Next generation phylogeography of cave and surface Astyanax mexicanus

Lyndon M. Coghill a, C. Darrin Hulsey b, Johel Chaves-Campos a, Francisco J. García de Leon c,⇑, Steven G. Johnson a

aDepartment of Biological Sciences, University of New Orleans, 2000 Lakeshore Drive, New Orleans, LA 70148, USA
bDepartment of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, 569 Dabney Hall, Knoxville, TN 37996, USA
cLaboratorio de Genética para la Conservación, Centro de Investigaciones Biológicas del Noroeste, P.O. Box 128, La Paz, B.C.S., Mexico

1. Introduction

The loss of phenotypes is best understood when placed within a clear geographic and evolutionary context (Arbogast and Kenagy, 2001; Avise et al., 1998; Hickerson et al., 2010). Intraspecific phylogeographies can clarify the evolutionary processes responsible for generating patterns of trait loss and also assess the number of independent origins of trait change (Avise, 1989; Hickerson et al., 2010; Leavitt et al., 2007). Astyanax mexicanus, a small characid fish with distinct cave and surface morphotypes native to the Nearctic ecozone, is an evolutionary model organism used extensively to study the loss of phenotypic characteristics (Gallo and Jeffery, 2012; Gross et al., 2009a, 2013; Gross and Wilkens, 2013; Pottin et al., 2011; Rasquin, 1951; Teyke, 1990; Voneida and Sligar, 1976).

While inferences concerning population structure of the various A. mexicanus populations have varied based on the markers used, researchers have discussed the divergence of populations within the context of contrasting populations as either phylogenetically “Young” or phylogenetically “Old”. This classification is based on the degree of troglomorphy (cave adaptation) exhibited and the major phylo-group to which these populations belong (Avise and Selander, 1972; Gross, 2012). Populations that have slightly reduced pigmentation and somewhat functional 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., 2004). This concept is consistent with the idea that a longer time of isolation will lead to a larger accumulation of mutations that could lead to the loss of traits having no functional advantage in cave environments (Protas et al., 2007). This variation in the degree of troglomorphy also adds support to the hypothesis that there have been multiple origins of extant cave populations, but there could easily be more than two invasions of A. mexicanus into cave habitats (Bradic et al., 2012; Dowling et al., 2002; Gross and Wilkens, 2013; Hohenlohe et al., 2010; Ornelas-Garcia et al., 2008; Strecker et al., 2004; Yoshizawa et al., 2012). Several phylogeographic studies of mitochondrial genes and a limited number of nuclear loci in A. mexicanus have broken the cave populations of A. mexicanus into three distinct regions. The Guatemala populations as well as the Micos populations are believed to be colonized by the Young epigean, or surface, and the El Abra populations are thought to have been colonized by the Old epigean stock (Fig. 1) (Bradic et al., 2012; Dowling et al., 2002; Gross and Wilkens, 2013; Hohenlohe et al., 2010; Ornelas-Garcia et al., 2008; Strecker et al., 2004; Yoshizawa et al., 2012). Mitochondrial studies by Dowling et al. (2002) and Strecker et al. (2004) found support for the multiple origin hypoth-
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 and potential for introgression of individual markers used to infer population structure (Gross, 2012; Hausdorf et al., 2011). Therefore, the resolution of the evolutionary history of A. mexicanus populations could benefit in several ways from the use of next generation sequencing datasets centered on single nucleotide polymorphisms (SNPs) to infer additional population subdivision. Because numerous phylogenetic hypotheses exist for this species, there is a prior framework to compare to NGS inferred histories (Gross, 2012; Hausdorf et al., 2011). Additionally, while introgression could occur with there is a prior framework to compare to NGS inferred histories (Gross, 2012; Hausdorf et al., 2011). 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 and potential for introgression of individual markers used to infer population structure (Gross, 2012; Hausdorf et al., 2011). Therefore, the resolution of the evolutionary history of A. mexicanus populations could benefit in several ways from the use of next generation sequencing datasets centered on single nucleotide polymorphisms (SNPs) to infer additional population subdivision. Because numerous phylogenetic hypotheses exist for this species, there is a prior framework to compare to NGS inferred histories (Gross, 2012). Additionally, while introgression could occur with any set of markers, a very large SNP dataset should provide a set of more robust inferences of the evolutionary history of A. mexicanus populations by helping to reduce signal masking from potential outlier loci (Emerson et al., 2010).

