Fall 12-20-2013

Statistical and Comparative Phylogeography of Mexican Freshwater Taxa in Extreme Aquatic Environments

Lyndon M. Coghill
University of New Orleans, lyndon.coghill@gmail.com

Follow this and additional works at: https://scholarworks.uno.edu/td

Part of the Biodiversity Commons, Bioinformatics Commons, Biology Commons, Evolution Commons, Other Ecology and Evolutionary Biology Commons, and the Terrestrial and Aquatic Ecology Commons

Recommended Citation
https://scholarworks.uno.edu/td/1724

This Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UNO. It has been accepted for inclusion in University of New Orleans Theses and Dissertations by an authorized administrator of ScholarWorks@UNO. The author is solely responsible for ensuring compliance with copyright. For more information, please contact scholarworks@uno.edu.
12-27-2013

Statistical and Comparative Phylogeography of Mexican Freshwater Taxa in Extreme Aquatic Environments

Lyndon M. Coghill

Follow this and additional works at: http://scholarworks.uno.edu/td
Statistical and Comparative Phylogeography of Mexican Freshwater Taxa in Extreme Aquatic Environments

Submitted to the Graduate Faculty of the University of New Orleans in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Conservation Biology

by
Lyndon M. Coghill
B.S. Western Illinois University, 2004

December, 2013
Table of Contents

List of Figures .................................................................................................................... iv

List of Tables ....................................................................................................................... v

Abstract .............................................................................................................................. vi

Introduction ..........................................................................................................................1

Chapter 1: Phylogeography and conservation genetics of a distinct lineage of sunfish in the Cuatro Ciéñegas valley of Mexico .....................................................................................11

Abstract .............................................................................................................................. 11

Introduction ....................................................................................................................... 12

Materials and Methods ................................................................................................... 16

Results ............................................................................................................................... 21

Discussion ......................................................................................................................... 28

Conclusions ....................................................................................................................... 30

Acknowledgements ......................................................................................................... 31

Chapter 2: Comparative phylogeography of aquatic species endemic to Cuatro Ciéñegas, Mexico ............................................................................................................................... 32

Abstract .............................................................................................................................. 32

Introduction ....................................................................................................................... 33

Materials and Methods ................................................................................................... 38

Results ............................................................................................................................... 42

Discussion ......................................................................................................................... 52

Conclusions ....................................................................................................................... 54

Acknowledgements ......................................................................................................... 55
Chapter 3: Next Generation Phylogeography of Cave and Surface Astyanax mexicanus

Abstract .........................................................................................................................56
Introduction ....................................................................................................................57
Materials and Methods ...................................................................................................61
Results ............................................................................................................................65
Discussion ......................................................................................................................71
Conclusions ....................................................................................................................75
Acknowledgements ........................................................................................................76
References .....................................................................................................................77
Vita ...................................................................................................................................101
List of Figures

*Lepomis megalotis* sampling map in the Cuatro Ciénegas basin ........................................ 14

*Lepomis megalotis* haplotype network .................................................................................. 23

*Lepomis megalotis* Bayesian gene tree .............................................................................. 26

Estimates of gene flow based on Bayesian inferences of migration rates and population
sizes for *Lepomis megalotis* ........................................................................................................ 28

Map showing the valley of Cuatro Ciénegas and the split between eastern and western
populations .................................................................................................................................... 35

Gene genealogies for the snail taxa ............................................................................................ 44

Gene genealogies for the fish taxa .............................................................................................. 46

The tMRCA for the six taxa that were included in the study ................................................... 47

Results of the msBayes ABC unconstrained analysis showing estimated number of
divergence events using simple rejection .................................................................................... 48

Joint probability of Omega and $E(t)$ for the msBayes ABC unconstrained analysis .......... 49

Posterior distribution of timing $(E(t))$ events of two diversification events ....................... 50

Map of Mexico showing the *Astyanax mexicanus* sampling locations ................................ 60

Maximum likelihood phylogenetic tree of *A. mexicanus* populations
based on COI sequences ............................................................................................................ 66

Constrained maximum likelihood phylogenetic trees of *Astyanax mexicanus* using
NGS SNP data and rooted with the COI tree ............................................................................ 68

Best tree from the unconstrained SNP analysis for *Astyanax mexicanus* ............................ 70
List of Tables

List of *Lepomis megalotis* sampling localities, GPS coordinates, and sample size (*n*)
from each locality ..................................................................................................................17

Summary of genetic differentiation by region of *Lepomis megalotis* ...........................21

Comparison of gene flow models using Bayes factors for *Lepomis megalotis* ..........27

List of all taxa, the gene used in molecular analysis, amplified gene length (bp), mutation rate
(%), sample size and generation time (in years) used for gene flow analysis for the comparative
phylogeography study ........................................................................................................41

Estimates of gene flow between populations of *Lepomis megalotis*, *Astyanax mexicanus* and
*Herichthys minckleyi* among east and west haplogroups within the Cuatro Ciénegas valley
............................................................................................................................................51
Abstract

Phylogeography aims to understand the processes that underlie the distribution of genetic variation within and among closely related species. Although the means by which this goal might be achieved differ considerably from those that spawned the field some thirty years ago, the foundation and conceptual breakthroughs made by Avise are nonetheless the same and are as relevant today as they were two decades ago. Namely, patterns of neutral genetic variation among individuals carry the signature of a species’ demographic past, and the spatial and temporal environmental heterogeneity across a species’ geographic range can influence patterns of evolutionary change. Aquatic systems throughout Mexico provide unique opportunities to study phenotypic plasticity and evolution in relation to climatic and environmental selective forces. There are several unique, often isolated aquatic environments throughout Mexico that have a history of geographic isolation and reconnection. The first study presented herein shows significant mitochondrial sequence divergence was also discovered between *L. megalotis* populations on either side of the Sierra de San Marcos that bisects the valley of Cuatro Ciénegas and that the populations in the valley are genetically distinct from those found outside of the valley. The second study recovered signals of two divergence events in Cuatro Ciénegas for six codistributed taxa, and reveals that both events occured in the Pleistocene during periods of increased aridity suggesting that climatic effects might have played a role in these species’ divergence. The final study presents an Illumina-based high-resolution species phylogeny for *Astyanax mexicanus* providing added support that there are multiple origins to cave populations and further clarifying the uniqueness of the Sabinos and Rio Subterráneo caves.

Keywords

phylogeography, phylogenetics, bioinformatics, mexico, fish, aquatic, conservation
INTRODUCTION

History

In its original form, phylogeography attempted to unite the fields of phylogenetics and population genetics to examine organismal data in the context of its geographical distribution (Avise et al. 1987). Phylogeography provides a framework to examine hypotheses about the relationships of geographic features and climatic events to spatial patterns of genetic differentiation, and ultimately how various mechanisms may drive speciation. Most of the original studies focused on a single taxon, although Avise et al. (1987) also explored the possibility that data from codistributed taxa could enhance inquiries about how geological or climatic events could have generated similar patterns across numerous taxa.

Historically, mitochondrial DNA (mtDNA) in animal taxa has been used extensively as the genetic marker of choice due to its lack of recombination, assumed neutrality and a smaller time expected to reconcile reciprocal monophyly between gene regions (Hickerson et al. 2010a). This perceived connection between geographic patterns and genetic lineages can be problematic (i.e., lineage sorting) but presented a powerful enticement for researchers in the field (Irwin 2002). However, many researchers inferred that distinct clades within a species represent real population divisions allowing the exploration of the history of individual clade-defined lineages within a species complex. While this type of study is desirable, many early studies ignored the need for an appropriate model to address any statistical uncertainty (Avise 2000, Knowles and Maddison 2002b, Hey and Machado 2003).

Shortly after the founding of phylogeography, the field of population genetics went through a major change as theoretical researchers developed coalescent theory (Hudson 1983,
Coalescent theory has allowed researchers to use only a sample of the potential alleles in a population, modeling gene histories regardless of the demographic history. This allows the estimation of parameters that are of interest to phylogeographic researchers such as divergence time, migration rates and population sizes (Wakeley 2009). Until these methods became more common, most researchers still relied on traditional descriptive techniques that reconcile geographic patterns of gene genealogies with known demographic history and stochastic events.

**Descriptive Phylogeography**

For many years phylogeographic studies were based on more qualitative and descriptive interpretations of a single locus genealogies and the belief that a geographically restricted monophyly was equivalent to a taxon’s isolation. To add depth to these interpretations, researchers often include summary statistics (Nei and Li 1979, Slatkin and Maddison 1989), such as $F$ statistics (Wright 1950) and permutation tests (Tajima 1989, Excoffier et al. 1992) to look at partitions of genetic variance within and between populations. However, researchers now realize that, by ignoring random variation in patterns of gene inheritance (coalescent stochasticity), this can lead to over interpretation of demographic and evolutionary processes (Edwards and Beerli 2000, Arbogast et al. 2002a, Knowles 2008, Templeton 2009a, b). The methods presented here for Chapter 1 involve using descriptive phylogeography to identify a unique lineage of sunfish in Mexico living in an extreme aquatic environment, and to estimate divergence times. However, these methods often suffer from large variation in time estimates, which can make it difficult to distinguish between multiple divergent events. Overlap in time estimates can lead to inaccurate interpretations such as multiple events being considered a single large divergent event or a single event being interpreted as multiple, closely occurring pulses of divergence (Riddle and Hafner...
2006a, Hickerson and Carnaval 2011). These challenges have led researchers to choose more statistically rigorous methods emphasizing simulation-based methods. These methods use parameterized coalescent models to estimate the parameters as well as to explore various alternative hypotheses as is demonstrated in Chapter 2 (Chamberlin 1890, Knowles and Maddison 2002b).

**Statistical and Comparative Phylogeography**

A key goal of phylogeographic studies is to determine how abiotic and ecological factors drive evolutionary processes (Avise 2000, Arbogast and Kenagy 2001a). One way to achieve that goal is to examine groups of organisms or communities of species in a given environment and determine if they show similar patterns of evolutionary change (Lapointe and Rissler 2005). Several methods of statistical phylogeography have been developed recently that allow these multiple-taxon analyses. Statistical phylogeography introduced the concept of using model-based coalescent approaches to estimate parameters of interest and to test specific hypotheses (Knowles 2001, Knowles and Maddison 2002b). Unfortunately, the computational limitations of these models became apparent for models that contain a large number of parameters (Hey and Nielsen 2007a, Nielsen and Beaumont 2009b). One possible alternative to this computational limitation is the use of approximate Bayesian computation (ABC), which simulates the data using parameter values randomly drawn from the prior distribution to approximate the posterior (Beaumont *et al.* 2002). Recent techniques using coalescent theory to build hierarchical Bayesian models for hypothesis testing have offered solutions to some of these historically problematic situations. These techniques allow researchers to account for stochastic processes and biologically meaningful variation such as population size and incorporate mutation rate variation across loci (Hickerson *et al.* 2006b, Barber and Klicka 2010a). Although the ABC-based
comparative methods are still relatively novel, they have been used to test for simultaneous
divergence times across a few co-distributed taxa (Voje et al. 2009b, Barber and Klicka 2010a,
Huang et al. 2011). They have also shown promise in detecting vicariance or colonization in
wide ranging species (Hickerson et al. 2006b, Hickerson and Meyer 2008a). Another strength of
these methods is that they are sensitive enough to analyze data sets with small sample sizes
offering a solution to researchers for whom collection of some taxa is logistically difficult
(Huang et al. 2011). In Chapter 2 the ABC-based method is used to delineate between multiple
divergence events of six aquatic taxa in the northern Mexican valley of Cuatro Ciénegas. As
demonstrated these techniques offer increased power and resolution compared to traditional
descriptive techniques.

