POPULATION GENOMICS AND CONSERVATION OF TEXAS CAVE AND SPRING SALAMANDERS (PLETHODONTIDAE: EURYCEA)

by

ANDREW BRIAN CORBIN

Submitted in partial fulfillment of the requirements

for the Degree of Doctor of Philosophy at

The University of Texas at Arlington

August, 2020

Arlington, Texas

Supervising Committee:
Paul Chippindale
Todd Castoe
Matthew Fujita
Matthew Walsh
Eric Smith

Acknowledgments

I am grateful to so many people who have helped me navigate the past 6 years of graduate school. Firstly, and most importantly, is the unwavering support from my loving wife Lacey who has been a constant source of encouragement and motivation. I also thank Todd Castoe and the members of the Castoe lab for many years of friendship and adventure – in no particular order, Jacobo Reyes-Velasco, Daren Card, Drew Schield, Rich Adams, Audra Andrew, Giulia Pasquesi, Blair Perry, Nicky Hales, Aundrea Westfall, Ricky Orton, and Zack Nicholakis. Also, a huge thanks to the many friends I've made at UTA, including Rachel and Eli Wostl, Matt Moseley, Goutam Sarker, Justin Jacobs, Will Budnick, Kris Row, Thor Larson, and James Titus-Mquillan. I would also like to thank my graduate committee for their guidance – Drs. Todd Castoe, Matt Fujita, Chris Nice, Matt Walsh, and Eric Smith. Thank you to Dr. Nick Pollock and Rachel Wostl for teaching me how to be a better instructor and mentor. Thank you to all of the dedicated people working to save Texas salamanders like Nate Bendik, Tom Devitt, Dee Ann Chamberlain, David Hillis, Dante Fenolio, Randy Gibson, and especially to Andy Gluesenkamp. Thank you to my undergraduate advisor, Dr. John Mull for your mentorship and inspiring me to pursue graduate school. Also, a huge thanks to all the Biology office personnel, Sherri Echols, Gloria Burlingham, Ashley Priest, Linda Taylor, Chris Magno, Kathleen Demuth, and Mallory Roelke. Also, thank you to various funding sources who made this work possible: The Beta Chapter of the Phi Sigma Society, The Foundation for the Conservation of Salamanders, and the Texas Parks and Wildlife Department. Finally, I would like to express my deepest appreciation to my Advisor Paul Chippindale for many years of guidance through the often difficult journey through graduate school.

Dedication

I dedicate this work first and foremost to my son Charlie. You are almost 1 year old, and I'm already counting down the days until your mom lets me buy your first pet salamander! You can make me smile or laugh even on the toughest days. I love you kid. I also dedicate this to my parents for encouraging my love of science and fascination for nature. Now seems like a good time to admit that I kept numerous spiders and bugs as pets that you didn't know about…. sorry, not sorry.

Abstract

POPULATION GENOMICS AND CONSERVATION OF TEXAS CAVE AND SPRING SALAMANDERS (PLETHODONITIDAE: EURYCEA)

Andrew B. Corbin, PhD

The University of Texas at Arlington, 2020

Supervising Professor: Paul T. Chippindale, PhD

Understanding species boundaries and gene flow among the cave and spring salamanders of the central Texas region (genus *Eurycea)* has been a challenge for many decades. Although previous research has greatly increased our understanding of these salamanders, there are still many unanswered questions involving the number of species in the group and the degree to which populations are connected. The plummeting cost of DNA sequencing has transformed the field of biology, including the field of population genetics. Before the recent advent of high-throughput sequencing, population genetic studies in *Eurycea* were limited to only a few genetic markers. Here, we leverage next-generation sequencing to test hypotheses of species boundaries, examine patterns of gene flow, and measure genetic diversity using thousands of genome-wide markers. I find multiple instances of hybridization between distantly related species, and evidence for an undescribed species of an extreme blind cave salamander. I also work with San Antonio Zoo, San Antonio, TX, to build a captive research population of the Texas blind salamander (*E. rathbuni*) and investigate patterns of diversity and generate preliminary estimates of effective population size for the species.

Table of Contents

Chapter 1. Introduction

Central Texas harbors an incredible array of organisms evolved for life in spring outflows and underground water-filled caves of the Edwards Aquifer. Some of the most fascinating organisms on the planet inhabit this system, including the cave and spring salamanders of the genus *Eurycea*, commonly known as brook salamanders due to their affinity for moist or aquatic habitats. Many of these species are listed as Threatened or Endangered at the Federal and State levels due to a plethora of threats to groundwater quantity and quality throughout central Texas. Although decades of research have illuminated many facets of the ecology, diversity, and biology of these animals, many important aspects of this group still remain unknown. Previously, researchers thought of populations of *Eurycea* as relicts of a more widely distributed species that now inhabit islands of moist aquatic habitats in an otherwise arid landscape. However, recent work using genetic data has largely overturned this belief, and it is clear that many populations are more connected than previously thought. In Chapter 2, I use new genomic approaches to understand these patterns of population connectivity, hybridization, introgression, and species boundaries within *Eurycea*.

Genomic approaches have transformed the way biologists understand many areas of biology, including understanding genetic diversity among populations of conservation concern. Because genomic data can be very computationally intensive to analyze, it is crucial that researchers have fast, reliable software for estimating population genomic parameters. In Chapter 2, I present the software program ThetaMater (implemented in R) that efficiently estimates the population genetic parameter theta, a measure of genetic diversity. This statistic not only provides useful information that can be used to compare diversity among populations but can also be used to

estimate effective population size. I use this program in Chapter 3 to further illuminate patterns of diversity in the Texas blind salamander.

Although many species of Texas *Eurycea* are relatively widespread throughout the region, the Texas blind salamander (*E. rathbuni*) is known only from a few sampling localities from the San Marcos Pool of the Edwards Aquifer. This elusive species was one of the original members of the Endangered Species Act of 1973 and is still listed as endangered to this day. Because a major water contamination event such as a chemical spill could be disastrous for the species, it is crucial that captive breeding populations be maintained for research and potential re-introduction purposes. Captive breeding programs of imperiled species should consider the population structure and genetic diversity of wild populations. For *E. rathbuni*, one unsolved question is the degree of isolation that exists between salamanders from different sampling localities. If these localities represent distinct, isolated populations, this should be considered when breeding these salamanders in captivity. If this species is largely panmictic, salamanders from different localities could be bred together without compromising the genetic integrity of the species. In Chapter 3, I explore the population structure and genetic diversity of *E. rathbuni* using emerging genomic techniques to generate the most robust population genetic assessment of the species to date, and I estimate effective population size for the species overall.

Ultimately, I use new and emerging genomic approaches to greatly increase our understanding of gene flow within Texas *Eurycea*. I find multiple instances of gene flow between distantly related species, evidence that an undescribed species of extreme blind salamander exists in the aquifer below the New Braunfels area, and provide estimates of genetic diversity within the Texas blind salamander (*E. rathbuni*), one of Texas' most iconic species. I hope that this work will also highlight the utility of genome-wide data in understanding rare and imperiled species.

Chapter 2. Mito-nuclear discordance and directional gene flow in central Texas cave and spring salamanders (Plethodontidae: Hemidactylini: *Eurycea*: *Paedomolge*)

Andrew B. Corbin[1], Andrew Gluesenkamp[2], Drew R. Schield[1], Daren C. Card[1], Rich Adams[1], Giulia I. M. Pasquesi[1], Matthew Moseley[1], Todd A. Castoe[1], and Paul T. Chippindale[1]

[1]Department of Biology, University of Texas at Arlington, Arlington, Texas 76019 USA

[2]Director of Conservation, Center for Conservation and Research, San Antonio Zoo, San Antonio, Texas 78212 USA

ABSTRACT

The Edwards Aquifer of central Texas is home to a diverse radiation of groundwater salamanders of the genus *Eurycea*. Despite decades of research, resolution of evolutionary relationships and species boundaries within this group remains a challenge. We used analyses of well over 100 mitochondrial DNA sequences together with thousands of nuclear SNPs (generated via restriction-site associated DNA sequencing) from 72 individuals and 20 localities to estimate population structure, gene flow, and species boundaries. Our results support the existence of multiple distinct species, and historical and ongoing gene flow between distantly related species. We found strong evidence of ongoing hybridization between *E. sosorum* and *E. waterlooensis* (approximately 10 million years divergent), and strong disagreement between mitochondrial and nuclear phylogenies with respect to species assignment of individuals within several populations. Our results also indicate that the geographic distribution of some species may be larger than previously thought and support the recognition of a recently discovered subterranean species in New Braunfels, TX. Collectively, our findings illustrate the value of incorporating large nuclear datasets in disentangling the intricacies of population structure and gene flow within this enigmatic group of salamanders.

1. INTRODUCTION

Understanding how hybridization shapes the evolution of previously isolated lineages is a longstanding issue in evolutionary biology. It is well known that hybridization is common in animals (e.g., Mallet 2005; Abbott et al. 2016). In some cases, hybridization may increase the mean fitness of a population by introducing novel alleles (Charlesworth and Charlesworth, 1987), potentially providing beneficial combinations of alleles, or increasing hybrid vigor (Lippman and Zamir, 2007; Shull, 1908). However, substantial genetic divergence between two hybridizing populations is commonly thought to reduce hybrid fitness due to the buildup of genetic incompatibilities (Ayala et al., 1974; Sasa et al., 1998). Thus, the more distantly related two species are, the less likely they are to produce viable offspring. If hybridization occurs, there may be strong selection against hybrid phenotypes, especially if the parent species are phenotypically dissimilar. One potential consequence of hybridization is mitochondrial introgression, where geographic disparities exist between shared nuclear and mitochondrial DNA between lineages termed 'mito-nuclear discordance'. Numerous molecular studies including both mitochondrial and nuclear markers have documented this phenomenon, which in many cases has provided clues regarding historical and evolutionary processes that have shaped patterns of gene flow among populations (Brennan et al., 2016; Fontenot et al., 2011; Gompert et al., 2008; Toews and Brelsford, 2012).

The Edwards Aquifer of central Texas is inhabited by a radiation of endemic, cave- and spring-dwelling aquatic salamander species (*Eurycea*: Plethodontidae) whose relationships and levels of genetic isolation are in some cases poorly understood. Nearly all members of the group are fully aquatic and paedomorphic, retaining larval features such as gills at sexual maturity (Chippindale and Wiens, 2005; Sweet, 1977). Species diversity was underrepresented prior to use of molecular

methods, largely due to a combination of morphological conservatism in surface (spring) dwellers and varying degrees of morphological convergence among subterranean populations (e.g., Wiens et al., 2003). One cave species was originally assigned to a separate genus (*Typhlomolge*) based on its extreme blind cave (troglobitic) features (*Typhlomolge rathbuni* Stejneger, 1896). In addition, the Comal blind salamander (*Eurycea tridentifera* Mitchell and Reddell, 1965) was placed by Wake (1966) in the genus *Typhlomolge*. Mitchell and Smith (1972) reallocated *E. tridentifera* to *Eurycea* but retained the genus *Typhlomolge* for the other two species. Potter and Sweet (1981) and Sweet (1977) continued to recognize *Typhlomolge* as a separate genus consisting of *T. rathbuni* and *T. robusta* (the latter redescribed by them based on a single specimen collected in 1951). Later systematic studies have decisively placed *Typhlomolge* within *Eurycea*, and revealed roughly three times the number of *Eurycea* species than were thought to exist in the central Texas region (Chippindale et al., 2000, 1993; Chippindale and Wiens, 2005; Devitt et al., 2019; Hillis et al., 2001).

Several clades, essentially equivalent to subgenera, have since been designated under an unranked taxonomic system (Hillis et al., 2001). Here we focus on two of these: (1) the *Blepsimolge* clade of surface- and cave-dwellers, specifically the *E. neotenes* complex (i.e., *E. latitans*, *E. neotenes*, and *E. pterophila*, plus *E. tridentifera*, recently reassigned to *E. latitans* by Devitt et al., (2019) plus the more genetically divergent surface species *E. nana* and *E. sosorum*, and (2) the *Typhlomolge* clade corresponding to the former genus *Typhlomolge* (excluding the former taxon *T. tridentifera*). Divergences between these two groups may be as ancient as 10 MYA (Bonett et al., 2013; Chippindale et al., 2000; Dawley, 2001; Wiens et al., 2003).

Since Hillis et al. (2001) re-examined the phylogenetic relationships among the central Texas *Eurycea* and confirmed the results of Chippindale et al. (2000) while adding the new species *E.*

waterlooensis from Barton Springs in Austin to the *Typhlomolge* clade, many new populations of salamanders have been discovered, including four localities of an extreme troglobite that belongs to the *Typhlomolge* clade in New Braunfels, TX. Also, several geographically intermediate populations have been found between *E. sosorum* (type locality Barton Springs, Austin, TX) and *E. nana* (type locality san Marcos Springs, San Marcos, TX). Because morphology is a usually poor diagnostic tool for species delimitation in this group, most attempts to identify species boundaries and examine gene flow among populations have used limited mitochondrial and nuclear markers (Bendik et al., 2013; Chippindale et al., 2000; Lucas et al., 2008). These studies sometimes yielded discordance between mitochondrial and nuclear datasets (i.e., mito-nuclear discordance), and the degree to which this discordance is driven by introgression versus incomplete lineage sorting is poorly understood (Bendik et al., 2013). In this study, we utilized recent advances in coalescent-based species delimitation tools to examine relationships among members of this group. Here, we combine analyses of mtDNA sequences and genome-wide nuclear SNP dataset to: (1) to describe the agreement between patterns derived from mitochondrial and nuclear genetic markers; (2) To examine patterns of gene flow and introgression between members of Texas *Eurycea*; (3) to evaluate phylogenetic relationships among major clades and determine the taxonomic status of a recently-discovered, putative species of extreme blind salamander (clade *Typhlomolge*) via coalescent Bayesian species delimitation. We found extensive mito-nuclear discordance, and multiple instances of hybridization between distantly related and morphologically dissimilar species.

2. MATERIALS AND METHODS

2.1 Data Generation and processing

We obtained tissue samples from 72 *Eurycea* from 30 localities in central Texas (Supplementary Table 1), which included tail clippings, liver tissue, or whole juvenile specimens for our RADseq dataset. Tissues were preserved by snap-freezing, in lysis buffer, or in 70% ethanol. All sampling procedures were approved by IACUC (approval to PTC) at the University of Texas at Arlington and state and federal permits for specimen and sample collection were held by AGG. For samples to be used for RADseq library preparation, we used phenol-choroform-isoamyl alchohol extraction. For samples to be sequenced for the mitochondrial gene cytochrome-b (*cyt-b*), DNA was isolated in one of four ways: using a Qiagen DNeasy extraction kit (Qiagen, Inc., Valencia, CA, USA), phenol-chloroform-isoamyl alcohol extraction, the STE method described by Hillis et al. (1996), or a modification of the Chelex extraction method (Walsh et al., 1991) as described by Chippindale et al. (2000). Purified genomic DNA extracts were quantified using a Qubit flourometer 2.0 (Life Technologies, Carlsbad, CA, USA).

Double-digest restriction site-associated DNA sequencing (ddRADseq) libraries were constructed for 72 individuals using a protocol modified from Peterson et al. (2012). In brief, genomic DNA was digested with two restriction enzymes, *Sbf1* (8bp recognition site) and *Sph1* (6 bp recognition site), which were selected to target approximately 30,000 genomic loci. Double-stranded indexed DNA adapters, which contain eight consecutive N's (unique molecular identifiers – UMIs), were ligated onto the cut sites for each fragment. Samples were pooled and then size selected to lengths of 302-360 bp using the Blue Pippin Prep (Sage Science, Beverly, MA, USA). Libraries were amplified through PCR with Phusion high-fidelity DNA polymerase (New England Biolabs) using primers that contain flow-cell binding sequences and indices

specific to each sub-pool. To verify size selection and calculate DNA concentration, sub-pools were analyzed on a Bioanalyzer (Agilent, Santa Clara, CA, USA). Samples were then combined based on molarity and sequenced using an Illumina HiSeq2500 platform using 100 bp paired-end reads.

To obtain mt sequence data, we sequenced an 865 bp fragment of the cytochrome *b* (*cyt-b*) gene for 132 individuals, including samples of *E. sosorum, E. waterlooensis*, *E. rathbuni*, *E.* sp. New Braunfels (putatively new *Typhlomolge* clade species), *E. latitans*, *E. neotenes*, *E. pterophila*, *E. troglodytes*, and *E. tonkawae* (as an outgroup) and *E. nana*. PCR amplification and sequencing cycling conditions are described in (Bendik et al., 2013). Briefly, PCR products were amplified with standard Taq polymerase (New England Biolabs or Promega), Hot Start *Ex-Taq* (Takara-mirus) or Phusion (New England Biolabs). Amplification for PCR and sequencing was performed with the primers listed in Supplementary Table 2. Typical reactions included 30 cycles and an annealing temperature of 50°C. PCR products were visualized using gel electrophoresis, and purified using Qiagen gel extraction kits (Qiagen, Inc., Valencia, CA, USA). Both strands of each amplicon were sequenced using BigDye on an ABI 3700 capillary sequencer (Life Science Technologies, Grand Island, NY, USA).

2.2 ddRADseq data processing

Raw reads were processed using the STACKS pipeline v1.37 (Catchen et al., 2013). PCR clones were removed with the clone_filter program in STACKS based on the UMIs, which were then trimmed away using the FASTX-Toolkit trimmer (HannonLab, 2014) (HannonLab, 2014). Trimmed reads were then demultiplexed into sample-specific read files using the process_radtags program in STACKS. Reads that lacked barcodes, digest cut sites, or that had

poor quality scores were discarded (-q option). Because the second reads were of considerably lower quality, only forward reads from the paired-end reads were used for further analysis.

To obtain population genetic information, we used pyRAD v3.0.61 (Eaton, 2014), which allows for indels when clustering sequence reads into orthologous loci. Because downstream analyses can be affected by pipeline parameters (Pante et al., 2015), we generated multiple initial datasets using different metrics within this pipeline to find an optimal dataset. Our aim was to generate a dataset that would give us the highest number of SNPS possible while filtering unnecessary paralogs. Specifically, there are two important parameters two considered: similarity value to be used for the alignment during within and across-sample clustering (Wclust), and (2) the minimum depth necessary to include a locus in the final dataset (MinCov). To explore how these parameters affect the number of SNPs and loci, we generated three datasets. First, Params_A with Wclust = .7 and MinCov = 36 (50% missing data allowed per locus), then Params_B with Wclust = .8 and MinCov = 54 (75% missing data allowed per locus), and finally Params_C with Wclust = .9 and MinCov = 64 (~90% missing data allowed per locus). All other parameters were held constant, with Mindepth = 5 (Minimum depth to form a cluster within and between samples), NQual = 20 (maximum number of sites with a quality score < 20), MaxSH = 15 (maximum number of individuals with a shared heterozygous site, which removes potential paralogs) which equates to ~20% of indivuduals in our dataset. We examined each dataset for the number of loci, SNPs, unlinked SNPs, and parsimony informative sites (Supplementary Table 3). Based on these results, we selected the Params_A for further processing and data analysis. For the STRUCTURE, BFD*, and RAXML phylogeny, we randomly selected one SNP per RAD locus to adhere to the assumption of free recombination between loci of the models.

2.3 Population structure and phylogenetic inferences

Population structure was estimated using STRUCTURE v2.3.4 (Pritchard et al., 2000), which includes an explicit population genetic model, and Discriminant Analysis of Principal Components (DAPC) (Jombart, 2008), which is model-free and uses principal component analysis PCA to cluster samples. Initial runs of STRUCTURE included MCMC chains of 50,000 generations (discarding 25% as burn-in) from K = 1-15 (each with three iterations) under a mixed ancestry model using sampling localities as putative population origins. Based on likelihood estimates in these initial runs, we determined that the most likely value of K was between 4-7 depending on which subsets of the data were analyzed (not surprisingly, analyses that included more taxa with higher levels of divergence generated higher values of K; see Figure 1). We then performed longer runs of STRUCTURE with MCMC chains of 500,000 generations (discarding 25% as burn-in) for K = 4-7 (each with three iterations), and used the ΔK method (Evanno et al., 2005) implemented in StructureHarvester (Earl and vonHoldt, 2012) to determine the most likely value of K as described above. We also generated three separate structure plots to evaluate the presence and extent of fine scale population structure for several populations of interest. Separate STRUCTURE plots were generated for the following subsets: (1) *E. sosorum* (at Barton Springs), as well as *E. sosorum* at newly discovered localities (Blowing Sink Cave, Cold Spring, Spillar Ranch Spring, Upper Taylor Spring, Zara Well; all were assigned to by Devitt et al., 2019 to *E. sosorum*); (2) members of the *Typhlomolge* clade, including the samples of unknown taxonomic status from New Braunfels, TX; (3) all other samples from the *Blepsimolge* clade excluding *E. nana* (see Supplementary Table 4 for samples included in each analysis). These plots were generated using the same general method as above, implementing a shorter MCMC chain (50,000 generations, discarding 25% as burn-in) over a

range of *K*, and then processed longer runs of the best supported value of *K* (MCMC chains of 500,000 generations). We performed DAPC in R (R Core team, 2015) using the ade4 (Dray et al., 2007) and adegenet (Jombart, 2008) packages. We generated several plots to visualize population structure across all samples and within select groups. First, we generated a plot including all individuals. To examine population structure at finer scales, we also generated three additional plots: (1) a plot including all samples excluding *E. nana*; (2) a plot including *E. waterlooensis* and *E. sosorum*; and (3) a plot including *E. latitans* and *E.* sp. New Braunfels to explore potential hybridization found in our STRUCTURE analysis. We optimized each plot by examining the '*a*-score', which assesses the optimal number of retained PCs by performing DAPC under different numbers of PCs and estimates re-assignment probabilities. To further explore patterns of variation, we estimated pairwise F_{ST} using the R package diveRsity (Keenan et al., 2013). We ran two separate analyses. First, we estimated diversity among *E. nana, E. sosorum* and all geographically intermediate populations to see if the intermediate populations were more similar to *E. nana* or *E. sosorum*. Next, we estimated pairwise F_{ST} between all members of *Typhlomolge* and *E. latitans* to determine whether *E.* sp. New Braunfels was more similar to *E. latitans* than other *Typhlomolge*.

