USING SNAKE GENOMES TO ILLUMINATE THE PATTERNS AND MECHANISMS OF RAPID ADAPTATION

by

DAREN CARTER CARD

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

AUGUST 2018

Acknowledgements

I am grateful to the many people who entered my life since I began this dissertation. Thank you to my close friends and lab mates in the Castoe Lab: Rich Adams, Blair Perry, Audra Andrew, Giulia Pasquesi, Nicky Hales, Andrew Corbin, and Jacobo Reyes-Velasco. Drew Schield deserves special recognition, as we both moved to the area and commenced our research at the same time, making him essentially an academic twin. This group has enriched my life greatly with a combination of intelligence, colleageality, compassion, patience, and encouragement, and my research would not have been possible without them. Several collaborators outside the lab have been particularly rewarding to work with, including the groups headed by Drs. Tereza Jezkova, Heath Blackmon, Maggie Hunter, Kristen Hart, Chad Montgomery, Scott Boback, and Warren Booth. Thank you to the many friends and colleagues I have met since arriving at UT-Arlington, especially Jill Castoe, Eli and Rachel Wostl, Alex Hall, James Titus McQuillan, and Kyle Shaney. To my committee members, Drs. Esther Betran, Jeff Demuth, Matt Fujita, and Matt Walsh: I am grateful for the encouragement, knowledge, and ideas that you provided, which greatly improved my research. Thank you to the administrative staff in the biology department – Sherri Echols, Gloria Burlingham, Ashley Priest, and, especially, Linda Taylor – for all of the help. Special thanks to Corey Roelke for helping me with the fun and rewarding experience of teaching others. I am eternally grateful for the support of my parents, Gary and Claire, who raised me to love science and strive to do my best in everything, and to my brothers, Jay and Andrew, the best friends that nature can provide. Thank you to my wife, Rachel, for whom this dissertation is dedicated; I look forward to continuing life's journey with you. Finally, no one will ever replace the unparalleled friendship and mentorship of my advisor, Todd Castoe, which I have internalized and will strive to replicate in my future.

July 10th, 2018

Dedication

To my wife, Rachel: thank you for your tireless support, patience, and inspiration as I pursued this endeavor.

Abstract

USING SNAKE GENOMES TO ILLUMINATE THE PATTERNS AND MECHANISMS OF RAPID ADAPTATION

Daren C. Card, PhD

The University of Texas at Arlington, 2018

Supervising Professor: Todd A. Castoe, PhD

One of the most important and interesting goals in evolutionary biology is to understand the mechanisms generating biodiversity and adaptive novelty. Ever-evolving genomic techniques have served as a catalyst for this work, enabling rapid increases in our knowledge of diverse taxa. By leveraging a combination of phylogeographic, population genetic, and comparative genomic methods, I established two snake systems with unique attributes that showed promise for increasing our understanding of important evolutionary questions related to local adaptation and convergence. Using sampling from several island populations of *Boa imperator* with similar adaptive phenotypes (e.g., reduced body size and craniofacial morphological shifts), I deduce that unique island phenotypes have evolved independently in at least three populations. Moreover, I explored the contribution of genetic drift and adaptation, as well as idiosyncratic versus convergent molecular evolution, in the evolution of morphological, physiological, and natural history traits shared across distinct island populations. I also investigated ecological shifts related to novel feeding ecology and climate within an invasive population of Burmese python (*Python molurus bivittatus*) and found evidence for extremely rapid adaptation of complex physiological traits related to these selective pressures. Collectively, this dissertation exemplifies the power that non-model snake species hold for understanding important evolutionary questions using novel genomics approaches.

Table of Contents

Chapter 1.

Introduction

Genomic approaches have revolutionized many areas of biology and are continuing to illuminate the links between genotype and phenotype, leading to paradigm-changing biological discoveries. Perhaps the biggest leaps in knowledge are coming from studies on traditional 'non-model' systems (i.e., organisms other than *Drosophila*, mouse, or human, for example) where genetic or molecular information is sometimes totally nonexistent. My dissertation has leveraged cutting-edge molecular biology to generate large amounts of genomic data, and powerful computational approaches have allowed me to discern key information about the biology of different organisms. Chapter 1, for example, used relatively low-coverage genome sequencing data from two non-model bird species and existing high-quality reference genome from relatively distantly related model organisms to produce reference-guided genome assemblies that provided significantly greater inferential power than traditional *de novo* assembly techniques.

Genomics, however, only provides an analytical framework, and major intellectual driver of this dissertation is understanding the evolutionary processes that generate biodiversity. This work has focused on populations of two widespread, generalist snake species – the Burmese python and boa constrictor – where recent isolation has led to the evolution of unique, adaptive phenotypes. My goal in both studies was to understand the genomic basis of recently-evolved adaptive traits.

As has been observed in many taxa, island populations of boa constrictor (snakes in the genus *Boa*) have evolved unique phenotypes after becoming isolated relatively recently (i.e., since the end last glacial maximum approximately 10,000 years ago). This constellation of traits, including reduced body size and shifts in craniofacial morphology, occurs on several geographically distinct islands and appears to be an adaptive response to unique island ecosystems. Evolutionary convergence towards remarkably similar eco-morphotypes across islands motivated several chapters of this dissertation focused on developing key genomic resources necessary for this system (Chapter 3), on understanding the structure and relationships between populations, including whether islands are evolutionarily independent (Chapter 4), and on understanding the genomic basis of convergent adaptive phenotypes across islands (Chapter 5). The final chapter of this dissertation focused on Burmese pythons (*Python molurus bivittatus*), which are native to Southeast Asia, but have recently become established as an invasive population in South Florida. These snakes have proliferated in a novel ecosystem where they have shifted their feeding ecology and contend with periodic freeze events, motivating a study of how extremely rapid and complex adaptation has occurred in this invasive population over just a few generations (Chapter 6).

Ideally this dissertation motivates further investigations of local adaptation in natural populations of non-model organisms, which hold great potential for helping biologists to understand how natural selection drives both divergent and convergent phenotypic evolution to maintain and generate biodiversity.

Chapter 2

Two low coverage bird genomes and a comparison of reference-guided versus *de novo* genome assemblies

Daren C. Card[1], Drew R. Schield[1], Jacobo Reyes-Velasco[1], Matthew K. Fujita[1], Audra L. Andrew[1], Sara J. Oyler-McCance[2], Jennifer A. Fike[2], Diana F. Tomback[3], Robert P. Ruggiero[4], and Todd A. Castoe[1]

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

[2] United States Geological Survey – Fort Collins Science Center, Fort Collins, CO, 80526 USA

[3] Department of Integrative Biology, University of Colorado Denver, Denver, CO, 80217 USA

[4] Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO, 80045 USA

ABSTRACT

As a greater number and diversity of high-quality vertebrate reference genomes become available, it is increasingly feasible to use these references to guide new draft assemblies for related species. Reference-guided assembly approaches may substantially increase the contiguity and completeness of a new genome using only low levels of genome coverage that might otherwise be insufficient for *de novo* genome assembly. We used low-coverage (~3.5-5.5x) Illumina paired-end sequencing to assemble draft genomes of two bird species (the Gunnison Sage-Grouse, *Centrocercus minimus*, and the Clark's Nutcracker, *Nucifraga columbiana*). We used these data to estimate *de novo* genome assemblies and reference-guided assemblies, and compared the information content and completeness of these assemblies by comparing CEGMA gene set representation, repeat element content, simple sequence repeat content, and GC isochore structure among assemblies. Our results demonstrate that even lower-coverage genome sequencing projects are capable of producing informative and useful genomic resources, particularly through the use of reference-guided assemblies.

INTRODUCTION

High quality sequencing, assembly, and annotation of vertebrate genomes have become feasible for non-traditional model species, as costs of sequencing decrease and analysis methods improve. The default method for generating initial genome assemblies for a species includes the use of *de novo* assembly algorithms that rely on sufficient overlap between sequencing reads to build larger contiguous sequences. This approach is fundamentally different from a reference-guided approach that utilizes existing contiguous sequences and sequence similarity between the target and reference species' genomes to assemble a genome. The availability of high quality reference genomes for a greater diversity of vertebrate species may enable inexpensive yet informative genomic resources to be generated for new species by leveraging information from existing high-quality genomes of related species. If there is a relatively high degree of synteny among related species, a reference-guided genome assembly approach may be capable of delivering more complete and biologically useful genome resources with far less data and computational effort than required for full *de novo* genome assembly. Thus, we may potentially achieve greater representation and understanding of genomic diversity across the tree of life through the use of high-quality genomes, complemented by the addition of lower-coverage genomes.

Among amniote vertebrates, birds possess among the smallest genomes and the lowest levels of repetitive elements (International Chicken Genome Sequencing Consortium, 2004; Shedlock et al., 2007; Warren et al., 2010). These two characteristics make their genomes relatively inexpensive to sequence and also make mapping and assembling genomic sequencing reads computationally more tractable. Bird genomes are also highly conserved at the chromosomal level, such that there is a high degree of synteny across chromosomes of divergent bird species

(Ellegren et al., 2012; Shetty, Kirby, Zarkower, & Graves, 2002; Vicoso, Kaiser, & Bachtrog, 2013). This karyotypic conservation facilitates ready transfer of information from one bird genome to another (Ansari, Takagi, & Sasaki, 1988; Ogawa, Murata, & Mizuno, 1998; Swathi Shetty, Griffin, & Graves, 1999) and justifies their use as a system to test a reference-guided genome assembly approach in this study. Birds are important model systems for a broad diversity of research, and having genomic information to facilitate these diverse research programs for all bird species would be ideal, which motivates the development of efficient and inexpensive means of assembling genomes and genomic resources. This raises the questions: 1) Can low coverage sequencing of new bird genomes be used to economically produce biologically valuable genome resources by leveraging existing complete genomes, and 2) How does the content of different types of biological features (e.g., genes, transposable elements, and GC-isochores) compare among low coverage *de novo*, low coverage reference-guided, and existing high-coverage high quality genomes?

In this study we use existing high-quality bird genomes from the Chicken (*Gallus gallus*; International Chicken Genome Sequencing Consortium, 2004) and the Zebra Finch (*Taeniopygia guttata*; Warren et al., 2010) to guide the assembly of two distantly related bird species, the Gunnison Sage-Grouse (*Centrocercus minimus*; Order Galliformes, Family Phasianidae, “Sage-Grouse” hereafter) and the Clark’s Nutcracker (*Nucifraga columbiana*; Order Passeriformes, Family Corvidae, “Clark’s Nutcracker” hereafter). For the purposes of this study, we define a high-quality reference genome as a genome with N50 contig lengths of >10kb that have been ordered and combined into supercontigs (or scaffolds). Ideally a high-quality genome would also have >200Mb scaffolds, which are mapped to physical chromosomes (as is the case with the two bird reference genomes used here). The Clark’s Nutcracker is an important seed disperser for two widely distributed Western North American conifers, whitebark pine (*Pinus albicaulis*) and limber pine (*P. flexilis*), which are declining due to the outbreaks of the mountain pine beetle (*Dendroctonus ponderosae*) and the invasive disease white pine blister rust (*Cronartium ribicola*; (Schoettle & Sniezko, 2007; Tomback & Achuff, 2010; Tomback, Arno, & Keane, 2001). Because the Clark's Nutcracker-mediated seed dispersal is key to maintaining viable populations of these imperiled pines (Barringer, Tomback, Wunder, & McKinney, 2012; Tomback, 1982), knowledge of population structure and dynamics of the Clark's Nutcrackers may provide important information relevant to management of these trees. The Gunnison Sage-Grouse is a geographically restricted species of grouse found south of the Colorado River in Colorado and Utah. The entire species consists of seven small populations ranging in size from 40 birds in the smallest population to roughly 2,500 in the largest (Gunnison Sage-Grouse Rangewide Steering Committee, 2005; Stiver, Apa, Remington, & Gibson, 2008). Most populations are isolated from one another and have low levels of genetic diversity (Oyler-McCance, St. John, Taylor, Apa, & Quinn, 2005). This species has been proposed for listing as threatened or endangered under the U.S. Endangered Species Act. The Sage-Grouse is in the order Galliformes along with the Chicken (*Gallus gallus*), for which a high quality genome is available (International Chicken Genome Sequencing Consortium, 2004). Similarly, the Clark's Nutcracker belongs in the order Passeriformes with the Zebra Finch (*Taeniopygia guttata*), for which there is also a high-quality genome (Warren et al., 2010). These available high-quality genomes from species related to our two species of interest present an opportunity to evaluate the utility and feasibility of reference-guided (versus *de novo*) assembly strategies.

Reference-guided genome assembly approaches have been used previously (e.g., Mellmann et al., 2011; Nishito et al., 2010; Parchman, Geist, Grahnen, Benkman, & Buerkle, 2010;

Schneeberger et al., 2011) and various pipelines currently exist for reference-guided assembly (e.g., MOSAIK – http://code.google.com/p/mosaik-aligner/; DNASTAR – http://www.dnastar.com/default.aspx). Indeed, many bacterial genomes have been generated with this approach (e.g., Mellmann et al., [2011]; Nishito et al., [2010]). The sequencing coverage in previous studies was, however, moderately high (>10x), and the reads were mapped to a guide genome of a very closely related species (e.g., a different strain of a species or a sister species in Schneeberger et al., [2011] and Parchman et al., [2010]). Here we evaluate the feasibility of using relatively low genomic coverage (~3.5 - ~5.5x) to assemble draft bird genomes using reference genomes from relatively distantly related species (>40 million years divergence between the species studied and the species' genomes used to guide the assembly; Ericson, Jansén, Johansson, & Ekman, 2005; Kan et al., 2010; Pereira & Baker, 2006; Phillips, Gibb, Crimp, & Penny, 2010). We hypothesized that with such low sequencing coverage, a traditional *de novo* assembly approach would yield a less contiguous genome with fragmentary biological features, but that a reference-guided approach might provide substantial gains in contiguity and the presence of intact biological features. Indeed, we find that the reference-guided approach substantially improves assembly and yields more informative genome assemblies as measured by most assessment metrics, indicating that this type of approach provides an economical alternative method for obtaining a preliminary estimate of genomic diversity and structure across a very large number of vertebrates.

MATERIALS AND METHODS

Ethics statement

Sage-Grouse blood was obtained from a single individual bird from Gunnison County, Colorado, USA, where no permit was required for trapping at the time of sampling. The trapping and sampling approach was approved and carried out by the Colorado Division of Wildlife. The Clark's Nutcracker muscle was sampled from an individual bird trapped near Logan, Utah, USA, which was kept as part of a long-term study at Northern Arizona University (IUCUC protocol 00-006) before its death from natural causes; the carcass was donated for genetic work by Alan Kamil (University of Nebraska) and Russell Balda (Northern Arizona University).

Preparation and sequencing of shotgun sequencing libraries

The methods used to prepare and sequence shotgun libraries of the Sage-Grouse and the Clark's Nutcracker were described previously (Castoe et al., 2012). Briefly, DNA was extracted from blood (Sage-Grouse) and muscle (the Clark's Nutcracker) samples using standard phenol-chloroform-isoamyl alcohol separation and the Wizard Genomic DNA Purification Kit (Promega) respectively. Illumina paired-end libraries were prepared by fragmenting genomic DNA using nebulization, ligation of "Y"-adapters, and size selection of libraries from agarose electrophoretic gels. The libraries, including adapters, had a mean size of 325 bp and were sequenced on the Illumina GAIIx platform with 120 bp paired-end reads. Raw sequence data were deposited in the NCBI Short Read Archive (SRA Accessions SRX468855 for the Sage-Grouse and SRX468897 for the Clark's Nutcracker).

De novo *draft genome assembly*

Raw read data were first demultiplexed and quality-trimmed to remove low quality reads and base calls in CLC Genomics Workbench using a modified Mott trimming algorithm and a parameter value limit of 0.05; ambiguous nucleotides were trimmed using a maximum number of ambiguities of two. *De novo* assembly was conducted in CLC Genomics Workbench using automatic word size and bubble size, and a minimum contig length of 200 bp. Paired read distances were automatically detected and contigs were scaffolded where possible. Following assembly, the reads were mapped back to the contigs using a mismatch cost = 2, insertion cost = 3, deletion cost = 3, length fraction = 0.5, and similarity fraction = 0.8; contigs were updated and gaps were filled.

Reference-guided draft genome assembly

We used the Chicken (*Gallus gallus* v. Galgal4; International Chicken Genome Sequencing Consortium, 2004) and the Zebra Finch (*Taeniopygia guttata* v. taeGut3.2.4; Warren et al., 2010) genomes to guide assembly of the Sage-Grouse and the Clark's Nutcracker, respectively. Quality trimmed reads from the two species in this study were mapped against their respective guide genome using CLC Genomics Workbench, with a mismatch cost = 2, insertion cost = 3, deletion cost = 3, length fraction = 0.5, and similarity fraction = 0.8, with paired distances automatically detected. A consensus sequence for each new species was exported using different thresholds of minimum coverage for reads mapping to the consensus (1x, 2x, and 5x). For example, a 1x reference-guided assembly denotes the consensus sequence at all positions where at least one read mapped. At positions where the threshold of minimum coverage was not met, an N ambiguity was inserted. At positions where disagreements in base calls were observed between reads (with disagreements representing at least 10% of the total reads at that position, and at least two reads supporting an alternative allele), an appropriate ambiguous nucleotide symbol was inserted.

Calculation of basic genome statistics and breaking of poly-N stretches

The reference-guided assemblies resulted in a mosaic of non-ambiguous regions interspersed with stretches of N ambiguities. Shorter stretches of N ambiguities are typical even in high quality scaffolded genome assemblies, but longer stretches (>500 bp) typically are not. Therefore, for the reference-guided assemblies we used a Perl script to break the consensus contigs at N ambiguity stretches of greater than 500 consecutive Ns. For the modified reference-guided assemblies and the *de novo* assembly, we assessed contiguity by calculating the frequency distribution of contig lengths and calculated standard statistics, such as the N50 contig length.

Analysis of CEGMA genes and repeat element content

To assess the completeness of each assembly with regard to gene content we used the CEGMA pipeline (Parra, Bradnam, & Korf, 2007), which searches assemblies for a set of core eukaryotic genes (CEGs) that are highly conserved and present in nearly all eukaryotes. The proportion of complete and partial CEGs (out of 248 possible) is taken as a measure of the completeness of the gene content of an assembly. The CEGMA pipeline was run on the *de novo* assembly, the three reference-guided assemblies, and the guide reference genomes.

Repeat elements often increase the difficulty of vertebrate genome assembly, and therefore might be underrepresented in lower-quality assemblies. We compared the repeat element content across all assemblies by annotating repeats using *RepeatMasker* (Smit, Hubley, & Green, 2013), using the standard "avian" *Repbase* repeat element library (Jurka et al., 2005). All other settings for *RepeatMasker* were set to default values.

A previous study quantified Single Sequence Repeat (SSR; also known as microsatellite) content in both of these bird species based on analysis of the raw unassembled Illumina reads (Castoe et al., 2012). We repeated the analysis on the *de novo* and reference-guided assemblies for both species to assess if SSR content varied among genome assemblies compared to the raw reads (which might indicate the under-representation of SSRs in certain assemblies). We used *Palfinder* v0.02.03 (Castoe et al., 2012) to identify SSRs across genome assemblies, with an SSR being classified as a stretch of 2-6mer tandem repeats that met a certain tandem repeat threshold: 6 tandem repeats for 2mers, 4 tandem repeats for 3mers, and 3 tandem repeats for 4mers, 5mers, and 6mers. For comparative purposes, we used the same methods to estimate SSR content in both reference genomes used, as well as the Turkey (*Meleagris gallopavo*; Dalloul et al., 2010) and the *Anolis* lizard (*Anolis carolinensis*; Alföldi et al., 2011) genomes.

Analysis of GC isochore structure

To examine whether such relatively low coverage genome assemblies could provide information about genomic GC isochores, we compared patterns of regional variation in nucleotide composition (e.g "isochores") between our reference-guided genomes and other high-quality vertebrate genomes. To do this, we estimated the standard deviation of GC content for genomic windows of varying sizes: 3-, 5-, 10-, 20-, 80-, 160-, and 320-kb. The expectation is that standard deviation will decrease as window sizes increase; based on a completely homogeneous genome, variation will halve as window sizes quadruples (International Human Genome

Sequencing Consortium, 2001). Deviations from this expectation indicate a genome with structural variation in GC content, as observed in mammals and birds but not in the *Anolis* lizard genome (Fujita, Edwards, & Ponting, 2011). In addition, we randomly sampled 3- and 5-kb windows from the Chicken genome to match the sample size in the Clark's Nutcracker to determine whether the sample size of the dataset was representational of genome-wide estimate of GC structure at these spatial scales. Patterns in GC variation, and how it declines as window size changes, can quantify the heterogeneity of GC content in a genome. For example, a genome that has a large GC content standard deviation for larger windows has significant nucleotide composition heterogeneity at a large spatial scale, indicative of strong isochore structure. Multiple mammal, bird, and reptile genomes were used to compare the compositional structure of genomes among vertebrates.

Variant analysis

We analyzed the relative frequencies of various types of heterozygous variants in the two bird genomes by mapping our quality-filtered Illumina reads back to the 1x reference-guided assemblies and by applying a Bayesian approach to determine the probability of heterozygosity at each position implemented in the Probabilistic Variant Detection tool in CLC Genomics Workbench. Heterozygous variants were filtered based on the following criteria: a minimum coverage of 4 reads, with at least two reads supporting a variant, and a variant probability of at least 80%. The analysis ignored non-specific matches, broken paired-end reads, and variants in non-specific regions, and required the presence of a variant in both the forward- and reverse-facing reads, and to expect a maximum of 2 variants per position. We further filtered these data to provide a more robust estimate of the heterozygosity using the following parameters and thresholds: read coverage greater than 5 reads, allele frequencies between 30% and 70%, forward and reverse reads both support the variant in at least 30% of the reads, and an average PHRED quality score of greater than 40.

Mitochondrial genome assembly

Mitochondrial genome reads were extracted from all reads prior to genome assembly, and used to reconstruct the mitochondrial genomes of both species for use in divergence time estimation between our target species and species used as genome references for each of our targets. The mitochondrial genome of each bird was identified by using *blast* (Altschul, Gish, Miller, Myers, & Lipman, 1990) to search for *de novo* assembled contigs using the consensus complete mitochondrial genome sequence from all members of the order Galliformes (Sage-Grouse), and a consensus for the family Corvidae (Clark's Nutcracker; Supplementary Tables 1-2). Contigs from the assembly that were matched by *blast* to the mitochondrial genome consensus sequences (of other previously sampled birds) were used to further assemble the mitochondrial genome. We created the assemblies by mapping the *blast* hits to the consensus mitochondrial genome sequence in CLC Genomics Workbench, using a mismatch cost = 2, insertion cost = 3, deletion cost = 3, length fraction = 0.5, and similarity fraction = 0.8. The consensus sequence was then exported using a minimum coverage threshold of 1x. At positions where the threshold of low coverage was not met, an N ambiguity code was inserted. We note that a separate study has recently conducted similar analyses using these data and deposited on NCBI nearly identical results (Barker, Oyler-McCance, & Tomback, 2013), and we therefore have not deposited our versions of these mitochondrial genome sequences in NCBI to avoid redundancy. We have, however, used our versions of these mitochondrial genomes for analysis because they were slightly more complete for some genes for the Sage-Grouse. Additionally, identification and removal of mitochondrial reads from the remaining data enable characterization of patterns solely from the nuclear genome of both species.

Mitochondrial gene phylogeny and divergence estimates

To accurately date divergence times between our target species and those that we used as guides for assembly, we obtained additional mitochondrial genomes from NCBI. We chose taxa to represent most avian lineages, with diverse representatives of the Galliformes, Passeriformes, and several outgroups (n = 20 taxa; see Fig. 9 and Supplementary Table 3), and specifically included taxa for which divergence times had been estimated previously (Ericson et al., 2005; Kan et al., 2010; Pereira & Baker, 2006; Phillips et al., 2010). Our phylogenetic analysis included sequences from 12 mitochondrial protein-coding genes (excluding ND6 and all non-coding loci; see Supplementary Table 3 for NCBI accession numbers). Annotated sequences from the mitochondrial genome of the Chicken were used as a reference to align and trim sequences. Complete mitochondrial protein sequences were then aligned using *Geneious* 6.1.6 (Biomatters Ltd.), followed by minor manual adjustment, and were concatenated using *Sequence Matrix 1.7.8* (Vaidya, Lohman, & Meier, 2011). Best-fit models of nucleotide evolution for each gene and codon position were estimated using Bayesian Information Criterion (BIC) in the program *PartitionFinder v1.1.1* (Lanfear, Calcott, Ho, & Guindon, 2012). The final alignment included a total of 10,845 bases for each species. A list of the best-fit models of nucleotide evolution used is included in the supplementary materials (Supplementary Table 4).

We estimated phylogenetic relationships using Bayesian Markov Chain Monte Carlo inference (BI) with all concatenated genes in *MrBayes version 3.2.1* (Ronquist & Huelsenbeck, 2003).

Analyses were conducted using 10^7 generations for each of two simultaneous runs, each with four chains (three heated and one cold) that were sampled every 1,000 generations. We estimated divergence times among taxa using BEAST 2 (Bouckaert et al., 2014; Drummond & Rambaut, 2007), and used the consensus tree resulting from *MrBayes* as a starting guide tree for BEAST 2 analyses. Divergence estimation in BEAST 2 used the concatenated mitochondrial gene set, with an HKY substitution model, a lognormal relaxed clock model, and a Yule process tree prior. We constrained nodes using dates obtained from previous mitochondrial divergence time estimates (Ericson et al., 2005; Kan et al., 2010; Pereira & Baker, 2006; Phillips et al., 2010). A list of calibration points used in the analysis is given in the supplementary materials (Supplementary Table 5). Two independent analyses were run for 5 x 10^6 generations, sampling every 1,000 generations. We used the program *Tracer* (Drummond & Rambaut, 2007) to confirm if the analyses had reach convergence based on likelihood and parameter value stationarity, and based on this discarded the first 10% of generations from each run as burn-in. We used the program *TreeAnnotator v. 1.7.4* (Drummond & Rambaut, 2007) to summarize parameter values of the samples from the posterior on the consensus tree.

RESULTS

Genome de novo *assemblies*

Assuming that the genome sizes of each species equaled the mean known genome size for their respective families (both 1.32 Gb; (Gregory, 2013; Gregory et al., 2007)), our genome sampling represents approximately 3.53x genome coverage of the Sage-Grouse and 5.41x for Clark's-Nutcracker (Table 1). A summary of the numbers of reads, total bases, and estimated genome sizes are given in Table 1. The *de novo* assembly of the Sage-Grouse totaled 309,822,517 bp, comprising 914,239 scaffolded contigs (Fig. 1A; Table 2). Most contigs were less than 1,000 bp in length (Fig. 1A), and the N50 contig size was 343 bp (Fig. 2A). The assembly consisted of 31.6% Adenine (A), 18.5% Cytosine (C), 19.0% Guanine (G), and 30.9% Thymine (T). The *de novo* assembly of the Clark's Nutcracker totaled 679,286,238 bp, comprising 1,457,264 scaffolded contigs (Fig. 1B; Table 2). While most contigs were again less than 1,000 bp in length, contig sizes tended to be slightly larger in the Clark's Nutcracker than in the Sage-Grouse (Figs. 1A-B). This slight shift upward in contig size is also observed in the larger N50 contig size in the Clark's Nutcracker (503 bp; Fig. 2B), as well as a higher maximum contig size (18,041 bp). The assembly consisted of 29.5% (A), 20.5% (C), 20.8% (G), and 29.0% (T).

Reference-guided assemblies

The total length of reference-guided assemblies for the Sage-Grouse were over 1 Gb, approximating the length of the Chicken reference genome, though a large fraction of this sequence consisted of "N" ambiguities due to low coverage and/or the number of reads mapping to the reference falling below set thresholds (Fig. 2C). When genome segments containing stretches of at least 500 N bases were removed, most remaining contigs were longer than 1,000 bp, with many being 10,000 bp or greater in the 1x reference-guided genome (Fig. 1C); this trend is also clear from the larger N50 contig sizes observed in the reference-guided assemblies (Fig. 2A; Table 2). The reference-guided assemblies for the Clark's Nutcracker showed trends similar to the Sage-Grouse in having substantial numbers of ambiguous bases comprising the reference-guided assemblies (Fig. 2D). The contigs that resulted from splitting stretches of at least 500 N bp were predominantly greater than 1,000 bp in length, with some contigs longer than 30 kb in the 1x reference-guided genome (Fig. 1D); N50 contig sizes for all three reference-guided assemblies were greater than 1,000 bp (Fig. 2B; Table 2). The *de novo* assembly, all reference-guided assemblies, and a chromosome annotated version of the 1x reference-guided assembly are available for each species from the Dryad Digital Repository (Card et al., 2014).

Presence of CEGMA genes in assemblies

We used CEGMA to assess the completeness of assemblies with respect to protein coding regions in both the *de novo* and the reference-guided genomes. *De novo* assemblies for both species had consistently far lower numbers of CEGMA genes identified (either partial or complete) compared to the reference-guided assemblies (Figs. 3A-3B), with the 1x reference-guided assemblies containing the most CEGMA genes (Fig. 3). It is notable that we observed substantial increases in CEGMA gene content with relatively minor changes in assembly length among the reference-guided assemblies with different read depth cutoffs (Figs. 2A-2B and 3). Comparing the two species, the Clark's Nutcracker assemblies showed systematically higher recoveries of CEGMA genes than the Sage-Grouse (Fig. 3), which parallels the higher coverage, longer contigs, and larger non-ambiguous assemblies in the Clark's Nutcracker.

Repeat element content

Because repetitive elements are notoriously difficult to assemble, we compared the abundance of repetitive elements in various genome assemblies. *A priori*, we assumed that poorly assembled or less completely assembled genomes would contain fewer annotated repetitive elements than higher-quality and more complete genomes. In general, this expectation holds in comparisons between the reference genomes and our *de novo* and reference-guided assembly genomes (Fig. 4). In the Sage-Grouse, the genome assembly with the most repetitive content was the 1x reference-guided assembly, followed by the *de novo* assembly (Fig. 4). In the Clark's

Nutcracker, which also had substantially more raw read data, the *de novo* assembly contained the greatest repeat element fraction compared to the reference-guided assemblies (Fig. 4). Neither the *de novo* or reference-guided assemblies, however, contained a similar amount of repeat elements as that in the respective reference genomes, indicating that much of the unassembled parts of the Clark's Nutcracker and the Sage-Grouse genomes may represent a biased failure to incorporate repeat elements.

Simple sequence repeat content

We estimated simple sequence repeat (SSR, or microsatellite) content of various assemblies to further examine qualitative and quantitative ways in which the *de novo* and reference-guided assemblies differed, and how they compared to high quality reference genomes. Because raw reads can also be used to identify SSR content (Castoe et al., 2012), we included analysis of unassembled reads in comparisons. Analogous to our findings with general repeat elements, we determined that the *de novo* assemblies contain the highest abundances of SSRs (Fig. 5). Also, unlike the general repeat element analysis, the SSR content estimates from the *de novo* assemblies are relatively similar to estimates in the high quality reference genomes, although the estimates derived from raw reads proved to be even better approximations to SSR densities observed in high-quality reference genomes (Fig. 5). Comparative analysis of SSR content across bird species indicates that genomic SSR content is relatively conserved among avian genomes, except for some variance in the abundance of 2-4mers (Fig. 6). In contrast to the conservation of the SSR landscape across bird species, the SSR landscape changes extensively between birds and the *Anolis* lizard, particularly in the abundance of 2-4mer SSRs (Fig. 6).

Genomic GC-isochore structure

Comparison of genomic GC-isochore structure across vertebrates is typically thought to require very well-assembled genomes, because it requires long contiguous regions of genome assemblies. We were interested to test if reference-guided genomes could be used for estimation of GC-isochore structure, and if they produced results that were reasonable compared to other related bird species. Overall, the *de novo* genome assemblies for both bird species did not contain enough contigs to adequately estimate GC content variation at large spatial scales. The 1x reference-guided assembly yielded the highest number of contigs at each window size and was used for subsequent comparison with other vertebrate genomes and with a randomly-sampled, proportionally reduced representation 3- and 5-kb contig sample from the Chicken. The distribution of GC content for the Sage-Grouse differed considerably from any other vertebrate genome, most likely because the estimate of GC isochore structure was unreliable for this species' assembly, which also had very low genome coverage and small contig sizes. However, the distribution for the Clark's Nutcracker was much more similar to that of other vertebrates, yet differed from the other bird genomes in having a slightly higher GC content and a more narrow distribution (Fig. 7A). To examine whether these differences are the consequence of the smaller sample sizes (73,158 and 35,090 3- and 5-kb windows, respectively, versus 338,120 and 202,814 3- and 5-kb windows, respectively, in the Chicken), we used a random subset of the Chicken genome windows to match the sample sizes of genomic windows available for the Clark's Nutcracker. We compared the GC distributions between the full and reduced sample sizes in the Chicken and found no difference (Kolmogorov-Smirnov test: $p = 0.5026$ for the 3-kb window size comparison and $p = 0.8398$ for the 5-kb window size comparison), indicating that such a reduced data set of genomic windows provides an adequate representation of the genome-wide

GC content distribution at 3- and 5-kb window sizes. This, together with the inference of no clear assembly bias in GC content (Supplementary Table 6), indicate that the GC distribution of the Clark's Nutcracker at the 3- and 5-kb window sizes is expected to accurately reflect the genomic GC content variation at these various spatial scales (Fig. 7B).

Variant detection

We examined variants with reasonable coverage thresholds to compare the relative frequencies of observed types of heterozygous variants between species. Overall, the relative levels of heterozygous variants for each bird were approximately equal, despite the Clark's Nutcracker having nearly double the number of each variant type when compared to the Sage-Grouse; this was expected due to the lower number of sites that met the criteria for calling heterozygous variants in the Sage-Grouse. Single nucleotide variants (SNVs) were most frequently observed with deletions also occurring regularly, and SNVs that represented transitions were much more frequently observed than transversions (Fig. 8). Multiple nucleotide variants (MNVs), insertions, and replacements were represented in lower frequencies in both genomes, but were similar in relative frequencies among the two species (Fig. 8).

Mitochondrial genome assemblies

The reference-guided mitochondrial genome assembly for the Sage-Grouse was incomplete and was likely related to the lower coverage available for this species; 59.08% of the mitochondrial genome was unresolved (and represented as ambiguities), and three of the 12 mitochondrial protein-coding loci used for phylogenetic analysis were essentially absent (and the remaining nine contained some ambiguous regions). Despite this partial assembly, these data provided an ample number of aligned sites to conduct phylogenetic analyses. The reference-guided mitochondrial genome for the Clark's Nutcracker was much more complete than the Sage-Grouse. Across the entire mitochondrial genome, only 8.69% of sites were ambiguous ("N"s). For the Clark's Nutcracker, all 12 protein-coding mitochondrial genes used for phylogenetic analysis were present and contained no ambiguous bases. Annotated versions of the assemblies are available from the Dryad Digital Repository (Card et al., 2014). Mitochondrial genome assembly and annotation was therefore more complete for the Clark's Nutcracker than for the Sage-Grouse, which may due to the relative amount of data combined with the density of mitochondria in the different tissue sources used for DNA extraction: blood in the case of the Sage-Grouse versus muscle tissue in the case of the Clark's Nutcracker (Barker et al., 2013).

Mitochondrial phylogeny and divergence dating of birds

Using the newly assembled mitochondrial genomes, we were able to estimate the phylogenetic relationships of the Clark's Nutcracker and the Sage-Grouse, as well as divergence times between these species and several other species of birds, including the two species used as reference genomes for guided assemblies. The Bayesian analysis recovered four major clades among the species sampled, which correspond to the major groups of birds, and all nodes received strong support (>95% posterior). We inferred that the Clark's Nutcracker formed a clade with the Rook (*Corvus frugilegus*), while the Sage-Grouse was nested in the Galliformes as sister species to the Hazel Grouse (*Bonasa bonasia*), and our divergence time estimates resulted in divergence ages similar to those of previous studies (Fig. 9; (Ericson et al., 2005; Kan et al., 2010; Pereira & Baker, 2006; Phillips et al., 2010)). Most importantly, we estimated that the Sage-Grouse split from its common ancestor with the Hazel Grouse approximately 27 million years ago (mya), while it split from the Chicken (*Gallus*) about 43 mya, and that the Clark's

Nutcracker diverged from its common ancestor with the Rook approximately 28 mya and from the Zebra Finch approximately 61 mya.

DISCUSSION

Our results demonstrate that substantial information can be extracted from lower-coverage genomic sampling projects, and that reference-guided assemblies provide much better representation of biologically important regions than *de novo* assemblies when genome coverage is low. We were surprised that reference-guided assembly approach was quite successful despite substantial divergence between target species and reference genome species (~40-60 mya; Fig. 9), and with fairly low levels of sequencing coverage (Table 1). While we suggest that higher coverage is preferable, our results provide an exciting proof of concept for an economical strategy to increase the diversity of vertebrate genome resources by using reference-guided assembly approaches. This strategy would be particularly useful for species that are somewhat closely related to those for which high-quality reference genomes are available. Such reference-guided low-coverage genomes do indeed fall short of the completeness of information contained in high-quality *de novo* assembled genomes, although our results indicate that compared to an alternative of having no information at all for a species, or to a highly fragmented *de novo* assembly from low-coverage data, reference-guided assemblies are capable of providing substantial biological information about the genome of a species at low cost.

While reference-guided genomes do appear to contain large amounts of biological information, the accuracy of this information is unknown, and probably dependent on the type of feature and the divergence between target and reference species. For example, estimates of most protein-coding genes are likely accurate given their conserved nature. More rapidly diverging genomic features or regions, such as transposable elements or other non-coding regions, may be more prone to inaccuracies in reference-guided assemblies. These inaccuracies will also increase with divergence between reference-target species, which may indeed lead to spurious contigs or nucleotide stretches that are not present in the actual garget genome. Thus, reference-guided genome estimates should be applied with the understanding that they may indeed be prone to inaccuracies and error, depending on reference-target sequence divergence. For this reason, it is also not wise to use one reference-guided assembly as a reference for a second reference-guided assembly, because errors and inaccuracies in assembly from one would be both perpetuated and compounded.

In both bird species analyzed here, reference-guided assemblies provided more complete representation of some important genomic features compared to *de novo* assemblies. The greatest difference in content among alternative assemblies was the number of CEGMA genes identified, with our *de novo* assemblies finding extremely few and reference-guided assemblies finding orders of magnitude more as coverage thresholds were lowered. This indicates that reference-guided approaches may be particularly useful for establishing genomic resources for gene-centric analyses. Repetitive elements tend to pose a particular challenge to *de novo* genome assembly in vertebrates (Li et al., 2010), and we expected repetitive element content to be higher (and more similar to reference genomes) in reference-guided versus *de novo* assemblies. This was not necessarily the case in our results, however, and, instead, both approaches seem to under-represent genomic repetitive element content, indicating that that these repetitive elements may be just as challenging for mapping (in reference-guided assembly) as they are for *de novo* assembly. Having more closely related reference genomes may substantially improve how well repeat element regions are assembled, as the ability to use a reference-guided approach to assemble these regions may be highly dependent on the degree of recent activity of repeat elements in a particular lineage. In contrast to major differences in repeat element content between new and reference genomes, and among assembly approaches, SSR estimates show little variation across these comparisons of different genome assembly approaches for a particular species (Fig 5). This finding also confirms the utility of analyses that have quantified SSR density and diversity using raw reads (Castoe et al., 2012), and indicates that read assembly gives no major advantage for identification and estimation of abundance of SSR loci on a genome-wide scale.

It is well established that avian genomes contain substantially less identifiable repetitive content than other vertebrate genomes, and are relatively depauperate in simple sequence repeats (SSRs) and transposable elements (International Chicken Genome Sequencing Consortium, 2004; Primmer, Raudsepp, Chowdhary, Møller, & Ellegren, 1997). Comparisons of the SSR content of avian and lizard genomes support this, confirming that bird genomes contain substantially less SSR content than does the lizard genome (Fig. 6); this trend was also observed in analogous comparisons to a snake genome sample (Castoe et al., 2012). It has been hypothesized that SSR evolution and turnover has been particularly slow in non-mammalian vertebrates (Shedlock et al., 2007), which is consistent with our findings of highly similar abundances of SSR loci across all bird genomes that we examined (Fig. 6), although this and other studies suggest this may not be the case in squamate reptiles like the *Anolis* lizard (Castoe et al., 2011, 2013).

Given previous evidence that the *Anolis* lizard essentially lacks the genomic GC-isochore structure present in birds and mammals (Fujita et al., 2011), interest in understanding the evolutionary dynamics of GC-isochore structure across vertebrates has increased (Castoe et al., 2013; Fujita et al., 2011; Shaffer et al., 2013; St John et al., 2012). Isochore structure is challenging to study with less than high-quality genome assemblies because it requires relatively long assembled regions of the genome. We therefore tested if reference-guided assemblies might provide a cost-effective alternative to the generation of high-quality genome assemblies for developing genomic resources for analysis of GC-isochore structure. While the sample sizes of windows were too small (20 windows of 320-kb in the Clark's Nutcracker) to confidently estimate variation in GC content at large spatial scales, we were able to estimate GC structure at smaller scales using the reference-guided assemblies. While this approach does not capture the full extent of isochore structure in a genome, we have observed previously that smaller windows still provide insight into GC content variation, especially when compared across vertebrates (Fig. 7A; (Fujita et al., 2011)). We found that variation in GC content at 3kb and 5kb window sizes for the Clark's Nutcracker resembled the structure known for other bird genomes (Fig. 7A). More interestingly, based on our sampling experiment, the Clark's Nutcracker assembly may be complete enough to capture the GC heterogeneity at these smaller spatial scales (Fig. 7B). This finding suggests that low (and therefore less-expensive) genome sequencing coverage, combined with a reference guided assembly approach, may hold great promise for economically providing novel insight into genomic GC heterogeneity across a large diversity of vertebrates.

Using reference-guided assemblies, we were able to establish that the relative proportions of certain variant classifications were very similar in both bird species, although the Clark's Nutcracker typically had about twice the number of each variant type (Fig. 8). This corresponds to the approximate genome coverage being about twice as high for the Clark's Nutcracker (Table 1). Thus, low coverage genome assemblies do appear to be useful for analysis of possible shifts in the proportions of certain types of heterozygous variants, and potentially for understanding shifts in genomic mutation spectra among lineages.

Among amniote vertebrates, birds are notable for their high levels of karyotypic conservation (Hansmann et al., 2009; Organ & Edwards, 2011; Takagi & Sasaki, 1974), genomic synteny (Nanda, Schlegelmilch, Haaf, Schartl, & Schmid, 2008; Pokorná et al., 2012), and low repeat element content (Ellegren, 2005; International Chicken Genome Sequencing Consortium, 2004). All these traits make bird genome assembly using *de novo* and reference-guided approaches more tractable, and indicate that among vertebrates, bird genomes may be a best-case scenario for the performance of reference-guided assembly approaches. It would therefore be interesting to investigate the utility of such lower-coverage reference-guided (versus *de novo*) assembly approaches in other lineages, such as mammals or non-avian reptiles. These lineages may have less conserved synteny and higher repeat element content, which implies that the amount of information available from a reference-guided approach may be more limited, and that the approach may only work well for more closely-related reference-target species pairs.

Until recently, only two high-quality and well-annotated bird genomes were available, the Chicken and the Zebra Finch (International Chicken Genome Sequencing Consortium, 2004; Warren et al., 2010), yet additional bird genomes have begun to emerge (Dalloul et al., 2010; Ellegren et al., 2012; Huang et al., 2013; Koren et al., 2012; Oleksyk et al., 2012; Qu et al., 2013; Rands et al., 2013; Shapiro et al., 2013; Zhan et al., 2013). Soon there will be approximately 50 additional high quality bird genomes completed as part of a Beijing Genomics – Genome 10K initiative (Erich Jarvis, pers. comm.). With so many diverse high-quality reference genomes available for birds expected in the near future, the reference-guided approach we test here may provide an attractive means of massively increasing knowledge of bird genome diversity with great economy. It is also notable that neither of the two bird species (or members of the same genera) will be included in these new 50 bird genomes, indicating that genome resources developed here will be highly useful and unique for the foreseeable future.

Not surprisingly, low-coverage reference-guided genome assemblies contain far less information than high-quality *de novo* assembled genomes. What is surprising is that such low-coverage reference-guided assemblies may yield substantial information about the genome of a species compared to a *de novo* assembly using the same data. Thus, approaches using low-coverage reference-guided assemblies, as well as other sample-sequencing approaches that sample <1x genome coverage (Castoe et al., 2011, 2012; Pagán et al., 2012; Sun et al., 2012) hold strong potential to contribute novel insight into vertebrate genomic diversity decades before it is feasible to obtain high-quality genomes from a large number of vertebrates. Such approaches may also be useful for initial surveys of genomic diversity across the tree of life, thereby guiding larger-scale, high-quality genome sampling of particular species that show genomic characteristics and features that are biologically interesting based on such preliminary studies.

ACKNOWLEDGMENTS

We thank the Korf lab, specifically Keith Bradnam, at University of California at Davis for running the CEGMA pipeline, and Kathryn Hall for computational assistance. The manuscript was improved by comments from anonymous reviewers. Any use of trade, product, or firm names in this publication is for descriptive purposes only and does not imply endorsement by the U.S. Government.

FIGURES

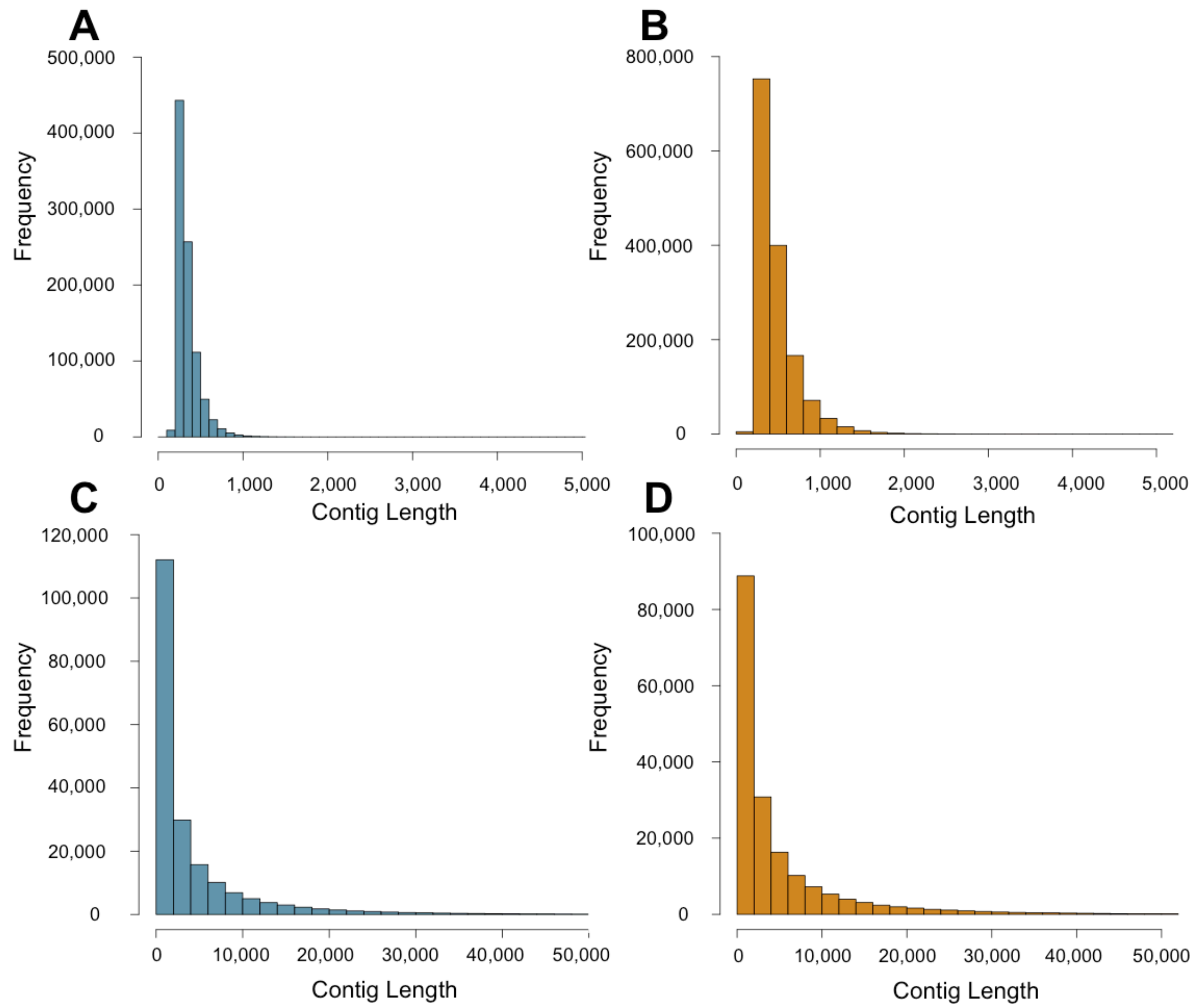


Figure 1. Genomic contig sizes based on various assembly strategies. Frequency histograms of contig sizes for (**A**) the Sage-Grouse *de novo* assembly, (**B**) the Clark's Nutcracker *de novo* assembly, (**C**) the Sage-Grouse reference-guided assembly (1x read coverage) split at $(N)_{500}$ motifs, and (**D**) the Clark's Nutcracker reference-guided assembly (1x read coverage) split at $(N)_{500}$.

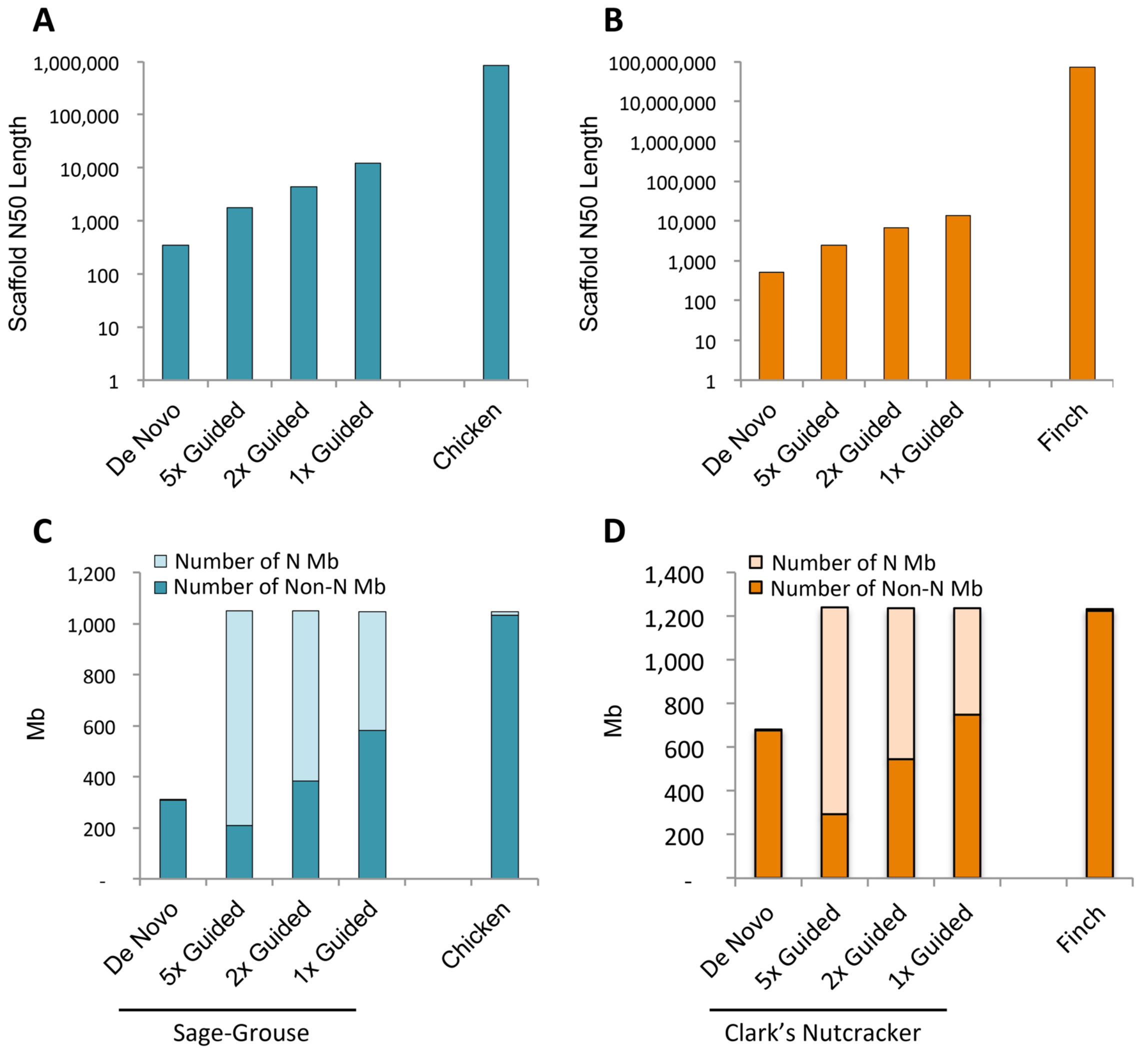


Figure 2. Comparison of N50 scaffold length and total assembly length for various assemblies. Histograms of the N50 scaffold length for new bird genomes with $(N)_{500}$ motifs removed and total genome sizes for guide genomes. (**A**) N50 contig length for the Chicken reference genome, the *de novo* Sage-Grouse genome, and each of the guided assembly genomes. (**B**) N50 scaffold length for the Zebra Finch reference genome, the *de novo* Clark's Nutcracker genome, and each of the Clark's Nutcracker guided assembly genomes. Note that the y-axis scales differ between panels A and B. (**C**) Total genome sizes for the Chicken reference genome, *de novo* Sage-Grouse, and three guided Sage-Grouse genomes at different read coverage levels. (**D**) Total genome sizes for the Zebra Finch reference, *de novo* Clark's Nutcracker, and three guided Clark's Nutcracker genomes at different read coverage levels.