Application of Next Generation Sequencing for Inference

The use of multilocus datasets to infer population and species history can help resolve
problems that can arise from coalescent stochasticity (Edwards 2009, Knowles 2009). For this
reason multilocus datasets have become the standard used by most researchers in contemporary
studies aided by the decreasing cost of sequencing technologies. The complication of this need
for larger datasets is that for many organisms, especially non-model organisms, researchers must
spend a significant amount of time identifying suitable markers for their study system (Thomson
et al. 2010). This has led many researchers to explore the possibility of next-generation
sequencing (NGS) as a way to generate large, multi-locus datasets in a more cost and time-
efficient way (Ekblom and Galindo 2010, Lerner and Fleischer 2010, Eaton and Ree 2013).
Despite the potential of NGS datasets to aid phylogeographic researchers, the major challenges
of sample preparation and the post-collection bioinformatics processing has hampered the use of
NGS data in phylogeographic studies (Mardis 2008, McCormack et al. 2011). The first challenge
has mostly been overcome by producing reduced representation libraries of the organismal genomes, for various several sequencing platforms and chemistries available that allow the creation of libraries for both model and non-model organisms (Baird et al. 2008, Hamady et al. 2008, Meyer et al. 2008, Emerson et al. 2010, Hohenlohe et al. 2010a, Glenn 2011, Neiman et al. 2011). Solutions to the second set of challenges are still very much in development. First, the logistical difficulties of large datasets are a primary issue for bioinformatics and computer network engineers. A traditional Sanger dataset will usually contain less than 1,000 sequences needing hundreds of megabytes in storage. Whereas a normal NGS dataset will likely possess millions of reads and require a terabyte or more of storage. Working with a dataset of this size would require extensive upgrades in computer hardware for many researchers. Another major challenge is that determining the quality of NGS data can also be difficult. Unlike Sanger sequencing which provide a chromatogram to directly assess the quality by showing the strength of each nucleotide’s signal, with NGS data you get an integer value that is an estimate of the confidence value of that call for each nucleotide. This can provide a challenge to researchers when attempting to determine orthologous versus paralogous loci, as in Sanger sequencing the chromatogram is used to detect the signal of more than one gene copy. NGS data is single stranded, and therefore you cannot detect paralogy until after the alignment stage, where it becomes possible to detect a signal that an excessive number of alleles have been included into a single locus, such as three alleles for a diploid locus.

Since most NGS reads are shorter, and they are pulled from the entire genome, alignment can be difficult. If there is a reference genome, or one of a closely related sister taxa, alignment can be conducted quickly using this reference. In the case of organisms without a completed genome, the alignment must be done *de novo*. This usually involves assembling the reads into
some type of cluster or stack that meet a percentage threshold of similarity, and then aligning these groups. There have recently been several software pipelines that have been developed to aid in the creation and alignment of clusters, though most still require some level of computer and coding proficiency to use effectively (Hird et al. 2011, Catchen et al. 2013).

Phylogeography is becoming more genomic in its nature, and in time this will allow the combination of phylogeography with currently disparate fields like ecological genomics (Sims et al. 2009, Emerson et al. 2010). Ecological genomics has utilized candidate and expressed genes to identify processes driving evolution, whereas phylogeography has traditionally focused on putatively neutral-evolving markers to answer similar questions (Stapley et al. 2010, Rice et al. 2011). This interest has already spurred similar analytical techniques between these two fields using restriction-digest based NGS based libraries to help identify outlier loci that may be under selection from large numbers of loci experiencing neutral divergence (Emerson et al. 2010, Gompert et al. 2010, Hohenlohe et al. 2010a, Holsinger 2010). These breakthroughs offer opportunities to begin using NGS data to continue looking at how adaptive speciation may be driving some of the phylogeographic patterns we observe in nature.

In the final chapter presented here, NGS data is used to take advantage of these novel techniques to examine the phylogeography of cave and surface populations of an increasingly popular model organism, *Astyanax mexicanus*. Using these large datasets has allowed the recovery of a higher resolution phylogeny than any previous study using more traditional markers. This phylogeny is then used to statistically test alternative hypotheses of independent cave origins. Finally, the number of transitions between cave and surface populations is estimated by mapping ancestral habitat states onto the best estimate of the cave and surface *A. mexicanus* phylogeny.
Empirical Applications

Aquatic environments with extreme thermoclines, hypoxic conditions or high salinity often have distinctive faunas and florae, and are a focus of conservation interest (Culver et al. 1995, Johnson 2005, Chaves-Campos et al. 2011a). Phylogenetic evidence suggests such regions are often invaded by taxa that were adapted to similar conditions elsewhere. But in other cases, inhabitants appear to be derived from forms that occurred at the edge of the extreme aquatic environment (Ibrahim et al. 1996, Roman and Darling 2007, Duckworth 2008). The invasion of extreme habitats from the periphery and the evolution of many derived characters in extreme environments are the result of an iterative process involving rounds of speciation, displacement, and adaptation to increasingly more severe environments. In fluctuating climatic conditions, where individual habitats are likely to fragment and later rejoin (i.e., periods of flooding interrupted by xeric periods), lineages initially confined to the periphery may evolve in isolates of their original habitat and then come into contact as these coalesce. Aquatic systems throughout Mexico provide unique opportunities to study phenotypic plasticity and evolution driven by these types of adaptations to “fringe” or extreme environments. There are several unique, often isolated aquatic environments throughout Mexico that have a history of geographic isolation and reconnection. This region has received a fair amount of attention as it serves as a zone of transition between the Neotropi and Neartic zones. This dissertation focuses on aquatic taxa living in Mexican subterranean aquatic cave systems and high salinity freshwater environments.
Phylogeography and Conservation

Stochastic geologic and climatic events can lead to isolation of organisms into extreme peripheral environments. These extreme environments often drive evolutionary changes and local adaptation to these specific conditions (Johnson 2005, Chaves-Campos et al. 2011a, Gross 2012). This can lead to the formation of cryptic species (Bickford et al. 2007, Alcántara-Rodríguez et al. 2012, Cooke et al. 2012, Wilson et al. 2012) causing some of these environments to have the highest levels of endemicity on the planet (Whitaker and Banfield 2005, Hendrickson et al. 2008a). Phylogeographic studies can be used to obtain a temporal context for major population subdivision seen across these extreme environments, and to facilitate inferences of the historical forces that have produced contemporary patterns of population structure (De Guia and Saitoh 2007). Determining the distinctiveness and age of populations especially in highly threatened habitats is essential to both managers and policy makers attempting to identify the population units most in need of conservation. Genetically identifying unique, persistent lineages of organisms can also address the impact that the loss of particular populations would have on overall biodiversity (Johnson 2005, Chaves-Campos et al. 2011a, Gross 2012).

The Cuatro Ciénegas valley in northern Mexico exhibits the highest level of endemism in North America, but the genetic distinctiveness of many species and populations within the valley remains unclear (Hendrickson et al. 2008a). Because of its biological uniqueness, Cuatro Ciénegas has been designated a National Protected Area by the Mexican Government, a RAMSAR site (intergovernmental treaty protected wetland) as well as an UNESCO World Heritage Biosphere Reserve (Hendrickson et al. 2008a, Souza et al. 2008a). This relatively small (~1500km$^2$) intermontane valley located in the Chihuahuan desert contains numerous aquatic
habitats and is home to more than 70 endemic species (Souza et al. 2008a). The valley is located in the center of an extremely arid region and virtually all of the endemic species are found within its more than 200 permanent pools, rivers, and lakes. These water bodies are isolated into several hydrologically distinct drainages that were historically separated from aquatic connections outside of the valley (Hendrickson et al. 2008a). The closest external drainage to the valley is the Río Salado de los Nadadores basin, but no natural aquatic connection exists between the two areas. However, several canals that carry water from the valley to agricultural land outside the valley have been constructed (Johnson 2005, Hendrickson et al. 2008a, Chaves-Campos et al. 2011a).

Subterranean species often have extreme adaptations to living in those environments compared to similar surface taxa (Poulson 1963, Tobler et al. 2006, Plath et al. 2007, Riesch et al. 2010, Gross 2012). Numerous recent studies have highlighted the genetic variation in cave organisms that are often morphological similar, but genetically and geographically different (Juan and Emerson 2010, Bradic et al. 2012b, Gross 2012, Niemiller et al. 2012). Cave ecosystems often provide long-term environmental stability compared to surface ecosystems and this can lead to high levels of endemism (Verovnik et al. 2003, Finston et al. 2007, Gross 2012). The fragmented nature and strong selective pressures of cave systems can lead to stable populations that are geographically isolated for long periods of time, promoting the development of cryptic species (Culver et al. 1995, Lefébure et al. 2006, Culver and Pipan 2009, Gross 2012). These populations are often morphologically indistinguishable from each other, but the genetic differentiation is often high (Lefébure et al. 2006, Finston et al. 2007, Bradic et al. 2012b). The discord between these two types of characters is due to convergence or parallel evolution (Culver and Pipan 2009). The high numbers of genetically distinct, but yet undiscovered species in these
environments highlights the need for continued phylogeographic studies to help identify unique lineages. Sadly, these ecosystems are labeled in most regions as extremely vulnerable with as many as 95% of their known endemic species threatened or endangered (Culver et al. 2000), emphasizing the need for further data to help aid in their management.
CHAPTER 1: Phylogeography and conservation genetics of a distinct lineage of sunfish in the Cuatro Ciénergas valley of Mexico

ABSTRACT

The valley of Cuatro Ciénergas, an aquatic oasis located in the Mexican Chihuahuan desert, exhibits the highest level of endemism in North America and is a Mexican National Protected Area. However, little is known about the evolutionary distinctiveness of several vertebrate species present in the Cuatro Ciénergas valley. We conducted a phylogeographic study using mitochondrial haplotypes from the centrarchid fish *Lepomis megalotis* to determine if the populations found within the valley were evolutionarily distinct from populations outside the valley. We also examined if there was evidence of unique haplotypes of this sunfish within the valley. Genetic divergence of *L. megalotis* suggests populations within the valley are evolutionarily unique when compared to *L. megalotis* outside the valley. Significant mitochondrial sequence divergence was also discovered between *L. megalotis* populations on either side of the Sierra de San Marcos that bisects the valley. Our results reinforce previous studies that suggest the organisms occupying aquatic habitats not only within Cuatro Ciénergas but also in each of the two lobes of the valley generally deserve independent consideration during management decisions.
INTRODUCTION

Phylogeographic analyses based on molecular markers are now widely used in conservation studies to identify unique evolutionary lineages. These analyses can clarify the evolutionary context of organismal diversification especially when combined with various geological and climatic events (Moritz 2002). Examination of the spatial patterns of intraspecific gene flow can also lead to the discovery of cryptic but genetically distinct populations (Suchard et al. 2005, Fujita et al. 2010, Heled and Drummond 2010). In addition, molecular phylogeographies can be used to obtain a temporal context for major population subdivision and facilitate inferences of the historical forces that have produced contemporary patterns of population structure (De Guia and Saitoh 2007). Determining the distinctiveness and age of populations especially in highly threatened habitats is essential to both managers and policy makers attempting to identify the population units most in need of conservation. Genetically identifying unique, persistent lineages of organisms can also address the impact that the loss of particular populations would have on overall biodiversity (Johnson 2005, Bradic et al. 2012b).

Within this framework, we examine the population structure and temporal divergence of long-eared sunfish populations, *Lepomis megalotis*, in a hotspot of aquatic endemcity.

The Cuatro Ciénegas valley exhibits the highest level of endemism in North America, but the genetic distinctiveness of many species and populations within the valley remains unclear (Hendrickson et al. 2008a). Because of its biological uniqueness, Cuatro Ciénegas has been designated a National Protected Area by the Mexican Government, a RAMSAR site (intergovernmental treaty protected wetland) as well as an UNESCO World Heritage Biosphere Reserve (Hendrickson et al. 2008a, Souza et al. 2008c). This relatively small (~1500km$^2$) intermontane valley located in the Chihuahuan desert contains numerous aquatic habitats and is
home to more than 70 endemic species (Souza et al. 2008c). The valley is located in the center of an extremely arid region and virtually all of the endemic species are found within its more than 200 permanent pools, rivers, and lakes. These water bodies are also isolated into several hydrologically distinct drainages that were historically separated from aquatic connections outside of the valley (Hendrickson et al. 2008a). The closest external drainage to the valley is the Río Salado de los Nadadores basin, but no natural aquatic connection exists between the two areas. However, several canals that carry water from the valley to agricultural land outside the valley have been constructed (Johnson 2005, Hendrickson et al. 2008a, Chaves-Campos et al. 2010). These man-made hydrologic connections could have provided an avenue for putatively non-endemic species such as the long-ear sunfish, Lepomis megalotis, to invade Cuatro Ciénegas and spread to numerous parts of the valley (Rafinesque 1820). Alternatively, the Cuatro Ciénegas lineage of Lepomis megalotis could be an endemic evolutionary lineage, and like many of the valley’s other aquatic species, it could show substantial phylogeographic substructure within the valley.