We tested for evidence of gene flow using TreeMix v1.12 (Pickrell and Pritchard, 2012). We allowed from 0-8 migration events between lineages and determined the most likely model by calculating the proportion of the variance in relatedness between populations by each migration model. Samples were assigned to populations based on the current taxonomy and, for some individuals, sampling locality.

Raw mitochondrial gene sequences (*cyt-b*; 865 bp) were edited using Sequencher v4.5 (Gene Codes Corp., Ann Arbor, MI, USA). Alignments were generated using MUSCLE (Edgar, 2004),

implemented in Geneious Prime v2020.1.2 (https://www.geneious.com) making manual adjustments as needed. We estimated a maximum likelihood tree using RAxML (Stamatakis, 2014) under the GTR_GAMMA model of evolution, and assessed tree support with 1000 bootstrap replicates. We specified *E. tonkawae* as an outgroup, and visualized the tree using FigTree v1.4.4 (https://github.com/rambaut/figtree/releases). We used PopArt v1.2.3 (Leigh and Bryant, 2015) to calculate and visualize a minimum spanning network.

2.4 Phylogenetics and species delimitation based on genome-wide variants

First, we estimated a maximum likelihood tree using all individuals. However, because it is well known that admixture can lead to misleading phylogenies (Cavender-Bares et al., 2015), we generated a second tree excluding any admixed individuals as indicated by the results of our STRUCTURE, DAPC, and TreeMix results indicate substantial admixture (see results below). For this analysis, we selected only one sample from each sampling locality at random, but only if the sample did not show any evidence of admixture in the STRUCTURE or DAPC analysis (See Supplementary Table 4 for samples included in the tree). Invariant sites were removed from the PHYLIP files using the phrynomics package (https://github.com/bbanbury/phrynomics) in R. Maximum likelihood phylogenetic tree reconstruction was performed in RAxML v8.0 according to the ASC_GTRGAMMA model of evolution, which corresponds to the GTR+G+ Lewis ascertainment bias correction (–asccorr=lewis). We conducted 1000 bootstrap replicates for each tree. Trees were visualized with FigTree v1.4.4.

To estimate species limits within the *Typhlomolge* clade, we performed Bayesian Species Delimitation (BSD). We generated a dataset including the *Typhlomolge* clade that included *E. rathbuni, E. waterlooensis,* and *E.* sp. New Braunfels. Because the BSD method must include at

least two groups, *E. nana* was included as an outgroup to test a single-species hypothesis (see Fig. 4 for hypotheses tested). Also, because recent gene flow may interfere with species tree estimation, individuals were excluded if they contained ~15% or greater ancestry coefficient from the other clade based on our STRUCTURE results. To reduce computational time and make sample sizes more even, we excluded some individuals from populations and species with larger sample sizes (see Supplementary Table 4 for sampling scheme). We tested competing hypotheses of species limits using the BFD* method (Leaché et al., 2014) in the SNAPP v1.2.4. plugin (Bryant et al., 2012) for BEAST2 v2.3.1 (Bouckaert et al., 2013). Path sampling for marginal likelihood estimation for each model included 48 steps, each with 1 X 10_5 generations plus a 10% burn-in. Comparisons were made for each competing model to identify the best-supported species model with Bayes factors using the Kass and Raftery (1995) framework. The best supported model was visualized using DensiTree v2.2.1 (Bouckaert, 2010).

3. RESULTS

3.1 Evaluating population structure and gene flow

Post clone-filtering, quality filtering, and alignment, we found that the pyrad dataset with wcliust = 0.7 and mincov = 36 (50% coverage) yielded the most loci and used this dataset for subsequent analyses. We recovered 24,286 SNPS and 6,201 loci across 72 samples. For the STRUCTURE dataset that included all samples, the ΔK method indicated the most likely value of $K = 4$. These groups correspond with four major clades/ancestral populations: the *Typhlomolge* clade, *E. sosorum* and geographically proximate populations, the *E. neotenes* complex, and *E. nana* (Fig. 1). It is notable that STRUCTURE did not discriminate among the various species within *Typhlomolge* or the *E. neotenes* complex but did distinguish *E. sosorum* from *E. nana*. The lack of distinction between the various species within the *Typhlomolge* clade and the *E. neotenes*

complex may be explained by known limitations of STRUCTURE, including variable sample sizes (Kalinowski, 2011). We then performed separate structure analyses each of the 3 subsets of the data: for the blind *Typhlomolge* clade, the *E. neotenes* complex, and *E. sosorum*. For the *Typhlomolge* clade, *E. sosorum, and E. neotenes* complex, we found that the optimal value of *K* = 5, *K* = 3, and *K* = 7, respectively (using the delta *K* method described in sec 2.2) (Fig. 1).

For the dataset that included all samples for the DAPC analysis, the K-means clustering analysis in ADEGENET indicated K = 7, with most cluster groups largely corresponding to the current taxonomy (Fig. 2). The DAPC analysis indicated that *E. nana* is highly differentiated from all other samples (Supplementary Figure S1). This is consistent with previous studies of *E. nana*, which is highly divergent from other Texas *Eurycea* based on allozyme analysis and morphology (Chippindale et al., 1998). For the DAPC analysis excluding *E. nana*, we recovered the same group assignments as seen in the DAPC run including all samples (Fig. 2; see Supplementary Fig. S1). Overall, the K-means clustering step in ADEGENET found that K = 6.

We found that two samples (one *E. sosorum* and one *E. waterlooensis)* share ~25% nuclear allelic content with the reciprocal species. To further explore these putative hybrids, we ran a separate DAPC analysis for *E. waterlooensis* and *E. sosorum* (including four nearby populations; Fig. 2B). These two putative hybrids formed a separate cluster to the exclusion of all other samples. We explored possible admixture between *E.* sp. New Braunfels and *E. latitans* (Fig. 2C) indicated by the STRUCTURE results and found that *E.* sp. New Braunfels individuals with varying degrees of admixture in the STRUCTURE analysis formed a separate cluster. The clustering steps for both of these DAPC analyses indicated a *K* = 3.

Overall, TreeMix produced a phylogeny very similar to those of previous studies (Bonett et al., 2013; Devitt et al., 2019; Hillis et al., 2001). The amount of variation explained by the model leveled at M = 3 (~98% of variation). These results suggest unidirectional gene flow from *E. latitans* (*Blepsimolge*) to *E.* sp. New Braunfels (*Typhlomolge*) (Fig. 3). Our results also indicate gene flow from *E. waterlooensis* (*Typhlomolge* clade) to *E. sosorum* (*Blepsimolge* clade). Overall, these results corroborate the STRUCTURE and DAPC results, indicating gene flow among these distantly related species. Pairwise F_{ST} results suggest substantial differentiation between *E. nana* and *E. sosorum* and indicate that all intermediate populations are more similar *E. sosorum* at Barton Springs than *E. nana* (Fig 4). We also find that *E.* sp. New Braunfels is more similar to the three *Typhlomolge* species than to *E. latitans* (in the *Blepsimolge* clade) (Fig. 5).

3.2 Phylogenetic estimates and species delimitation

The reduced RAxML phylogeny shows that the *Typhlomolge* and *Blepsimolge* clades are both monophyletic with 100% bootstrap support (Fig. 6). The tree also indicates that the newly-discovered localities of *E. sosorum* between San Marcos and Barton Springs form a monophyletic group, and are sister to *E. sosorum*. The *E. neotenes* complex was generally poorly resolved and suffered from poor bootstrap support, likely due to incomplete sampling of these species and complex evolutionary relationships between these species. In addition, members of this group are recently diverged from one another and likely have experienced hybridization (e.g., Bendik et al., 2013). *Eurycea pterophila* and *E. neotenes* did not form a monophyletic group in this analysis, and *E. latitans* fell within the *E. pterophila* group. Within the *Typhlomolge* clade, we found that *E.* sp. New Braunfels is sister to *E. rathbuni*, with *E. waterlooensis* sister to both of these groups. It should be noted that only Panther Canyon Well

was represented in this analysis, as both animals from Mission Valley Bowling Alley Well and the troglobitic Hueco Springs sample showed substantial evidence of hybridization.

The RAxML phylogeny including all samples showed an unexpected topology, where some members of *E. latitans* (from Honey Creek Cave, Preserve Cave, and Badweather Pit) and *E.* sp. New Braunfels (samples from Mission Valley Bowling Alley Well and Hueco Springs) are near the root of the tree, and the *E. neotenes* complex is sister to all other *Blepsimolge* with *E. nana* sister to *E. sosorum*. These patterns largely disagree with the reduced tree and may be the result of including admixed individuals. Bayesian species delimitation found strong statistical support for the three species model analysis of *Typhlomolge*, suggesting that populations in the New Braunfels area represent an undescribed species (Fig. 7). These results largely agree with our STRUCTURE and DAPC results.

In our mitochondrial phylogeny and minimum spanning haplotype network, we found strong support for two major mitochondrial clades that correspond to the *Typhlomolge* and *Blepsimolge* clades with strong bootstrap support (Fig. 8). Although these two clades are clearly distinct, there are a few instances where members of one clade are nested within the other. For example, two *E. sosorum* group with *E. waterlooensis*, and one *E. waterlooensis* is nested within *E. sosorum* where these two species live in sympatry at Barton Springs. Also, we find that one *E. rathbuni* is nested within the *E. neotenes* complex. Within the *Blepsimolge* clade, the mitochondrial phylogeny shows that *E. nana* shares a nearly identical mt haplotype with *E. sosorum* populations between San Marcos Springs and Barton Springs, and that this haplotype is found in many *E. sosorum* at Barton Springs. We also find that *E. sosorum* at Barton Springs is highly variable and contains haplotypes of upstream populations that are similar to *E. nana*. Within the *Typhlomolge* clade, there is no clear distinction between species, as many *E. rathbuni* share

haplotypes with *E. waterlooensis.* Also, *E.* sp. New Braunfels is nested within *E. rathbuni,* yet shows relatively little differentiation and forms a monophyletic clade. In the minimum spanning network, they are clearly very similar to some *E. rathbuni,* but there are relatively few changes between each other compared to *E. rathbuni.*

DISCUSSION

We find strong evidence of hybridization between the *Typhlomolge* and *Blepsimolge* clades of central Texas *Eurycea*. Although they diverged roughly 10 MYA (Bonett et al., 2013), we found evidence of recent or ongoing gene flow between *E. waterlooensis* (*Typhlomolge*) and *E. sosorum* (*Blepsimolge*) at Barton Springs where these two species are essentially sympatric (Fig 3.) Also, we detected evidence of admixture between the putative undescribed species of blind salamander (*Typhlomolge*) at New Braunfels and *E. latitans* (*Blepsimolge*). These findings are striking, not only because *Blepsimolge* and *Typhlomolge* diverged so long ago, but because of the extreme morphological differences between these species (although some *E. latitans* exhibit an extreme troglobitic morphology, while other *E. latitans* exhibit a range from subterranean to surface). This also provides additional evidence that when lineages are subject to extreme selective environments (such as surface versus cave), morphological convergence may mislead phylogeny (Bendik et al., 2013; McGaugh et al., 2020; Wiens et al., 2003).

4.1 Localized hybridization of distantly related lineages and mito-nuclear discordance

Previous studies using allozymes, mitochondrial DNA, and/or RADseq data have been instrumental in establishing species boundaries and conservation units in central Texas *Eurycea* (Bendik et al., 2013; Chippindale et al., 2000; Devitt et al., 2019; Hillis et al., 2001). Some phylogenetic studies of this group indicated that species may be isolated due to relatively high

genetic divergences (Lucas et al., 2008) and the assumption that groundwater was slow-moving and that springs and underground water reservoirs within the aquifer were inaccessible from one another (Hauwert, 2016). Recent hydrogeologic studies (dye trace and geochemistry studies) have since overturned these assumptions about groundwater connectivity and flow (Hauwert et al., 2004; Hauwert, 2016; Johnson et al., 2012), and more recent genetic work began to show that populations may not be as isolated as previously thought. For example, Chippindale (2009) discovered that some *E. rathbuni* exhibited mt haplotypes that were extremely divergent from those seen in other *Typhlomolge,* and very closely related to members of the *Blepsimolge* clade, including surface salamanders at Comal and nearby Hueco Springs. Here, we find evidence of the same mito-nuclear discordance. Bendik et al. (2013) conducted a detailed morphological and genetic analysis (using only mt DNA) which showed that the *E. neotenes* complex (*Blepsimolge*) may exhibit hybridization and species were likely over split (e.g., recognition of *E. tridentifera* as separate from *E. latitans*, confirmed using nearly identical RADSeq methods by Devitt et al. (2019).

Our results highlight several prominent instances in which nuclear and mitochondrial inferences of species identity and relatedness disagree. The first of these relates to the phylogenetic affinity of populations that are geographically intermediate between topotypical *E. sosorum* (Barton Springs, Austin, Travis County) and *E. nana* (San Marcos Springs, San Marcos, Hays County). Our nuclear dataset suggests that these populations are nearly indistinguishable from topotypical *E. sosorum* (Barton Springs), while they have one of the same mt haplotypes as *E. nana* (haplotype variation in *E. nana* is extremely minor, and *E. nana* is very divergent based on nuclear data, including allozymes [Chippindale et al., 2000, 1998]; Fig. 8). Although the cause of this incongruence is unknown, incomplete lineage sorting may be one explanation (Bendik et al.,

2013). However, this seems unlikely, as the mt haplotype found in all geographically intermediate locations is nearly identical to one of the mt haplotypes found in *E. nana*, and the haplotype found in ~%70 of *E. sosorum* at Barton Springs is quite divergent from those of the *E. nana* group, the intermediate populations, and the other roughly 30% of *E. sosorum* at Barton Springs, which have the *E. nana* haplotype. This pattern of fixation of the *E. nana* mitochondrial haplotype in the geographically intermediate populations despite zero evidence of nuclear introgression is strongly suggestive of mitochondrial introgression following a fairly recent hybridization event (reviewed in Toews and Brelsford (2012). These patterns of gene flow indicated by mito-nuclear discordance and our nuclear RADseq analyses indicate that gene flow generally follows the flow of groundwater through the Aquifer, and that the *E. nana* haplotype swept northward through nearby populations of *E. sosorum*, became fixed in these populations, and now exist in ~33% of *E. sosorum* at Barton Springs where this segment of the Aquifer terminates.

The incongruence seen between *E. waterlooensis* and *E. rathbuni* (neither is mitochondrially monophyletic) is more suggestive of incomplete lineage sorting given that each species' mt haplotype is at ~50% frequency in the reciprocal species (Fig. 8). In addition, the putative new species in the *Typhlomolge* clade from New Braunfels appears as sister in the mt tree to a subset of *E. rathbuni*. *Eurycea rathbuni* itself contains two main mt haplotype groups (individuals with haplotypes from either group are often found in microsympatry), and two individuals show mt haplotypes very similar to those of members of the Blepsimolge clade that include *E. neotenes*, *E. pterophila*, *E.* sp Pedernales, and *E. latitans*. This suggests that *E. rathbuni* may have undergone past events of fragmentation and reconnection (very plausible given the complex history of the Edwards Aquifer (Musgrove et al., 2019) and in addition, has experienced mt

introgression from a member of the *Blepsimolge* clade (*E.pterophila*, the most likely candidate, occurs in nearby springs and caves).

Our results highlight the importance of analyzing nuclear and organellar genetic data in tandem when studying gene flow in natural populations, and that discordance between these datasets can reveal patterns of incomplete lineage sorting and hybridization (Toews and Brelsford, 2012). Our results show that distantly related *Eurycea* with strongly contrasting morphologies can readily hybridize in the wild. It is well documented that time since divergence between lineages is positively correlated with pre-and post-zygotic reproductive barriers (Coyne and Orr, 2004; Edmands, 2002; Sasa et al. 1998), however amphibians display an exceptional ability to hybridize between distant relatives (Aarntzen et al. 2009; Pramuk et al 2007). In one extreme example, two toad species (*Bufo bufo* and *Bufo viridis)* that diverged >36MYA can produce viable offspring (Duda, 2008; Portik and Papenfuss, 2015). Among plethodontid salamanders, (Wiens et al., 2006) show that rare instances of hybridization have been observed between lineages that diverged about 11.4 MYA, however they also document that species pairs living in sympatry and able to hybridize shared a common ancestor no earlier than about 8.57 MYA. Because *Typhlomolge* and *Blepsimolge* diverged roughly 10 MYA, our results indicate that the hybridization between some members of *Blepsimolge* and *Typhlomolge* may represent the most divergent plethodontid species living in sympatry that are known to be able to produce viable hybrids. Our results further highlight the ability of distantly related amphibian taxa to successfully hybridize in nature.

Two individuals of *Eurycea sosorum* (*Blepsimolge* clade) from Barton Springs (one of which, a captive, wild-caught specimen, shows a morphology intermediate between that of *E. sosorum* and *E. waterlooensis* and about 25% nuclear *E. waterlooensis* ancestry based on RAD data)

possess the main *E. waterlooensis* mt haplotype. Conversely, one *E. waterlooensis* possesses the predominant *E. sosorum* haplotype. Both species are essentially sympatric at Barton Springs; *E. sosorum* occurs primarily on the surface but also uses subterranean habitat (e.g., Chippindale et al., 1993). *Eurycea waterlooensis* (*Typhlomolge* clade) is almost exclusively subterranean (e.g., Hillis et al., 2001). Despite approximately 10 MY of divergence, it appears that hybridization between the two is not uncommon, although species integrity is clearly maintained.

The overall picture is one of extensive mitochondrial introgression across species boundaries that in some cases are ancient, although most species (based on nuclear data, and morphological analyses from other studies) appear to maintain distinct identities.

4.3 Phylogenetic considerations

Fully resolving species boundaries in central Texas *Eurycea* still remains difficult task given the dynamic nature of the hydrology of the Edwards Aquifer and the inaccessibility of much of the habitat of these salamanders. Thanks to considerable efforts on behalf of city, state, and federal agencies many new populations of central Texas *Eurycea* have been discovered, and existing populations/species are better understood. Here, we find several notable taxonomic conclusions. Within the *Blepsimolge* clade, our results indicate that *E.* sp. Hope springs and *E.* sp. Roy Creek Spring (*E.* sp. Pedernales) may represent a distinct species (also noted by Devitt et al., 2019). We also find strong evidence that *E. nana* is a distinct species and does not show evidence of ongoing hybridization with any population or species represented in our data (consistent with even relatively "primitive" allozyme data presented by Chippindale et al., (1998); this is a unique species with an extremely limited geographic range that shows high levels of nuclear divergence from all other Texas *Eurycea*). . Within the *Typhlomolge* clade, it seems apparent that *E.* sp. New

Braunfels represents a distinct species, sister to *E. rathbuni* but exhibiting considerable levels of hybridization with a member of the *Blepsimolge* clade, despite roughly 10 MY of divergence. It has been suggested (informally) that *E.* sp. New Braunfels may simply represent a range extension of *E. rathbuni.* We find that *E.* sp. New Braunfels is not only a distinct species (Fig. 7), but it also apparently readily hybridizes with *E. latitans* (Fig. 3). The full extent of this hybridization is still not fully known due to limited access to sampling sites throughout the region and very limited sampling of the species.

The relationships inferred by our phylogeny (including all samples) disagree with the currently recognized taxonomy, and displays a laddered tree topology, likely caused by admixture between *E. latitans* and *E.* sp. New Braunfels. It is well known that evolutionary processes such as incomplete lineage sorting, admixture, and morphological convergence can mislead phylogenetic inferences (Bendik et al., 2013; Wiens et al., 2003). Notably, admixture can cause distortions of true evolutionary relationships as reflected in phylogenies (Baroni et al., 2006), as reflected by relationships within our *E. neotenes* complex.

Recently, (Devitt et al., 2019) performed a robust population genomic analysis and delimited primarily within *Septentriomolge* (northern clade) and western *Blepsimolge* (*E. troglodytes* complex) and found strong support for several cryptic species. They also suggested several notable taxonomic changes within the *E. neotenes* complex. Specifically, they recommended that that *E. tridentifera* be subsumed into *E. latitans* (in agreement with Bendik, et al., 2013), and that surface salamanders at Comal Springs be considered *E. pterophila* (previously considered *E. nana* by Sweet (1982), but *E. neotenes* by Chippindale et al. (2000), Lucas et al. (2008), and Bendik et al. (2013)).

We concur with Devitt et al. (2019) in recognizing populations at Upper Taylor Spring, Spillar Ranch Spring, Blowing Sink Cave, Cold Spring, and Zara Monitoring Well as *E. sosorum*. This is supported by our phylogenetic inferences, (Fig. 6), lower F_{ST} compared to nearby *E. nana* (Fig. 4), and our DAPC analysis (Fig. 2).

4.4 Conservation Implications

Salamanders living in the Edwards Aquifer face a multitude of threats to the quality and quantity of groundwater on which they depend (Bendik et al., 2013; Burri et al., 2019; Chippindale and Price, 2005; Chippindale and Wiens, 2005), and systems such as the Edwards Aquifer system are intrinsically vulnerable to water contamination due to thin soils, focused recharge, and rapid water flow paths (White, 1988). In addition, central Texas is one of the fastest growing regions in the USA. Increasing demands on groundwater resources and urbanization have had a negative impact on groundwater quality and quantity throughout the region (Musgrove et al., 2019). For example, at Comal Springs, researchers have found a doubling of NO_3-N levels in recent decades (Musgrove et al., 2016). Also, studies of salamander abundance in northern species of Texas *Eurycea* found that increasing levels of urbanization negatively impact salamander abundance (Bendik et al., 2014; Bowles and Arsuffi, 1993; Devitt et al., 2019). Because these salamanders face a multitude of threats, understanding population connectivity, species boundaries, and gene flow is crucial for effective species management.