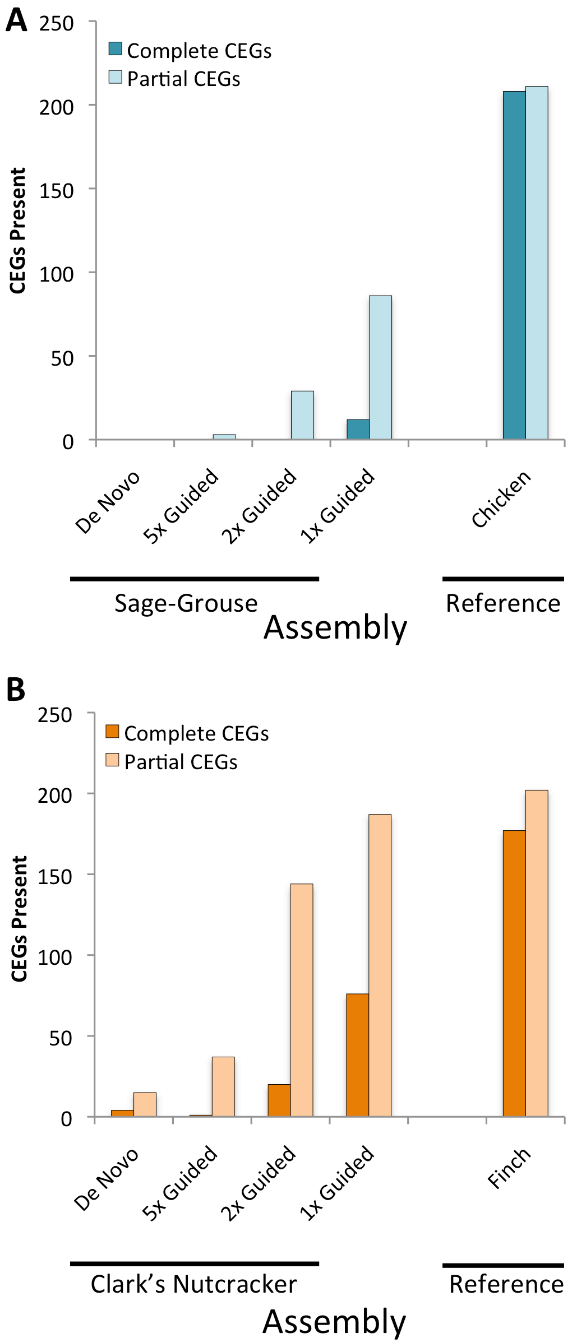


Figure 3. Comparison of Core Eukaryotic Genes identified in various new and reference genome assemblies. Histogram of the number of complete and partial ultraconserved CEGs obtained from the CEGMA pipeline. Maximum number of CEGs is 248. (**A**) The *de novo* assembly and three guided genome assemblies for the Sage-Grouse at different read depth thresholds, plus the guide genome the Chicken. (**B**) The *de novo* assembly and three guided genome assemblies for the Clark's Nutcracker at different read depth thresholds, plus the guide genome the Zebra Finch.

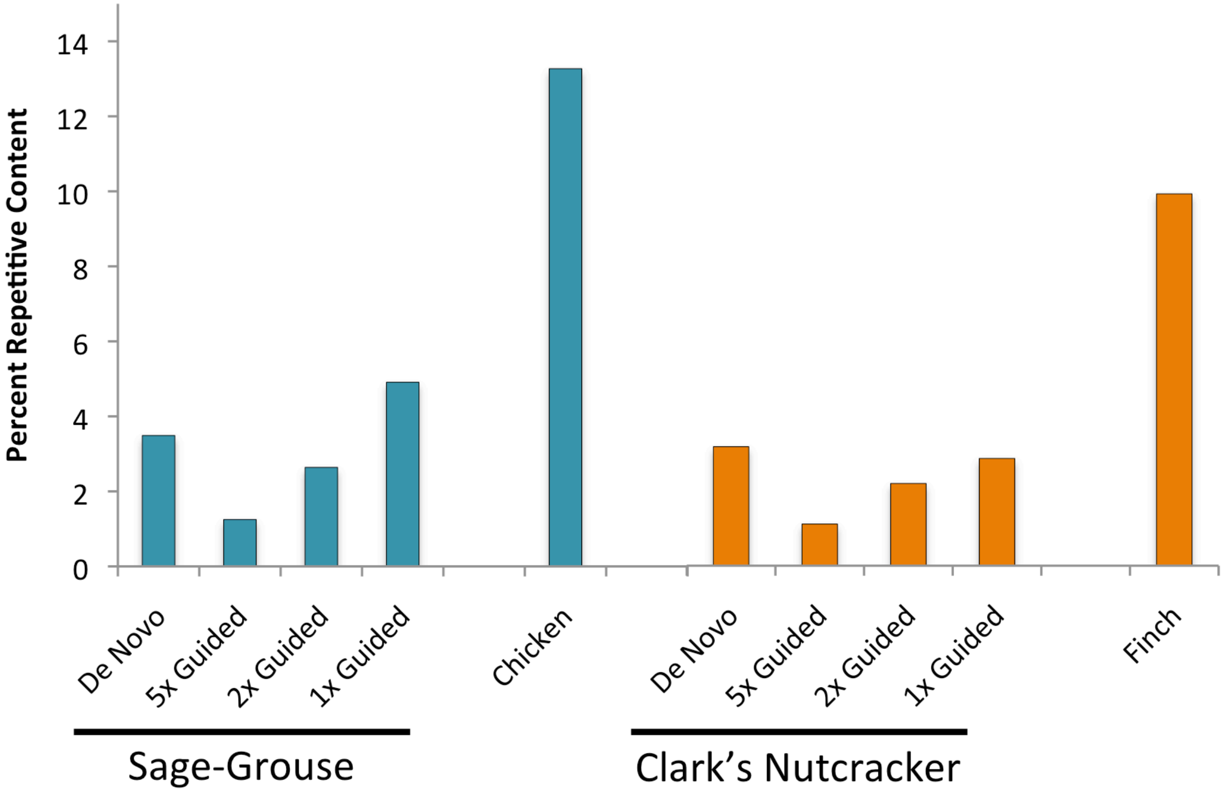


Figure 4. Percent of the genome identified as repetitive elements by RepeatMasker. Histograms of percent repetitive content for all assemblies and the reference genomes of both species. Repetitive content was estimated using RepeatMasker.

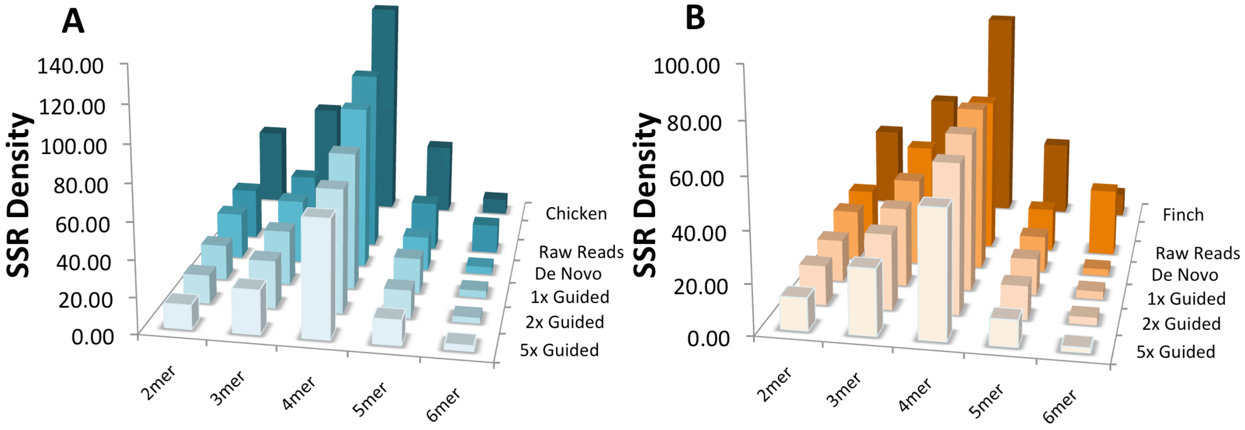


Figure 5. Genomic simple sequence repeat (SSR) density in raw reads and various genome assemblies. Histograms of the simple sequence repeat (SSR) density of sequence is given for raw sequence reads, each of the assembly genomes, and reference genomes for (**A**) the Sage-Grouse and (**B**) the Clark's Nutcracker. Density for each motif length is the number of motif loci per Mb.

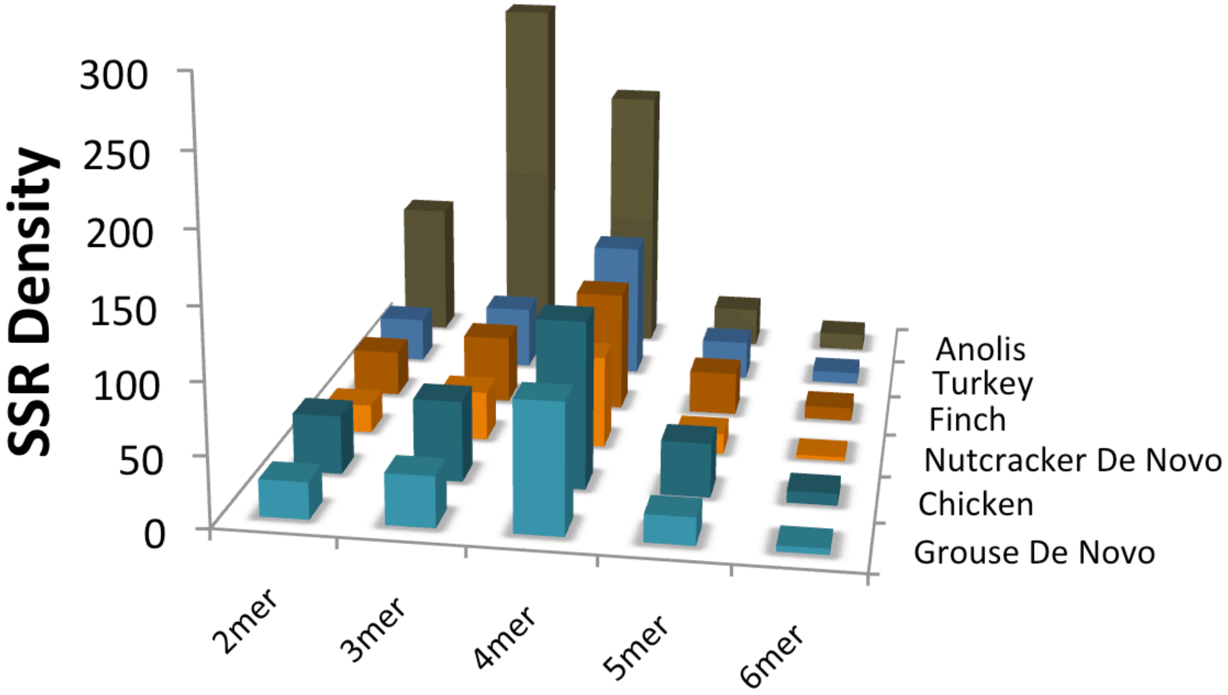


Figure 6. Genomic simple sequence repeat (SSR) density across select amniote vertebreate genomes. Histograms of SSR density for each *de novo* assembly and its respective reference genome, and for the Turkey (*Meleagris gallopavo*) and the Anolis Lizard (*Anolis carolinensis*) genome assemblies. Density for each motif length is the number of motif loci per Mb.

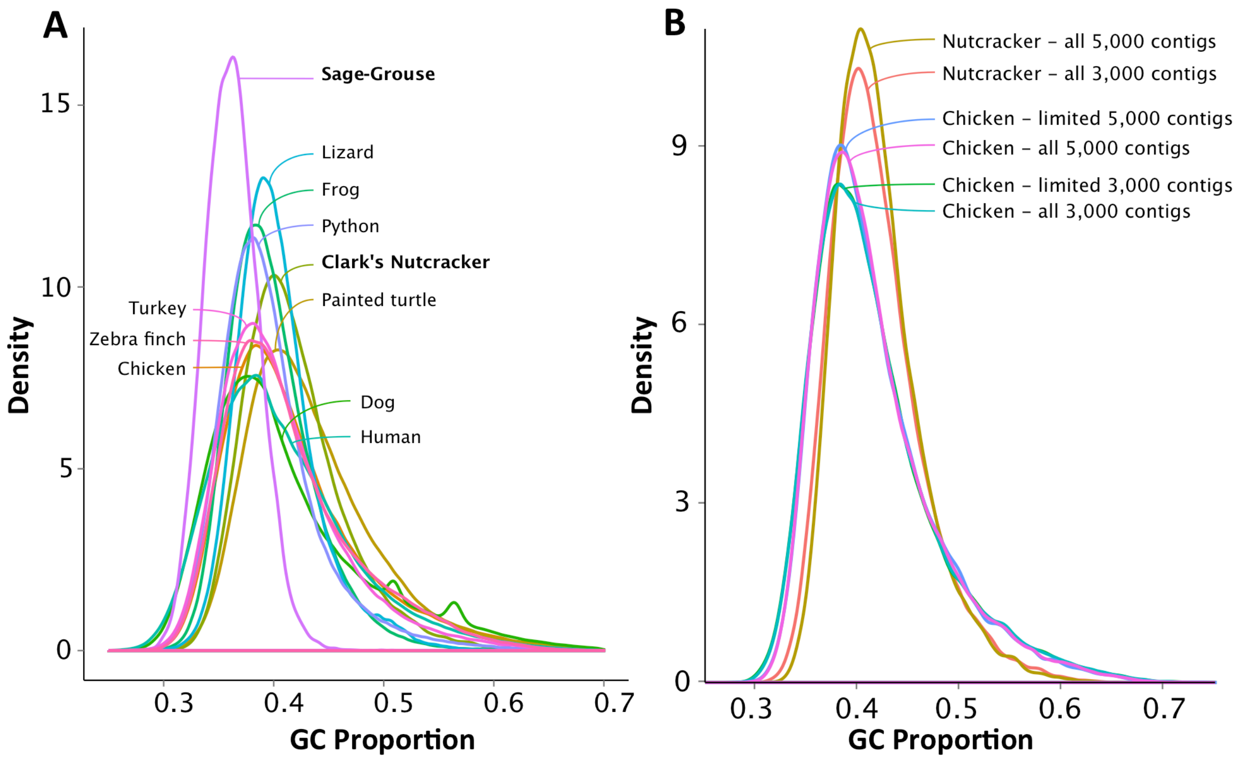


Figure 7. Genomic GC isochore structure among amniote vertebrates, and in draft genomes. (**A**) GC isochore structure plot of 1x guided assemblies for both bird species, their reference genomes, and other select amniote vertebrate genomes using a 3 kb window size. (**B**) GC isochore structure plot comparison of 1x the Clark's Nutcracker guided assembly and the reference the Chicken genome. All contigs at both a 3,000 and a 5,000 bp window were used for the Clark's Nutcracker (n = 73,158 and n = 30,090 contigs respectively). All contigs (referred to as "all" in figure) or a random selection equal to the number of contigs in the Clark's Nutcracker assembly ("limited") for both the 3,000 and 5,000 bp window were used in the comparison.

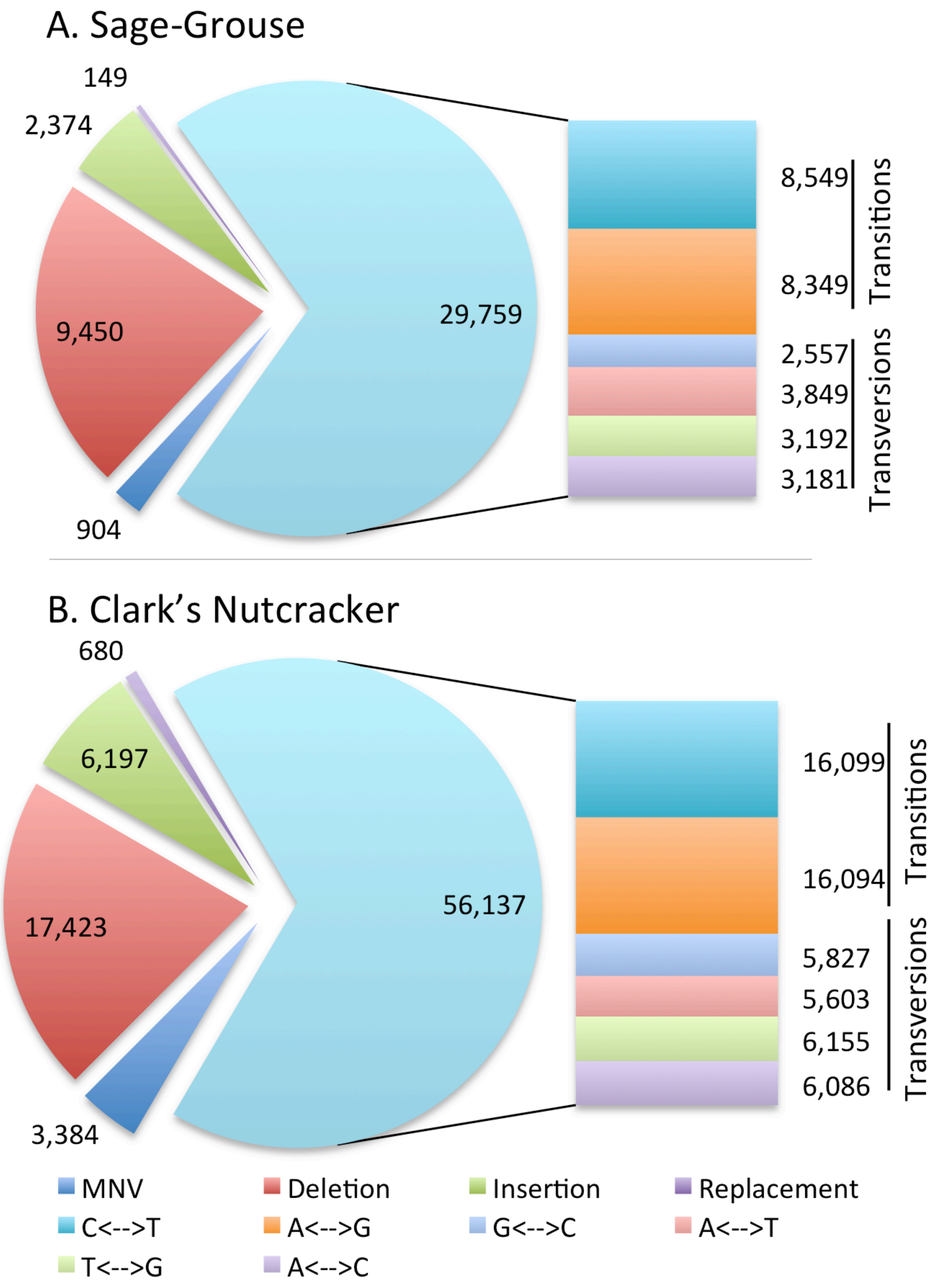


Figure 8. Heterozygous variant composition for the Sage-Grouse and the Clark's Nutcracker. Pie chart includes Single Nucleotide Variants (SNV), Multiple Nucleotide Variants (MNV), Insertions, Deletions, and Replacements. SNVs are further annotated in a bar graph form according to all possible transitions. Key provides color-coding for each variant.

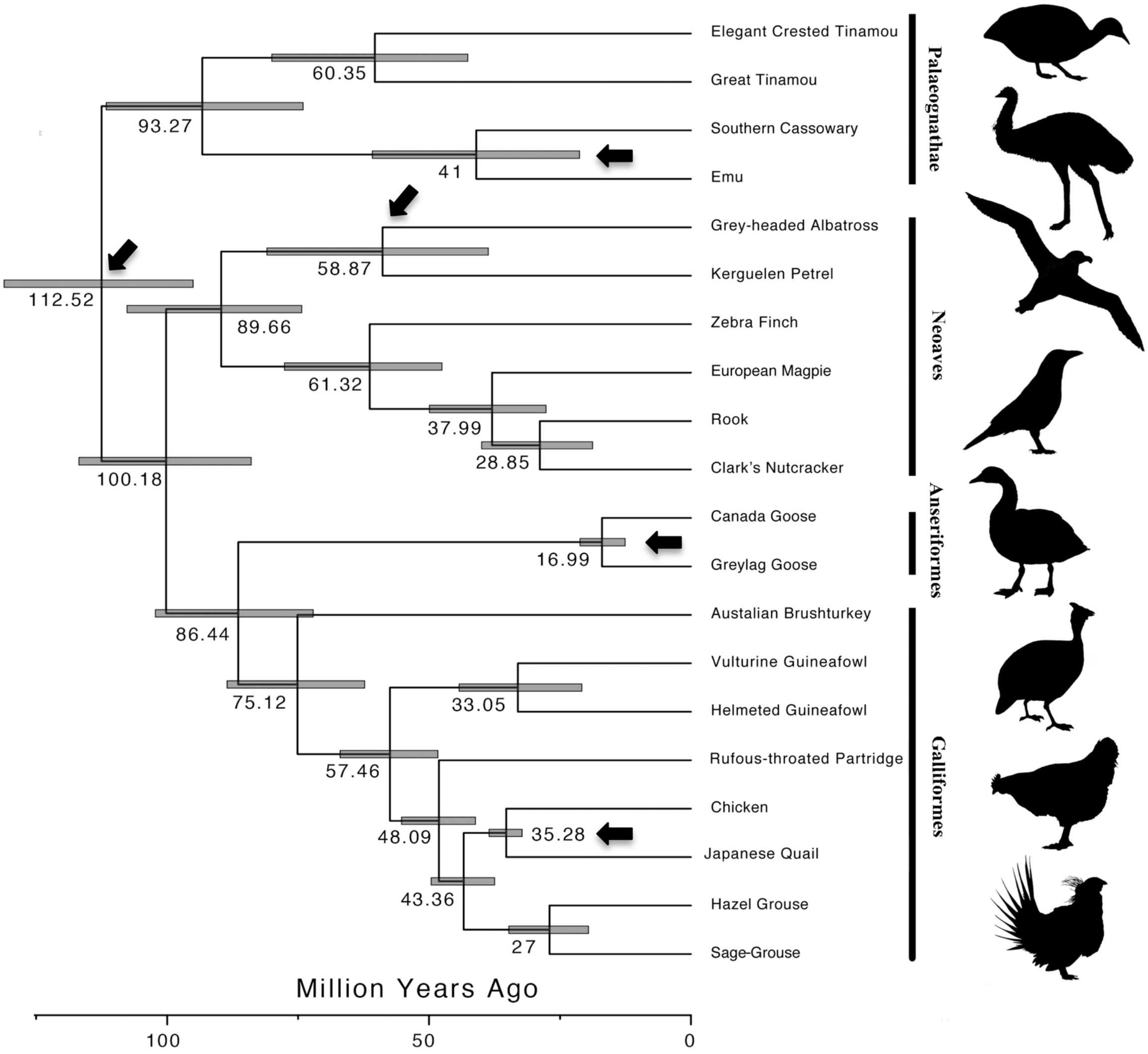


Figure 9. Estimated divergence times among birds, including focal and reference genome species. Bayesian relaxed clock estimate of divergence times among several bird lineages based on 12 mitochondrial protein-coding genes, with 95% credibility intervals shown as shaded bars at nodes. Dark arrows represent calibration points used in the analysis.

TABLES

Table 1. Summary of raw genome sequence data used.

Species	Reads	Total Bp	Estimated genome size (Gb)	Estimated fold coverage
Sage-Grouse	39,582,844	4,662,514,211	1.32	3.53
Clark’s Nutcracker	60,573,448	7,135,441,227	1.32	5.41

Table 2. Summary of genome assembly statistics from various assembly approaches.

	Sage-Grouse				Clark's Nutcracker			
		Reference-guided				Reference-guided		
	De novo	Coverage >1x	Coverage >2x	Coverage >5x	*De novo*	Coverage >1x	Coverage >2x	Coverage >5x
% N Bases	0.04	44.35	63.36	79.99	0.09	39.55	55.86	76.34
N50 - No Break	343	90,198,103	90,394,695	90,527,046	503	65,905,513	73,959,172	74,132,310
N50 - Break 500	--	12,125	4,447	1,804	--	13,369	6,765	2,409
Complete CEGs	0	12	0	0	4	76	20	1

The terms 'no break' and 'break 500' refer to whether or not contigs were broken up by deleting regions that contained stretches of 500 or more ambiguous ("N") nucleotides, and CEGs refer to core eukaryotic genes.

SUPPLEMENTARY MATERIAL

Supplementary Table 1. Species and NCBI accessions used to guide the Sage-Grouse mitochondrial genome reconstruction.

Species	NCBI Accession Number
Acryllium vulturinum	NC014180
Alectoris chukar	FJ752426
Alectura lathami	NC007227
Arborophila gingica	FJ752425
Arborophila rufipectus	FJ194942
Arborophila rufogularis	NC020584
Bambusicola fytchii	FJ752423
Bambusicola thoracica	EU165706
Coturnix chinensis	AB073301
Coturnix japonica	AP003195
Crossoptilon auritum	JF937589
Crossoptilon crossoptilon	HQ891119
Francolinus pintadeanus	EU165707
Gallus gallus	NC001323
Gallus lafayettei	AP003325
Gallus sonneratii	AP006741
Gallus varius	AP003324
Ithaginis cruentus	JF921875
Lophophorus lhuysii	GQ871234
Lophophorus sclateri	FJ752432
Lophura ignita	AB164627
Lophura nycthemera	EU417810
Meleagris gallopavo	EF153719
Numida meleagris	NC006382
Pavo muticus	EU417811
Perdix dauurica	FJ752431
Phasianus colchicus	FJ752430
Phasianus versicolor	AB164626
Polyplectron bicalcaratum	EU417812
Pucrasia macrolopha	FJ752429
Syrmaticus ellioti	AB164624
Syrmaticus humiae	AB164625
Syrmaticus reevesii	AB164623
Syrmaticus soemmerringi	AB164622
Tetraophasis obscurus	JF921876
Tetraophasis szechenyii	FJ752428
Tetrastes bonasia	NC020591
Tragopan temminckii	FJ752427

Supplementary Table 2. Species and NCBI accessions used to guide the Clark's Nutcracker mitochondrial genome reconstruction.

Species	NCBI Accession Number
Lanius tephronotus	JX486029
Cyanopica cyanus	JN108020
Corvus frugilegus	NC002069
Urocissa erythrorhyncha	JQ423932
Podoces hendersoni	GU592504
Pica pica	HQ915867
Oriolus chinensis	JQ083495

Supplementary Table 3. Species and NCBI accessions used for phylogeny and divergence estimation.

Species	NCBI Accession Number
Corvus frugilegus	NC002069
Dromaius novaehollandiae	NC002784
Tinamus major	NC002781
Eudromia elegans	NC002772
Casuarius casuarius	NC002778
Branta canadensis	NC007011
Pterodroma brevirostris	NC007174
Alectura lathami	NC007227
Diomedea chrysostoma	AP009193
Anser anser	NC011196
Pica pica	HQ915867
Coturnix japonica	AP003195
Numida meleagris	NC006382
Acryllium vulturinum	NC014180
Arborophila rufogularis	NC020584
Tetrastes bonasia	NC020591
Gallus gallus	NC001323
Taeniopygia guttata	NC007897

Supplementary Table 4. Best-fit models of nucleotide evolution for mitochondrial genes used in phylogenetic analyses.

Gene	Codon Position	Model	Gene	Codon Position	Model
ATP6	1st	GTR+Γ	**ND1**	1st	JC
ATP6	2nd	GTR+Γ	**ND1**	2nd	JC
ATP6	3rd	HKY+I+Γ	**ND1**	3rd	HKY
ATP8	1st	HKY+Γ	**ND2**	1st	HKY+Γ
ATP8	2nd	HKY+Γ	**ND2**	2nd	HKY+Γ
ATP8	3rd	HKY+Γ	**ND2**	3rd	HKY+Γ
CO1	1st	HKY+I+Γ	**ND3**	1st	HKY+Γ
CO1	2nd	HKY+I+Γ	**ND3**	2nd	HKY+Γ
CO1	3rd	HKY+Γ	**ND3**	3rd	HKY+Γ
CO2	1st	HKY+Γ	**ND4**	1st	HKY+Γ
CO2	2nd	HKY+Γ	**ND4**	2nd	HKY+Γ
CO2	3rd	GTR+I+Γ	**ND4**	3rd	GTR+Γ
CO3	1st	HKY+Γ	**ND4L**	1st	HKY+Γ
CO3	2nd	HKY+Γ	**ND4L**	2nd	HKY+Γ
CO3	3rd	HKY+Γ	**ND4L**	3rd	HKY+I+Γ
CytB	1st	HKY+I+Γ	**ND5**	1st	HKY+Γ
CytB	2nd	HKY+I+Γ	**ND5**	2nd	HKY+Γ
CytB	3rd	HKY+Γ	**ND5**	3rd	HKY+I+Γ

Supplementary Table 5. Calibration points used in the divergence time analysis.

	Distribution	Mean (mya)	StDev (mya)
Anser-Branta	Normal	14.5	2.7
Archosauria	Normal	243	3.6
Aves	Normal	93.5	17
Coturnix-Gallus	Normal	35	1.7
Neoaves	Normal	91.9	7.8

Supplementary Table 6. Percent GC in new (and reference) genome assemblies.

	Assembly	Percent GC Content
Sage-Grouse	*De novo*	37.5
	1x Guided	38.9
	2x Guided	38.6
	5x Guided	38.0
	Chicken	41.8
Clark’ s Nutcracker	*De novo*	41.4
	1x Guided	41.4
	2x Guided	41.6
	5x Guided	42.0
	Zebra Finch	41.4

CITATIONS

Alföldi, J., Palma, F. D., Grabherr, M., Williams, C., Kong, L., Mauceli, E., … Lindblad-Toh, K. (2011). The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, *477*(7366), 587. https://doi.org/10.1038/nature10390

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ansari, H. A., Takagi, N., & Sasaki, M. (1988). Morphological differentiation of sex chromosomes in three species of ratite birds. *Cytogenetic and Genome Research*, *47*(4), 185–188. https://doi.org/10.1159/000132545

Barker, F. K., Oyler-McCance, S., & Tomback, D. F. (2013). Blood from a turnip: tissue origin of low-coverage shotgun sequencing libraries affects recovery of mitogenome sequences. *Mitochondrial DNA*, *26*(3), 384–388. https://doi.org/10.3109/19401736.2013.840588

Barringer, L. E., Tomback, D. F., Wunder, M. B., & McKinney, S. T. (2012). Whitebark Pine Stand Condition, Tree Abundance, and Cone Production as Predictors of Visitation by Clark’s Nutcracker. *PLOS ONE*, *7*(5), e37663. https://doi.org/10.1371/journal.pone.0037663

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., … Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology*, *10*(4), e1003537. https://doi.org/10.1371/journal.pcbi.1003537

Card, D. C., Schield, D. R., Reyes-Velasco, J., Fujita, M. K., Andrew, A. L., Oyler-McCance, S. J., … Castoe, T. A. (2014). Data from: Two low coverage bird genomes and a comparison of reference-guided versus de novo genome assemblies. *Dryad*. https://doi.org/10.5061/dryad.qn1n2

Castoe, T. A., Hall, K. T., Mboulas, G., L, M., Gu, W., Koning, D., … Pollock, D. D. (2011). Discovery of Highly Divergent Repeat Landscapes in Snake Genomes Using High-Throughput Sequencing. *Genome Biology and Evolution*, *3*, 641–653. https://doi.org/10.1093/gbe/evr043

Castoe, T. A., Koning, A. P. J. de, Hall, K. T., Card, D. C., Schield, D. R., Fujita, M. K., … Pollock, D. D. (2013). The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proceedings of the National Academy of Sciences*, *110*(51), 20645–20650. https://doi.org/10.1073/pnas.1314475110

Castoe, T. A., Poole, A. W., Koning, A. P. J. de, Jones, K. L., Tomback, D. F., Oyler-McCance, S. J., … Pollock, D. D. (2012). Rapid Microsatellite Identification from Illumina Paired-End Genomic Sequencing in Two Birds and a Snake. *PLOS ONE*, *7*(2), e30953. https://doi.org/10.1371/journal.pone.0030953

Dalloul, R. A., Long, J. A., Zimin, A. V., Aslam, L., Beal, K., Blomberg, L. A., … Reed, K. M. (2010). Multi-Platform Next-Generation Sequencing of the Domestic Turkey (*Meleagris gallopavo*): Genome Assembly and Analysis. *PLOS Biology*, *8*(9), e1000475. https://doi.org/10.1371/journal.pbio.1000475

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*, 214. https://doi.org/10.1186/1471-2148-7-214

Ellegren, H. (2005). The avian genome uncovered. *Trends in Ecology & Evolution*, *20*(4), 180–186. https://doi.org/10.1016/j.tree.2005.01.015

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., … Wolf, J. B. W. (2012). The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, *491*(7426), 756–760. https://doi.org/10.1038/nature11584

Ericson, P. G. P., Jansén, A., Johansson, U. S., & Ekman, J. (2005). Inter-generic relationships of the crows, jays, magpies and allied groups (Aves: Corvidae) based on nucleotide sequence data. *Journal of Avian Biology*, *36*(3), 222–234. https://doi.org/10.1111/j.0908-8857.2001.03409.x

Fujita, M. K., Edwards, S. V., & Ponting, C. P. (2011). The Anolis Lizard Genome: An Amniote Genome without Isochores. *Genome Biology and Evolution*, *3*, 974–984. https://doi.org/10.1093/gbe/evr072

Gregory, T. R. (2013). *Animal Genome Size Database*. Retrieved from http://www.genomesize.com

Gregory, T. Ryan, Nicol, J. A., Tamm, H., Kullman, B., Kullman, K., Leitch, I. J., … Bennett, M. D. (2007). Eukaryotic genome size databases. *Nucleic Acids Research*, *35*(suppl_1), D332–D338. https://doi.org/10.1093/nar/gkl828

Gunnison Sage-Grouse Rangewide Steering Committee. (2005). *Gunnison Sage-grouse Rangewide Conservation Plan* (p. 359). Denver, CO, USA: Colorado Division of Wildlife. Retrieved from http://cpw.state.co.us/learn/Pages/GunnisonSagegrouseConservationPlan.aspx

Hansmann, T., Nanda, I., Volobouev, V., Yang, F., Schartl, M., Haaf, T., & Schmid, M. (2009). Cross-Species Chromosome Painting Corroborates Microchromosome Fusion during Karyotype Evolution of Birds. *Cytogenetic and Genome Research*, *126*(3), 281–304. https://doi.org/10.1159/000251965

Huang, Y., Li, Y., Burt, D. W., Chen, H., Zhang, Y., Qian, W., … Li, N. (2013). The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nature Genetics*, *45*(7), 776–783. https://doi.org/10.1038/ng.2657

International Chicken Genome Sequencing Consortium. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, *432*(7018), 695. https://doi.org/10.1038/nature03154

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, *110*(1–4), 462–467. https://doi.org/10.1159/000084979

Kan, X.-Z., Li, X.-F., Lei, Z.-P., Chen, L., Gao, H., Yang, Z.-Y., … Qian, C.-J. (2010). Estimation of divergence times for major lineages of galliform birds: Evidence from complete mitochondrial genome sequences. *African Journal of Biotechnology*, *9*(21), 3073–3078.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., … Phillippy, A. M. (2012). Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotechnology*, *30*(7), 693–700. https://doi.org/10.1038/nbt.2280

Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. *Molecular Biology and Evolution*, *29*(6), 1695–1701. https://doi.org/10.1093/molbev/mss020

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., … Wang, J. (2010). The sequence and *de novo* assembly of the giant panda genome. *Nature*, *463*(7279), 311–317. https://doi.org/10.1038/nature08696

Mellmann, A., Harmsen, D., Cummings, C. A., Zentz, E. B., Leopold, S. R., Rico, A., … Karch, H. (2011). Prospective Genomic Characterization of the German Enterohemorrhagic *Escherichia coli* O104:H4 Outbreak by Rapid Next Generation Sequencing Technology. *PLOS ONE*, *6*(7), e22751. https://doi.org/10.1371/journal.pone.0022751

Nanda, I., Schlegelmilch, K., Haaf, T., Schartl, M., & Schmid, M. (2008). Synteny conservation of the Z chromosome in 14 avian species (11 families) supports a role for Z dosage in avian sex determination. *Cytogenetic and Genome Research*, *122*(2), 150–156. https://doi.org/10.1159/000163092

Nishito, Y., Osana, Y., Hachiya, T., Popendorf, K., Toyoda, A., Fujiyama, A., … Sakakibara, Y. (2010). Whole genome assembly of a natto production strain *Bacillus subtilis* natto from very short read data. *BMC Genomics*, *11*, 243. https://doi.org/10.1186/1471-2164-11-243

Ogawa, A., Murata, K., & Mizuno, S. (1998). The location of Z- and W-linked marker genes and sequence on the homomorphic sex chromosomes of the ostrich and the emu. *Proceedings of the National Academy of Sciences*, *95*(8), 4415–4418.

Oleksyk, T. K., Pombert, J.-F., Siu, D., Mazo-Vargas, A., Ramos, B., Guiblet, W., … Martinez-Cruzado, J.-C. (2012). A locally funded Puerto Rican parrot (*Amazona vittata*) genome sequencing project increases avian data and advances young researcher education. *GigaScience*, *1*, 14. https://doi.org/10.1186/2047-217X-1-14

Organ, C. L., & Edwards, S. V. (2011). Major Events in Avian Genome Evolution. In G. Dyke & G. Kaiser (Eds.), *Living Dinosaurs: The Evolutionary History of Modern Birds* (pp. 325–337). John Wiley & Sons Ltd. Retrieved from https://onlinelibrary.wiley.com/doi/10.1002/9781119990475.ch13

Oyler-McCance, S. J., St. John, J., Taylor, S. E., Apa, A. D., & Quinn, T. W. (2005). Population Genetics of Gunnison Sage-Grouse: Implications for Management. *The Journal of Wildlife Management*, *69*(2), 630–637.

Pagán, H. J. T., Macas, J., Novák, P., McCulloch, E. S., Stevens, R. D., & Ray, D. A. (2012). Survey Sequencing Reveals Elevated DNA Transposon Activity, Novel Elements, and Variation in Repetitive Landscapes among Vesper Bats. *Genome Biology and Evolution*, *4*(4), 575–585. https://doi.org/10.1093/gbe/evs038

Parchman, T. L., Geist, K. S., Grahnen, J. A., Benkman, C. W., & Buerkle, C. A. (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics*, *11*, 180. https://doi.org/10.1186/1471-2164-11-180

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, *23*(9), 1061–1067. https://doi.org/10.1093/bioinformatics/btm071

Pereira, S. L., & Baker, A. J. (2006). A Mitogenomic Timescale for Birds Detects Variable Phylogenetic Rates of Molecular Evolution and Refutes the Standard Molecular Clock. *Molecular Biology and Evolution*, *23*(9), 1731–1740. https://doi.org/10.1093/molbev/msl038

Phillips, M. J., Gibb, G. C., Crimp, E. A., & Penny, D. (2010). Tinamous and Moa Flock Together: Mitochondrial Genome Sequence Analysis Reveals Independent Losses of Flight among Ratites. *Systematic Biology*, *59*(1), 90–107. https://doi.org/10.1093/sysbio/syp079

Pokorná, M., Giovannotti, M., Kratochvíl, L., Caputo, V., Olmo, E., Ferguson-Smith, M. A., & Rens, W. (2012). Conservation of chromosomes syntenic with avian autosomes in squamate reptiles revealed by comparative chromosome painting. *Chromosoma*, *121*(4), 409–418. https://doi.org/10.1007/s00412-012-0371-z

Primmer, C. R., Raudsepp, T., Chowdhary, B. P., Møller, A. P., & Ellegren, H. (1997). Low Frequency of Microsatellites in the Avian Genome. *Genome Research*, *7*(5), 471–482. https://doi.org/10.1101/gr.7.5.471

Qu, Y., Zhao, H., Han, N., Zhou, G., Song, G., Gao, B., … Lei, F. (2013). Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau. *Nature Communications*, *4*, 2071. https://doi.org/10.1038/ncomms3071

Rands, C. M., Darling, A., Fujita, M., Kong, L., Webster, M. T., Clabaut, C., … Ponting, C. P. (2013). Insights into the evolution of Darwin's finches from comparative analysis of the *Geospiza magnirostris* genome sequence. *BMC Genomics*, *14*, 95. https://doi.org/10.1186/1471-2164-14-95

Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), 1572–1574. https://doi.org/10.1093/bioinformatics/btg180

Schneeberger, K., Ossowski, S., Ott, F., Klein, J. D., Wang, X., Lanz, C., … Weigel, D. (2011). Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proceedings of the National Academy of Sciences*, *108*(25), 10249–10254. https://doi.org/10.1073/pnas.1107739108

Schoettle, A. W., & Sniezko, R. A. (2007). Proactive intervention to sustain high-elevation pine ecosystems threatened by white pine blister rust. *Journal of Forest Research*, *12*(5), 327–336. https://doi.org/10.1007/s10310-007-0024-x

Shaffer, B., Minx, P., Warren, D. E., Shedlock, A. M., Thomson, R. C., Valenzuela, N., … Wilson, R. K. (2013). The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biology*, *14*, R28. https://doi.org/10.1186/gb-2013-14-3-r28

Shapiro, M. D., Kronenberg, Z., Li, C., Domyan, E. T., Pan, H., Campbell, M., … Wang, J. (2013). Genomic Diversity and Evolution of the Head Crest in the Rock Pigeon. *Science*, *339*(6123), 1063–1067. https://doi.org/10.1126/science.1230422

Shedlock, A. M., Botka, C. W., Zhao, S., Shetty, J., Zhang, T., Liu, J. S., … Edwards, S. V. (2007). Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proceedings of the National Academy of Sciences*, *104*(8), 2767–2772. https://doi.org/10.1073/pnas.0606204104

Shetty, S., Kirby, P., Zarkower, D., & Graves, J. a. M. (2002). DMRT1 in a ratite bird: evidence for a role in sex determination and discovery of a putative regulatory element. *Cytogenetic and Genome Research*, *99*(1–4), 245–251. https://doi.org/10.1159/000071600

Shetty, Swathi, Griffin, D. K., & Graves, J. A. M. (1999). Comparative Painting Reveals Strong Chromosome Homology Over 80 Million Years of Bird Evolution. *Chromosome Research*, *7*(4), 289–295. https://doi.org/10.1023/A:1009278914829

Smit, A. F. A., Hubley, R., & Green, P. (2013). *RepeatMasker Open-4.0*. Retrieved from http://repeatmasker.org/

St John, J. A., Braun, E. L., Isberg, S. R., Miles, L. G., Chong, A. Y., Gongora, J., … Ray, D. A. (2012). Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biology*, *13*, 415. https://doi.org/10.1186/gb-2012-13-1-415

Stiver, J. R., Apa, A. D., Remington, T. E., & Gibson, R. M. (2008). Polygyny and female breeding failure reduce effective population size in the lekking Gunnison sage-grouse. *Biological Conservation*, *141*(2), 472–481. https://doi.org/10.1016/j.biocon.2007.10.018

Sun, C., Shepard, D. B., Chong, R. A., López Arriaza, J., Hall, K., Castoe, T. A., … Mueller, R. L. (2012). LTR Retrotransposons Contribute to Genomic Gigantism in Plethodontid Salamanders. *Genome Biology and Evolution*, *4*(2), 168–183. https://doi.org/10.1093/gbe/evr139

Takagi, N., & Sasaki, M. (1974). A phylogenetic study of bird karyotypes. *Chromosoma*, *46*(1), 91–120. https://doi.org/10.1007/BF00332341

Tomback, D. F. (1982). Dispersal of Whitebark Pine Seeds by Clark's Nutcracker: A Mutualism Hypothesis. *Journal of Animal Ecology*, *51*(2), 451–467. https://doi.org/10.2307/3976

Tomback D. F., & Achuff P. (2010). Blister rust and western forest biodiversity: ecology, values and outlook for white pines. *Forest Pathology*, *40*(3-4), 186–225. https://doi.org/10.1111/j.1439-0329.2010.00655.x

Tomback, D. F., Arno, S. F., & Keane, R. E. (2001). The compelling case for management and intervention. In D. F. Tomback, S. F. Arno, & R. E. Keane (Eds.), *Whitebark Pine Communities: Ecology And Restoration* (pp. 3–25). Island Press.

Vaidya, G., Lohman, D. J., & Meier, R. (2011). SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics*, *27*(2), 171–180. https://doi.org/10.1111/j.1096-0031.2010.00329.x

Vicoso, B., Kaiser, V. B., & Bachtrog, D. (2013). Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. *Proceedings of the National Academy of Sciences*, *110*(16), 6453–6458. https://doi.org/10.1073/pnas.1217027110

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., … Wilson, R. K. (2010). The genome of a songbird. *Nature*, *464*(7289), 757–762. https://doi.org/10.1038/nature08819

Zhan, X., Pan, S., Wang, J., Dixon, A., He, J., Muller, M. G., … Bruford, M. W. (2013). Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nature Genetics*, *45*(5), 563–566. https://doi.org/10.1038/ng.2588

Chapter 3.

A high-quality annotation for the *Boa constrictor* reference genome

Daren C. Card[1], Giulia I. M. Pasquesi[1], Blair W. Perry[1], and Todd A. Castoe[1,*]

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

ABSTRACT

Boa constrictor and closely related *Boa* species represent a widespread group of snakes found across diverse habitats in North, Central, and South America. These typically large, heavily-bodied snakes possess several interesting natural history characteristics that make them valuable model systems for a broad spectrum of biological questions. Although a well-assembled genome sequence is available for this group, the utility of this genome assembly is currently limited by the lack of any annotation. We created a *de novo*, *Boa*-specific repeat library and combined this resource with existing tetrapod and snake repeat libraries to annotate genomic repeat element content for the *Boa constrictor* genome. Our repeat annotation demonstrates that approximately 32% of the *Boa* genome is composed of identifiable repeat elements, and analyses of the timing of transposable element family expansion indicates three distinct temporal periods of element proliferation. We generated RNAseq data from 10 tissue types and used these data to produce a transcriptome assembly, which we combined with existing protein models from other squamate reptiles to produce a well-supported protein annotation comprised of 19,178 genes. We inferred protein identity for approximately 97% of these genes using several databases and identified 7,398 one-to-one orthologs shared between the *Boa* genome and genomes of four other squamate reptiles. Our comprehensive repeat and gene annotation greatly expands the utility of the *Boa constrictor* reference genome, which now represents the highest quality and most contiguous snake reference genome available.

INTRODUCTION

Highly contiguous genome assemblies are inherently valuable for a broad spectrum of research questions, yet most of the utility of genomes derives from the annotation of genes and repetitive elements. The *Boa constrictor* genome assembly was created as part of the Assemblathon2 genome assembly competition and is currently the best assembled (most contiguous) snake genome (Bradnam et al., 2013). However, this reference genome lacks any annotation of genes or repetitive elements, which has limited the utility of this resource, particularly for biologically-driven research questions. For example, while the Assemblathon2 paper has been frequently cited as an example of genome assembly practices, citation metrics indicate that few citations are from research groups attempting to use this genome as a resource for investigating biologically motivated research. This limitation is unfortunate because Boas, and snakes in general, represent increasingly important model systems for investigating a variety of biological questions.

The broadly ranging genus *Boa* includes substantial population diversity and at least three distinct species (Card et al., 2016; Reynolds, Niemiller, & Revell, 2014; Suárez-Atilano, Burbrink, & Vázquez-Domínguez, 2014). Populations have colonized several offshore islands in Central American, where they have become dwarfed in size (Boback, 2005, 2006; Henderson, Waller, Micucci, Puorto, & Bourgeois, 1995). This adaptation to island environments appears to have occurred multiple times (Card et al., 2016), which has led to this species becoming a model for studying the genetic basis of rapid and complex convergent phenotypic evolution. Ecologically, *Boa* species are large snakes that employ a sit-and-wait, infrequently feeding life history strategy that has led boas, and other snakes like pythons, to evolve adaptations to significantly downregulate their metabolism, physiology, and even organ mass while fasting. These snakes then rapidly upregulate these features upon feeding, leading to unparalleled upregulation of metabolism and physiological states, and rapid tissue and organ regeneration upon feeding (Andrade, Toledo, Abe, & Wang, 2004; Secor, Stein, & Diamond, 1994; Secor, 2008; Secor & Diamond, 1995, 1998). Most physiological and genomic research on this interesting and medically relevant phenotype has focused on Burmese pythons (Andrew et al., 2015, 2017; Castoe et al., 2013; Lignot, Helmstetter, & Secor, 2005), but could be extended to *Boa* species in a powerful comparative framework if an annotated reference genome were available. *Boa* also represents an emerging model for studying the evolution of sex chromosomes. Population genomic data for *Boa* recently demonstrated that at least some species of boas and pythons appear to have XY sex determination, overturning the long-held belief that all snake species possessed ZW sex determination (Gamble et al., 2017). Interestingly, comparisons between *Boa* and the Burmese python indicate that sex chromosomes may have evolved independently from different ancestral autosomes (Gamble et al., 2017), and while the genome assembly has been useful for deciphering this phenomenon, future research to investigate sex chromosome evolution and its biological implications in snakes is currently limited by the lack of an annotated *Boa* reference genome. These examples represent some of the many possible research topics that could be assisted by the creation of a high-quality annotation for the *B. constrictor* reference genome. In the following sections we briefly describe the existing *Boa* reference genome composition, describe the creation of a genome annotation for the *Boa*, and demonstrate the quality and utility of this resource through several additional analyses.

MATERIALS & METHODS

Characterizing the existing Boa constrictor genome assembly

A high-quality reference genome has been assembled for *Boa constrictor* as part of the Assemblathon2 project, which focused exclusively on evaluating genome contiguity and quality in competing assemblies generated using different methods (Bradnam et al., 2013). For the purpose of genome annotation, we used a single *B. constrictor* genome assembly ('*snake assembly 7C*' produced by the SGA team) – the assembly that was ranked the highest based on a thorough analysis of 10 assembly metrics (Bradnam et al., 2013). We evaluated assembly quality using BUSCO v. 2.0.1 (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015), which was not available when the genome was originally generated. BUSCO is an informative technique for evaluating genome assembly completeness that searches for evolutionarily-informed sets of highly conserved genes found broadly across particular clades of organisms (Simão et al., 2015), based on OrthoDB (Waterhouse, Tegenfeldt, Li, Zdobnov, & Kriventseva, 2013). BUSCO was run using the genome mode with default parameters and the Tetropoda library of conserved genes derived from 55 tetrapod species (OrthoDB version 9 (Zdobnov et al., 2017)).

To provide an initial characterization of biologically-relevant genomic composition information from the *Boa* genome, we estimated GC content from non-overlapping 50 kb windows (regions of > 25% N gap sequence were excluded), from the full CDS sequences of annotated genes (see below for details on how these annotations were produced), and at third codon positions (i.e., GC3) in annotated genes. The distribution of k-mers produced from large amounts of genomic sequencing coverage is useful for estimating total genome size. To estimate this genomic characteristic we used the equivalent of approximately 40x genome coverage of quality-trimmed sequences described above to produce k-mer counts from 19mers, 23mers, and 27mers using jellyfish v. 2.2.3 (Marçais & Kingsford, 2011). Based on the resulting k-mer count tables, we estimated genome size for each k-mer using GCE v. 1.0.0 (Liu et al., 2013).