Within the valley, the Sierra de San Marcos demarcates a deep genetic subdivision for several species. This mountain splits the valley into eastern and western partitions (Fig. 1).
Figure 1 The Cuatro Ciénegas basin, Río Salado de los Nadadores, and the valley’s general location in Northern Mexico. The inset shows an enlarged diagram of the valley geography, and labels the various sampling locations with dots. Several major man-made canals that now occur within the valley and that all ultimately meet and flow northeast into the Río Salado de los Nadadores are depicted with dotted lines. Alabama sampling location is not shown.
The two endemic pupfish (*Cypriodon* spp.), largemouth bass within the valley (*Micropterus* spp.), one of the endemic aquatic snails (*Mexipyrgus churinceanus*), and the endemic freshwater shrimp (*Palaemonetes suttkusi*) all show patterns of geographic isolation on either side of this Sierra (Rodríguez-Martínez 2004, Johnson 2005, Carson and Dowling 2006, Chaves-Campos *et al.* 2010). However, other species with a relatively high capacity for dispersal like the endemic box turtle (*Terrapene coahuila*) exhibit little population structure within the valley (McGaugh 2012). Most of the species that show high levels of population genetic structure are obligate aquatic species, and those that show little differentiation are capable of crossing small parts of dry land. However, although *L. megalotis* is restricted to aquatic habitats, it does have an extensive range outside the valley. Therefore, this sunfish might be predicted to show limited genetic structure within the Cuatro Ciénegas valley and could even exhibit little divergence between populations found inside and outside of the valley.

*Lepomis megalotis* is one of several species that are found both within Cuatro Ciénegas and in the adjacent Río Salado drainage that ultimately drains into the Río Grande (Miller *et al.* 2006) (Fig. 1). Like the large-mouth bass, *Micropterus salmoides*, that also occurs in both areas, *L. megalotis* could have easily been introduced into the Rio Salado or Cuatro Ciénegas due to their popularity as a game fish (Lee *et al.*, 1980; Near *et al.*, 2004). The native range of *L. megalotis* in North America extends from Ohio to the Gulf of Mexico (Lee *et al.* 1980, Near *et al.* 2004), and its native range is believed to include parts of Northeastern Mexico (Miller *et al.* 2006). However, no studies have examined whether populations in Mexico represent divergent entities, and the wide-ranging *L. megalotis* species complex could exhibit substantial genetic structure in many parts of its range (Husemann *et al.* 2012).
The primary goal of this study was to determine whether the *Lepomis megalotis* populations found within the Cuatro Ciénegas basin are genetically unique and should receive increased conservation attention. In order to investigate this idea, three specific questions were examined. First, we asked whether *L. megalotis* mitochondrial haplotypes from within the valley are highly divergent from haplotypes outside of the Cuatro Ciénegas basin. Second, we determined whether populations within the valley show phylogeographic structure. Third, we tested several gene flow models to determine whether contemporarily isolated populations of *L. megalotis* in Cuatro Ciénegas exhibit evidence of recent gene flow.

**MATERIALS AND METHODS**

**Ethics Statement**

This study was conducted in Mexico as a part of an international, multi-taxa study and was approved by the Mexican Government and SEMARNAT (The Ministry of Environment and Natural Resources for Mexico) which approved all field and laboratory protocols under (Permit No. N°DAPA/2/130409/0961 and DAN-01202).

**Sampling and Laboratory Procedures**

Samples of *L. megalotis* were collected in June 2009 and August 2010 from several sites in the Cuatro Ciénegas basin as well as several locations from outside the valley. Within the Cuatro Ciénegas basin, we sampled 6 sites that spanned the geographic breadth of the valley (Fig 1). Sample locations, sample size and GPS coordinates are given in Table 1. Samples of *L. megalotis* collected outside the valley were obtained from the Río Salado drainage directly
outside of the valley and also from Texas and Alabama. In total, tissue samples from 77 individuals were examined.

Table 1 List of *Lepomis megalotis* sampling localities, GPS coordinates, and sample size (n) from each locality. From the left, columns show the name of the sampling locations in both Mexico and the U.S., the GPS coordinates of the sites, and the sample size examined from each location.

<table>
<thead>
<tr>
<th>Sampling Location</th>
<th>Latitude</th>
<th>Longitude</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juan Santos, Cuatro Ciénegas, Coahuila</td>
<td>26°53.859′N</td>
<td>102° 08.807′W</td>
<td>17</td>
</tr>
<tr>
<td>Poza Bonita, Cuatro Ciénegas, Coahuila</td>
<td>26° 50.232′N</td>
<td>102° 08.438′W</td>
<td>24</td>
</tr>
<tr>
<td>Pozas Azules, Cuatro Ciénegas, Coahuila</td>
<td>26° 49.730′N</td>
<td>102° 01.683′W</td>
<td>11</td>
</tr>
<tr>
<td>Río Mesquites, Cuatro Ciénegas, Coahuila</td>
<td>26° 55.378′N</td>
<td>102° 6.753′W</td>
<td>11</td>
</tr>
<tr>
<td>Río Salado, Coahuila</td>
<td>27° 02.059′N</td>
<td>101° 43.300′W</td>
<td>1</td>
</tr>
<tr>
<td>Tío Candido, Cuatro Ciénegas, Coahuila</td>
<td>26° 52.225′N</td>
<td>102° 04.740′W</td>
<td>4</td>
</tr>
<tr>
<td>Brazos River, Texas</td>
<td>30° 53.016′N</td>
<td>95° 17.3591′W</td>
<td>6</td>
</tr>
<tr>
<td>Uphapee Creek, Alabama</td>
<td>32° 28.053′N</td>
<td>85° 47.059′W</td>
<td>4</td>
</tr>
</tbody>
</table>

For all 77 individuals, we sequenced three mtDNA genes comprising 2839bp (ND2: 1047, Cytb: 1140 and COI: 652). First, DNA was extracted in the laboratory from fin tissue using Qiagen Blood and Tissue Kit (Qiagen). Primers for the three gene regions used were taken from published studies: Cytochrome \( b \) F: CTGCCCTCAAAACATTITCA R: GGGTTGGGGGAGAATAAGGCTAA, 53°C, (Strecker *et al.* 2004a); Cytochrome \( c \) oxidase subunit I, F: TCAAACAAACCACAAAGACATTGAC, R: TCGACTAATCATAAAGATATCGGCAC, 54°C (Ward *et al.* 2005); NADH dehydrogenase 2, F: CTACCTGAAGAGATATCGGCAC, R: CGCGTTTAGCTGTAACTAA, 55°C, (Kocher *et al.* 1995). Amplifications were carried out in a BioRAD iCycler Gradient thermocycler and conditions generally consisted of an initial denaturation step of 94 °C (2.0 min) followed by 35
cycles between 54-60 °C (30 s), and 72 °C (1.5 min). A final incubation of 72 °C for 4 min was added to ensure complete extension of products. Positively amplified DNA was then purified using an enzymatic combination of 1 μl of Exonuclease I (10.0 U/μl) and 1 μl shrimp alkaline phosphatase (2.0 U/μl) per 10μl of PCR product. Treated PCR products were sequenced at the W.M. Keck Conservation and Molecular Genetics Laboratory at the University of New Orleans using the same primers utilized for amplification. Complete gene sequences were assembled from individual reactions using the program Geneious version 5.3.6 (Biomatters 2010). All sequences were submitted to GenBank (LCXXXXXX-LCXXXXXX). Additional sequences for the outgroup comparison were collected from GenBank (AY517741, JN027026 and AY828969).

**Population Structure**

Following previous studies (Carson and Dowling 2006, Johnson *et al.* 2007, Chaves-Campos *et al.* 2010) regions were initially defined based on geographic boundaries based on the position of the sampling sites relative to the Sierra (Fig. 1). Pozas Azules, at the far southeastern edge of the valley, was defined as a unique region based on its geographic isolation. The Rio Mesquites and Tio Candido along the eastern edge of the Sierra were grouped to form the “Eastern” region based on the genetic structure of other organisms (Carson & Dowling, 2006; Chaves-Campos *et al.*, 2010). Pozas Bonita and Juan Santos were the locations sampled along the western side of the Sierra and make up the “Western” region. Because of their isolation from other bodies of water and distance from one another, these sites were initially treated as independent regions. In order to evaluate population structure, we performed an Analysis of Molecular Variance (AMOVA) on these four regions (Pozas Azules, Eastern, Juan Santos and
Poza Bonita) using ARLEQUIN 3.5 to examine differences among the sampled regions within
the valley (Excoffier and Lischer 2010).

**Phylogeographic Analysis**

For the phylogeographic analyses, sequences were aligned with MUSCLE, and a
haplotype network was constructed using the median joining method (Bandelt *et al.* 1999)
Unique haplotypes were coded according to the regions mentioned above, and jMODELTEST
2.0 was used to choose the best fitting, least-parameter rich model of sequence evolution based
on Bayesian Information Criterion (Darriba *et al.* 2012b). The program BEAST v 1.7.4 was then
used to simultaneously estimate a gene tree and the divergence of haplotypes among regions
(Drummond *et al.* 2012). We partitioned and applied the appropriate model of molecular
evolution to each gene (HKY+G for COI and ND2, and GTR+G for Cytb). An uncorrelated log
normal relaxed clock was used to estimate divergence times based on a fossil-calibrated split
between *L. megalotis* and *L. marginatus* of 1.72 ± 0.83 million years (Near *et al.* 2003). The
relaxed clock, uncorrelated lognormal model allows simultaneous estimation of phylogeny and
divergence times (Drummond *et al.* 2006b). Two primary analyses were conducted. The first
constrained the individuals from Cuatro Ciénegas to be a monophyletic clade. The second
allowed all individuals to be assigned to any particular clade during the analysis. Each analysis
was run for 10,000,000 generations starting with a random starting tree, constant size coalescent
prior, and a burn-in of at least 1,000,000 (Hastings 1970, Rambaut and Drummond 2007). The
analysis was repeated three times to confirm the robustness of the topology and divergence time
estimates (Rambaut and Drummond 2007).
For our two models examining population structure in Cuatro Ciénegas, the BEAST output was inspected and analyses of Bayes Factors were performed using Tracer 1.5. This allowed us to examine the posterior distributions, to check for convergence, and to confirm that the effective sample size for each parameter exceeded 200 (Rambaut and Drummond 2007). Posterior probabilities and the “maximum clade credibility tree” were calculated using TreeAnnotator 1.5.4 (Rambaut and Drummond 2007).

**Gene Flow Analysis**

The grouping pattern and splitting order of divergent populations recovered in the BEAST gene tree were used to estimate gene flow under the coalescent in MIGRATE-N 3.5.1 (Beerli and Felsenstein 2001, Beerli 2006, Beerli and Palczewski 2010). Four migration models were tested: (1) bi-directional gene flow between all 3 well supported clades (Pozas Azules, Eastern and Western) recovered from the phylogeographic analysis, (2) two populations divided strictly by the Sierra de San Marcos (Eastern and Western), (3) a split between Pozas Azules and the remainder of the valley, (4) a panmictic model assuming open gene flow between all populations. Our MIGRATE-N 3.2.6 analyses were implemented with default parameters except for modifications to run-length, heating, and relative mutation rate that were specific to the different migration models. To calculate marginal likelihoods for the model comparisons, we used a heating scheme of 1.00, 1.50, 3.00, and 1,000,000.00. After the runs were completed, results of each model were compared using Bayes Factors calculated from the model probabilities as described in MIGRATE-N (Kass and Raftery 1995, Beerli and Palczewski 2010).
RESULTS

Population Structure

Mitochondrial haplotype diversity among populations of *L. megalotis* was substantial. A total of 26 unique haplotypes were recovered from the 77 individuals sampled across the entire valley (Fig. 2). Distances among these unique haplotypes ranged from 0.1% to 7.4%. The AMOVA (Table 2) showed that the haplotypes were not homogeneously distributed in the Cuatro Ciénegas valley: 74% of sequence variation is due to differences among regions (Pozas Azules, Eastern Valley, Juan Santos and Poza Bonita), while the remaining 26% is due to differences found within those regions ($F_{ST} = 0.92, P < 0.001$). Pairwise $F_{ST}$ values between all of the regions were high (>0.71), and most were significantly different from zero. The exceptions were comparisons among the Eastern Valley (Río mesquites and Tio Candido), Juan Santos and Poza Bonita, which had lower $F_{ST}$ values (<0.58).