Hybridization between species of conservation concern presents difficult practical issues. On the one hand, hybridization between species may dilute the 'purity' of a species. Alternatively, natural hybridization may increase genetic diversity by introducing novel alleles, which may increase the resilience of populations. Here, we find support for recognition of an undescribed

species of blind salamander (clade *Typhlomolge*) found at four localities in New Braunfels area (Comal Springs, Mission Valley Bowling Club Well, Hueco Springs, and Panther Canyon Monitoring Well), Texas. Since the discovery of the first specimens of this putatively undescribed species in 2003, there have been sporadic efforts by researchers to collect specimens at these locations. Despite well over a decade of trapping (including bi-weekly sampling attempts for over a year), only seven specimens have been collected. Given the apparently miniscule range of this putative species and the plethora of threats to the habitat and health of the salamanders, we believe this population group deserves immediate conservation attention. A formal description is in preparation by several of the authors of this paper.

4.5 Conclusions

The Edwards Aquifer terminates at the Barton Springs system in Austin, TX, where *E. sosorum* and *E. waterlooensis* live in sympatry. Our results indicate that gene flow generally follows the flow of groundwater, and that *E. sosorum* at Barton Springs seems to be highly variable, likely due to gene flow from upstream populations of *E. sosorum* and from natural hybridization with *E. waterlooensis.* This is consistent with previous morphological studies and is consistent with the idea that genetic variation may be accumulating at the end of the aquifer in this species. We also find evidence of a previously undescribed species of *Typhlomolge* in the New Braunfels area that readily hybridizes with *E. latitans,* however the nature of this hybridization is less clear and warrants further research. Overall, our analyses and approach highlight the power of using multiple genetic datasets (i.e., mitochondrial and nuclear) when disentangling relationships and patterns of gene flow in natural populations.

REFERENCES

Abbott, R.J., Barton, N.H., Good, J.M., 2016. Genomics of hybridization and its evolutionary consequences. Mol. Ecol. 25, 2325–2332. doi:10.1111/mec.13685

Ayala, F.J., Tracey, M.L., Hedgecock, D., Richmond, R.C., 1974. Genetic differentiation during the speciation process in *Drosophila*. Evolution (N.Y). 28, 576–592. doi:10.2307/2407283

Baroni, M., Semple, C., Steel, M., 2006. Hybrids in real time. Syst. Biol. 55, 46–56. doi:10.1080/10635150500431197

Bendik, N.F., Meik, J.M., Gluesenkamp, A.G., Roelke, C.E., Chippindale, P.T., 2013. Biogeography, phylogeny, and morphological evolution of central Texas cave and spring salamanders. BMC Evol. Biol. 13, 1–18.

Bendik, N.F., Sissel, B.N., Fields, J.R., O'Donnell, L.J., Sanders, M.S., 2014. Effect of urbanization on abundance of jollyville plateau salamanders (*Eurycea tonkawae*). Herpetol. Conserv. Biol. 9, 206–222.

Bonett, R.M., Steffen, M.A., Lambert, S.M., Wiens, J.J., Chippindale, P.T., 2013. Evolution of paedomorphosis in plethodontid salamanders: Ecological correlates and re-evolution of metamorphosis. Evolution (N.Y). doi:10.1111/evo.12274

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T.G., Wu, C.-H., Xie, D., Suchard, M. a, Rambaut, A., Drummond, A.J., 2013. Beast2: A software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 10, 1003537. doi:10.1371/journal.pcbi.1003537

Bouckaert, R.R., 2010. DensiTree: Making sense of sets of phylogenetic trees. Bioinformatics 26, 1372–1373. doi:10.1093/bioinformatics/btq110

Bowles, D.E., Arsuffi, T.L., 1993. Karst aquatic ecosystems of the Edwards Plateau region of central Texas, USA: A consideration of their importance, threats to their existence, and efforts for their conservation. Aquat. Conserv. Mar. Freshw. Ecosyst. 3, 317–329. doi:10.1002/aqc.3270030406

Brennan, I.G., Bauer, A.M., Jackman, T.R., 2016. Mitochondrial introgression via ancient hybridization, and systematics of the Australian endemic pygopodid gecko genus *Delma*. Mol. Phylogenet. Evol. 94, 577–590. doi:10.1016/j.ympev.2015.10.005

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N.A., Roychoudhury, A., 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29, 1917–1932. doi:10.1093/molbev/mss086

Burri, N.M., Weatherl, R., Moeck, C., Schirmer, M., 2019. A review of threats to groundwater quality in the anthropocene. Sci. Total Environ. 684, 136–154. doi:10.1016/j.scitotenv.2019.05.236

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., Cresko, W.A., 2013. Stacks: An analysis tool set for population genomics. Mol. Ecol. 22, 3124–3140. doi:10.1111/mec.12354

Cavender-Bares, J., González-Rodríguez, A., Eaton, D.A.R., Hipp, A.A.L., Beulke, A., Manos, P.S., 2015. Phylogeny and biogeography of the American live oaks (*Quercus* subsection *Virentes*): a genomic and population genetics approach. Mol. Ecol. 24, 3668–3687. doi:10.1111/mec.13269

Charlesworth, D., Charlesworth, B., 1987. Inbreeding depression and its evolutionary consequences. Annu. Rev. Ecol. Syst. 18, 237–268. doi:10.1146/annurev.es.18.110187.001321

Chippindale, P.T., 2009. Population genetic analysis of the Texas blind salamander, *Eurycea rathbuni*. Endangered Species Program for. TPWD Final Report, 1–26.

Chippindale, P.T., Price, a H., Hillis, D.M., 1998. Systematic status of the San Marcos salamander, *Eurycea nana* (Caudata: Plethodontidae). Copeia 1998, 1046–1049.

Chippindale, P.T., Price, A.H., 2005. Diversity and conservation of plethodontid salamanders of the genus *Eurycea* in the Edwards Plateau region of central Texas, in: Lannoo, M. (Ed.), Status and Conservation of North American Amphibians Vol. 1. University of California Press.

Chippindale, P.T., Price, A.H., Hillis, D.M., 1993. A new species of perennibranchiate salamander (*Eurycea*: Plethodontidae) from Austin, Texas. Herpetologica 49, 248–259.

Chippindale, P.T., Price, A.H., Wiens, J.J., Hillis, D.M., 2000. Phylogenetic relationships and systematic revision of central Texas Hemidactyliine plethodontid salamanders. Herpetol. Monogr. 14, 1–80.

Chippindale, P.T., Wiens, J.J., 2005. Re-evolution of the larval stage in the plethodontid salamander genus *Desmognathus*. Herpetol. Rev. 36, 113–117.

Coyne, J.A., Orr, H.A., 2004. Speciation. Sunderland, MA.

Dawley, E.M., 2001. The biology of plethodontid salamanders. Copeia. doi:10.1643/0045-8511(2001)001[1162:]2.0.CO;2

Devitt, T.J., Wright, A.M., Cannatella, D.C., Hillis, D.M., 2019. Species delimitation in endangered groundwater salamanders: Implications for aquifer management and biodiversity conservation. Proc. Natl. Acad. Sci. U. S. A. 116, 2624–2633. doi:10.1073/pnas.1815014116

Dray, S., Dufour, A.B., Chessel, D., 2007. The ade4 Package — II : Two-table and K -table Methods. R News 7, 47–52. doi:10.1159/000323281

Duda, M., 2008. First record of a natural male hybrid of *Bufo* (*Pseudepidalea*) *viridis* LAURENTI, 1768 and *Bufo* (*Bufo*) *bufu* LINNEUS, 1758 from Austria. Herpetoozoa 20, 184–186.

Earl, D.A., vonHoldt, B.M., 2012. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. 4, 359–361. doi:10.1007/s12686-011-9548-7

Eaton, D.A.R., 2014. PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30, 1844–1849. doi:10.1093/bioinformatics/btu121

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113. doi:10.1186/1471-2105-5-113

Edmands, S., 2002. Does parental divergence predict reproductive compatibility? Trends Ecol. Evol. 17, 520–527. doi:10.1016/S0169-5347(02)02585-5

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. doi:10.1111/j.1365-294X.2005.02553.x

Fontenot, B.E., Makowsky, R., Chippindale, P.T., 2011. Nuclear-mitochondrial discordance and gene flow in a recent radiation of toads. Mol. Phylogenet. Evol. 59, 66–80. doi:10.1016/j.ympev.2010.12.018

Gompert, Z., Forister, M.L., Fordyce, J. a., Nice, C.C., 2008. Widespread mito-nuclear discordance with evidence for introgressive hybridization and selective sweeps in *Lycaeides*. Mol. Ecol. 17, 5231–5244. doi:10.1111/j.1365-294X.2008.03988.x

HannonLab, 2014. FASTX toolkit. Cold Spring Harb. Lab. Cold Spring Harb. NY. URL http://hannonlab.cshl.edu/fastx_toolkit/

Hauwert, N., Johns, D., Hunt, J., 2004. Flow systems of the Edwards Aquifer Barton Springs Segment interpreted from tracing and associated field studies. Edwards Water Resources. Central Texas, South Texas Geol. Soc. Austin Geol. Soc.

Hauwert, N.M., 2016. Stream recharge water balance for the Barton Springs Segment of the Edwards Aquifer. J. Contemp. Water Res. Educ. 24–49. doi:10.1111/j.1936-704x.2016.03228.x

Hillis, D.M., Chamberlain, D.A., Wilcox, T.P., Chippindale, P.T., 2001. A new species of subterranean blind salamander (Plethodontidae: Hemidactyliini: *Eurycea*: *Typhlomolge*) from Austin, Texas, and a systematic revision of central Texas paedomorphic salamanders. Herpetologica 57, 266–280.

Hillis, D.M., Mable, B.K., Larson, A., Davis, S.K., Zimmer, A., 1996. Nucleic acids IV: sequencing and cloning, in: Hillis, D., Moritz, C., Mable, B. (Eds.), Molecular Systematics. Sunderland: Sinauer Associates, pp. 321–381.

Johnson, S., Schindel, G., Veni, G., Hauwert, N., Hunt, B., Smith, B., Gary, M., 2012. Tracing groundwater flowpaths in the vicinity of San Marcos Springs, Texas, Edwards Aquifer Authority. 12-01, 82-85.

Jombart, T., 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24, 1403–5. doi:10.1093/bioinformatics/btn129

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795. doi:10.2307/2291091

Keenan, K., Mcginnity, P., Cross, T.F., Crozier, W.W., Prodöhl, P.A., 2013. DiveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. Methods Ecol. Evol. 4, 782–788. doi:10.1111/2041-210X.12067

Leaché, A.D., Fujita, M.K., Minin, V.N., Bouckaert, R.R., 2014. Species delimitation using genome-wide SNP Data. Syst. Biol. 63, 534–542. doi:10.1093/sysbio/syu018

Leigh, J.W., Bryant, D., 2015. POPART: Full-feature software for haplotype network construction. Methods Ecol. Evol. 6, 1110–1116. doi:10.1111/2041-210X.12410

Lippman, Z.B., Zamir, D., 2007. Heterosis: revisiting the magic. Trends Genet. 23(2), 60-66. doi:10.1016/j.tig.2006.12.006

Lucas, L.K., Gompert, Z., Ott, J.R., Nice, C.C., 2008. Geographic and genetic isolation in spring-associated *Eurycea* salamanders endemic to the Edwards Plateau region of Texas. Conserv. Genet. 10, 1309–1319. doi:10.1007/s10592-008-9710-2

Mallet, J., 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20(5), 229-237. doi:10.1016/j.tree.2005.02.010

McGaugh, S.E., Weaver, S., Gilbertson, E.N., Garrett, B., Rudeen, M.L., Grieb, S., Roberts, J., Donny, A., Marchetto, P., Gluesenkamp, A.G., 2020. Evidence for rapid phenotypic and behavioural shifts in a recently established cavefish population. Biol. J. Linn. Soc. 129, 143–161. doi:10.1093/biolinnean/blz162

Mitchell, R.., Reddell, J.R., 1965. *Eurycea tridentifera*, a new species of troglobitic salamander from Texas and a reclassification of *Typhlomolge rathbuni*. Texas J. Sci. 12–27.

Mitchell, R.W., Smith, R.E., 1972. Some aspects of the osteology and evolution of the neotenic spring and cave salamanders (*Eurycea*), Plethodontidae) of central Texas. Texas J. Sci. 343–362.

Musgrove, M., Opsahl, S.P., Mahler, B.J., Herrington, C., Sample, T.L., Banta, J.R., 2016. Source, variability, and transformation of nitrate in a regional karst aquifer: Edwards aquifer, central Texas. Sci. Total Environ. 568, 457–469. doi:10.1016/j.scitotenv.2016.05.201

Musgrove, M., Solder, J.E., Opsahl, S.P., Wilson, J.T., 2019. Timescales of water-quality change in a karst aquifer, south-central Texas. J. Hydrol. X. 4, 100041. doi:10.1016/j.hydroa.2019.100041

Pante, E., Abdelkrim, J., Viricel, A., Gey, D., France, S.C., Boisselier, M.C., Samadi, S., 2015. Use of RAD sequencing for delimiting species. Heredity (Edinb). 114, 450–459. doi:10.1038/hdy.2014.105

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7, e37135. doi:10.1371/journal.pone.0037135

Pickrell, J.K., Pritchard, J.K., 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8, e1002967. doi:10.1371/journal.pgen.1002967

Portik, D.M., Papenfuss, T.J., 2015. Historical biogeography resolves the origins of endemic Arabian toad lineages (*Anura: Bufonidae*): Evidence for ancient vicariance and dispersal events with the Horn of Africa and South Asia. BMC Evol. Biol. 15, 152.doi:10.1186/s12862-015-0417-y

Potter, F.E., Sweet, S.S., 1981. Generic boundaries in Texas cave salamanders, and a redescription of *Typhlomolge robusta* (Amphibia: Plethodontidae). Copeia 64–75. doi:10.2307/1444041

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959. doi:10.1111/j.1471-8286.2007.01758.x

R Core team, 2015. R: A language and environment for statistical computing. R. Foundation for Statistical Computing. Vienna, Austria. URL http//www.R-project.org/.

Sasa, M.M., Chippindale, P.T., Johnson, N.A., 1998. Patterns of postzygotic isolation in frogs. Evolution (N.Y). 52, 1811–1820. doi:10.2307/2411351

Shull, G.H., 1908. The composition of a field of maize. J. Hered. 4, 296. doi:10.1093/jhered/os-4.1.294

Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. doi:10.1093/bioinformatics/btu033

Stejneger, L., 1896. Description of a new genus and species of blind tailed batrachians from the subterranean waters of Texas. Proc. Natl. Museum 18, 619–621.

Sweet, S.S., 1982. A Distributional analysis of epigean populations of *Eurycea neotenes* in central Texas, with comments on the origin of troglobitic populations. Herpetologica 38, 430–444.

Sweet, S.S., 1977. Natural metamorphosis in *Eurycea neotenes*, and the generic allocation of the Texas Eurycea 33, 364–375.

Toews, D.P.L., Brelsford, A., 2012. The biogeography of mitochondrial and nuclear discordance in animals. Mol. Ecol. 21, 3907–30. doi:10.1111/j.1365-294X.2012.05664.x

Wake, D.., 1966. Comparative osteology and evolution of the lungless salamanders, family Plethodontidae. Mem. South Calif. Aca. Sci. 1–111.

Walsh, P.S., Metzger, D.A., Higuchi, R., 1991. Chelex-100 as a medium for simple extraction of DNA for PCR- based typing from forensic material. Biotechniques 10, 506–513. doi:10.2144/000113897

White, W.B., 1988. Geomorphology and hydrology of karst terrains, 1st ed. Oxford University Press. doi:10.5860/choice.26-2715

Wiens, J.J., Chippindale, P.T., Hillis, D.M., 2003. When are phylogenetic analyses misled by convergence? A case study in Texas cave salamanders. Syst. Biol. 52, 501–514. doi:10.1080/10635150390218222

Wiens, J.J., Engstrom, T.N., Chippindale, P.T., 2006. Rapid diversification, incomplete isolation, and the "speciation clock" in North American salamanders (Genus *Plethodon*): testing the hybrid swarm hypothesis of rapid radiation. Evolution 60, 2585–2603. doi:10.1111/j.0014-3820.2006.tb01892.x

TABLES AND FIGURES

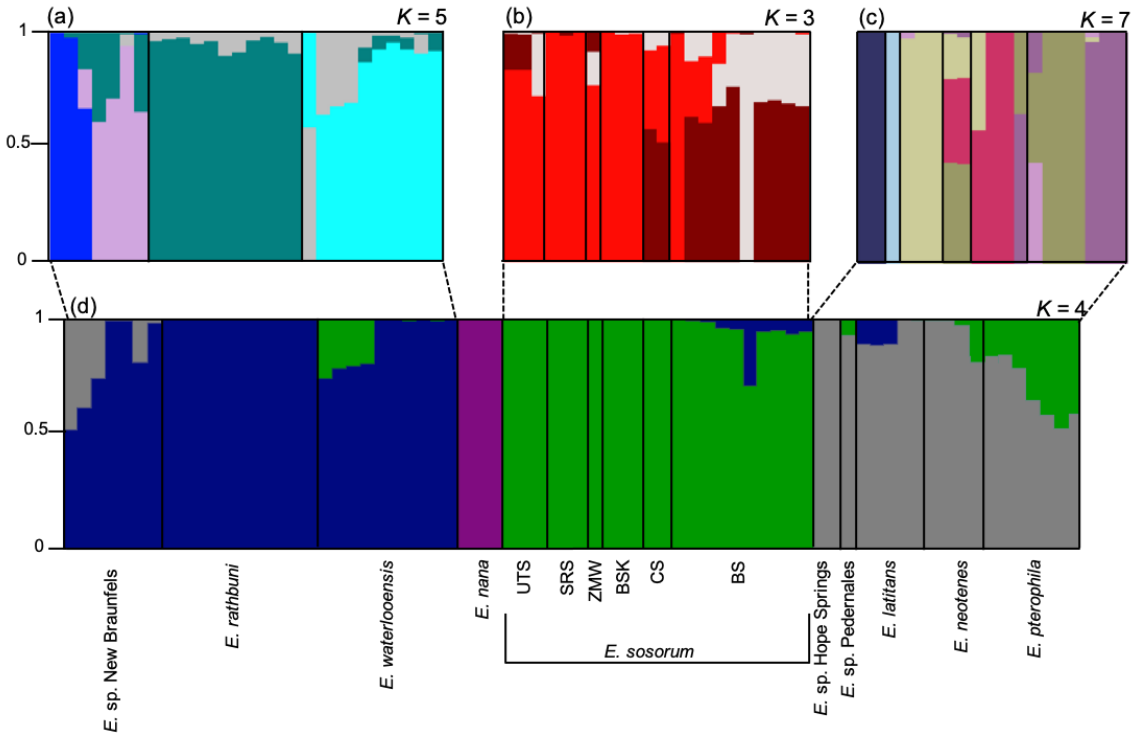


Figure 1. Results from four independent STRUCTURE analyses showing ancestry proportions for all 72 samples. We show the plots corresponding to the optimal value of *K* as determined by the Δ*K* method (See section 2.3). Abbreviated localities of *E. sosorum* are as follows: UTS = Upper Taylor Springs, SRS = Spillar Ranch Spring, SMZ = Zara Monitoring Well, BSK = Blowing Sink Cave, CS = Cold Spring, BS = Barton Springs). (a) *Typhlomolge* clade (K = 5); (b) *E. sosorum* and nearby newly discovered populations (*K* = 3); (c) *E. neotenes* complex (*K* = 7); (d) all samples (*K* = 4).

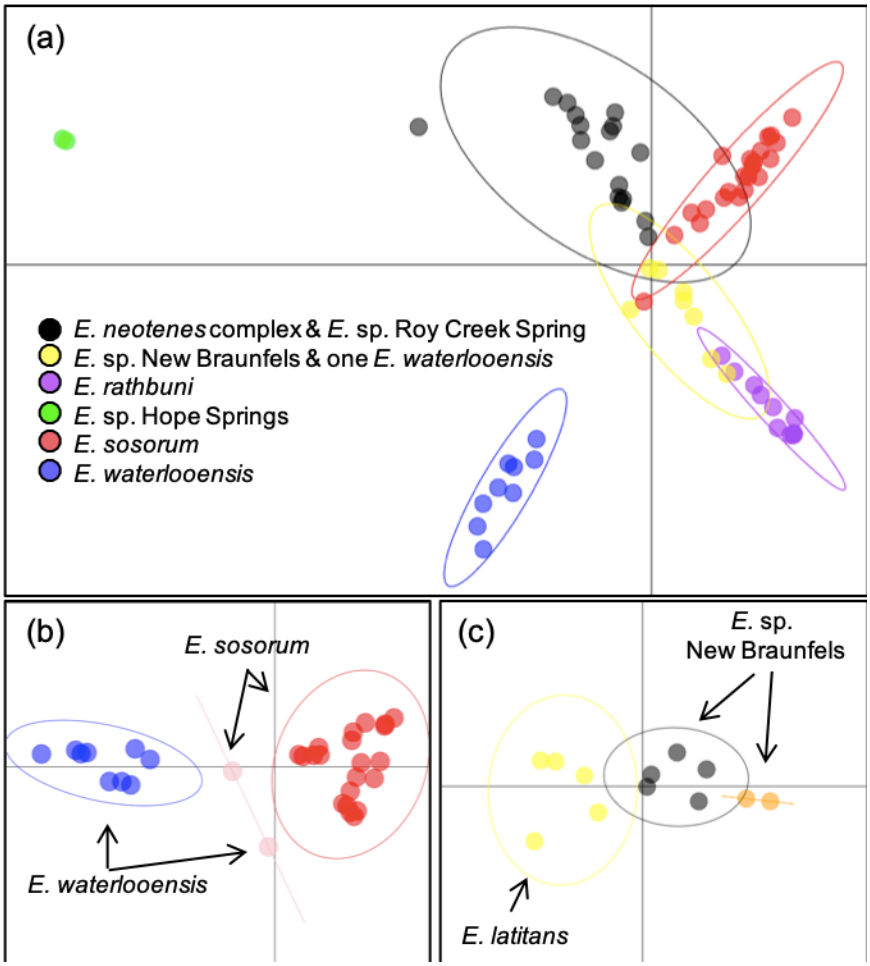


Figure 2. Discriminant analysis of principal components (DAPC). (a) All samples excluding *E. nana* (K = 6). (b) *E. sosorum* and *E. waterlooensis* (K = 3) (c) *E.* sp. New Braunfels and *E. latitans* (K = 3). The one *E. waterlooensis* in the *E.* sp. New Braunfels group (yellow) of panel (a) corresponds to the sample showing ~25% ancestry from *E. sosorum* as determined by our STRUCTURE analysis (Fig. 1d) and is one of the two samples in the intermediate group (pink) of panel (b). The second pink sample in panel b is a *E. sosorum* that shows ~25% ancestry in our STRUCTURE plot (Fig. 1d). The intermediate samples in panel c (black) show evidence of co-ancestry with *E. latitans* in our STRUCTURE analysis (Fig. 1d).