We also estimated heterozygosity using approximately 40x genome coverage of short-insert Illumina reads that were used to assemble the genome (see Supplementary Table 2 for information on sequence read files). We quality trimmed reads using Trimmomatic v. 0.33 (Bolger, Lohse, & Usadel, 2014) with the settings: LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:36. Quality-trimmed reads were mapped to the *Boa* genome using the MEM algorithm of BWA v. 0.7.12-r1039 (Li & Durbin, 2009) and default settings. We followed the GATK Best-Practices recommendations (DePristo et al., 2011; Van der Auwera et al., 2013) to quality-control mapped reads and call variants using SAMtools (Li et al., 2009), Picard v. 1.95, and GATK v. 3.8-0-ge9d806836 (McKenna et al., 2010). Briefly, duplicate reads were excluded and regions around InDels were realigned with default settings. Variants were called using HaplotypeCaller and we filtered to exclude SNPs within 3 bp of an InDel (option: -g 3), clusters of InDels within 10 bp (option: -G 10), variant Phred quality scores below 30 (QUAL < 30), variants with a read depth of less than 25 or greater than 110 (half or double the average coverage, respectively), and variants not passing a series of stringent hard filters (documented at https://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set): QD<2, FS>60.0, MQ<40.0, MQRankSum<-12.5, or ReadPosRankSum<-8.

Using genome-wide heterozygous variation, it is possible to infer the distribution of allele coalescence along thousands of portions of the genome and to use such information to estimate historical population sizes. The Pairwise Sequential Markovian Coalescent (PSMC) model (Li & Durbin, 2011) was used to infer historical demography based on a heterozygous consensus sequence produced using only SNP genotypes (N = 846,905). We performed the analysis with the following parameters -N25 -t15 -r5 -p "4+25*2+4+6". We also used a bootstrapping approach to estimate the variation in demographic estimates, which was carried out by randomly sampling with replacement 5 Mb genomic segments to match the total sequence size of the empirical genome. This sampling was repeated 100 times and PSMC was run with the same parameters as above on each bootstrapped dataset. We rescaled time in units of years using a mutation rate of 2.0 x 10^{-9} mutations/site/generation and a generation time of 3 years.

Finally, we quantified the genome-wide distribution in heterozygosity and the location and putative impacts of genetic variation. To quantify genome-wide heterozygosity, we calculated the proportion of sites that were heterozygous in non-overlapping 50 kb windows (regions with >25% of positions with coverage below 25 or above 110 or with > 25% N gap sequence were excluded) and estimated the mean and variance in heterozygosity genome-wide. We used the Ensembl Variant Effect Predictor (McLaren et al., 2016) to determine the sequence ontology (SO) and impact of sequence variation based on annotations of protein-coding genes (see below for details on how these annotations were produced). We categorized variation as either SNPs or InDels and quantified the number of each type of variant that fell into 21 SO categories.

Tissue sampling and RNAseq library preparation, sequencing, and assembly

To produce a dataset of expressed genes in *Boa*, we obtained samples from 10 tissue types (see Supplementary Table 4 for more information). Blood samples from male and female *B. constrictor* collected as part of a previously published project were also obtained from NCBI (Vicoso, Emerson, Zektser, Mahajan, & Bachtrog, 2013; Supplementary Table 4). The nine other tissues were collected from two snakes obtained from commercial breeders and were preserved in RNAlater before being stored at -80°C. Total RNA was extracted from 25 mg tissue subsamples using Trizol reagent (Invitrogen). mRNAseq libraries were constructed using an

Illumina TruSeq RNAseq kit that employed poly-A selection, RNA fragmentation, cDNA synthesis, and adapter ligation. Multiplexed RNAseq tissue libraries were combined in equal molar ratios, quantified using a BioAnalyzer (Agilent Technologies), and were sequenced using an Illumina HiSeq2000 and 100 bp paired-end sequencing. We used Trinity v. r20140717 (Haas et al., 2013) with default parameters and internal Trimmomatic quality trimming to assembly the Illumina reads into transcript contigs. We performed a BUSCO analysis on single isoforms of assembled transcripts, as outlined above, but in the transcriptome mode.

Repeat annotation

A multi-step process was used to annotate repetitive content in the *Boa* genome. First, we constructed a *de novo Boa*-specific repeat library using RepeatModeler v. 1.0.8 (Smit & Hubley, 2008), which uses RepeatScout (Price, Jones, & Pevzner, 2005) and RECON (Bao & Eddy, 2002) to identify repetitive genomic regions using k-mer abundances and all-to-all mapping, respectively. We used CENSOR (Kohany, Gentles, Hankus, & Jurka, 2006) and BLAST v. 2.2.27+ (Altschul, Gish, Miller, Myers, & Lipman, 1990) searches against Repbase release 20150807 (Bao, Kojima, & Kohany, 2015; Jurka et al., 2005) and a custom 12 snake repeat library (see below) to further curate the resulting repeat library and assign unknown repeats to appropriate families. We ran RepeatMasker v. 4.0.6 (Smit, Hubley, & Green, 2013) serially to identify genome-wide repetitive sequences. The genome was first masked with a custom library to properly annotate BovB/CR1 LINE elements, which was necessary because of a previously recognized misannotation due to the BOVB_VA chimeric element that exists in Repbase (Castoe, Hall, et al., 2011). Repbase release 20150807 was then used to mask tetrapod elements. We also combined our *de novo Boa* repeat library with a previously-published library containing elements from sample genome sequencing of 12 snakes (Castoe et al., 2013): *Leptotyphlops dulcis*, *Typhlops reticulatus*, *Anilius scytale*, *Boa constrictor* (from previous low-coverage 454 sequencing), *Casarea dussumieri*, *Python molurus*, *Loxocemus bicolor*, *Sibon nebulatus*, *Thamnophis sirtalis*, *Agkistrodon contortrix*, *Crotalus atrox*, and *Micrurus fulvius*. We used this library to perform two rounds of masking with the first round using the subset of elements that could be assigned to specific repeat families (i.e., known; 7,745 elements) and the second round used the remaining, unassigned (i.e., unknown; 2,934 elements) repeat elements. This sequential mapping strategy prioritizes known, curated repeats while also accounting for potentially unique or more divergent repeats.

We were also interested in inferring the relative timing of activity for different TE families and subfamilies. We used the assumption that TE copies that expanded more recently are less divergent from their consensus than copies that reached fixation in the more distant past. Based on this assumption, we used the `calcDivergenceFromAlign.pl` script included in RepeatMasker to calculate the CpG-corrected Kimura 2-parameter divergence between all individual TE copies and their consensus sequence for each TE subfamily. We split divergence levels into bins of 1% evolutionary distance, and for each bin we calculated the proportion of the genome masked with a given TE subfamily to visualize temporal patterns of TE activity.

Gene annotation

Genes were annotated using MAKER v. 2.31.8 (Holt & Yandell, 2011) using an iterative process. For the first MAKER run, we extracted complex repeat annotations from the RepeatMasker annotation described above and provided them to MAKER using the "rm_gff" option in the MAKER control file. This allowed these complex repeats to be properly masked prior to gene annotation, and we also instructed MAKER to soft mask simple repeats by setting the “model_org” option in the MAKER control file to “simple”. Several forms of gene evidence were used to construct gene models. First, the *de novo* transcriptome assembled using Trinity was supplied as EST evidence. Second, we used protein sequences for gene models from three other squamate species – *Anolis carolinensis* (NCBI GCF_000090745.1; Alföldi et al., 2011), *Python molurus bivittatus* (NCBI GCF_000186305.1; Castoe et al., 2013), and *Thamnophis sirtalis* (NCBI GCF_001077635.1; Castoe, Bronikowski, et al., 2011; Perry et al., In Review) – as protein homology evidence. The “est2genome” and “protein2genome” options were turned on for the first MAKER run so that gene models would be constructed directly from the above evidence data.

The resulting gene models were used to train the gene prediction software SNAP (Korf, 2004) and Augustus (Stanke, Steinkamp, Waack, & Morgenstern, 2004; Stanke & Waack, 2003). For SNAP, we only used gene models of length 50 or greater amino acids and with a max AED threshold of 0.25. For Augustus, we extracted the genomic regions containing the transcript models and 1 kb of upstream and downstream sequence. We used BUSCO in the genome mode with the conserved Tetrapoda genes, but also specified the “--long" option, which uses Augustus self-training to optimize the gene prediction parameters. This has the effect of training Augustus on over 1,000 gene models constructed from the initial MAKER run. The resulting gene models were inputted into a second MAKER run alongside the repeat, EST, and protein evidence from the first MAKER run, but with the “est2genome” and “protein2genome” options turned off. Based on these settings, MAKER uses the gene models produced from SNAP and Augustus as the final annotation, with evidence supplied by the empirical EST and protein data. The resulting gene models were extracted and used to train SNAP and Augustus again, as above, and the resulting trained gene prediction models were then used for a third MAKER run that was otherwise identical to the second MAKER run. This iterative process has the effect of improving gene prediction parameter settings and thus the resulting gene annotation models produced from MAKER. We visually evaluated the resulting gene models from the third MAKER run alongside the empirical EST and protein data and found that SNAP produced gene models that were poorly supported by the empirical data. Because of this, we re-ran a third round of MAKER that excluded the SNAP gene prediction parameters in favor of those produced in Augustus, and after further evaluation of the overlap between empirical transcript and protein evidence and gene models, these gene models were considered the final gene annotation.

Finally, we identified the mitochondrial genome by using BLASTn (threshold 1e-10) and an existing whole mitochondrial genome sequence for *B. constrictor imperator* downloaded from NCBI as a query (accession AM236348.1; Douglas, Janke, & Arnason, 2006). We annotated this mitochondrially-derived scaffold using the MITOS webserver (Bernt et al., 2013) with default settings.

We assessed the quality of the annotation using BUSCO, as outlined above but in the protein mode. We also used a custom script to quantify the number and length of exons and introns, and other basic information about gene annotations. We used the protein sequences to annotate gene models based on homology to several outside sources. We used InterProScan v. 5.27-66.0 (Jones et al., 2014) to match proteins against the InterPro database version 66.0 (Mitchell et al., 2015) and BLAST with the e-value threshold set to 1e-5 to match proteins against UniProt/SwissProt release 2017-11-22 (The UniProt Consortium, 2017). We also performed both reciprocal best

BLAST and stringent unidirectional BLAST searches using a custom script between the protein annotations for *Boa* and annotated proteins from *Python*, *Thamnophis*, *Anolis*, and Human (International Human Genome Sequencing Consortium, 2001) obtained from NCBI. We used e-value cutoffs of 1e-5 for the reciprocal best BLASTp searches and 1e-8 for the stringent unidirectional BLASTp searches. When summarizing these searches, we prioritized the results of the reciprocal best BLASTp over the one-way BLASTp, as these reflect higher confidence in homology between *Boa* proteins and proteins of the other species.

Identification of squamate orthologs

To further evaluate our gene annotation, we identified gene families and orthology between protein sets from the *Boa* and four other squamate species obtained from NCBI: *Anolis*, *Python*, *Protobothrops mucrosquamatus* (Aird et al., 2017), and *Thamnophis*. The gene sets for each species were filtered to retain the longest coding sequence for each annotated gene and we also removed genes with protein sequences <50 amino acids in length. The resulting dataset ranged from 18,565 (*Thamnophis*) to 20,015 (*Protobothrops*) protein sequences (Supplementary Table 7). We used OrthoMCL (Fischer et al., 2002; Li, Stoeckert, & Roos, 2003) to group proteins into families based on homology and identify orthologous protein sequences across species. We automated the OrthoMCL analysis using the OrthoMCL Pipeline (https://github.com/apetkau/orthomcl-pipeline) and summarized the results to understand the numbers of different types of homologs across these species.

Evaluating cross-tissue gene expression

The presence of RNAseq data for several *Boa* tissues provides the opportunity to examine cross-tissue gene expression patterns. We quality-trimmed the raw paired-end Illumina data for each tissue using Trimmomatic with the settings LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:36, and over 95% of paired reads were retained (Supplementary Table 4). Quality trimmed reads were mapped using STAR v. 2.5.2b (Dobin et al., 2013) and the transcript features produced from our gene annotation to produce a first-pass mapping. We collected the junctions for all samples and then ran a second mapping pass for all samples. This two-pass method produces alignments with high sensitivity to splice junctions. We used HTSeq (Anders, Pyl, & Huber, 2015) to produce raw expression counts for each gene. Data were normalized using TMM normalization (Robinson & Oshlack, 2010) implemented in the R (v. 3.4.3; R Core Team, 2018) package edgeR (v. 3.20.9; McCarthy, Chen, & Smyth, 2012; Robinson, McCarthy, & Smyth, 2010), and were converted to units of counts per million (CPM). We visualized cross-tissue gene expression patterns using a heatmap, with expression scaled independently for each gene and tissue expression profiles clustered by similarity based on the complete linkage method.

Data Availability

The existing *B. constrictor* reference genome is available from the *GigaScience* database (Bradnam et al., 2013) and raw reads used to construct this genome were already available through NCBI (accession ERP002294). The new and existing RNA sequencing reads from each tissue library have been deposited at NCBI (accession SRP148755). Supplementary data and results files are provided in a figshare repository.

RESULTS & DISCUSSION

Boa constrictor genome assembly composition and heterozygosity

Detailed information on genome quality is available from Bradnam *et al.* (2013); because contig and scaffold N50 values are commonly reported and understood assembly metrics, we have included these statistics for this assembly in Table 1. Our BUSCO analysis identified 3,694 of 3,950 total conserved Tetrapoda genes (93.5%) as complete in the *Boa* genome assembly (Table 1 and Supplementary Table 1). Less than 1% of these complete genes were duplicated, and only 256 genes (6.5%) were inferred to be fragmented or absent from the assembly (Supplementary Table 1). These results indicate the *Boa* genome assembly is high quality, further confirming findings reported in Bradnam *et al.* (2013).

Another shortcoming of the existing description of the *Boa* genome assembly, in addition to its lack of an annotation, is that no biological features of the genome sequence itself were identified or analyzed. To address this, we performed several analyses that quantify various aspects of the genome sequence itself. Mean genome-wide GC content based on 27,936 windows was 40.2%, but varied markedly with a minimum of 33.0% and a maximum of 62.7% (Fig. 1A and Table 1). GC content in the CDS regions and at third codon positions, in particular, was higher on average (48.4% and 51.1%, respectively) and more variable (range 26.7%-80.2% and 15.8%-100%, respectively; Fig. 1A and Table 1). Our k-mer-based analysis of genome size found similar estimates of genome size across three k-mer sizes that provide consistent support for a total genome size of approximately 1.3 Gb (Table 1 and Supplementary Fig. 1). This estimate is appreciably smaller than estimated genome size of 1.75 Gb derived from static cell fluorometry (De Smet, 1981), but is closer to the total assembly length (1.44 Gb) and within the range of estimates for Squamate reptiles provided by the Animal Genome Size Database (Gregory, 2018; Gregory et al., 2007), indicating that our estimate is credible.

We used 979,326 biallelic SNPs to infer the historical demography of this species and to evaluate the potential functional impacts of genomic variation by quantifying the composition and location of heterozygous variation. Our PSMC analysis indicated that effective population size (N_e) has varied cyclically over the approximately 2.5 million years of population history captured by PSMC, ranging from a low of about 30,000 to a recent high of about 250,000 (Fig. 1B). Mean heterozygosity based on results from 27,736 50 kb windows was 9.2 x 10^{-4} with a standard deviation of 4.1 x 10^{-4} (Table 1 and Supplementary Fig. 2). As expected, the vast majority of variation falls well outside coding regions, with only 0.45% of variation being classified as moderate or high impact (Fig. 1C and Supplementary Table 3). We further evaluated the length of InDel variants in coding regions and found that most were less than 6 bp in length and, expectedly, had lengths that were multiples of three, which does not disrupt the reading frame (Fig. 1D). Collectively, these analyses provided important foundational genomic characteristics that were lacking from the original genome description.

Description of de novo transcriptome assembly

The *de novo* transcriptome assembly contained 475,359 transcripts representing 374,608 Trinity genes (average of 1.27 isoforms per gene). Mean transcript length was 837 bp and transcript N50 was 1,732 bp. Only 2,481 (62.8%) of tetrapod BUSCOs were found in the transcriptome assembly, with 1,743 (44.1%) and 738 (18.7%) being single-copy and duplicated, respectively. 844 (21.4%) of BUSCOs are fragmented and 625 (15.8%) are missing (Supplementary Table 3). The moderate recovery of complete BUSCO genes from transcripts alone suggests that while our transcriptome assembly did include transcripts for many genes, a substantial subset of genes were not included in our empirical transcript set – this is likely due to the fact that we did not include all body tissues, and also lacked tissues from developmental samples (e.g., embryonic stages).

Boa repeat element landscapes and historical activity

We identified approximately 2.6 million repetitive elements that collectively comprised 31.61% (ca. 439 Mb) of the *Boa* genome. Identifiable transposable elements (TEs) account for 29.6% of the assembly, while simple sequence repeats (microsatellites) represent 2.35% of the assembly (Fig. 2 and Supplementary Table 5). LINE elements are most abundant in the *Boa* genome (12.8%), with DNA transposons (5.2%), LTR elements (2.3%), non-LTR elements (1.1%), and Penelope-like elements (1.0%) also comprising significant portions of the genome (Fig. 2 and Supplementary Table 5). Like other reptile species, the *Boa* genome is dominated by L2, CR1, and BovB LINE elements (Fig. 2A and Supplementary Table 5), whereas DNA elements are noticeably less represented; this feature is shared between *Boa* and the Burmese python – the most closely related species with an annotated genome. Our analyses suggest that the *Boa* genome underwent three major waves of TE amplification: a first major expansion of LINE elements (L2 in particular) was followed by a long-lasting reduction in L2 activity and concomitant increase in DNA elements transposition, whereas more recent and likely ongoing TE activity appears to be mostly restricted to BovB LINEs and MIR SINEs (Fig. 2B). These general dominance in expansion of LINE elements are shared with other squamate genomes analyzed to date (Pasquesi et al., In Review).

Details of Boa gene annotation

The resulting gene annotation included a total of 19,178 gene models. Mean gene length was approximately 17 kb while mean CDS length was 1,455 bp (Table 2 and Supplementary Fig. 3). On average, each gene was approximately structured into 9 exons of 340 bp in length and 8 introns of 2,150 bp in length (Table 2, and Supplementary Fig. 3). The mitochondrial genome BLASTn search identified a single ~11.8 kb hit and several additional hits on *Boa* genome scaffold "scaffold-4019", clearly identifying this sequence as a mitochondrially-derived scaffold. This scaffold is 17,048 bp in length (query sequence was 16,607 bp) but contains a 1,450 bp N gap sequence that corresponds to control region 1 (snakes possess duplicate control regions; Dong & Kumazawa, 2005; Jiang et al., 2007; Kumazawa, Ota, Nishida, & Ozawa, 1996). All protein-coding (N=13), ribosomal (N=2), and tRNA (N=22) genes were annotated. Genome-wide, 3,159 (79.9%) of BUSCOs were complete, with most (3,114) being single-copy. About 1.1% of BUSCOs were duplicated, 14.2% of BUSCOs were fragmented, and 5.9% of BUSCOs were missing (Table 2 and Supplementary Table 3). 83.7% of proteins were annotated using Swiss-Prot, 89.9% were matched to elements in the InterPro database, and 82.4% of proteins were matched against HMM models from PFAM. We were also able to ascribe gene ontology and PANTHER pathway IDs to 68.5% and 92.4% genes, respectively (Supplementary Table 6). For all comparisons with *Python*, *Thamnophis*, *Anolis*, and Human over 90% of proteins were matched against homologs from the reference species, with about 68% to 76% of proteins confidently assessed based on reciprocal best BLAST searches (Supplementary Table 6). Only 3.3% of annotated proteins were left without assignment based on either protein databases or homology with other vertebrate species (Supplementary Table 6). These results indicate that we created a high-quality gene annotation for the *Boa* genome and our successful homology-based identification of genes provides a significant resource for those investigating various biological questions in this group and snakes in general.

Ortholog and gene family classifications across Squamata

Greater than 85% of genes for each species were grouped into protein families, and the number of families ranged from approximately 13,000 to 14,000 (Fig. 3A, Supplementary Table 8). The average gene family size across species ranged from 1.20 to 1.32 genes per family and the maximum gene family size ranged from 27 to 340 (Supplementary Table 8). 7,398 genes were complete one-to-one orthologs across all species (i.e., were not missing in any species) and between 2,428 and 3,264 additional genes were one-to-one orthologs across two or more species (Fig 3A). *Anolis* contained the most unique paralogs not found in any other species (N = 523) and *Thamnophis* contained the least (N = 38; Fig. 3A). The low number of unclustered genes and the high number of inferred one-to-one orthologs indicates that the *B. constrictor* gene annotation is relatively high quality. Moreover, our sets of orthologs span most snake diversity and can therefore be used in future investigations of gene family and protein evolution.

Patterns of cross-tissue gene expression

Greater than 90% of all reads were mapped in nine out of the 10 tissues, with the remaining tissue (brain) having a mapping rate of 75% (Supplementary Table 4), indicating high-quality RNAseq data and a well-constructed gene annotation. Testes contained the highest number of expressed (CPM > 2) genes (N = 14,031), while muscle contained the lowest (N = 7,896; Fig. 3B). Alternatively, muscle possessed the highest average CPM expression level (126.1 CPM; genes with CPM < 2 excluded), while testes possessed the lowest (79.1 CPM; Fig 3B). Some tissues shared very similar expression patterns, including muscle and skin, blood and spleen, and stomach and small intestine (Fig 3B). We found 6,561 genes that were broadly expressed across all tissues examined (Fig 3B). These results provide a preliminary look at cross-tissue gene expression patterns that can be further refined in future investigations of *Boa* physiology.

Conclusion

In summary, we described biological characteristics of the existing *Boa constrictor* genome sequence and produced high-quality repeat and gene annotations that will be a valuable resource for researchers studying many interesting aspects of snake and vertebrate biology.

ACKNOWLEDGEMENTS

This work was supported by startup funding from the University of Texas Arlington to TAC, an NSF Award to TAC (NSF IOS-1655735), and an NSF Doctoral Dissertation Improvement Grant to DCC and TAC (NSF DEB-1501747). We would like to thank all anonymous reviewers for their comments on this manuscript.

FIGURES

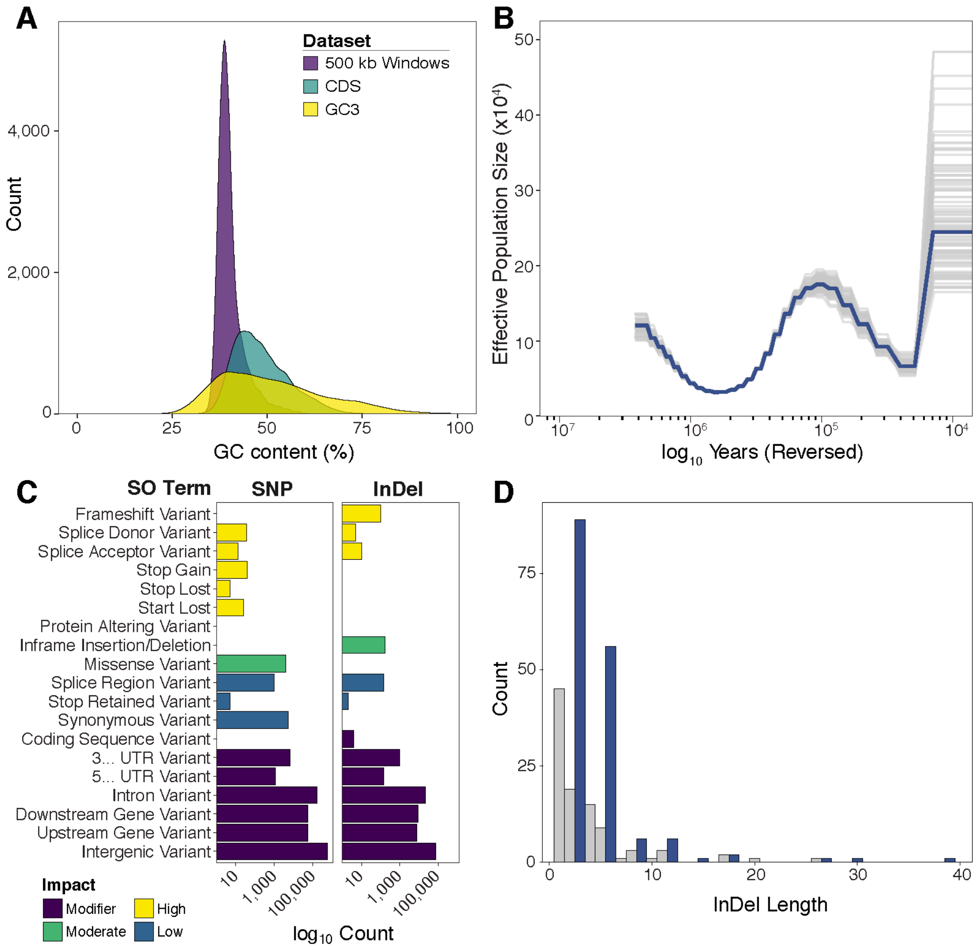


Figure 1. Summary of genomic composition. (A) Distributions of GC content in 500kb windows genome-wide, CDS sequences, and 3rd codon positions (GC3). **(B)** Historical demography of *B. constrictor* based on genome-wide SNPs. Gray lines indicate bootstrap replicates while the blue line resulted from the empirical dataset. Time was scaled to years assuming a generation time of 3 years and a mutation rate of 2.0x10^9 mutations/site/generation. **(C)** Sequence ontologies of genome-wide SNP (left) and InDel (right) variation color coded by biological impact. **(D)** Distribution of InDel lengths in coding regions with grey bars representing shifts that are not divisible by 3 and blue bards indicating InDels that do not disrupt the downstream reading frame.

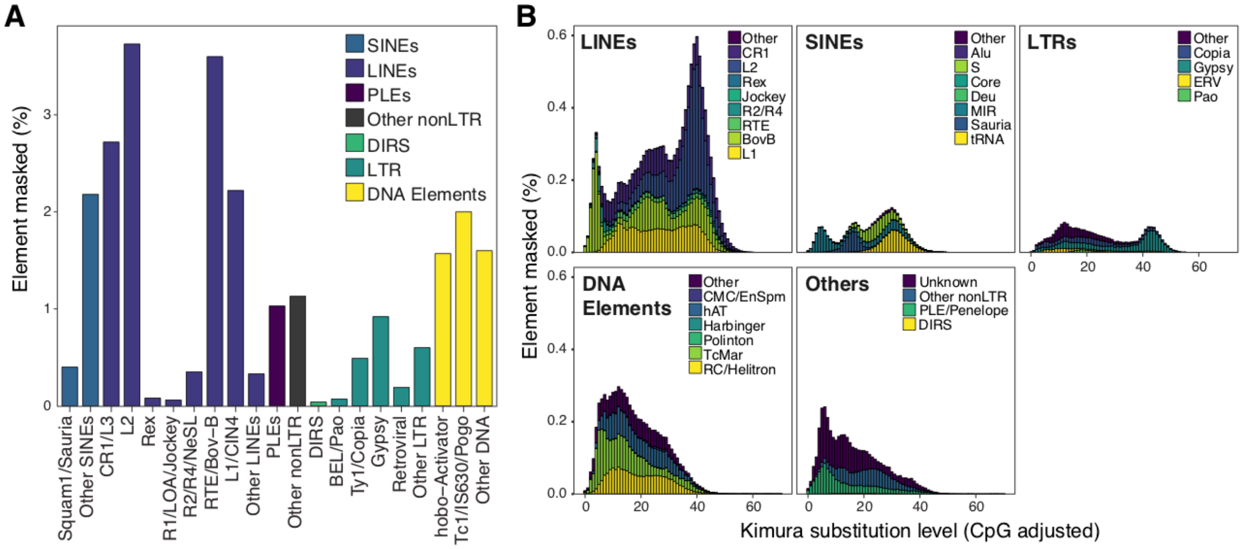


Figure 2. Summary of repetitive element annotations. (A) Composition of repeat element families genome-wide. Bars and family names are color coded based on major TE classifications. **(B)** Age distributions of major TE classifications and families based on the substitution levels.

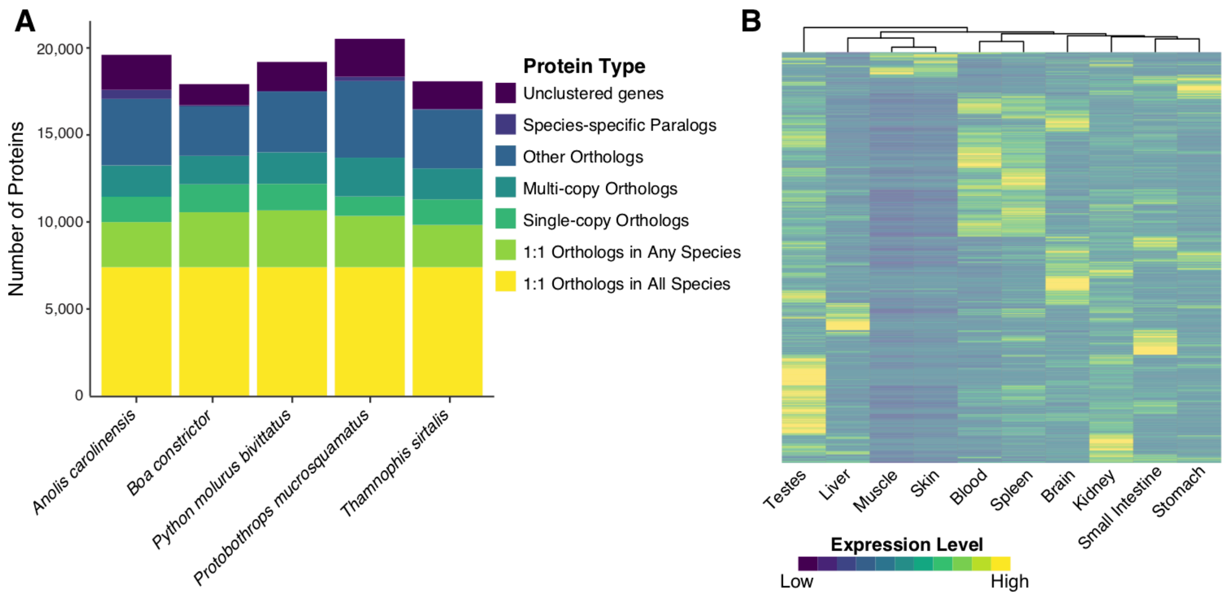


Figure 3. Overview of gene annotation. (A) Classifications of protein homology between *B. constrictor* and 4 other squamate species inferred from OrthoMCL. **(B)** Heatmap of gene expression across 10 sampled tissue types clustered by genome-wide expression profiles (N = 15,404 expressed genes total). Gene expression was scaled individually for each gene.

TABLES

Table 1. Characteristics of the existing *Boa constrictor* genome assembly.

Statistic	**Measure**
Contig N50 (bp)	4,505,203
Scaffold N50 (bp)	29,326
Mean (±SD) GC Content in 50 kb windows	40.2% (±3.6%)
Mean (±SD) GC Content in CDS regions	48.4% (±7.1%)
Mean (±SD) GC Content in third codon positions	51.1% (±13.9%)
Mean (±SD) Heterozygosity	9.2×10^{-4} ($\pm 4.1 \times 10^{-4}$)
Complete BUSCO genes (%)	3,694 (93.5%)
Mean k-mer genome size estimate	1.30 Gbp

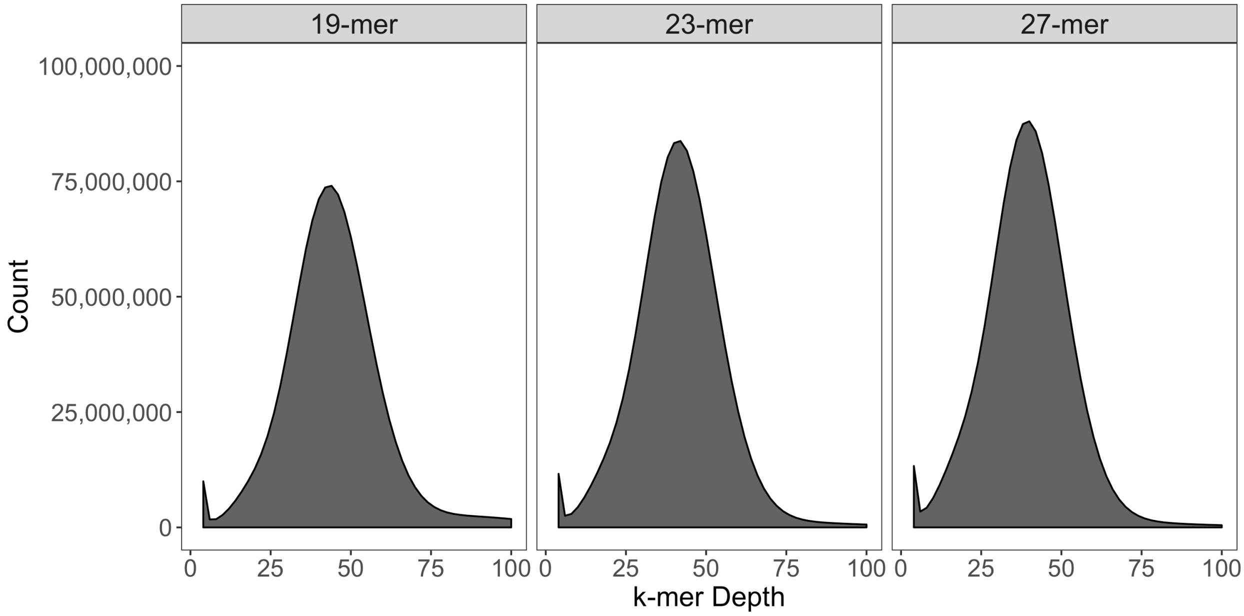
Note: SD, standard deviation

Table 2. Gene annotation statistics.

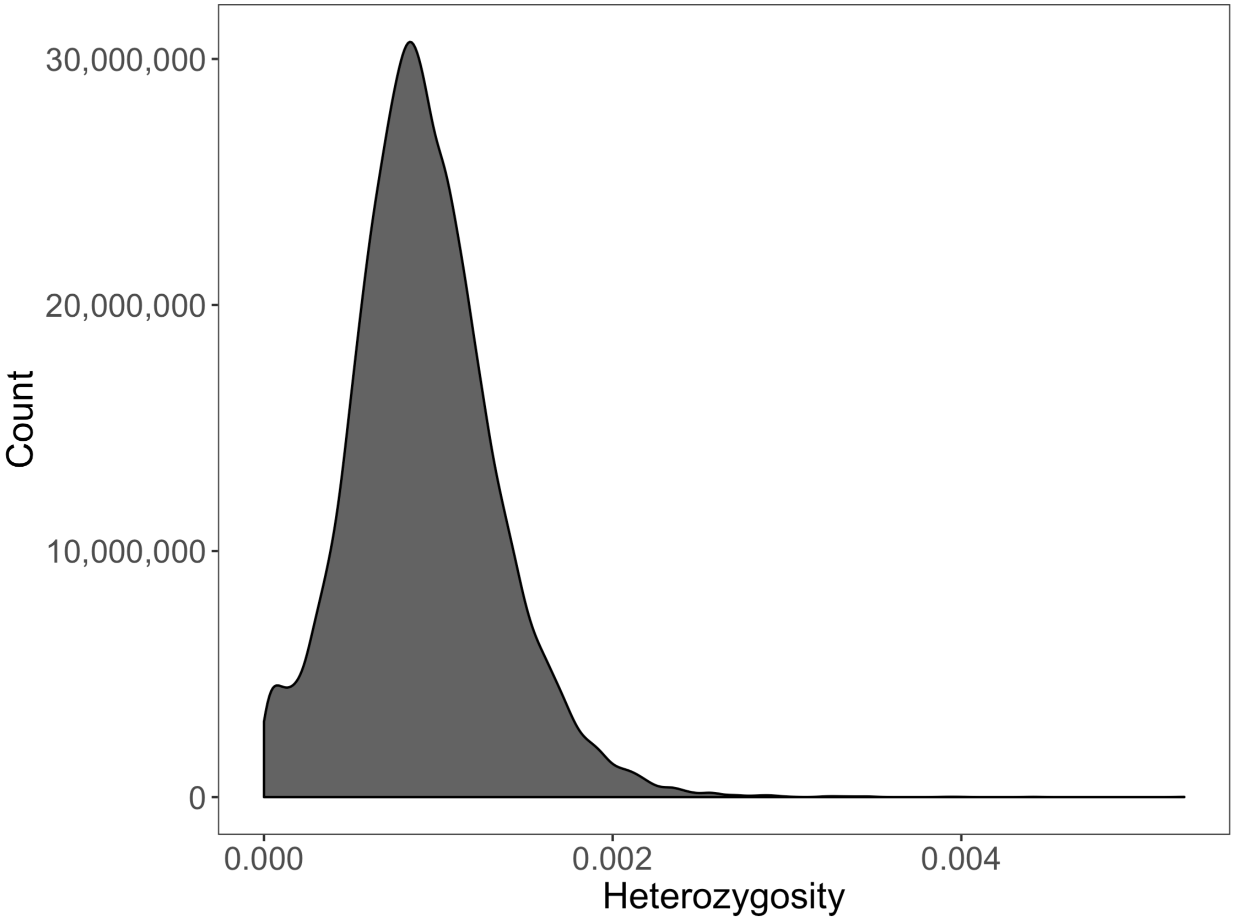
Statistic	**Measure**
Annotated transcripts/proteins	19,178
Mean (±SD) gene length (bp)	17,117.1 (±16,266)
Mean (±SD) CDS length (bp)	1,455.1 (±1,460.7)
Mean (±SD) exons per gene	9.05 (±8.68)
Mean (±SD) exon length (bp)	340.7 (±365)
Mean (±SD) introns per gene	8.05 (±8.68)
Mean (±SD) intron length (bp)	2,154.1 (±1,498.7)
Complete BUSCO genes (%)	3,159 (79.9%)

Note: SD, standard deviation

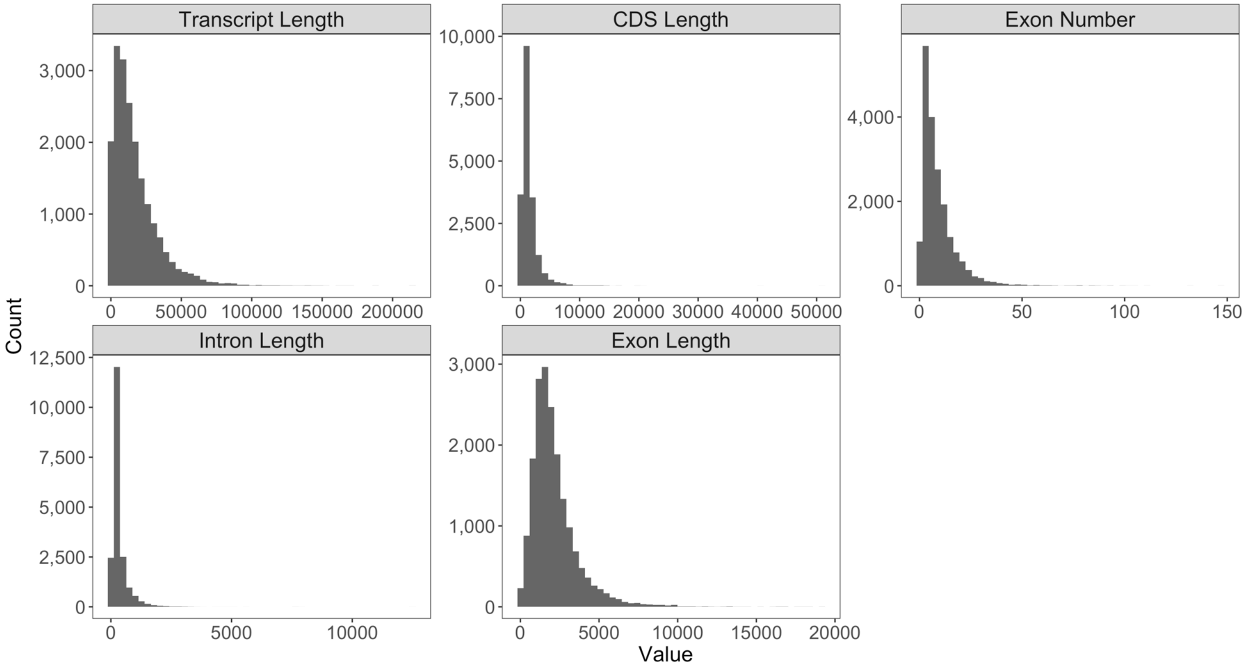
SUPPLEMENTARY MATERIAL



Supplementary Figure 1. Distributions of 19-mers, 23-mers, and 27-mers used to estimate the genome size of *Boa constrictor*.



Supplementary Figure 2. Distribution of heterozygosity across non-overlapping 500kb windows.



Supplementary Figure 3. Distribution of gene structure characteristics across annotated genes.

Supplementary Table 1. Results of BUSCO analyses using the Tetrapoda database (N = 3,950 BUSCOs) of the genome, transcriptome, and gene annotation.

Sequence Set	**Complete**	**Complete Single-copy**	**Complete Duplicated**	**Fragmented**	**Missing**
Genome	3,694 (93.5%)	3,669 (92.9%)	25 (0.6%)	135 (3.4%)	121 (3.1%)
Transcriptome	2,481 (62.8%)	1,743 (44.1%)	738 (18.7%)	844 (21.4%)	625 (15.8%)
Protein Annotations	3,159 (79.9%)	3,114 (78.8%)	45 (1.1%)	560 (14.2%)	231 (5.9%)

Supplementary Table 2. Summary of genomic sequencing reads mapped to the *Boa constrictor* genome.

NCBI Run Accession	Insert Size	Raw PE Reads	Quality-trimmed PE Reads	Mapped PE Reads
ERR234359	400	34,584,029	34,416,309	34,276,044
ERR234360	400	34,127,072	33,952,566	33,812,577
ERR234361	400	34,449,005	34,262,618	34,123,346
ERR234362	400	33,967,757	33,808,987	33,672,096
ERR234363	400	34,234,072	34,084,659	33,946,097
ERR234364	400	34,550,177	34,403,113	34,264,050
ERR234365	400	34,590,978	34,448,550	34,308,454
ERR234366	400	34,752,322	34,582,976	34,438,496
ERR234367	400	35,956,546	35,790,957	35,645,856
ERR234368	400	35,802,014	35,633,437	35,487,522

Supplementary Table 3. Genomic locations of SNPs and InDels in the *Boa constrictor* genome.

Variant Type	SO accession	SO term	Count	Impact
SNP	SO:0001628	Intergenic Variant	593,408	Modifier
	SO:0001631	Upstream Gene Variant	57,242	Modifier
	SO:0001632	Downstream Gene Variant	58,617	Modifier
	SO:0001627	Intron Variant	175,971	Modifier
	SO:0001623	5’ UTR Variant	1,178	Modifier
	SO:0001624	3’ UTR Variant	6,560	Modifier
	SO:0001819	Synonymous Variant	5,434	Low
	SO:0001583	Missense Variant	4,179	Moderate
	SO:0002012	Start Lost	26	High
	SO:0001578	Stop Lost	5	High
	SO:0001587	Stop Gain	40	High
	SO:0001567	Stop Retained Variant	5	Low
	SO:0001630	Splice Region Variant	1,018	Low
	SO:0001574	Splice Acceptor Variant	13	High
	SO:0001575	Splice Donor Variant	36	High
InDel	SO:0001628	Intergenic Variant	71,519	Modifier
	SO:0001631	Upstream Gene Variant	7,561	Modifier
	SO:0001632	Downstream Gene Variant	8,416	Modifier

SO:0001627	Intron Variant	21,741	Modifier
SO:0001623	5’ UTR Variant	140	Modifier
SO:0001624	3’ UTR Variant	937	Modifier
SO:0001589	Frameshift Variant	100	High
SO:0001821/2*	Inframe Insertion/Deletion	163	Moderate
SO:0002012	Start Lost	1	High
SO:0001587	Stop Gain	1	High
SO:0001567	Stop Retained Variant	2	Low
SO:0001630	Splice Region Variant	140	Low
SO:0001574	Splice Acceptor Variant	10	High
SO:0001575	Splice Donor Variant	5	High
SO:0001580	Coding Sequence Variant	4	Modifier
SO:0001818	Protein Altering Variant	1	Moderate

* Note: Because InDels cannot be ascribed as either insertions or deletions, we have reported the combination of two sequence ontology terms.

Supplementary Table 4. Summary of RNA sequencing reads.

NCBI Run	Organ	Raw PE Reads	Quality-trimmed	Mapped PE
SRR7206975	Muscle	6,104,026	5,942,622	11,447,549
SRR7206974	Muscle	6,080,052	5,922,951	
SRR7206973	Small intestine	4,209,137	4,104,664	7,543,439
SRR7206972	Small intestine	4,196,586	4,094,982	
SRR7206971	Liver	7,784,321	7,576,382	13,563,790
SRR7206970	Liver	7,759,817	7,558,021	
SRR7206969	Kidney	3,844,237	3,746,864	6,927,890
SRR7206968	Kidney	3,830,034	3,735,184	
SRR7206977	Skin	4,126,300	4,011,760	7,588,709
SRR7206976	Skin	4,110,431	3,999,323	
SRR7206965	Stomach	4,118,406	4,015,626	7,300,853
SRR7206964	Stomach	4,104,984	4,006,145	
SRR7206967	Brain	6,902,817	6,189,598	4,623,644
SRR7206966	Testes	6,619,650	6,000,965	5,373,905
SRR7206963	Spleen	8,128,893	7,356,245	6,457,744
SRR941243	Blood (Male)	12,985,828	11,552,988	22,248,033
SRR941236	Blood (Female)	12,985,828	11,594,239	

Supplementary Table 5. Full summary of repetitive elements annotated in the *Boa constrictor* genome.

	# elements	length masked (bp)	% of genome sequence	% of masked elements
Total masked	2,635,008	438,665,981	31.61	100.00
Total interspersed repeats	1,943,114	410,724,575	29.60	73.74
Retroelements	1,124,019	276,381,763	19.25	42.66
SINEs	260,879	35,884,431	2.59	9.90
Squam1/Sauria	24,609	5,591,218	0.40	0.93
Other SINEs	236,270	30,293,213	2.18	8.97
LINEs	574,159	178,057,797	12.83	21.79
CR1-Like	298,405	86,966,977	6.27	11.32
CR1/L3	135,482	37,713,269	2.72	5.14
L2	182,283	51,719,221	3.73	6.92
Rex	2,702	1,048,001	0.08	0.10
R1/LOA/Jockey	3,200	892,892	0.06	0.12
R2/R4/NeSL	12,481	4,873,749	0.35	0.47
RTE/BovB	154,680	50,013,724	3.60	5.87
L1/CIN4	80,593	30,794,376	2.22	3.06
Other LINEs	121,796	13,710,446	0.33	4.62
Other nonLTR	82,241	15,630,018	1.13	3.12
DIRS	3,866	498,882	0.04	0.15
PLEs	85,022	14,292,039	1.03	3.23
LTR elements	113,986	31,519,714	2.27	4.33
BEL/Pao	2,133	936,820	0.07	0.08
Ty1/Copia	25,148	6,868,658	0.49	0.95
Gypsy	28,572	12,813,803	0.92	1.08
Retroviral	16,204	2,643,460	0.19	0.61
Other LTR	41,929	8,256,973	0.60	1.59

DNA transposons	476,127	71,838,415	5.18	18.07
hobo-Activator	178,421	21,853,263	1.57	6.77
Tc1-IS630-Pogo	111,126	27,704,889	2.00	4.22
En-Spm	2,044	419,877	0.03	0.08
MuDR-IS905	2,100	450,198	0.03	0.08
PiggyBac	2,450	207,930	0.01	0.09
Tourist/Harbinger	3,280	222,129	0.02	0.12
P elements	5,847	1,241,052	0.09	0.22
Rolling-circles	2,883	762,608	0.05	0.11
SPIN	-	-	0.00	0.00
Other DNA	167,976	18,976,469	1.37	6.37
Unclassified	342,968	54,571,520	3.93	13.02
Total interspersed repeats	1,943,114	410,724,575	29.60	73.74
Small RNA	3,946	329,376	0.02	0.15
Satellites	3,518	742,792	0.05	0.13
Simple repeats	615,353	32,541,197	2.35	23.35
Low complexity	69,077	3,616,028	0.26	2.62

Supplementary Table 6. Functional annotation of predicted genes in the *Boa constrictor* genome.

	Database	Number	Percent (%)
Total		19,178	100
	Swiss-Prot	16,054	83.71
	InterPro	17,238	89.88
	PFAM	15,800	82.39
	GO	13,143	68.53
Annotated	**PANTHER**	17,718	92.39
	Python*	17,948 (14,649)	93.59 (76.38)
	Thamnophis*	17,627 (13,444)	91.91 (70.10)
	Anolis*	18,010 (13,748)	93.91 (71.69)
	Human*	17,871 (13,002)	93.18 (67.79)
Unannotated		627	3.27

* Format: All matches (reciprocal best-BLAST matches)

Supplementary Table 7. Gene annotation characteristics for species used in orthology analysis.

Species	Genome Version	Number of annotated proteins	Number of original genes	Average gene length (bp)
Anolis carolinensis	AnoCar2.0	34,826	19,367	52,482.3
Boa constrictor	SGA (snake 7C) assembly	19,178	19,178	17,117.1
Python molurus bivittatus	Python_molurus_bivittatus-5.0.2	26,040	18,750	26,172.9
Protobothrops mucrosquamatus	P.Mucros_1.0	22,660	20,015	26,367.9
Thamnophis sirtalis	Thamnophis_sirtalis-6.0	25,180	18,565	27,013.1

Supplementary Table 8. Summary statistics of gene families in 5 squamate species based on OrthoMCL.