Table 2 Summary of genetic differentiation by region of *Lepomis megalotis*. Pairwise $F_{ST}$ values are presented below the diagonal. The corresponding $P$-values of significance from zero are presented above the diagonal.

<table>
<thead>
<tr>
<th></th>
<th>Pozas Azules</th>
<th>Eastern Valley</th>
<th>Juan Santos</th>
<th>Poza Bonita</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pozas Azules</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Eastern Valley</td>
<td>0.78</td>
<td>&lt;0.0001</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Juan Santos</td>
<td>0.92</td>
<td>0.58</td>
<td>&lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>Poza Bonita</td>
<td>0.88</td>
<td>0.71</td>
<td>0.41</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
The haplotype network analysis recovered several unique haplotype clusters within Cuatro Ciénegas that largely fell along sampling localities. These were: 1) Western Valley (Juan Santos and Poza Bonita). 2) Río Mesquites and Tío Candido (Eastern Valley), as well as 3) Pozas Azules in the southeastern lobe of the valley.
Figure 2 Haplotype network generated using a median-joining method. Pie graphs are proportional to the haplotype frequencies. Branch lengths are roughly proportional to the number of mutational steps between nodes. The number of steps is shown near each branch.
**Phylogeographic Analysis**

The BEAST analyses identified major phylogeographic structure within the valley. There was a clear division between populations within the valley and those found outside the valley with a posterior probability support of 1.0 and an estimated divergence time of 1.75 million years (Fig. 3). All of the individuals within the valley share a most recent common ancestor. However, within the Cuatro Ciénegas valley there was support for splitting *L. megalotis* into three distinct phylo-groups. The timing of the oldest split recovered suggests that the Eastern and Western populations (Poza Bonita and Juan Santos) of the valley diverged approximately 0.55 million years ago (posterior probability support of 0.99). Within the Eastern valley clade, the individuals from Tio Candido and Rio Mesquites clustered together. These populations were inferred to have diverged from the Pozas Azules clade approximately 0.40 million years ago (posterior probability support 0.98). The Juan Santos and Poza Bonita groups form a distinct clade on the Western side of the valley that is well supported. There also was a phylogeographic split between Juan Santos and Poza Bonita individuals based on the haplotype network and AMOVA results. However, the posterior probability support for this divergence was low (0.64). While most of the individuals were found in only one geographically defined clade, three individuals from the Eastern Valley did fall out within the Pozas Azules clade. Four individuals from Pozas Azules also fell out in the primarily Eastern clade. There was also a few shared haplotypes between the Poza Bonita clade and Juan Santos clade. Two Juan Santos individuals grouped with Poza Bonita and one individual from Poza Bonita grouped with the primarily Juan Santos haplotypes. However, it is important to note that no haplotypes were shared between the Poza Azules + Eastern clade and the Western Clade. Additionally, Bayes factor analyses supported the
monophyly of the Cuatro Ciénegas valley clade with a log_{10} Bayes factor value of 3.2 indicating monophyly is highly (1000) times more likely than non-monophyly.
Figure 3 Bayesian gene tree estimated from 77 individuals using 2839bp (ND2: 1047, Cytb: 1140 and COI: 652) of the mitochondrial genome. Geographically isolated regions within the valley are highlighted with shading for emphasis. An * denotes posterior probability support greater than 0.98.
Gene Flow

The MIGRATE-N 3.5.1 (Beerli 2006) results suggest that levels of gene flow were overall fairly minor across the valley with most populations experience less than 1 migrant per generation (Fig. 4). The highest levels of inferred migration were found between Pozas Azules and the Eastern populations. However, the median levels of migration between even these two populations were still quite low with approximately 2 migrants between these populations per generation. Overall, the gene flow analysis supports high levels of genetic structure and low levels of migration. With a probability of 0.912 (Table 3), the Bayes factor analyses suggested that among the models tested, the model defining three distinct populations (Pozas Azules, Eastern Valley and Western Valley) is the best-supported characterization of *L. megalotis* population subdivision within Cuatro Ciénegas.

Table 3  Comparison of gene flow models using Bayes Factors. Between 3 populations, eastern valley, western valley and Pozas Azules. Eastern valley populations (E) and western valley populations (W), between Pozas Azules (PA) and the rest of the valley (V) and a complete panmictic single population. Estimates of model probabilities derived from using summarized log marginal likelihoods and natural log Bayes factors. Model of the highest probability is reported in bold. Harmonic means are reported but were not used in the analysis, as the variance in the harmonic mean is generally too large to recover the best model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Structure</th>
<th>Bezier lML</th>
<th>Harmonic lML</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3 Pops</td>
<td>-3022</td>
<td>-3241</td>
<td>0.912</td>
</tr>
<tr>
<td>B</td>
<td>2 Pops (E&amp;W)</td>
<td>-3221</td>
<td>-3065</td>
<td>0.058</td>
</tr>
<tr>
<td>C</td>
<td>2 Pops (V&amp;PA)</td>
<td>-3111</td>
<td>-2781</td>
<td>0.030</td>
</tr>
<tr>
<td>D</td>
<td>Panmictic</td>
<td>-3267</td>
<td>-2930</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Figure 4 Estimates of gene flow based on Bayesian inferences of migration rates and population sizes. M is the estimated migration rate, scaled for the appropriate mutation rate between the different population clusters. Box plots represent values from the lower (25%) to upper (75%) quartiles with the median value marked as a +. The lines extend from the 2.5% to 97.5% percentile.

**DISCUSSION**

The findings presented here suggest that the populations of *L. megalotis* within the valley are highly divergent from *L. megalotis* populations found outside Cuatro Ciénegas. The level of divergence observed between populations found inside the valley and outside the valley leads us to infer that *L. megalotis* did not invade the valley in the past 0.5 million years. Populations of *L. megalotis* also exhibit a substantial amount of genetic differentiation and little gene flow among the various populations examined within the valley. The observed genetic differentiation has a number of conservation implications for this fish and its unique habitat (Vignieri et al. 2006, Sinclair et al. 2011).
All individuals from Cuatro Ciénegas form a monophyletic clade that likely split from the *L. megalotis* populations outside the valley approximately 1.0 – 2.3 million years ago. This substantial divergence from other *L. megalotis* populations mirrors what has been found in other species present both inside the valley and in the Río Salado, the closest watershed to Cuatro Ciénegas (Chaves-Campos *et al.* 2010). This result also is consistent with suggestions by Smith (1984) who reported that the deserts of North America experienced a cycle of heavy precipitation between 1.3 and 3.2 Myr ago that might have led to connections between Cuatro Ciénegas and external drainages. Importantly, the timeframe of genetic divergence within *L. megalotis* suggests that the populations of *L. megalotis* in the Cuatro Ciénegas basin are likely native to this region and evolutionarily distinct from populations found outside the valley. More focused conservation efforts for this distinctive Cuatro Ciénegas lineage of sunfish should be considered (Du Toit 2010).

We also found that populations within the valley show high levels of phylogeographic structure and relatively ancient population divergence. Our molecular clock estimates indicate that the Pozas Azules region likely split from other populations in the Eastern region approximately 400,000 years ago. Additionally, the populations from the Poza Azules + Eastern region of the valley split from the Western region of the valley approximately 550,000 years ago. These results mirror what has been found for a number of other taxa that show very high levels of divergence between the regions of the Cuatro Ciénegas valley found on either side of the Sierra de San Marcos (Rodríguez-Martínez 2004, Johnson 2005, Carson and Dowling 2006, Chaves-Campos *et al.* 2010). Additionally, although the eastern lobe of the valley is currently receiving a substantial amount of conservation attention and protection of habitats, the western lobe of the valley is not (Hendrickson *et al.* 2008a). If it is a management priority to preserve the
unique fauna of Cuatro Ciénegas, the genetically distinct lineages of organisms and the habitats on the western lobe of the valley should receive greater conservation consideration (Du Toit 2010).

Despite the genetic isolation between the three major population clusters of the valley, we did infer that there are low levels of gene flow among some locations. These low levels of gene flow could be signatures of more recent aquatic corridors that existed during wet cycles of the Holocene around approximately 11,000 years ago (Castiglia and Fawcett 2006). Another possible explanation are rare flood events, such as hurricanes, which can flood much of the valley floor and could facilitate movement among otherwise disjunct locations (Chaves-Campos et al. 2010). It is also possible that the canal systems built within the last hundred years that connects the pools and streams near the Rio Mesquites and Pozas Azules could be allowing gene flow between long isolated regions (Hendrickson et al. 2008a, Chaves-Campos et al. 2010). This canal-mediated mixing is supported by the fact that a small number of haplotypes from the Eastern valley region were recovered in the Pozas Azules region and vice versa despite otherwise substantial divergence between these two populations (Fig 2). Other studies have recovered similar patterns (Chaves-Campos et al. 2010) suggesting that these canal systems could be facilitating genetic mixing of evolutionarily distinct populations of aquatic organisms within the valley (Crispo et al. 2011).

CONCLUSIONS

Most of Cuatro Ciénegas is currently managed as a single conservation unit. Our results, combined with other phylogeographic studies within the valley, indicate that Cuatro Ciénegas is made up of several historically independent regions that are inhabited by distinctive genetic
lineages. Management efforts should begin to account for how distinctive the faunas of the
different lobes of the valley are. Populations of *L. megalotis* within the valley are also quite
genetically distinct from populations found outside the valley, and this should reinforce the
general recognition that the valley contains a highly unique vertebrate fauna (Hulsey *et al.* 2006,
Howeth *et al.* 2008, Hulsey *et al.* 2008, McGaugh 2012). We also found evidence consistent with
human-mediated habitat changes in the form of canals putting evolutionarily unique populations
of *Lepomis megalotis* at risk (Hawkins *et al.* 2000, Wesner and Belk 2012). The continued
increases in water use in and around Cuatro Ciénegas could result in the irrevocable loss of one
of North America’s most distinctive faunas whose genetic differentiation we are only now
coming to fully appreciate.

ACKNOWLEDGEMENTS

We thank Dean Hendrickson for help in Cuatro Ciénegas, Phillip Hollingsworth for field
assistance, DESUVALLE A.C. (especially Leonardo Vásquez, Eduardo Cordero and Alma
Zertuche) and PRONATURA A. C. (Isabel Morán and Perla Vásquez) for logistic support. We
also thank Martin Husemann for assistance with tissue collection. The University of Tennessee
provided support to C. D. Hulsey, Knoxville and S. G. Johnson, L. M. Coghill and Johel Chaves-
Campos were supported by The University of New Orleans. CIBNOR provided support to F. J.
García de León. We thank the Mexican government for providing us with permits (Permiso de Pesca de Fomento N°DAPA/2/130409/0961 and DAN-01202).
CHAPTER 2: Comparative phylogeography of aquatic species endemic to Cuatro Ciénegas, Mexico

ABSTRACT

Understanding contemporary patterns of genetic differentiation across co-distributed taxa requires knowing whether a single or multiple vicariant events have driven these patterns. We examined the effect of the Sierra de San Macros Mountain on the patterns and timing of genetic diversification of six native aquatic taxa occurring in Cuatro Ciénegas, Mexico. Using approximate Bayesian computation (ABC), we tested whether the mountain forms a barrier to these aquatic taxa, and estimated how long populations on the eastern and western regions of the valley have been isolated. We recovered signals of two divergence events in the valley suggesting that the three snail taxa diverged first followed by a more recent divergence event in the fish taxa. Both events occur in the Pleistocene during periods of increased aridity suggesting that climatic effects might have played a role in driving these vicariant events.
INTRODUCTION

Phylogeography attempts to characterize genealogical lineages and patterns within a spatial framework (Avise et al. 1998, Arbogast and Kenagy 2001b, Hickerson et al. 2010b). For more than twenty years, researchers have recognized that multiple co-distributed taxa can have similar patterns of genetic variation that reflect divergence due to geographical features and climatic events (Cracraft 1994, Arbogast and Kenagy 2001b, Hickerson and Carnaval 2011). The field of comparative phylogeography explores whether co-distributed taxa have experienced simultaneous divergence events (Bermingham and Moritz 1998, Craig et al. 2010, Beatty and Provan 2011). Traditional attempts at examining questions of concordant patterns of genetic variation and geographic features have relied primarily upon descriptive techniques. These studies used tools common to single-species phylogeographic studies such as the distribution of a taxon and branching time estimates that employed fixed molecular clocks recovered from fossil calibrations (Avise 1989, Rosenberg and Nordborg 2002, Papadopoulou et al. 2010). However, these methods often suffer from large variation in temporal estimates that can make it difficult to discriminate between single and multiple vicariant events. This overlap in time estimates can also lead to inaccurate interpretations such as multiple events being considered a single large event, or alternatively, a single event being inaccurately interpreted as multiple, closely occurring pulses of divergence (Riddle and Hafner 2006b, Hickerson and Carnaval 2011).

