Species characterization and hybrid investigation in juvenile spiny lizards (Sceloporus spp.) by genetic sequencing

Malia Loustalot
malialou@hawaii.edu

Methods Details (supplementary information)

DNA extraction and PCR amplification for ND1 and TRAF6 genes

These steps first remove DNA from the tissue of each individual (DNA extraction), then generate thousands to millions of copies of specific parts of the DNA, ND1 gene and TRAF6 gene regions in this case (PCR amplification). I obtained tissue from a small (1 mm) length of the distal portion of the tail, and preserved the tissue in 95% ethanol. I extracted genomic DNA using an alcohol and salts. Partial mitochondrial NADH subunit 1 (ND1) and nuclear TRAF6 genes were amplified (copied) using the following PCR protocols that define the amounts of DNA and reagents used: For genes ND1 and TRAF6, a 50 µL reaction contained: 10 µL of 20 ng/ µL DNA template, 5 µL of each primer at a concentration of 10 µM, 25 µL of MyTaq^{TM} Red Mix and 5 µL water. The amplification conditions for ND1 consisted of 30 sec of denaturing at 95°C, 30 sec of primer annealing at 58°C, and 120 sec of extension at 72°C for 34 cycles. The amplification temperatures for TRAF6 consisted of 30 sec of denaturing at 95°C, 30 sec of primer annealing at 55°C, and 120 sec of extension at 72°C for 34 cycles. DNA amplification (copying) is necessary for DNA sequencing, which requires thousands of copies of the sequence to accurately know the gene sequence.

Table 1. Primers used in this study

Gene	Source	Primer name: sequence (5' - 3')	Direction	Reference
ND1	Mitochondrial	16dR: CTA CGT GAT CTG AGT TCA GAC CGG AG	Forward	Leaché, (2010)
		tMet: ACC AAC ATT TTC GGG GTA TGG GC	Reverse	
TRAF6	Nuclear	TRAF6_f1: ATG CAG AGG AAT GAR YTG GCA CG	Forward	Wiens, Kuczynski, Arif, and Reeder, (2010)
		TRAF6_r2: AGG TGG CTG TCR TAY TCY CCT TGC	Reverse	

PCR product verification and sequencing for ND1 and TRAF6 genes

This step is necessary to make sure that the specific gene regions of interest (ND1 and TRAF6) were copied successfully and can be sequenced accurately. I ran PCR products on a 2% agarose

Tris-borate-EDTA gel to verify successful amplification of target gene fragments. This involves placing the DNA in one end of a gel and passing an electric current through it, to separate the DNA fragments based on size as smaller fragments move faster through the gel. Once I verified successful amplification of all unknown samples as well as the possible hybrid sample by comparing them to a DNA fragment of known size, I combined 4 µL of PCR product (the many copies of DNA from the PCR machine) with 2 µL of ExoSap (a reagent that helps prepare the sample for sequencing) and incubated for 15 min at 37°C followed by 15 min at 80°C on a thermocycler. Upon completion, 4 µL of the PCR/ExoSap mixture, 2 µL of either the primer (small piece of DNA that helps initiate the sequencing reaction), and 4 µL of water were combined and submitted to the University of Hawaii at Manoa Advanced Studies in Genomics, Proteomics and Bioinformatics (ASGPB) facility at the University of Hawai‘i at Mānoa for Sanger sequencing.

Sequence data processing for ND1 and TRAF6 genes

This step is where the gene sequences are checked and aligned together so they can be compared. I imported sequence data into the Geneious v8.1.9 computer program (Ammundsen & Duran, 2015), and trimmed 5’ and 3’ ends (both sides) using a built-in algorithm with a 0.05 error probability limit, I aligned the reads using the Geneious alignment algorithm by automatically determining strand direction, conducting a global alignment with free end gaps, and a 65% cost matrix. These settings are described in the program, and define how the gene sequences for each individual are aligned together (see figures 4 and 5 for example sections of the gene alignments). I left open gap penalty and gap extension penalty at default values of 12 and 3, respectively. Following alignment of forward and reverse sequences, I visually examined the entire length of each sequence, I corrected base calls when necessary, and removed low-quality regions, to generate a consensus sequence for each of the genes. My output nexus file had a total of 990 alignment sites for the ND1 gene and 547 for the TRAF6 gene.

