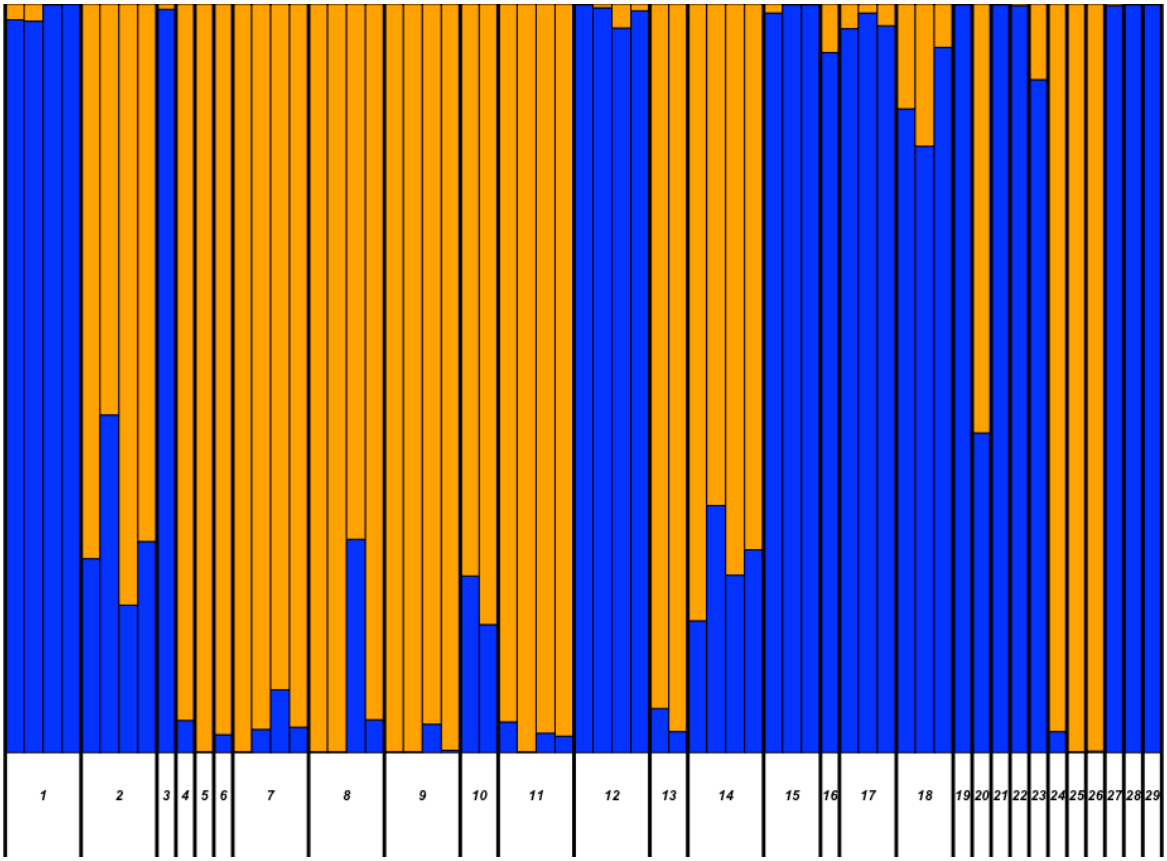

Figure 16. MavericK admixture assignment plot for K=2. Individuals are represented by each column and the bold vertical lines represent locations as indicated by the numbers. Blue indicates assignment to *Pseudacris triseriata,* and orange indicates assignment to *P. feriarum.*