In this study, we first estimated the evolutionary structure of cave and surface A. mexicanus populations using a mitochondrial phylogeny. Then, we generated a 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 estimated the number of transitions between cave and surface populations using a mapping of ancestral habitat states onto our best estimate of the cave and surface A. mexicanus phylogeny. Previous phylogeographic studies of A. mexicanus have used a small number of markers that could be strongly subject to sampling bias and afford relatively little power to diagnose and resolve alternative hypotheses concerning the evolutionary history of this model system (Luikart et al., 2003). In contrast, the huge number of loci we score from across the genome could provide much more accurate estimates of population structure, offer a more robust understanding of how troglomorphy is related to population history, and identify previously overlooked invasions of A. mexicanus into caves.

2. Material and methods

2.1. COI sequencing

Four to six individuals were collected from sixteen populations covering almost the entire range of A. mexicanus, from southern Mexico to the Río 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 (Río 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 TCAACCAACCAAAAGACACCTGACAC and FISHR1 TAGACTTCTGGGTGGCCAAAGAATCA. 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. All original sequences were submitted to GenBank (accession Nos. KMO43792–KMO43827).

2.2. NGS 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; Hölsinger, 2010). Previously extracted DNA was digested with the enzymes MseI and 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.

2.3. 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 were discarded with an overall error rate greater than one percent in their quality score and any sequences that were overrepresented (i.e. repetitive reads that made up more than one percent of the overall reads). All remaining reads were checked to make sure they did not exceed 20% variation in bases between any position and contained less than 20% ‘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 checked and only retained if they contained the entire barcode and an intact cut site. Any read that contained an uncalled base was discarded, and sequence reads that had a quality score of below 90% (raw phred score of 10 or under) using a “sliding window” approach were discarded.

Identical reads were combined into stacks and pairwise sequence divergence among stacks was used to group loci that were no more than one nucleotide divergent from another locus (Emerson et al., 2010). This method combines the observed frequency and depth of coverage into a single test to determine if a nucleotide is fixed within a population, based on a likelihood ratio test. Using the read counts of alternative nucleotides, we tested whether the frequency of the most common nucleotide is significantly larger than the threshold allele frequency \( \hat{p} = 0.5 \) using a likelihood ratio test of the read counts of alternative nucleotides. The population consensus sequence was assigned the most-observed nucleotide if both \( p \) (nucleotide frequency) \( \geq \hat{p} \) and \( 2 \mathcal{L}_R > \text{likelihood of the observed read counts} \) (significance level \( \alpha = 0.05 \)). 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.

2.4. Phylogenetic analysis

Three different datasets were generated for the phylogenetic analysis: (1) the COI (2) the NGS SNPs including all SNPs that were fixed–within, and variant among populations and (3) a second NGS SNP dataset that included only the variable SNPs. The COI dataset was first aligned using MUSCLE (Edgar, 2004). Phylogenetic analysis on the COI dataset was conducted in PhyML version 3.0 (Guindon and Gascuel, 2003). Using Akaike’s information criterion (AIC) and Bayesian information criterion (BIC) for the datasets respectively and jModelTest version 2.0 was used to determine the appropriate model of nucleotide evolution (Darriba et al., 2012). Both sets of criteria agreed that for the COI dataset an HKY + G model was appropriate while for the SNP dataset that included both fixed and variable sites, a GTR model was used that included a gamma distribution to incorporate rate variation among the variable SNP sites. For the SNP dataset including only the variable sites, an HKY model was used.