Statistical phylogeography introduced the concept of using model-based coalescent approaches to estimate parameters of interest and to test specific hypotheses (Knowles and Maddison 2002a). Unfortunately, the computational limitations of these models became apparent for models that contain a large number of parameters (Hey and Nielsen 2007b, Nielsen and Beaumont 2009a). One possible alternative to this computational limitation is the use of
approximate Bayesian computation (ABC), which simulates the data using parameter values randomly drawn from the prior distributions in order to approximate the posterior (Beaumont et al. 2002). Recent techniques using coalescent theory to build hierarchical Bayesian models for hypothesis testing have offered solutions to some of these problems. These techniques allow researchers to account for stochastic processes and biologically meaningful variation such as population size and incorporate mutation rate variation across loci (Hickerson et al. 2006a, Barber and Klicka 2010b). Although the ABC-based comparative methods are still relatively novel, they have been successfully used to test for simultaneous divergence times across co-distributed taxa (Voje et al. 2009a, Barber and Klicka 2010b, Huang et al. 2011). They have also shown promise in detecting vicariance or colonization in wide ranging species (Hickerson et al. 2006a, Hickerson and Meyer 2008b). Another strength of these methods is that they are sensitive enough to analyze data sets with small sample sizes offering a solution to researchers for whom collection of some taxa is logistically difficult (Huang et al. 2011).

The Cuatro Ciénegas valley exhibits the highest level of endemism in North America, but the phylogeographic history of many species and populations within the valley remains unclear (Hendrickson et al. 2008b). This relatively small (~1500km²) inter-montane valley located in the Chihuahuan desert contains numerous aquatic habitats and is home to more than 70 endemic species (Souza et al. 2008b). The valley is located in the center of an extremely arid region and virtually all of the endemic species are found within its more than 200 permanent pools, rivers, and lakes. These water bodies are also isolated into several hydrologically distinct drainages that are often physically separated from each other (Fig. 1).
Figure 1. Map showing the valley of Cuatro Ciénegas, and the split between what are considered to be eastern and western populations from the valley.
The eastern side of the valley is made up of several pools (Tio Candido, Tierra Blanca and Mojarral) that are all within the drainage of the Rio Mesquites river system. In the far southeast corner, the Pozas Azules region is isolated both by geographic distance and by elevation from the other drainages. The western side of the valley is dominated by two large pool systems (Juan Santos and Churince), which account for much of the aquatic habitat. There is evidence that, on more than one occasion, the hydrology of the valley has been altered allowing potential aquatic connections between these currently isolated regions. Previous studies have found that two endemic snail species, \( Mexipyrgus churinceanus \) and \( Nymphophilus minckleyi \), the two endemic pupfish (genus \( Cyprinodon \)) and native populations of softshell turtles (\( Apalone atra \)) exhibit genetic divergence between the eastern and western lobes of the valley (Rodríguez-Martínez 2004, Johnson 2005, Carson and Dowling 2006, Chaves-Campos et al. 2011b, McGaugh 2012). However, additional studies were unable to recover this same pattern of divergence in other species like the endemic box turtle (\( Terrapene coahuila \)), another endemic snail (\( Mexithauma quadripaludium \)) and shrimp species (\( Palaemonetes suttkusi \)). For example, in the case of \( P. suttkusi \) and \( M. quadripaludium \), populations within Pozas Azules were well differentiated from the rest of the valley without much support for divergence into the eastern and western phylo-groups found within the other taxa (Howeth et al. 2008, Chaves-Campos et al. 2010, Chaves-Campos et al. 2011b). While these studies have highlighted the observed genetic divergence and its potential relationship to putative drainage boundaries within the valley, as of yet, no study has quantitatively evaluated whether similar events might have resulted in shared phylogeographic patterns.
Most comparative phylogeographic studies have traditionally assumed that disparate coalescent times reveal divergences between taxa or lineages (Avise 1989, Arbogast et al. 2002b). Coalescent approaches can often reveal large variation in gene trees regardless of whether multiple taxa have experienced similar population divergence events driven by the same vicariant event (Hudson 1991). In the case of Cuatro Ciénegas, several studies have shown divergence between the east and west sides of the valley that seems to be centered to some degree at the tip of the Sierra del San Marcos (Rodríguez-Martínez 2004, Johnson 2005, Carson and Dowling 2006, Chaves-Campos et al. 2011b, McGaugh 2012). However, these studies have focused on phylogeographic patterns within single species instead of analyzing divergence simultaneously, and the few coalescent-based approaches employed have returned variable estimates of divergence ages. Due to the overlap in variation in depths of tMRCA, these estimates alone will not allow the determination of the number of divergence events that likely have occurred within the valley. However, relatively recent statistical tools provide frameworks that allow the joint estimation of the number of divergent events structuring the phylogeography of co-distributed taxa (Hickerson et al. 2007, Barber and Klicka 2010b). Here we employ msBayes (Hickerson et al. 2007, Huang et al. 2011), a hierarchical approximate Bayesian computation (ABC) pipeline, to test for simultaneous intraspecific diversification in co-distributed aquatic species in the desert valley of Cuatro Ciénegas. To address these ideas we asked several questions. First, we tested whether all six taxa examined experienced a single simultaneous or multiple divergent events. Second, we used a combination of approximate Bayesian computation and a tMRCA analysis to assign various taxa to the appropriate divergent event and to date those events. Finally, we performed a gene-flow analysis to determine whether
divergence estimates recovered were the effect of climate (i.e.: arid regions prohibiting aquatic species movement) or dispersal limited (the species have behaviorally limited ranges).

MATERIALS AND METHODS

Taxa

Six taxa with distributions throughout the Cuatro Ciénegas valley were examined. Three species of fish, covering 3 families were used: *Lepomis megalotis* (Centrarchidae), *Hericthys minckleyi* (Cichlidae) and *Astyanax mexicanus* (Characidae). Three invertebrates that are endemic to the valley were also used: *Nymphophilus minckleyi* (Hydrobiidae), *Mexipyrgus churinceanus* (Hydrobiidae) and *Mexithauma quadripaludium* (Hydrobiidae). All of these species are endemic to the valley except for *Astyanax mexicanus* (Hendrickson et al. 2008b).

Molecular data

New mitochondrial data sets were generated for the species *Astyanax mexicanus* and *Lepomis megalotis* from several sites in the Cuatro Ciénegas basin. Both sides of the Sierra de San Marcos were sampled covering the two major drainages consisting of spring-fed pools and stream outflows located on the eastern and western side of the mountains. Samples were preserved in ethanol and DNA was extracted in the laboratory from muscle or fin tissue using Qiagen Blood and Tissue Kit (Qiagen). Sequence data for 652bp from the mitochondrial cytochrome c oxidase subunit I (COI) gene were obtained using standard molecular laboratory protocols for mitochondrial data. For the new data the forward primer (5’-
TCAACCAACCACAAAGACATTGGCAC-3’) and for reverse (5’-
TCGACTAATCATAAAGATATCGGCAC-3’) from (Ward et al. 2005) were used. All samples
were deposited in GenBank after the study was complete (LCXXXXXXX – LCXXXXXXXX).

Data from existing data sets were gathered from GenBank or the authors of the
corresponding publications. *Mexipyrgus churinceanus* data was used from Johnson (2005)
(AY851303–AF851361) and comprised 697bp of the cytochrome b gene. For all other samples
data was gathered from Chaves-Campos et al. (2011b) (GU321684 - GU321898) and comprised
652bp of cytochrome b for *Hericthys minckleyi*, 620bp of COI for *Nymphophilus minckleyi*, and
634bpm of COI for *Mexithauma quadripaludium*.

**Time to most recent common ancestor estimations (tMRCA)**

To assess temporal divergences in the gene tree of each species, a Bayesian MCMC
approach was used to estimate the time-to-most-recent-common-ancestor (tMRCA) using
BEAST version 1.7.4 (Drummond et al. 2012). The best fitting, least parameter rich model of
sequence evolution based on hierarchical likelihood-ratio tests performed in jMODELTEST 2.0
was chosen for each mitochondrial dataset (Darriba et al. 2012b). Due to the uncertainty of
mtDNA substitution rates, we used an uncorrelated exponential molecular clock (Drummond et
al. 2006a). For each analysis, there were 10,000,000 steps in the MCMC chain using a UPGMA
as the starting tree and a burn-in of 1,000,000. Posterior probabilities and tMRCA were
estimated for all nodes in the resulting trees using TreeAnnotator version 1.7.4 (Drummond et al.
2012).
Simultaneous diversification

We used the ABC msBayes (Hickerson et al. 2007) pipeline (Hickerson et al. 2006a) to examine two primary hypotheses: (1) that all organisms in the valley experienced the same simultaneous divergence event and (2) the Sierra de San Macros has historically always been an impassable barrier to the taxa of Cuatro Ciénegas. In msBayes, one can allow specification of independent demographic parameters for each taxon. Additionally, this pipeline is thought to perform well with sample sizes as small as one individual sampled per population (Huang et al. 2011). We used the default lower and upper limit of the uniform prior for the population parameter, $\theta$, which is determined from the maximum average pairwise distance, $\pi$, observed within subpopulations. The ratio of Watterson’s (Watterson 1975) $\theta$ to $\pi$ ($\theta_w$) was used to set the prior for maximum ancestral population size ($\theta_{anc-max}$). The prior for the number of possible divergence events was set to the maximum number of possible events, equal to the number of codistributed taxa analyzed ($\Upsilon = 6$). Posterior distributions and their 95 per cent quartiles for the number of divergent events ($\Upsilon$) were estimated using 1000 draws of 500,000 simulated replicates.

Since an initial test suggested more than one event (i.e. mode of $\Upsilon > 1$ and $\Omega >> 0$), a second analysis was conducted to determine the number of taxa that diverged during each event and the timing of those events. In this analysis $\Upsilon$ was constrained to the number of divergence events indicated in the unconstrained analysis, in this case ($\Upsilon = 2$). The posterior mode of divergence time across taxa ($E(t)$) was converted to years using methods suggested by (Arbogast et al. 2006) based on the formula $t(0.5\theta_{max}/\mu)$ and estimated mutation rates for each taxa recovered from the literature (Table 1). A mtDNA mutation rate of 2.0% per million years was
used for all fish taxa in accordance with several other studies (Shulman and Bermingham 1995, Bowen et al. 2001, Costedoat and Gilles 2009). Generation times for each of the fish taxa were as follows: *Astyanax mexicanus* 0.5 years (Jeffery 2009), *Lepomis megalotis* 2.0 years (Kawamura et al. 2006) *Herichthys minckleyi* 3.0 years (Buchanan 1971). For each of the three snail taxa a mutation rate of 1.76% per million years was used for both *Mexithauma quadripaludium* and *Nymphophilus minckleyi* based on general hydrobiid snail COI calibrations (Wilke 2003) and 1.83% per million years Cytochrome b for *Mexipyrgus churinceanus* (Hershler et al. 2007) respectively. For each of the snail species the generation time was estimated at four months based on estimates available for other hydrobiid snails (Broekhuizen et al. 2001, Negovetic and Jokela 2001). The pipeline msBayes does not currently have the functionality to assign taxa to divergence events with any confidence values. In order to address this need, we followed methods presented by other researchers using the relative depths of a tMRCA analysis to place the taxa in the most likely divergence event (Barber and Klicka 2010b).