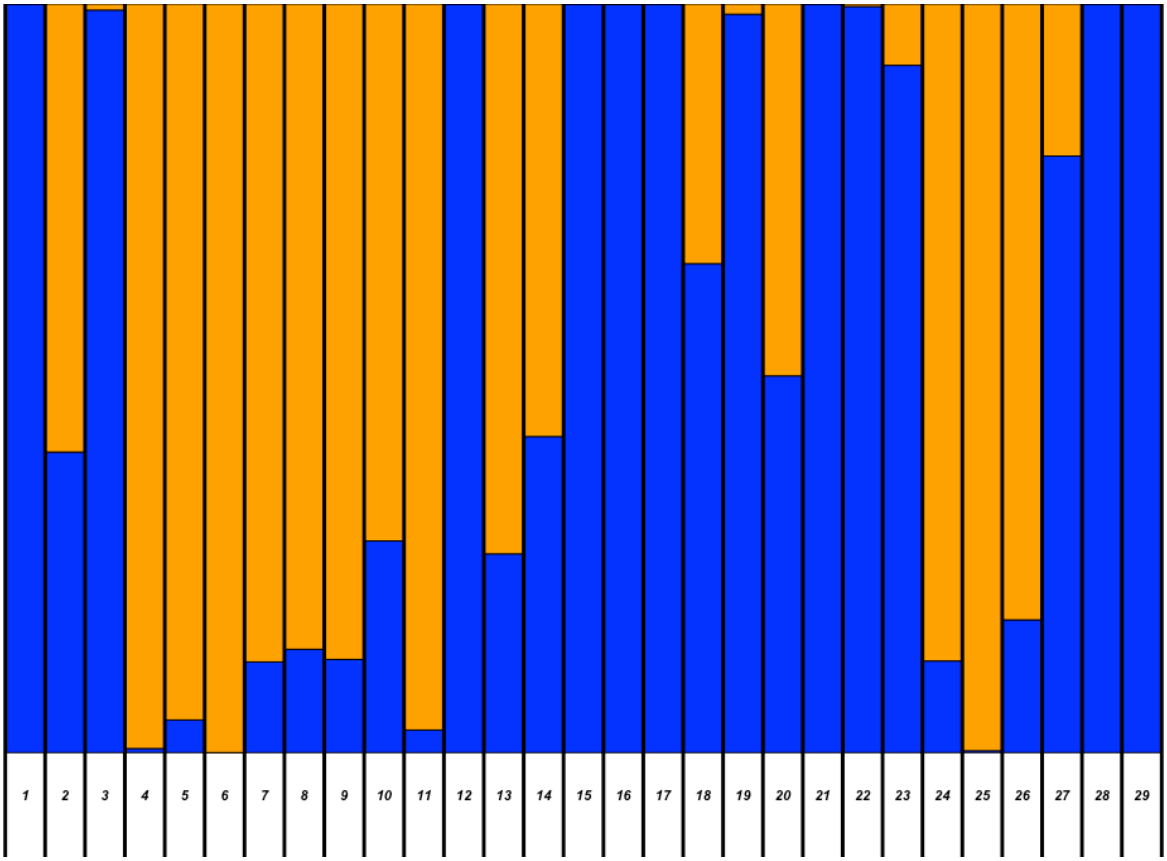

Figure 17 Construct admixture assignment plot for K=2. Locations are represented by each column as indicated by the numbers. Blue indicates assignment to *Pseudacris triseriata,* and orange indicates assignment to *P. feriarum.*