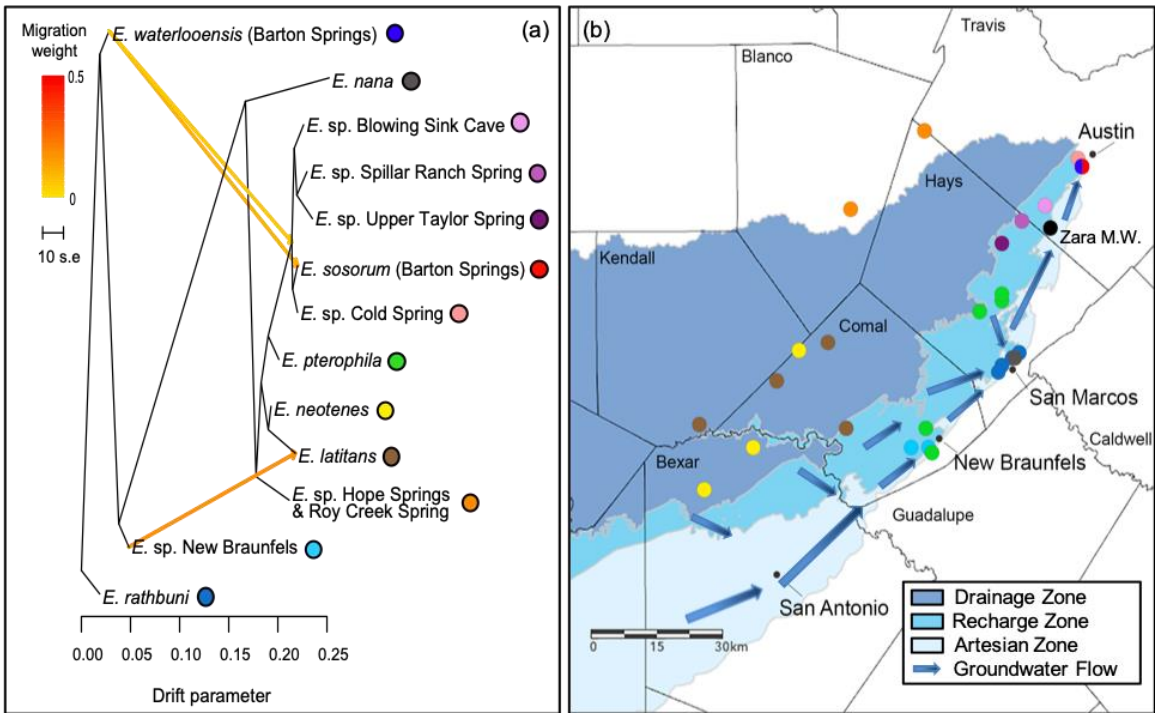


Figure 3. Treemix plot and sampling localities. (a) Phylogeny of populations/species, where all tips contain $n \geq 2$ individuals, and $M = 3$. The first migration is from *E. latitans* to *E.* sp. New Braunfels, and the second two migration arrows are from *E. waterlooensis* to *E. sosorum*. (b) Shows the geographic distribution of the populations represented in panel (a), also showing Zara Monitoring Well (Zara M.W.) that is not represented in Treemix due to having only one sample from this locality.

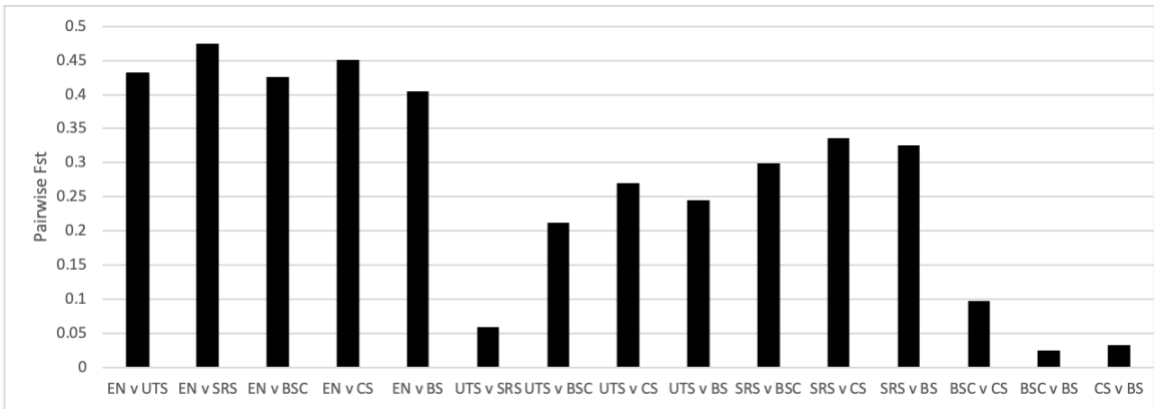


Figure 4. Pairwise Wier & Cockerham F_{ST} estimates between *E.* nana and *E. sosorum* using the diffCalc function within the diveRsity package in R. EN = *E. nana*, UTS = Upper Taylor Spring, SRS = Spillar Ranch Spring, BSC = Blowing Sink Cave, CS = Cold Spring, BS = Barton Springs. All comparisons between *E. nana* and *E. sosorum* are higher than comparisons within *E. sosorum* populations.

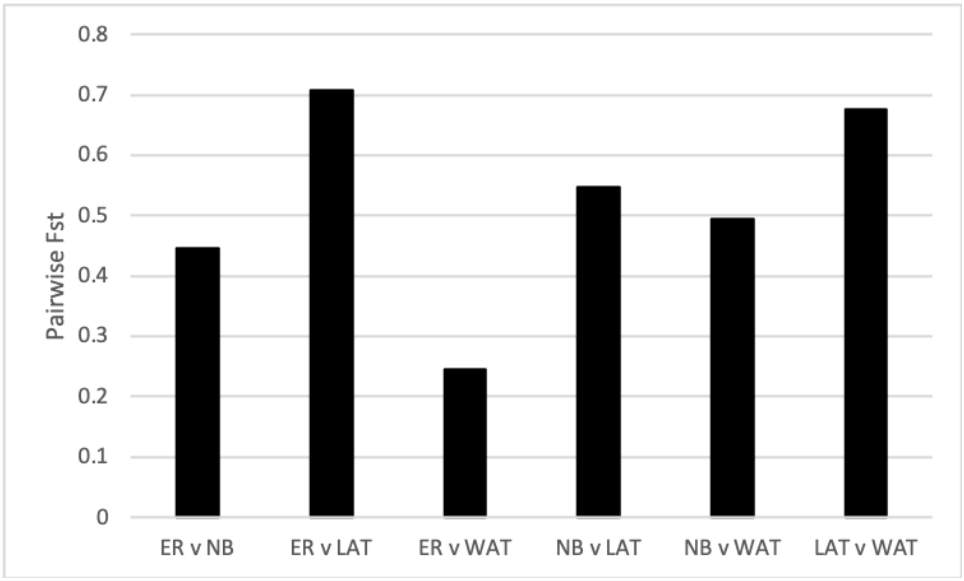


Figure 5. Pairwise Wier & Cockerham FST estimates of *E. ratubuni*, *E. latitans*, *E. waterlooensis* and *E.* sp. New Braunfels, using the diffCalc function within the diveRsity package in R. 100 bootstrap replicates were used for the calculation. ER = *E. rathbuni*, NB = *E.* sp. New Braunfels, LAT = *E. latitans*, WAT = *E. waterlooensis*.

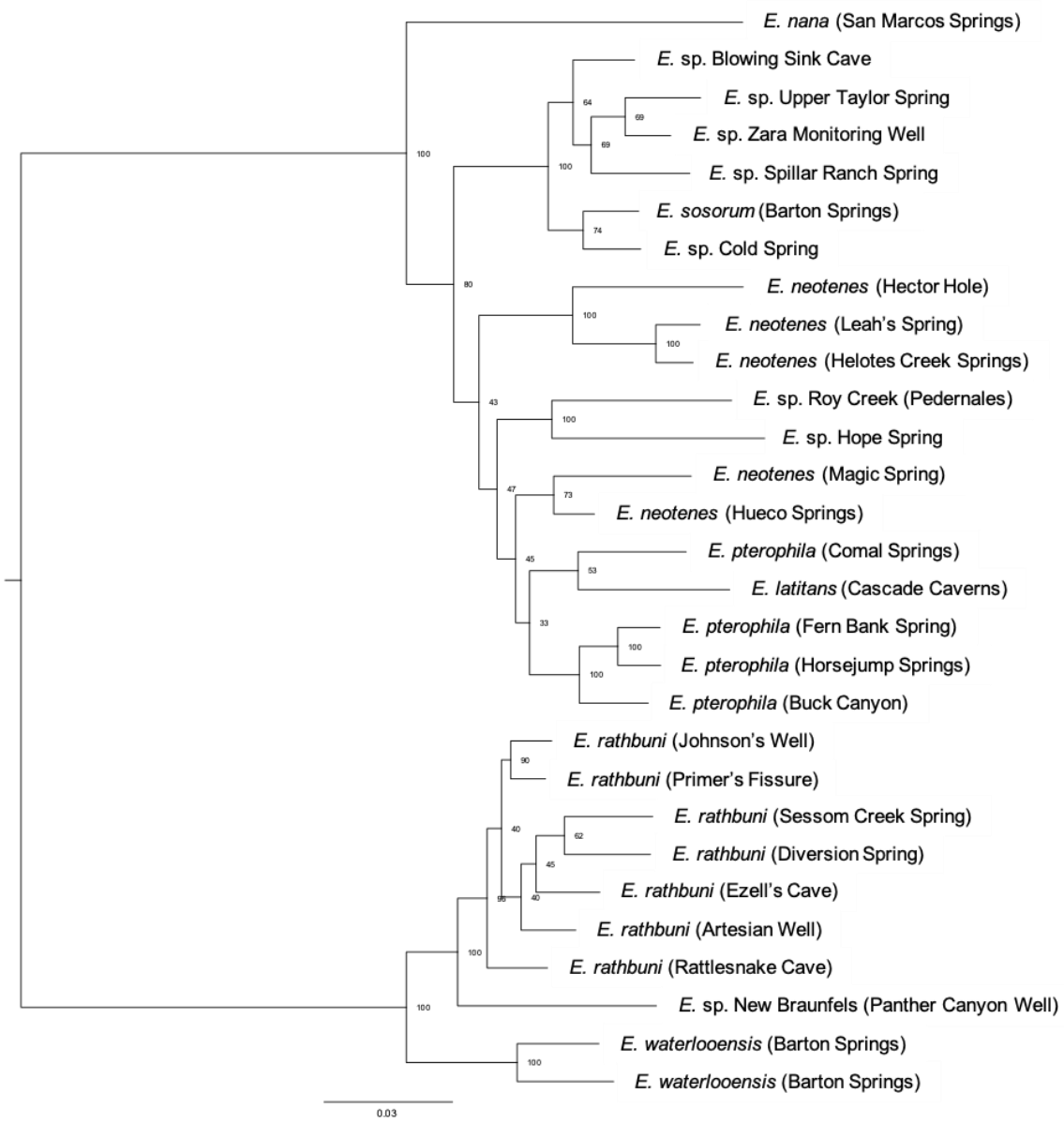
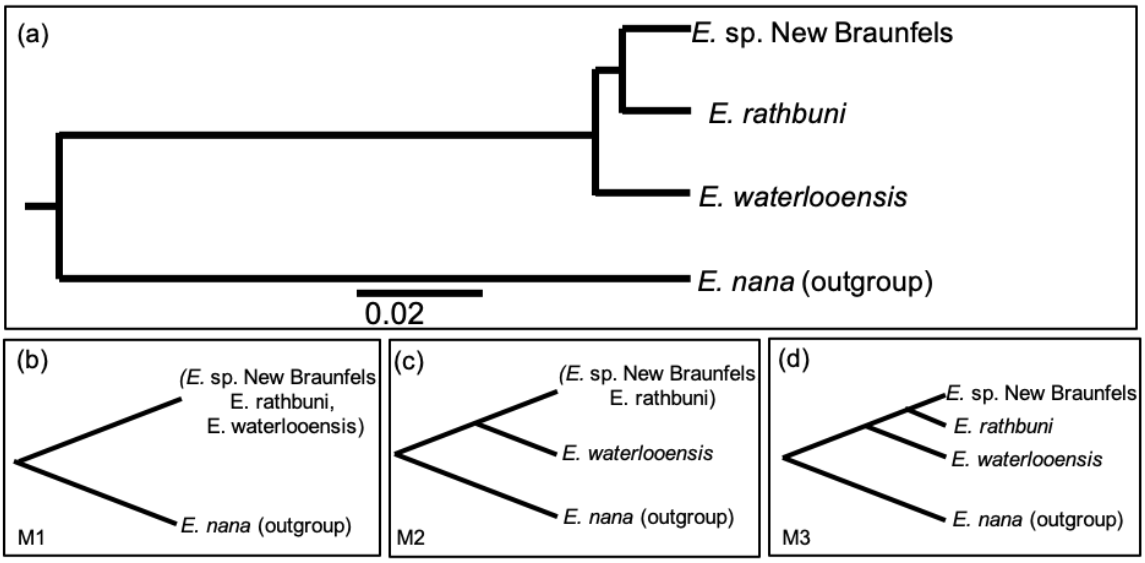


Figure 6. Maximum likelihood tree generated in RAxML including 29 individuals from our dataset, including at least one exemplar from each population or sampling locality. We did not include any samples showing evidence of admixture based on our STRUCTURE analysis (Fig. 1) or our DAPC analysis (Fig. 2). Nodes are numbered with bootstrap support generated using 1000 bootstrap replicates and the GTR_GAMMA model of nucleotide evolution. Scale bar indicates substitutions per site.



Model number	Number of Species in Tree	Rank	Marginal Likelihood	Bayes Factor
M1	2	3	-33341	24700
M2	3	2	-30314	18646
M3	**4**	**1**	**-20991**	**N/A**

Figure 7. Bayesian species delimitation for *Typhlomolge* with *E. nana* (*Blepsimolge*) used as an outgroup. The three species model with the highest marginal likelihood (M3) was best supported, indicating that *E.* sp. New Braunfels is a distinct species sister to *E. rathbuni.* (A) Species tree of the best supported model (M3) depicting that *E.* sp. New Braunfels is sister to *E. rathbuni* with branch lengths drawn to scale according to scale bar (expected number of mutations per site). (B) Model 1 hypothesis. (C) Model 2 hypothesis. (D) Model 3 hypothesis.

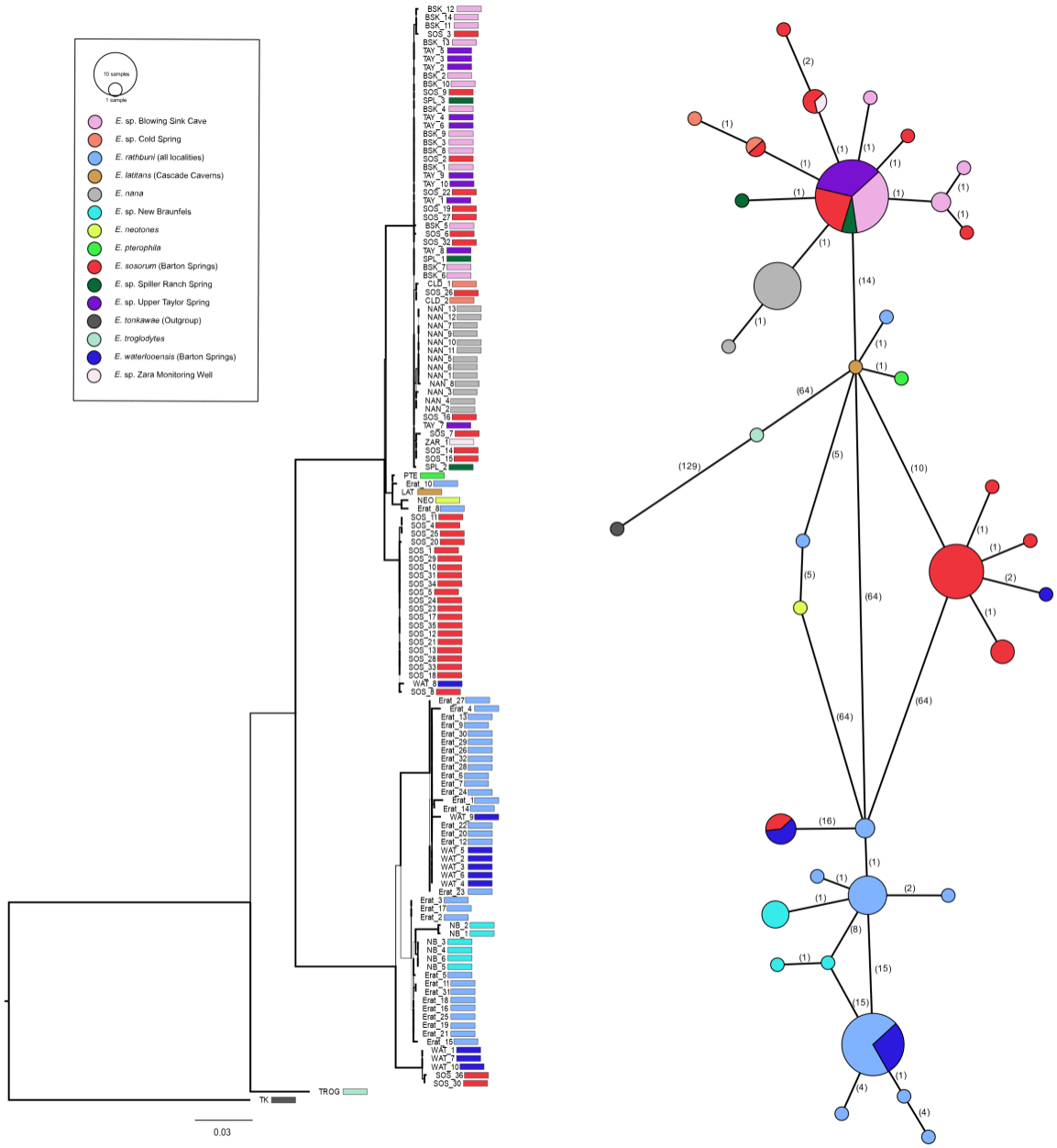
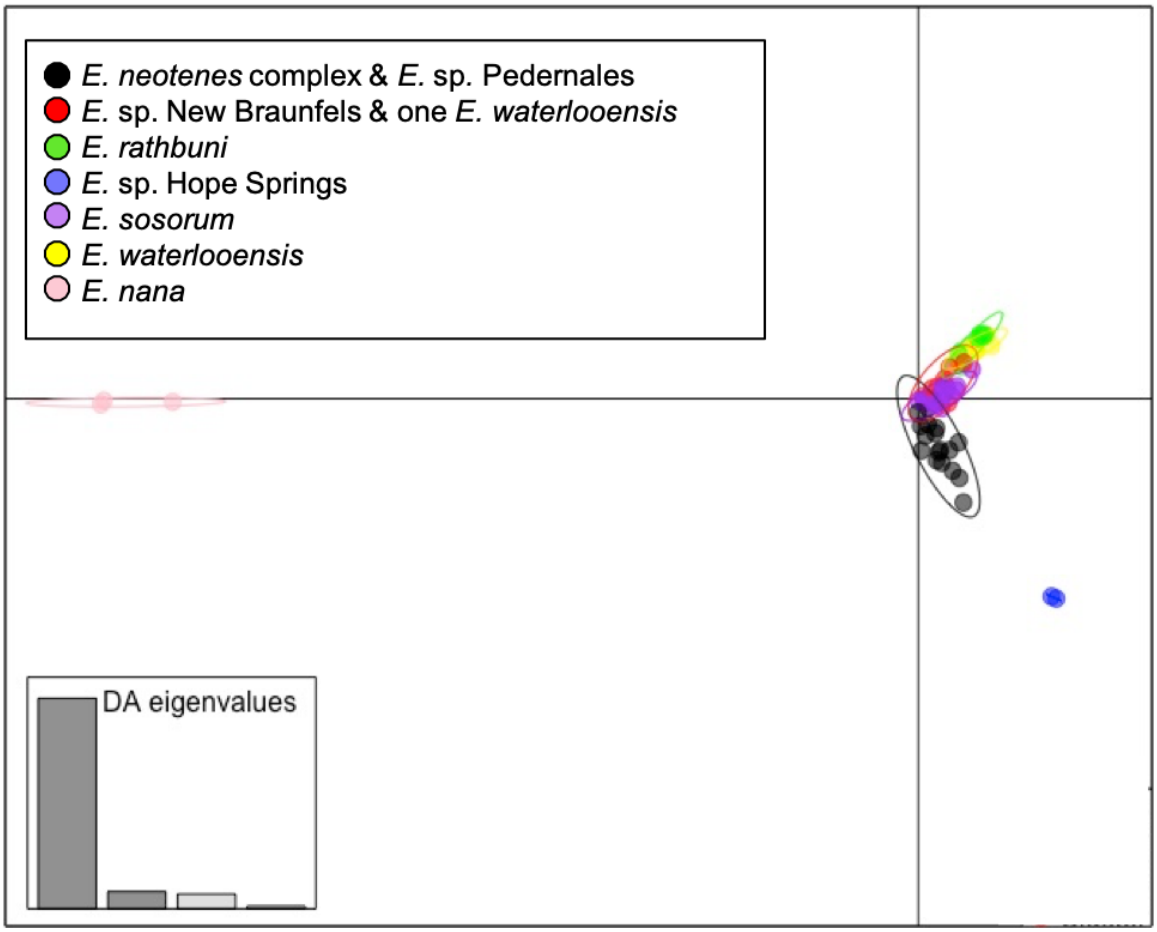
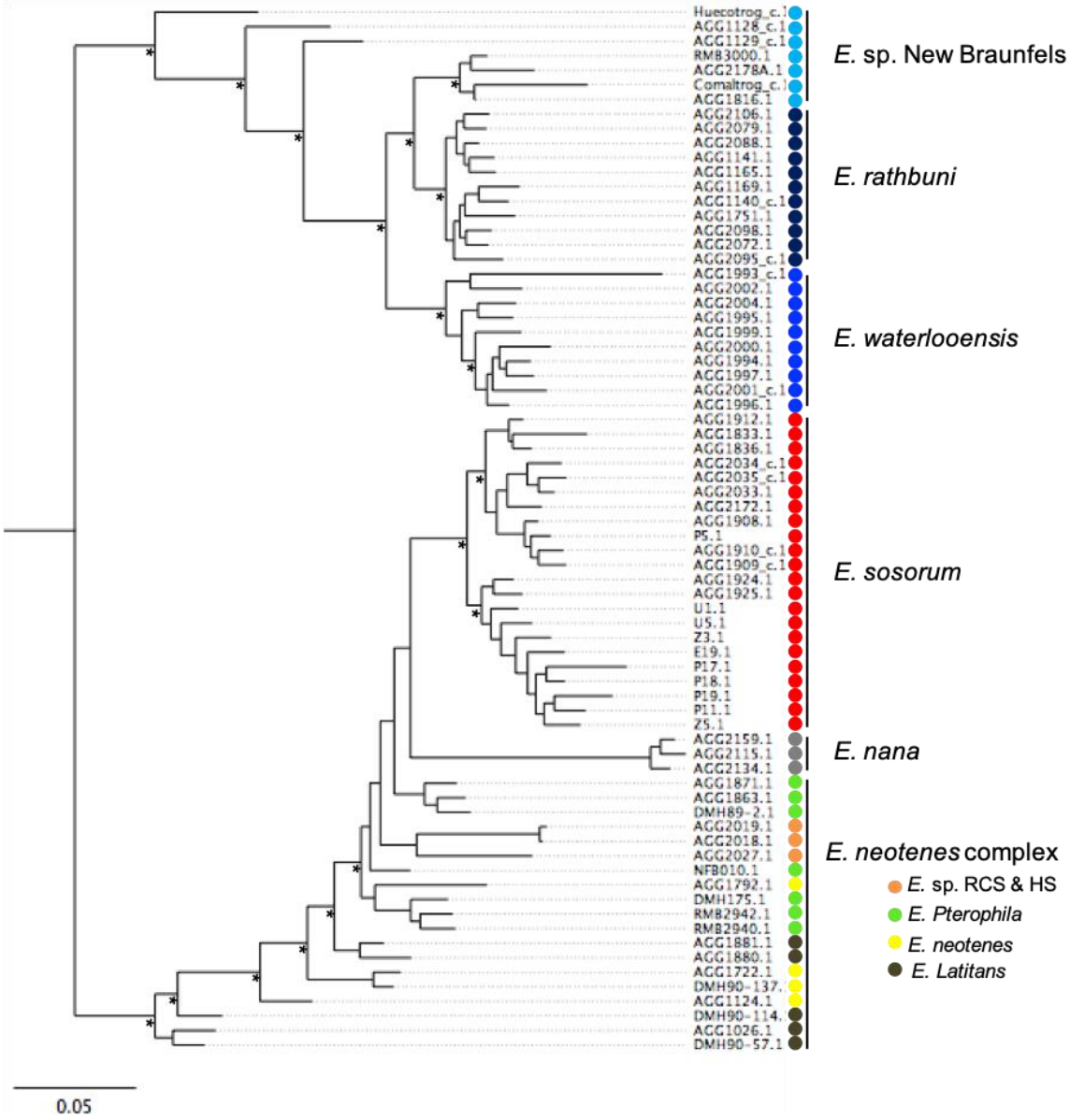