We built the phylogenetic trees under both likelihood and parsimony models. For the dataset that contained only variable sites, a maximum parsimony analysis was conducted using PAUP with 100 replicates using reconnection branch swapping and 200 bootstrap replicates (Swofford, 2003). For the dataset containing both fixed and variable sites, maximum likelihood and topological empirical Bayesian analyses were used in PhyML. In both analyses, a Bayesian model transformation of a likelihood ratio test (aBayes) methodology providing posterior probability support was performed. In order to test the various hypotheses, corresponding to the possible population structure of these groups with the large NGS dataset, an unconstrained phylogeny was compared to the following constrained phylogenies: (1) constraining cave populations into two monophyletic groups (Young 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. A well-supported tree in this case would suggest a single origin for cave populations. In order to compare the likelihood scores of the best tree reconstructed without constraint versus those trees with monophyly constraints, Treefinder was used to perform an approximately unbiased test to assess the confidence of trees (Jobb et al., 2004; Shimodaira, 2002).

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 (Bollback, 2006; Huelsenbeck et al., 2003; 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).

3. Results

3.1. 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 recovered the same tree topology. The topology displayed reasonably strong support for previous hypotheses of the grouping of the El Abra population cluster (Fig. 2). 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. However, support for both of these clades is weak.

3.2. SNP-based tree

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

![Fig. 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 locations labeled in Fig. 1.](image-url)
300,080 ± 54,973 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 2728 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 No. SRR1425170).

The trees generated using the variable and non-variable SNP dataset for the constrained analyses (Young and Old groups, surface and cave) showed relatively poor resolution (Fig. 3). The unconstrained analysis allowed the resolution of 3 major clades in A. mexicanus with strong support for nearly every node (Figs. 4 and 5). The parsimony analysis using only the variable SNP data recovered a well-supported tree that is generally in agreement with the tree recovered from the combined dataset (Fig. 5). In general, both of these methods resulted in trees with very similar topologies. The only note-worthy difference is the placement of the Tapijulapa (Grijalva-Usumacinta Basin) clade that falls out with Rio Tzendales (Grijalva-Usumacinta Basin), Sabinos and Teapa (Grijalva-Usumacinta Basin) populations in the parsimony analysis whereas in the likelihood analysis, Tapijulapa falls out on a clade with El Zapotal (Tuxpán Basin) and Catemaco (Papaloapan basin).

The hypothesis that the Young and Old constrained tree is the correct tree was rejected (Approximately Unbiased test, $p < 0.02$). Notably, all other topology comparison tests implemented in Treefinder (ELW, BP, KH, SH and WSH tests) led to the same conclusion ($p < 0.04$ per comparison). 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 ($p < 0.02$ in all cases).

For the best-supported tree (unconstrained hypothesis), the estimated proportional likelihood of the ancestral character state (cave or surface) was calculated and is represented in Fig. 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 and Chica cave populations. These populations cluster together with the surface population of Troncones (Soto La Marina Basin) 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. It also clusters with other Southern surface populations suggesting support for a geographically sympatric clade. Finally, the Rio Subterrâneo cave population forms a well-supported clade with the rest of the Micos group surface populations (Pánico basin).

### 4. Discussion

The current study of A. mexicanus phylogeography provides three new insights into the evolutionary history of colonization of cave environments. First, we recovered a very well-supported phylodyssey providing more robust inferences of relatedness among populations. Second, we were able to rigorously test, through constrained clade analysis, long-standing hypotheses about phylegetically Old versus Young populations. Lastly, we recovered further evidence supporting previous morphological assertions that the Sabinos Cave population is unique and not part of the El Abra population cluster.