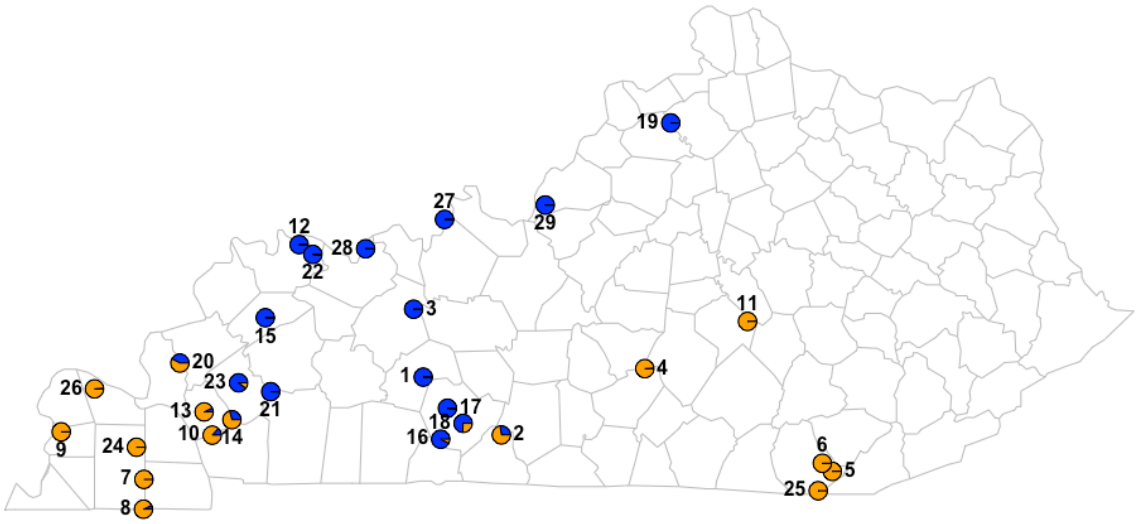

Figure 18 STRUCTURE pie chart admixture assignment for the sampled locations of Kentucky. Blue indicates assignment to *Pseudacris triseriata,* and orange indicates assignment to *P. feriarum.*

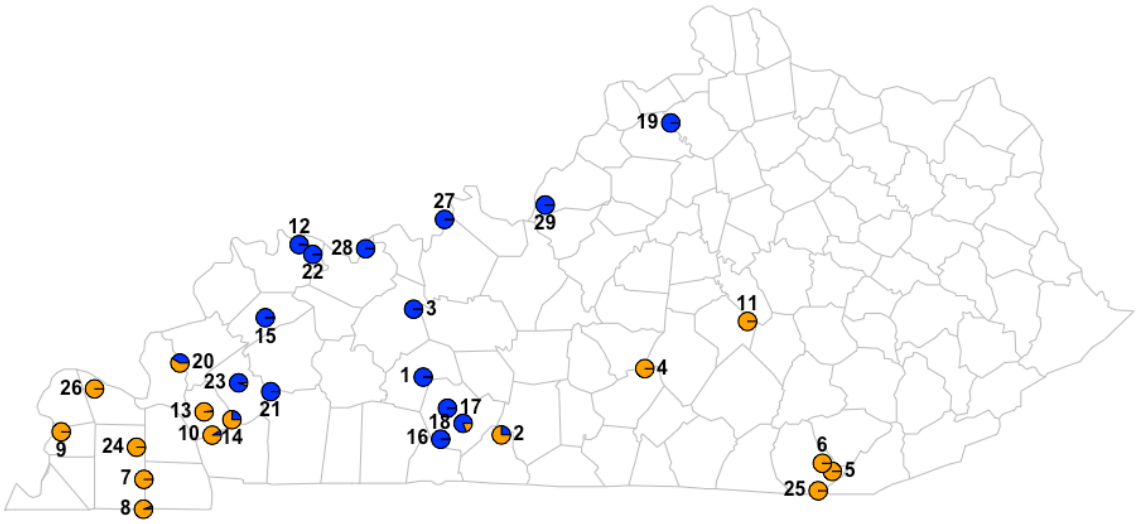

Figure 19 FastStructure pie chart admixture assignment for the sampled locations of Kentucky. Blue indicates assignment to *Pseudacris triseriata,* and orange indicates assignment to *P. feriarum.*