<table>
<thead>
<tr>
<th>Species</th>
<th>Gene</th>
<th>Length</th>
<th>Mutation Rate</th>
<th>N</th>
<th>Generation Time</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Astyanax mexicanus</em></td>
<td>COI</td>
<td>652</td>
<td>2.00</td>
<td>82</td>
<td>0.5</td>
</tr>
<tr>
<td><em>Lepomis megalotis</em></td>
<td>COI</td>
<td>652</td>
<td>2.00</td>
<td>77</td>
<td>2.0</td>
</tr>
<tr>
<td><em>Herichthys minckleyi</em></td>
<td>Cytb</td>
<td>652</td>
<td>2.00</td>
<td>47</td>
<td>3.0</td>
</tr>
<tr>
<td><em>Nymphophilus minckleyi</em></td>
<td>COI</td>
<td>620</td>
<td>1.76</td>
<td>56</td>
<td>0.33</td>
</tr>
<tr>
<td><em>Mexipyrgus churinceanus</em></td>
<td>Cytb</td>
<td>697</td>
<td>1.83</td>
<td>81</td>
<td>0.33</td>
</tr>
<tr>
<td><em>Mexithauma quadripaludium</em></td>
<td>COI</td>
<td>634</td>
<td>1.76</td>
<td>58</td>
<td>0.33</td>
</tr>
</tbody>
</table>
Gene-flow

To determine whether divergence estimates recovered were effects of climate (i.e.: arid regions prohibiting aquatic species movement) or dispersal limited (the species have behaviorally limited ranges), we estimated gene flow for the three fish taxa with potentially much higher rates of dispersal. The rooted and dated BEAST gene tree for each of the three fish taxa was used in IMa2 to estimate gene flow between the divergent populations using the isolation with migration model (Hey 2010). The model of substitution implemented in the BEAST analysis for each taxon was combined with an MCMC simulation of genealogies to model the effective number of female gene copies that would be predicted to migrate between populations. A series of likelihood ratio tests were used to test for a significant difference from zero. Initial priors were used relative to the population mutation parameters, 4N_eµ. The ancestral genetic diversity θ (theta) was calculated using the Watterson method (Watterson 1975) in ARLEQUIN 3.5 (Excoffier and Lischer 2010). Estimated ancestral N_e was then calculated with the equation θ = 2N_eµ, using the appropriate mutation rate and generation time for each taxa (Table 1). We used 20 heated chains with a burn-in of 120,000 trees out of our total 12,000,000 trees.

RESULTS

Time to most recent common ancestor estimations (tMRCA)

The effective sample size (the number of effectively independent draws from the posterior distribution to which the Markov chain is equivalent) for each parameter exceeded 200 in nearly all cases so we assume the trees were well sampled. Mean values in years-before-present (YBP) in the tMRCA fell into two discrete groups with significant overlap. The tMRCA
detected an initial grouping of dates for the snails. The tMRCA for *Mexipyrgus churinceanus* was 2,400,000 (84,000 – 4,790,000), *Nymphophilus minckleyi* at 2,170,000 (72,000 – 4,420,000) and *Mexithauma quadripaludium* at 2,140,000 (78,000 – 4,420,000) (Fig. 2).
Figure 2. Gene genealogies for the snail taxa *Mexipyrgus churinceanus* cytochrome b haplotypes, and for *Mexithauma quadripaludium* and *Nymphophilus minckleyi* COI haplotypes. Asterisks represent Bayesian a posteriori support probabilities that exceeded 0.95. The scale bar at the bottom left is proportional to branch length, measured as the number of DNA substitutions per site.
The Beast analysis recovered a well-supported split between the east and west samples populations of the valley for *M. quadripaludium* and *M. churinceanus*. In the case of *N. minckleyi*, the individuals from the western valley form a single well-supported clade, but the individuals from the eastern valley were split into two very divergent clades.

A second group contained all of the fish taxa with tMRCA for *Astyanax mexicanus* at 490,000 (11,000 – 960,000), *Herichthys minckleyi* at 510,000 (12,000 – 1,050,000) years and *Lepomis megalotis* at 550,000 (23,000 – 1,300,000) (Fig. 3). In the case of the various fish taxa, *H. minckleyi* and *A. mexicanus* showed a lack of structure throughout the valley. However, in *L. megalotis* the individuals from the western side of the valley form a clade with strong support. All of the coalescent time ranges for the taxa examined had some degree of overlap in their confidence intervals (Fig. 4).
Figure 3. Gene genealogies for the fish taxa, *Hericthys minckleyi* cytochrome b haplotypes, and for *Lepomis megalotis* and *Astyanax mexicanus* COI haplotypes. Asterisks represent Bayesian a posteriori support probabilities that exceeded 0.95. The scale bar at the bottom left is proportional to branch length, measured as the number of DNA substitutions per site.
Simultaneous diversification

The ratio of means of $\theta_w$ and $\pi$ was close to one ($12.8/13.4 = 0.95$), thus we interpreted the maximum ancestral population was close to $\theta_{\text{max}}$. The first ABC analysis detected evidence of non-simultaneous divergence ($\gamma=2.1$ and $\Omega = 0.24$) with strong support (Bayes Factor = 36.11) among the 6 taxa (Fig. 5, Fig. 6a, Fig. 6b).

**Figure 4.** The tMRCA for the 6 taxa included in the study. Squares are the mean values and lines are the 95% highest posterior densities (HPD). Scale bar is in millions of years before present (YBP).
Figure 5. Results of the msBayes ABC unconstrained analysis. Light grey bars represent the prior distribution of the number of possible events and were set to the number of taxa ($\bar{v} = 6$). Dark bars are the posterior distribution of the divergence events. The mode was ($\bar{v} = 2$).
Figure 6. Results of the msBayes ABC unconstrained analysis. The authors of msBayes recommend using both $\gamma$ and $\Omega (=\text{Var}(\tau)/\text{E}(\tau))$ to evaluate the relative strengths of each hypothesis. A low value (approx. 0) of $\Omega$ suggests the data fit a simultaneous model. (A) pure omega analysis ($\Omega = 0.24$) suggesting the data do not support a simultaneous divergence for all taxa. (B) joint density of omega and $\text{E}(t)$ with the probability of getting the respective omega and $\text{E}(t)$ values given the data.
A further analysis with the number of possible divergence events constrained to two (¥ = 2) suggested three taxa diverged in a more recent event, while the three remaining taxa diverged in an older, second event (Fig. 7). The tMRCA analysis suggests that the three species of snail (*Nymphophilus minckleyi*, *Mexipyrgus churinceanus* and *Mexithauma quadripaludium*) split during the older event, while the three species of fish (*Lepomis megalotis*, *Herichthys minckleyi* and *Astyanax mexicanus*) diverged during a more recent event. The mode divergence times ($E(t)$) when converted to YBP were 550,000 (12,000 – 980,000) for the more recent event 1 and 2,400,000 (1,500,000 – 4,100,000) for the older event 2 (Fig. 7).

![Figure 7](image)

**Figure 7.** Posterior distribution of timing ($E(t)$) events of two diversification events. The mode divergence times for the two events were 550,000 (12,000 – 980,000) for event 1 and 2,400,000 (1,500,000 – 4,100,000) for event 2.
Gene-flow

IMa2 detected significant overall rates of migration for all three fish taxa in both directions, with the exception of *Herichthys minckleyi* moving from the eastern side of the valley to the western side of the valley (Table 2). The highest rates were detected for *Astyanax mexicanus* (4.90 immigrant females per generation, *P* < 0.01) while the lowest rates were detected for *Herichthys minckleyi* (0.50 immigrant females per generation, *P* = 0.73). Sample size parameter estimates exceeded 200 in all three runs. IMa2 results suggest that, for all three fish taxa, the east and west lobes of the valley last experienced gene flow approximately 320,000 YBP (highest posterior probability support) and 700,000 years (mean posterior probability value).

**Table 2** Estimates of gene flow between populations of *Lepomis megalotis, Astyanax mexicanus* and *Herichthys minckleyi* among east and west haplogroups within the Cuatro Ciénegas valley. The estimates and 95% confidence intervals are shown for the number of immigrant females per generation as determined from IMa2 listed to and from each side of the valley. Populations with significant migration rates (*P* < 0.05) based on a likelihood ratio test with a significant deviation from zero are shown are indicated with an *asterisk* (*).
DISCUSSION

The test of simultaneous divergence detected two pulses of divergence among the six taxa. The tMRCA analysis found similar results with more overlap than the ABC method implemented, but msBayes tests for simultaneous divergence while allowing the variation in effective population size ($N_e$) and growth rate among taxa. Effective population size has a large impact on coalescent times; this makes it critical that the estimation uncertainty of this parameter is accounted for (Arbogast et al. 2002b). However, both methods consistently suggest the timing of these two pulses occurred around 550,000 and 2,400,000 years ago, respectively, thus placing these events around the middle to last Pleistocene.

Previous work indicated that climate can have a strong effect on species distributions and this effect might have been particularly strong during the Pleistocene (Smith et al. 1984, Mack et al. 1994, Hewitt 2000). This cycle of high levels of precipitation and temperature likely offered multiple opportunities for the taxa within Cuatro Ciénegas to move around followed by arid periods in which the taxa would diverge. Evidence suggests that as time progressed further into the Pleistocene, many regions of North America might have experienced more severe periods of aridity (Metcalfe et al. 2000). Our findings indicate that snail species experienced a much older divergence than the fish taxa. The three species of snails in this study are all obligate aquatic organisms with very limited dispersal capacity indicated by strong signatures of genetic structure (Johnson 2005, Chaves-Campos et al. 2011b). The fish species examined for this study have a much larger capacity for dispersal despite inconsistent patterns of genetic structure (Witt and Marzolf 1954, Swanson et al. 2008, Bradic et al. 2012b). Two of the species examined in this study, *H. minckleyi* and *A. mexicanus* exhibited a clear lack of structure in the tMRCA analysis whereas *L. megalotis* showed a strong signature of structure. The ABC analysis indicates that an
initial divergence event occurred 2,400,000 YBP. This estimate is could reflect structure that resulted from a period of increased aridity in North American deserts from 3,200,000 to 2,560,000 YBP Smith et al. (1984) Mack et al. (1994). It is then possible that another brief period of high precipitation about 550,000 YBP may have allowed the secondary contact of highly dispersing fish populations. However, the limited dispersal of the snail species might have prohibited further migration among valley populations because the wet period didn’t last long enough. This increase in precipitation could have created aquatic corridors, which allowed further migration of the fish taxa examined here among drainages within the valley. This subsequent movement and gene flow might have masked the signature of the initial divergence event. After this period, during the mid to late Pleistocene the area entered an extended period of increased aridity (Metcalfe et al. 2000). This period roughly coincides with the timing of the second divergence event recovered (Fig. 4, Fig. 7).

Climate could drive changes in habitat or habitat accessibility for many species (Zink 2002, Chaves-Campos et al. 2010). However, since the first divergence event only seems to have limited the dispersal of the snails, and was not detected in the fish taxa with higher rates of dispersal (Witt and Marzolf 1954, Swanson et al. 2008, Bradic et al. 2012b), the split in snail could be the signature of limited dispersal or an ecological limitation. The Isolation with Migration analysis we conducted to provide time estimates for gene flow for the fish species helped us clarify between these possibilities. The results suggest that all species of fish experienced some level of historic gene flow, with 320,000 YBP (highest posterior probability support) and 700,000 YBP (mean posterior probability value), which are in agreement with the ABC method estimates of divergence for the fish taxa. This suggests the second divergence
event detected by msBayes is an actual split for the fish taxa, as gene flow seems to have been interrupted for these species at approximately the same time.