BLAST analysis for ND1 and TRAF6 genes

The BLAST tool compares a sequence of DNA to all sequences in a NCBI (National Center for Biotechnology Information) database to identify the species from which the DNA sequence is most similar to. I used the NCBI BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) tool to identify the possible hybrid individual by measuring similarity in the genetic sequence of the unknown individuals to those of known species.

Model testing and phylogenetic analyses for ND1 and TRAF6 genes

This step tests different models of how the DNA is evolving, so that a phylogenetic analysis (which looks at how the individuals are related to each-other evolutionarily) can be conducted. I used the jmodeltest version 2 software (Darriba, Taboada, Doallo, & Posada, 2012) to select a phylogenetic model among those implemented in Mr Bayes v3.2.5 (Huelsenbeck & Ronquist, 2001). The best-fit model of nucleotide evolution was HKY (which allows the frequencies of each of the four nucleotides to be different in the gene fragment, with one rate of evolution between them) with Gamma distributed rates (meaning there is a difference in the rate of evolution of nucleotide in each codon—the set of three nucleotides in a gene that determines

each amino acid in protein) across sites for the ND1 gene, and an HKY model with no rate variation across sites for TRAF6. To make phylogenetic trees, I ran MrBayes version 3.2.5 with two runs and four chains for 10,000,000 generations, discarding the first 25% as burn-in. MCMC convergence was inspected using Tracer v1.6.0 (Rambaut, Drummond, Xie, Baele, & Suchard, 2018). The resulting phylogenetic tree was then visualized using FigtTee version 1.4.4 (Rambaut, 2009).

Bench protocol for genomic data

I extracted genomic DNA following the same protocol outlined above. I included these samples in a larger ddRAD (double-digest restriction-enzyme associated DNA) library of *Sceloporus* species, following the methods outlined in Peterson, Weber, Kay, Fisher, & Hoekstra (2012) with some modifications. I used SbfI and NlaIII restriction enzymes for the restriction digest (to cut the DNA at specific locations), selected fragments in the range of 400-550 bp during the size selection step, ran 8 cycles of PCR, and standardized DNA concentration before pooling. The genomic library was sequenced at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley.

Bioinformatics protocol and phylogenetics for genomic data

I demultiplexed (separated individuals from a single data file) raw reads using STACKS version 2.52 (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013) with the following options in the program: recover barcodes with up to one mismatch, remove any reads with an uncalled base, discard reads with low quality scores, and truncate to 95 bp. I then loaded demultiplexed reads into ipyrad version 0.9.42 (Eaton & Overcast, 2020) for further processing. I used a reference-based approach using the Western Fence lizard genome sequenced by Harris, Banbury, and Leache (2014). I ran ipyrad with default settings, but I elected to keep only the nucleotides shared by at least 50% of the samples in the dataset. My output file had a total of 93,109 nucleotides. I used the jmodeltest version 2 program (Darriba et al., 2012) to evaluate phylogenetic model fit of those implemented in MrBayes v3.2.5 (Huelsenbeck & Ronquist, 2001). The resulting model was the general-time-reversible model of DNA substitution (GTR) (which allows the frequencies of each of the four nucleotides to be different in the gene fragment, with different rates of evolution between them) with gamma-distributed rate variation among sites and a proportion of invariable sites (meaning there is a difference in the rate of evolution of nucleotide in each codon—the set of three nucleotides in a gene that determines each amino acid in protein, and some sites do not vary). I ran the MrBayes program in four runs each with four heated chains for 10,000,000 generations, discarding the first 25% as burnin. MCMC convergence was inspected using Tracer version 1.6.0 (Rambaut et al., 2018). The resulting phylogenetic tree was then visualized using FigTree version 1.4.4 (Rambaut, 2009). In order to evaluate levels of heterozygosity between samples and gather more evidence about the possible hybrid, I processed the data using ipyrad (Eaton & Overcast, 2020) and consulted the per-sample heterozygosity output file and recorded the per-sample heterozygosity and error estimates for each individual. I followed the analysis recommendations and default settings specified by the program authors (Eaton & Overcast, 2020), but retained loci shared by 50% of samples in the dataset for analysis.