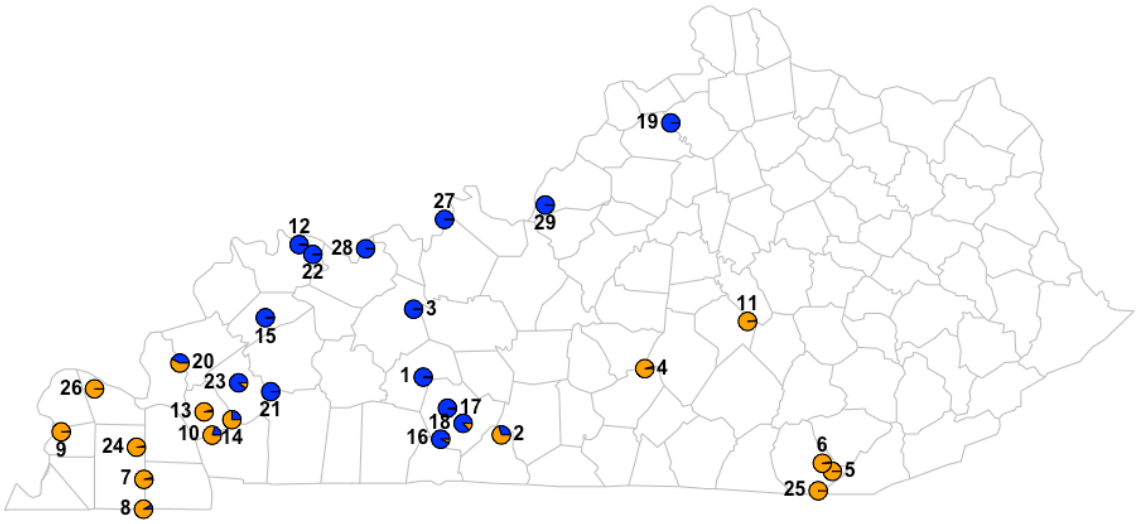

Figure 20 MavericK pie chart admixture assignment for the sampled locations of Kentucky. Blue indicates assignment to *Pseudacris triseriata,* and orange indicates assignment to *P. feriarum.*

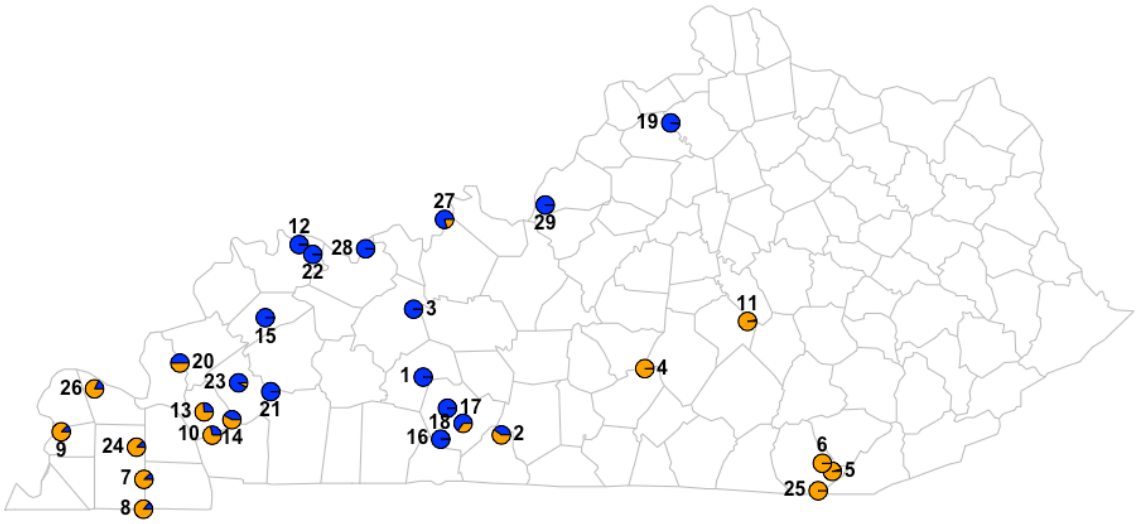

Figure 21 Construct pie chart admixture assignment for the sampled locations of Kentucky. Blue indicates assignment to *Pseudacris triseriata,* and orange indicates assignment to *P. feriarum.*