CONCLUSIONS

Our ABC analyses suggest that two major events have split taxa between the east and west side of the Sierra de San Marcos in Cuatro Ciénegas. Both of these events appear to have occurred during the early to mid-Pleistocene, suggesting climate variation might have played a strong role in the splits. This study provides a comparative statistical phylogeography study of diversification in the valley of Cuatro Ciénegas and as more loci and more taxa are examined, the study should provide testable hypothesis for future work in this unique ecosystem. There appears to be strong isolating factors among populations of the various taxa that inhabit Cuatro Ciénegas. This makes the fact that Cuatro Ciénegas has one of the highest levels of endemicity in North America less surprising as it suggests that there have been repeated episodes of climatically driven isolation with the Cuatro Ciénegas valley capable of generating distinct lineages. These patterns further emphasize the uniqueness of this ecosystem and its fragile nature, as many of the taxa in the valley seem to be quite sensitive to environmental change. Recent canal systems built are beginning to connect once isolated parts of the valley (Hendrickson et al. 2008b). If this continues, these distinctive faunas whose genetic differentiation we are only now beginning to recognize will be extremely threatened.
ACKNOWLEDGEMENTS

We thank Dean Hendrickson for assistance in Cuatro Ciénegas. Phillip Hollingsworth for field assistance, DESUVALLE A.C. (especially Leonardo Vásquez, Eduardo Cordero and Alma Zertuche) and PRONATURA A. C. (Isabel Morán and Perla Vásquez) for logistic support. We also thank Martin Husemann for assistance with tissue collection. The University of Tennessee provided support to C. D. Hulsey, Knoxville and S. G. Johnson, L. M. Coghill and Johel Chaves-Campos were supported by The University of New Orleans. CIBNOR provided support to F. J. García de León. We thank the Mexican government for providing us with permits (Permiso de Pesca de Fomento N°DAPA/2/130409/0961, 230401-613-03 and DAN-01202 and DAN 02939).
Chapter 3: Next Generation Phylogeography of Cave and Surface *Astyanax mexicanus*

**ABSTRACT**

The regression of phenotypic characters is a common evolutionary pattern in cave organisms. *Astyanax mexicanus* (the blind Mexican cavefish) has become a model organism for studying the genetic basis for developmental and phenotypic regression. Here we present a study using both mitochondrial DNA and Illumina-based next generation sequence data to reconstruct a high-resolution species phylogeography for *Astyanax*. We provide added resolution to the origin of cave populations from the Sabinos and Rio Subterráneo caves. Our results suggest that the Sabinos cave population is part of a unique cave lineage unlike any other extant populations and that the Rio Subterráneo cave population is actually part of the older invasion into caves. Additionally, we find support that cave populations have at least four independent origins.
INTRODUCTION

The evolution of novel phenotypes can often be best understood when placed within a geographic and evolutionary context (Avise et al. 1998, Arbogast and Kenagy 2001b, Hickerson et al. 2010b). Intraspecific phylogeography can decipher spatial and temporal patterns of population structure (i.e., genetic differences within and among populations) and clarify the evolutionary processes driving novel phenotypes (Avise 1989, Leavitt et al. 2007, Hickerson et al. 2010b). Markers such as mitochondrial, chloroplast DNA, microsatellites, or combinations of these markers have been instrumental in providing initial inferences about the spatial and evolutionary origin of novel phenotypes (Dowling et al. 2002, Leavitt et al. 2007, Chaves-Campos et al. 2010, Gaither et al. 2011, Hou et al. 2012). Now, next-generation sequencing (NGS) tools and genome-wide SNP detection pipelines are being increasingly used to identify patterns of variation in model organisms and their close relatives that exhibit these novel phenotypes (Baird et al. 2008, Hohenlohe et al. 2010b, Etter et al. 2011). Using a large SNP dataset generated via a reduced representation library, we examined several phylogeographic hypotheses proposed for divergence in surface and cave forms of the Mexican tetra Astyanax mexicanus.

Astyanax mexicanus is a small characid fish that is used extensively as a model in evolutionary research due to its distinct cave and surface morphotypes (Rasquin 1951, Voneida and Sligar 1976, Teyke 1990, Gross et al. 2009a, Pottin et al. 2011, Gallo and Jeffery 2012, Gross et al. 2013, Gross and Wilkens 2013). There is a consensus among researchers that there have been multiple colonization events of the caves from various surface populations (Gross 2012). While the proposed structure of the various ancestral populations has varied slightly based on the markers used, the literature has generally settled on contrasting populations as either
“phylogenetically young” or “phylogenetically old” based on their degree of troglomorphy (Avise and Selander 1972, Gross 2012). The degree of troglomorphy has been used as an indicator of the age of cave populations. Populations that have reduced but not absent pigmentation and visual systems have classically been labeled as “young” populations, whereas those populations with significantly reduced or absent pigmentation or visual systems are classified as “old” populations (Strecker et al. 2004b). This concept is consistent with the idea that a longer time of isolation will lead to a larger accumulation of neutral mutations (Protas et al. 2007).

Several phylogeographic studies of mitochondrial genes and a limited number of nuclear loci in Astyanax have broken the cave populations of Astyanax up into three distinct regions known as the Guatemala and Micos populations colonized by the new epigean, or surface, stock, and the El Abra populations colonized by the old epigean stock (Fig. 1) (Dowling et al. 2002, Strecker et al. 2004b, Ornelas-García et al. 2008, Hohenlohe et al. 2010b, Bradic et al. 2012a, Yoshizawa et al. 2012, Gross and Wilkens 2013). Mitochondrial studies by Dowling et al. (2002) and Strecker et al. (2004b) found support for the multiple origin hypothesis inferring at least two independent origins for cave populations, which they termed “Phylogenetically Old” and “Phylogenetically Young”. Later work with more markers found similar results suggesting at least two independent origins of cave populations that invaded cave habitats approximately eight million and two million years ago respectively (Ornelas-García et al. 2008, Hausdorf et al. 2011, Bradic et al. 2012a). Initial studies thought that the Trans-Mexican Volcanic Belt (TMVB) might have driven this divergence in invasion patterns (Ornelas-García et al. 2008). Recent studies on comparable taxa have shown the TMVB serves as an effective barrier to fish taxa and drives divergence patterns of taxa found both north and south of the TMVB (Hulsey et al. 2010). Most
of the studies agree that there were multiple independent invasions into cave habitats, but there is still extensive incongruence among these datasets (Dowling et al. 2002, Strecker et al. 2004b, Ornelas-García et al. 2008, Hohenlohe et al. 2010b, Bradic et al. 2012a, Yoshizawa et al. 2012, Gross and Wilkens 2013). A robust phylogeography of the cave and surface forms of *A. mexicanus* remains elusive not only due to high levels of convergence between cave populations but also likely due to the limited power of markers used to infer population structure (Hausdorf et al. 2011, Gross 2012).
Figure 1 Map of Mexico showing the *A. mexicanus* sampling locations examined in this study and their corresponding historic population clusters. Surface populations (numbers) and Cave populations (letters) are labeled and the morphology of a typical individual is shown.
The populations of *A. mexicanus* present an excellent model to explore the use of NGS SNP datasets to infer population subdivision. First, numerous phylogenies exist for this species, and these phylogenies provide a framework to compare to the NGS phylogeny (Gross 2012). Second, while SNPs are subject to convergence, a very large genome-wide SNP dataset should overcome any issues of convergence due to the robust sampling of markers from throughout the genome (Emerson *et al.* 2010). In this study, we first generated a phylogeny using the mitochondrial gene *COI* to generate an initial estimate of how evolutionarily independent cave populations are from one another and from surface populations. Then, we generated a NGS SNP based phylogeny that was rooted using the mitochondrial *COI* gene. With this phylogeny we explored several constrained versus unconstrained phylogenies to statistically test alternative hypotheses of independent cave origins. Finally, we explored the number of transitions between cave and surface populations by mapping ancestral habitat states onto our best estimate of the cave and surface *A. mexicanus* phylogeny.

**MATERIALS AND METHODS**

**COI Sequencing**

Four to six individuals were collected from 16 populations covering the natural latitudinal range of *Astyanax mexicanus*, from southern Mexico to the Rio Grande, including representatives from each of the major previously identified cave population groups: Guatemala (Molino cave), El Abra (Pachón cave, Sabinos cave and Chica cave) and Micos (Rio Subterráneo cave) (Fig. 1). Genomic DNA was extracted from fin clips preserved in ethanol for each individual using a DNeasy Blood and Tissue Kit (Qiagen). A 655-bp segment of COI was amplified from four to
six individuals from each of 16 populations of *A. mexicanus* using primers FISHF1 TCAACCAACCACAAAGACATTGGCAC and FISHR1 TAGACTTCTGGTGCCAAAGAATCA. Sequences were then compared, and a consensus sequence was generated for each population using Geneious version 6.0.1 (Biomatters 2010). One sequence of COI for *Astyanax belizanus* was pulled from GenBank (FJ439401) to use as an outgroup (Ornelas-García et al. 2008).

**Amplified Restriction Fragment Library**

To build an Illumina library we used a protocol that amplifies a reduced representation restriction fragment library (Gompert et al. 2010, Holsinger 2010). Previously extracted DNA was digested with EcoRI (New England Biolabs). Ninety-two unique barcodes, ranging in size from 14-bp to 16-bp, and Illumina adapter sequences were then ligated to the fragments using T4 DNA ligase (New England Biolabs). Subsets of these fragments were then amplified via PCR using Iproof High Fidelity DNA polymerase (BioRad). The amplified fragments were size selected, at the 400-bp range, in agarose and purified using a QIAquick gel extraction kit (Qiagen). The resulting products were subsequently pooled into an Illumina sequencing library and the library was sequenced as one lane on an Illumina GAIIX sequencer.

**Detection of Informative SNPs**

A complete reference genome for *A. mexicanus* is not yet available to help align sequence reads. Therefore, we used a slightly modified version of the multistep approach suggested by Emerson et al. (2010) to assign a consensus sequence to each population at each locus and align
those consensus sequences across populations. All reads were scored using FASTQC to initially check for the quality of reads (Andrews et al. 2012). Sequences with an overall error rate greater than one percent in their overall quality score and any sequences that were overrepresented (i.e. repetitive reads that made up more than one percent of the overall reads) were discarded. All remaining reads were checked to make sure they didn’t exceed 20 percent variation in bases between any position and contained less than 20 percent ‘N’ content before proceeding. The remaining raw sequence reads were cleaned and processed using the Stacks pipeline version 1.01 (Catchen et al. 2013). Sequence reads were also checked to see if they contained the entire barcode and an intact cut site and, if so, those reads were retained for the process_radtags program in Stacks (Catchen et al. 2013). A second level of quality control was performed where any read that contained an uncalled base was discarded. A final level of quality control was performed using a “sliding window” approach where if sequence reads had a quality score of below 90% (raw phred score of 10 or under) it was discarded.

Identical reads were combined into stacks, and pairwise sequence divergence among stacks was used to group stacks into loci. A locus for our purposes was defined as a group of stacks, in which there is another stack that is no more than one nucleotide divergent (Emerson et al. 2010). To determine if a nucleotide is fixed within a population, we performed a likelihood ratio test, based on the read counts of alternative nucleotides, to test whether the frequency of the most common nucleotide is significantly larger than the threshold (\( \hat{\rho} = 0.5 \)) as suggested by Emerson et al. (2010). A consensus sequence was assigned to each population using the nucleotide occurring most frequently in the population. These loci were then aligned among the populations, and any locus present in at least two populations was kept for the phylogenetic analysis.
**Phylogenetic Analysis**

The two datasets (COI and SNP) were first aligned using MUSCLE (Edgar 2004a). All analyses were performed in PhyML version 3.0 (Guindon and Gascuel 2003). Using Akaike’s information criterion (AIC) and Bayesian information criterion (BIC) for the appropriate dataset, jModelTest version 2.0 was used to determine the appropriate model of nucleotide evolution (Darriba et al. 2012a). Both sets of criteria agreed that for the COI dataset an HKY +G model was appropriate. For the SNP dataset a GTR +G model was used. A Bayesian model transformation of a likelihood ratio test (aBayes) methodology providing posterior probability support was performed.

We wanted to test the various hypotheses corresponding to the possible population structure of these groups with the large NGS dataset. We compared our unconstrained phylogeny to the following constrained phylogenies: 1) constraining cave populations into two monophyletic groups (New and Old) based on delineations in previous literature and 2) constraining all cave and surface populations as two monophyletic groups in order to further test the support of multiple origins of the cave populations (Dowling et al. 2002, Strecker et al. 2004b, Ornelas-García et al. 2008, Hohenlohe et al. 2010b, Bradic et al. 2012a, Yoshizawa et al. 2012, Gross and Wilkens 2013). A well-supported tree in this case would suggest a single origin for cave populations. We used Treefinder to compare the likelihood scores of the best tree reconstructed without constraint versus those trees with monophyly constraints (Jobb et al. 2004).

To further explore the number of origins of cave populations, the ancestral states of the phenotypes were mapped onto the single best phylogeny recovered from the SNP dataset (Huelsenbeck et al. 2003, Bollback 2006, Hulsey et al. 2013). Evolution between habitats was
inferred using Mesquite version 2.75 with a maximum likelihood model for reconstruction and equal transition probabilities between the two habitat types. These parameters were then used to infer the proportional likelihood of the two ancestral habitats at each node (Lewis 2001, Maddison and Maddison 2011).