Figure 8. Maximum likelihood tree of 865bp of mitochondrial gene cytochrome-b generated in RAxML and minimum spanning haplotype network generated in PopArt. The RAxML tree was generated with 1000 bootstrap replicates under the GTR_GAMMA model of nucleotide evolution, and lines are weighted by bootstrap support with thicker lines indicating higher bootstrap support. *Eurycea tonkawae* (from the Northern *Septentriomolge* clade) was used as an outgroup. Numbers in parentheses within the haplotype network indicate the number of changes between haplotypes or haplotype groups.

SUPPLEMENTARY FIGURES



Supplementary Figure 1. Discriminant analysis of principal components (DAPC) of all samples where the optimal number of groups was 7. *Eurycea nana* is highly differentiated from all other samples in this analysis, as corroborated by our STRUCTURE analysis (Fig. 1).



Supplementary Figure 2. Maximum likelihood tree generated in RAxML including all 72 individuals from our RADseq dataset. Invariant sites were removed prior to analysis, tree support was assessed via 1000 bootstrap replicates, and we used the GTR_GAMMA model of nucleotide evolution with the Lewis Ascertainment bias correction to account for including only invariable sites. Scale bar indicates substitutions per site. Sample names in the tree correspond to sample names in Supplementary Table 1. Asterisks indicate >95% bootstrap support. Blue dots represent *E. latitans* that appears to share ancestry with *E.* sp. New Braunfels based on our STRUCTURE and DAPC results (Fig. 1 & 2). RCS = Roy Creek Spring, HS = Hope Springs.

SUPPLEMENTARY TABLES

Supplementary Table 1. List of samples included in this study with population/species/sampling locality assignments.

Sample	Species	Sampling locality
AGG1026	*E. latitans*	Preserve Cave
AGG1124	*E. neotenes*	Hector Hole
AGG1128	*E*. sp. New Braunfels	Mission Valley Bowling Club Well
AGG1129	*E*. sp. New Braunfels	Mission Valley Bowling Club Well
AGG1140	*E. rathbuni*	Rattlesnake Cave
AGG1141	*E. rathbuni*	Ezell's Cave
AGG1165	*E. rathbuni*	Rattlesnake Cave
AGG1169	*E. rathbuni*	Ezell's Cave
AGG1722	*E. neotenes*	Leah's Spring
AGG1751	*E. rathbuni*	Sessom Creek Spring
AGG1792	*E. neotenes*	Magic Spring
AGG1816	*E*. sp. New Braunfels	Panther Canyon Well
AGG1833	*E. sosorum*	Blowing Sink Cave
AGG1836	*E. sosorum*	Blowing Sink Cave
AGG1863	*E. pterophila*	Horsejump Springs
AGG1871	*E. pterophila*	Buck Canyon
AGG1880	*E. latitans*	Cascade Caverns
AGG1881	*E. latitans*	Cascade Caverns
AGG1908	*E. sosorum*	Upper Taylor Spring
AGG1909	*E. sosorum*	Upper Taylor Spring
AGG1910	*E. sosorum*	Upper Taylor Spring
AGG1912	*E. sosorum*	Blowing Sink Cave
AGG1924	*E. sosorum*	Cold Spring
AGG1925	*E. sosorum*	Cold Spring
AGG1993	*E. waterlooensis*	Barton Springs
AGG1994	*E. waterlooensis*	Barton Springs
AGG1995	*E. waterlooensis*	Barton Springs
AGG1996	*E. waterlooensis*	Barton Springs
AGG1997	*E. waterlooensis*	Barton Springs
AGG1999	*E. waterlooensis*	Barton Springs
AGG2000	*E. waterlooensis*	Barton Springs
AGG2001	*E. waterlooensis*	Barton Springs
AGG2002	*E. waterlooensis*	Barton Springs
AGG2004	*E. waterlooensis*	Barton Springs

AGG2018	*E.* sp. Hope Springs	Hope Springs
AGG2019	*E*. sp. Hope Springs	Hope Springs
AGG2027	*E.* sp. Pedernales	Roy Creek Spring
AGG2033	*E. sosorum*	Spillar Ranch Spring
AGG2034	*E. sosorum*	Spillar Ranch Spring
AGG2035	*E. sosorum*	Spillar Ranch Spring
AGG2072	*E. rathbuni*	Rattlesnake Cave
AGG2079	*E. rathbuni*	Johnson's Well
AGG2088	*E. rathbuni*	Primer's Fissure
AGG2095	*E. rathbuni*	Diversion Spring
AGG2098	*E. rathbuni*	Diversion Spring
AGG2106	*E. rathbuni*	Artesian Well
AGG2115	*E. nana*	San Marcos Springs
AGG2134	*E. nana*	San Marcos Springs
AGG2159	*E. nana*	San Marcos Springs
AGG2172	*E. sosorum*	Zara Monitoring Well
AGG2178	*E*. sp. New Braunfels	Panther Canyon Well
Comaltrog	*E*. sp. New Braunfels	Comal Springs
DMH175	*E. pterophila*	Comal Springs
DMH892	*E. pterophila*	Fern Bank Spring
DMH90114	*E. latitans*	Badweather Pit
DMH90137	*E. neotenes*	Helotes Creek Springs
DMH9057	*E. latitans*	Honey Creek Cave
E19	*E. sosorum*	Barton Springs
Huecotrog	*E*. sp. New Braunfels	Hueco Springs
NFB010	*E. pterophila*	Hueco Springs
P11	*E. sosorum*	Barton Springs
P17	*E. sosorum*	Barton Springs
P18	*E. sosorum*	Barton Springs
P19	*E. sosorum*	Barton Springs
P5	*E. sosorum*	Barton Springs
RMB2940	*E. pterophila*	Comal Springs
RMB2942	*E. pterophila*	Comal Springs
RMB3000	*E*. sp. New Braunfels	Panther Canyon
U1	*E. sosorum*	Barton Springs
U5	*E. sosorum*	Barton Springs
Z3	*E. sosorum*	Barton Springs
Z5	*E. sosorum*	Barton Springs

Supplementary Table 2. Primers used for amplification of the mitochondrial cytochrome *b* (cyt-b) gene.

Primer name	Primer sequence (5'-3')
PGLU	GAARAAYCANTRTTGTATTCAAC
PGLU-TAT**	GAARAAYCANTRTTGTATTCAACTAT
MVZ-15	GAACTAATGGCCCACACWWTACGNAA
HEM-CB1-5'	CCATCCAACATCTCAGCATGATGAAA
CYTBTN5Fv2	CATATTTAGGRGAAACACTTGTTCA
CYTBTYPHmR	GTCKGGGYTAGAATTAATTCCTG
EurTXCRThr	GYCAATGTTTTTCTAAACTACAACAGCATC

Supplementary Table 3. Data processing statistics for each 'params file' tested in pyrad.

Metric	Params_A	Params_B	Params_C
Parsimony Informative Sites	14897	3503	1104
Total SNPs	5718	1514	538
Total Unlinked SNPs	63142	21752	8398

Supplementary Table 4. List of datasets and the samples included in specified analyses. Samples demarcated with an x indicates that the sample was included in the analysis. The main STRUCTURE plot (Fig. 1d), initial RAxML tree (Supplementary Fig. 2), and initial DAPC plot (Supplementary Fig. 1) were generated with all individuals. Treemix included all but Zara Well Monitoring Well.

		Dataset									
Sample	Species	STR (a)	STR (b)	STR (c)	DAPC (a)	DAPC (b)	DAPC (c)	BSD	Fst (a)	Fst (b)	RAxML
AGG1026	*E. latitans*			x	x		x		x		
AGG1124	*E. neotenes*			x	x					x	
AGG1128	*E.* sp. New Braunfels	x			x		x		x		
AGG1129	*E.* sp. New Braunfels	x			x		x		x		
AGG1140	*E. rathbuni*	x			x				x		
AGG1141	*E. rathbuni*	x			x				x		x
AGG1165	*E. rathbuni*	x			x				x		x
AGG1169	*E. rathbuni*	x			x				x		
AGG1722	*E. neotenes*			x	x					x	
AGG1751	*E. rathbuni*	x			x				x		x
AGG1792	*E. neotenes*			x	x						
AGG1816	*E.* sp. New Braunfels	x			x		x		x		
AGG1833	*E. sosorum*		x		x	x				x	
AGG1836	*E. sosorum*		x		x	x				x	
AGG1863	*E. pterophila*			x	x						x
AGG1871	*E. pterophila*			x	x						x
AGG1880	*E. latitans*			x	x						x
AGG1881	*E. latitans*			x	x						
AGG1908	*E. sosorum*		x		x	x				x	x
AGG1909	*E. sosorum*		x		x	x				x	

AGG1910	*E. sosorum*		x		x	x				x		
AGG1912	*E. sosorum*		x		x	x				x		x
AGG1924	*E. sosorum*		x		x	x				x		x
AGG1925	*E. sosorum*		x		x	x				x		
AGG1993	*E. waterlooensis*	x			x	x			x			
AGG1994	*E. waterlooensis*	x			x	x		x	x			
AGG1995	*E. waterlooensis*	x			x	x			x			
AGG1996	*E. waterlooensis*	x			x	x		x	x			
AGG1997	*E. waterlooensis*	x			x	x			x			
AGG1999	*E. waterlooensis*	x			x	x		x	x			x
AGG2000	*E. waterlooensis*	x			x	x		x	x			x
AGG2001	*E. waterlooensis*	x			x	x			x			
AGG2002	*E. waterlooensis*	x			x	x			x			
AGG2004	*E. waterlooensis*	x			x	x			x			
AGG2018	*E.* sp. Hope Springs			x	x							
AGG2019	*E.* sp. Hope Springs			x	x							
AGG2027	*E.* sp. Pedernales			x	x							
AGG2033	*E. sosorum*		x		x	x				x		
AGG2034	*E. sosorum*		x		x	x				x		x
AGG2035	*E. sosorum*		x		x	x				x		
AGG2072	*E. rathbuni*	x			x				x			
AGG2079	*E. rathbuni*	x			x			x	x			x
AGG2088	*E. rathbuni*	x			x			x	x			x
AGG2095	*E. rathbuni*	x			x				x			x
AGG2098	*E. rathbuni*	x			x			x	x			
AGG2106	*E. rathbuni*	x			x			x	x			x
AGG2115	*E. nana*							x		x		x
AGG2134	*E. nana*							x		x		
AGG2159	*E. nana*							x		x		

AGG2172	*E. sosorum*		x		x	x						x
AGG2178	*E.* sp. New Braunfels	x			x		x	x	x			x
Comaltrog	*E.* sp. New Braunfels	x			x		x		x			
DMH175	*E. pterophila*			x	x							
DMH892	*E. pterophila*			x	x							x
DMH90114	*E. latitans*			x	x		x		x			
DMH90137	*E. neotenes*			x	x							x
DMH9057	*E. latitans*			x	x		x		x			
E19	*E. sosorum*		x		x	x				x		
Huecotrog	*E.* sp. New Braunfels	x			x		x		x			
NFB010	*E. pterophila*			x	x							
P11	*E. sosorum*		x		x	x				x		
P17	*E. sosorum*		x		x	x				x		
P18	*E. sosorum*		x		x	x				x		
P19	*E. sosorum*		x		x	x				x		
P5	*E. sosorum*		x		x	x				x		
RMB2940	*E. pterophila*			x	x							
RMB2942	*E. pterophila*			x	x							
RMB3000	*E.* sp. New Braunfels	x			x		x	x	x			
U1	*E. sosorum*		x		x	x				x		x
U5	*E. sosorum*		x		x	x				x		
Z3	*E. sosorum*		x		x	x				x		
Z5	*E. sosorum*		x		x	x				x		

Supplementary Table 5. List of PCR primers used during the amplification step of ddRADseq library generation. Primer one is a universal primer used for all sequences.

Index	Sequence
PCR1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC*G
PCR2_Idx_1_ATCACG	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTG*C
PCR2_Idx_2_CGATGT	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTG*C
PCR2_Idx_3_TTAGGC	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTG*C
PCR2_Idx_4_TGACCA	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTG*C
PCR2_Idx_5_ACAGTG	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTG*C
PCR2_Idx_6_GCCAAT	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTG*C
PCR2_Idx_7_CAGATC	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTG*C
PCR2_Idx_8_ACTTGA	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTG*C
PCR2_Idx_9_GATCAG	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTG*C
PCR2_Idx_10_TAGCTT	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTG*C
PCR2_Idx_11_GGCTAC	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTG*C
PCR2_Idx_12_CTTGTA	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTG*C

.

Chapter 3. ThetaMater: Bayesian estimation of population size parameter θ from genomic data

Richard H. Adams[1], Drew R. Schield[1], Daren C. Card[1], Andrew Corbin[1], and Todd A. Castoe[1, §]

[1]Department of Biology, 501 S. Nedderman Dr., The University of Texas at Arlington, Arlington, TX 76010, USA

ABSTRACT

Summary: We describe ThetaMater, an open source R package comprising a suite of functions for efficient and scalable Bayesian estimation of the population size parameter h from genomic data.

Availability and implementation: ThetaMater is available at GitHub (https://github.com/radamsRHA/ThetaMater).

INTRODUCTION

The population size parameter $\theta = 4N_e\mu$ ($2N_e\mu$ for haploid organisms) reflects the mutation- drift balance occurring within a population with an effective size of N_e individuals and a mutation rate of μ per site per generation. As a measure of genetic diversity, θ represents the expected number of segregating sites observed between a pair of homologous sequences sampled from a given population (Wakeley 2008). Given an estimate of mutation rate, information about θ can be leveraged to obtain an estimate of the effective population size N_e. θ is a therefore a fundamental parameter of population genetics and is useful for understanding the degree to which neutral processes shape patterns of genetic variation in nature. Quantifying genetic diversity is also important to conservation biology, and thus estimates of θ provide critical insight into the genetic health of endangered species for informed conservation practices (Crandall et al. 1999).

Numerous methods and genetic models have been developed to estimate θ from genetic data (see Wang, 2005 for examples). As any estimate obtained from a single locus or a small set of loci entails substantial uncertainty, large genome-scale datasets offer opportunity to estimate θ with high accuracy and precision. However, few likelihood-based methods are currently scalable to such massive datasets (>106 loci, >10kb/locus), are often restricted to using a single or small set of diploid genomes, are restricted to a specific type of sequence data (i.e., whole genomes vs. reduced representation), or require users to make assumptions about generation time and mutation rates. For example, most implementations of the popular pairwise-sequential Markov coalescent model (PSMC) require whole-genome data and that users provide a mutation rate assumed to be identical across all loci (Li and Durbin 2011), while other methods are restricted to using individual diploid genomes (Haubold, B. et al. 2010). There are many genealogy-based methods for estimating demographic parameters (Felsenstein 1992; Kuhner et al. 1995), but these are intractable for genomic

datasets that include many individuals. Furthermore, no current methods provide a statistical framework for leveraging estimates of θ to filter potentially spurious loci from datasets (i.e., paralogs). Accordingly, there is major need for efficient and scalable likelihood-based methods for estimating θ from diverse genomic datasets.

IMPLEMENTATION

The R package ThetaMater was written in R and C++, and requires the R package MCMCpack (Martin et al. 2011) to simulate posterior probability distributions of θ. At the core of ThetaMater is the infinite-sites likelihood function (Watterson, 1975), which describes the probability distribution of observing k segregating sites in a sample size of n sequences obtained from a locus of size l. The likelihood of a genomic dataset under a given value of θ is then computed as a product of the individual-locus specific likelihoods (or summation of log- likelihoods), each with an associated number of segregating sites k, sample size n and length l (see manual for model description). We have further expanded this approach to incorporate a discretized-gamma model of among-locus rate variation to accommodate rate variation and to characterize the genomic landscape of among-locus rate variation by estimating the gamma shape parameter (Yang 1997). Importantly, our method provides a user-friendly framework for efficient estimation of θ and substitution rate variation that is scalable to diverse genome-scale datasets (>106 loci) with larger samples sizes (>10 genomes), while accounting for uncertainty within a likelihood-based framework. Our method collapses datasets into sets of unique patterns, such that under many conditions, there is almost no limit to the number of loci that can be used to estimate θ within minutes on a desktop computer. Unlike other methods restricted to a particular format, ThetaMater includes functions for converting a variety of widely-used alignment formats into usable input, including whole-genome sequences, reduced-representation data (i.e., RADseq, sequence capture), and single or multilocus Sanger sequenced

datasets. Finally, ThetaMater includes a posterior predictive simulator (PPS) that allows users to leverage estimates of θ to identify loci with evidence of model violations, such as selection (Adams et al. 2016) or paralogy.