Species	Total genes	Genes in families	Unclustered genes	Families	Unique families	Genes per family	Max gene family size
Anolis carolinensis	19,360	17,596	1,764	13,411	523	1.28	340
Boa constrictor	19,005	16,712	2,293	13,825	68	1.20	27
Python molurus bivittatus	18,741	17,532	1,209	14,024	49	1.24	154
Protobothrops mucrosquamatus	20,002	18,348	1,654	13,785	245	1.32	187
Thamnophis sirtalis	18,559	16,491	2,068	13,097	38	1.25	127

CITATIONS

Aird, S. D., Arora, J., Barua, A., Qiu, L., Terada, K., & Mikheyev, A. S. (2017). Population Genomic Analysis of a Pitviper Reveals Microevolutionary Forces Underlying Venom Chemistry. *Genome Biology and Evolution*, *9*(10), 2640–2649. https://doi.org/10.1093/gbe/evx199

Alföldi, J., Palma, F. D., Grabherr, M., Williams, C., Kong, L., Mauceli, E., … Lindblad-Toh, K. (2011). The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, *477*(7366), 587. https://doi.org/10.1038/nature10390

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Andrade, D. V., Toledo, L. F. D., Abe, A. S., & Wang, T. (2004). Ventilatory compensation of the alkaline tide during digestion in the snake *Boa constrictor*. *Journal of Experimental Biology*, *207*(8), 1379–1385. https://doi.org/10.1242/jeb.00896

Andrew, A. L., Card, D. C., Ruggiero, R. P., Schield, D. R., Adams, R. H., Pollock, D. D., … Castoe, T. A. (2015). Rapid changes in gene expression direct rapid shifts in intestinal form and function in the Burmese python after feeding. *Physiological Genomics*, *47*(5), 147–157. https://doi.org/10.1152/physiolgenomics.00131.2014

Andrew, A. L., Perry, B. W., Card, D. C., Schield, D. R., Ruggiero, R. P., McGaugh, S. E., … Castoe, T. A. (2017). Growth and stress response mechanisms underlying post-feeding regenerative organ growth in the Burmese python. *BMC Genomics*, *18*, 338. https://doi.org/10.1186/s12864-017-3743-1

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*, 11. https://doi.org/10.1186/s13100-015-0041-9

Bao, Z., & Eddy, S. R. (2002). Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Research*, *12*(8), 1269–1276. https://doi.org/10.1101/gr.88502

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., … Stadler, P. F. (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, *69*(2), 313–319. https://doi.org/10.1016/j.ympev.2012.08.023

Boback, S. M. (2005). Natural History and Conservation of Island Boas (*Boa constrictor*) in Belize. *Copeia*, *2005*(4), 879–884. https://doi.org/10.1643/0045-8511(2005)005[0879:NHACOI]2.0.CO;2

Boback, S. M. (2006). A Morphometric Comparison of Island and Mainland Boas (*Boa constrictor*) in Belize. *Copeia*, *2006*(2), 261–267. https://doi.org/10.1643/0045-8511(2006)6[261:AMCOIA]2.0.CO;2

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bradnam, K., Fass, J., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., … Howard, J. (2013). Assemblathon 2 assemblies. *GigaScience Database*. https://doi.org/10.5524/100060

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., … Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1), 1–31. https://doi.org/10.1186/2047-217X-2-10

Card, D. C., Schield, D. R., Adams, R. H., Corbin, A. B., Perry, B. W., Andrew, A. L., … Castoe, T. A. (2016). Phylogeographic and population genetic analyses reveal multiple species of *Boa* and independent origins of insular dwarfism. *Molecular Phylogenetics and Evolution*, *102*, 104–116. https://doi.org/10.1016/j.ympev.2016.05.034

Castoe, T. A., Bronikowski, A. M., Brodie III, E. D., Edwards, S. V., Pfrender, M. E., Shapiro, M. D., … Warren, W. C. (2011). A proposal to sequence the genome of a garter snake (*Thamnophis sirtalis*). *Standards in Genomic Sciences*, *4*(2), 257. https://doi.org/10.4056/sigs.1664145

Castoe, T. A., Hall, K. T., Mboulas, G., L, M., Gu, W., Koning, D., … Pollock, D. D. (2011). Discovery of Highly Divergent Repeat Landscapes in Snake Genomes Using High-Throughput Sequencing. *Genome Biology and Evolution*, *3*, 641–653. https://doi.org/10.1093/gbe/evr043

Castoe, T. A., Koning, A. P. J. de, Hall, K. T., Card, D. C., Schield, D. R., Fujita, M. K., … Pollock, D. D. (2013). The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proceedings of the National Academy of Sciences*, *110*(51), 20645–20650. https://doi.org/10.1073/pnas.1314475110

De Smet, W. H. O. (1981). The nuclear Feulgen-DNA content of the vertebrates (especially reptiles), as measured by fluorescence cytophotometry, with notes on the cell and the chromosome size. *Acta Zoologica et Pathologica Antverpiensia*, (76), 119–167.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dong, S., & Kumazawa, Y. (2005). Complete Mitochondrial DNA Sequences of Six Snakes: Phylogenetic Relationships and Molecular Evolution of Genomic Features. *Journal of Molecular Evolution*, *61*(1), 12–22. https://doi.org/10.1007/s00239-004-0190-9

Douglas, D. A., Janke, A., & Arnason, U. (2006). A mitogenomic study on the phylogenetic position of snakes. *Zoologica Scripta*, *35*(6), 545–558. https://doi.org/10.1111/j.1463-6409.2006.00257.x

Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., Iodice, J. B., … Stoeckert, C. J. (2002). Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc. https://doi.org/10.1002/0471250953.bi0612s35

Gamble, T., Castoe, T. A., Nielsen, S. V., Banks, J. L., Card, D. C., Schield, D. R., … Booth, W. (2017). The Discovery of XY Sex Chromosomes in a Boa and Python. *Current Biology*, *27*(14), 2148-2153.e4. https://doi.org/10.1016/j.cub.2017.06.010

Gregory, T. R. (2018). *Animal Genome Size Database*. Retrieved from http://www.genomesize.com/

Gregory, T. R., Nicol, J. A., Tamm, H., Kullman, B., Kullman, K., Leitch, I. J., … Bennett, M. D. (2007). Eukaryotic genome size databases. *Nucleic Acids Research*, *35*(suppl_1), D332–D338. https://doi.org/10.1093/nar/gkl828

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494. https://doi.org/10.1038/nprot.2013.084

Henderson, R. W., Waller, T., Micucci, P., Puorto, G., & Bourgeois, R. W. (1995). Ecological correlates and patterns in the distribution of Neotropical boines (Serpentes: Boidae): a preliminary assessment. *Herpetological Natural History*, *3*, 1.

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*, 491. https://doi.org/10.1186/1471-2105-12-491

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Jiang, Z. J., Castoe, T. A., Austin, C. C., Burbrink, F. T., Herron, M. D., McGuire, J. A., … Pollock, D. D. (2007). Comparative mitochondrial genomics of snakes: extraordinary substitution rate dynamics and functionality of the duplicate control region. *BMC Evolutionary Biology*, *7*, 123. https://doi.org/10.1186/1471-2148-7-123

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, *110*(1–4), 462–467. https://doi.org/10.1159/000084979

Kohany, O., Gentles, A. J., Hankus, L., & Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, *7*, 474. https://doi.org/10.1186/1471-2105-7-474

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59. https://doi.org/10.1186/1471-2105-5-59

Kumazawa, Y., Ota, H., Nishida, M., & Ozawa, T. (1996). Gene rearrangements in snake mitochondrial genomes: highly concerted evolution of control-region-like sequences duplicated and inserted into a tRNA gene cluster. *Molecular Biology and Evolution*, *13*(9), 1242–1254. https://doi.org/10.1093/oxfordjournals.molbev.a025690

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493. https://doi.org/10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Lignot, J.-H., Helmstetter, C., & Secor, S. M. (2005). Postprandial morphological response of the intestinal epithelium of the Burmese python (*Python molurus*). *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *141*(3), 280–291. https://doi.org/10.1016/j.cbpb.2005.05.005

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., … Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *ArXiv:1308.2012 [q-Bio]*. Retrieved from http://arxiv.org/abs/1308.2012

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–4297. https://doi.org/10.1093/nar/gks042

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., … Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*, 122. https://doi.org/10.1186/s13059-016-0974-4

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., … Finn, R. D. (2015). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research*, *43*(D1), D213–D221. https://doi.org/10.1093/nar/gku1243

Pasquesi, G. I. M., Adams, R. H., Card, D. C., Schield, D. R., Corbin, A. B., Perry, B. W., … Shortt, J. A. (In Review). Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nature Communications*.

Perry, B. W., Card, D. C., McGlothlin, J. W., Pasquesi, G. I. M., Adams, R. H., Schield, D. R., … Castoe, T. A. (In Review). Molecular adaptations for sensing and securing prey, and insight into amniote genome diversity, from the garter snake genome. *Genome Biology and Evolution*.

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, *21*(suppl_1), i351–i358. https://doi.org/10.1093/bioinformatics/bti1018

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Reynolds, R. G., Niemiller, M. L., & Revell, L. J. (2014). Toward a Tree-of-Life for the boas and pythons: Multilocus species-level phylogeny with unprecedented taxon sampling. *Molecular Phylogenetics and Evolution*, *71*, 201–213. https://doi.org/10.1016/j.ympev.2013.11.011

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*, R25. https://doi.org/10.1186/gb-2010-11-3-r25

Secor, S. M., Stein, E. D., & Diamond, J. (1994). Rapid upregulation of snake intestine in response to feeding: a new model of intestinal adaptation. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, *266*(4), G695–G705. https://doi.org/10.1152/ajpgi.1994.266.4.G695

Secor, S. M. (2008). Digestive physiology of the Burmese python: broad regulation of integrated performance. *Journal of Experimental Biology*, *211*(24), 3767–3774. https://doi.org/10.1242/jeb.023754

Secor, S. M., & Diamond, J. (1995). Adaptive responses to feeding in Burmese pythons: pay before pumping. *Journal of Experimental Biology*, *198*(6), 1313–1325.

Secor, S. M., & Diamond, J. (1998). A vertebrate model of extreme physiological regulation. *Nature*, *395*(6703), 659–662. https://doi.org/10.1038/27131

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smit, A. F. A., & Hubley, R. (2008). RepeatModeler Open-1.0 (Version 1.0.8). Retrieved from http://www.repeatmasker.org

Smit, A. F. A., Hubley, R., & Green, P. (2013). RepeatMasker Open-4.0 (Version 4.0.6). Retrieved from http://www.repeatmasker.org

Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research*, *32*(suppl_2), W309–W312. https://doi.org/10.1093/nar/gkh379

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, *19*(suppl_2), ii215–ii225. https://doi.org/10.1093/bioinformatics/btg1080

Suárez-Atilano, M., Burbrink, F., & Vázquez-Domínguez, E. (2014). Phylogeographical structure within *Boa constrictor imperator* across the lowlands and mountains of Central America and Mexico. *Journal of Biogeography*, *41*(12), 2371–2384. https://doi.org/10.1111/jbi.12372

The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169. https://doi.org/10.1093/nar/gkw1099

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline: The Genome Analysis Toolkit Best Practices Pipeline. In A. Bateman, W. R. Pearson, L. D. Stein, G. D. Stormo, & J. R. Yates (Eds.), *Current Protocols in Bioinformatics* (pp. 11.10.1-11.10.33). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471250953.bi1110s43

Vicoso, B., Emerson, J. J., Zektser, Y., Mahajan, S., & Bachtrog, D. (2013). Comparative Sex Chromosome Genomics in Snakes: Differentiation, Evolutionary Strata, and Lack of Global Dosage Compensation. *PLOS Biology*, *11*(8), e1001643. https://doi.org/10.1371/journal.pbio.1001643

Waterhouse, R. M., Tegenfeldt, F., Li, J., Zdobnov, E. M., & Kriventseva, E. V. (2013). OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research*, *41*(D1), D358–D365. https://doi.org/10.1093/nar/gks1116

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simão, F. A., Ioannidis, P., … Kriventseva, E. V. (2017). OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research*, *45*(D1), D744–D749. https://doi.org/10.1093/nar/gkw1119

Chapter 4.

Phylogeographic and population genetic analyses reveal multiple species of *Boa* and independent origins of insular dwarfism

Daren C. Card[1], Drew R. Schield[1], Richard H. Adams[1], Andrew B. Corbin[1], Blair W. Perry[1], Audra L. Andrew[1], Giulia I. M. Pasquesi[1], Eric N. Smith[1], Tereza Jezkova[2], Scott M. Boback[3], Warren Booth[4], and Todd A. Castoe[1]

[1] Department of Biology, 501 S. Nedderman Drive, University of Texas at Arlington, Arlington, TX, 76019, USA.

[2] Department of Ecology & Evolutionary Biology, University of Arizona, P.O. Box 210088, Tucson, AZ, 85721, USA.

[3] Department of Biology, P.O. Box 1773, Dickinson College, Carlisle, PA, 17013, USA.

[4] Department of Biological Science, 800 South Tucker Drive, University of Tulsa, Tulsa, OK, 74104, USA.

ABSTRACT

Boa is a neotropical genus of snakes historically recognized as monotypic despite its expansive distribution. The distinct morphological traits and color patterns exhibited by these snakes, together with the wide diversity of ecosystems they inhabit, collectively suggest that the genus may represent multiple species. Morphological variation within *Boa* also includes instances of dwarfism observed in multiple offshore island populations. Despite this substantial diversity, the systematics of the genus *Boa* has received little attention until very recently. In this study we examined the genetic structure and phylogenetic relationships of *Boa* populations using mitochondrial sequences and genome-wide SNP data obtained from RADseq. We analyzed these data at multiple geographic scales using a combination of phylogenetic inference (including coalescent-based species delimitation) and population genetic analyses. We identified extensive population structure across the range of the genus *Boa* and provide multiple lines of support for three widely-distributed clades roughly corresponding with the three primary land masses of the Western Hemisphere. We also find both mitochondrial and nuclear support for independent origins and parallel evolution of dwarfism on offshore island clusters in Belize and Cayos Cochinos Menor, Honduras.

INTRODUCTION

Widespread, generalist species are powerful model systems for understanding how diverse ecological factors may drive regional patterns of species divergence and diversification (e.g., (Brouat, Chevallier, Meusnier, Noblecourt, & Rasplus, 2004; Fields, Reisser, Dukić, Haag, & Ebert, 2015; Hull, Hull, Sacks, Smith, & Ernest, 2008). The snake family Boidae includes several examples of such systems, with species occupying wide distributions and encompassing a broad range of latitudes, altitudes, and ecosystems (Henderson, Waller, Micucci, Puorto, & Bourgeois, 1995). Modern Boid snake distributions are the result of numerous vicariance events associated with the fragmentation of Gondwana, and thus these snakes have been cited as a classic example of the role that plate tectonics plays in shaping species distributions (Bauer, 1993; Noonan & Chippindale, 2006a, 2006b; Rage, 1988, 2001). Recent studies have also examined the phylogenetic relationships among certain Boid lineages, and collectively have identified evidence for previously unrecognized diversity (Colston et al., 2013; Hynková, Starostová, & Frynta, 2009; Reynolds, Niemiller, & Revell, 2014; Suárez-Atilano, Burbrink, & Vázquez-Domínguez, 2014).

Boa constrictor, the sole species historically comprising the monotypic genus *Boa*, occurs almost continuously from southern South America through northern Mexico. Multiple studies have placed *Boa constrictor* as sister to the Neotropical clade containing *Corallus*, *Eunectes*, and *Epicrates* (Burbrink, 2005; Noonan & Chippindale, 2006a). Numerous subspecies have been described, yet there have been substantial differences in taxonomic recognition among studies. Mainland subspecies include *B. c. amarali* (Bolivia, Paraguay, and southern Brazil; Stull, 1932), *B. c. constrictor* (South America), *B. c. eques* (Piura, Peru; Eydoux et al., 1841), *B. c. imperator* (Central and North America; Daudin, Buffon, Daudin, Sève, & Sonnini, 1802), *B. c. longicauda* (Tombes, Peru; Price & Russo, 1991), *B. c. melanogaster* (Ecuador; Langhammer, 1983), *B. c. occidentalis* (Argentina and Bolivia; Philippi, 1873), and *B. c. ortonii* (northwest Peru; Cope, 1877). In addition to mainland taxa, multiple island populations have been identified as distinct subspecies, including *B. c. nebulosa* (Lazell, 1964) from Dominica, *B. c. orophias* (Linné, 1758) from St. Lucia, *B. c. sabogae* (Barbour, 1906) from the Pearl Islands of Panama, and *B. c. sigma* (Smith, 1943) from the Tres Marías islands of Mexico. These subspecies are mostly recognized based on approximate geographic range and morphological traits (O'Shea, 2007). The Argentine boa (*B. c. occidentalis*), for instance, tends to be dark-colored or black, with white patterning; this color combination is quite distinct from other subspecies. Striking color morphs are also found among island subspecies (e.g., hypomelanism in *B. c. sabogae*) and populations. Much of the diversity in *B. constrictor* color and pattern morphs is known, mostly anecdotally, from the pet trade, where these snakes are popular. Moreover, while mainland *B. c. imperator* in Central and Northern America are long and large-bodied, several Central American islands consist of populations composed entirely of dwarfed individuals (e.g., Cayos Cochinos and Crawl Cay). Limited work with these populations (i.e., common garden experiments) and knowledge from the pet trade indicates that the dwarfed phenotype is heritable and apparently coincides with a shift towards arboreality likely driven by selection imposed by the availability of migratory birds, a primary food source for the snakes on these small islands (Boback, 2005, 2006; Boback & Carpenter, 2007).

Despite examples of morphologically and geographically distinct *B. constrictor* populations, population-level analyses of the species have been entirely lacking until recently. Hynková *et al*. (2009) used data from the mitochondrial cytochrome B locus and found evidence of two major clades, one restricted to South America and one comprising populations in Central and North America. Reynolds *et al*. (2014) used multiple mitochondrial and nuclear genes from two invasive Puerto Rican samples (also examined in the context of mainland populations by Reynolds *et al*. (2013) to further examine the genus *Boa*. This resulted in the splitting of *B. constrictor* (*sensu lato*) into two species: *B. constrictor* from South America and *B. imperator* from Central and North America. Suárez-Atilano *et al*. (2014) identified two additional distinct clades in Northern-Central America using dense sampling and data from two genes (mitochondrial cytochrome b and nuclear ornithine decarboxylase) and 10 microsatellites. Given these suggestions of unrecognized species within the genus, and the recently variable taxonomy of the group, we refer to all populations in the genus *Boa* (*B. constrictor*, *sensu lato*) as the *Boa* complex hereafter. Despite this recent progress, major gaps in our knowledge of the diversification of the *Boa* complex remain, as previous studies have lacked robust population-level sampling across the entire distribution, and from Central American island populations in particular. Furthermore, conclusions from previous studies were also limited to relatively small sets of molecular markers and were based largely on mitochondrial gene sequences.

Here we explore population genetic boundaries, population structure, and phylogenetic relationships across the *Boa* complex, with a focus on Northern-Central American populations that remain taxonomically unresolved, including expanded sampling from multiple dwarfed island populations. We used both mitochondrial and nuclear SNP datasets to address four major aims: (1) to characterize the degree of congruence between genetic markers (mitochondrial versus nuclear) in defining lineages of *Boa*; (2) to determine the number of species that should be recognized within the genus *Boa*; (3) to understand the fine-scale population structure and genetic diversity existing among *Boa* lineages and quantify levels of gene flow that may exists between major *Boa* clades; and (4) to investigate the potential for independent origins of dwarfism in a number of *Boa* island lineages.

MATERIALS AND METHODS

Population sampling and DNA extraction

We extracted DNA from seventy-seven *Boa* samples that were obtained from one of three sources: (1) preserved tissues from vouchered specimens at the University of Texas at Arlington

Amphibian and Reptile Diversity Research Center; (2) blood or scale samples obtained from wild-caught individuals (and progeny) from Belize that are maintained in a colony at Dickinson College; and (3) shed skin samples from commercial breeders with confident provenance (see Supplementary Tables 1-2 for details). DNA was extracted from blood or tissue using either a Zymo Research Quick-gDNA Miniprep kit (Zymo Research, Irvine, CA, USA) according to the manufacturer's protocol or a standard phenol-chloroform-isoamyl alcohol extraction.

Mitochondrial locus amplification and sequencing

Primers L14910 and H16064 (Burbrink, Lawson, & Slowinski, 2000) were used to amplify the mitochondrial cytochrome b gene (cyt-b; 1112 bp). Cycling conditions included 40 cycles with a 45°C annealing temperature and standard *Taq* polymerase (New England BioLabs Inc., Ipswitch, MA, USA). PCR products were visualized using gel electrophoresis and purified using Agencourt AMPure XP beads (Beckman Coulter, Inc., Irving, TX, USA) according to manufacturer's protocols. Sanger sequencing reactions were conducted using ABI BigDye, and visualized on an ABI 3730 capillary sequencer (Life Technologies, Grand Island, NY, USA) using the amplification primers.

Forward and reverse sequence chromatographs for individual samples were aligned and quality trimmed using Geneious 6.1.6 (Biomatters Ltd., Auckland, NZ). New sequences were combined with previously published cyt-b sequences for *Boa* (Hynková et al., 2009; Suárez-Atilano et al., 2014; see Supplementary Table 2 for full details on sampling) and outgroup species obtained from GenBank (see Supplementary Table 3). Mitochondrial nucleotide sequences for all samples were aligned using Muscle v. 3.8.31 (Edgar, 2004), with manual adjustments and trimming to exclude samples with sequence lengths shorter than 500 bp. We also excluded samples with uncertain localities from GenBank based upon descriptions in Hynková *et al*. (2009). The samples included in individual analyses described below are indicated in Supplementary Table 4.

RADseq data preparation and sequencing

Forty-nine samples from North and Central American and two samples from South American populations were sequenced using double digest Restriction-site Associated DNA sequencing (RADseq hereafter), using the protocol of Peterson *et al*. (2012). *Sbf*I and *Sau*3AI restriction enzymes were used to digest genomic DNA, and double-stranded adapters containing unique barcodes and unique molecular identifiers (UMIs; eight consecutive random nucleotides prior to the ligation site) were ligated to digested DNA per sample. Following adapter ligation, samples were pooled into groups of eight and were size selected for fragments ranging from 590 to 640bp using the Blue Pippin (Sage Science, Beverly, MA, USA); this size range was chosen to target roughly 20,000 loci, based on preliminary estimates from an *in silico* digestion of the *Boa constrictor* reference genome (Bradnam et al., 2013). Sub-pools were pooled again based on quantification of samples on a Bioanalyzer (Agilent, Santa Clara, CA, USA) using a DNA 7500 chip. Final pools were sequenced using 100 bp paired-end reads on an Illumina HiSeq 2500 (Illumina Inc., San Diego, CA, USA).

RADseq data analysis and variant calling

Raw Illumina reads from RADseq library sequencing were first filtered using the clone_filter program from the Stacks pipeline (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011; Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013), which excludes PCR replicates using the UMIs, which were subsequently trimmed away using the FASTX Toolkit trimmer v. 0.0.13 (Gordon & Hannon, 2010). Trimmed reads were processed using the process_radtags function with the “rescue” feature activated in Stacks, which parses reads by barcode, confirms the presence of restriction digest cut sites, and discards reads lacking these features. Parsed reads were quality trimmed using Trimmomatic v. 0.32 (Bolger, Lohse, & Usadel, 2014) and were aligned to the reference *B. constrictor* genome (Assemblethon2 team SGA assembly; Bradnam et al., 2013) using BWA v. 0.7.9 (Li & Durbin, 2009) with default settings (see Supplementary Table 5 for information on the number of quality-filtered and mapped reads). We identified single nucleotide polymorphisms (SNPs) using SAMtools and BCFtools v. 1.2 (Li, 2011; Li et al., 2009). We used default parameters for SNP calling (ignoring indels) and used VCFtools v. 0.1.14 (Danecek et al., 2011) to construct a stringently filtered dataset where sites were excluded that did not have a minimum Phred score of 20, that had >2 alleles per individual, that possessed a minor allele frequency <5%, or that contained >25% missing data across individuals after low confidence genotypes (Phred score < 20) were coded as missing data. This dataset was further filtered such that only the first SNP within a 50 kb window was used, to adhere to model assumptions in downstream analyses regarding independence of SNPs. This stringently filtered SNP dataset contained 1,686 SNPs and we used custom Python and R scripts to format datasets for several downstream analyses.

Estimating phylogenetic relationships and divergence times across Boa

We used the cyt-b alignment to estimate phylogenetic relationships and infer divergence times among *Boa* lineages using a fossilized birth-death model. This model removes the need for *a priori* node constraints and infers divergence times by integrating fossil dates into the lineage diversification and extinction model (Heath, Huelsenbeck, & Stadler, 2014; Stadler, 2010). This model was implemented in BEAST v. 2.2.1 (Bouckaert et al., 2014) using the Sampled

Ancestors add-on package (Gavryushkina, Welch, Stadler, & Drummond, 2014). Fossils and associated dates (the average of the minimum and maximum dates in the age range) were acquired from the Paleobiology Database (http://paleobiodb.org), PaleoDB (http://paleodb.org), and from previous estimates of Boid divergence dates (Colston et al., 2013; Noonan & Chippindale, 2006, 2006; Suárez-Atilano et al., 2014; see Supplementary Table 6 for full details). We specified a strict molecular clock and an HKY nucleotide substitution model with no codon partitioning to ensure proper mixing and convergence after experimenting with more complex models that showed signs of poor mixing and convergence. We performed the analysis using a total of 2.5 x 10^8 MCMC generations, sampling every 5000 generations, and discarded the first 20% as burn-in, based on likelihood stationarity visualized using Tracer v. 1.6 (Drummond & Rambaut, 2007). Phylogenetic trees were visualized and manipulated in R v. 3.2.0 (R Core Team, 2015) using the ape v. 3.3 (Paradis, Claude, & Strimmer, 2004) and strap v. 1.4 (Bell & Lloyd, 2014) packages.

To further characterize the relationships among mitochondrial haplotypes and their frequencies within our dataset, we constructed a median-joining haplotype network using Network v. 4.613 (Bandelt, Forster, & Röhl, 1999). For this analysis, the mitochondrial alignment was further trimmed to eliminate any missing data located at the alignment ends (total alignment length was 878 bp). We used a recommended weighted transition:transversion ratio of 2:1 (per the Network manual) and used the maximum parsimony network method to minimize the number connections among haplotypes.

Mitochondrial estimates of haplotype diversity and inter-clade gene flow among Boa *populations*

We assessed landscape-level patterns of genetic differentiation across the collective geographic range covered by our sampling, and individually on ranges occupied by the three major resolved population clusters (see Results section 3.1 for details). For this analysis we used only mitochondrial samples associated with precisely known collection localities (i.e., localities with geographic coordinate data or reliable descriptions for which coordinates could be well estimated; see Supplementary Table 4 for assignments) and applied a previously described methodology (Jezkova et al., 2015; Schield et al., 2015) that interpolates mitochondrial genetic distances across a geographic landscape and colors geographic regions based on the interpolated level of interpopulation genetic distance.

We used IMa2 (Hey & Nielsen, 2007) to estimate parameters of the isolation-migration model (Hey & Nielsen, 2004) between multiple island and mainland population pairs, and between populations east and west of the Isthmus of Tehuantepec (see Supplementary Table 4 for population assignments). We estimated burn-in to occur prior to 3.75 x 10^6 generations based on trial runs, and our full analyses included a total of 1.5 x 10^7 post burn-in MCMC generations, with sampling every 100 generations, and four independent runs per population comparison. We found these run times to be sufficient based on chain mixing and convergence, and parameter effective sample sizes >1000 for all parameters in each run. We rescaled parameter estimates into demographic units using generation time of three years (Lindemann, 2009) and a mitochondrial mutation rate estimate from Castoe *et al*. (2007).

Population genetic analyses of nuclear SNP data

We estimated the phylogenetic relationships among samples by inferring a maximum likelihood (ML) phylogeny using RAxML v. 8.1.20 (Stamatakis, 2014) with a GTR + Γ nucleotide substitution model with estimated base frequencies and 1000 bootstrap replicates (sensu Cariou *et al*. [2013]). We visualized the resulting phylogeny and assessed bootstrap support using FigTree v. 1.4.2 (Rambaut, 2015).

We used NGSadmix (Skotte, Korneliussen, & Albrechtsen, 2013) and Entropy (Gompert et al., 2014), which are both similar to Structure (Pritchard, Stephens, & Donnelly, 2000), but leverage genotype likelihoods to infer admixture proportions across all samples and to investigate how ancestry may be partitioned under different numbers of assumed source populations (i.e., values of *K* population clusters). We conducted 10 independent runs for each value of *K* ranging from 1 to 11 and used the Δ*K* method (Evanno, Regnaut, & Goudet, 2005) to estimate the highest supported *K* value (i.e., the most likely number of source populations). Parallel runs were summarized using CLUMPP v. 1.1.2 (Jakobsson & Rosenberg, 2007) with the 'greedy' algorithm. Based on these results, we ran Entropy on a more targeted range of *K* from 1 to 8. We ran two MCMC chains for each value of *K* with 15,000 iterations per chain, with sampling every 5 iterations. We eliminated the first 20% of samples as burn-in and confirmed proper mixing and convergence before using Deviance Information Criteria (DIC) to determine the best-supported *K* value.

Based on the inferred genetic clustering of populations provided by NGSadmix and Entropy, we inferred population summary statistics for Central and North America populations. We used Stacks v. 1.34 (Catchen et al., 2011, 2013) to estimate nucleotide diversity (*π*), heterozygosity

(*H*), and the inbreeding coefficient (F_{IS}) at each locus, and determined the total number of private alleles per population. We also compared pairwise allelic differentiation (F_{ST}) between populations. This analysis was performed on a single Stacks-derived dataset (distinct from above-described SNP datasets) that we constructed from mapped RADseq data using the ref_map.pl tool and a minimum stack depth of 3. This dataset was filtered to allow for up to 50% missing data and retained loci with a minimum per-individual stack (i.e., read) depth of 10, resulting in 44,041 RAD loci.

We also tested for nuclear evidence of gene flow between major *Boa* lineages using TreeMix v. 1.12 (Pickrell & Pritchard, 2012). This analysis was conducted using population delineations informed from the results of several inferences (see Results and Supplementary Table 4). We allowed from zero to 12 migration events between lineages and calculated the fraction of the variance in relatedness between populations that is explained by each migration model.

Genome-wide Bayesian species delimitation of Boa

We used a subset of the total RADseq sampling to perform coalescent Bayesian species delimitation analysis (n = 33 samples; Supplementary Table 7). This subset was chosen to exclude individuals that contained higher levels of missing data (e.g., from low numbers of mapped reads), that when excluded did not result in major geographic/phylogenetic sampling gaps. We perform Bayes factor species delimitation using the BFD* method (Leaché, Fujita, Minin, & Bouckaert, 2014) implemented using the SNAPP (Bryant, Bouckaert, Felsenstein, Rosenberg, & RoyChoudhury, 2012) plugin for BEAST2. Overall, we tested three competing species models, including two “two species” models that lump either Central and North American populations (Model A) or Central and South American populations (Model B) into a single monophyletic species, and a third three species model that designates North, Central, and South American populations each as distinct species (Model C; Fig. 6 & Supplementary Table 7). These three models were informed by recent work (Hynková et al., 2009; Reynolds et al., 2014; Suárez-Atilano et al., 2014), and by our mitochondrial and nuclear analyses (see Results sections 1 and 3). For all three species models, we conducted path sampling for a total of 14 steps (100,000 MCMC steps, 10,000 burn-in steps each) to estimate marginal likelihoods for each competing model. Bayes factor support was compared between models to identify the best-supported species model. We visualized the best-supported species tree posterior from the final path sampling step (minus a 10% burn-in) using DensiTree v. 2.2.1 (Bouckaert, 2010).

RESULTS

Mitochondrial patterns of population structure, relationships, and divergence timing

The mitochondrial cyt-b alignment contained 305 total in-group samples and 1059 aligned bases. There were a total of 301 polymorphic sites and 250 total informative sites across the alignment. Phylogenetic inference in BEAST 2 resolved deeper relationships among *Boa* samples with high support (defined as >95% posterior support hereafter), but recent nodes received far less posterior support (Fig. 1). There was high posterior support for a sister relationship between a clade comprising *Boa* samples from Colombia and the remaining populations of *Boa.* Following this basal split, the core *Boa* radiation contains a highly supported split between South and Northern-Middle America (Fig. 1-2). Within the South American clade, there is also high support for two Ecuadorian samples being sister to the rest of the clade. A clade of Argentinian samples is resolved as the sister group to all other remaining samples, which includes individuals from Peru, Brazil, Guyana, and Surinam.

Among Northern-Central American sampling, we found strong support for two mitochondrial clades. One clade includes samples from nuclear Central America, including localities that extend from northern South America through the Isthmus of Panama to the Isthmus of Tehuantepec, and along the Gulf coast of Mexico. The second clade includes samples west of the Isthmus of Tehuantepec, along the Pacific coast of Mexico (Fig. 1-2). Samples from Oaxaca, Mexico, located at the boundary between these two clades, fall into both of these two large clades, indicating a potential zone of introgression between these lineages in this region. Among island populations sampled, individuals from the Cay islands of Belize fall within one subclade of the Central American clade, while samples from Cayos Cochinos Menor in Honduras clustered with mainland samples from another subclade within the Central American clade. The split between these two Central American subclades is highly supported (see inset of Fig. 1).

We estimated the oldest split between the *Boa* clade containing Colombian samples and the rest of the *Boa* complex to have occurred almost 20 million years ago (Mya; 95% highest posterior density [HPD] = ca. 16 to 22.4 Mya) with a subsequent split between the North American and Northern-Central American clades occurring approximately 16 Mya (95% HPD = ca. 13.0 to 17.8 Mya). Within the well-resolved South American clade, we estimated the split between the Argentinian clade and its sister lineage to have occurred ca. 8 Mya (95% HPD = ca. 6.2 to 9.9 Mya). Other well-resolved divergences (i.e., > 95% posterior support) within the South American clade ranged from ca. 6 to 2 Mya. The split between the two Northern-Central American clades is estimated to have occurred 14 Mya (95% HPD = ca. 11.6 to 15.9 Mya), with subsequent splits in both lineages ranging from 5 to 10 Mya. The well-supported divergence between the two clades containing dwarfed island populations are estimated to have occurred 5

Mya (95% HPD = ca. 3.6 to 6.1 Mya), and 95% HPD ranges indicate that individual island divergences occurred within the past 1 My (Fig. 1).

Landscape patterns of mitochondrial diversity and admixture across populations

Pairwise mitochondrial genetic distance interpolations highlight several regions across the distribution of the genus *Boa* that contain particularly high genetic diversity. In South America, there is a region of high genetic diversity in Colombia, which coincides with the distribution of a deeply divergent lineage of Colombian *Boa* mitochondrial haplotypes that are sister to all *Boa* lineages in our mitochondrial tree (Fig. 3A). In Central America, regions of northern Honduras contain high average pairwise genetic distances (> 0.02). In North America, areas along the Pacific coast of Mexico also show average pairwise genetic distances higher than 0.02 (Fig. 3A). These results are corroborated by our haplotype network analysis, which indicated high levels of haplotype diversity in the North American and Central American clades overall, including these populations specifically (Fig. 3B). We also found high haplotype diversity within the South American clade. North American populations along the Pacific coast of Mexico show haplotype diversity patterns similar to South American populations, which coincide with the high levels of landscape genetic distances observed in the region (Fig. 3B).

Estimates of gene flow inferred using mitochondrial data and the Isolation-Migration model show evidence of gene flow from mainland populations to islands (approximately 1 – 20 migrants per generation; Supplementary Fig. 1A-C). In contrast, all three mainland-island comparisons provided no evidence of migration from any island to its respective mainland population. We also found no evidence of migrants shared between populations east and west of the Isthmus of Tehuantepec (Supplementary Fig. 1D), which contrasts with the phylogenetic findings that indicate admixture across the isthmus.

Patterns of population structure and relationships from nuclear SNP data

We recovered an average of 1.96 million quality-filtered (1.74 million mapped) Illumina reads per sample (Supplemental Table 5). Overall, three separate analyses – phylogenetic reconstruction using RAxML, admixture analyses from NGSadmix and Entropy, and inferences of population splits and mixtures using TreeMix – provide strong nuclear support for three distinct clades of *Boa*. The maximum likelihood analysis of concatenated SNPs inferred strong support for three major continental clades, mirroring the results from mitochondrial analyses (Fig. 4A), but also revealed considerable intra-clade lineage diversity, including two well-supported clades in Central America that each include island populations. In the Northern-Central American clade, analyses largely confirmed observations from the mitochondrial data. Populations along the Pacific coast of Mexico and Guatemala are distinct from those in the rest of Central America based on phylogenetic results (Fig. 4A). One major discordance between our mitochondrial and nuclear phylogenies, however, was that samples from the Pacific coast of Guatemala phylogenetically clustered with North American samples in Mexico (Fig. 4A), which conflicts with the Central American assignment evident in the mitochondrial phylogeny (Fig. 1).

The Δ*K* test of NGSadmix results supported an optimal model with two source populations, which divides the two Northern-Central American clades, with samples from South America clustering more closely with North America samples (Supplementary Fig. 2; Supplementary Table 8). Similar patterns of population assignment and ancestry proportions were obtained from the results of population clustering using the Bayesian framework implemented in Entropy, though these analyses favor an optimal model of $K = 8$ source populations based on comparisons of DIC values (Fig. 4B; Supplementary Fig. 3; Supplementary Table 8). Results from our population clustering analyses largely agree with phylogenetic results, even as additional source populations are allowed, and assignments to additional population clusters are intuitive given sampling geography (Supplementary Figs. 2-3).

Population allelic differentiation inferred from nuclear SNPs is high between both the South America to Central America, and North America to Central America pairwise population comparisons (mean $F_{ST} = 0.179 \pm 0.300$ standard deviation [SD] and $F_{ST} = 0.133 \pm 0.197$ SD, respectively; Fig. 3C). We examined broad intra-clade genetic diversity in Central America using genome-wide SNP data and found modest levels of nucleotide diversity (mean = 0.136 ± 0.155 SD) and heterozygosity (mean = 0.100 ± 0.143 SD; Fig. 3C). Similar measures were observed in the North American *Boa* clade, as mean (± SD) nuclear nucleotide diversity and heterozygosity were 0.131 (± 0.174) and 0.104 (± 0.173), respectively (Fig. 3C). We found greater levels of nucleotide diversity and heterozygosity in the South American clade (mean = 0.316 ± 0.430 SD and 0.281 ± 0.428, respectively; Fig. 3C) than in either of the northern clades. Inbreeding coefficients were relatively high in Central America (mean = 0.163 ± 0.331 SD) compared to both South America (mean = 0.053 ± 0.229 SD) and North America (mean = 0.077 ± 0.228 SD). We observed 2,059 private alleles in the South American clade, 7,210 private alleles in the Central American clade, and 1,683 private alleles in the North American clade.

We found strong additional support for independent island population establishment (from the mainland) beyond the evidence already presented from phylogenetic and population clustering analyses (see above and Figs. 1 and 3). Moderately high population allelic differentiation (F_{ST}) is evident between pairwise comparisons of island and mainland populations, and varied from an average of 0.045 to 0.058 (Supplementary Fig. 4). F_{ST} estimates are lower between islands in Belize than between any pairwise comparison between islands in Belize and Cayos Cochinos Menor in Honduras (Supplementary Fig. 4), which is consistent with the large geographic distance between these two distinct island systems (ca. 200 km straight-line distance across ocean). Substantially different levels of heterozygosity, nucleotide diversity, and inbreeding coefficients were observed between island and mainland populations, but these intra-population statistics are consistent across island populations (Supplementary Fig. 4). We found that the mainland populations across Central America collectively contained the highest number of private alleles (6,403), while Crawl Cay, Belize and Cayos Cochinos Menor, Honduras contained modest numbers of private alleles (1,847 and 1,248, respectively), and Lagoon and West Snake Cays in Belize contained relatively few private alleles (489 and 270, respectively).

Overall, TreeMix produced a phylogeny which was similar to that based on nuclear phylogenetic analysis (Fig. 5). The amount of variance explained by the model plateaued at M=2 migration events, which explained about 99% of the variance in the dataset. The analysis supported an admixture event from a Central American population to the Guatemalan population of the North American clade. The second supported migration event was from a population ancestral to the Guatemalan population in North America to the mainland population of the Belize clade in Central America (Fig. 5C). These admixture events comprise a high (48%) and low (6%) portion of the recipient population ancestry, respectively (Fig. 5C).

Results of Bayesian species delimitation

Marginal likelihood estimation and Bayes factor comparison of three competing species models found strong statistical support for a three species model that delineated *Boa* samples into a North American, Central American, and South American species (Model C, *ln*(Marginal Likelihood) = -34,278.01; Fig. 6). These three species designations largely coincide with our phylogenetic and population genetic analyses that show substantial lineage independence and divergence of these clades.

DISCUSSION

Evidence for extensive lineage diversity and three species of Boa

Our results provide evidence from both mitochondrial and nuclear data that there are at least three well-differentiated species within the genus *Boa.* These three lineages correspond approximately to the three major landmasses of the Western Hemisphere inhabited by boas: North America (the Pacific coast of Mexico), Central America (including the Gulf coast of Mexico), and South America (Fig. 1-2). Mitochondrial data indicate a sharp division between individuals in the South American and Central American clades that appears to occur at the junction of lower Central America and South America. The transition from the Central American to the North American clade appears to be more diffuse, as mitochondrial haplotypes near the Isthmus of Tehuantepec in Oaxaca, Mexico fall in both the Central American and North American clades, suggesting potential gene flow between clades in this region. These same general patterns have been observed by Hynková *et al*. (2009) and, with much greater resolution, by Suárez-Atilano *et al*. (2014), whose mitochondrial datasets have been included in our own analyses. With our additional sampling of this region, we observed similar patterns and find additional evidence of mitochondrial admixture localized to areas surrounding the Isthmus of Tehuantepec.

Our nuclear SNP sampling, although geographically focused on Central America and Mexico, provides further support for three distinct species-level lineages of *Boa*. Our maximum likelihood analysis of the concatenated SNP alignment yielded a similar topology to that obtained from the more geographically well-sampled mitochondrial data (except for the deep divergence of some Colombian lineages from the mitochondrial data, discussed below). Multiple genetic clustering analyses indicate that at least three major genetic clusters exist within *Boa*, including a strong distinction between central-northern Mexican and Central American populations consistent with our North American and Central American mitochondrial clades. Based on our nuclear SNP data and mitochondrial IMa2 results, we found minimal evidence of admixture between lineages on either side of the Isthmus of Tehuantepec, which was somewhat surprising given indications of admixture from the mitochondrial data and previous results from microsatellites presented by Suárez-Atilano *et al*. (2014). Landscape diversity estimates based on mitochondrial data also indicate a pattern of high diversity in this region, highlighting the confluence of two highly distinct lineages there. Additional investigation with greater sampling from this region would help to establish the extent to which these populations are introgressing and the precise geographic boundaries of this apparent admixture zone.

A recent formal taxonomic revision of *Boa constrictor* (*sensu lato*) was conducted by Reynolds *et al*. (2014) in the context of a broad scale analysis of all Boid and Pythonid snakes. In their study they used two pet trade individuals from Puerto Rico, which had previously been examined in a continental context (Reynolds et al., 2013), to split the genus *Boa* into *B. constrictor* and *B.*

imperator. Our sampling encompassed these two samples, and interestingly, we find that one individual clusters with the enigmatic *Boa* mitochondrial lineage containing samples from Colombia, which is sister to all other populations of *Boa* in our mitochondrial trees. The second sample, however, clusters with samples from Mexico. The fact that these samples, and many from Hynková *et al*. (2009), were from the pet trade is problematic because true sample provenance may be unclear or possibly erroneous. Nonetheless, the finding some Colombian samples form a lineage sister to all other boa in the mitochondrial phylogeny populations requires further investigation to determine if these are indeed mitochondrial sequences (versus nuclear inserts of mitochondrial genes; NUMTs; see (Hazkani-Covo, Zeller, & Martin, 2010) for a review), deep coalescence of ancient mitochondrial haplotypes, or if these populations do indeed represent a fourth divergent lineage of *Boa*. These questions, however, fall outside the scope of the present study due to a lack of high-quality samples with known locality data from Colombia. Future studies that incorporate nuclear SNP sampling for Colombian and other South American samples would be valuable for further investigating patterns of *Boa* diversity.

Given both our nuclear and mitochondrial results, as well as previous work indicating the likelihood of multiple species-level lineages of *Boa* (Hynková et al., 2009; Reynolds et al., 2014; Suárez-Atilano et al., 2014), we were interested in explicitly testing three alternative models of species recognition for *Boa* lineages. Bayes Factor delimitation of the genome-wide SNP dataset rejected both of the alternative two species hypotheses that lumped either Central and North American clades (Model A) or Central and South American clades (Model B) into single species. Instead, Bayes factor comparisons overwhelmingly supported a three species model for the genus *Boa* in which North, Central and South American clades each represent distinct species (Model C). These results are highly consistent with our analyses of mitochondrial and nuclear variation and provide yet another level of support for the recognition of at least three species within the genus *Boa*.

Both our mitochondrial and nuclear analyses indicate that taxonomic revisions are necessary within the genus *Boa*. This genus has previously been recognized as monotypic, *Boa constrictor*, with 7 recognized subspecies (Uetz & Etzold, 1996; Uetz, Hošek, & Hallermann, 2015). Based on mitochondrial data, Reynolds *et al*. (Reynolds et al., 2014) elevated the subspecies *B. c. imperator*, comprising populations in Central and North America, to *B. imperator*. This change was previously suggested by Hynková *et al*. (2009). Suárez-Atilano *et al*. (2014) described greater population diversity and divergence across North and Central American populations, and concluded that the two major lineages in this region comprise evolutionary significant units, though did not make taxonomic recommendations. Our population clustering analyses, phylogenetic inference, and coalescent-based species delimitation methods spanning both mitochondrial and nuclear datasets provide multiple lines of evidence for three major lineages within the genus *Boa*. We recognize the South American lineage as *B. constrictor* and the Central American lineage (including South American populations in the Choco of Colombia and Ecuador [and probably Peru], and North American populations along the Gulf coast of Mexico [west of the Isthmus of Tehuantepec]) as *B. imperator*, in line with previous taxonomic discussions (Hynková et al., 2009; Reynolds et al., 2014; Suárez-Atilano et al., 2014). We recognize the North American lineage, comprising Mexican populations along the Pacific coast west of the Isthmus of Tehuantepec, as *B. sigma* (Smith, 1943). The taxon *Constrictor c. sigma* was described based on three specimens from María Madre Island, Tres Marías Islands, Nayarit, Mexico by Smith (Smith, 1943); types: CAS 58681, USNM 24672 46484 [holotype]). The description notes that this population has the highest ventral counts of any other *Boa* population in Mexico, this character difference serving as diagnostic for the new taxon. Smith apparently was unaware that Slevin (Slevin, 1926) had mentioned the presence of the same taxon for María Magdalena Island. Zweifel (1960) reported on an American Museum expedition to the Tres Marías Islands and found seven more individuals, including specimens from María Madre, María Magdalena, and María Cleofas. In this publication Zweifel argues for the recognition of *B. c. sigma* as a junior synonym of *Boa c. imperator* based on expanded variation of ventral scale counts in the Tres Marías populations, which overlaps that found on the mainland (253 – 260 vs. 225 – 253 in the mainland of Mexico [including Pacific and Atlantic populations]). The Tres Marías population barely overlaps with the mainland in ventral counts, by one in nine specimens versus 41 from the mainland (given by Smith [1943]). Although, we lack genetic sampling from the Tres Marías Islands, given our finding of a distinct species found in Western Mexico, *B. sigma* is the only available name we can unambiguously apply to a population within this North American lineage. Our taxonomic recommendation is to recognize *B. sigma* (Smith, 1943) as full species, encompassing the Western Mexico lineage. Finally, we acknowledge that further population-level investigations and analyses of morphology should be conducted to reinforce this recommendation.

Divergence time estimates and historical biogeography

Boid snakes in general, and the genus *Boa* in particular, are considered to be South American in origin, based on Gondwanan vicariance models of boine biogeography (e.g., Noonan & Chippindale [2006, 2006]), which are also consistent with early boid fossils from Colombia (Head et al., 2009) and a highly diverse boid radiation in South America (Burbrink, 2005; Noonan & Chippindale, 2006). Using the newly-developed FBD model of divergence time estimation, we estimated the divergence between South American and Northern-Central American lineages at approximately 16 Mya (95% HPD = ca. 13.0 to 17.8 Mya), well earlier than findings from Suárez-Atilano *et al.* (2014), which place the split at 7.4 Mya (95% HPD = ca. 6.2 to 9.9 Mya). This divergence time substantially predates the historically recognized date of the closure of the Isthmus of Panama (estimated to occur ca. 5 Mya; Haug & Tiedemann, 1998; Haug, Tiedemann, Zahn, & Ravelo, 2001; Keigwin, 1982; Ravelo, Andreasen, Lyle, Olivarez Lyle, & Wara, 2004), but also falls prior to a newly articulated date for the closure of the Isthmus of Panama (13 – 15 Mya; Montes et al., 2015). This suggests that boas may have successfully colonized Central America before the Panamanian land bridge was formed, an inference that is consistent with a Miocene *Boa* fossil known from Panama that was dated at 19.3 Mya (Head, Rincon, Suarez, Montes, & Jaramillo, 2012). Similarly, divergence times between other major *Boa* clades are also older than in previous estimates, as the split between the two major Northern-Central American clades is estimated to have occurred shortly after boas presumably colonized this landmass, at approximately 14 Mya (95% HPD = ca. 11.6 to 15.9 Mya). It is notable that this split may represent two coastal expansion fronts that moved northward through Central America, which were isolated by transcontinental mountain ranges. Even within the Central American clade, we find relatively deep divergences (ca. 5 – 10 Mya) among subclades, and thus significant population diversity that may warrant further investigation and taxonomic recognition, that indicates a long history of *in situ Boa* evolution in Central America. Lastly, mito-nuclear discordances in phylogenetic (including divergence timing) estimates have been recognized (see Toews & Brelsford [2012] for a review) and divergence estimates from a single gene is known to be difficult (Arbogast, Edwards, Wakeley, Beerli, & Slowinski, 2002; Graur & Martin, 2004), facts that we acknowledge. However, given the concordance between our divergence estimates and limited fossil evidence, we believe our estimates are reasonable and may be even more realistic than the much younger divergence estimates from previous studies of Boid snakes (Noonan & Chippindale, 2006, 2006; Suárez-Atilano et al., 2014).

Support for independent insular dwarfism in Central American Boa

Our results provide evidence that dwarf forms of boas that occur on multiple islands off the coast of Central America – from coastal islands in Belize and on Cayos Cochinos Menor in Honduras – have independent evolutionary origins. With regard to community assembly, this is not surprising, as it has been established that offshore islands are usually populated by the most common mainland species (Burbrink, McKelvy, Pyron, & Myers, 2015). However, it is particularly exciting that the dwarfed phenotype appears to be a product of convergent evolution, whereby similar insular ecosystems have independently selected for similar dwarf phenotypes. Mitochondrial haplotypes of individuals from these two separate island groups cluster within distinct highly-supported clades that are estimated to have diverged from one another approximately 5 Mya. A similar pattern is observed in our nuclear SNP-based phylogeny, where we find strong nodal support for the split between these two larger Central American clades, each of which includes one of the two groups of islands. Patterns observed from our SNP-based population cluster analyses also resolve these two population groups into separate distinct clusters, though there is some evidence of admixture across islands and Central American mainland source populations under various population models that we speculate represents standing genetic variation from the adjacent mainland populations more than recent gene flow (especially between islands).

Isolation-Migration analyses indicate that gene flow between the island and adjacent mainland populations is essentially unidirectional, from mainland to island in each of the two island systems. The broad posterior estimate on gene flow indicates a great deal of uncertainty in the degree of gene flow between island and mainland populations and is likely a product of small sample sizes, data from a single mitochondrial gene, and the confounding effects of multiple historical periods of gene flow and isolated with sea level change. Patterns of diversity in nuclear SNPs also indicate small effective population sizes on these islands that have likely allowed drift to substantially alter allele frequencies to the extent that pairwise allelic divergence (F_{ST}) is quite high between each island and mainland pair. This pattern is consistent with small empirical estimates of population sizes on the Belize islands (Boback, 2005) and on Cayos Cochinos Menor (Reed et al., 2007). Collectively our results support the hypothesis that evolutionary processes, including the evolution of dwarf phenotypes, have occurred in parallel between the two independent island population groups.

While drift is likely driving the majority of genetic differentiation in these island populations, it is likely that a subset of genetic differentiation observed between island and mainland populations may also be due to selection associated with these unique island ecosystems, which includes selection driving the evolution of dwarfism and other specialized phenotypes on these islands (Boback, 2005, 2006; Boback & Montgomery, 2003). Indeed, common garden experiments using dwarfed snakes from several Belize islands indicates that selection has favored genetic changes that are apparently causing dwarfism (Boback & Carpenter, 2007), a scenario also supported by the maintenance and breeding of dwarfed *Boa* from Cayos Cochinos and elsewhere in the pet trade. Beyond these two island systems, *Boa* populations exist on at least 50 near offshore islands (Henderson et al., 1995), and other known (but unsampled) populations of island dwarf populations exist from islands that are more widely geographically separated from those in our study. Collectively, this suggests that there is very likely to be more than two independent instances where island dwarfism evolved, though the proportion explained by genetic underpinnings versus phenotypic plasticity remains to be explored.