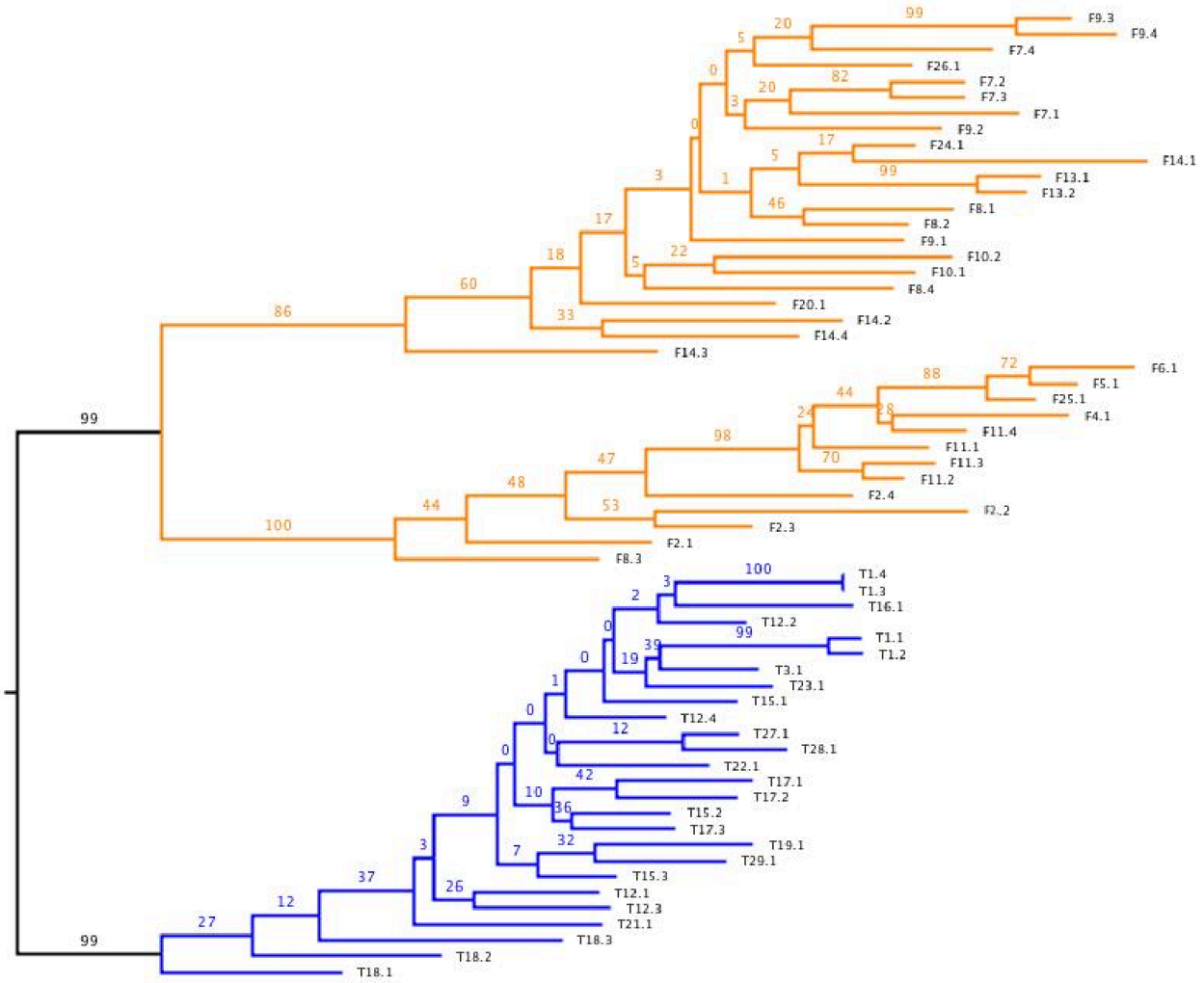

Figure 22 Maximum likelihood tree generated by RAxML. Bootstrap values are indicated for each node. Each branch is coded by is species assignment from STRUCTURE as well as its location and individual ID. Branches are colored by their species assignment by STRUCTURE. Blue indicates assignment to *P. triseriata,* and orange indicates assignment to *P. feriarum.*