ThetaMater includes three Bayesian Markov Chain Monte Carlo (MCMC) simulation models for estimating posterior distributions of θ: M1 (ThetaMater.M1) assumes no among-locus rate variation, M2 (ThetaMater.M2) estimates θ using a fixed α parameter, and M3 (ThetaMater.M3) estimates the joint posterior distribution of θ and α. We implement a gamma prior distribution for both θ and α with user-specified shape and scale parameters, and users can specify the number of rate classes used to approximate the distribution. The posterior predictive simulator function (ThetaMater.PPS) is directly integrated with the results from the three Bayesian models.

BIOLOGICAL APPLICATION

As a demonstration, we applied ThetaMater on a previously published RADseq dataset (2051 loci; Schield et al., 2017). We conducted Bayesian estimation of using ThetaMater.M1 for the empirical dataset before and after filtering loci with ThetaMater.PPS (Fig. 1A). We also simulated a large genomic dataset comprised of 106 loci (2kb each), sampling 20 genomes from a population with θ = 0.002 and among-locus rate variation = 0.5 (Fig. 1B). We specified the shape and scale parameters of the prior distribution at 10 and 0.0001 for the empirical example, and set prior parameters to 20 and 0.0001 for θ and 5 and 0.01 for α in the simulated analysis.

We ran the MCMC chain for a total of generations and discarded 10% as burn in. PPS were run using the unfiltered posterior distribution, simulating a single locus for all 10_4 generations present in the post-burn in MCMC samples using ThetaMaterPPS.

ThetaMater analysis of the unfiltered RADseq dataset suggested a mean θ estimate of 0.0019, corresponding to N_e = 47,500 assuming a mutation rate of 10^{-8} (Fig. 1A, red). PPS based on this posterior distribution identified 3 loci with a significant excess of mutations, and these loci were filtered prior to reanalysis with ThetaMater. The posterior distribution of Ne inferred was centered around 45,000 individuals after removing these potentially spurious loci (Fig. 1A, blue). ThetaMater analysis of the simulated data returned the simulated parameter values with high probability (Fig. 1B).

ThetaMater is optimized for diverse datasets, including single diploid genome analyses, multi-genome data, reduced-representation data, and single or multilocus alignments. ThetaMater assumes free recombination between loci, no recombination within loci, error-free SNP calls, and neutral evolution. We encourage all users to carefully consider these assumptions prior to analysis with ThetaMater (see manual). Given the user-friendly framework and tractability of ThetaMater, we expect ThetaMater to be useful for a variety of applications, including population biology, comparative genomics, and conservation biology.

ACKNOWLEDGMENTS AND FUNDING

This work has been supported by a University of Texas at Arlington Phi Sigma Society Grant to R.H.A, startup funds to T.A.C., and NSF DDIG grants to D.R.S & T.A.C (NSF DEB-1501886) and D.C.C. & T.A.C (NSF DEB-1501747).

FIGURES

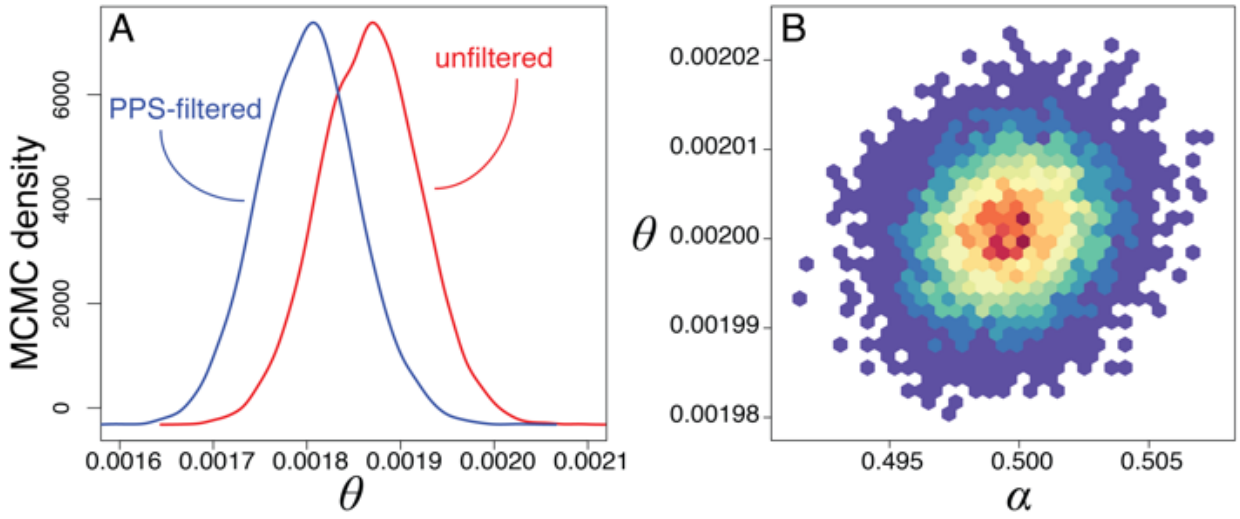


Fig. 1. (A) Empirical posterior estimates of $\boldsymbol{\theta}$ before (red) and after (blue) filtering with Thetamater.PPS, and (B) the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ for the simulated dataset showing highest densities (warm colors) at the true simulated values ($\boldsymbol{\theta}$ = 0.002, $\boldsymbol{\alpha}$ = 0.5).

REFERENCES

Adams,R. et al. (2016) GppFst: genomic posterior predictive simulations of FST and dXY for identifying outlier loci from population genomic data. Bioinformatics, 3, 1414–1415.

Crandall,K. et al. (1999) Effective population sizes: missing measures and missing concepts. Anim. Conserv., 2, 317–319.

Felsenstein,J. (1992) Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res., 59, 139–147.

Haubold,B. et al. (2010) mlRho—a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. Mol. Ecol., 19, 277–284.

Kuhner,M.K. et al. (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics, 140, 1421–1430.

Li,H. and Durbin,R. (2011) Inference of human population history from individual whole-genome sequences. Nature, 475, 493–496.

Martin,A.D. et al. (2011) MCMCpack: Markov Chain Monte Carlo in R. J. Stat. Softw., 42, 1–21.

Schield,D.R. et al. (2017) Insight into the roles of selection in speciation from genomic patterns of divergence and introgression in secondary contact in venomous rattlesnakes. Ecol. Evol., 7, 3951–3966.

Wakeley,J. (2008) Coalescent Theory: An Introduction. Roberts and Company, Greenwood Village, CO.

Wang,J. (2005) Estimation of effective population sizes from data on genetic markers. Philos. Trans. R. Soc. Lond. B. Biol. Sci., 360, 1395–1409.

Watterson,G.A. (1975) On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol., 7, 256–276.

Yang,Z.H. (1997) On the estimation of ancestral population sizes of modern humans. Genet. Res., 69, 111–116.

APPENDIX

The effective population size parameter θ

The population size parameter θ reflects the effects of genetic drift and mutation on patterns of genetic variation within a diploid population (for a haploid population) with an effective size of individuals and a mutation rate of per site per generation. If two homologous sequences are sampled at random from a population, describes the expected number of segregating sites observed between these two sequences. θ is a fundamental measure of genetic diversity in populations and is thus an informative parameter used in many population genetic models. The R package ThetaMater provides a Bayesian framework to estimate both θ and (shape of among- locus rate variation) parameters from a variety of genetic datasets, including haploid or diploid genomic data from single or multiple individuals, reduced-representation genomic data (e.g., RADseq, sequence capture), and single or multilocus Sanger sequence data (and variations of these datasets). ThetaMater implements three different functions that can be used to estimate these parameters within a Bayesian framework:

- ThetaMater.M1: estimate without among-locus variation
- ThetaMater.M2: estimate with a fixed parameter of rate variation and a user-defined number of locus rate classes
- ThetaMater.M3: estimate both and the shape parameter given a user-defined number of rate classes

The likelihood function implemented by ThetaMater

The three functions (ThetaMater.M1, ThetaMater.M2, ThetaMater.M3) simulate posterior probability distributions of effective population size parameters for a given dataset. These functions employ the likelihood function $P(S = k|l, n; \theta)$ to compute the probability of observing k segregating sites in a

sample size of n from a locus with length l for a given value of θ. These methods compute the likelihood of a given dataset as a summation of the log- transformed likelihoods across all loci. See the following publications for more information about this model, its derivation, applications, and similar models:

- Tavaré, Simon. "Line-of-descent and genealogical processes, and their applications in population genetics models." Theoretical population biology 26.2 (1984): 119-164.
- Watterson, G.A. On the number of segregating sites in genetical models without recombination. Theoretical population biology 1975.
- Wakeley, John. "Coalescent theory." Roberts & Company (2009).
- Hein, Jotun, Mikkel Schierup, and Carsten Wiuf. Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford University Press, USA, 2004.
- Takahata, Naoyuki, and Yoko Satta. "Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences." Proceedings of the National Academy of Sciences 94.9 (1997): 4811-4815.
- Takahata, Naoyuki, Yoko Satta, and Jan Klein. "Divergence time and population size in the lineage leading to modern humans." Theoretical population biology 48.2 (1995): 198- 221.
- Yang, Ziheng. "On the estimation of ancestral population sizes of modern humans." Genetical research 69.02 (1997): 111-116.

Below is the formula for the likelihood function described in these papers that is central to the three ThetaMater functions:

$$P(S = k|l, n; \theta) = \int_0^{\infty} P\,(S = k|t) f_T(t) dt$$

$$P(S = k|l, n; \theta) = (\frac{l\theta}{2})^k \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \frac{i-1}{2} \int_0^{\infty} \frac{(t^k\, exp\{\frac{(-\theta + i - 1)t}{2}\})}{k!}\, dt$$

$$P(S = k|l, n; \theta) = (\frac{l\theta}{2})^k \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \frac{i-1}{2} (\frac{2}{\theta + i - 1})^{(k+1)}$$

$$P(S = k|l, n; \theta) = \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \frac{i-1}{\theta + i - 1} (\frac{\theta}{\theta + i - 1})^k$$

For a dataset consisting of x loci, each an observed number of segregating sites k5, number of bases l5, and number of sequences sampled n5, we can sum the likelihoods of the individual loci to get the likelihood of the entire dataset under a given value of θ:

$$L(D|\theta) = \sum_{i=1}^{x} log(P(S = k_i|l_i, n_i; \theta)$$

APPLICATIONS, ASSUMPTIONS, AND LIMITATIONS OF THETAMATER

Understanding the assumptions of ThetaMater and the underlying coalescent model are critical to the appropriate use of ThetaMater. Importantly, ThetaMater assumes that there is no recombination within individual loci and free recombination between loci (i.e., no linkage).

Furthermore, all loci are assumed to have evolved under strictly neutral evolution. These are fundamental assumptions of the coalescent model and the likelihood function implemented in ThetaMater. This can been seen in the form of the likelihood equation provided above: the likelihood of an entire dataset is a summation of the log-likelihoods across loci that are assumed to be

genetically unlinked. In other words, the genealogy and number of segregating sites observed at each locus is assumed to be independently and identically distributed (i.i.d).

To explore the potential effects of one such model violation (unrecognized recombination) in datasets, we simulated loci using the software msprime under 6 different recombination rates:

(2e-9, 2e-8, 2e-7, 2e-6, 2e-5, 2e-4), using a sample size of 5 gene copies per 10kb locus, and with each dataset consisting of 10k loci. See Step 8: "Recombination & ThetaMater" for a plot of these analyses for each recombination rate and a dataset without recombination. In general, ThetaMater appears largely unaffected by recombination, as the posterior distribution of each analysis is largely centered around the true simulation value (θ = 0.008). ThetaMater assumes that all loci are genetically unlinked, and at the request of a reviewer, we conducted a simulation of human chromosome 1 to evaluate the effects of linkage on ThetaMater estimates (See Step 9: "Linkage & ThetaMater"). Under extreme scenarios of linkage, ThetaMater appeared to be biased towards larger values, but for more realistic conditions, ThetaMater appears to be robust to linkage. Nonetheless, these are complex subjects, and we recommend users to explore all potential violations of the model (including selection and recombination/linkage) prior to using ThetaMater.

As estimates from any one locus entail significant uncertainty, ThetaMater allows researchers to take full advantage of large, genomic datasets when estimating and provides a distribution of plausible values for parameter estimates while accounting for uncertainty. Users can also use an estimate of the shape of among-locus rate variation (ThetaMater.M1) or estimate the shape of among-locus rate variation (ThetaMater.M2) to account for among-locus rate variation when estimating , as well as characterize the genomic landscape of rate variation. The posterior predictive simulator included in ThetaMater allows users to identify potential outlier loci from the genomic distribution of genetic variation, whether due to issues of orthology (see Step 7), or other violations of model assumptions,

such as selection (see GppFst R package, Adams 2017). ThetaMater also includes several functions for simulating datasets under the neutral coalescent model. Briefly, datasets are simulated under the infinite-sites model of mutation according to the protocol described in Wakeley (2008: pg. 255).

Users can estimate locus-specific s for each locus within a dataset to characterize among-locus estimates of θ, or leverage all loci to estimate a single, population-wide estimate. For single locus-based estimates, reflects the time to the most common ancestor among a sample of sequences. This is because the average time for 2 copies to reach a common ancestor is equal to 2N generations (~4N generations for larger sample sizes). Thus, users can characterize differences in TRMCA (locus-specific) among loci for a number of different applications, such as understanding what evolutionary processes may be at work across the genome. For example, a short TMRCA (i.e., small effective population size) may indicate the effects of positive selection, while an older TMRCA (i.e., large effective population size) may indicate balancing selection (or other processes).

LINKAGE & THETAMATER

ThetaMater assumes that all loci are genetically unlinked (i.e., free recombination between loci). At the request of a reviewer, we conducted a simulation analysis to evaluate the effects of linkage on posterior estimates derived via ThetaMater. We simulated a sample of human chromosome 1 (length = 248,956,000bp) with three different recombination rates (2e-8, 2e-9, 2e-10) and a sample size of 10 individuals. We also simulated a single dataset without recombination at all (i.e., $\rho = 0$, such that the entire chromosome was linked). We randomly sampled 1000bp loci every 100kb, with resembles "reduced-representation" sampling, such as RADseq and sequence capture data. We used the following command in msprime: msprime.simulate (sample_size=10, N_e=10000, length=248956000, recombination_rate= ρ, mutation_rate=2e-8) The results of these analyses are plotted below for each recombination rate. We included a python script (SimulateChr1.v2.py) to generate these results. As

you can see, in all cases with recombination (ρ = 2e-8, 2e-9, 2e-10), the posterior distribution of was centered near the true simulated value (θ = 0.0008), suggesting that ThetaMater is likely robust to linkage in these conditions. However, in the most extreme simulation in which ρ = 0 (no recombination at all), we did find that ThetaMater was biased towards a larger value than the true simulated value. Under extreme scenarios of recombination (all loci are genetically linked), ThetaMater may be biased, but under realistic conditions, ThetaMater appears robust to linkage. These simulations are not necessarily conclusive for all scenarios, and we encourage users to explore all potential violations of the coalescent model prior to using ThetaMater (i.e., recombination, linkage, selection). If there is some concern for model violations, one can simulate datasets (as we have done with msprime) to explore other potential violations, including linkage and selection using similar approaches to those presented here.

Chapter 4. Genetic diversity and effective population size of the Texas blind salamander (*Eurycea rathbuni*) from Central Texas

Andrew B. Corbin[1], Andrew Gluesenkamp[2], Paul T. Chippindale[1]

[1]Department of Biology, University of Texas at Arlington, Arlington, Texas 76019 USA

[2]Director of Conservation, Center for Conservation and Research, San Antonio Zoo, San Antonio, Texas 78212 USA

ABSTRACT

The Texas blind salamander (*Eurycea rathbuni*) is among Texas's most iconic species and was one of the first species to be listed under the Endangered Species Act. This species is imperiled primarily due to anthropogenic impacts on its habitat as well as groundwater quality and quantity. Because this species is known only from the San Marcos pool of the Edwards Aquifer in central Texas, a single event such as chemical contamination could be catastrophic to this species. To safeguard *E. rathbuni,* it is essential to collect and maintain individuals from the wild and to understand the genetic diversity, populations structure, and effective population size (N_e) of the species. In this study, we collect individuals from wild populations of *E. rathbuni* and the research population at San Antonio Zoo, San Antonio TX, to measure the genetic diversity and effective population size of the species in the wild and in captivity. In addition, we estimate the number of founders required to maintain adequate levels of genetic diversity in captive populations. To achieve these goals, we leverage restriction-site associated DNA sequencing (RADseq) to target thousands of genetic markers per individual. This study provides the most robust measure of population structure, genetic diversity, and N_e of wild *E. rathbuni* to date, establishing a baseline for long-term genetic monitoring of the captive and wild populations. Finally, this study provides data-based guidance for captive breeding strategies that may benefit captive assurance colonies such as those currently maintained by US Fish and Wildlife Service.

INTRODUCTION

The Edwards Aquifer is inhabited by groundwater salamanders (genus *Eurycea*) that are restricted and highly adapted to this unique system of springs and water-filled caves. Due to the sensitivity of this habitat, many of these species are of great conservation concern. The diversity of salamanders of the genus *Eurycea* in the region is remarkable (Bendik et al., 2013; Chippindale et al., 1993; Hillis et al., 2001; Wiens et al., 2003). The Texas blind salamander (*Eurycea rathbuni*) is an extreme cave-dweller that occupies one of the smallest known ranges of any amphibian in North America and was one of the first species to be listed under the Endangered Species Act (U.S. Congress 1973). This species faces many threats and is identified as a Species of Greatest Conservation Need (SGCN) in Texas, listed as Endangered at the state and Federal levels, and ranked as G1 ("critically imperiled") under the global NatureServe ranking system. These threats include groundwater withdrawal, chemical spills, recharge contamination, and aquifer degradation from lack of recharge protection in this rapidly urbanizing region (Bendik et al., 2014; Bowles and Arsuffi, 1993; Burri et al., 2019). Recent research has also demonstrated that species living in the aquifer in the central Texas region are greatly threatened by extinction within the next century (Devitt et al., 2019).

Captive breeding is an important tool used by conservation biologists. Captive breeding populations are intended to provide opportunities for research on imperiled species and to provide stock for reintroduction or supplementation of wild populations. There are currently two captive assurance colonies for *E. rathbuni* maintained by the USFWS. One is at the San Marcos Aquatic Research Center (SMARC) in San Marcos, TX, and the other at the Uvalde National Fish Hatchery, Uvalde, TX that are maintained for potential re-introduction purposes. One main

goal of this study is to establish a complementary captive research population of *E. rathbuni* at San Antonio Zoo (SAZ).

It is crucial that captive breeding strategies are informed by population connectivity because, if isolated populations harbor unique fixed adaptations, breeding animals among unique populations could lead to the loss of important genetic adaptations through outcrossing (Charlesworth and Charlesworth, 1987). Alternatively, if the wild population is largely panmictic but captive animals are segregated based on the assumption that sampling localities depict isolated populations, this could result in a 'bottleneck' situation where captive populations lose valuable genetic diversity due to inbreeding (Wang et al., 1999). We seek to provide guidance for captive breeding strategies that are informed by patterns of diversity found in wild populations.

Previous studies based on allozyme data, mt sequences, and limited nuclear markers (e.g. Chippindale et al. 2000; Chippindale 2009) indicate high levels of genetic polymorphism in *E. rathbuni*, and *Eurycea* in the region display a complex evolutionary history marked by separation and reconnection (Bendik et al., 2013; Devitt et al., 2019). One unresolved question is the degree to which populations of *E. rathbuni* are hydrologically connected, and whether current collection sites represent distinct populations or if substantial genetic structure exists among localities. Although studies of hydrology of the San Marcos Pool of the Edwards Aquifer indicate high levels of hydrologic connectivity among sites (Ogden et al., 1986; Russell, 1976; Schindel and Gary, 2017), researchers and resource managers still know very little about the structure of the San Marcos Pool of the Aquifer where *E. rathbuni* resides and the degree to which salamanders may move among sampling sites (although gene flow can still occur even if salamanders do not move between sites).

Estimates of effective population size (N_e) using molecular data can be a powerful tool in the struggle to understand rare and imperiled species (Culver et al., 2008; Olsen et al., 2016), and serve as an important complement to consensus population size (N_c; Nunziata and Weisrock 2018). However, there are currently no reliable estimates of N_e or N_c for *E. rathbuni.* Krejca and Gluesenkamp (2007) conducted a mark-recapture study at three sites. Over the course of a year of bi-weekly sampling, they captured 12 individuals and had zero recaptures. These numbers highlight the difficulty of conducting population studies on this species due to the inaccessibility of *E. rathbuni* habitat and generally low detection probability. Thus, estimates of N_e based on molecular data could greatly enhance our understanding of this species.

Although many studies have advanced our understanding of *E. rathbuni* population genetics and connectivity (Chippindale, 2009; Devitt et al., 2019; Hillis et al., 2001; Krejca and Gluesenkamp, 2007; Potter Jr. and Sweet, 1981; Wiens et al., 2003), these studies used traditional marking techniques, limited genetic markers (allozymes or a small number of nuclear or mitochondrial sequences), or were limited by sample size. Recent advances in DNA sequencing allow researchers to collect hundreds to thousands of unlinked genetic markers. Here, we used double digest restriction-site associated DNA sequencing (ddRADseq; Peterson et al. 2012), allowing a fine-scale investigation of levels and patterns of variation and providing a baseline for monitoring captive and wild populations.