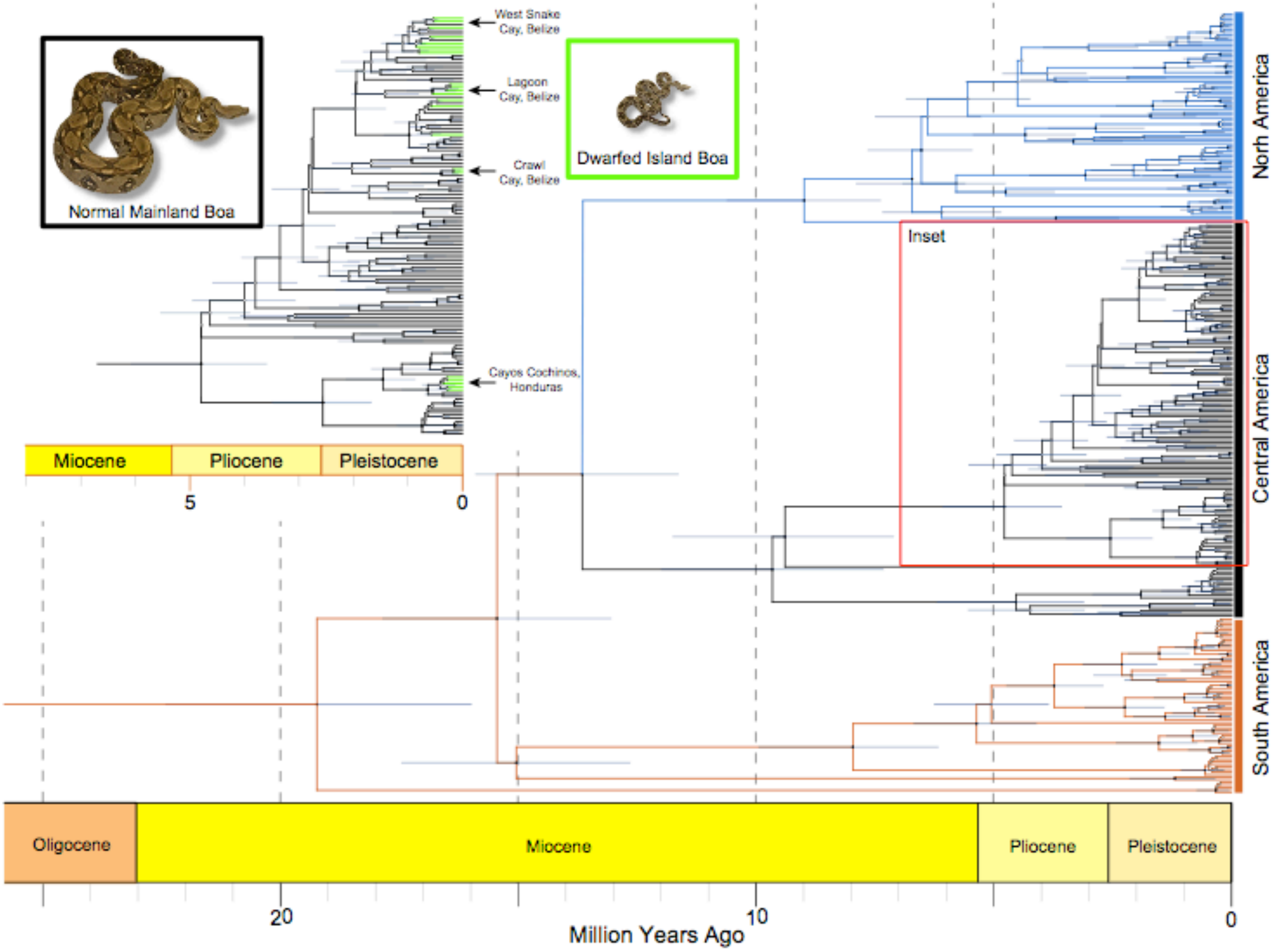
Conclusions

Our genome-wide nuclear and single-locus mitochondrial datasets both identified extensive population structure across the range of the genus *Boa*. Multiple lines of evidence indicate that there are (at least) three widely distributed clades, and each clade roughly corresponds to three major landmasses of the Western Hemisphere – North, Central, and South America. Our data also confirm results and taxonomic suggestions from previous studies, and further warranted the recognition of a third species in the genus *Boa*, *B. sigma*, corresponding to the North American clade. Additional studies using molecular data would be desirable to further test the hypothesis that the Mexican island populations from which the type specimens of *B. sigma* originate (Tres Marías) represent the same taxon as adjacent mainland *Boa* populations. Expanded sampling for South American *Boa* populations, especially those in Colombia where mitochondrial lineage diversity is high, would also be important for addressing outstanding questions about lineage diversity in *Boa*. Lastly, our data suggest two apparently independent instances of the evolution of dwarfism in *Boa* populations inhabiting offshore islands (in Belize and Cayos Cochinos Menor, Honduras) implicating substantial morphological convergence among these populations.

ACKNOWLEDGMENTS

We thank Timothy Colston and Dr. Chad Montgomery for tissue samples and the numerous researchers at the University of Texas at Arlington who collected and deposited *Boa* tissue samples at the Amphibian and Reptile Diversity Research Center, which formed the basis of this work. Funding for fieldwork was provided in part by NSF grants DEB-0613802 and 0102383 to J. A. Campbell (whom we thank for making tissues available), and NSF DEB-0416160 and Bioclon grants to ENS. Collecting permits in Mexico were issued by the Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT) to V. León-Regañon and O. Flores-Villela, and we thank them for their help while working in Mexico. We thank Kyle Shaney, Nicole Hales, and Nicole Proctor for their help in the laboratory, and Chris Nice for assistance with SNP calling and filtering. We thank Mark Liu for the photographs used in Figure 1. Support was provided from startup funds from the University of Texas at Arlington to TAC and from a National Science Foundation Doctoral Dissertation Improvement Grant (DEB 1401747) to TAC and DCC.

FIGURES



Normal Mainland Boa
West Snake Cay, Belize
Lagoon Cay, Belize
Crawl Cay, Belize
Cayos Cochinos, Honduras
Dwarfed Island Boa
Miocene
Pliocene
Pleistocene
5
0
Inset
North America
Central America
South America
Oligocene
Miocene
Pliocene
Pleistocene
20
10
0
Million Years Ago

Figure 1. Phylogenetic patterns of population division within the genus *Boa*. BEAST2 cladogram inferred using the Fossilized Birth-Death model with node bars reflecting the 95% HPD. Branches have been colored and annotated to reflect the broad geographic assignments of the major BCSC clades. The inset figure provides a high resolution view of Central American populations that contain island dwarf populations, with branches to these samples highlighted in bright green. Node symbols are colored according to posterior support: black = >95%, grey = 75% – 95%, white = 50% – 75%, and no symbols = <50%.

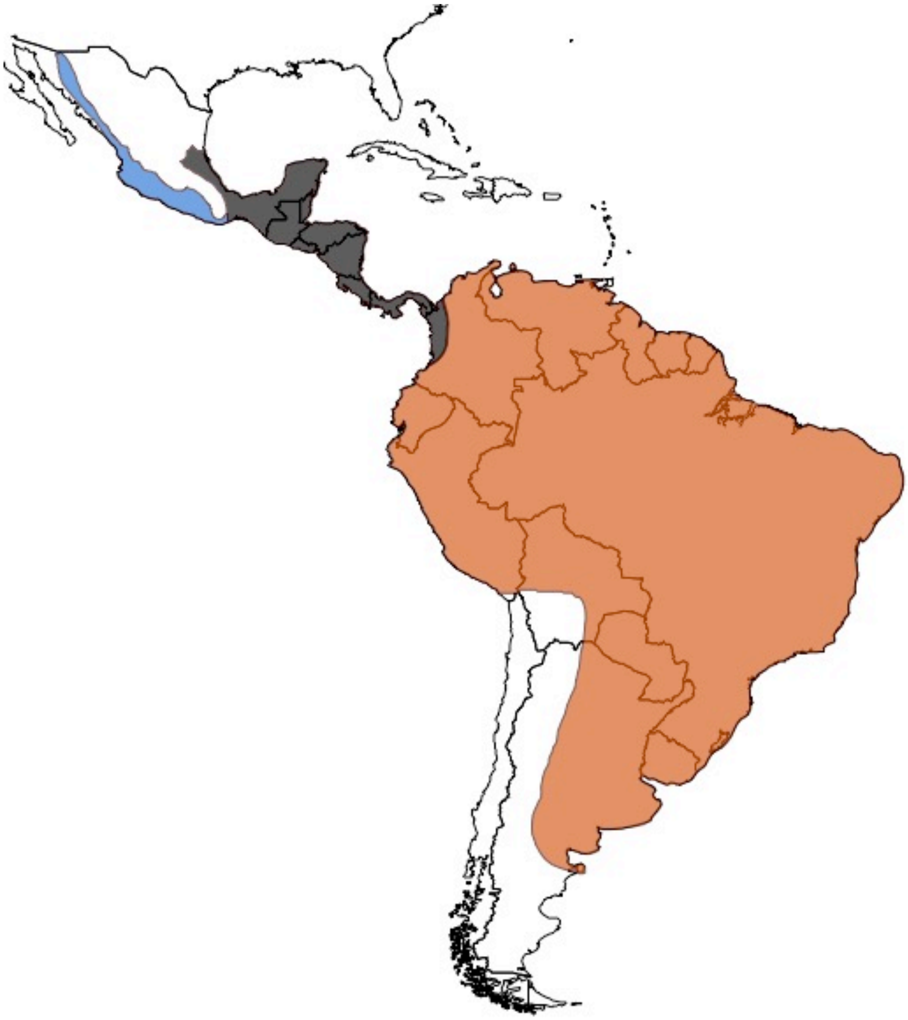


Figure 2. Geographic delimitation of major clades within the genus *Boa*. The three major of *Boa* snakes are localized roughly to the three major New World landmasses: South America, Central America (including parts of Colombia and the Gulf Coast of Mexico), and North America (the Pacific Coast of Mexico to the west of the Isthmus of Tehuantepec). Geographic ranges are colored to correspond to major clades outlined in Fig. 1 and 3.

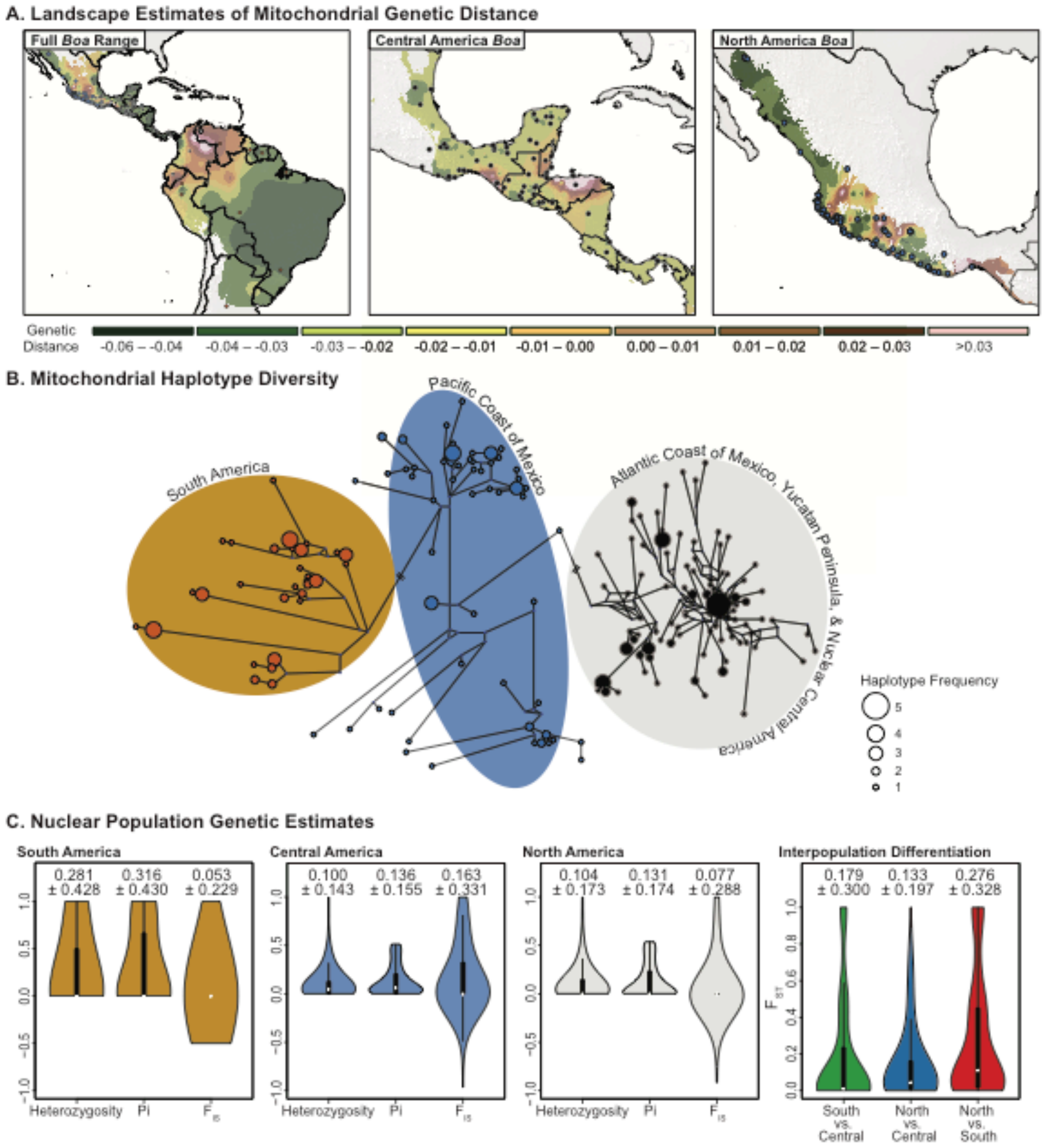


Figure 3. Landscape patterns of mitochondrial genetic diversity and estimates of interpopulation gene flow. (**A**) Residual pairwise mitochondrial genetic distances interpolated across landscape for all *Boa* clades, the Central American clade, and the North American clade. (**B**) Median-joining haplotype network inferred using cyt-b haplotypes, with major geographic assignments indicated. (**C**) Violin plots of genome-wide estimates of interpopulation genetic statistics (*Pi*, *Heterozygosity,* and F_{IS}) for South America, Central America, and North America, and of interpopulation genetic differentiation (F_{ST}) between each pairwise clade. For each violin plot, the white point indicates the median value and the black box indicates the interquartile range. The mean and standard deviations are reported above each respective violin plot.

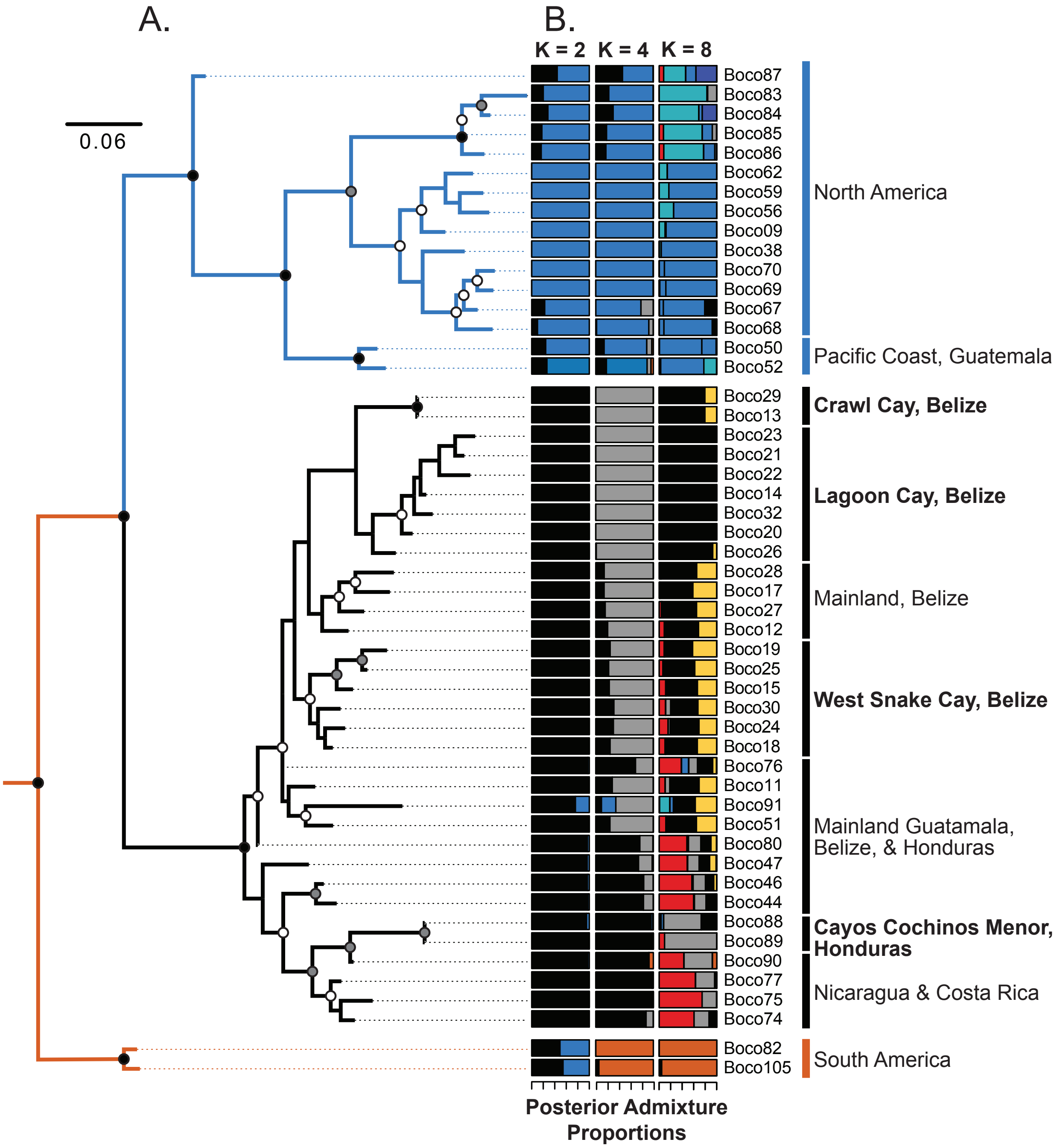


Figure 4. Population structuring and relationships inferred from nuclear RADseq data. (**A**) Maximum likelihood phylogeny inferred from RAxML analysis of the nuclear SNP alignment with a topology, and color annotations, mirroring that of the mitochondrial phylogenies. Nodes symbols are colored according to bootstrap support: black = >95%, grey = 75% – 95%, white = 50% – 75%, and no symbols = <50%. (**B**) Admixture graphs *K* = 2, *K* = 4, and *K* = 8 allowed source populations inferred in Entropy.

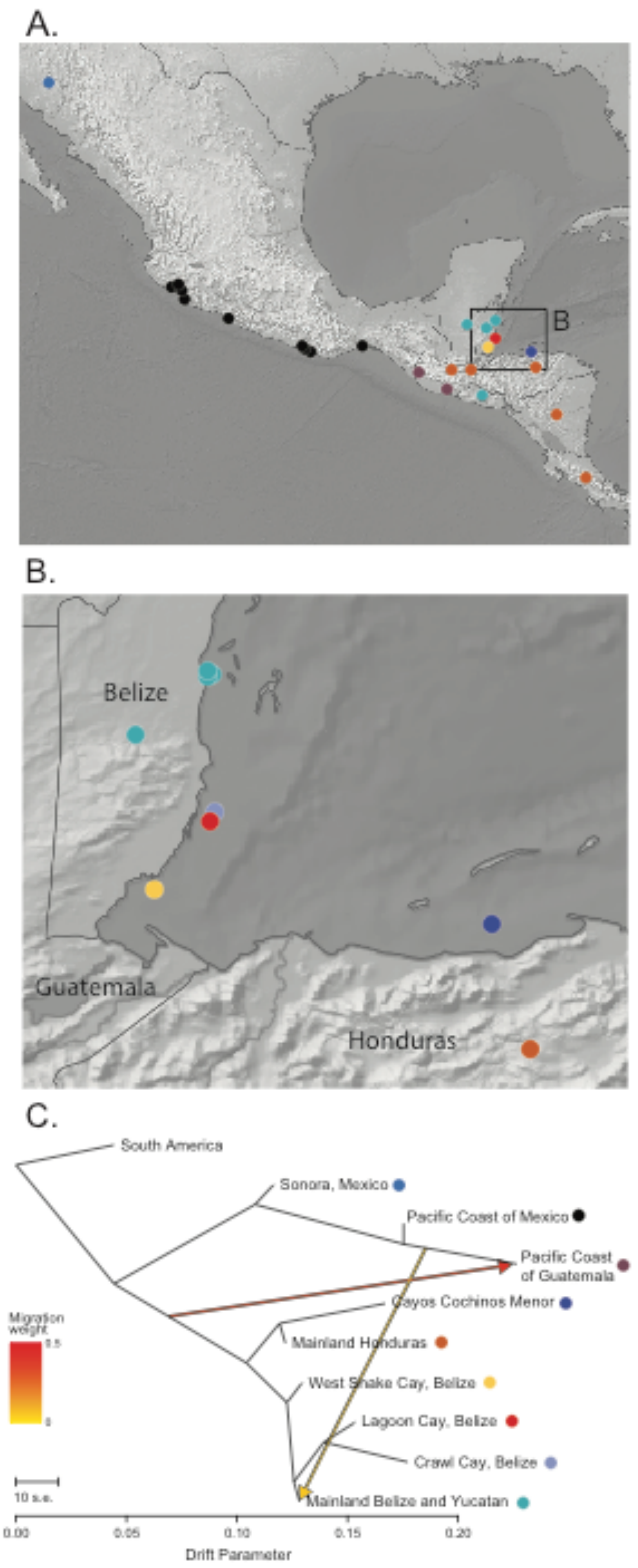

A.
B
B.
Belize
Guatemala
Honduras
C.
South America
Sonora, Mexico ●
Pacific Coast of Mexico ●
Pacific Coast of Guatemala ●
Cayos Cochinos Menor ●
Mainland Honduras ●
West Snake Cay, Belize ●
Lagoon Cay, Belize ●
Crawl Cay, Belize ●
Mainland Belize and Yucatan ●
Migration weight
0.5
0
10 s.e.
0.00
0.05
0.10
0.15
0.20
Drift Parameter

Figure 5. Nuclear patterns of population divergence and gene flow from TreeMix. (**A**) Map of Northern-Central American nuclear sampling with samples color coded by population assignment (inferred from Fig. 4). (**B**) Fine-scale map of island and adjacent mainland sampling. (**C**) TreeMix population tree for the more stringent nuclear SNP dataset, which mirrors the topology observed in Figure 3A. The populations are color coded according to major population assignment in A and B. The drift parameter is ten times the average standard error of the estimated entries in the sample covariance matrix. Migration arrows are colored according to a weight that represents the fraction alleles in the descendent population that originated in the parental population. A model with two migration edges received the highest support – one from the Pacific Coast of Mexico and Guatemala to the Yucatan region and mainland Belize, and one from Central America to the Pacific Coast of Guatemala.

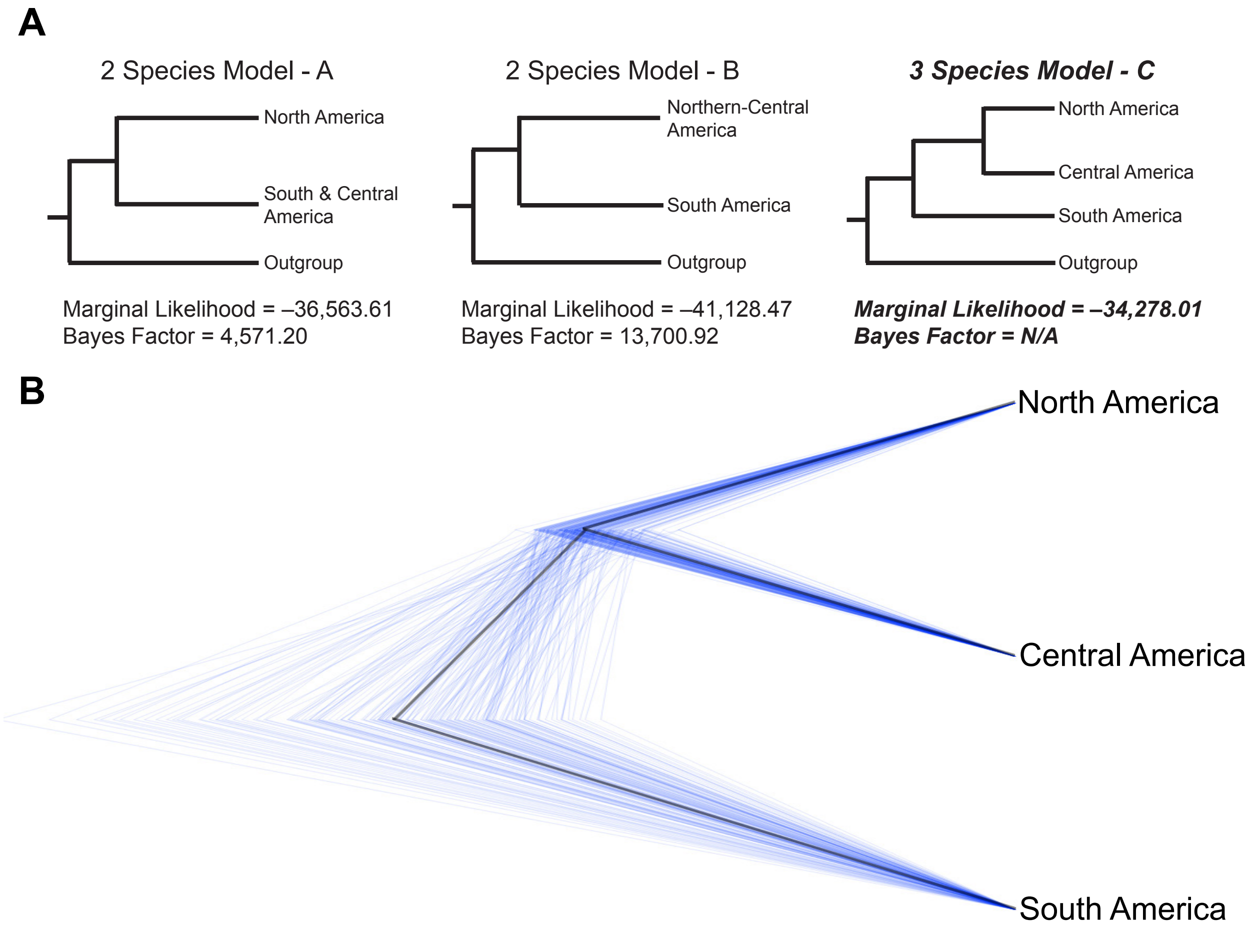
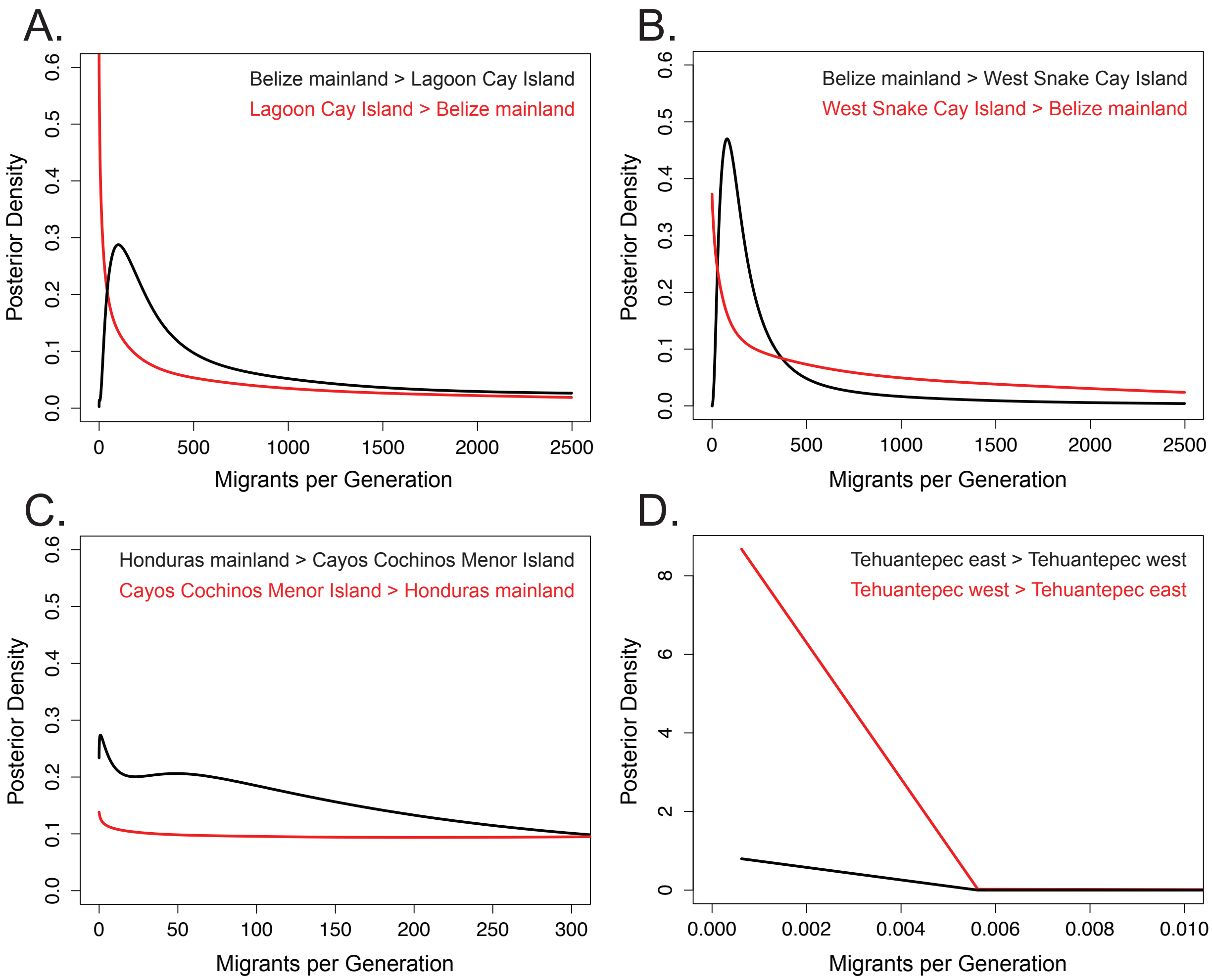
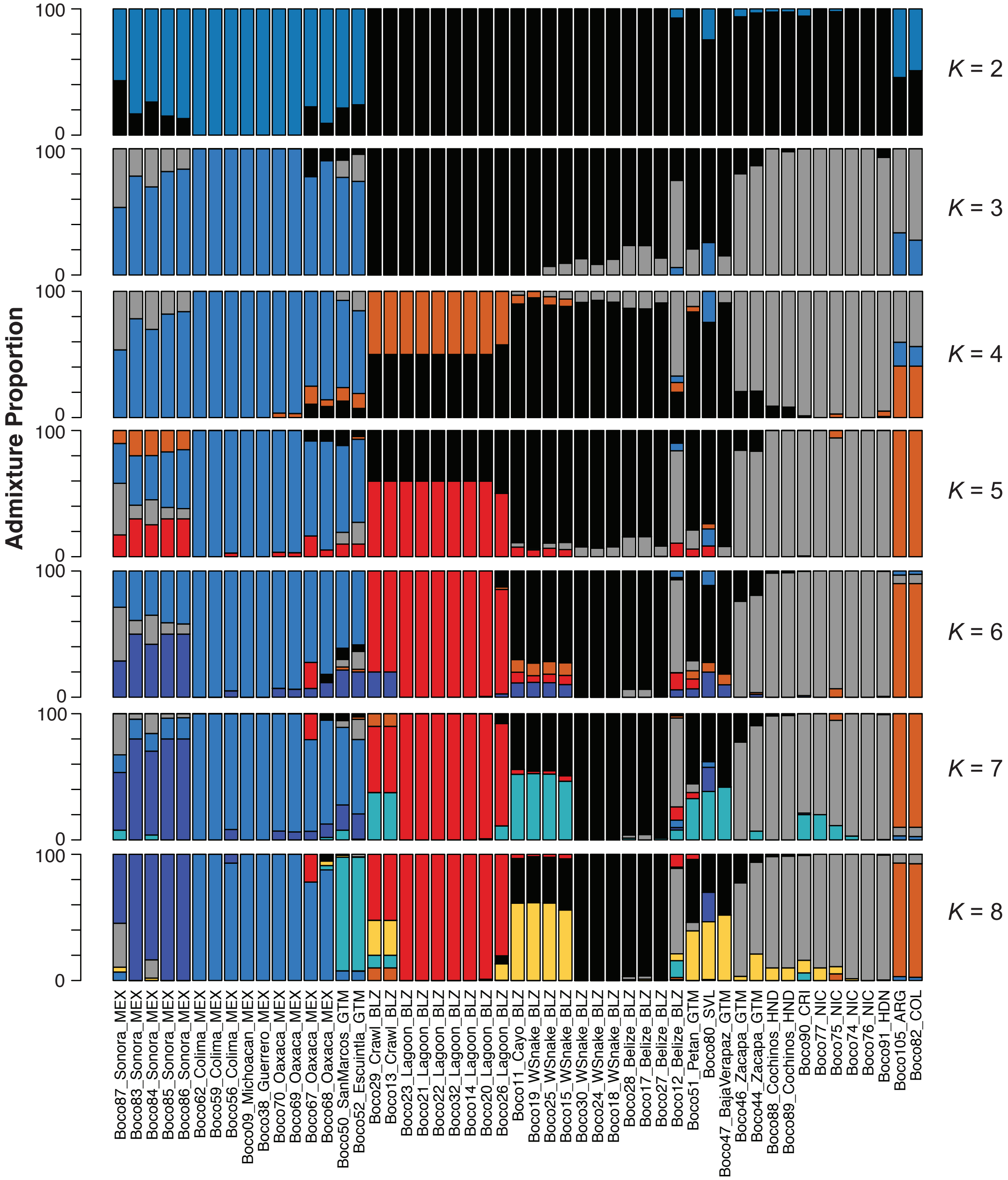


Figure 6. Results from Bayes Factor comparisons of alternative species models. (**A**) Simplified trees showing the species model hypotheses tested using the BFD* framework and the Bayes Factor support obtained under each model. Outgroups are displayed only to aid comprehension and were not incorporated into any of the models. The best-supported species model and associated support values are bolded and italicized. (**B**) DensiTree of posterior estimates of the highest-supported species tree.

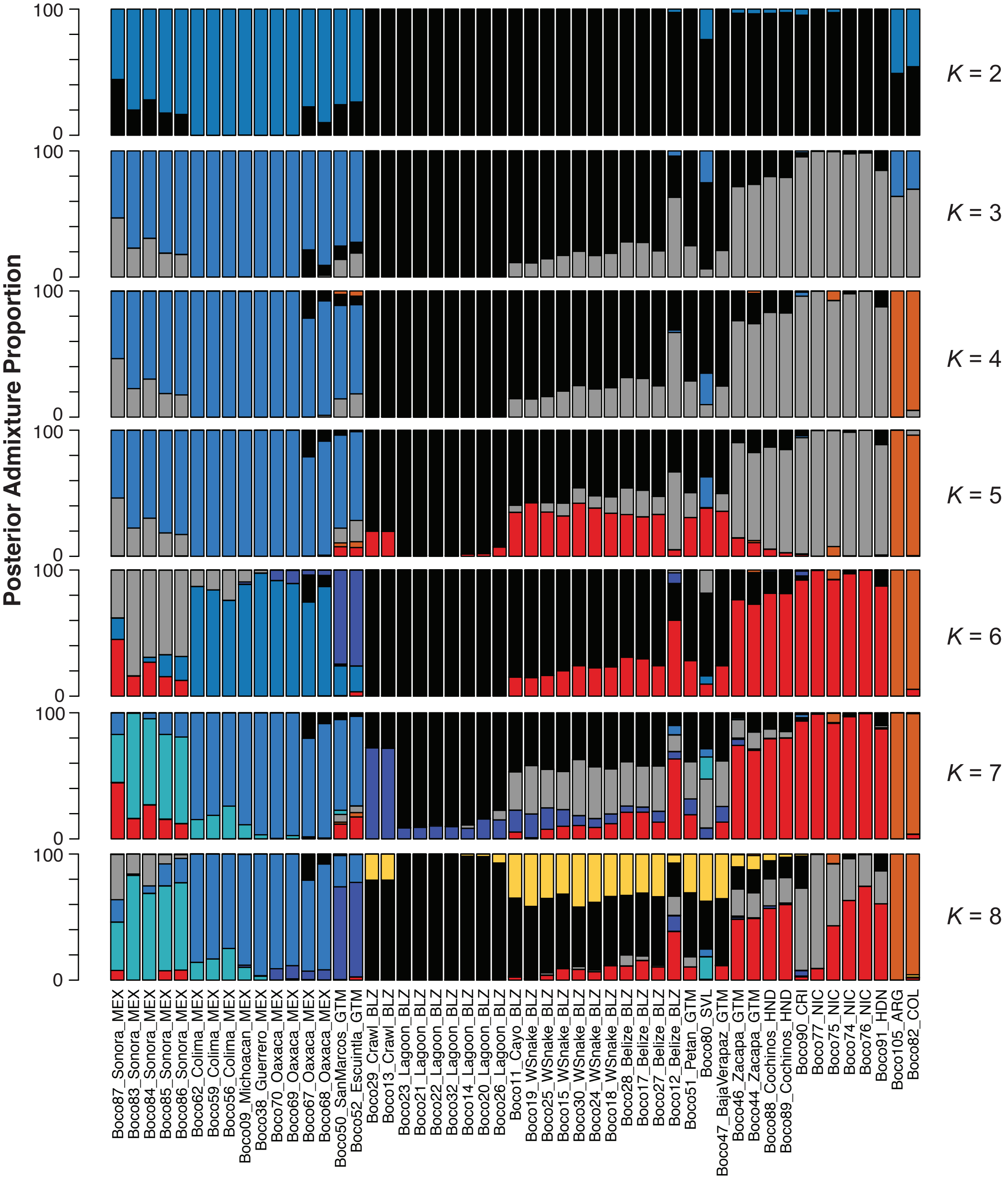
SUPPLEMENTARY MATERIAL



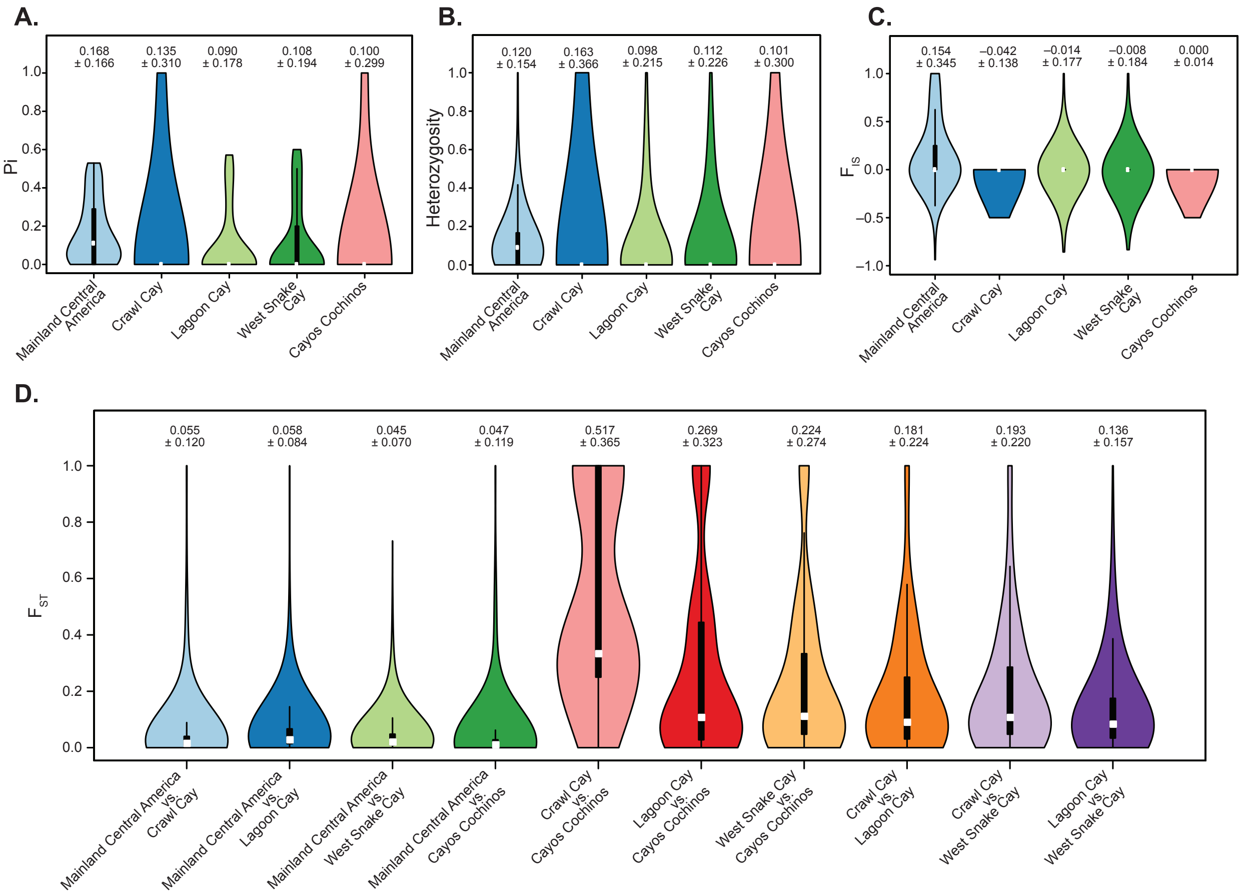
Supplementary Figure 1. Mitochondrial patterns of gene flow based on the Isolation-Migration model. Posterior density estimates of reciprocal migration parameters from the Isolation-Migration model for (**A**) Lagoon Cay island and mainland Belize, (**B**) West Snake Cay island and mainland Belize, (**C**) Cayos Cochinos Menor island and mainland Honduras, and (**D**) populations to the east and west of the Isthmus of Tehuantepec in Mexico.



Supplementary Figure 2. Admixture graphs for all *K* = 2 – 8 allowed source populations inferred using the stringently filtered SNP dataset and NGSadmix. Samples are labeled based on the sample ID and the coarse locality information for the sample. In most cases, the province of the sample is included and “Crawl”, “Lagoon”, “WSnake” and “Cochinos” refer to Crawl, Lagoon, and West Snake Cays in Belize and to Cayos Cochinos Menor in Honduras, respectively. Nation names are abbreviated according to the ISO three-letter country codes.



Supplementary Figure 3. Admixture graphs for all *K* = 2 – 8 allowed source populations inferred using the stringently filtered SNP dataset and Entropy. Samples are labeled based on the sample ID and the coarse locality information for the sample. In most cases, the province of the sample is included and "Crawl", "Lagoon", "WSnake" and "Cochinos" refer to Crawl, Lagoon, and West Snake Cays in Belize and to Cayos Cochinos Menor in Honduras, respectively. Nation names are abbreviated according to the ISO three-letter country codes.



Supplementary Figure 4. Violin plots of genome-wide estimates of *Pi* (**A**), *Heterozygosity* (**B**), and F_{IS} (**C**) for island and mainland populations in Central America and of F_{ST} (**D**) between each pairwise clade. For each violin plot, the white point indicates the median value and the black box indicates the interquartile range. The mean and standard deviations for each plot are evident above respective violin plots.

Supplementary Table 1. Voucher or collector identifiers for all new samples included as part of this work. Captive or pet animals lacking vouchers are encoded by '--'. UTA-ARDRC = University of Texas at Arlington Amphibian and Reptile Diversity Research Center.

Sample ID	Sample Provenance	Voucher Location	Collector/Accession ID
Boco02	Field	UTA-ARDRC	JAC 27533
Boco03	Field	UTA-ARDRC	JAC 27539
Boco04	Field	UTA-ARDRC	JAC 27537
Boco05	Field	UTA-ARDRC	JAC 27719
Boco09	Field	UTA-ARDRC	JAC 27911
Boco10	Field	UTA-ARDRC	JAC 27960
Boco11	Captive, pedigreed colony originating from field	--	SB03-34
Boco12	Captive, pedigreed colony originating from field	--	SB02-33
Boco13	Captive, pedigreed colony originating from field	--	SB02-2
Boco14	Captive, pedigreed colony originating from field	--	SB02-16
Boco15	Captive, pedigreed colony originating from field	--	SB02-28
Boco16	Captive, pedigreed colony originating from field	--	SB02-19
Boco17	Captive, pedigreed colony originating from field	--	SB02-18
Boco18	Captive, pedigreed colony originating from field	--	SB02-21
Boco19	Captive, pedigreed colony originating from field	--	SB02-26
Boco20	Captive, pedigreed colony originating from field	--	SB02-17
Boco21	Captive, pedigreed colony originating from field	--	SB02-12
Boco22	Captive, pedigreed colony originating from field	--	SB03-24
Boco23	Captive, pedigreed colony originating from field	--	SB03-23
Boco24	Captive, pedigreed colony originating from field	--	SB03-15
Boco25	Captive, pedigreed colony originating from field	--	SB03-8
Boco26	Captive, pedigreed colony originating from field	--	SB02-14
Boco27	Captive, pedigreed colony originating from field	--	SB37-11
Boco28	Captive, pedigreed colony originating from field	--	SB02-38
Boco29	Captive, pedigreed colony originating from field	--	SB02-1
Boco30	Captive, pedigreed colony originating from field	--	SB02-29
Boco32	Captive, pedigreed colony originating from field	--	SB02-15

Boco34	Field	UTA-ARDRC	JAC 30049
Boco35	Field	UTA-ARDRC	JAC 30066
Boco36	Field	UTA-ARDRC	JAC 27534
Boco37	Field	UTA-ARDRC	JAC 27619
Boco38	Field	UTA-ARDRC	JAC 27744
Boco39	Field	UTA-ARDRC	JAC 27762
Boco40	Field	UTA-ARDRC	JAC 27795
Boco41	Field	UTA-ARDRC	JAC 27906
Boco42	Field	UTA-ARDRC	JAC 27907
Boco43	Field	UTA-ARDRC	ENS 9615
Boco44	Field	UTA-ARDRC	MSM 64
Boco45	Field	UTA-ARDRC	MSM 375
Boco46	Field	UTA-ARDRC	MSM 65
Boco47	Field	UTA-ARDRC	JAC 19389
Boco48	Field	UTA-ARDRC	ENS 11066
Boco49	Field	UTA-ARDRC	ENS 11091
Boco50	Field	UTA-ARDRC	JAC 20093
Boco51	Field	UTA-ARDRC	JAC 21085
Boco52	Field	UTA-ARDRC	MSM 763
Boco53	Field	UTA-ARDRC	JAC 27920
Boco54	Field	UTA-ARDRC	JAC 27913
Boco55	Field	UTA-ARDRC	JAC 27956
Boco56	Field	UTA-ARDRC	JAC 28130
Boco58	Field	UTA-ARDRC	JAC 30042
Boco59	Field	UTA-ARDRC	JAC 30045
Boco60	Field	UTA-ARDRC	JAC 30061
Boco62	Field	UTA-ARDRC	JAC 30623
Boco63	Field	UTA-ARDRC	JAC 30456
Boco64	Field	UTA-ARDRC	JAC 30488
Boco68	Field	UTA-ARDRC	ENS 12060
Boco74	Locality pet skin shed	--	WB-NIC-2006-F1

Boco75	Locality pet skin shed	--	WB-NIC-2010a
Boco76	Locality pet skin shed	--	WB-NIC-2010b
Boco77	Locality pet skin shed	--	WB-BS-01
Boco79	Locality pet skin shed	--	WB-BS-ES01
Boco80	Locality pet skin shed	--	WB-BS-EL02
Boco81	Locality pet skin shed	--	WB-BWC-F1
Boco82	Locality pet skin shed	--	WB-HAR-M1
Boco83	Locality pet skin shed	--	WB-SONHL-2006a
Boco84	Locality pet skin shed	--	WB-SONHL-2006b
Boco85	Locality pet skin shed	--	WB-SONHL-2010a
Boco86	Locality pet skin shed	--	WB-SONHL-2010b
Boco87	Locality pet skin shed	--	WB-SONHL-2011a
Boco88	Locality pet skin shed	--	WB-BS-CC01
Boco89	Locality pet skin shed	--	WB-BS-CC02
Boco90	Locality pet skin shed	--	WB-CRT+-2007a
Boco91	Locality pet skin shed	--	WB-IK_LB-M1
Boco92	Field	Chad Montgomery (Truman State University)	467570162B
Boco102	Field	Timothy Colston (University of Mississippi)	TJC 928
Boco105	Locality pet skin shed	--	WB-BS-BCO-F1

Supplementary Table 2. All samples used for this study, along with the source of the data and the corresponding locality information. Missing or unsampled data is encoded by '--'. In some cases, coordinates were approximated for analyses by using coordinates at approximately the middle of the country or state/province (indicated with * following the coordinates) or using coordinates inferred from recorded locality data (indicated with ** following the coordinates).