The genome-wide SNP phylogeny recovered a very well-supported phylogeny that suggests a similar evolutionary history as previous work in relation to the 3 major cave groups of Guatemala, El Abra and Micos (Bradic et al., 2012; Dowling et al., 2002; Gross and Wilkens, 2013; Hohenlohe et al., 2010; Ornelas-García et al., 2008; Streeker et al., 2004; Yoshizawa et al., 2012; Esquivel-Bobadilla, 2011). This dataset also included additional northern surface populations relative to the better studied southern cave and surface populations. Most of the clades formed well-supported groups that are geographically located near one another, but the results suggest the Trocones population (Soto La Marina Basin) is
closely related to the Pachón and Chica cave populations in the El Abra group. It would generally be expected for Trocones to cluster with the other Guatemala surface populations, where these results suggest the Molino cave population clusters instead with Cuatro Ciénegas and San Fernando. The surface populations on the western side of El Abra (the Micos Cave populations) cluster together with the Rio Subterráneo cave population. The more southern populations contain another notable exception with El Zapotal and Tapijulapa (both of Grijalva-Usumacinta basins) clustering together, despite the fact that the Catemaco population geographically falls between them. The remaining far southern surface populations, Catemaco, El Zapotal (Tuxpán and Papaloapan basins) Rio Tzendales, Tapijulapa and Teapa (Grijalva-Usumacinta basins) form a well-supported clade. The most surprising 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 (Figs. 1 and 5). This geographically anomalous clustering further suggests the Sabinos population shares a more ancient ancestral origin with the other populations of the Grijalva-Usumacinta basins.

Using hypotheses proposed in previous work, further tests were conducted to determine whether populations can be divided into phylogenetically Old and Young populations through constrained clade analysis (Avise and Selander, 1972; Gross, 2012). Constraining the cave populations into clades based on the degree of troglomorphy produced a tree that had much weaker support than the unconstrained analysis. This is very apparent in the Young clade that includes Rio Subterráneo and Molino Cave in the constrained analysis. Support for this clade is very low when constrained, and when these populations are allowed to group based on phylogenetic similarity, instead of morphological similarity, they fall out in well-supported separate clades. Traditionally, degree of troglomorphy has been used as a proxy or assumed to be correlated with time since colonization. However, our findings suggest that phenotypic adaptations to cave environments likely reach fixation in populations at an uneven rate, and therefore, the level of trait regression is not likely a robust indicator of time since colonization of cave habitats.

Earlier studies have suggested two major colonization events: a Young event, containing the Guatemala and Micos populations, and an Old event containing the populations from the El Abra region (Fig. 1). Additionally it has been suggested that each of these colonization waves possibly gave rise to multiple independent invasions into cave habitats (Bradic et al., 2012; Hausdorf et al., 2011; Yoshizawa et al., 2012). In order to further test this multiple origin hypothesis for cave populations, a test for a single ancestral origin for all cave populations was used in a constrained phylogenetic context. The findings agree with previous work that there has been multiple origins of A. mexicanus cave populations in Mexico (Bradic et al., 2012; Dowling et al., 2002; Gross, 2012; Hausdorf et al., 2011; Ornelas-García et al., 2008; Strecker et al., 2004; Yoshizawa et al., 2012). Constraining all cave populations to a single clade produced very low nodal support, suggesting that the clade is not a true representation of the phylogenetic history of these populations.

The unconstrained tree suggests that evolution into caves has occurred several times, and the signature recovered suggests two previously undetected colonization events. These findings suggest that the Micos populations are more closely related to the El Abra surface populations than to the Guatemala surface populations suggesting that the Micos populations are either part of the Old colonization event, or form a unique, previously undiscovered colonization event (Fig. 4). Another possibility is that since the 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 cluster in a well-supported clade with neighboring surface populations as reported here. 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

Fig. 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 likelyhood approximate likelihood ratio test support, and the empirical Bayes posterior probabilities.

Fig. 5. Maximum parsimony tree of A. mexicanus using only the variable sites between populations from the SNP dataset. Bootstrap percentages with values > 50 are shown behind nodes.

![Diagram](image-url)
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, making this scenario seem unlikely (Bradic et al., 2012; Hulsey et al., 2012; Hulsey et al., 2011; Yoshizawa et al., 2012). We also thank Dean Hendrickson and Gil Rosenthal for some samples. Finally, we thank Nathan Heidenreich for assistance with graphical design.

Acknowledgements

The University of Tennessee provided support to C.D. Hulsey, Knoxville and S.G. Johnson. L.M. Coghill and Jehel 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, DGOP-A.05003.181010-5003, DGOPA.00570.280108–0291, and DAN-0120 and DAN 0293).

References