We will use emerging high-throughput DNA sequencing to achieve three specific aims: 1) estimate diversity metrics among and within collection sites of *E. rathbuni*, as well as compare the diversity seen within the SAZ population to diversity among wild populations; 2) examine population structure and phylogenetic relatedness among individuals; 3) estimate N_e for the species overall.

METHODS

Texas blind salamander collection and sampling

Collection of *E. rathbun*i was conducted under permits issued to AGG, and all sampling and sample processing followed IACUC protocols from UTA. Collection of live *E. rathbuni* was conducted by SAZ staff using hand nets, mop heads, bottle traps, and minnow traps at Johnson's Well, Primer's Fissure, Rattlesnake Cave, and Rattlesnake Well (Supplementary Table 1). Of the 14 salamanders that were captured, six were sent live to SAZ and seven were returned to the wild after collection of a small tail clip. We also extracted DNA from 40 additional *E. rathbuni* tail clips that were taken in 2014 from live animals housed at the SMARC that were captured from the wild across several localities of *E. rathbuni* (see localities listed in Supplementary Table 1).

DNA preparation and sequencing

We extracted DNA from tail clips using a phenol-chloroform-isoamyl method due to its ability to provide high-quality genomic DNA. Samples were quantified using a Qubit fluorometer 2.0 (Life Technologies, Carlsbad, CA, USA), and a minimum of 500ng per sample was used for library preparation. We used a modified protocol of Peterson et al. (2012) to generate ddRADseq libraries. Briefly, samples were digested with *Sbf1* (8bp recognition site) and *Sph1* (6 bp recognition site) restriction enzymes according to manufacturer protocols. We then cleaned the samples with Ampure magnetic beads (Invitrogen), and ligated double-stranded indexed adapters containing 8 consecutive Ns (unique molecular identifiers, UMIs) used to identify and remove PCR clones post-sequencing. Samples were pooled into groups of 8, ensuring that no two indices were pooled into the same group. Samples were then size selected to 302-360bp (58bp window) using the Blue Pippin Prep (Sage Science, Beverly, MA, USA). Pooled, size selected groups

were then amplified using through PCR using Phusion high fidelity polymerase (New England Biolabs) for 25 cycles. Groups were quantified using a Bioanalyzer (Agilent, Santa Clara, CA, USA) to verify size-selection and calculate DNA concentration. Groups were combined into a single library base on molarity and sequenced on an Illumina platform using 150bp paired-end reads.

Bioinformatics and data analysis

Raw reads were first processed using the clone_filter program in STACKS v2.5.3 (Rochette et al., 2019), using the 8bp UMIs to filter PCR clones, and then removed UMIs using the fastx_trimmer function from the FASTX-Toolkit v0.0.13 (hannonlab.schl.edu/fastx_toolkit). We then ran the process_radtags command in STACKS v2.53 (Rochette et al., 2019) to demultiplex trimmed reads using the –clean function to remove reads with uncalled bases, the -q function to filter reads with low quality scores, and the -r function to rescue barcodes with no more than one mismatch.

To optimize datasets used in further analyses, we follow the *r80loci* guidelines described in (Paris et al., 2017). Our overall aim is to optimize parameters to include as many orthologous, polymorphic loci as possible while excluding paralogs and erroneous sequences. Briefly, we ran the denovo_map.pl pipeline while varying the value of M (the number of mismatches allowed among putative alleles in ustacks), n (the number of mismatches allowed among putative loci in cstacks). Because m=3 (minimum number of raw reads required to form a putative allele in ustacks) has been found to be suitable for most biological datasets with reasonable coverage (Paris et al., 2017), we held this parameter constant for all analyses. We set M = n for each parameter, as recommended by (Paris et al., 2017). We ran the pipeline 7 times, with values of M

and n = [1-7], generating 7 datasets. We then compared the number of loci for each run and the number of loci including at least one SNP. We examined the difference in the number of loci among each consecutive run to find the maximum number of loci optimal value for M and n as described in (Paris et al., 2017). We then examined the number of loci present in each individual after sstacks and removed any samples with fewer than 500 loci.

After parameter optimization, we apply further filters to produced final datasets using the populations program in STACKS under differing values of -p (minimum number of populations a locus must be present in to process a locus) and -r (minimum number of individuals within a population to process a locus). We have 6 sampling localities represented in our dataset, however we excluded Diversion Spring due to low sample size (n=3), so we ran three runs p = 1, 3, and 5. For each of these values of -p, we varied -r (percent of samples required to be present within each population) to 0.50 and 0.75. For each run, we examined the number of loci and SNPs. We used the –fstats flag to generate F-statistics for these datasets in STACKS. F-statistics were calculated using an AMOVA F_{ST} method described in (Weir, 1997). To compare F_{ST} of SAZ animals to wild populations, we compared F_{ST} between the animals kept at SAZ to the individuals that were captured and released, generated two datasets, using -r = 0.50 and setting -p to 1 and 3. To generate a dataset for phylogenetic analysis and to examine population structure, we created a new population map file listing each sample as a unique population to avoid biasing the dataset with sampling locality assumptions. We also used the --write-random-snp option to include only one SNP per locus in cases where two or more SNPs exist on a single locus for, thus avoiding including linked SNPs. We set -p = 1 and varied the amount of missing data allowed using the -r option (0.25, 0.5, and 0.75) to generate three datasets, and we examined the number

of SNPs and loci in each dataset. These datasets were used to generate phylogenetic trees and analyse the population structure.

To examine population structure, we implemented the program STRUCTURE (Pickrell and Pritchard, 2012) without assuming prior sampling locality information. We used values of K = 1-7 with an MCMC length of 1 million and a 10% burnin. We determined the most optimal value of K using the Evanno method (Evanno et al., 2005) implemented in Structure Harvester (Earl and vonHoldt, 2012). We then used CLUMPAK (Kopelman et al., 2015) to summarize and visualize results. We used RAxML (Stamatakis, 2014) to estimate a maximum likelihood phylogenetic tree under the GTR_GAMMA model of nucleotide evolution. We used the Lewis ascertainment bias correction to correct for having only variable site and assessed tree support using 1000 bootstrap replicates.

To generate estimates of N_e for *E. rathbuni* using the program NeEstimator v2.0 (Do et al., 2014) under the linkage disequilibrium method (Hill, 1981), we generated a dataset including the 40 individuals with the greatest number of loci regardless of sampling locality. For this, we generated a separate population map denoting that all individuals belong to one population, and set -r to 0.25, 0.5, and 0.75 to vary the amount of missing data allowed in the calculation. We included only one SNP per locus, avoiding including known linked sites that can confound results, although there are other factors that can contribute to uneven allele sorting which can cause linkage disequilibrium among SNPs (Slatkin, 2008). We used Pcrit cutoffs to consider the effect of excluding rare alleles, which is important in Ne estimates. We set Pcrit 0.02 (recommended by Waples and Do 2010), 0.3, and 0.05.

Because our data generated in this Chapter displayed unusually high amounts of allelic drop out, (see Results and Discussion), we used 11 *E. rathbuni* samples from Chapter 2 of this dissertation to estimate N_e using Watterson's theta (θ; Watterson 1975). To generate this dataset, we required that a locus be present in 8 out of 11 individuals, and extracted only one SNP per locus, as sites must be unlinked to accurately estimate θ. Estimates of θ using the following method performs most accurately with large numbers of SNPs and is robust to smaller sample sizes (Richard Adams, personal communication).

We estimated N_e using the equation ($\theta = 4N_e\mu$) by estimating θ using the R package *ThetaMater* (Adams et al., 2018). *ThetaMater* uses an infinite-sites likelihood function to estimate the posterior probability of theta based on the probability of observing k segregating sites in a sample size of n sequences obtained from each locus (Watterson, 1975). We used the *Read.AllelesFile* function to read pyRAD output into the program directly. To further filter the data, we use the *ThetaMater.PPS* function to filter loci with excessive mutations (i.e. paralogs or loci under selection). Bayesian estimation of θ was conducted using the *ThetaMater.M1* MCMC for a total of 1×10^6 generations and discarded 10% of generations as burn in. Following Schield et al. (2018), to calculate N_e we divide the median of the posterior distribution of θ by the generation time, and assume the previously estimated substitution rate of frogs ($\mu = 0.776 \times 10^{-9}$) as estimated by Sun et al. (2015). Based on observed patterns of reproduction in captivity it takes a minimum of 2-3 years from hatching for salamanders to breed successfully, although there is considerable variability and mating is highly sporadic (Andrew Gluesenkamp, personal communication). Because are currently no reliable estimates of generation time in wild populations, we estimated N_e under each generation time of 1-10.

RESULTS AND DISCUSSION

Data processing

We obtained 799,559,860 raw reads across both lanes of sequencing, and we recovered 26% of reads after filtering PCR cloned reads. Running the denovo pipeline under m = 3, we found very little difference in the number of polymorphic loci (loci containing at least one SNP) with each increasing number of M = n (Supplementary Figure 1), indicating that multiple values of M = n could be considered appropriate for further use. We moved forward using M = n of 5 for each dataset prepared in this study. Upon examining the number of loci per individual after running the denovo pipeline, we excluded four individuals were excluded due to low coverage. The datasets used to estimate diversity metrics included 5695 loci for the very relaxed dataset, and 978 loci for the more stringent dataset. Values of F_{ST} were very similar between both datasets (Table 1). For the datasets to be used for population structure F_{ST} and phylogenetic estimates, we find that the dataset allowing 75% missing data recovered 2,503 SNPS and allowing 50% missing data recovered 372 SNPs. However, the dataset allowing 25% missing data recovered only 4 SNPs and was not used for further analysis. This amount of allelic dropout was unexpected given the relatively high level of loci retention of the RADseq analyses in Chapter 2, and other studies that have used almost identical laboratory methods in salamanders (Devitt et al., 2019; O'Connell et al., 2019).

Genetic diversity

The datasets requiring 75% of individuals to be present to call a locus (-r = 0.75) did not yield enough loci to estimate F_{ST} values (most values were 'nan'). The datasets requiring that 5 localities be present also failed to generate enough loci for these estimates, so we compared the

results of the two remaining datasets where -r = 0.50, and -p = 1 and 3. We found very little difference between the pairwise comparisons despite observing substantial differences in missing data (Table 1). This is somewhat surprising, given that setting -p = 1 is a very relaxed setting and allows a substantial amount of missing data, which can confound F_{ST} estimates (Gautier et al., 2013). Sampling localities display low to moderate divergence between them, despite STRUCTURE showing little to no population structure (Fig. 1). This could indicate that populations are somewhat divergent despite STRUCTURE's inability to detect population structure. The Evanno method found that the optimal value of $K = 2$. Because there seems to be no apparent pattern among sampling localities, we expect that this is due to the limitations of this method, and that $K = 1$ may be the most appropriate value of K.

We find that $F_{ST} = 0.101$ between SAZ and the captured and released (wild) populations. Ideally, F_{ST} would be as close to zero as possible between captive and wild populations to ensure that captive populations mirror that of wild populations. Global diversity statistics indicate that heterozygosity within each sampling locality is very low compared to expected heterozygosity. This can indicate higher than expected levels of inbreeding, and our F_{IS} estimates corroborate this (although considerable variation exists in this estimate; Table 2). Overall, it appears that this species may be highly panmictic and considerable inbreeding may occur. Increased levels of inbreeding can occur in small populations (O'Grady et al., 2006) or if the population is undergoing a genetic bottleneck (Keller et al., 1994), and can be accentuated if there are overlapping generations (Felsenstein, 1971). Although little information exists on the longevity of wild *E. rathbuni*, this species is presumed to be quite long-lived (potentially several decades). Some specimens captured as large adults in the 1990s are still alive (Randy Gibson, SMARC, personal communication). Given that generations likely overlap, this could be one potential

cause of the increased levels of observed heterozygosity (compared to expected heterozygosity) and F_{IS} seen in this dataset, however future research should explore other potential causes of these patterns such as a potential genetic bottleneck due to a recent or ongoing population decline.

Phylogenetics and population structure

For our STRUCTURE analysis, we found an optimal value of $K = 2$, although the results of each run are shown. We found little evidence of population structure among *E. rathbuni* collection sites based on our STRUCTURE analysis (Fig. 1). Diversion Spring is the only sampling locality with distinct structure (at higher values of *K*). However, this may be due to the low sample size for this locality (n = 3) compared to others. It is well documented that uneven sample sizes can confound STRUCTURE analyses (Shringarpure and Xing, 2014), so this result is unsurprising. Our phylogenetic analyses failed to discriminate among sampling localities, and we find very low bootstrap support for both RAxML trees (50% allowed missing data, and 75% allowed missing data; Fig. 2), further indicating a lack of population structure among sampling localities.

N_e estimates

We found that the most relaxed dataset yielded 3,783 SNPS while allowing 75% missing data did not provide precise measurements of N_e (infinite values under each Pcrit value), nor did the most stringent dataset allowing only 25% missing data. Allowing 50% missing data offered the best trade-off between dataset completeness and number of SNPs. We found that using a Pcrit of 0.02 produced an infinite N_e estimate and confidence intervals, however Pcrit of 0.03 suggests that $N_e = 942.3$ (95CI = 180.2 – infinite), and a Pcrit of 0.05 estimates $N_e = 315.7$ (95CI = 107.2-infinite) (Table 3). ThetaMater estimated that $\theta = 0.002536134$ (stdev = 3.86×10^{-5}) using the 11

E. rathbuni samples from Chapter 2. Increasing generation time will, by definition, decrease estimates of N_e. Setting generation time to 3 years resulted in N_e of 272,351 while a generation time of 10 years resulted in an N_e of 81,705 (Fig. 3).

Ideally, N_e should be interpreted alongside mark-recapture studies and estimates of N_c (Nunziata and Weisrock, 2018). Unfortunately, the only mark-recapture study that has been conducted on *E. rathbuni* involved very few individuals despite intensive efforts (Krejca and Gluesenkamp, 2007). The inaccessible nature of *E. rathbuni* habitat and low population densities at accessible sites prohibit reliable estimates of consensus population size via traditional mark-recapture studies. Therefore, estimates of N_e based on molecular data have the potential to provide crucial information regarding effective population size.

Here, we found that generating robust estimates of N_e using the ddRADseq data generated in this Chapter was greatly hindered by unexpectedly high amounts of missing data. We expect that this is either due to laboratory errors (most likely low quality starting DNA, or errors in size selection), but could also be caused by the inherent nature of the biological system (Paris et al., 2017) or bioinformatic processing. In any case, future studies should generate RADseq libraries using more starting DNA, a wider size-selection window, and more sequencing coverage to explore possible causes of allelic dropout observed here. The results of this chapter ultimately represent an experiment using ddRADseq data that show high levels of missing data, and we recommend interpreting these results cautiously. We are currently planning to re-prepare the ddRADseq libraries to address these issues. This would ideally produce a dataset with thousands of loci even when using stringent filtering steps. This will allow more appropriate estimates of N_e using the LD and Watterson's θ methods as implemented above, and more robust estimates of F-statistics and population structure.

REFERENCES

Adams, R.H., Schield, D.R., Card, D.C., Corbin, A., Castoe, T.A., 2018. ThetaMater: Bayesian estimation of population size parameter θ from genomic data. Bioinformatics 34, 1072–1073. doi:10.1093/bioinformatics/btx733

Bendik, N.F., Meik, J.M., Gluesenkamp, A.G., Roelke, C.E., Chippindale, P.T., 2013. Biogeography, phylogeny, and morphological evolution of central Texas cave and spring salamanders. BMC Evol. Biol. 13, 1–18.

Bendik, N.F., Sissel, B.N., Fields, J.R., O'Donnell, L.J., Sanders, M.S., 2014. Effect of urbanization on abundance of jollyville plateau salamanders (Eurycea tonkawae). Herpetol. Conserv. Biol. 9, 206–222.

Bowles, D.E., Arsuffi, T.L., 1993. Karst aquatic ecosystems of the Edwards Plateau region of central Texas, USA: A consideration of their importance, threats to their existence, and efforts for their conservation. Aquat. Conserv. Mar. Freshw. Ecosyst. 3, 317–329. doi:10.1002/aqc.3270030406

Burri, N.M., Weatherl, R., Moeck, C., Schirmer, M., 2019. A review of threats to groundwater quality in the anthropocene. Sci. Total Environ. 684, 136–154. doi:10.1016/j.scitotenv.2019.05.236

Charlesworth, D., Charlesworth, B., 1987. Inbreeding depression and its evolutionary consequences. Annu. Rev. Ecol. Syst. 18, 237–268. doi:10.1146/annurev.es.18.110187.001321

Chippindale, P.T., 2009. Population genetic analysis of the Texas blind salamander, Eurycea rathbuni. Endangered Species Program TPWD. Final Report, 1–26.

Chippindale, P.T., Price, A.H., Hillis, D.M., 1993. A new species of perennibranchiate salamander (Eurycea : Plethodontidae) from Austin, Texas 49, 248–259.

Chippindale, P.T., Price, A.H., Wiens, J.J., Hillis, D.M., 2000. Phylogenetic relationships and systematic revision of central Texas hemidactyliine plethodontid salamanders. Herpetol. Monogr. 14, 1–80.

Culver, M., Hedrick, P.W., Murphy, K., O'Brien, S., Hornocker, M.G., 2008. Estimation of the bottleneck size in Florida panthers. Anim. Conserv. 11, 104–110. doi:10.1111/j.1469-1795.2007.00154.x

Devitt, T.J., Wright, A.M., Cannatella, D.C., Hillis, D.M., 2019. Species delimitation in endangered groundwater salamanders: Implications for aquifer management and biodiversity conservation. Proc. Natl. Acad. Sci. U. S. A. 116, 2624–2633. doi:10.1073/pnas.1815014116

Do, C., Waples, R.S., Peel, D., Macbeth, G.M., Tillett, B.J., Ovenden, J.R., 2014. NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Mol. Ecol. Resour. 14, 209–214. doi:10.1111/1755-0998.12157

Earl, D.A., vonHoldt, B.M., 2012. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. 4, 359–361. doi:10.1007/s12686-011-9548-7

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. doi:10.1111/j.1365-294X.2005.02553.x

Felsenstein, J., 1971. Inbreeding and variance effective numbers in populations with overlapping generations. Genetics 68, 581–597.

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., Cornuet, J.M., Estoup, A., 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol. Ecol. 22, 3165–3178. doi:10.1111/mec.12089

Hill, W.G., 1981. Estimation of effective population size from data on linkage disequilibrium. Genet. Res. 38, 209–216. doi:10.1017/S0016672300020553

Hillis, D.M., Chamberlain, D.A., Wilcox, T.P., Chippindale, P.T., 2001. A new species of subterranean slind salamander (Plethodontidae: Hemidactyliini: Eurycea: Typhlomolge) from Austin, Texas, and a systematic revision of central Texas paedomorphic salamanders. Herpetologica 57, 266–280.

Keller, L.F., Arcese, P., Smith, J.N.M., Hochachka, W.M., Stearns, S.C., 1994. Selection against inbred song sparrows during a natural population bottleneck. Nature 372, 356–357. doi:10.1038/372356a0

Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., Mayrose, I., 2015. Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. Mol. Ecol. Resour. 15, 1179–1191. doi:10.1111/1755-0998.12387

Krejca, J., Gluesenkamp, A., 2007. Mark recapture study of E. rathbuni at three sites in San Marcos, Texas. Endangered Species Program TPWD. Final Report, 1-22.

Nunziata, S.O., Weisrock, D.W., 2018. Estimation of contemporary effective population size and population declines using RAD sequence data. Heredity. 120, 196–207. doi:10.1038/s41437-017-0037-y

O'Connell, K.A., Mulder, K.P., Maldonado, J., Currie, K.L., Ferraro, D.M., 2019. Sampling related individuals within ponds biases estimates of population structure in a pond-breeding amphibian. Ecol. Evol. 9, 3620–3636. doi:10.1002/ece3.4994

O'Grady, J.J., Brook, B.W., Reed, D.H., Ballou, J.D., Tonkyn, D.W., Frankham, R., 2006. Realistic levels of inbreeding depression strongly affect extinction risk in wild populations. Biol. Conserv. 133, 42–51. doi:10.1016/j.biocon.2006.05.016

Ogden, A., Quick, R., Rothermel, S., 1986. Hydrochemistry of the Comal, Hueco, and San Marcos springs, Edwards Aquifer, Texas, in: Abbott, P.L., Woodruff, C.M. (Eds.), The Balcones Ecarpment. Geological Society of America, San Antonio, pp. 51–54.

Olsen, J.B., Kinziger, A.P., Wenburg, J.K., Lewis, C.J., Phillips, C.T., Ostrand, K.G., 2016. Genetic diversity and divergence in the fountain darter (Etheostoma fonticola): implications for conservation of an endangered species. Conserv. Genet. 17, 1393–1404. doi:10.1007/s10592-016-0869-7

Paris, J.R., Stevens, J.R., Catchen, J.M., 2017. Lost in parameter space: a road map for stacks. Methods Ecol. Evol. 8, 1360–1373. doi:10.1111/2041-210X.12775

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7, e37135. doi:10.1371/journal.pone.0037135

Pickrell, J.K., Pritchard, J.K., 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8, e1002967. doi:10.1371/journal.pgen.1002967

Potter Jr., F.E., Sweet, S.S., 1981. Generic boundaries in Texas cave salamanders, and a redescription of Typhlomolge robusta (Amphibia: Plethodontidae). Am. Soc. Ichthyol. Herpetol. 64–75. doi:10.2307/1444041

Rochette, N.C., Rivera-Colón, A.G., Catchen, J.M., 2019. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol. Ecol. 28, 4737–4754. doi:10.1111/mec.15253

Russell, W.H., 1976. Distribution of troglobitic salamanders in the San Marcos area Hays County, Texas (BITE Report 7601). Austin, Texas.