Sample ID	**NCBI Accessions**		**Citation**	**Locality**	**Country: State/Province**	**Decimal Degree Latitude (WGS84)**	**Decimal Degree Longitude (WGS84)**
	Mito. cyt-b	**Nuclear RADseq**					
U69746	U69746	--	Campbell, 1997	--	--	--	--
AY575035	AY575035	--	Hynková et al., 2009	--	Mexico: Michoacan	19.3*	-101.34*
EU273605	EU273605	--	Hynková et al., 2009	--	El Salvador	--	--
EU273606	EU273606	--	Hynková et al., 2009	--	Belize: Crawl Cay	--	--
EU273607	EU273607	--	Hynková et al., 2009	--	El Salvador	--	--
EU273608	EU273608	--	Hynková et al., 2009	--	Belize: Crawl Cay	--	--
EU273609	EU273609	--	Hynková et al., 2009	--	Colombia: Bei Choco	--	--
EU273611	EU273611	--	Hynková et al., 2009	--	Colombia: Bei Choco	--	--
EU273613	EU273613	--	Hynková et al., 2009	--	Honduras: "Hog Island"	15.957*	-86.5*
EU273614	EU273614	--	Hynková et al., 2009	--	Costa Rica	--	--
EU273615	EU273615	--	Hynková et al., 2009	--	Nicaragua	--	--
EU273616	EU273616	--	Hynková et al., 2009	Cancún	Mexico: Quintana Roo	21.157*	-86.886*
EU273617	EU273617	--	Hynková et al., 2009	--	El Salvador	13.771*	-89.207*
EU273618	EU273618	--	Hynková et al., 2009	--	Honduras: Utila	--	--
EU273619	EU273619	--	Hynková et al., 2009	Tuxtla de Gutiérrez	Mexico: Chiapas	16.76*	-93.105*
EU273620	EU273620	--	Hynková et al., 2009	--	Guatemala: Escuintla	14.313*	-90.776*
EU273622	EU273622	--	Hynková et al., 2009	--	Mexico: "Sonora"	--	--
EU273623	EU273623	--	Hynková et al., 2009	Iquitos	Peru: Loreto	-3.755*	-76.26*
EU273624	EU273624	--	Hynková et al., 2009	Iquitos	Peru: Loreto	-3.755*	-76.26*
EU273625	EU273625	--	Hynková et al., 2009	Iquitos	Peru: Loreto	--	--
EU273626	EU273626	--	Hynková et al., 2009	Iquitos	Peru: Loreto	-3.755*	-76.26*
EU273627	EU273627	--	Hynková et al., 2009	Iquitos	Peru: Loreto	-3.755*	-76.26*

EU273628	EU273628	--	Hynková et al., 2009	Iquitos	Peru: Loreto	--	--
EU273629	EU273629	--	Hynková et al., 2009	--	Guyana	4.872*	-58.951*
EU273630	EU273630	--	Hynková et al., 2009	--	Surinam	4.058*	-55.885*
EU273631	EU273631	--	Hynková et al., 2009	--	Guyana	4.872*	-58.951*
EU273632	EU273632	--	Hynková et al., 2009	--	Guyana	4.87*	-58.95*
EU273633	EU273633	--	Hynková et al., 2009	--	Guyana	4.872*	-58.951*
EU273634	EU273634	--	Hynková et al., 2009	Iquitos	Peru: Loreto	--	--
EU273635	EU273635	--	Hynková et al., 2009	Tarapoto	Peru: San Martín	-6.496*	-76.37*
EU273636	EU273636	--	Hynková et al., 2009	Iquitos	Peru: Loreto	--	--
EU273637	EU273637	--	Hynková et al., 2009	--	Peru	--	--
EU273638	EU273638	--	Hynková et al., 2009	--	Peru	--	--
EU273639	EU273639	--	Hynková et al., 2009	--	Guyana	4.872*	-58.951*
EU273640	EU273640	--	Hynková et al., 2009	--	Peru	--	--
EU273641	EU273641	--	Hynková et al., 2009	--	S Brazil	-25.793*	-51.221*
EU273642	EU273642	--	Hynková et al., 2009	--	S Brazil	-25.793*	-51.221*
EU273643	EU273643	--	Hynková et al., 2009	--	Peru	--	--
EU273644	EU273644	--	Hynková et al., 2009	--	Brazil	--	--
EU273645	EU273645	--	Hynková et al., 2009	--	Brazil	--	--
EU273646	EU273646	--	Hynková et al., 2009	--	Guyana	4.872*	-58.951*
EU273647	EU273647	--	Hynková et al., 2009	--	Guyana	4.872*	-58.951*
EU273648	EU273648	--	Hynková et al., 2009	--	Surinam	--	--
EU273649	EU273649	--	Hynková et al., 2009	--	Surinam	--	--
EU273651	EU273651	--	Hynková et al., 2009	--	Argentina	-34.999*	-64.923*
EU273652	EU273652	--	Hynková et al., 2009	--	Brazil	--	--
EU273653	EU273653	--	Hynková et al., 2009	Marajó	Brazil: Pará	-0.898*	-49.801*
EU273654	EU273654	--	Hynková et al., 2009	Marajó	Brazil: Pará	-0.898*	-49.801*
EU273655	EU273655	--	Hynková et al., 2009	--	Brazil	--	--
EU273656	EU273656	--	Hynková et al., 2009	--	Belize: Crawl Cay	--	--
EU273657	EU273657	--	Hynková et al., 2009	--	Belize: Crawl Cay	--	--
EU273658	EU273658	--	Hynková et al., 2009	--	Colombia	--	--
EU273659	EU273659	--	Hynková et al., 2009	--	Ecuador	-1.262*	-78.548*

EU273660	EU273660	--	Hynková et al., 2009	--	Ecuador	-1.262*	-78.548*
EU273661	EU273661	--	Hynková et al., 2009	--	Guyana	4.872*	-58.951*
EU273662	EU273662	--	Hynková et al., 2009	--	Guyana	--	--
EU273664	EU273664	--	Hynková et al., 2009	--	Nicaragua	--	--
EU273665	EU273665	--	Hynková et al., 2009	--	Panama: Saboga Island	8.622**	-79.06**
EU273666	EU273666	--	Hynková et al., 2009	--	El Salvador	--	--
GQ300883	GQ300883	--	Hynková et al., 2009	--	Colombia	--	--
GQ300884	GQ300884	--	Hynková et al., 2009	--	Colombia	--	--
GQ300887	GQ300887	--	Hynková et al., 2009	--	Colombia	--	--
GQ300894	GQ300894	--	Hynková et al., 2009	--	Brazil	--	--
GQ300895	GQ300895	--	Hynková et al., 2009	--	Brazil	--	--
GQ300896	GQ300896	--	Hynková et al., 2009	--	Guyana	4.872*	-58.951*
GQ300897	GQ300897	--	Hynková et al., 2009	--	Surinam	--	--
GQ300898	GQ300898	--	Hynková et al., 2009	--	Guyana	4.872*	-58.951*
GQ300899	GQ300899	--	Hynková et al., 2009	--	Surinam	--	--
GQ300900	GQ300900	--	Hynková et al., 2009	--	Peru	--	--
GQ300901	GQ300901	--	Hynková et al., 2009	--	Peru	--	--
GQ300902	GQ300902	--	Hynková et al., 2009	Iquitos	Peru: Loreto	--	--
GQ300903	GQ300903	--	Hynková et al., 2009	--	Guyana	4.872*	-58.951*
GQ300904	GQ300904	--	Hynková et al., 2009	--	Surinam	4.058*	-55.885*
GQ300905	GQ300905	--	Hynková et al., 2009	--	Surinam	--	--
GQ300906	GQ300906	--	Hynková et al., 2009	--	Surinam	--	--
GQ300907	GQ300907	--	Hynková et al., 2009	--	Guyana	4.872*	-58.951*
GQ300908	GQ300908	--	Hynková et al., 2009	--	Guyana	4.872*	-58.951*
GQ300909	GQ300909	--	Hynková et al., 2009	--	Guyana	4.872*	-58.951*
GQ300910	GQ300910	--	Hynková et al., 2009	--	Surinam	--	--
GQ300911	GQ300911	--	Hynková et al., 2009	--	Argentina	-34.999*	-64.923*
GQ300912	GQ300912	--	Hynková et al., 2009	--	Argentina	-34.999*	-64.923*
GQ300913	GQ300913	--	Hynková et al., 2009	--	Argentina	-34.999*	-64.923*
GQ300914	GQ300914	--	Hynková et al., 2009	--	Argentina	-34.999*	-64.923*
GQ300915	GQ300915	--	Hynková et al., 2009	--	Argentina	-34.999*	-64.923*

GQ300916	GQ300916	--	Hynková et al., 2009	--	Argentina	-34.999*	-64.923*
GQ300917	GQ300917	--	Hynková et al., 2009	--	Belize: Crawl Cay	--	--
GQ300918	GQ300918	--	Hynková et al., 2009	--	Nicaragua	--	--
GQ300919	GQ300919	--	Hynková et al., 2009	--	Honduras: "Hog Island"	15.957**	-86.5**
GQ300920	GQ300920	--	Hynková et al., 2009	--	Honduras: "Hog Island"	15.957**	-86.5**
GQ300922	GQ300922	--	Hynková et al., 2009	--	Costa Rica	--	--
GQ300923	GQ300923	--	Hynková et al., 2009	--	Costa Rica	--	--
GQ300924	GQ300924	--	Hynková et al., 2009	--	Costa Rica: Canuita	9.735**	-82.844**
GQ300925	GQ300925	--	Hynková et al., 2009	--	Costa Rica	--	--
GQ300926	GQ300926	--	Hynková et al., 2009	Tuxtla de Gutiérrez	Mexico: Chiapas	16.76*	-93.105*
GQ300927	GQ300927	--	Hynková et al., 2009	Tuxtla de Gutiérrez	Mexico: Chiapas	--	--
GQ300928	GQ300928	--	Hynková et al., 2009	Tuxtla de Gutiérrez	Mexico: Chiapas	16.76*	-93.105*
GQ300929	GQ300929	--	Hynková et al., 2009	--	Mexico: "Sonora"	--	--
GQ300930	GQ300930	--	Hynková et al., 2009	--	Mexico: "Sonora"	--	--
GQ300931	GQ300931	--	Hynková et al., 2009	--	Belize: Crawl Cay	--	--
GQ300932	GQ300932	--	Hynková et al., 2009	--	Mexico: "Sonora"	--	--
GQ300933	GQ300933	--	Hynková et al., 2009	--	Mexico: "Sonora"	--	--
GQ300934	GQ300934	--	Hynková et al., 2009	--	Mexico: "Sonora"	--	--
GQ300935	GQ300935	--	Hynková et al., 2009	Cancún	Mexico: Quintana Roo	21.157*	-86.886*
JX026897	JX026897	--	Reynolds et al., 2013	--	Puerto Rico	--	--
JX026898	JX026898	--	Reynolds et al., 2013	--	Puerto Rico	--	--
HQ399514	HQ399514	--	Rivera et al., 2011	--	Unkonwn	--	--
KJ621415	KJ621415	--	Suárez-Atilano et al., 2014	Rondônia	Brazil: Nova Brasilia	-11.150	-61.57
KJ621416	KJ621416	--	Suárez-Atilano et al., 2014	Alamos	Mexico: "Sonora"	29.212	-110.136
KJ621417	KJ621417	--	Suárez-Atilano et al., 2014	Beldiraguato	Mexico: Sinaloa	25.221	-107.610
KJ621418	KJ621418	--	Suárez-Atilano et al., 2014	Mazatlan	Mexico: Sinaloa	23.406	-106.506
KJ621419	KJ621419	--	Suárez-Atilano et al., 2014	Acaponeta	Mexico: Sinaloa	22.351	-103.314
KJ621420	KJ621420	--	Suárez-Atilano et al., 2014	El Naranjo	Mexico: Colima	19.159	-104.269
KJ621421	KJ621421	--	Suárez-Atilano et al., 2014	Manzanillo	Mexico: Colima	19.056	-104.269
KJ621422	KJ621422	--	Suárez-Atilano et al., 2014	Manzanillo	Mexico: Colima	19.101	-104.295
KJ621423	KJ621423	--	Suárez-Atilano et al., 2014	Puerto Vallarta	Mexico: Jalisco	20.676	-105.239

KJ621424	KJ621424	--	Suárez-Atilano et al., 2014	Puerto Vallarta	Mexico: Jalisco	20.676	-105.239
KJ621425	KJ621425	--	Suárez-Atilano et al., 2014	Puerto Vallarta	Mexico: Jalisco	20.676	-105.239
KJ621426	KJ621426	--	Suárez-Atilano et al., 2014	Puerto Vallarta	Mexico: Jalisco	20.733	-105.295
KJ621427	KJ621427	--	Suárez-Atilano et al., 2014	Puerto Vallarta	Mexico: Jalisco	19.885	-105.341
KJ621428	KJ621428	--	Suárez-Atilano et al., 2014	Melaque	Mexico: Jalisco	19.435	-104.672
KJ621429	KJ621429	--	Suárez-Atilano et al., 2014	Melaque	Mexico: Jalisco	19.456	−104.657
KJ621430	KJ621430	--	Suárez-Atilano et al., 2014	Puerto Vallarta	Mexico: Jalisco	19.996	-105.315
KJ621431	KJ621431	--	Suárez-Atilano et al., 2014	Mascota	Mexico: Jalisco	18.436	-103.531
KJ621432	KJ621432	--	Suárez-Atilano et al., 2014	La Huerta	Mexico: Jalisco	19.593	-105.041
KJ621433	KJ621433	--	Suárez-Atilano et al., 2014	El Limón	Mexico: Jalisco	19.807	-104.918
KJ621434	KJ621434	--	Suárez-Atilano et al., 2014	Puerto Escondido	Mexico: Jalisco	20.355	-105.317
KJ621435	KJ621435	--	Suárez-Atilano et al., 2014	Aquila	Mexico: Michoacan	18.585	−103.558
KJ621436	KJ621436	--	Suárez-Atilano et al., 2014	Apatzingan	Mexico: Michoacan	19.206	−102.614
KJ621437	KJ621437	--	Suárez-Atilano et al., 2014	Solera de Agua	Mexico: Michoacan	18.023	−102.472
KJ621438	KJ621438	--	Suárez-Atilano et al., 2014	Playa Azul	Mexico: Michoacan	18.195	−103.053
KJ621439	KJ621439	--	Suárez-Atilano et al., 2014	Coalcomán	Mexico: Michoacan	18.396	−103.509
KJ621440	KJ621440	--	Suárez-Atilano et al., 2014	Lazaro Cardenas	Mexico: Michoacan	20.269	−105.319
KJ621441	KJ621441	--	Suárez-Atilano et al., 2014	Marauta	Mexico: Michoacan	18.228	−103.188
KJ621442	KJ621442	--	Suárez-Atilano et al., 2014	Ayutla	Mexico: Guerrero	17.142	−99.540
KJ621443	KJ621443	--	Suárez-Atilano et al., 2014	Atoyac	Mexico: Guerrero	16.971	−99.890
KJ621444	KJ621444	--	Suárez-Atilano et al., 2014	Atoyac	Mexico: Guerrero	17.369	−100.200
KJ621445	KJ621445	--	Suárez-Atilano et al., 2014	Puerto Escondido	Mexico: Oaxaca	15.898	−97.063
KJ621446	KJ621446	--	Suárez-Atilano et al., 2014	Puerto Escondido	Mexico: Oaxaca	15.898	−97.063
KJ621447	KJ621447	--	Suárez-Atilano et al., 2014	Huatulco	Mexico: Oaxaca	16.650	−98.669
KJ621448	KJ621448	--	Suárez-Atilano et al., 2014	Huatulco	Mexico: Oaxaca	16.650	−98.669
KJ621449	KJ621449	--	Suárez-Atilano et al., 2014	Puerto Escondido	Mexico: Oaxaca	16.650	−98.669
KJ621450	KJ621450	--	Suárez-Atilano et al., 2014	Puerto Escondido	Mexico: Oaxaca	16.650	−98.669
KJ621451	KJ621451	--	Suárez-Atilano et al., 2014	Puerto Escondido	Mexico: Oaxaca	16.650	−98.669
KJ621452	KJ621452	--	Suárez-Atilano et al., 2014	San Juan de los Cues	Mexico: Oaxaca	18.068	−98.068
KJ621453	KJ621453	--	Suárez-Atilano et al., 2014	San Agustín Loxicha	Mexico: Oaxaca	15.700	−96.5
KJ621454	KJ621454	--	Suárez-Atilano et al., 2014	Tlachicón	Mexico: Oaxaca	15.724	−96.637

KJ621455	KJ621455	--	Suárez-Atilano et al., 2014	Tlachicón	Mexico: Oaxaca	15.732	−96.493
KJ621456	KJ621456	--	Suárez-Atilano et al., 2014	San José Chacalapa	Mexico: Oaxaca	15.844	−96.464
KJ621457	KJ621457	--	Suárez-Atilano et al., 2014	San José Chacalapa	Mexico: Oaxaca	16.434	−98.321
KJ621458	KJ621458	SRS1141623	Suárez-Atilano et al., 2014 (cyt-b)/This study (RADseq)	Pinotepa Nacional	Mexico: Oaxaca	16.123	−97.712
KJ621459	KJ621459	SRS1141626	Suárez-Atilano et al., 2014 (cyt-b)/This study (RADseq)	Pinotepa Nacional	Mexico: Oaxaca	16.239	−97.792
KJ621460	KJ621460	--	Suárez-Atilano et al., 2014	Santiago Jamiltepec	Mexico: Oaxaca	16.250	−97.801
KJ621461	KJ621461	SRS1141624	Suárez-Atilano et al., 2014 (cyt-b)/This study (RADseq)	Santiago Jamiltepec	Mexico: Oaxaca	15.961	−97.376
KJ621462	KJ621462	--	Suárez-Atilano et al., 2014	Pochutla	Mexico: Oaxaca	15.882	−96.485
KJ621463	KJ621463	--	Suárez-Atilano et al., 2014	San Gabriel Mixtepec	Mexico: Oaxaca	16.887	−98.903
KJ621464	KJ621464	--	Suárez-Atilano et al., 2014	El Camarón	Mexico: Oaxaca	16.400	−95.650
KJ621465	KJ621465	--	Suárez-Atilano et al., 2014	Sto. Domingo Tehuantepec	Mexico: Oaxaca	16.383	−95.268
KJ621466	KJ621466	--	Suárez-Atilano et al., 2014	Huatulco	Mexico: Oaxaca	15.82	−96.001
KJ621467	KJ621467	--	Suárez-Atilano et al., 2014	Sierra Mazateca	Mexico: Oaxaca	18.221	−96.687
KJ621468	KJ621468	--	Suárez-Atilano et al., 2014	Teotitlan del Camino	Mexico: Oaxaca	18.247	−97.155
KJ621469	KJ621469	--	Suárez-Atilano et al., 2014	Eloxotitlán	Mexico: Oaxaca	18.488	−96.854
KJ621470	KJ621470	--	Suárez-Atilano et al., 2014	Huautla	Mexico: Morelos	18.448	−98.987
KJ621471	KJ621471	--	Suárez-Atilano et al., 2014	Huautla	Mexico: Morelos	18.458	−99.026
KJ621472	KJ621472	--	Suárez-Atilano et al., 2014	Huautla	Mexico: Morelos	18.458	−99.026
KJ621473	KJ621473	--	Suárez-Atilano et al., 2014	El Cielo	Mexico: Tamulipas	23.045	−99.229
KJ621474	KJ621474	--	Suárez-Atilano et al., 2014	Barra del Tordo	Mexico: Tamaulipas	22.913	−97.949
KJ621475	KJ621475	--	Suárez-Atilano et al., 2014	Ejido la Concepción	Mexico: San Luis Potosí	21.688	−98.800
KJ621476	KJ621476	--	Suárez-Atilano et al., 2014	San Andrés Tuxtla	Mexico: Veracruz	18.613	−95.070
KJ621477	KJ621477	--	Suárez-Atilano et al., 2014	Los Chimalapas	Mexico: Veracruz	17.159	−94.229
KJ621478	KJ621478	--	Suárez-Atilano et al., 2014	Boca del Río	Mexico: Veracruz	19.106	−96.115
KJ621479	KJ621479	--	Suárez-Atilano et al., 2014	Boca del Río	Mexico: Veracruz	19.106	−96.115
KJ621480	KJ621480	--	Suárez-Atilano et al., 2014	Boca del Río	Mexico: Veracruz	19.106	−96.115
KJ621481	KJ621481	--	Suárez-Atilano et al., 2014	Boca del Río	Mexico: Veracruz	19.106	−96.115
KJ621482	KJ621482	--	Suárez-Atilano et al., 2014	Boca del Río	Mexico: Veracruz	19.106	−96.115

KJ621483	KJ621483	--	Suárez-Atilano et al., 2014	Boca del Río	Mexico: Veracruz	19.106	−96.115
KJ621484	KJ621484	--	Suárez-Atilano et al., 2014	Boca del Río	Mexico: Veracruz	19.106	−96.115
KJ621485	KJ621485	--	Suárez-Atilano et al., 2014	Boca del Río	Mexico: Veracruz	19.106	−96.115
KJ621486	KJ621486	--	Suárez-Atilano et al., 2014	Yumkaa'	Mexico: Tabasco	18.008	−92.825
KJ621487	KJ621487	--	Suárez-Atilano et al., 2014	Yumkaa'	Mexico: Tabasco	18.008	−92.825
KJ621488	KJ621488	--	Suárez-Atilano et al., 2014	Yumkaa'	Mexico: Tabasco	18.008	−92.825
KJ621489	KJ621489	--	Suárez-Atilano et al., 2014	El espino	Mexico: Tabasco	18.242	−92.831
KJ621490	KJ621490	--	Suárez-Atilano et al., 2014	Centla	Mexico: Tabasco	18.413	−92.919
KJ621491	KJ621491	--	Suárez-Atilano et al., 2014	Comunidad Emiliano	Mexico: Tabasco	17.737	−91.762
KJ621492	KJ621492	--	Suárez-Atilano et al., 2014	Cd. Del Cármen	Mexico: Campeche	18.227	−91.829
KJ621493	KJ621493	--	Suárez-Atilano et al., 2014	Cd. Del Cármen	Mexico: Campeche	18.227	−91.829
KJ621494	KJ621494	--	Suárez-Atilano et al., 2014	Cd. Del Cármen	Mexico: Campeche	18.227	−91.829
KJ621495	KJ621495	--	Suárez-Atilano et al., 2014	Xpujil–Bel Ha	Mexico: Campeche	18.573	−89.409
KJ621496	KJ621496	--	Suárez-Atilano et al., 2014	Cuxtal	Mexico: Yucatán	20.911	−89.611
KJ621497	KJ621497	--	Suárez-Atilano et al., 2014	Cuxtal	Mexico: Yucatán	20.911	−89.611
KJ621498	KJ621498	--	Suárez-Atilano et al., 2014	Cuxtal	Mexico: Yucatán	20.911	−89.611
KJ621499	KJ621499	--	Suárez-Atilano et al., 2014	Cuxtal	Mexico: Yucatán	20.911	−89.611
KJ621500	KJ621500	--	Suárez-Atilano et al., 2014	Cuxtal	Mexico: Yucatán	20.911	−89.611
KJ621501	KJ621501	--	Suárez-Atilano et al., 2014	Cuxtal	Mexico: Yucatán	20.911	−89.611
KJ621502	KJ621502	--	Suárez-Atilano et al., 2014	Cuxtal	Mexico: Yucatán	20.911	−89.611
KJ621503	KJ621503	--	Suárez-Atilano et al., 2014	Cuxtal	Mexico: Yucatán	20.911	−89.611
KJ621504	KJ621504	--	Suárez-Atilano et al., 2014	Cuxtal	Mexico: Yucatán	20.911	−89.611
KJ621505	KJ621505	--	Suárez-Atilano et al., 2014	Cuxtal	Mexico: Yucatán	20.911	−89.611
KJ621506	KJ621506	--	Suárez-Atilano et al., 2014	Cuxtal	Mexico: Yucatán	20.911	−89.611
KJ621507	KJ621507	--	Suárez-Atilano et al., 2014	Cuxtal	Mexico: Yucatán	20.911	−89.611
KJ621508	KJ621508	--	Suárez-Atilano et al., 2014	Mani	Mexico: Yucatán	20.389	−89.373
KJ621509	KJ621509	--	Suárez-Atilano et al., 2014	Piste	Mexico: Yucatán	20.719	−88.610
KJ621510	KJ621510	--	Suárez-Atilano et al., 2014	Mérida	Mexico: Yucatán	20.964	−89.616
KJ621511	KJ621511	--	Suárez-Atilano et al., 2014	Mérida	Mexico: Yucatán	20.964	−89.616
KJ621512	KJ621512	--	Suárez-Atilano et al., 2014	Mérida	Mexico: Yucatán	20.964	−89.616
KJ621513	KJ621513	--	Suárez-Atilano et al., 2014	Mérida	Mexico: Yucatán	20.964	−89.616

KJ621514	KJ621514	--	Suárez-Atilano et al., 2014	Mérida	Mexico: Yucatán	20.964	−89.616
KJ621515	KJ621515	--	Suárez-Atilano et al., 2014	Mérida	Mexico: Yucatán	20.964	−89.616
KJ621516	KJ621516	--	Suárez-Atilano et al., 2014	Valladolid	Mexico: Yucatán	20.693	−88.199
KJ621517	KJ621517	--	Suárez-Atilano et al., 2014	Felipe Carrillo Puerto	Mexico: Quintana Roo	20.14	−88.301
KJ621518	KJ621518	--	Suárez-Atilano et al., 2014	Felipe Carrillo Puerto	Mexico: Quintana Roo	20.14	−88.301
KJ621519	KJ621519	--	Suárez-Atilano et al., 2014	El Triunfo	Mexico: Chiapas	15.354	−92.589
KJ621520	KJ621520	--	Suárez-Atilano et al., 2014	Antigua	Guatemala	16.048	−90.066
KJ621521	KJ621521	--	Suárez-Atilano et al., 2014	Santa Inés Chicar	Guatemala	14.431	−89.631
KJ621522	KJ621522	--	Suárez-Atilano et al., 2014	Asunción Mita	Guatemala	15.962	−88.618
KJ621523	KJ621523	--	Suárez-Atilano et al., 2014	El Rosario	Guatemala: Zacapa	14.972	−89.526
KJ621524	KJ621524	--	Suárez-Atilano et al., 2014	Escuintla	Guatemala: Izabal	15.738	−88.579
KJ621525	KJ621525	--	Suárez-Atilano et al., 2014	Antigua	Guatemala	14.616	−90.567
KJ621526	KJ621526	--	Suárez-Atilano et al., 2014	Bocas del Toro	Panama	9.237	−82.342
KJ621527	KJ621527	--	Suárez-Atilano et al., 2014	Canal Zone	Panama	9.101	−79.699
KJ621528	KJ621528	--	Suárez-Atilano et al., 2014	Barro Colorado	Panama	9.333	−79.912
KJ621529	KJ621529	--	Suárez-Atilano et al., 2014	Paraíso	Panama	9.031	−79.610
KJ621530	KJ621530	--	Suárez-Atilano et al., 2014	Cayos Cochino Pequeño	Honduras: Islas de la Bahía	15.958	−86.464
KJ621531	KJ621531	--	Suárez-Atilano et al., 2014	Isla De Roatan	Honduras: Islas de la Bahía	16.315	−86.537
KJ621532	KJ621532	--	Suárez-Atilano et al., 2014	San Francisco Menéndez	El Salvador: Ahuachapán	13.867	−89.983
KJ621533	KJ621533	--	Suárez-Atilano et al., 2014	San Francisco Menéndez	El Salvador: Ahuachapán	18.823	−89.943
KJ621534	KJ621534	--	Suárez-Atilano et al., 2014	San Francisco Menéndez	El Salvador: Ahuachapán	13.867	−89.983
KJ621535	KJ621535	--	Suárez-Atilano et al., 2014	Arambala	El Salvador: Morazán	13.767	−88.129
KJ621536	KJ621536	--	Suárez-Atilano et al., 2014	Río Tortuguero	Costa Rice: Limón	10.583	−83.517
KJ621537	KJ621537	--	Suárez-Atilano et al., 2014	Río Tortuguero	Costa Rice: Limón	10.572	−83.517
Boco02	KX150438	--	This study	HWY 51 between Iguala and Tlapehuala	Mexico: Guerrero	18.25702	-100.49908
Boco03	KX150439	--	This study	HWY 51 between Cd.	Mexico: Guerrero	18.5689	-100.84668

				Altamirano and Huetamo			
Boco04	--	--	This study	HWY 51 between Cd. Altamirano and Huetamo	Mexico: Michoacan	18.57488	-100.78448
Boco05	--	--	This study	HWY 200 between Atoyac de Alvarez and Zihuatanejo	Mexico: Guerrero	17.15576	-100.51598
Boco09	KX150421	SRS1141657	This study	HWY 200 between La Placita and Maruata	Mexico: Michoacan	18.55801	-103.60585
Boco10	KX150444	--	This study	Road from Comala to Minatitlan	Mexico: Colima	19.38787	-104.06107
Boco11	KX150375	SRS1141656	This study	--	Belize: Cayo	17.15428	-88.67733
Boco12	--	SRS1141655	This study	--	Belize: Belize	17.50428	-88.19586
Boco13	KX150377	SRS1141654	This study	--	Belize: Crawl Cay	16.599068	-88.219541
Boco14	KX150382	SRS1141653	This study	--	Belize: Lagoon Cay	16.631967	-88.20894
Boco15	KX150392	SRS1141652	This study	--	Belize: West Snake Cay	16.191679	-88.571827
Boco16	KX150393	--	This study	--	Belize: West Snake Cay	16.191679	-88.571827
Boco17	KX150383	SRS1141651	This study	--	Belize: Belize	17.49572	-88.22369
Boco18	KX150397	SRS1141650	This study	--	Belize: West Snake Cay	16.191679	-88.571827
Boco19	KX150396	SRS1141649	This study	--	Belize: West Snake Cay	16.191679	-88.571827
Boco20	KX150400	SRS1141648	This study	--	Belize: Lagoon Cay	16.631967	-88.20894
Boco21	KX150386	SRS1141647	This study	--	Belize: Lagoon Cay	16.631967	-88.20894
Boco22	KX150385	SRS1141646	This study	--	Belize: Lagoon Cay	16.631967	-88.20894
Boco23	KX150387	SRS1141645	This study	--	Belize: Lagoon Cay	16.631967	-88.20894
Boco24	KX150398	SRS1141644	This study	--	Belize: West Snake Cay	16.191679	-88.571827
Boco25	KX150391	SRS1141643	This study	--	Belize: West Snake Cay	16.191679	-88.571827
Boco26	KX150401	SRS1141642	This study	--	Belize: Lagoon Cay	16.631967	-88.20894
Boco27	KX150395	SRS1141641	This study	--	Belize: Belize	17.535829	-88.235735
Boco28	KX150390	SRS1141640	This study	--	Belize: Belize	17.516032	-88.199182
Boco29	KX150376	SRS1141639	This study	--	Belize: Crawl Cay	16.599068	-88.219541
Boco30	KX150389	SRS1141638	This study	--	Belize: West Snake Cay	16.191679	-88.571827
Boco32	KX150388	SRS1141637	This study	--	Belize: Lagoon Cay	16.631967	-88.20894

Boco34	KX150442	--	This study	Road from Comala to Minatitlan	Mexico: Colima	19.31932	-103.93382
Boco35	KX150430	--	This study	Road from Comala to Minatitlan	Mexico: Colima	19.30380	-103.81249
Boco36	--	--	This study	HWY 51 between Cd. Altamirano and Huetamo	Mexico: Michoacan	18.4767	-100.72076
Boco37	KX150440	--	This study	HWY 51 between Huetamo and El Limon de Papatzingan	Mexico: Michoacan	19.27947	-100.80436
Boco38	KX150420	SRS1141636	This study	HWY 200 between Atoyac de Alvarez and Zihuatanejo	Mexico: Guerrero	17.61819	-101.45058
Boco39	KX150418	--	This study	HWY 134 from Ixtapa to Cd. Altamirano	Mexico: Guerrero	17.91146	-101.33376
Boco40	KX150419	--	This study	HWY 134 from Ixtapa to Cd. Altamirano	Mexico: Guerrero	17.66888	-101.57269
Boco41	KX150437	--	This study	HWY 200 between La Placita and Maruata	Mexico: Michoacan	18.50881	-103.57672
Boco42	--	--	This study	HWY 200 between La Placita and Maruata	Mexico: Michoacan	18.47123	-103.54637
Boco43	--	--	This study	--	Mexico: Oaxaca	17.05423*	-96.71323*
Boco44	KX150380	SRS1141635	This study	Cabañas, El Arenal, El Zarco	Guatemala: Zacapa	15.078426*	-89.43639*
Boco45	KX150378	--	This study	--	Guatemala: Huehuetenango	15.587991*	-91.67607*
Boco46	KX150379	SRS1141634	This study	Cabañas, El Arenal, El Zarco	Guatemala: Zacapa	15.078426*	-89.43639*
Boco47	KX150381	SRS1141633	This study	--	Guatemala: Baja Verapaz	15.07875*	-90.41252*
Boco48	KX150373	--	This study	--	Venezuela	6.423749*	-66.58973*
Boco49	KX150374	--	This study	--	Venezuela	6.423749*	-66.58973*
Boco50	KX150413	SRS1141632	This study	Malacaton, Finca San Ignacio	Guatemala: San Marcos	14.94583	-92.025
Boco51	KX150384	SRS1141631	This study	7.4 mi N Tikal on road to Uaxactún	Guatemala: Petén	17.3025	-89.63444

Boco52	KX150412	SRS1141630	This study	Brito, Finca El Caobanal	Guatemala: Escuintla	14.11367	-90.6295
Boco53	KX150435	--	This study	HWY 200 between La Placita and Maruata	Mexico: Michoacan	18.47473	-103.54273
Boco54	KX150436	--	This study	HWY 200 between La Placita and Maruata	Mexico: Michoacan	18.49786	-103.57076
Boco55	KX150441	--	This study	Road from Colima to Minatitlan	Mexico: Colima	19.41521	-104.0118
Boco56	KX150443	SRS1141629	This study	Road from Colima to Minatitlan	Mexico: Colima	19.17588	-104.25472
Boco58	KX150428	--	This study	Road from Comala to Minatitlan	Mexico: Colima	19.31468	-103.84741
Boco59	KX150431	SRS1141628	This study	Road from Comala to Minatitlan	Mexico: Colima	19.30975	-103.89030
Boco60	KX150432	--	This study	Road from Comala to Minatitlan	Mexico: Colima	19.32201	-103.93701
Boco62	KX150429	SRS1141627	This study	Road from HWY 54 to Ixtlahuacan; side road to Jiliotupa	Mexico: Colima	19.00815	-103.75595
Boco63	KX150433	--	This study	Road from Comala to Minatitlan	Mexico: Colima	19.27894	-103.75465
Boco64	KX150434	--	This study	Road from HWY 54 to Ixtlahuacan	Mexico: Colima	19.05072	-103.78416
Boco68	KX150424	SRS1141625	This study	--	Mexico: Oaxaca	16.24965	-94.80138
Boco74	KX150402	SRS1141622	This study	--	Nicaragua	12.865416	-85.207229
Boco75	KX150408	SRS1141621	This study	--	Nicaragua	12.865416*	-85.20723*
Boco76	KX150404	SRS1141620	This study	--	Nicaragua	12.865416*	-85.20723*
Boco77	KX150406	SRS1141619	This study	--	Nicaragua	12.865416*	-85.20723*
Boco79	KX150403	--	This study	--	El Salvador	13.794185*	-88.89653*
Boco80	KX150407	SRS1141618	This study	--	El Salvador	13.794185*	-88.89653*
Boco81	KX150405	--	This study	--	Colombia	4.570868*	-74.29733*
Boco82	--	SRS1141617	This study	--	Colombia	4.570868*	-74.29733*
Boco83	KX150416	SRS1141616	This study	--	Mexico: Sonora	29.297226*	-110.3309*
Boco84	KX150415	SRS1141615	This study	--	Mexico: Sonora	29.297226*	-110.3309*
Boco85	KX150414	SRS1141614	This study	--	Mexico: Sonora	29.297226*	-110.3309*

Boco86	KX150417	SRS1141613	This study	--	Mexico: Sonora	29.297226*	-110.3309*
Boco87	--	SRS1141612	This study	--	Mexico: Sonora	29.297226*	-110.3309*
Boco88	--	SRS1141611	This study	--	Honduras: Cayos Cochinos Menor	15.97212**	-86.4756**
Boco89	KX150410	SRS1141610	This study	--	Honduras: Cayos Cochinos Menor	15.97212**	-86.4756**
Boco90	KX150409	SRS1141609	This study	--	Costa Rica	9.748917	-83.753428
Boco91	KX150399	SRS1141608	This study	--	Honduras	15.199999	-86.241905
Boco92	KX150411	--	This study	--	Honduras: Cayos Cochinos Menor	--	--
Boco102	KX150394	--	This study	--	Mexico	--	--
Boco105	--	SRS1141607	This study	--	Argentina	--	--

Supplementary Table 3. Outgroup species used in the mitochondrial phylogenetic analysis. NCBI Genbank accession for the cyt-b sequence and the citation where the data was originally used are also included.

NCBI Mitochondrial cyt-b Accession	Citation	Species
U69751	Campbell, 1997	*Candoia aspera*
U69754	Campbell, 1997	*Candoia carinata*
U69777	Campbell, 1997	*Epicrates cenchria*
U69808	Campbell, 1997	*Eunectes murinus*
U69812	Campbell, 1997	*Eryx colubrinus loveridgei*
U69823	Campbell, 1997	*Eryx johnii*
U69839	Campbell, 1997	*Liasis mackloti savuensis*
U69851	Campbell, 1997	*Morelia spilota*
U69853	Campbell, 1997	*Python molurus*
U69866	Campbell, 1997	*Sanzinia madagascariensis*
JX576179	Colston et al., 2013	*Corallus caninus*
HM348832	Colston, 2010	*Corallus annulatus*
KC329924	Reynolds et al., 2013	*Chilabothrus chrysogaster*
KC329931	Reynolds et al., 2013	*Chilabothrus monensis granti*
KC329953	Reynolds et al., 2013	*Eunectes notaeus*
HQ399504	Rivera et al., 2011	*Epicrates crassus*
AY099985	Slowinski and Lawson, 2002	*Calabaria reinhardtii*
AY099986	Slowinski and Lawson, 2002	*Charina bottae*
AY099989	Slowinski and Lawson, 2002	*Exiliboa placata*
AY099993	Slowinski and Lawson, 2002	*Loxocemus bicolor*

Supplementary Table 4. A summary of which ingroup samples were used for individual mitochondrial and nuclear genetic analyses. Data not included in an analysis – due either to not being collected or being excluded (see Materials and Methods) – is encoded by ‘--'.

Sample ID	Mitochondrial clade	Sample Assignment for mtDNA Analyses		Sample Assignment for Nuclear Analyses
		Landscape Diversity	IMa2	TreeMix
U69746	Central America	--	Mainland Honduras	--
AY575035	North America	North America	West Tehuantepec	--
EU273605	Central America	--	--	--
EU273606	Central America	--	--	--
EU273607	Central America	--	--	--
EU273608	Central America	--	--	--
EU273609	Central America	--	--	--
EU273611	Central America	--	--	--
EU273613	Central America	Central America	Cayos Cochinos	--
EU273614	Central America	--	Mainland Honduras	--
EU273615	Central America	--	Mainland Honduras	--
EU273616	Central America	Central America	Mainland Belize	--
EU273617	Central America	Central America	--	--
EU273618	Central America	--	--	--
EU273619	Central America	Central America	East Tehuantepec	--
EU273620	Central America	Central America	East Tehuantepec	--
EU273622	North America	--	--	--
EU273623	South America	South America	--	--
EU273624	South America	South America	--	--
EU273625	South America	--	--	--
EU273626	South America	South America	--	--
EU273627	South America	South America	--	--
EU273628	South America	--	--	--

EU273629	South America	South America	--	--
EU273630	South America	South America	--	--
EU273631	South America	South America	--	--
EU273632	South America	South America	--	--
EU273633	South America	South America	--	--
EU273634	South America	--	--	--
EU273635	South America	South America	--	--
EU273636	South America	--	--	--
EU273637	South America	--	--	--
EU273638	South America	--	--	--
EU273639	South America	South America	--	--
EU273640	South America	--	--	--
EU273641	South America	South America	--	--
EU273642	South America	South America	--	--
EU273643	South America	--	--	--
EU273644	South America	--	--	--
EU273645	South America	--	--	--
EU273646	South America	South America	--	--
EU273647	South America	South America	--	--
EU273648	South America	--	--	--
EU273649	South America	--	--	--
EU273651	South America	South America	--	--
EU273652	South America	--	--	--
EU273653	South America	South America	--	--
EU273654	South America	South America	--	--
EU273655	South America	--	--	--
EU273656	Central America	--	--	--
EU273657	Central America	--	--	--
EU273658	South America	--	--	--
EU273659	South America	Central America	--	--
EU273660	South America	Central America	--	--

EU273661	South America	South America	--	--
EU273662	South America	--	--	--
EU273664	Central America	--	--	--
EU273665	Central America	Central America	--	--
EU273666	Central America	--	--	--
GQ300883	South America	--	--	--
GQ300884	South America	--	--	--
GQ300887	South America	--	--	--
GQ300894	South America	--	--	--
GQ300895	South America	--	--	--
GQ300896	South America	South America	--	--
GQ300897	South America	--	--	--
GQ300898	South America	South America	--	--
GQ300899	South America	--	--	--
GQ300900	South America	--	--	--
GQ300901	South America	--	--	--
GQ300902	South America	--	--	--
GQ300903	South America	South America	--	--
GQ300904	South America	South America	--	--
GQ300905	South America	--	--	--
GQ300906	South America	--	--	--
GQ300907	South America	South America	--	--
GQ300908	South America	South America	--	--
GQ300909	South America	South America	--	--
GQ300910	South America	South America	--	--
GQ300911	South America	South America	--	--
GQ300912	South America	South America	--	--
GQ300913	South America	South America	--	--
GQ300914	South America	South America	--	--
GQ300915	South America	South America	--	--
GQ300916	South America	South America	--	--

GQ300917	Central America	--	--	--
GQ300918	Central America	--	--	--
GQ300919	Central America	Central America	Cayos Cochinos Menor	--
GQ300920	Central America	Central America	Cayos Cochinos Menor	--
GQ300922	Central America	--	Mainland Honduras	--
GQ300923	Central America	--	Mainland Honduras	--
GQ300924	Central America	Central America	Mainland Honduras	--
GQ300925	Central America	--	Mainland Honduras	--
GQ300926	Central America	Central America	East Tehuantepec	--
GQ300927	Central America	--	East Tehuantepec	--
GQ300928	Central America	Central America	East Tehuantepec	--
GQ300929	North America	--	--	--
GQ300930	North America	--	--	--
GQ300931	Central America	--	--	--
GQ300932	North America	--	--	--
GQ300933	North America	--	--	--
GQ300934	North America	--	--	--
GQ300935	Central America	Central America	--	--
JX026897	South America	--	--	--
JX026898	Central America	--	--	--
HQ399514	South America	--	--	--
KJ621415	South America	South America	--	--
KJ621416	North America	North America	--	--
KJ621417	North America	North America	--	--
KJ621418	North America	North America	--	--
KJ621419	North America	North America	--	--
KJ621420	North America	North America	--	--
KJ621421	North America	North America	--	--
KJ621422	North America	North America	--	--
KJ621423	North America	North America	--	--
KJ621424	North America	North America	--	--

KJ621425	North America	North America	--	--
KJ621426	North America	North America	--	--
KJ621427	North America	North America	--	--
KJ621428	North America	North America	--	--
KJ621429	North America	North America	--	--
KJ621430	North America	North America	--	--
KJ621431	North America	North America	West Tehuantepec	--
KJ621432	North America	North America	--	--
KJ621433	North America	North America	--	--
KJ621434	North America	North America	--	--
KJ621435	North America	North America	West Tehuantepec	--
KJ621436	North America	North America	West Tehuantepec	--
KJ621437	North America	North America	West Tehuantepec	--
KJ621438	North America	North America	West Tehuantepec	--
KJ621439	North America	North America	West Tehuantepec	--
KJ621440	North America	North America	West Tehuantepec	--
KJ621441	North America	North America	West Tehuantepec	--
KJ621442	North America	North America	West Tehuantepec	--
KJ621443	North America	North America	West Tehuantepec	--
KJ621444	North America	North America	West Tehuantepec	--
KJ621445	North America	North America	West Tehuantepec	--
KJ621446	North America	North America	West Tehuantepec	--
KJ621447	Central America	Central America	East Tehuantepec	--
KJ621448	Central America	Central America	East Tehuantepec	--
KJ621449	Central America	Central America	East Tehuantepec	--
KJ621450	Central America	Central America	East Tehuantepec	--
KJ621451	Central America	Central America	East Tehuantepec	--
KJ621452	Central America	Central America	East Tehuantepec	--
KJ621453	Central America	Central America	East Tehuantepec	--
KJ621454	Central America	Central America	East Tehuantepec	--
KJ621455	Central America	Central America	East Tehuantepec	--

KJ621456	North America	North America	West Tehuantepec	--
KJ621457	North America	North America	West Tehuantepec	--
KJ621458	North America	North America	West Tehuantepec	Pacific Coast of Mexico
KJ621459	North America	North America	West Tehuantepec	--
KJ621460	North America	North America	West Tehuantepec	--
KJ621461	North America	North America	West Tehuantepec	Pacific Coast of Mexico
KJ621462	Central America	Central America	East Tehuantepec	--
KJ621463	North America	North America	West Tehuantepec	--
KJ621464	Central America	Central America	East Tehuantepec	--
KJ621465	Central America	Central America	East Tehuantepec	--
KJ621466	Central America	Central America	East Tehuantepec	--
KJ621467	Central America	Central America	East Tehuantepec	--
KJ621468	North America	Central America	West Tehuantepec	--
KJ621469	North America	Central America	West Tehuantepec	--
KJ621470	North America	North America	West Tehuantepec	--
KJ621471	North America	North America	West Tehuantepec	--
KJ621472	North America	North America	West Tehuantepec	--
KJ621473	Central America	Central America	--	--
KJ621474	Central America	Central America	--	--
KJ621475	Central America	Central America	--	--
KJ621476	Central America	Central America	East Tehuantepec	--
KJ621477	Central America	Central America	East Tehuantepec	--
KJ621478	Central America	Central America	East Tehuantepec	--
KJ621479	Central America	Central America	East Tehuantepec	--
KJ621480	Central America	Central America	East Tehuantepec	--
KJ621481	Central America	Central America	East Tehuantepec	--
KJ621482	Central America	Central America	East Tehuantepec	--
KJ621483	Central America	Central America	East Tehuantepec	--
KJ621484	Central America	Central America	East Tehuantepec	--
KJ621485	Central America	Central America	East Tehuantepec	--
KJ621486	Central America	Central America	East Tehuantepec	--

KJ621487	Central America	Central America	East Tehuantepec	--
KJ621488	Central America	Central America	East Tehuantepec	--
KJ621489	Central America	Central America	East Tehuantepec	--
KJ621490	Central America	Central America	East Tehuantepec	--
KJ621491	Central America	Central America	East Tehuantepec	--
KJ621492	Central America	Central America	--	--
KJ621493	Central America	Central America	Mainland Belize	--
KJ621494	Central America	Central America	--	--
KJ621495	Central America	Central America	--	--
KJ621496	Central America	Central America	--	--
KJ621497	Central America	Central America	Mainland Belize	--
KJ621498	Central America	Central America	Mainland Belize	--
KJ621499	Central America	Central America	Mainland Belize	--
KJ621500	Central America	Central America	Mainland Belize	--
KJ621501	Central America	Central America	Mainland Belize	--
KJ621502	Central America	Central America	--	--
KJ621503	Central America	Central America	--	--
KJ621504	Central America	Central America	--	--
KJ621505	Central America	Central America	Mainland Belize	--
KJ621506	Central America	Central America	Mainland Belize	--
KJ621507	Central America	Central America	Mainland Belize	--
KJ621508	Central America	Central America	Mainland Belize	--
KJ621509	Central America	Central America	Mainland Belize	--
KJ621510	Central America	Central America	Mainland Belize	--
KJ621511	Central America	Central America	Mainland Belize	--
KJ621512	Central America	Central America	Mainland Belize	--
KJ621513	Central America	Central America	--	--
KJ621514	Central America	Central America	--	--
KJ621515	Central America	Central America	Mainland Belize	--
KJ621516	Central America	Central America	--	--
KJ621517	Central America	Central America	Mainland Belize	--

KJ621518	Central America	Central America	Mainland Belize	--
KJ621519	Central America	Central America	East Tehuantepec	--
KJ621520	Central America	Central America	East Tehuantepec	--
KJ621521	Central America	Central America	East Tehuantepec	--
KJ621522	Central America	Central America	Mainland Belize / East Tehuantepac	--
KJ621523	Central America	Central America	Mainland Belize / East Tehuantepac	--
KJ621524	Central America	Central America	Mainland Belize / East Tehuantepac	--
KJ621525	Central America	Central America	--	--
KJ621526	Central America	Central America	--	--
KJ621527	Central America	Central America	--	--
KJ621528	Central America	Central America	--	--
KJ621529	Central America	Central America	--	--
KJ621530	Central America	Central America	--	--
KJ621531	Central America	Central America	--	--
KJ621532	Central America	Central America	--	--
KJ621533	Central America	Central America	--	--
KJ621534	Central America	Central America	--	--
KJ621535	Central America	Central America	--	--
KJ621536	Central America	Central America	--	--
KJ621537	Central America	Central America	--	--
Boco02	North America	North America	West Tehuantepec	--
Boco03	North America	North America	West Tehuantepec	--
Boco04	North America	North America	West Tehuantepec	--
Boco05	North America	North America	West Tehuantepec	--
Boco09	North America	North America	--	Pacific Coast of Mexico
Boco10	North America	North America	--	--
Boco11	Central America	Central America	--	Mainland Belize
Boco12	--	--		Mainland Belize
Boco13	Central America	Central America	--	Crawl Cay, Belize
Boco14	Central America	Central America	Lagoon Cay	Lagoon Cay, Belize
Boco15	Central America	Central America	West Snake Cay	West Snake Cay, Belize

Boco16	Central America	Central America	West Snake Cay	--
Boco17	Central America	Central America	Mainland Belize	Mainland Belize
Boco18	Central America	Central America	West Snake Cay	West Snake Cay, Belize
Boco19	Central America	Central America	West Snake Cay	West Snake Cay, Belize
Boco20	Central America	Central America	Lagoon Cay	Lagoon Cay, Belize
Boco21	Central America	Central America	Lagoon Cay	Lagoon Cay, Belize
Boco22	Central America	Central America	Lagoon Cay	Lagoon Cay, Belize
Boco23	Central America	Central America	Lagoon Cay	Lagoon Cay, Belize
Boco24	Central America	Central America	West Snake Cay	West Snake Cay, Belize
Boco25	Central America	Central America	West Snake Cay	West Snake Cay, Belize
Boco26	Central America	Central America	Lagoon Cay	Lagoon Cay, Belize
Boco27	Central America	Central America	Mainland Belize	Mainland Belize
Boco28	Central America	Central America	Mainland Belize	Mainland Belize
Boco29	Central America	Central America	--	Crawl Cay, Belize
Boco30	Central America	Central America	West Snake Cay	West Snake Cay, Belize
Boco32	Central America	Central America	Lagoon Cay	Lagoon Cay, Belize
Boco34	North America	North America	--	--
Boco35	North America	North America	--	--
Boco36	North America	North America	West Tehuantepec	--
Boco37	North America	North America	West Tehuantepec	--
Boco38	North America	North America	West Tehuantepec	Pacific Coast of Mexico
Boco39	North America	North America	West Tehuantepec	--
Boco40	North America	North America	West Tehuantepec	--
Boco41	North America	North America	West Tehuantepec	--
Boco42	North America	North America	West Tehuantepec	--
Boco43	North America	North America	West Tehuantepec	--
Boco44	Central America	Central America	East Tehuantepec	Mainland Honduras
Boco45	Central America	Central America	East Tehuantepec	--
Boco46	Central America	Central America	East Tehuantepec	Mainland Honduras
Boco47	Central America	Central America	East Tehuantepec	Mainland Honduras
Boco48	South America	South America	--	--

Boco49	South America	South America	--	--
Boco50	Central America	Central America	East Tehuantepec	Pacific Coast of Guatemala
Boco51	Central America	Central America	Mainland Belize / East Tehuantepac	Mainland Belize
Boco52	Central America	Central America	East Tehuantepec	Pacific Coast of Guatemala
Boco53	North America	North America	West Tehuantepec	--
Boco54	North America	North America	West Tehuantepec	--
Boco55	North America	North America	--	--
Boco56	North America	North America	--	Pacific Coast of Mexico
Boco58	North America	North America	--	--
Boco59	North America	North America	--	Pacific Coast of Mexico
Boco60	North America	North America	--	--
Boco62	North America	North America	--	Pacific Coast of Mexico
Boco63	North America	North America	--	--
Boco64	North America	North America	--	--
Boco67	--	--		Pacific Coast of Mexico
Boco68	North America	North America	West Tehuantepec	Pacific Coast of Mexico
Boco74	Central America	Central America	--	Mainland Honduras
Boco75	Central America	Central America	Mainland Honduras	Mainland Honduras
Boco76	Central America	Central America	--	Mainland Belize
Boco77	Central America	Central America	Mainland Honduras	Mainland Honduras
Boco79	Central America	Central America	--	Mainland Honduras
Boco80	Central America	Central America	Mainland Honduras	Mainland Belize
Boco81	Central America	Central America	--	--
Boco82	--	--		South America
Boco83	North America	North America	--	Sonoroa, Mexico
Boco84	North America	North America	--	Sonoroa, Mexico
Boco85	North America	North America	--	Sonoroa, Mexico
Boco86	North America	North America	--	Sonoroa, Mexico
Boco87	--	--	--	Sonoroa, Mexico
Boco88	--	--	--	Cayos Cochinos Menor
Boco89	Central America	Central America	Cayos Cochinos Menor	Cayos Cochinos Menor

Boco90	Central America	Central America	Mainland Honduras	Mainland Honduras
Boco91	Central America	Central America	Mainland Belize	Mainland Belize
Boco92	Central America	Central America	Cayos Cochinos Menor	--
Boco102	Central America	--	Mainland Belize	--
Boco105	--	--	--	South America

Supplementary Table 5. Statistics on the number of high-quality and mapped Illumina reads for each sample with nuclear data.

Sample ID	Quality-Filtered Reads	Mapped Reads
KJ621458	335,722	332,335
KJ621459	102,190	100,999
KJ621461	1,836,254	1,822,825
Boco09	4,621,367	1,152,374
Boco11	4,151,765	4,126,169
Boco12	341,589	338,965
Boco13	3,897,770	3,872,780
Boco14	191,169	189,709
Boco15	165,194	163,622
Boco17	2,741,293	2,724,617
Boco18	155,548	154,393
Boco19	5,522,205	5,489,923
Boco20	3,949,352	3,923,964
Boco21	2,665,905	2,648,657
Boco22	796,759	790,712
Boco23	1,776,243	1,765,571
Boco24	118,939	117,984
Boco25	3,210,145	3,191,312
Boco26	876,802	863,888
Boco27	3,694,367	3,667,776
Boco28	1,493,224	1,483,028
Boco29	3,630,407	3,606,147
Boco30	300,644	298,076
Boco32	6,407,649	6,366,966
Boco38	902,765	891,872
Boco44	4,638,183	4,612,291
Boco46	1,084,228	1,075,665
Boco47	208,908	206,925
Boco50	2,587,882	2,571,523
Boco51	6,879,632	6,836,845
Boco52	3,184,668	3,158,504
Boco56	737,416	532,816
Boco59	3,751,377	2,043,073
Boco62	1,975,947	1,912,495
Boco67	102,190	100,999
Boco68	117,557	116,571
Boco74	217,137	208,521
Boco75	939,037	930,605
Boco76	506,572	32,531

Boco77	534,656	205,384
Boco80	64,759	40,410
Boco82	3,411,460	2,587,027
Boco83	2,135,507	1,757,646
Boco84	344,433	339,925
Boco85	232,992	204,372
Boco86	158,807	156,728
Boco87	165,679	150,458
Boco88	223,648	43,761
Boco89	2,364,398	2,333,287
Boco90	6,607,528	4,412,125
Boco91	4,450,781	3,180,681
Boco105	729,290	718,664

Supplementary Table 6. Fossil data, including Paleobiology database collection number, date used, clade constrained, and relevant citation, used for divergence dating under the Fossilized Birth-Death model.