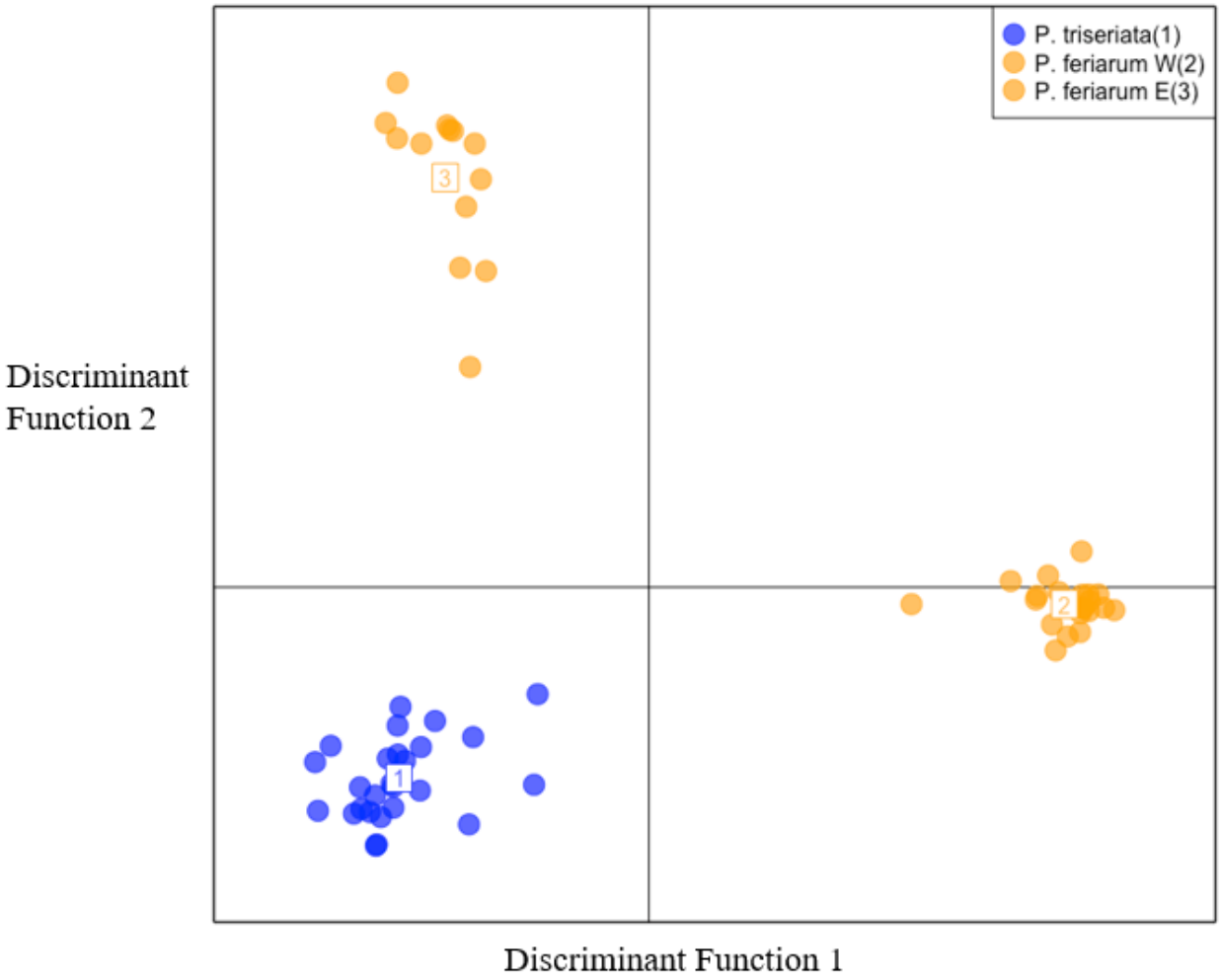

Figure 23 Discriminant Analysis of Principle Components of data indicating three genetic clusters. Points indicate individuals within the analysis. Individuals are colored by their species assignment by STRUCTURE. Blue indicates assignment to *P. triseriata,* and orange indicates assignment to *P. feriarum.*