Schield, D.R., Adams, R.H., Card, D.C., Corbin, A.B., Jezkova, T., Hales, N.R., Meik, J.M., Perry, B.W., Spencer, C.L., Smith, L.L., García, G.C., Bouzid, N.M., Strickland, J.L., Parkinson, C.L., Borja, M., Castañeda-Gaytán, G., Bryson, R.W., Flores-Villela, O.A., Mackessy, S.P., Castoe, T.A., 2018. Cryptic genetic diversity, population structure, and gene flow in the Mojave rattlesnake (Crotalus scutulatus). Mol. Phylogenet. Evol. 127, 669–681. doi:10.1016/j.ympev.2018.06.013

Schindel, G.M., Gary, M., 2017. Hypogene Processes in the Balcones Fault Zone Segment of the Edwards Aquifer of South-Central Texas, in: Klimchouk, A., Palmer, A.N., De Waele, J., Auler, A.S., Audra, P. (Eds.), Cave and karst systems of the world. Springer, Cham, pp. 647–652. doi:10.1007/978-3-319-53348-3_41

Shringarpure, S., Xing, E.P., 2014. Effects of sample selection bias on the accuracy of population structure and ancestry inference. G3 Genes, Genomes, Genet. 4, 901–911. doi:10.1534/g3.113.007633

Slatkin, M., 2008. Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. 9, 477–485. doi:10.1038/nrg2361

Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. doi:10.1093/bioinformatics/btu033

Sun, Y.-B., Xiong, Z.-J., Xiang, X.-Y., Liu, S.-P., Zhou, W.-W., Tu, X.-L., Zhong, L., Wang, L., Wu, D.-D., Zhang, B.-L., Zhu, C.-L., Yang, M.-M., Chen, H.-M., Li, F., Zhou, L., Feng, S.-H., Huang, C., Zhang, G.-J., Irwin, D., Hillis, D.M., Murphy, R.W., Yang, H.-M., Che, J., Wang, J., Zhang, Y.-P., 2015. Whole-genome sequence of the Tibetan frog Nanorana parkeri and the comparative evolution of tetrapod genomes. Proc. Natl. Acad. Sci. U. S. A. 112, 1257–1262. doi:10.1073/pnas.1501764112

U.S. Congress, 1973. Endangered Species Act of 1973, An Act to provide for the conservation of endangered and threatened species of fish, wildlife, and plants, and for other purposes.

Wang, J., Hill, W.G., Charlesworth, D., Charlesworth, B., 1999. Dynamics of inbreeding depression due to deleterious mutations in small populations: mutation parameters and inbreeding rate. Genet. Res. 74, 165–178. doi:10.1017/S0016672399003900

Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7, 256–276. doi:10.1016/0040-5809(75)90020-9

Weir, B.S., 1997. Genetic Data Analysis II. Biometrics. doi:10.2307/2533134

Wiens, J.J., Chippindale, P.T., Hillis, D.M., 2003. When are phylogenetic analyses misled by convergence? A case study in Texas cave salamanders. Syst. Biol. 52, 501–514. doi:10.1080/10635150390218222

FIGURES

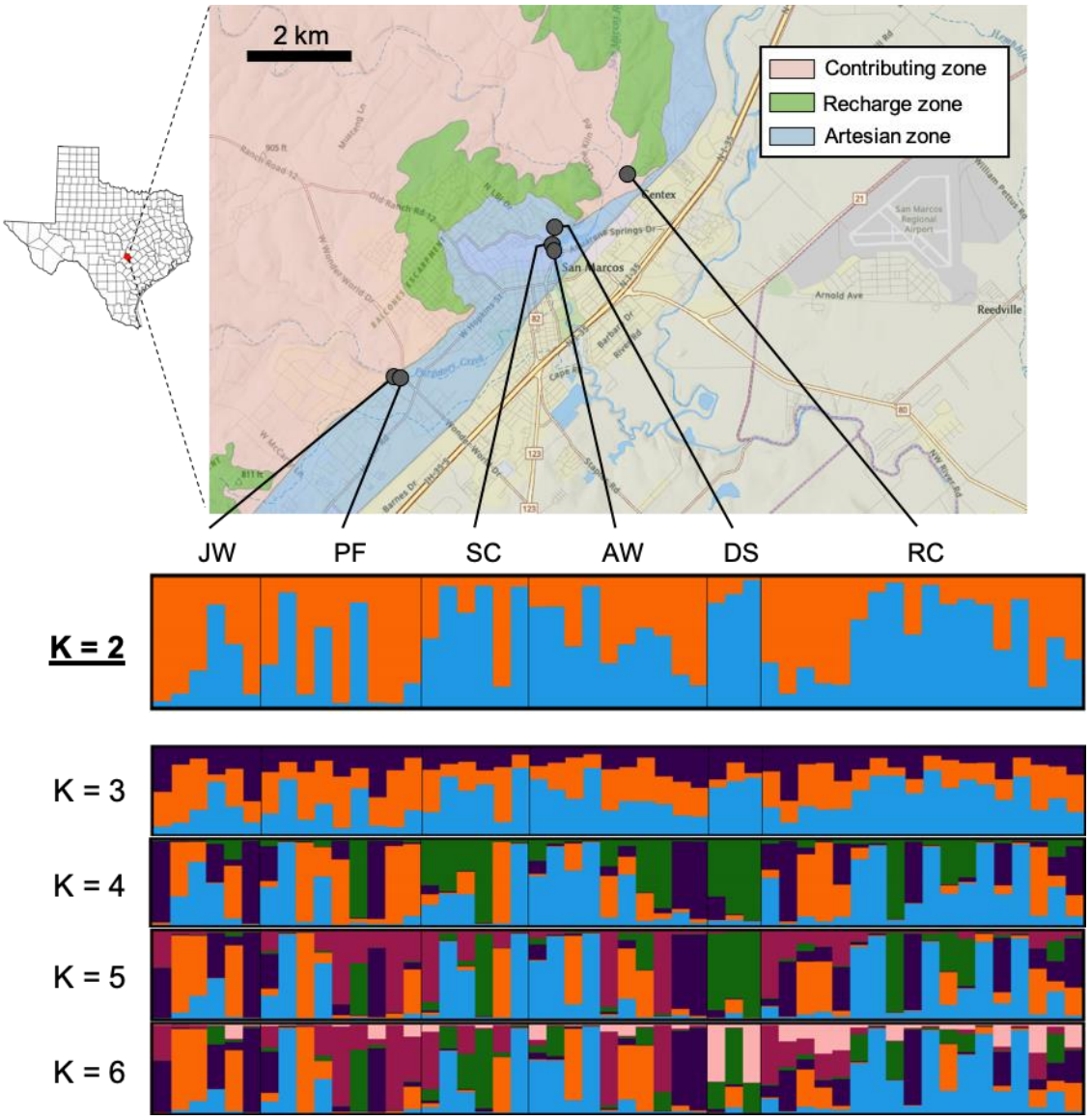


Figure 1. STRUCTURE plots showing *K* = 2-6 with localities displayed on a map of the San Marcos area, Hays Co., TX. JW = Johnson's Well, PF = Primer's Fissure, SC = Sessom Creek Spring, AW = Artesian Well, DS = Diversion Spring, RC = Rattlesnake Cave and Rattlesnake Well. Plot indicates that sampling localities do not show evidence of distinct population structure. Determining the optimal value of *K* using the Evanno method (see Methods) cannot find that *K* = 1 to be the optimal number of *K*. We suspect that *K* = 1 may be the most appropriate value of *K* due to the lack of sub-structure at *K* = 2, the Evanno method found *K* = 2 to be the optimal value of *K*.

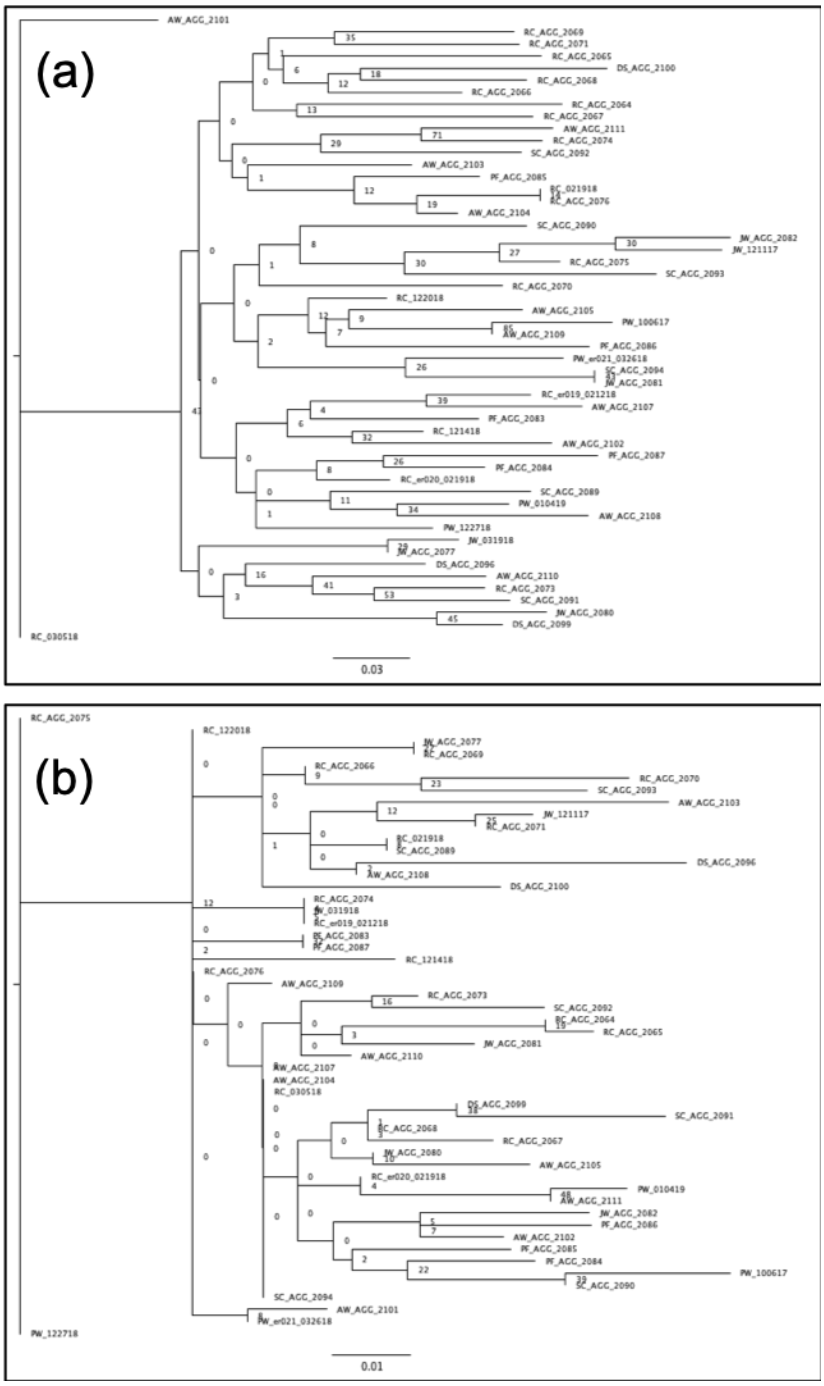


Figure 2. Maximum likelihood trees generated in RAxML using 1000 bootstrap replicates under the GTR_GAMMA model of nucleotide evolution. Scale bar represents number of changes per site. (a) allowing 75% missing data; (b) allowing 50% missing data. Sampling localities are included in sample names (JW = Johnson's Well, PF = Primer's Fissure, SC = Sessom Creek Spring, AW = Artesian Well, DS = Diversion Spring, RC = Rattlesnake Cave & Well). Poor bootstrap support may indicate the lack of differentiation within *E. rathbuni.*

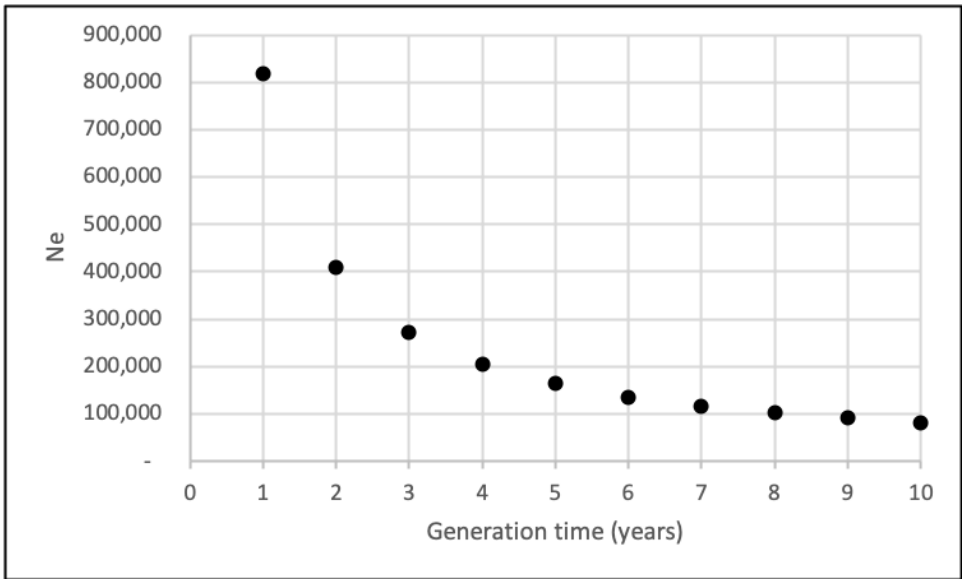


Figure 3. Estimates of N_e using Watterson's theta, estimated using *ThetaMater* across generation times of 1-10.

TABLES

Table 1. Pairwise AMOVA Fst using two different values of -p (number of populations that must be present in order to call a locus in STACKS). Requiring -p = 1 included 5695 loci, and -p = 3 included 978 loci. Northeast AW = Artesian Well, JW = Johnson's Well, PF = Primer's Fissure, RC = Rattlesnake Cave & Well, SC = Sessom Creek Spring.

Pairwise Comparison	p = 1	p = 3
AW x JW	0.0837476	0.0852053
AW x PF	0.0808147	0.0770739
AW x RC	0.0532448	0.0549660
AW x SC	0.1183710	0.0968649
JW x PF	0.0750572	0.0772519
JW x RC	0.0575693	0.0582094
JW x SC	0.1488730	0.1232330
PF x RC	0.0578019	0.0546952
PF x SC	0.1104940	0.1043020
RC x SC	0.0675670	0.0642233

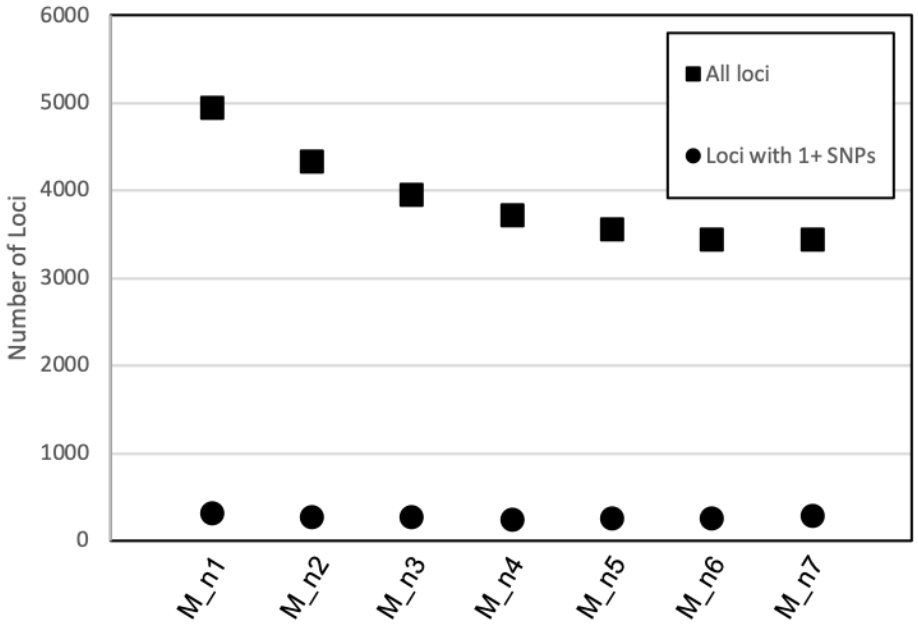
Table 2. Global diversity metrics for each sampling locality estimated using only variable sites.

Sampling Locality	Observed Het.	Variation (obs het)	Expected Het.	Variation (exp het)	Pi	Variation (Pi)	Fis	Variation (Fis)
Artesian Well	0.03569	0.00728	0.16200	0.03067	0.17683	0.03668	0.38502	0.22267
Johnson's Well	0.02381	0.00709	0.16521	0.03735	0.18675	0.04779	0.36571	0.23489
Primer's Fissure	0.04030	0.00774	0.14723	0.02912	0.16193	0.03522	0.31664	0.20499
Rattlesnake Cave & Well	0.06842	0.01273	0.16563	0.02316	0.17445	0.02575	0.386	0.20801
Sessom Creek Spring	0.03703	0.01299	0.11516	0.03236	0.13419	0.04444	0.20458	0.1622

Table 3. Estimates of effective population size (N_e) using NeEstimatorv2, including only the 40 samples with the highest coverage, yielding 387 loci.

Low Allele Frequency	0.05	0.03	0.02	0
Estimated Ne^	315.7	942.3	infinite	infinite
lower 95% Ci for Ne^	107.2	180.2	638.5	1279.4
upper 95% Ci for Ne^	infinite	infinite	infinite	infinite

SUPPLEMENTARY MATERIAL



Supplementary Figure 1. Number of loci and number of variable loci at each value of M = n examined in STACKS to find the optimal number of M = n for this dataset. The number of loci plateaus at M = n of 5, however the number of variable loci stays almost constant, indicating that multiple values of M = n may be appropriate as the number of loci containing at least one SNP was the same.

Supplementary Table 1. Samples of *E. rathbuni* included in this study, and whether they are in the captive breeding program or if they were released after taking a tail clip. Samples not demarcated as captive or released were previously collected and housed at the USFWS San Marcos Aquatic Resource Center (SMARC).

Sample ID	Sampling Location	Captive or Released
AW_AGG_2101	Artesian Well	SMARC
AW_AGG_2102	Artesian Well	SMARC
AW_AGG_2103	Artesian Well	SMARC
AW_AGG_2104	Artesian Well	SMARC
AW_AGG_2105	Artesian Well	SMARC
AW_AGG_2107	Artesian Well	SMARC
AW_AGG_2108	Artesian Well	SMARC
AW_AGG_2109	Artesian Well	SMARC
AW_AGG_2110	Artesian Well	SMARC
AW_AGG_2111	Artesian Well	SMARC
DS_AGG_2096	Diversion Spring	SMARC
DS_AGG_2099	Diversion Spring	SMARC
DS_AGG_2100	Diversion Spring	SMARC
JW_031918	Johnson's Well	Released
JW_121117	Johnson's Well	Released
JW_AGG_2077	Johnson's Well	SMARC
JW_AGG_2080	Johnson's Well	SMARC
JW_AGG_2081	Johnson's Well	SMARC
JW_AGG_2082	Johnson's Well	SMARC
PF_AGG_2083	Primer's Fissure	SMARC
PF_AGG_2084	Primer's Fissure	SMARC
PF_AGG_2085	Primer's Fissure	SMARC
PF_AGG_2086	Primer's Fissure	SMARC
PF_AGG_2087	Primer's Fissure	SMARC
PW_010419	Primer's Fissure	Captive
PW_100617	Primer's Fissure	Released
PW_122718	Primer's Fissure	Released
PW_er021_032618	Primer's Fissure	Captive
RC_021918	Rattlesnake Cave	Released
RC_030518	Rattlesnake Cave	Released
RC_121418	Rattlesnake Cave	Released
RC_122018	Rattlesnake Cave	Captive
RC_AGG_2064	Rattlesnake Well	SMARC

RC_AGG_2065	Rattlesnake Well	SMARC
RC_AGG_2066	Rattlesnake Well	SMARC
RC_AGG_2067	Rattlesnake Cave	SMARC
RC_AGG_2068	Rattlesnake Cave	SMARC
RC_AGG_2069	Rattlesnake Cave	SMARC
RC_AGG_2070	Rattlesnake Cave	SMARC
RC_AGG_2071	Rattlesnake Cave	SMARC
RC_AGG_2073	Rattlesnake Cave	SMARC
RC_AGG_2074	Rattlesnake Cave	SMARC
RC_AGG_2075	Rattlesnake Cave	SMARC
RC_AGG_2076	Rattlesnake Cave	SMARC
RC_er019_021218	Rattlesnake Cave	Captive
RC_er020_021918	Rattlesnake Cave	Captive
SC_AGG_2089	Sessom Creek Spring	SMARC
SC_AGG_2090	Sessom Creek Spring	SMARC
SC_AGG_2091	Sessom Creek Spring	SMARC
SC_AGG_2092	Sessom Creek Spring	SMARC
SC_AGG_2093	Sessom Creek Spring	SMARC
SC_AGG_2094	Sessom Creek Spring	SMARC

Supplementary Table 2. Global diversity metrics for each sampling locality estimated using all sites, variable and fixed.

Sampling Locality	Observed Het.	Variation (obs het)	Expected Het.	Variation (exp het)	Pi	Variation (Pi)	Fis	Variation (Fis)
Artesian Well	0.00082	0.00020	0.00374	0.00130	0.00408	0.00155	0.00888	0.00847
Johnson's Well	0.00073	0.00023	0.00509	0.00196	0.00576	0.00251	0.01128	0.01120
Primer's Fissure	0.00114	0.00026	0.00416	0.00142	0.00457	0.00171	0.00895	0.00854
Rattlesnake Cave & Well	0.00153	0.00039	0.00371	0.00112	0.00391	0.00124	0.00865	0.00792
Sessom Creek Spring	0.00091	0.00035	0.00283	0.00111	0.00330	0.00152	0.00503	0.00498