Paleobiology DB Collection No.	Date Used (Mya)	Date Estimate Range (Mya)	Location	Formation	Stage	Clade	Citation
N/A	3.00	-	Mexico: Baja California Sur	-	Late/Upper Pliocene	Boa PAC	Miller, W.E., 1980. The late Pliocene Las Tunas local fauna from southernmost Baja California, Mexico. *Journal of Paleontology 54*, 762-805.
N/A	7.00	-	Panama: Panama Canal Basin	Las Cascadas	Early/Lower Miocene	Boa GYAC	Head, J.J., 2012. Fossil evidence for earliest Neogene American faunal interchange: Boa (Serpentes, Boinae) from the early Miocene of Panama. *Journal of Vertebrate Paleontology 32(6)*, 1328-1334.
13346	48.25	46.2 - 50.3	USA: Alabama: Covington Co.	Tallahatta	Bridgerian	Boinae	Holman, J.A., Case, G.R., 1988. Reptiles from the Eocene Tallahatta Formation of Alabama. *Journal of Vertebrate Paleontology 8(3),* 328-333.
16786	43.3	40.4 - 46.2	USA: California: Ventura Co.	Sespe	Uintan	Boinae	Golz, D.J., Lillegraven, J.A., 1977. Summary of known occurrences of terrestrial vertebrates from Eocene strata of southern California. *Rocky Mountain Geology* 15(1), 43-65.
18597	18.2	16.0 - 20.4	USA: Florida: Gilchrist Co.	Alachua	Hemingfordian	Boinae	White, T.E., 1942. The Lower Miocene mammal fauna of Florida. *Bulletin of the Museum of Comparative Zoology* 92(1), 1-49.
27311	19.25	17.5 - 21.0	Argentina: Chubut	Sarmiento	Colhuehuapian	Boinae	Albino, A.M., 1996. Snakes from the Miocene of Patagonia (Argentina) Part I: The Booidea. *Neues Jahrbuch für Geologie und Paläontologie* 199(3), 417-434.
28629	16.435	16.0 - 16.9	Germany: Bavaria	-	MN 4	Boinae	Szyndlar, Z., Schleich, H.H., 1993. Description of Miocene snake from

							Petersbuch 2 with comments on the lower and middle Miocene ophidian faunas of southern Germany. *Stuttgarter Beitrage zur Naturkunde, Series B. Geologie und Palaontologie* 192, 1-47.
36712	25.715	23.0 - 28.4	France: Quercy	-	MP 28	Boinae	Crochet, J.Y., 1974. Les Insectivores des Phosphorites du Quercy. *Palaeovertebrata* 6(1-2), 109-159.
39295	51.9	37.2 - 48.6	Argentina: Chubut	Sarmiento	Middle Eocene	Boinae	Simpson, G.G., 1937. New reptiles from the Eocene of South America. *American Museum Novitates* 927, 1-3.
39662	35.55	33.9 - 37.2	United Kingdom: England	Headon Beds	Late/Upper Eocene	Boinae	Wood, S., 1844. Record of the discovery of an Alligator with several new Mammalia in the Freshwater Strata at Hordwell. *Annals and Magazine of Natural History* 14, 349-351.
48091	61	48.6 - 58.7	Brazil: Rio de Janeiro	Itaboraian	Late/Upper Paleocene	Boinae	Rage, J.C., 1998. Fossil snakes from the Palaeocene of São José de Itaboraí, Brazil. Part I. Madtsoiidae, Aniliidae. *Palaeovertebrata* 27(3-4), 109-144.
48173	17.985	16.0 - 20.0	Czech Republic: Ústí nad Labem	Most	Orleanian	Boinae	Ivanov, M., 2002. The oldest known Miocene snake fauna from Central Europe: Merkur-North locality, Czech Republic. *Acta Palaeontologica Polonica* 47(3), 513-534.
55602	9.433	7.2 - 11.6	Brazil: Acre	Solimões	Tortonian	Boinae	Villanueva, J.B., Souza-Filho, J.P., Negri, F.R., 1990. Novos achados de cetaceos longirrostros no Neogeno do Acre, Brasil. *Boletim do Museu Paraense Emilio Goeldi, Ciencias da Terra* 2, 59-64.
60213	14.81	13.7 - 16.0	Germany: Bavaria	-	MN 5	Boinae	Szyndlar, Z., Rage, J.C., 2003. Non-erycine Booidea from the Oligocene and Miocene of Europe. 1-109.

60215	25.565	23.0 - 28.4	France	-	Chattian	Boinae	Szyndlar Z., Rage, J.C., 2003. Non-erycine Booidea from the Oligocene and Miocene of Europe. 1-109.
67386	9.433	7.2 - 11.6	Brazil: Amazonas	Solimões	Tortonian	Boinae	Cozzuol, M.A., 2006. The Acre vertebrate fauna: Age, diversity, and geography. *Journal of South American Earth Sciences* 21(3), 185-203.
106271	60.2	58.7 - 61.7	Colombia: Guajira	Cerrejón	Middle Peleocene	Boinae	Head, J.J., Bloch, J.I., Hastings, A.K., Bourque, J.R., Cadena, E.A., Herrera, F.A., Polly, P.D., Jaramillo, C.A., 2009. Giant boid snake from the Paleocene neotropics reveals hotter past equatorial temperatures. *Nature* 457(7357), 715-717.
106304	0.3965	0.0 - 0.8	Argentina: Corrientes	Toropí	Lujanian	Boinae	Albino, A.M., Carlini, A.A., 2008. First Record of Boa constrictor (Serpentes, Boidae) in the Quaternary of South America. *Journal of Herpetology* 42(1), 82-88.
134954	14.895	13.7 - 16.0	Germany: Bavaria	Upper Freshwater Molasse	Langhian	Boinae	Ivanov, M., Böhme, M., 2011. Snakes from Griesbeckerzell (Langhian, Early Badenian), North Alpine Foreland Basin (Germany), with comments on the evolution of snake faunas in Central Europe during the Miocene Climatic Optimum. *Geodiversitas* 33(3), 411-449.
136898	24.27	20.4 - 28.4	Turkey: Van	Mendikdere	Chattian	Boinae	Szyndlar, Z., Hösgör, I., 2012. Boine snake Bavarioboa from the Oligocene/Miocene of eastern Turkey with comments on connections between European and Asiatic snake faunas. *Acta Palaeontologica Polonica* 57(3), 667-671.
138666	16.435	16.0 - 16.9	Czech Republic: Karlovy Vary	-	MN 4	Boinae	Szyndlar, Z., 1987. Snakes from the Lower Miocene Locality of Dolnice (Czechoslovakia). *Journal of Vertebrate Paleontology* 7(1), 55-71.

144649	1.29985	0.0 - 2.6	Bahamas: New Providence Island	-	Pleistocene	Boinae	Pregill, G.K., 1982. Fossil Amphibians and Reptiles from New Providence Island, Bahamas. *Smithsonian Contributions to Paleobiology* 48, 8-21.
144663	0.00585	0.0 - 0.0	Bahamas: Great Abaco Island	-	Holocene	Boinae	Steadman, D.W., Franz, R., Morgan, G.S., Albury, N.A., Kakuk, B., Broad, K., Franz, S.E., Tinker, K., Pateman, M.P., Lott, T.A., Jarzen, D.M., 2007. Exceptionally well preserved late Quaternary plant and vertebrate fossils from a blue hole on Abaco, The Bahamas. *Proceedings of the National Academy of Sciences* 104(50), 19897-19902.
167432	25.715	23.0 - 28.4	Tanzania: Mbeya	Nsungwe	Late/Upper Oligocene	Boinae	McCartney, J.A., Stevens, N.J., O’Connor, P.M., 2014. The Earliest Colubroid-Dominated Snake Fauna from Africa: Perspectives from the Late Oligocene Nsungwe Formation of Southwestern Tanzania. *PLoS ONE* 9(3), e90415.
92733	0.00585	0.0 - 0.0	Madagascar: Toliara	-	Holocene	Acrantophis	Burney, D.A., Vasey, N., Godfrey, L.R., Jungers, W.L., Ramarolahy, M.F., Raharivony, L.L., 2008. New findings at Andrahomana Cave, southeastern Madagascar. *Journal of Cave and Karst Studies* 70(1), 13-24.
26723	15.804	11.6 - 20.0	France: Rhône-Alpes	-	Orleanian	Eryx	Ivanov, M., 2000. Snakes of the lower/middle Miocene transition at Vieux Collonges (Rhône, France), with comments on the colonisation of western Europe by colubroids. *Geodiversitas* 22(4), 559-588.
26752	8.4705	5.3 - 11.6	Spain: Granada	-	MN 12	Eryx	Szyndlar, Z., Schleich, H.H., 1994. Two species of the genus Eryx (Serpentes: Boidae; Erycinae) from the Spanish Neogene with comments on the past

							distribution of the genus in Europe. *Amphibia-Reptilia* 15(3), 233-248.
34370	4.2665	3.2 - 5.3	Turkey: Ankara	-	Ruscinian	Eryx	Sen, S.,Bouvrain, G., Geraads, D., 1998. Pliocene vertebrate locality of Calta, Ankara, Turkey. 12. Paleoecology, biogeography and biochronology. *Geodiversitas* 20(3), 497-510.
48175	7.498	7.2 - 7.8	Ukraine: Odessa	-	MN 12	Eryx	Szyndlar, Z., Zerova, G.A., 1992. Miocene snake fauna from Cherevichnoie (Ukraine, USSR), with description of a new species of Vipera. *Neues Jahrbuch für Geologie und Paläontologie, Abhandlungen* 184(1), 87-99.
48446	10.154	8.7 - 11.6	Ukraine: Khmel'nitsk'yi	-	Vallesian	Eryx	Szyndlar, Z., Zerova, G.A., 1990. Neogene cobras of the genus Naja (Serpentes: Elapidae) of East Europe. *Annalen des Naturhistorischen Museums in Wien* 91A, 53-61.
56729	3.9605	2.6 - 5.3	Armenia	-	Pliocene	Eryx	Kharabadze, E., 1997. Fossil snake localities of the Caucasus. *Bulletin of the Georgian Academy of Sciences* 156(1), 151-154.
75600	8.4705	5.3 - 11.6	Mongolia: Övörkhangai	Loh	Late/Upper Miocene	Eryx	Ziegler, R., Dahlmann, T., Storch, G., 2007. Marsupialia, Erinaceomorpha and Soricomorpha (Mammalia). In G. Daxner-Höck (ed.), *Oligocene-Miocene Vertebrates from the Valley of Lakes (Central Mongolia): Morphology, phylogenetic and stratigraphic implications. Annalen des Naturhistorischen Museums in Wien* 108A, 53-164.
136068	4.4665	3.6 - 5.3	Hungary: North Hungary	-	Early/Lower Pliocene	Eryx	Golz, D.J., Lillegraven, J.A., 1977. Summary of known occurrences of terrestrial vertebrates from Eocene

							strata of southern California. *Rocky Mountain Geology* 15(1), 43-65.
136269	16.435	16.0 - 16.9	Spain: Castile-La Mancha	Córcoles	MN 4	Eryx	Szyndlar, Z., Alférez, F., 2005. Iberian snake fauna of the early/middle Miocene transition. *Revista Española de Herpetología* 19, 57-70.
17878	18.2	16.0 - 20.4	USA: South Dakota: Bennett Co.	Rosebud	Hemingfordian	Charina	Green, M., Martin, J.E., 1976. Peratherium (Marsupialia: Didelphidae) from the Oligocene and Miocene of South Dakota. Athlon, essays on palaeontology in honour of Loris Shano Russel, pp.155-168.
18141	14.785	13.6 - 16.0	USA: Nebraska: Brown Co.	Valentine	Barstovian	Charina	Voorhies, M.R., 1990. Vertebrate paleontology of the proposed Norden Reservoir Area, Brown, Cherry and Keya Paha counties, Nebrask*a*. *Technical Report, Division of Archeological Research, Department of Anthropology, University of Nebrask*a 82-09.
18198	7.6	4.9 - 10.3	USA: Texas: Lipscomb Co.	-	Hemphillian	Charina	Schultz, G.E., 1990. Stop 15: Early Hemphillian faunas of the Texas and Oklahoma panhandles. In T. C. Gustavson (ed.), Tertiary and Quaternary stratigraphy and vertebrate paleontology of parts of northwestern Texas and eastern New Mexico; *Guidebook - Bureau of Economic Geology, University of Texas at Austin* 95-103.
20062	3.35	1.8 - 4.9	USA: Washington: Adams Co.	Ringold	Blancan	Charina	Tedford, R.H., Gustafson, E.P., 1977. First North American record of the extinct panda *Parailurus*. *Nature* 265, 621-623.
26758	13.789	11.6 - 16.0	USA: Wyoming: Fremont Co.	Split Rock	Middle Miocene	Charina	Holman, J.A., 1976. Snakes of the Split Rock Formation (middle Miocene), central Wyoming. 32(4), 419-426.

21831	3.094	2.6 - 3.6	Tanzania	Vogel River Series	Piacenzian	Pythonidae	Leakey, M.D., Harris, J.M., 1987. Laetoli: a Pliocene Site in Northern Tanzania. *Clarendon Press, Oxford, Great Britain.*
21855	17.319	11.6 - 23.0	Namibia	-	Early/Lower Miocene	Pythonidae	Pickford, M., Senut, B., Mein, P., Gommery, D., Morales, J., Soria, D., Nieto, M., Ward, J., 1996. Preliminary results of new excavations at Arrisdrift, middle Miocene of southern Namibia. Comptes rendus de l'Académie des sciences. Série 2. Sciences de la terre et des planètes, 322(11), pp.991-996.
22258	3.9605	2.6 - 5.3	Uganda	Warwire	Pliocene	Pythonidae	Tassy, P., 1994. Fossil proboscideans, Mammalia, from the Western Rift,Uganda. *Geology and palaeobiology of the Albertine rift valley* 2, 217-257.
22462	12.809	2.6 - 23.0	Uganda	Nkondo	Miocene	Pythonidae	Bailon, S., Rage, J.C., 1994. Neogene and Pleistocene squamates from the Western Rift,Uganda. *Geology and paleobiology of the Albertine rift valley, B.Senut(ed.)* 2, 129-135.
22469	1.2935	0.8 - 1.8	Tanzania	Olduvai	Calabrian	Pythonidae	Greenwood, P.H., Todd, E.J., 1970. Fish remains from Olduvai. *Fossil Vertebrates of Africa* 2, 225-241.
22596	2.197	1.8 - 2.6	Tanzania	Olduvai	Gelasian	Pythonidae	Todd, N.E., 1996. Dissertation, personal communication.
22628	1.6845	0.8 - 2.6	Tanzania	Olduvai	Early/Lower Pleistocene	Pythonidae	Rage, J.C., 1973. Fossil snakes from Olduvai,Tanzania. In L. S. B. Leakey, R. J. B. Sauvage, and S. C. Coryndon (eds.), *Fossil Vertebrates from Africa* 3, 1-6.
24181	7.098	2.6 - 11.6	Chad	-	Late/Upper Miocene	Pythonidae	Brunet, M., 2000. Chad: discovery of a vertebrate fauna close to the Mio-Pliocene boundary. *Journal of Vertebrate Paleontology* 20(1), 205-209.
26723	15.804	11.6 - 20.0	France: Rhône-Alpes	-	Orleanian	Pythonidae	Ivanov, M., 2000., Snakes of the lower/middle Miocene transition at Vieux Collonges (Rhône, France), with comments on the colonisation of

							western Europe by colubroids. *Geodiversitas* 22(4), 559-588.
28420	13.789	11.6 - 1.0	Australia: Queensland	-	Middle Miocene	Pythonidae	Flannery, T., Archer, M., 1987. Hypsiprymnodon bartholomaii (Potoroidae: Marsupialia), a new species from the Miocene Dwornamor Local Fauna and a reassessment of the phylogenetic position of H. mochatus. In M. Archer (ed.), *Possums and Opossums: Studies in Evolution* 2, 749-758.
32050	3.9605	2.6 - 5.3	Tanzania: Eastern Drift Valley	Vogel River Series	Pliocene	Pythonidae	Leakey, M.D., Harris, J.M., 1987. Laetoli: a Pliocene Site in Northern Tanzania. *Clarendon Press, Oxford, Great Britain.*
47021	19.5	16.0 -23.0	Saudi Arabia	Dam	Early/Lower Miocene	Pythonidae	Thomas, H., Sen, S., Khan, M., Battail, B., Ligabue, G., 1982. The Lower Miocene Fauna of Al-Sarrar (Eastern Province, Saudi Arabia). *Atlal* 5(3a), 109-136
48630	13.789	11.6 - 16.0	Morocco: Tadla-Azilal	-	Astaracian	Pythonidae	Rage, J.C., 1976. Les Squamates du Miocène de Bèni Mellal, Maroc. *Géologie Méditerranéenne* 3(2), 57-70
51266	4.4665	3.6 - 5.3	Australia: Queensland	Allingham	Early/Lower Pliocene	Pythonidae	Scanlon, J.D., 2001. Montypythonoides: the Miocene snake Morelia riversleighensis (Smith & Plane, 1985) and the geographical origin of pythons. *Memoirs of the Association of Australasian Palaeontologists* 25, 1-35
59839	6.2895	5.3 - 7.2	Chad	-	Messinian	Pythonidae	Vignaud, P., Duringer, P., Mackaye, H.T., Likius, A., Blondel, C., Boisserie, J.R., de Bonis, L., Eisenmann, V., Etienne, M.E., Geraads, D., Guy, F., 2002. Geology and paleontology of the Upper Miocene Toros-Menalla hominid locality, Chad. *Nature* 418, 152-155.
83304	0.00585	0.0 - 0.0	Niger	-	Holocene	Pythonidae	Sereno, P.C., Garcea, E.A., Jousse, H., Stojanowski, C.M., Saliège, J.F., Maga, A., Ide, O.A., Knudson, K.J., Mercuri,

							A.M., Stafford Jr, T.W., Kaye, T.G., 2008. Lakeside cemeteries in the Sahara: 5000 years of Holocene population and environmental change. *PLoS ONE* 3(8), e2995.
106497	19.5	16.0 - 23.0	Australia: Queensland	System B	Early/Lower Miocene	Pythonidae	Roberts, K.K., Archer, M., Hand, S.J. & Godthelp, H., 2007. New Genus and Species of Extinct Miocene Ringtail Possums (Marsupialia: Pseudocheiridae). *American Museum Novitates* 3560, 1-15.
135038	14.895	13.7 - 16.0	Germany: Bavaria	Upper Freshwater Molasse	Langhian	Pythonidae	Ivanov, M., Böhme, M., 2011. Snakes from Griesbeckerzell (Langhian, Early Badenian), North Alpine Foreland Basin (Germany), with comments on the evolution of snake faunas in Central Europe during the Miocene Climatic Optimum. *Geodiversitas* 33(3), 411-449.
135708	17.319	11.6 - 23.0	Australia: Queensland	-	Early/Lower Miocene	Pythonidae	Muirhead, J., 1992. A specialised thylacinid, Thylacinus macknessi, (Marsupialia: Thylacinidae) from Miocene deposits of Riversleigh, northwestern Queensland. *Australian Mammalogy* 15, 67-76.
136919	16.435	16.0 - 16.9	France: Midi-Pyrenees	-	MN 4	Pythonidae	Rage, J.C.,Bailon, S., 2005. Amphibians and squamate reptiles from the late early Miocene (MN 4) of Béon 1 (Montréal-du-Gers, southwestern France). *Geodiversitas* 27(3), 413-441.
137618	13.789	11.6 - 16.0	Australia: Northern Territory	Camfield Beds	Middle Miocene	Pythonidae	Schwartz, L.R.S., 2006. Miralinidae (Marsupialia: Phalangeroidea) from northern Australia, including the youngest occurrence of the family. *Alcheringa* 30(2), 343-350.
143209	4.4665	3.6 - 5.3	Australia: Queensland	Allingham	Early/Lower Pliocene	Pythonidae	Willis, P.M.A., Mackness, B.S., 1996. Quinkana babarra, a new species of ziphodont mekosuchine crocodile from the Early Pliocene Bluff Downs local

							fauna, northern Australia with a revision of the genus. *Proceedings of The Linnean Society of New South Wales* 116, 143-151.
143470	25.715	23.0 - 28.4	Australia: Queensland	Carl Creek Limestone	Late/Upper Oligocene	Pythonidae	Willis, P.M.A., 1997. New crocodilians from the late Oligocene White Hunter Site, Riversleigh, northwestern Queensland. *Memoirs of the Queensland Museum* 41(2), 423-438.
151091	13.789	11.6 - 16.0	Australia: Queensland	-	Middle Miocene	Pythonidae	Pian, R., Archer, M., Hand, S.J., 2013. A New, Giant Platypus, Obdurodon tharalkooschild, sp. nov. (Monotremata, Ornithorhynchidae), from the Riversleigh World Heritage Area, Australia. *Journal of Vertebrate Paleontology* 33(6), 1255-1259.

Supplementary Table 7. Sample assignments used for the three models compared using Bayesian Species Delimitation.

Sample ID	Broad Location	Species Assignment Number		
		2 Species Model - A	2 Species Model - B	3 Species Model - C
KJ621458	North America	1	1	1
KJ621459	North America	1	1	1
Boco09	North America	1	1	1
Boco11	Central America	2	1	2
Boco12	Central America	2	1	2
Boco13	Central America	2	1	2
Boco14	Central America	2	1	2
Boco15	Central America	2	1	2
Boco23	Central America	2	1	2
Boco26	Central America	2	1	2
Boco28	Central America	2	1	2
Boco29	Central America	2	1	2
Boco30	Central America	2	1	2
Boco38	North America	1	1	1
Boco44	Central America	2	1	2
Boco46	Central America	2	1	2
Boco47	Central America	2	1	2
Boco50	North America	1	1	1
Boco51	Central America	2	1	2
Boco52	North America	1	1	1
Boco56	North America	1	1	1
Boco62	North America	1	1	1
Boco74	Central America	2	1	2
Boco75	Central America	2	1	2
Boco77	Central America	2	1	2
Boco82	South America	2	3	3
Boco84	North America	1	1	1
Boco86	North America	1	1	1
Boco87	North America	1	1	1
Boco89	Central America	2	1	2
Boco90	Central America	2	1	2
Boco91	Central America	2	1	2
Boco105	South America	2	3	3

Supplementary Table 8. Summary table of highest supported number of source populations (*K*) deduced using the Δ*K* framework for NGSadmix and a DIC framework for Entropy. The values indicated the highest support for particular analyses are bolded.

Populations (K)	DeltaK	DIC
1	NA	264,866.71
2	**15,236,468.09**	193,827.99
3	66,662.37	175,259.30
4	0.40	167,239.80
5	3.26	151,609.41
6	0.38	152,286.56
7	0.51	130,360.14
8	1.61	**123,853.94**

CITATIONS

Arbogast, B. S., Edwards, S. V., Wakeley, J., Beerli, P., & Slowinski, J. B. (2002). Estimating Divergence Times from Molecular Data on Phylogenetic and Population Genetic Timescales. *Annual Review of Ecology and Systematics*, *33*(1), 707–740. https://doi.org/10.1146/annurev.ecolsys.33.010802.150500

Bandelt, H. J., Forster, P., & Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, *16*(1), 37–48. https://doi.org/10.1093/oxfordjournals.molbev.a026036

Barbour, T. (1906). Vertebrata from the Savanna of Panama. Reptilia; Amphibia. *Bulletin of The Museum of Comparative Zoology*, *46*, 211–230. https://doi.org/10.5962/bhl.part.13044

Bauer, A. M. (1993). African-South American relationships: a perspective from the Reptilia. In P. Goldblatt (Ed.), *Biological relationships between Africa and South America* (pp. 244–288). New Haven, CT: Yale University Press.

Bell, M. A., & Lloyd, G. T. (2014). strap: an R package for plotting phylogenies against stratigraphy and assessing their stratigraphic congruence. *Palaeontology*, *58*(2), 379–389. https://doi.org/10.1111/pala.12142

Boback, S. M. (2005). Natural History and Conservation of Island Boas (*Boa constrictor*) in Belize. *Copeia*, *2005*(4), 879–884. https://doi.org/10.1643/0045-8511(2005)005[0879:NHACOI]2.0.CO;2

Boback, S. M. (2006). A Morphometric Comparison of Island and Mainland Boas (*Boa constrictor*) in Belize. *Copeia*, *2006*(2), 261–267. https://doi.org/10.1643/0045-8511(2006)6[261:AMCOIA]2.0.CO;2

Boback, S. M., & Carpenter, D. M. (2007). Body size and head shape of island *Boa constrictor* in Belize: environmental versus genetic contributions. In R. W. Henderson & R. Powell (Eds.), *Biology of the Boas and Pythons* (pp. 102–117). Eagle Mountain, UT: Eagle Mountain Publishing.

Boback, S. M. (2003). Body Size Evolution in Snakes: Evidence from Island Populations. *Copeia*, *2003*(1), 81–94. https://doi.org/10.1643/0045-8511(2003)003[0081:BSEISE]2.0.CO;2

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., … Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology*, *10*(4), e1003537. https://doi.org/10.1371/journal.pcbi.1003537

Bouckaert, R. R. (2010). DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics*, *26*(10), 1372–1373. https://doi.org/10.1093/bioinformatics/btq110

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., … Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1), 1–31. https://doi.org/10.1186/2047-217X-2-10

Brouat, C., Chevallier, H., Meusnier, S., Noblecourt, T., & Rasplus, J. -Y. (2004). Specialization and habitat: spatial and environmental effects on abundance and genetic diversity of forest generalist and specialist *Carabus* species. *Molecular Ecology*, *13*(7), 1815–1826. https://doi.org/10.1111/j.1365-294X.2004.02206.x

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Molecular Biology and Evolution*, *29*(8), 1917–1932. https://doi.org/10.1093/molbev/mss086

Burbrink, F. T. (2005). Inferring the phylogenetic position of *Boa constrictor* among the Boinae. *Molecular Phylogenetics and Evolution*, *34*(1), 167–180. https://doi.org/10.1016/j.ympev.2004.08.017

Burbrink, F. T., Lawson, R., & Slowinski, J. B. (2000). Mitochondrial DNA phylogeography of the polytypic North American rat snake (*Elaphe obsoleta*): a critique of the subspecies concept. *Evolution*, *54*(6), 2107–2118. https://doi.org/10.1111/j.0014-3820.2000.tb01253.x

Burbrink, F. T., McKelvy, A. D., Pyron, R. A., & Myers, E. A. (2015). Predicting community structure in snakes on Eastern Nearctic islands using ecological neutral theory and phylogenetic methods. *Proc. R. Soc. B*, *282*(1819), 20151700. https://doi.org/10.1098/rspb.2015.1700

Cariou, M., Duret, L., & Charlat, S. (2013). Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecology and Evolution*, *3*(4), 846–852. https://doi.org/10.1002/ece3.512

Castoe, T. A., Spencer, C. L., & Parkinson, C. L. (2007). Phylogeographic structure and historical demography of the western diamondback rattlesnake (*Crotalus atrox*): A perspective on North American desert biogeography. *Molecular Phylogenetics and Evolution*, *42*(1), 193–212. https://doi.org/10.1016/j.ympev.2006.07.002

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes, Genomes, Genetics*, *1*(3), 171–182. https://doi.org/10.1534/g3.111.000240

Catchen, J. M., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Colston, T. J., Grazziotin, F. G., Shepard, D. B., Vitt, L. J., Colli, G. R., Henderson, R. W., … Burbrink, F. T. (2013). Molecular systematics and historical biogeography of tree boas (*Corallus* spp.). *Molecular Phylogenetics and Evolution*, *66*(3), 953–959. https://doi.org/10.1016/j.ympev.2012.11.027

Cope, E. D. (1877). Synopsis of the Cold Blooded Vertebrata, Procured by Prof. James Orton during His Exploration of Peru in 1876-77. *Proceedings of the American Philosophical Society*, *17*(100), 33–49.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Daudin, F. M., & Sonnini, C. S. (1802). *Histoire naturelle, générale et particulière, des reptiles : ouvrage faisant suite à l'Histoire naturelle générale et particulière, composée par Leclerc de Buffon, et rédigée par C.S. Sonnini*. A Paris : De l'Imprimerie de F. Dufart,. Retrieved from https://www.biodiversitylibrary.org/bibliography/60678

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*, 214. https://doi.org/10.1186/1471-2148-7-214

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

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, *14*(8), 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Eydoux, F., Souleyet, L. F. A., Bevalet, A.-G., Chazal, A., Gerbe, Z., Delahaye, C., … National Capital Shell Club (Washington, D. C.). (1841). *Voyage autour du monde exécuté pendant les années 1836 et 1837 sur la corvette la Bonite*. Paris, A. Bertrand. Retrieved from http://archive.org/details/Voyageautourdum00Eydo

Fields, P. D., Reisser, C., Dukić, M., Haag, C. R., & Ebert, D. (2015). Genes mirror geography in *Daphnia magna*. *Molecular Ecology*, *24*(17), 4521–4536. https://doi.org/10.1111/mec.13324

Gavryushkina, A., Welch, D., Stadler, T., & Drummond, A. J. (2014). Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration. *PLOS Computational Biology*, *10*(12), e1003919. https://doi.org/10.1371/journal.pcbi.1003919

Gompert, Z., Lucas, L. K., Buerkle, C. A., Forister, M. L., Fordyce, J. A., & Nice, C. C. (2014). Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular Ecology*, *23*(18), 4555–4573. https://doi.org/10.1111/mec.12811

Gordon, A., & Hannon, G. (2010). *Fastx-toolkit.*

Graur, D., & Martin, W. (2004). Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends in Genetics*, *20*(2), 80–86. https://doi.org/10.1016/j.tig.2003.12.003

Haug, G. H., & Tiedemann, R. (1998). Effect of the formation of the Isthmus of Panama on Atlantic Ocean thermohaline circulation. *Nature*, *393*(6686), 673–676. https://doi.org/10.1038/31447

Haug, G. H., Tiedemann, R., Zahn, R., & Ravelo, A. C. (2001). Role of Panama uplift on oceanic freshwater balance. *Geology*, *29*(3), 207–210. https://doi.org/10.1130/0091-7613(2001)029<0207:ROPUOO>2.0.CO;2

Hazkani-Covo, E., Zeller, R. M., & Martin, W. (2010). Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. *PLOS Genetics*, *6*(2), e1000834. https://doi.org/10.1371/journal.pgen.1000834

Head, J. J., Bloch, J. I., Hastings, A. K., Bourque, J. R., Cadena, E. A., Herrera, F. A., … Jaramillo, C. A. (2009). Giant boid snake from the Palaeocene neotropics reveals hotter past equatorial temperatures. *Nature*, *457*(7230), 715–717. https://doi.org/10.1038/nature07671

Head, J. J., Rincon, A. F., Suarez, C., Montes, C., & Jaramillo, C. (2012). Fossil evidence for earliest Neogene American faunal interchange: *Boa* (Serpentes, Boinae) from the early Miocene of Panama. *Journal of Vertebrate Paleontology*, *32*(6), 1328–1334. https://doi.org/10.1080/02724634.2012.694387

Heath, T. A., Huelsenbeck, J. P., & Stadler, T. (2014). The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, *111*(29), E2957–E2966. https://doi.org/10.1073/pnas.1319091111

Henderson, R. W., Waller, T., Micucci, P., Puorto, G., & Bourgeois, R. W. (1995). Ecological correlates and patterns in the distribution of Neotropical boines (Serpentes: Boidae): a preliminary assessment. *Herpetological Natural History*, *3*, 1.

Hey, J., & Nielsen, R. (2004). Multilocus Methods for Estimating Population Sizes, Migration Rates and Divergence Time, With Applications to the Divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, *167*(2), 747–760. https://doi.org/10.1534/genetics.103.024182

Hey, J., & Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences*, *104*(8), 2785–2790. https://doi.org/10.1073/pnas.0611164104

Hull, J. M., Hull, A. C., Sacks, B. N., Smith, J. P., & Ernest, H. B. (2008). Landscape characteristics influence morphological and genetic differentiation in a widespread raptor (*Buteo jamaicensis*). *Molecular Ecology*, *17*(3), 810–824. https://doi.org/10.1111/j.1365-294X.2007.03632.x

Hynková, I., Starostová, Z., & Frynta, D. (2009). Mitochondrial DNA Variation Reveals Recent Evolutionary History of Main *Boa constrictor* Clades. *Zoological Science*, *26*(9), 623–631. https://doi.org/10.2108/zsj.26.623

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, *23*(14), 1801–1806. https://doi.org/10.1093/bioinformatics/btm233

Jezkova, T., Riddle, B. R., Card, D. C., Schield, D. R., Eckstut, M. E., & Castoe, T. A. (2015). Genetic consequences of postglacial range expansion in two codistributed rodents (genus

Dipodomys) depend on ecology and genetic locus. *Molecular Ecology*, *24*(1), 83–97. https://doi.org/10.1111/mec.13012

Keigwin, L. (1982). Isotopic Paleoceanography of the Caribbean and East Pacific: Role of Panama Uplift in Late Neogene Time. *Science*, *217*(4557), 350–353. https://doi.org/10.1126/science.217.4557.350

Langhammer, J. K. (1983). A new subspecies of *Boa constrictor*, *Boa constrictor melanogaster*, from Ecuador (Serpentes: Boidae). *Tropical Fish Hobbyist*, *32*(4), 70–79.

Lazell, J. D. (1964). The Lesser Antillean representative of *Bothrops* and *Constrictor*. *Bulletin of the Museum of Comparative Zoology at Harvard College.*, *132*, 245–273.

Leaché, A. D., Fujita, M. K., Minin, V. N., & Bouckaert, R. R. (2014). Species Delimitation using Genome-Wide SNP Data. *Systematic Biology*, *63*(4), 534–542. https://doi.org/10.1093/sysbio/syu018

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lindemann, L. (2009). Boa constrictor (*Boa constrictor*). Retrieved June 18, 2018, from http://animaldiversity.org/accounts/Boa_constrictor/

Linné, C. von. (1758). *Systema naturae per regna tria naturae: secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis.* Holmiae : Impensis Direct. Laurentii Salvii,. Retrieved from https://www.biodiversitylibrary.org/bibliography/542

Montes, C., Cardona, A., Jaramillo, C., Pardo, A., Silva, J. C., Valencia, V., … Niño, H. (2015). Middle Miocene closure of the Central American Seaway. *Science*, *348*(6231), 226–229. https://doi.org/10.1126/science.aaa2815

Noonan, B. P., & Chippindale, P. T. (2006). Dispersal and vicariance: The complex evolutionary history of boid snakes. *Molecular Phylogenetics and Evolution*, *40*(2), 347–358. https://doi.org/10.1016/j.ympev.2006.03.010

Noonan, B. P., & Chippindale, P. T. (2006). Vicariant Origin of Malagasy Reptiles Supports Late Cretaceous Antarctic Land Bridge. *The American Naturalist*, *168*(6), 730–741. https://doi.org/10.1086/509052

O'Shea, M. (2007). *Boas and Pythons of the World*. Princeton, NJ: Princeton University Press.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, *20*(2), 289–290. https://doi.org/10.1093/bioinformatics/btg412

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLOS ONE*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Philippi, R. (1873). Über die Boa der westlichen Provinzen der Argentinischen Republik. *Zeitschrift Für Gesammten Naturwissenschaften, Berlin*, *41*, 127–130.

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS Genetics*, *8*(11), e1002967. https://doi.org/10.1371/journal.pgen.1002967

Price, R., & Russo, P. (1991). Revisionary comments on the genus *Boa* with the description of a new subspecies of *Boa constrictor* from Peru. *The Snake*, *23*(1), 29–35.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(2), 945–959.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rage, J.-C. (1988). Gondwana, Tethys, and terrestrial vertebrates during the Mesozoic and Cainozoic. *Geological Society, London, Special Publications*, *37*(1), 255–273. https://doi.org/10.1144/GSL.SP.1988.037.01.18

Rage, J.-C. (2001). Fossil snakes from the Palaeocene of São José de Itaboraí, Brazil. Part II. Boidae. *Palaeovertebrata*, *30*(3–4), 111–150.

Rambaut, A. (2015). FigTree (Version 1.4.2).

Ravelo, A. C., Andreasen, D. H., Lyle, M., Olivarez Lyle, A., & Wara, M. W. (2004). Regional climate shifts caused by gradual global cooling in the Pliocene epoch. *Nature*, *429*(6989), 263–267. https://doi.org/10.1038/nature02567

Reed, R. N., Boback, S. M., Montgomery, C. E., Green, S., Stevens, Z., & Watson, D. (2007). Ecology and conservation of an exploited insular population of *Boa constrictor* (Squamata: Boidae) on the Cayos Cochinos, Honduras. In R. W. Henderson & R. Powell (Eds.), *Biology of the Boas and Pythons* (pp. 289–403). Eagle Mountain, UT: Eagle Mountain Publishing.

Reynolds, R. G., Niemiller, M. L., & Revell, L. J. (2014). Toward a Tree-of-Life for the boas and pythons: Multilocus species-level phylogeny with unprecedented taxon sampling. *Molecular Phylogenetics and Evolution*, *71*, 201–213. https://doi.org/10.1016/j.ympev.2013.11.011

Reynolds, R. G., Puente-Rolón, A. R., Reed, R. N., & Revell, L. J. (2013). Genetic analysis of a novel invasion of Puerto Rico by an exotic constricting snake. *Biological Invasions*, *15*(5), 953–959. https://doi.org/10.1007/s10530-012-0354-2

Schield, D. R., Card, D. C., Adams, R. H., Jezkova, T., Reyes-Velasco, J., Proctor, F. N., … Castoe, T. A. (2015). Incipient speciation with biased gene flow between two lineages of the Western Diamondback Rattlesnake (*Crotalus atrox*). *Molecular Phylogenetics and Evolution*, *83*, 213–223. https://doi.org/10.1016/j.ympev.2014.12.006

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating Individual Admixture Proportions from Next Generation Sequencing Data. *Genetics*, *195*(3), 693–702. https://doi.org/10.1534/genetics.113.154138

Slevin, J. R. (1926). Expedition to the Revillagigedo Islands, Mexico, in 1925, III. Notes on a collection of reptiles and amphibians from the Tres Marias and Revillagigedo Islands, and the West Coast of Mexico, with description of a new species of *Tantilla*. *Proceedings of the California Academy of Sciences, 4th Series.*, *15*, 195–207.

Smith, H. M. (1943). Summary of the collections of snakes and crocodilians made in Mexico under the Walter Rathbone Bacon Traveling Scholarship. *Proceedings of the United States National Museum*, *93*(3169). Retrieved from http://repository.si.edu//handle/10088/16420

Stadler, T. (2010). Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, *267*(3), 396–404.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stull, O. G. (1932). Five new subspecies of the family Boidae. *Occasional Papers of the Boston Society of Natural History*, *8*.

Suárez-Atilano, M., Burbrink, F., & Vázquez-Domínguez, E. (2014). Phylogeographical structure within *Boa constrictor imperator* across the lowlands and mountains of Central America and Mexico. *Journal of Biogeography*, *41*(12), 2371–2384. https://doi.org/10.1111/jbi.12372

Toews, D. P. L., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, *21*(16), 3907–3930. https://doi.org/10.1111/j.1365-294X.2012.05664.x

Uetz, P., & Etzold, T. (1996). The EMBL/EBI Reptile Database. *Herpetological Review*, *27*(4), 174–175.

Uetz, P., Hošek, J., & Hallermann, J. (2015). *The reptile database*. Retrieved from http://www.reptile-database.org/

Zweifel, R. G. (1960). Results of the Puritan-American Museum of Natural History Expedition to Western Mexico. 9, Herpetology of the Tres Marías Islands. *Bulletin of the AMNH*, *119*(2). Retrieved from http://digitallibrary.amnh.org/handle/2246/1974

Chapter 5.

Genomic basis of convergent island phenotypes in boa constrictors

Daren C. Card[1], Drew R. Schield[1], Blair W. Perry[1], Richard H. Adams[1], Andrew B. Corbin[1], Giulia I.M. Pasquesi[1], Kristopher Row[1], Juan M. Daza[2], Warren Booth[3], Chad E. Montgomery[4], Scott M. Boback[5], and Todd A. Castoe[1,*]

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

[2] Instituto de Biología, Universidad de Antiochia, 67th Street No. 53 – 108, Medellín, Colombia

[3] Department of Biological Science, University of Tulsa, 800 South Tucker Drive, Tulsa, OK, 74104, USA

[4] Department of Biology, Truman State University, 100 E. Normal Ave., Kirksville, MO, 63501 USA

[5] Department of Biology, Dickinson College, P.O. Box 1773, Carlisle, PA, 17013, USA

ABSTRACT

Major biological paradigms have been developed by studying the diversification of island flora and fauna. The unique ecological conditions and isolation of island systems make island fauna well-suited for studying rapid and convergent evolution in ecology, physiology, body size, and other natural history characteristics under strong local selection on islands. Here we use complementary genomic approaches to understand the contribution of genetic drift and adaptation, as well as idiosyncratic versus convergent molecular evolution, in the evolution of morphological, physiological, and natural history traits shared across distinct island populations of *Boa imperator*. We used high-density restriction-site associated DNA sequencing to establish evidence for the independent evolution of insular traits within three island populations and used demographic analyses to infer the relative roles of drift and selection in shaping genomic differentiation between island and mainland lineages. We also use whole-genome resequencing data to identify regions of unique island-specific allelic fluctuation that contain genes with phenotypically-relevant mutations, and these genes display statistical enrichment for molecular phenotypes associated with island traits. By intersecting gene sets from distinct insular populations, we also identify genes with significant associations with phenotypes across islands, including four candidate genes putatively underlying body size reduction in all three islands. The molecular pathway-level correspondence between our implicated genes and genes already deduced as important in other model and non-model systems indicates that convergent molecular mechanisms are capable of impacting similar traits in convergent and divergent fashions across diverse animal taxa.

INTRODUCTION

Island systems have been fundamental to the development of numerous disciplines in evolutionary biology, including colonization of novel habitats (Diamond, 1972), selection and migration dynamics (King, 1987), and adaptive radiation (Losos, Warheitt, & Schoener, 1997; Seehausen, 2006), due primarily to their geographic isolation, ecological simplicity, assemblages of unique and derived taxa, and replication. Island fauna often exhibit unique phenotypes due to their isolation and the ecological simplicity or uniqueness of island environments, which is known as the island syndrome (Adler & Levins, 1994; Lomolino, Riddle, Whittaker, & Brown, 2010). Variation in body size between mainland and island populations, for example, is well-known and widespread in diverse island fauna. This phenomenon, termed the island rule (Foster, 1964), describes how small species of vertebrates tend to grow larger on islands (i.e., gigantism) and larger species tend to become smaller on islands (i.e., dwarfism; Foster, 1964; Lomolino et al., 2013; Lomolino, Sax, Palombo, & Geer, 2012). Island dwarfism has received considerable attention, and the shift in body size is thought to arise from adaption in response to limited resources on islands (Boback & Guyer, 2003; Köhler & Moyà-Solà, 2009) or ecological character displacement (Grant & Grant, 2006). In addition to body size, many other phenotypic and ecological traits have been shown to undergo major shifts in island populations. For example, coloration (King, 1987) and reproductive output (Covas, 2012) have been show to undergo island-specific adaptations in various island populations of vertebrates.

In this study, we investigate the genomic basis for repeated evolution of highly distinct island eco-morphotypes found in multiple island populations of Central American boas (Figure 1A-C). Snakes in the genus *Boa* are widespread throughout the New World and are well known for their large size and particularly robust phenotype. *Boa imperator*, which is found throughout Central America (Card et al., 2016; Hynková, Starostová, & Frynta, 2009; Reynolds, Niemiller, & Revell, 2014; Suárez-Atilano, Burbrink, & Vázquez-Domínguez, 2014), has colonized dozens of off-shore islands (Henderson, Waller, Micucci, Puorto, & Bourgeois, 1995; Porras, 1999), including several off the coasts of Belize and Honduras, though the number of independent dispersal and colonization events is unknown. Many of these islands lie on the continental shelf and became isolated when sea levels rose at the end of the last glacial maximum (6.5 kya; Gischler, 2014; Mazzullo, 2006).

Based on detailed studies of natural history from the most thoroughly studied populations in Belize, island *B. imperator* vary significantly in morphology and ecology from nearby mainland populations. In both Belize and Honduras, overall body size is much smaller on islands and the ratio of body mass to overall length is particularly reduced on islands (Boback, 2005, 2006; Boback & Carpenter, 2007; Figure 1D). Further, relative tail length is greater on islands and various degrees of snout attenuation or craniofacial divergence is apparent across Belize islands (Boback, 2006; Figure 1E). The evolution of more slender snakes with longer tails corresponds with features of snake arboreality (Lillywhite & Henderson, 2002; Shine, 1983). Snout attenuation is also well known in snake species that prey upon fast-moving prey, as it aids in visual hunting (Henderson & Binder, 1980). Indeed, island boas from Belize are largely arboreal and feed on one of a few available prey species that are significantly smaller and more fast-moving than typical mainland prey items – adult snakes subsist primarily on migratory passerine birds (the largest of the prey options; Boback, 2005; Lillywhite & Henderson, 2002). Island boas also vary in coloration and can be lighter or darker than normal mainland populations, depending on the island population (Porras, 1999; Figure 1F). Finally, island populations have reduced litter sizes and produce offspring with lower masses and shorter bodies than mainland populations (Boback, 2005). These traits are all apparently heritable outside of natural island conditions (Boback & Carpenter, 2007), suggesting a genetic basis for trait variation.

Here we leverage phenotypically differentiated island populations in Belize and Honduras, and nearby mainland populations, to answer four main questions about the evolution of genomic and phenotypic variation in island populations: (1) What is the demographic history that has given rise to these island populations?, (2) Do patterns of insular genomic differentiation support selection, in addition to genetic drift, contributing to island population evolution?, (3) Can genomic variation be linked to phenotypic differences between island and mainland populations, and is such variation shared across multiple island populations?, and (4) Considering the many derived phenotypes shared across island populations, what role has convergent evolution played in the parallel evolution of these phenotypes?

MATERIALS & METHODS

Population sampling and DNA extraction

We obtained tissue from 44 *Boa* samples from populations in Central America, including seven island samples each from West Snake Cay and Lagoon Cay in Belize, four samples from the adjacent mainland of Belize, 15 island samples from Cayos Cochinos in Honduras, five samples from the adjacent mainland populations in Honduras and Nicaragua, and four "outgroup" samples to each island-mainland pair obtained from Guatemala and El Salvador. West Snake Cay and Lagoon Cay are located approximately 5 – 10 km off the cost of Belize and are separated by approximately 60 km. Cayos Cochinos is located approximately 15 km off the coast of Honduras, approximately 200 km or greater from the other two islands, and is comprised of two sister islands: Cayos Cochinos Menor and Cayos Cochinos Major. Tissue was in the form of blood samples obtained from wild-caught individuals from Belize that are maintained in a colony at Dickinson College, blood samples obtained from snakes on Cayos Cochinos, skin shed samples obtained from direct descendants of wild-caught *Boa* from Central America, and samples of preserved liver or muscle from vouchered specimens at the University of Texas at Arlington Amphibian and Reptile Diversity Research Center (see Supplementary Table 1 for details). DNA was extracted from using either a Zymo Research Quick-gDNA Miniprep kit (Zymo Research, Irvine, CA, USA) according to the manufacturer's protocol or a standard phenol-chloroform-isoamyl alcohol extraction.

Restriction-site associated DNA library preparation and sequencing

We used double digest Restriction-site Associated DNA sequencing (RADseq hereafter), following the protocol of Peterson *et al.* (2012). *Pst*I and *Sau*3AI restriction enzymes were used to digest genomic DNA, and to the resulting fragments we ligated to double-stranded adapters containing barcodes and unique molecular identifiers (UMIs; eight consecutive random nucleotides prior to the ligation site). Samples were pooled into groups for efficient size selection for fragments ranging from 570 to 690bp using a Blue Pippin (Sage Science, Beverly, MA, USA), a range that was expected to yield approximately 200,000 loci based on an *in silico* digestion of the *Boa constrictor* reference genome (Bradnam et al., 2013). We used a Bioanalyzer (Agilent, Santa Clara, CA, USA) to quantify and pool libraries, which were sequenced using 100 bp paired-end reads on an Illumina HiSeq 2500 (Illumina Inc., San Diego, CA, USA).

RADseq data analysis and variant calling

We used the clone_filter module from the Stacks v. 1.42 pipeline (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011; Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013) to filter out PCR replicates based on raw read UMIs, which were subsequently trimmed off using the FASTX Toolkit trimmer v. 0.0.13 (A. Gordon & Hannon, 2010). The process_radtags module from Stacks was used to parse reads by index, and default options were used except that the "rescue" feature was activated and the restriction digest site check was disabled. Parsed reads were filtered for RADseq adapter and primer sequences and were quality trimmed using Trimmomatic v. 0.33 (Bolger, Lohse, & Usadel, 2014) using the settings LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:36. We used NextGenMap (Sedlazeck, Rescheneder, & von Haeseler, 2013) with default settings to map the quality-trimmed reads to the *B. constrictor* reference genome (Assemblethon2 team SGA assembly; Bradnam et al., 2013).

We identified single nucleotide polymorphisms (SNPs) and short insertions/deletions (InDels) using the 'GATK Best Practices' workflow (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013). We used the GATK pipeline to perform local indel realignment (with default settings) and joint genotyping from individual GVCFs using HaplotypeCaller to infer variants. We filtered the resulting variants using samtools (Li, 2011; Li et al., 2009) and vcftools (Danecek et al., 2011), as follows. We excluded SNPs within 3 bp of an InDel and clusters of InDels within 10 bp windows. Variants with a PHRED quality score below 30, a read depth less than 500 or greater than 100 (approximately half or double the average coverage of 221), and variants not passing a series of stringent hard filters: QD<2, FS>60.0, MQ<40.0, MQRankSum<-12.5, or ReadPosRankSum<-8. We also required variants to be biallelic and coded genotypes as missing data when individual sample coverage fell below 5x. ThetaMater (Adams et al., 2018) was used to identify loci with excess variation indicative of read mapping derived from paralogous regions, which was indicated by three variants or greater in a single RAD locus. The resulting dataset contained 187,221 variants.

Evaluating demographic models to assess island population independence

Previous work has indicated that the island populations in Belize (Lagoon and West Snake Cays) and in Honduras (Cayos Cochinos) fall into different Central American clades with significant genetic divergence (Card et al., 2016), but the independence of individual islands in each of these two clades (e.g., two islands in Belize) has not been previously assessed. To understand population genetic structure across samples collected for this study, including allelic differentiation among islands within a region, we generated a population tree using SNAPP (Bryant, Bouckaert, Felsenstein, Rosenberg, & RoyChoudhury, 2012) and the population assignments are provided in Supplementary Table 1. In SNAPP, we ran the MCMC for a total of 10 million generations, sampling every 1,000 generations and assessed posterior convergence and stationarity using Tracer (Drummond & Rambaut, 2007). We discarded the first 25% of generations as burn-in and used the remaining MCMC samples to produce a maximum clade credibility consensus with median node heights.

To estimate the demographic histories of island populations, we analyzed the two-dimensional allele frequency spectrum (2D AFS) and tested models of different demographic scenarios using δaδi (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009). We tested eight competing models with various numbers of estimated population size, migration, and divergence time parameters – seven of these models involved population splitting, and we tested a single model scenario without inter-island population divergence (Supplementary Tables 2-3). We tested these models in two parallel analyses: one between the Lagoon and West Snake Cays off the coast of Belize, and one between Cayos Cochinos Menor and Major off the coast of Honduras. For each analysis, we down-sampled to 10 alleles per population in δaδi, which retained 3,565 variants for downstream analysis in the Lagoon and West Snake Cay comparison and 2,228 variants for downstream analysis in the Cayos Cochinos Menor and Major comparison. We then used δaδi to fit each of the eight demographic scenarios to the 2D AFS and used the Nelder-Mead method (Nelder & Mead, 1965) to generate 20 sets of parameter perturbations over a maximum of 50 iterations. We then performed two additional parameter optimization steps using the highest scoring parameter estimates per model from each previous round. The 2D AFS was simulated for each optimized parameter set using a [40,50,60] grid size. Log-likelihoods of model fit were then estimated using the multinomial approach and we assessed the fit of each model using the Akaike Information Criterion (AIC) using the log-likelihood of the highest scoring replicate per model. These analyses were performed using modified two-population δaδi model scripts initially reported in Portik et al. (2017).