RESULTS

MtDNA

A 652-bp segment from 16 populations (92 individuals) of *A. mexicanus* was used to infer the phylogenetic relatedness of various cave and surface populations (Kolaczkowski and Thornton 2008). Both our Bayesian Markov chain Monte Carlo and maximum likelihood framework agreed on the topology of the tree recovered. The topology displayed reasonably strong support for previous hypotheses of the grouping of the El Abra population cluster (Fig. 2) (Dowling *et al.* 2002, Strecker *et al.* 2004b, Ornelas-García *et al.* 2008, Hausdorf *et al.* 2011, Bradic *et al.* 2012a, Gross 2012, Yoshizawa *et al.* 2012). The Rio Subterráneo cave populations and Molino cave populations formed clades representing the Micos and Guatemala clusters respectively that have been reported in previous work (Dowling *et al.* 2002, Strecker *et al.* 2004b, Ornelas-García *et al.* 2008, Hausdorf *et al.* 2011, Bradic *et al.* 2012a, Gross 2012, Yoshizawa *et al.* 2012). However, mitochondrial-based support for both of these clades was weak.
Figure 2 Maximum likelihood phylogenetic tree of *A. mexicanus* populations based on COI sequences with *A. belizanus* used as an outgroup. Node support from the approximate likelihood ratio test support and empirical Bayes posterior probabilities (expressed as %), respectively are shown for nodes that have at least one support value greater than 50. Population names are those used in previous publications where possible and correspond to the geographic sampling location.
SNP-Based Tree

Genomic libraries were created from pooling four to six unique barcoded individuals from each of the 16 different populations at cut sites throughout the genome. One lane of GAIIx Illumina sequencing resulted in more than 11 million sequence reads of which more than 4.5 million were retained after quality control filtering. An average of 300,080 ± 54,973 SE sequence reads were recovered for each population. From within each population 53,033 ± 2733 SE stacks were identified across 20,180 ± 1831 loci. A total of 2,728 SNPs were recovered that were fixed within at least two populations and variable among populations. All raw sequence reads are available at the National Center for Biotechnology Information Short Read Archive (accession nos. XXXXX- XXXXX).

The trees generated using the constrained analyses for the new and old groups and the surface and cave populations showed relatively poor and resolution (Fig. 3). The unconstrained analysis using the NGS SNP dataset resolved 3 major clades in A. mexicanus with strong support for nearly every node (Fig. 4).
Figure 3 Maximum likelihood phylogenetic trees of *A. mexicanus* using NGS SNP data and rooted with the *COI* tree (Fig 2). (A) A test constraining "phylogenetic old" and "phylogenetic new" populations, and tree (B) is the constrained test of cave and surface populations and tree. Node support is given as the maximum likelihood approximate likelihood ratio test support, and the empirical Bayes posterior probabilities. Black and grey bars identify the groups constrained for the comparative analysis. Population names are those used in previous publications where possible and correspond to the geographic sampling location.
The hypothesis that the old and young groups are monophyletic was rejected (Approximately Unbiased test, p < 0.02). Notably, all the other topology comparison tests implemented in Treefinder (ELW, BP, KH, SH and WSH tests) lead to the same conclusion (p < 0.04). The hypothesis of the cave and surface populations forming monophyletic groups in a constrained tree was also rejected (Approximately Unbiased test, p < 0.01) with all the other topology comparison tests implemented in Treefinder (ELW, BP, KH, SH and WSH tests) reaching the same conclusion (p < 0.02).

For the best-supported tree (unconstrained hypothesis), the estimated proportional likelihood of the ancestral character state (cave or surface) was calculated and is represented on Figure 4 as a pie diagram at each node. The Molino cave population clustered together with other northern populations with strong support (Fig. 4). This isolation of the Molino cave population and northern population clade is historically supported as part of the Guatemala group. Within the other well-supported clades, we recovered signatures of three independent colonization events into caves (Fig. 4). The first of these is with the El Abra group containing Pachón cave population and Chica cave population. These populations cluster together with the surface population of Trocones with strong support. The two cave populations form a separate well-supported clade. The Sabinos cave population, which has traditionally been grouped with other El Abra populations, forms another, unique colonization event into caves. Finally, the Rio Subterráneo cave population forms a well-supported clade with the rest of the Micos group surface populations.
Figure 4 The best tree from the unconstrained SNP analysis. Each node is a pie diagram estimate of the proportional likelihood of the two ancestral habitats (Black = Cave and White = Surface). Node support is given as the maximum likelihood approximate likelihood ratio test support, and the empirical Bayes posterior probabilities. Population names are those used in previous publications where possible and correspond to the geographic sampling location.
DISCUSSION

The NGS SNP analysis confirms multiple origins of *A. mexicanus* cave populations. The additional resolution of this approach revealed four distinct origins of cave populations. We have also recovered further evidence supporting previous morphological assertions that the Sabinos Cave population is unique and not part of the El Abra population cluster. Recent work on mitochondrial and nuclear loci and the addition of more recent microsatellite studies have concluded that the cave populations have originated from at least two independent colonization events (Dowling *et al.* 2002, Strecker *et al.* 2004b, Ornelas-García *et al.* 2008, Hausdorf *et al.* 2011, Bradic *et al.* 2012a, Gross 2012, Yoshizawa *et al.* 2012). The first, and older event includes cave populations located in the El Abra region, while the second and younger event includes the populations in the Guatemala and Micos populations (Fig. 1). Additional research has suggested that each of these colonization waves possibly gave rise to multiple independent invasions into cave habitats (Hausdorf *et al.* 2011, Bradic *et al.* 2012a, Yoshizawa *et al.* 2012). As we discuss below, our findings suggest that evolution into caves has occurred several times, and we recover the signature of two previously undetected colonization events.

Geographic Phylogeny

The genome-wide SNP phylogeny recovered mostly the same pattern as previous work in relation to the 3 major cave groups of Guatemala, El Abra and Micos (Dowling *et al.* 2002, Strecker *et al.* 2004b, Ornelas-García *et al.* 2008, Hohenlohe *et al.* 2010b, Bradic *et al.* 2012a, Yoshizawa *et al.* 2012, Gross and Wilkens 2013). Our data include many additional surface populations and offer some degree of resolution to their phylogenetic structure. For example, most of the clades form well-supported groups that are geographically located near each other, but we found the Trocones population falling out closely related to the Pachón cave population.
and Chica cave population in the El Abra group. The fact that the Molino cave population is located south of Trocones, but clusters instead with Cuatro Ciénegas and San Fernando is unexpected based on geographical location. We might have expected Trocones to cluster with the other Guatemala surface populations. The surface populations on the western side of El Abra (the Micos Cave populations) clusters together with the Rio Subterráneo cave population as expected. The more southern populations contain another notable exception with El Zapotal and Tapijulapa clustering together, despite the fact that the Catemaco population geographically falls between them. As expected, the remaining far southern surface populations, Catemaco, Rio Tzendales and Teapa form a clade. The most unexpected result is that the Sabinos cave population clusters together with Rio Tzendales and all of the other southern populations, despite being located in the El Abra region (Fig. 1).

Multiple Cave Origins

Our findings agree with previous work that there have been multiple origins of *A. mexicanus* cave populations (Dowling *et al.* 2002, Strecker *et al.* 2004b, Ornelas-García *et al.* 2008, Hausdorf *et al.* 2011, Bradic *et al.* 2012a, Gross 2012, Yoshizawa *et al.* 2012). These studies suggested two major colonization events, a young event, containing the Guatemala and Micos populations, and an old event containing the El Abra populations. However, our findings suggest that the Micos populations are more closely related to the El Abra populations than the Guatemala populations (Fig. 4). This result suggests that the Micos populations are either part of the old colonization event, or form a unique, previously undiscovered colonization event. Another possibility is that since our Micos population sample came from the Rio Subterráneo cave population, the opening in this cave to the surface could possibly allow introgression
between surface and cave populations (Gross 2012). This might cause this population to fall out in a well-supported clade with neighboring surface populations as we find. While some work has reported variation in the regressive phenotypes over time in some cave populations, and records of flood events moving large numbers of surface individuals into Micos cave populations exist, few if any intermediate phenotypes have been observed (Langecker et al. 1991, Strecker et al. 2012). Hausdorf et al. (2011) reported that most of the surface individuals are purged from the population due to an inability to compete for limited resources with the better-adapted cave individuals. Additionally, the nuclear SNP dataset used for this study is not affected by mitochondrial introgression that has been suggested by previous authors (Hausdorf et al. 2011, Bradic et al. 2012a, Hailer et al. 2012, Yoshizawa et al. 2012). Our results also find that the other morphologically similar (12 ribs), older populations form a well-supported clade that is quite divergent from the clade containing the Rio Subterráneo cave population. Under these circumstances it is reasonable to infer that the Micos population forms a unique, yet undiscovered colonization event.

Additionally, we found further support for the findings of Dowling et al. (2002) and Hausdorf et al. (2011) that individuals from the Sabinos cave population are unique and possibly established by a now extinct ancient surface population. Previous work have found that fish from the Sabinos cave population exhibit a very unique morphology not found in any other A. mexicanus populations in northern Mexico suggesting they are a distinct population (Dowling et al. 2002). More specifically this population is mostly closely related to the Rio Tzendales population found on the border of southern Mexico and Guatemala. This finding supports suggestions by Dowling et al. (2002) that these fish are likely an ancient ancestral population that is more closely related to fish in Costa Rica and southern Mexico than any other extant
northern Mexico populations. Morphologically Sabinos individuals have 11 ribs, while individuals from the other populations in El Abra, Guatemala and Micos groups all have 12 ribs (Dowling et al. 2002). Other genetic studies have placed this lineage as more closely related to species from far southern Mexico and Costa Rica (Dowling et al. 2002, Ornelas-García et al. 2008). Our results indicate that the Sabinos population forms a well-supported cluster with far southern Mexico populations, specifically Rio Tzendales found near the border of Mexico and Guatemala (Fig 1). Their unique morphology, lack of gene flow with neighboring populations Strecker et al. (2012) and genetic distinctiveness from the rest of the El Abra group, suggest that this population is a relic of an old ancestral surface population that is no longer found in Mexico.

Some studies have found a higher level of genetic variability in cave populations than in sympatric surface populations (Hausdorf et al. 2011, Bradic et al. 2012a, Strecker et al. 2012). This suggests that a reasonable amount of gene flow occurs between different cave populations and between surface and cave populations (Bradic et al. 2012a). Gene flow between cave and surface forms would obscure the true patterns of descent making interpretation of population history difficult. Some recent work, and ideally additional studies in the future will recover the genes responsible for the morphological variation in different cave forms to help determine if they independently derived (Protas et al. 2005, Gross et al. 2009b). However, this analysis provides the most robust marker sampling to date and it supports four independent origins of the cave forms (Fig 4).
CONCLUSIONS

The use of the NGS SNP dataset has provided added clarity to the phylogeny of *A. mexicanus*. The resolution gained from using approximately 2,700 SNPs in the nuclear genome provided evidence for several alternatives that were inconsistently resolved in previous studies. Additionally, the power of this data has allowed us to uncover a potentially cryptic clade, with a high level of confidence that the placement in the phylogeny was not recovered by previous studies (Dowling et al. 2002, Strecker et al. 2004b, Ornelas-García et al. 2008, Strecker et al. 2012). A high-resolution phylogeny will help us to determine with more clarity the underlying patterns driving the regressive evolutionary change in this species. Finally, we hope this data set and phylogeny will provide a stronger evolutionary and genetic foundation for future studies that examine the genomic basis of trait evolution in *A. mexicanus*.
ACKNOWLEDGEMENTS

The University of Tennessee provided support to C. D. Hulsey, Knoxville and S. G. Johnson, L. M. Coghill and Johel Chaves-Campos were supported by The University of New Orleans.

CIBNOR provided support to F. J. García de León. We thank the Mexican government for providing us with permits (Permiso de Pesca de Fomento N°DAPA/2/130409/0961, 230401-613-03, DGOPA.0503.181010.-5003, DGOPA.00570.280108.-0291, and DAN-01202 and DAN 02939).
REFERENCES


desert night lizard *Xantusia vigilis* species complex (Squamata: Xantusiidae). Molecular Ecology 16:4455-4481.


VITA

The author was born in Florida. He later moved to Illinois where he spent much of his life. He obtained a Bachelor’s degree in Zoology with minors in Microbiology and Geology from Western Illinois University in 2004. He joined the lab of Professor Michael A. Romano at Western Illinois University where he conducted his Master’s research in population genetics from 2004 to 2007. He joined the University of New Orleans biology graduate program to pursue his PhD working in phylogeography and evolutionary ecology. He became a member of Professor Steven G. Johnson’s lab and research group in 2007.