RADseq-based calculation of population genetic statistics and identification of signatures of selection

Using our RADseq variant dataset, we thinned variants to reduce confounding effects of linkage by keeping the first variant within a 10 kb window and allowed for 25% missing data across all samples when calculating population genetic statistics. We calculated Weir and Cockerham's (1984) measure of F_{ST} between each island population and the adjacent mainland populations, and between the two mainland populations, using the pegas package (v. 0.10; Paradis, 2010) in R v. 3.4.1 (R Core Team, 2018). To test for evidence that selection, in addition to genetic drift, contributed to population divergence, we explicitly tested whether a null model of neutral genetic drift alone explained divergence between pairs of populations using GppFst (Adams, Schield, Card, Blackmon, & Castoe, 2017). GppFst conducts posterior predictive simulations (PPS) of F_{ST} under a model of divergence between two populations with subsequent evolution only through mutation and drift. To conduct the PPS, we estimated the divergence times and population parameters for each pair of populations ($\tau_{\mathrm{pop1\text{-}pop2}}$, θ_{pop1}, θ_{pop2}, $\theta_{\mathrm{pop1\text{-}pop2}}$) via Markov chain Monte Carlo (MCMC) sampling implemented in SNAPP (Bryant et al., 2012) using variant data produced with the same thinning and missing data constraints outlined above. We ran the MCMC for a total of 10 million generations, sampling every 1,000 generations and assess posterior convergence and stationarity using Tracer (Drummond & Rambaut, 2007). We retained an appropriate number of post burn-in MCMC generations to match the number of loci in each comparison, using these samples to generate a PPS F_{ST} distribution under neutrality. For each MCMC step, we simulated 10 independent loci with lengths drawn from the empirical locus length distribution under a JC69 model using the R package phybase (Liu & Yu, 2010) with random sampling of individuals according to the empirical distributions of locus allele counts for each population. Using these PPS data, we set a conservative threshold of 97.5% F_{ST} percentile, where all variants with F_{ST} values greater than this threshold were considered to be under selection. We then computed the probability of observing the number of empirical variants given the counts of F_{ST} values above the 97.5% percentile obtained from the PPS simulations.

Whole genome resequencing library preparation, sequencing, and data processing

We augmented our RADseq variant dataset by conducting whole genome resequencing (WGS) on a subset of individuals to enable identification of putatively causal genetic variants in selected candidate genes. We targeted 10-15x genomic coverage per individual for two island samples from each Belize island (Lagoon and West Snake Cays), four island samples from Cayos Cochinos, Honduras, and for 6 mainland samples each from the Belize and Honduras clades (N= 20 samples across all populations). We shotgun genome sequencing libraries were produced using either a KAPA HyperPlus or an Illumina Nextera library preparation kit, following the manufacturers protocols. Samples were pooled in equal molar ratios into combined libraries, which were sequenced on the Illumina HiSeq X platform. We followed essentially the same data analysis process outlined above for the RADseq data, but with duplicated reads resulting from PCR being filtered away following mapping to the *Boa* reference genome using the Picard MarkDuplicates tool. Final variants were called using HaplotypeCaller based on the GVCF files of individual samples, and we filtered variants using identical settings to the RADseq data, except that loci with total read depths less than 500 or greater than 125 (half or double the average coverage of 250) were excluded. The resulting variant dataset contained 8,146,817 variants.

Quantifying parallel island allele frequency fluctuation from WGS data

For each island population we calculated the allele frequency fluctuation based on our WGS dataset between the island population and the adjacent mainland population, allowing up to 10% missing data. We took a sliding window approach to evaluate whether extreme allele frequency fluctuations occurred in parallel in two or more populations. The maximum allele frequency change for non-overlapping 10 kb windows was recorded for each population, and we identified windows with allele frequency changes of 0.90 or greater in each population (i.e., instances were an allele was fixed, or nearly fixed, in an island population relative to the mainland). We also quantified the number of windows in which multiple islands experienced an allele frequency shift of ≥ 0.90 and assessed the degree of overlap among multiple islands using the Jaccard index. To better understand whether more overlap was observed than is expected by random chance, we randomly selected windows for each population at the same frequencies observed in the empirical dataset and measured the Jaccard index of overlap. By permutating this analysis 100 times, we established a distribution of expected Jaccard indices under a null model of no parallel evolution, and we compared our empirical results to this null distribution.

Predicting the effects of coding variation estimated from WGS data

We expected that variants in protein coding regions that are unique to islands may contain variants in genes that cause phenotypically relevant island-specific traits, like dwarfism. Based on variants from our whole genome resequencing data and gene models from the *B. constrictor* genome annotation, we used the Variant Effect Prediction (VEP v. 91.1; McLaren et al., 2016) program to identify the locations and infer the relative consequences and impacts of all identified variants according to established Sequence Ontologies (SOs). Similarly, we also ran PROVEAN

v. 1.1.5 (Choi, 2012; Choi, Sims, Murphy, Miller, & Chan, 2012), with the dependencies CD-HIT (v. 4.6; Li & Godzik, 2006) and BLAST (v. 2.2.28+; Altschul, Gish, Miller, Myers, & Lipman, 1990), to estimate the relative likelihood of a phenotypic impact of coding variants, based on evolutionary conservation inferred from the NCBI non-redundant protein database (downloaded 29 January, 2018). We used protein sequences from the *Boa* genome annotation gene models and variants were summarized from the results of the VEP analysis and encoded based on recommendations from the Human Genome Variation Society. Variants resulting in the gain or loss of stop codons or where one or more amino acid was left as unresolved in VEP (i.e., 'X' residues resulting from frameshift variants) were excluded because they are incompatible with PROVEAN. Following the recommendations of the creators of PROVEAN, we used a threshold of -2.5 for binary classification of deleterious (-2.5 or below) versus neutral (above -2.5) variation.

Integrating RADseq and WGS data to identify genomic regions with loci that may underlie island phenotypes

We combined information from across analyses in several ways to identify genomic regions that show strong evidence of containing loci involved with island morphological evolution. We anchored our inferences to the WGS dataset given the density of variants and the potential for identifying functionally-relevant variation (e.g., variants in coding regions). For each island population and for pairwise comparisons between island populations we retained 10 kb window that exhibited a large (0.9 or greater) allele frequency fluctuation between each island and mainland population. We did a similar analysis between the two mainland populations from Belize and Honduras. We searched up to 100 kb in either direction of any windows with high allelic differentiation for annotated protein-coding genes and extracted functionally-relevant variants that met three conditions: (1) showed a 0.75 or greater allele frequency fluctuation in the target population, (2) was annotated as a high impact coding variant according to the VEP analysis, and (3) was annotated as a moderate coding impact variant (i.e., non-synonymous variants) in the VEP analysis and also had a deleterious PROVEAN score. Finally, for each comparison involving island populations we filtered the resulting gene sets to eliminate genes that were also detected in regions of extreme (i.e., ≥0.90) allele frequency fluctuation in a comparison of the two mainland populations (Belize and Honduras) to eliminate hits in regions where allele frequencies fluctuated due to population processes other than selection, as our GppFst analyses indicated that selection has not driven divergence between mainland populations (see Results and Discussion section for more details).

We further performed a similar analysis as described above on pairs of island populations, dictating that the window-based extreme allele frequency had to occur in both populations of interest in order for a window to be further considered (i.e., an "and" statement between populations). As above, we scanned for nearby genes and kept functionally-relevant variants using identical criteria except that coding-region variants with a 0.75 or greater allele frequency fluctuation were kept if they occurred in at least one of the two populations and not in the mainland versus mainland comparison. Finally, we were also interested in understanding whether there were any regions of extreme allelic differentiation that were shared across all three island populations, which may contain genes important for convergent genetic evolution. Given our null expectation that parallel patterns of allele frequency fluctuation from all three island populations should be rarely observed or totally unobserved, we less stringently combined datasets to isolate potential regions involved with island phenotypes across all three islands. We identified windows and neighboring genes in the same manner as other comparisons, but we related the stringency of our filtering of coding-region variants such that all variants with a moderate or high VEP impact classification were retained (i.e., no filtering by coding variant allele frequency or PROVEAN scores). Because this filtering scheme retained more genes and coding variants than we expected *a priori*, we also produced a second dataset following the same full filtering steps that were used to produce single-island and island-pair subsets (details are noted above). In both of these datasets we did not exclude genes that were also found in the comparison of the two mainland populations given that the numbers of genes were small enough to evaluate candidates on a case-by-case basis.

Functional analysis of genes putatively involved in the evolution of island phenotypes

We inferred the identity of all genes identified by our filtering approach using reciprocal and one-way stringent best-BLAST matches between the B. constrictor gene annotation and human. Human gene symbols were extracted for each gene and where necessary we translated gene symbols to human or mouse Ensembl identifiers using BioMart with Ensembl release 92. For each gene set, we determined whether there was enrichment for particular gene ontology (GO) terms, KEGG pathways, and mammalian phenotypes. We used WebGestalt (Zhang, Kirov, & Snoddy, 2005) to identify enriched GO terms or KEGG pathways using an overrepresentation enrichment analysis with the full human protein-coding set as the background reference set and a minimum number of genes for a category of 2. We evaluated enrichment in phenotypic terms associated with both mouse and human using modPhEA (Weng & Liao, 2017), with our search covering all phenotypic levels and using the full set of reference genome genes for the background. For all enrichment analyses, p-values were corrected based on the method of Benjamini and Hochberg (1995) and we retained terms, pathways, or phenotypes with FDR p-values less than 0.1 as significantly enriched. We also extracted all mouse phenotype data associated with gene sets using the Mouse Genome Informatics (Smith, Blake, Kadin, Richardson, & Bult, 2018) batch query. We visualized the relative frequencies of different mouse phenotype terms using frequency histograms to ascertain whether our gene sets contain any signal for phenotypes that fit expectations derived from our knowledge of phenotypic evolution in island populations.

RESULTS AND DISCUSSION

Demographic analyses of genomic variation support independent dwarfism on Belize islands

Boas have colonized at least 43 islands across Central America, yet the exact number of independent island colonization events is unknown. We previously demonstrated that island populations from Belize and Honduras cluster into distinct Central American clades with significant divergence (Card et al., 2016). However, the question remains of whether distinct island populations in Belize and Honduras, which are relatively close geographically, have evolved independently following isolation from the mainland. We therefore conducted demographic analysis of two island populations from the barrier island system of Belize (Lagoon and West Snake Cays) and from the twin island system of Cayos Cochinos (Menor and Major). Our SNAPP analysis produced a consensus tree where all nodes were resolved with 100% posterior support, indicating well-defined and genetically distinct populations (Figure 2A). In Belize, the estimated median θ for each mainland population (0.0102 for mainland clade 1 and 0.0235 for mainland clade 2) was at least double the median θ for each island population (0.0011 and 0.0034 for Lagoon and West Snake Cays, respectively). Median divergence between any pairwise population comparison within the Belize or Honduras clades ranged from τ=0.0001 to τ=0.0002 coalescent units. In Honduras, the median θ for Cayos Cochinos Menor was similar to that of the nearby mainland population (0.0161 and 0.0191, respectively), while θ for Cayos Cochinos Major was less than half as large (0.0065). We detected effectively no divergence between the two Cayos Cochinos islands (median τ=0), suggesting that migration between these two islands (separated by less than 2 km) is relatively high, and the island populations had a median divergence from the mainland of τ=0.0002 coalescent units. The median divergence between the Belize and Honduras clades was τ=0.0005 coalescent units.

We also performed demographic modeling of these island populations using the 2D AFS in δaδi to compare models with and without gene flow. Demographic modeling of the Lagoon and West Snake island populations resulted in a best-fit model of population divergence without migration, further indicating that the two populations have evolved independently since they colonized these islands (Fig. 2B; Supplementary Table 2). Unscaled population size estimates suggest similar effective population sizes on these two islands, but with a slightly greater effective size for the West Snake population, and a relatively recent divergence time. In contrast to evidence for independent colonization and evolution of the Belize islands, we found that a best-fit model of secondary contact with asymmetric gene flow was supported for the two Cayos Cochinos islands (Fig. 2C; Supplementary Table 3), indicating gene flow in both directions between the islands of Menor and Major, but with a higher migration rate from Major to Menor. Population size estimates in this comparison also indicate a larger population size on the island of Major. Overall, demographic analyses highlight the independent colonization and evolution of the

Belize island populations and the ongoing gene flow between the two Cayos Cochinos islands off the coast of Honduras. These results support the hypothesis that the two Belize island populations have evolved in isolation from one another since divergence. This brings the confirmed number of independent dwarf boa populations to three: Lagoon Cay, West Snake Cay, and Cayos Cochinos. These analyses also highlight very small effective population sizes on these small islands; in Belize, snakes inhabit cays that range in size from approximately 5 to 25 hectares and previous estimates of census population sizes were estimated at ≤100 individuals (Boback, 2005).

Divergence, demography, and the roles of selection versus drift on islands

Given the small relative effective population sizes of island boa populations, we tested whether genetic drift alone could explain patterns of genetic divergence on islands, or if there was evidence that natural selection also contributed to island population divergence. The results of our GppFst PPS analyses (Figure 3 and Supplementary Table 4) indicate that the average expected neutral allelic differentiation between each island and mainland population pair was for lower on the Belize islands (Lagoon Cay mean F_{ST} =0.07 and median F_{ST}=0.04; West Snake Cay mean F_{ST} =0.03 and median F_{ST}=0.005) than Cayos Cochinos (mean F_{ST}=0.12 and median F_{ST} =0.08). The allelic differentiation between the two mainland populations was relatively small compared to the island-mainland population pairs (mean F_{ST}=0.07 and median F_{ST} =0.03). Based on our PPS, neutral genetic drift between populations was capable of producing measures of F_{ST} that were as high as 0.75 to 1.0, depending on the comparison, but these extreme measures were quite rare (less than 5% of PPS loci had F_{ST} greater than 0.5; Figure 3).

Our GppFst analysis allowed us to assess evidence of genome-wide selection while taking background patterns of neutral genetic drift into account. The 97.5% quantile threshold for empirical F_{ST} values ranged from 0.35 to 0.75 in the island-mainland comparisons (Figure 3 and Supplementary Table 4). In each of these comparisons, the top 2.5% tail of empirical F_{ST} values contained significantly more loci than expected in the absence of selection (binomial test $p<0.05$ in all island-mainland comparison). With the exception of the West Snake Cay vs. Mainland Belize comparison, each other island-to-mainland comparisons had at least a two-fold excess of observed loci with extreme F_{ST} in the empirical dataset than neutral simulations, with the West Snake Cay vs. Mainland Belize comparison having an excess of ~25 loci. In contrast, the Mainland Belize versus Mainland Honduras comparison, had no significant excess of loci in the top 2.5% tail, with the number of expected loci due to drift almost exactly matching the number of observed loci (binomial test $p=0.42$; Figure 3 and Supplementary Table 4). While drift is capable of producing high F_{ST} values (Nosil, Funk, & Ortiz-Barrientos, 2009), we found evidence that natural selection has occurred in island populations, leading to a greater than expected number of loci with high allelic differentiation on islands.

Single-island candidate regions contain genes with links to island-specific phenotypes

To identify functional genomic links between genes and phenotypes important for adaptive evolution in island lineages we used our WGS data to identify genes in genomic regions with high allele frequency changes in island populations. We found 4,278, 3,848, and 6,887 10-kb windows with extreme allelic fluctuations (≥0.90) in the Lagoon Cay, West Snake Cay, and Cayos Cochinos populations, respectively (Figure 4A-C, E), and 6,678 such windows between the two mainland populations (Figure 4D-E). Genomic windows identified in the Lagoon Cay population were associated with 105 nearby genes (within 100 kb upstream/downstream) with 339 impactful coding variants. We used 98 of these genes with confident human homologs to test for functional enrichment against full genome protein-coding gene backgrounds, but did not find enrichment for any GO terms, KEGG pathways, or mouse knockout phenotypes. One human phenotype (dicarboxylic aciduria [HP:0003215]) was enriched and may be related to reduced body sizes on islands, as dicarboxylic aciduria is associated with non-ketotic hypoglycemia (Divry et al., 1983; Duran, Klerk, Wadman, Bruinvis, & Ketting, 1984; Rhead, Amendt, Fritchman, & Felts, 1983; Figure 5 and Supplementary Table 5). In West Snake Cay, 95 genes were located near genomic windows with extreme allelic fluctuation and included 287 phenotypically-relevant coding variants. We found enrichment for one mouse and one human enriched phenotype, abnormal liver cholesterol level (MP:0012776) and arterial calcification (HP:0003207), respectively, providing a potential link between these genomic regions and the reduced body size found on this island (Boback, 2005; Figure 5 and Supplementary Table 6). We also observed enriched mouse phenotypes tied to reproduction (abnormal ovarian folliculogenesis [MP:0001130]), which may be linked to the significantly reduced litter sizes observed in island populations (Boback, 2005; Figure 5 and Supplementary Table 6). In regions surrounding high allele frequency fluctuations in the Cayos Cochinos population we found 177 nearby genes with 884 associated coding variants. These regions contained the greatest number of enriched mouse phenotypes, but lacked enriched GO terms, KEGG pathways, and human phenotypes. Several of the enriched mouse phenotypes were related to tooth morphology (long incisors [MP:0004831], abnormal lower/upper incisor morphology [MP:0030136/MP:0030137], and macrodontia [MP:0030091]), which may be linked to a shift in prey type (Figure 5 and Supplementary Table 7). Several additional phenotypes tie these regions to the distinct pigmentation that is characteristic of island systems (abnormal skin pigmentation [MP:0002095], abnormal coat/hair pigmentation [MP:0002075], and white spotting [MP:0002938]; Figure 5 and Supplementary Table 7). Moreover, there is a broad enriched mouse phenotype related to body size (growth/size/body region phenotype [MP_0005378]) and another linked to abnormal tail morphology (MP:0002111), and each may correspond to differences in island boa body size and tail length (Boback, 2005, 2006; Figure 5 and Supplementary Table 7). There were also many enriched phenotypes in both West Snake Cay and Cayos Cochinos linked to immunity, and while immune function has not been analyzed in island populations, immune-related genes are often detected in evolutionary comparisons between isolated lineages (Fumagalli et al., 2011; Hurst & Smith, 1999; McTaggart, Obbard, Conlon, & Little, 2012; Obbard, Welch, Kim, & Jiggins, 2009; Schlenke & Begun, 2003). Our results therefore indicate that regions of high island allelic fluctuation appear to be enriched for genes relevant to island phenotypes. However, most phenotypes were mutually exclusive between islands in this analysis, indicating that independent molecular mechanisms may have led to convergent island phenotypes across islands.

Evidence for convergent allelic shifts in protein-coding genes with associations with island boa phenotypes

To better understand whether phenotypic convergence between independent island boa population is a product of convergent molecular evolution, we intersected regions of extreme allele frequency fluctuation between pairs of islands and stringently filtered nearby genes based on the impact of putative coding variation. Among genome-wide 10-kb windows, we found that Lagoon and West Snake Cays shared 238 windows, Lagoon Cay and Cayos Cochinos shared 285 windows, and West Snake Cay and Cayos Cochinos shared 259 windows (Figure 4E). For all between-island comparisons, the degree of overlap in genomic windows is significantly higher than expected based on randomly permutated datasets (Figure 4F), consistent with the hypothesis that convergent molecular evolution driven by selection may underlie phenotypic shifts shared among island populations. In the comparison between Lagoon and West Snake Cays, we found several mouse phenotypes associated with immunity-related inflammation, which may be linked to immune system adaptation, and with circulating thyroid hormone (thyroxine) levels (Figure 6 and Supplementary Table 8). Altered hormone levels are *a priori* expected to play a role in body size, and reduced thyroxine, specifically, is known to reduce bone growth and leads to shorter long bones and decreased body weight in rats (Choi, Ryu, Roh, & Bae, 2018). Thyroxine also positively regulates growth hormone (GH), and reductions of thyroxine can depress GH secretion, thereby depressing growth (Amit et al., 1991; Root, Shulman, Root, & Diamond, 1986). Our comparison between Lagoon Cay and Cayos Cochinos yielded multiple enriched human phenotypes related to spinal cord abnormalities and a set of mouse phenotypes all related to monocyte morphology and abundance, providing another instance of enriched immunity-related phenotypes (Figure 6 and Supplementary Table 9). We also found three mouse phenotypes related to eye structure and development and two enriched phenotypes linked to mesoderm, including somite, development (Figure 6 and Supplementary Table 9). The latter two phenotypes are broad phenotypic categories, which complicates linking them directly to boa island phenotypes. Finally, when comparing West Snake Cay and Cayos Cochinos we found enriched mouse phenotypes that clustered into two broad phenotypic classifications. A strong signal consisting of eight phenotypes related to brain development and morphology was evident (Figure 6 and Supplementary Table 10), although it is unclear how exactly these phenotypes may be relevant to island populations. We also recovered an enriched mouse phenotype related to abnormal circulating LDL cholesterol levels (Figure 6 and Supplementary Table 10), which has a much more logical and concrete connection to the body composition observe on islands, where individuals appear to deposit fat more quickly than on the mainland. Overall, pairwise comparisons between islands produced several statistically-enriched mouse and human phenotypes with logical connections to known island boa phenotypes, though enrichment analyses lose important individual-gene context that is important in understanding the links between genotype and phenotype.

To understand how specific genes identified in our pairwise comparisons may play a role in the evolution of island traits, we assessed the mouse phenotypes related to each gene. Across pairwise comparisons, several genes stood out for their connection to mouse phenotypes that show strong parallels with island boa traits. Mutations in many of these genes lead to decreased body size or weight (SPTB, TG, GBA, SYTL4, LRP6, MYO10, CSPG4, ACE, PFAS, DTNNA1, and LIPA) and appear to impact these phenotypes through one or more processes, such as cholesterol levels and body fat (e.g., LIPA, CSPG4, LRP6, SPTB, and GBA), insulin signaling (e.g., EOGT and SYTL4), and/or skeletal growth (e.g., TG and LRP6). Many of these same genes, and others, are also linked to reproduction (ACE, PCYT1B, NOBOX, and TYK2), craniofacial morphology (LIMA1, LRP6, and PFAS), and pigmentation (MYO10 and CTNNA1). Overall, while these genes were originally detected due to their proximity to genomic windows with shared allelic fluctuation across pairs of islands, a small proportion of associated coding variants displayed parallel allele frequency fluctuations across these same islands. One variant within the gene TG showed a relatively high allele frequency fluctuation in both Lagoon (0.5) and West Snake (0.75) Cays while another variant in this same gene had a high allele frequency fluctuation only in Cayos Cochinos (0.875). A variant in GBA showed parallel allelic fluctuations in Lagoon (0.75) and West Snake (0.5) Cays while a variant in LRP6 shows parallel allelic fluctuations in Lagoon Cay (0.8) and Cayos Cochinos (0.875). Three variants in CSPG4 show high allele frequency fluctuation on Cayos Cochinos (0.625 or greater), while two of these also have a high allele frequency fluctuation in the Lagoon Cay population (0.75 or greater). Finally, the gene ACE contains three coding-region variants that fluctuate in allele frequency by at least 0.5 in both West Snake Cay and Cayos Cochinos. In all cases, the allelic fluctuation at these variant sites is much lower in the comparison between the two mainland populations and all five of these genes were within 2 Mb (three were within 200 kb) of a RAD locus that is under selection. In conclusion, regions of shared allele frequency fluctuation contain genes with associated phenotypes that are easily linked to the traits we find across island boa populations, and a subset of these coding regions contain variants with patterns of allele frequency fluctuation that suggest convergent evolution via natural selection.

Candidate genes with links to island phenotypes identified in regions of shared allelic fluctuation across three island systems

Following our pairwise island comparisons, we interrogated regions of the genome that show extreme allele frequency fluctuations (0.9 or greater) across all three island populations, with 20 10-kb regions showing parallel allele frequency fluctuations of this magnitude (Figure 4E). As with the pairwise island comparisons, the number of overlapping windows exceeds our neutral expectation deduced from permutation analyses (Figure 4F). These regions contained 36 genes, which were not enriched for any GO biological processes or KEGG pathways but did show enrichment for several human and mouse phenotypes. Three human phenotypes were enriched, including peripheral primitive neuroectodermal neoplasm (HP:0030067), missing ribs (HP:0000921), and iris hypopigmentation (HP:0007730; Supplementary Table 11). The last enriched human phenotype is particularly interesting given that eye color is known to vary across islands. Only a single mouse phenotype was enriched, abnormal litter size (MP:0001933; Supplementary Table 11), which corresponds well with the reduced fecundity that has been observe on islands (Boback, 2005). When we further evaluated the full set of mouse phenotypes associated with these 36 genes, we observed that multiple genes are associated with reproduction (FAAP20, MAK, and SYCP2L), body fat and metabolism (BHMT, ELOVL2, and MYLIP), skeletal development and body growth (ARSB, ATXN1, DNAJC10, EEF1AKMT1, GCM2, IFT88, PRDM5, PTPRS, SCIN, and SKI), and pigmentation (PRKCZ and SLC5A8). The striking correspondence between this gene set and many of the phenotypes that are known to differ drastically between island and mainland populations (Boback, 2005, 2006; Boback & Carpenter, 2007; Reed et al., 2007) provides several candidate genes that putatively underlie island phenotypic shifts.

To restrict the three-island gene set to genes that have the greatest chance of directly driving island phenotypes, we produced a stringently filtered set of genes and coding-region variants. This filtering resulted in four total boa genes, and we were able to confidently assign human homologs for three of these genes: PTPRS, MYLIP, and DMGDH. A thorough review of literature indicates that these genes may play an important role in island adaptations.

Protein tyrosine phosphatase receptor type S (PTPRS) and other members of the protein tyrosine phosphatase family modulate signal transduction through the de-phosphorylation of tyrosine residues, thus influencing numerous cellular processes essential for proper embryonic development and growth (Hale, ter Steege, & den Hertog, 2017). The knockout of PTPRS in mice causes a significant reduction in circulating levels of insulin-like growth factor 1 (IGF-1) and growth hormone (GH) due to disruption of GH-secreting somatotroph cell differentiation and improper development of the pituitary gland (Batt, Asa, Fladd, & Rotin, 2002; Elchebly et al., 1999). GH secretion stimulates the production and release of IGF-1, which then directs a variety of cellular processes related to cellular proliferation and organismal growth. IGF-1 can therefore play a major role in determining body size. For example, deficiency of circulating IGF-1 correlates with reduced body size in mammals (Baker, Liu, Robertson, & Efstratiadis, 1993; K. A. Woods, Camacho-Hübner, Barter, Clark, & Savage, 1997; Katie A. Woods, Camacho-Hübner, Savage, & Clark, 1996), and a single haplotype of IGF-1 has been identified as a major determinant of reduced body size in small dog breeds (Sutter et al., 2007). Accordingly, mice in which PTPRS is knocked out exhibit reduced body size and weight, general retardation of growth, and decreased litter size (Elchebly et al., 1999). Additionally, the inactivation of PTPRS in mice results in alterations to BMP and WNT signaling pathways, resulting in improper maxillary and mandibular development and changes to craniofacial morphology (Stewart, Uetani, Hendriks, Tremblay, & Bouchard, 2013). It is therefore possible that changes to the PTPRS gene may explain a large degree of the hallmark phenotypes associated with these dwarf snakes, likely due to a reduction or change in the regulation of GH/IGF-1 and other major growth and development pathways. PTPRS was inferred to be the most likely homolog for three boa successive gene models, which likely represent three isoforms of the same gene. Across these gene models, 12 non-synonymous variants were identified, though only one high impact InDel (protein residue 220) was retained in the heavily-filtered dataset. This high-impact InDel has an allele frequency fluctuation of 1.0 in the West Snake Cay population but does not vary between pairwise comparisons between the other populations (Figure 7A). A second moderate-impact variant (protein residue 434) with a non-deleterious PROVEAN score displays a Lagoon-specific extreme allele frequency fluctuation of 0.9. Moreover, a selected RAD locus lies approximately 200 kb away from this gene. Collectively, all evidence points towards this gene being a high-quality candidate gene underlying island-specific phenotypes in at least the two Belize island populations.

Myosin Regulatory Light Chain Interacting Protein (MYLIP), also known as E3 Ubiquitin ligase-inducible degrader of the low-density lipoprotein receptor, is an important regulator of lipoprotein metabolism. Human GWAS studies have identified MYLIP in screens for low density lipoprotein cholesterol and total cholesterol (Global Lipids Genetics Consortium et al., 2013; Surakka et al., 2015; Weissglas-Volkov et al., 2011). Similarly, mice with null mutations in the gene encoding MYLIP show a number of phenotypes, including those linked to cholesterol levels, lipid regulation, and body fat mass, as well as others linked to behavior and hyperactivity (Smith et al., 2018). The rarity and seasonality of prey, together with the less massive, more slender phenotypes of island boas suggest that substantial difference in metabolism and fat storage may have evolved between island and mainland populations. MYLIP contains a single putatively deleterious, non-synonymous coding variant (protein residue 360) and shows relatively high allele frequency fluctuation in the West Snake Cay and Cayos Cochinos populations (Figure 7B). No selected RAD loci reside on the genomic scaffold that contains MYLIP.

A third gene with shared non-synonymous variants and highly differentiated island alleles is also functionally linked to endocrine signaling and GH, growth, and fat metabolism. Dimethylglycine

Dehydrogenase (DMGDH) is an enzyme involved in the catabolism of choline, leading to the breakdown of dimethylglycine (DMG) to glycine. A loss of function mutation in this gene in mice leads to decreased circulating thyroxine (Smith et al., 2018), resulting in reduced bone growth and body weight (Choi et al., 2018), and to depressed GH secretion leading to suppressed growth (Amit et al., 1991; Root et al., 1986). In addition to impacts on growth, human genome-wide association studies have identified DMGDH as being significantly associated with increased plasma insulin, increased insulin resistance, and an increased incidental risk of diabetes and cardiovascular diseases (Adeva-Andany et al., 2018; Magnusson et al., 2015), linking DMGDH to its impacts on glucose and fat metabolism. Three non-synonymous coding variants were observed in DMGDH, though only one (protein residue 271) has a deleterious PROVEAN score. This variant has high allelic fluctuation restricted to West Snake Cay and very low fluctuation in the other pairwise comparisons (Figure 7C), despite the fact that it lies near a window of parallel extreme allelic variation across all three island populations. A RAD locus under selection was found on the same scaffold at approximately 600 kb away. Overall, though coding variation only fluctuates greatly in one island population, this gene is still a viable candidate given its role in modulating growth.

A striking pattern that emerged in the three candidate genes discussed above is the lack of parallel protein coding variant allele fluctuation across all three island populations. This is despite the close proximity of these genes to genomic windows that did show parallel allele frequency changes across these populations. Further data is needed to determine whether these patterns are a product of the low sample size in our WGS dataset, where stochasticity in sequencing coverage or stringent variant filtering could explain the lack of congruence between loci. However, as we were interested in identifying whether any genes do show evidence of high allele frequency fluctuations on islands, we revisited our less stringently filtered gene set and manually scanned for variants where the three island populations showed a relatively high allele frequency fluctuation (0.5 or greater), while the two mainland island populations showed a low allele frequency fluctuation (0.1 or lower). One gene, arylsulfatase B (ARSB; which is adjacent to DMGDH), met these criteria and deserves further characterization as a putative candidate gene underlying island-specific phenotypes (Figure 7C). ARSB is associated with abnormal caudal vertebrae morphology, head and nose morphology, fat/triglyceride levels, and decreased birth and adult body size in mouse (Smith et al., 2018). Moreover, it is the causative gene for the human disorder mucopolysaccharidosis type VI, which is a lysosomal storage disorder resulting from a deficiency of arylsulfatase B. The disease is characterized by several phenotypes, including stiff joints, cardiac abnormalities, swollen liver and spleen (hepatosplenomegaly), and bone development issues (dysostosis multiplex; Azevedo et al., 2004). Remarkably, the disease is also associated with short stature and with facial dysmorphism (Azevedo et al., 2004), which have obvious associations with key phenotypic shifts that occur across island boa populations. Similar phenotypes have also been noted in dogs and are caused by mutations to the orthologous gene (Wang et al., 2018). Reduced expression of ARSB in human prostate cancer tissues has been linked to downstream increases in Wnt/B-catenin signaling (Bhattacharyya et al., 2017), a pathway that is important for proper development (Logan & Nusse, 2004). Interestingly, ARSB directly reduces the expression of Dickkopf Wnt signaling inhibitor DKK3, which is a negative regulator of Wnt signaling, through a possible interaction with LDL-receptor related protein (LRP) 5/6 (Kawano et al., 2006; Ueno et al., 2013; Veeck & Dahl, 2012), one of which was discussed above due to parallel high allele frequency shifts in Lagoon Cay and Cayos Cochinos.

Links between key island phenotypes and genetic variation across diverse lineages

Many phenotypic traits vary greatly across nature, between both large groupings like species but also potentially in closely-related populations. Perhaps the most noticeable and commonly-cited phenotype that can vary widely is body size, which has been most extensively studied in a controlled fashion in humans and other model and domestic organisms (Kemper, Visscher, & Goddard, 2012). From humans alone, it has been estimated that ~50 genes have some effect on size, though only a handful have been found to influence height consistently (Gudbjartsson et al., 2008; Lettre et al., 2008; Visscher, 2008; Weedon et al., 2008; Yang et al., 2010) . In dogs, IGF1 is known to play a major role in body size (Sutter et al., 2007), as it can in humans (Becker et al., 2013), though other more alleles in different genes have been identified that are associated with canine body size (Boyko et al., 2010). Work on livestock domesticates has also produced candidate alleles associated with body size (Bouwman et al., 2018; Chung et al., 2018; Fink et al., 2017; Kemper et al., 2012) and a deletion in a single gene has been linked to dwarfism in rabbit breeds (Carneiro et al., 2017). Despite the ubiquity of island size dimorphism across diverse taxa worldwide, genetic studies of this phenomenon has also been focused on populations of model organisms, like humans from Sardinia (Zoledziewska et al., 2015) and mice from Gough island (Gray et al., 2015; Parmenter et al., 2016). Though increasing numbers of genes associated with subtle body size differences, a common thread of many studies of body size is that many genes of large effect appear to have some regulatory effect on the insulin-like growth factor I (IGF-1)/growth hormone (GH) pathway. We, too, implicate two genes, PTPRS and DMGDH, that also interact with IGF-1/GH and putatively underlie body size differences between island and mainland boa populations, though variants in these genes only appear to fluctuate widely in the West Snake Cay population. The results of our work and previous studies on body size, which often find adaptation in genes related to IGF-1/GH, do, however, implicate molecular pathways as the level at which molecular convergence typically takes place and leads to similar phenotypic changes. Similar results have been documented for complex phenotypes in other organisms (Bergey et al., 2018; Gallant et al., 2014; Larter, Dunbar-Wallis, Berardi, Smith, & Purugganan, 2018; Pinto et al., 2014; Soy et al., 2016).

Island population also vary in the degree of snout attenuation, head width, and eye size, which is likely an adaptation that aids in visual hunting of divergent island food sources (Lillywhite & Henderson, 2002) and may also aid in the arboreal ecology of these populations (Lillywhite & Henderson, 2002; Shine, 1983). Wnt signaling, has been implicated in craniofacial development (Brugmann et al., 2010; Samantha A. Brugmann et al., 2007; Kurosaka, Iulianella, Williams, & Trainor, 2014; Schmidt & Patel, 2005), and two candidates genes identified in our analysis, ARSB and PTPRS, have links to the Wnt pathway. Wnt signaling may therefore be an important hub for molecular convergence to mediate adaptive phenotypic convergence. Indeed, the pathway appears to underlie adaptive craniofacial variation in a classic case of adaptive radiation: the extremely rapid evolution of thousands of species of cichlids in Lake Malawi (Parsons, Taylor, Powder, & Albertson, 2014). Natural phenotypic variation in adaptive traits, especially when replicated in the manner we observe in some island taxa, present a powerful opportunity to better understand how both convergent and divergent genetic changes propagate through molecular pathways to alter complex phenotypes.

Island systems are ecologically constrained due to limited land area and habitat and their isolation, leading to significant resource limitations. Island boas subsist on significantly smaller prey items, including migrating passerine birds that can lead to long periods of fasting (Boback, 2005). Such limitations impose significant energetic restrictions on these snakes, which appears to manifest in divergent body fat deposition and reduced litter sizes in island population (Boback, 2005). Accordingly, we find enriched phenotypes related to body fat and cholesterol (e.g., abnormal liver cholesterol and arterial calcification) and to reproduction (e.g., abnormal ovarian folliculogenesis) that appear to be mediating these phenotypic shifts, implicating genes like MYLIP as important for energetic adaptation across islands.

Coloration varies significantly between island and mainland populations, with island snakes having either lighter or darker coloration, depending on the island they come from (Boback & Siefferman, 2010; Porras, 1999). Eye color is also known to differ across islands and interestingly, we find a gene with links to retinal pigmentation (SLC5A8; Babu et al., 2011). Our analyses also implicate PRKCZ as a potentially important mediator of pigmentation differences between islands. The protein kinase C pathways appears to regulate melanogenesis by activating tyrosinase, the enzyme that catalyzes melanin synthesis (D'Mello, Finlay, Baguley, & Askarian-Amiri, 2016; Gordon & Gilchrest, 1989). This pathway operates separately from other pathways regulating melanin, including the MC1R pathway that has been previously implicated in local color adaptation in mice (Hoekstra, Hirschmann, Bundey, Insel, & Crossland, 2006; Nachman, Hoekstra, & D'Agostino, 2003; Steiner, Weber, & Hoekstra, 2007) and lizards (Rosenblum, Hoekstra, & Nachman, 2007; Rosenblum, Römpler, Schöneberg, & Hoekstra, 2010). Therefore, separate pathways may be used idiosyncratically to produce convergent phenotypes in different taxa.

Conclusion

Despite the ubiquity of island size dimorphism across diverse taxa worldwide, little work has focused on the underlying genetic and molecular basis of this phenotypic shift. Similarly, many other island-specific morpho- and eco-types have been identified yet few studies have explored the molecular basis of these traits. Given the general paradigm that there is a preponderance of genetic "solutions" for a phenotype, we do not expect molecular convergence, but we note that numerous recent studies are challenging this assertion (Castoe et al., 2009; Hohenlohe et al., 2010; Jones et al., 2012). We find no clear evidence of broad convergent evolution based on any enrichment-based approaches, which instead highlights the surprising uniqueness of functional or regulatory classes of genes that differentiate single islands. We did observe some evidence that pairs of islands contained overlapping sets of genes with mouse or human phenotypes that closely parallel the observed phenotypes in island populations. Our results collectively suggest that despite remarkably similar phenotypes across island populations, parallel evolution largely driven by unique and island-specific evolutionary trajectories rather than dominated by convergent molecular evolution.

ACKNOWLEDGMENTS

Support was provided from startup funds from the University of Texas at Arlington to TAC, an NSF Doctoral Dissertation Improvement Grant award (NSF DEB-1501747) to DCC and TAC, funding from the Beta Phi chapter of the Phi Sigma society to DCC, and a Lewis and Clark Award from the American Philosophical Society to DCC.

FIGURES

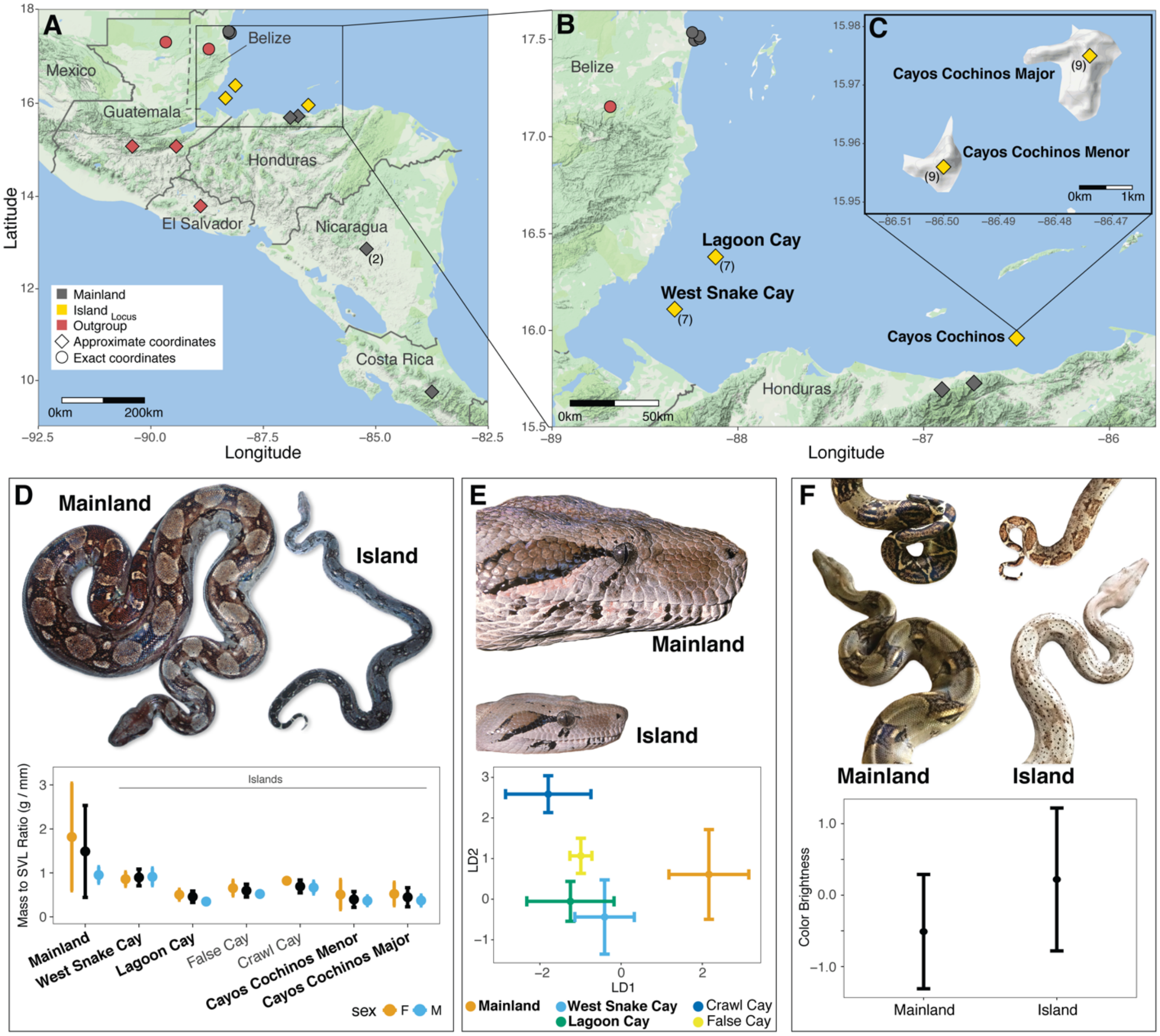


Figure 1. Summary of island boa study system. (**A-C**) Geographic representation of the location of sampling of populations (with sample sizes), including the island populations located on Lagoon and West Snake Cays, Belize and Cayos Cochinos, Honduras. (**D-F**) Overview of the phenotypic differences known between island and mainland boa populations, including body size (**D**), craniofacial morphology (**E**), and coloration (**F**). Phenotypic data are replotted based on data from previous studies (Boback, 2005, 2006; Boback & Carpenter, 2007; Reed et al., 2007) and collected from sampling of the Cayos Cochinos population conducted in 2016.

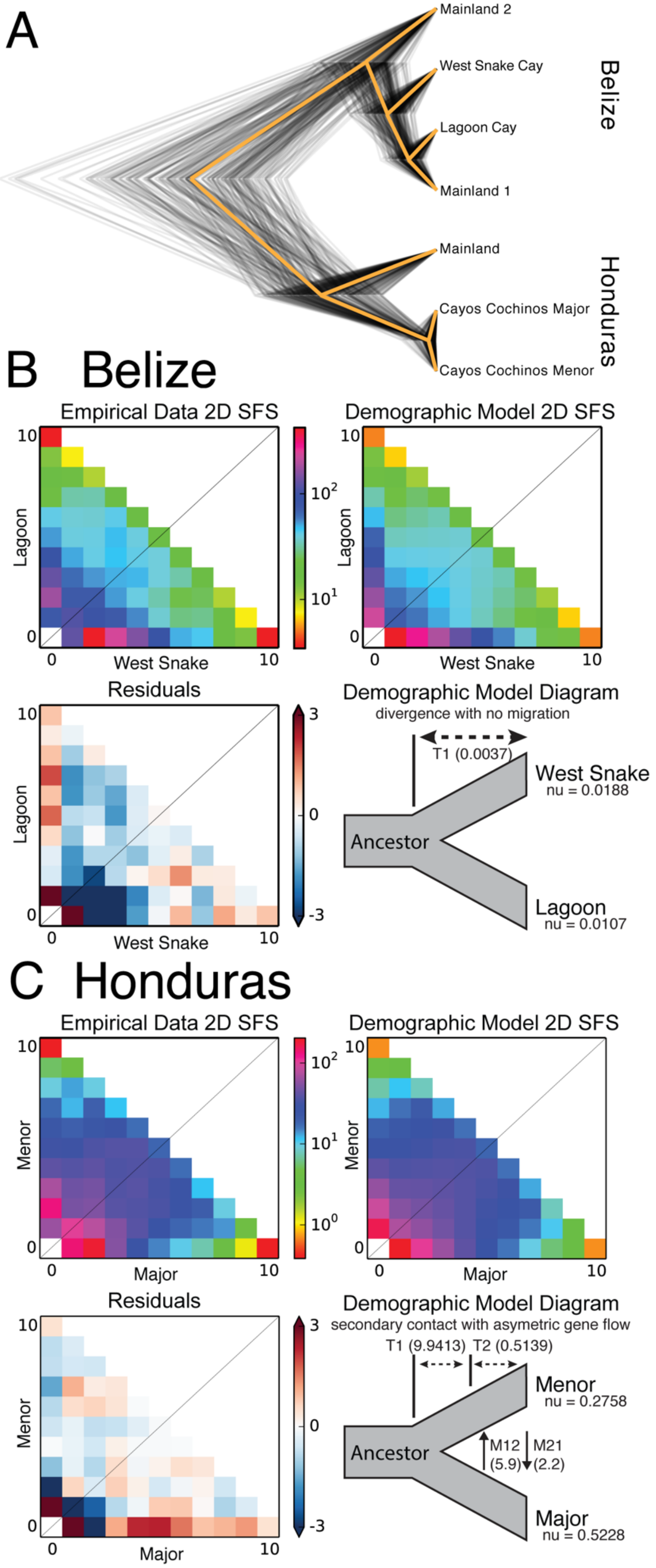
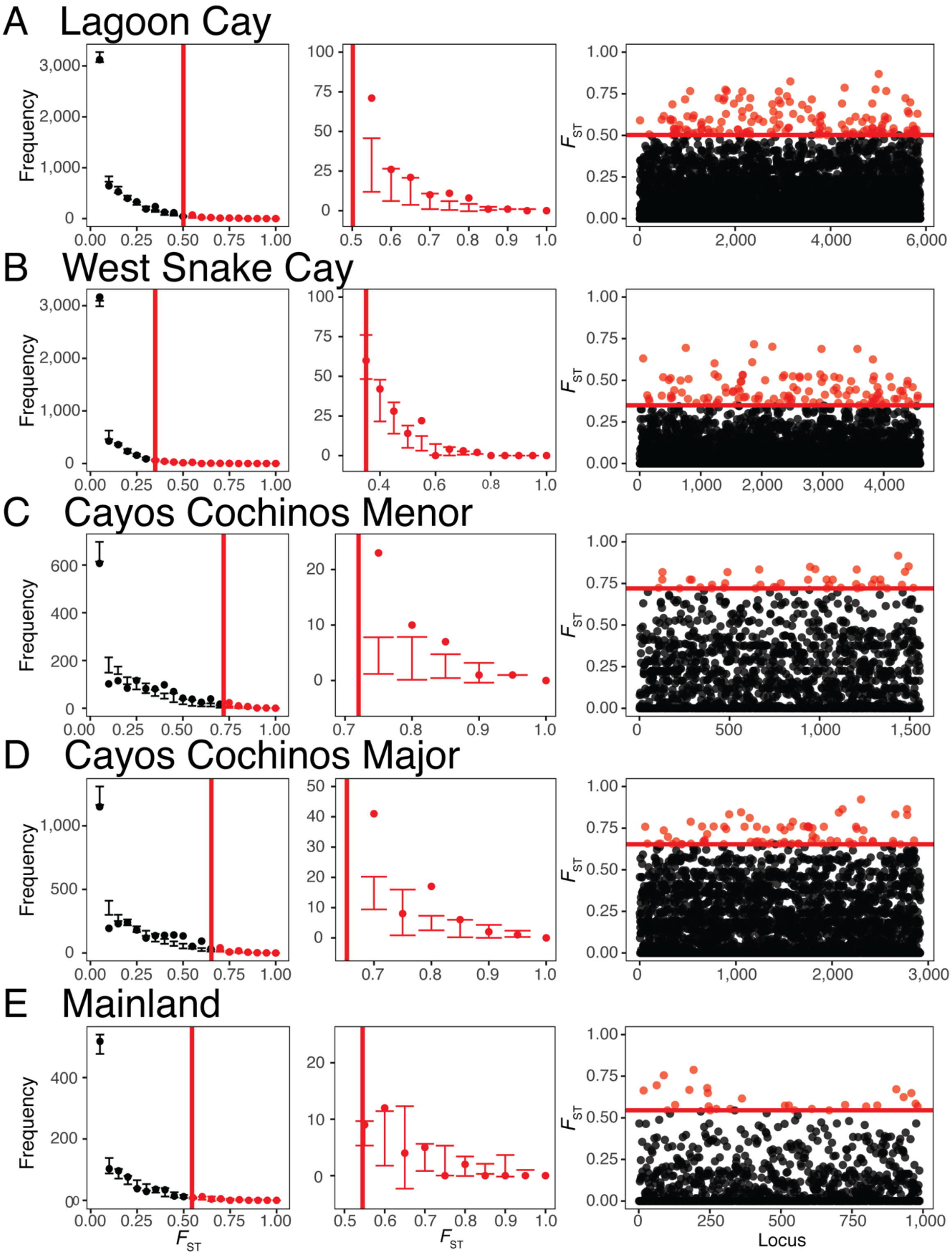

A
Mainland 2
West Snake Cay
Lagoon Cay
Mainland 1
Belize
Mainland
Cayos Cochinos Major
Cayos Cochinos Menor
Honduras
B Belize
Empirical Data 2D SFS
Demographic Model 2D SFS
Lagoon
West Snake
0
10
10^2
10^1
Residuals
3
0
-3
Demographic Model Diagram
divergence with no migration
T1 (0.0037)
West Snake
nu = 0.0188
Ancestor
Lagoon
nu = 0.0107
C Honduras
Empirical Data 2D SFS
Demographic Model 2D SFS
Menor
Major
10^2
10^1
10^0
Residuals
Demographic Model Diagram
secondary contact with asymetric gene flow
T1 (9.9413) T2 (0.5139)
Menor
nu = 0.2758
Ancestor
M12 (5.9)
M21 (2.2)
Major
nu = 0.5228

Figure 2. Demographic analysis of island population establishes three independent instances of the evolution of dwarfism on islands. (**A**) DensiTree showing posterior topologies estimated from SNAPP, with the consensus population phylogeny indicated in orange. (**B**) Results of δaδi 2D SFS analysis comparing plausible demographic relationships between Lagoon and West Snake Cays in Belize, which supports a model of divergence with no subsequent migration. (**C**) Results of δaδi 2D SFS analysis comparing plausible demographic relationships between the two Cayos Cochinos populations, which results in a best-supported model of ongoing gene flow between islands.



A Lagoon Cay
B West Snake Cay
C Cayos Cochinos Menor
D Cayos Cochinos Major
E Mainland
Frequency
3,000
2,000
1,000
0
0.00
0.25
0.50
0.75
1.00
100
75
50
25
0.5
0.6
0.7
0.8
0.9
1.0
F_{ST}
0
2,000
4,000
6,000
1,000
3,000
600
400
200
20
10
500
1,500
50
40
30
Locus

Figure 3. Evidence for genomic diversity stemming from natural selection versus neutral genetic drift in island populations. Panels present the distributions of F_{ST} values from pairwise comparisons between island and mainland population pairs (**A-D**) and between the two mainland populations (**E**). Left-most panel provides the full F_{ST} distributions while the middle panel focuses on the top 2.5% tail of the distribution (indicated in red). Whisker plots represent the 95% confidence interval that resulted from 10 GppFst PPS runs while points show the empirical frequency of F_{ST} bins. The right-most plot are Manhattan plots of empirical F_{ST} values from genome-wide RAD loci, with the top 2.5% of values (and the corresponding F_{ST} threshold) represented in red. Statistically significant excess frequencies were observed in the top 2.5% tail of F_{ST} values in comparisons between island and mainland population pairs (**A-D)**, while the same threshold did not yield excess frequencies in the comparison between mainland populations (**E**). These findings indicate that natural selection, on top of drift, had impacted allelic differentiation between island and mainland populations, but not between the two mainland populations.

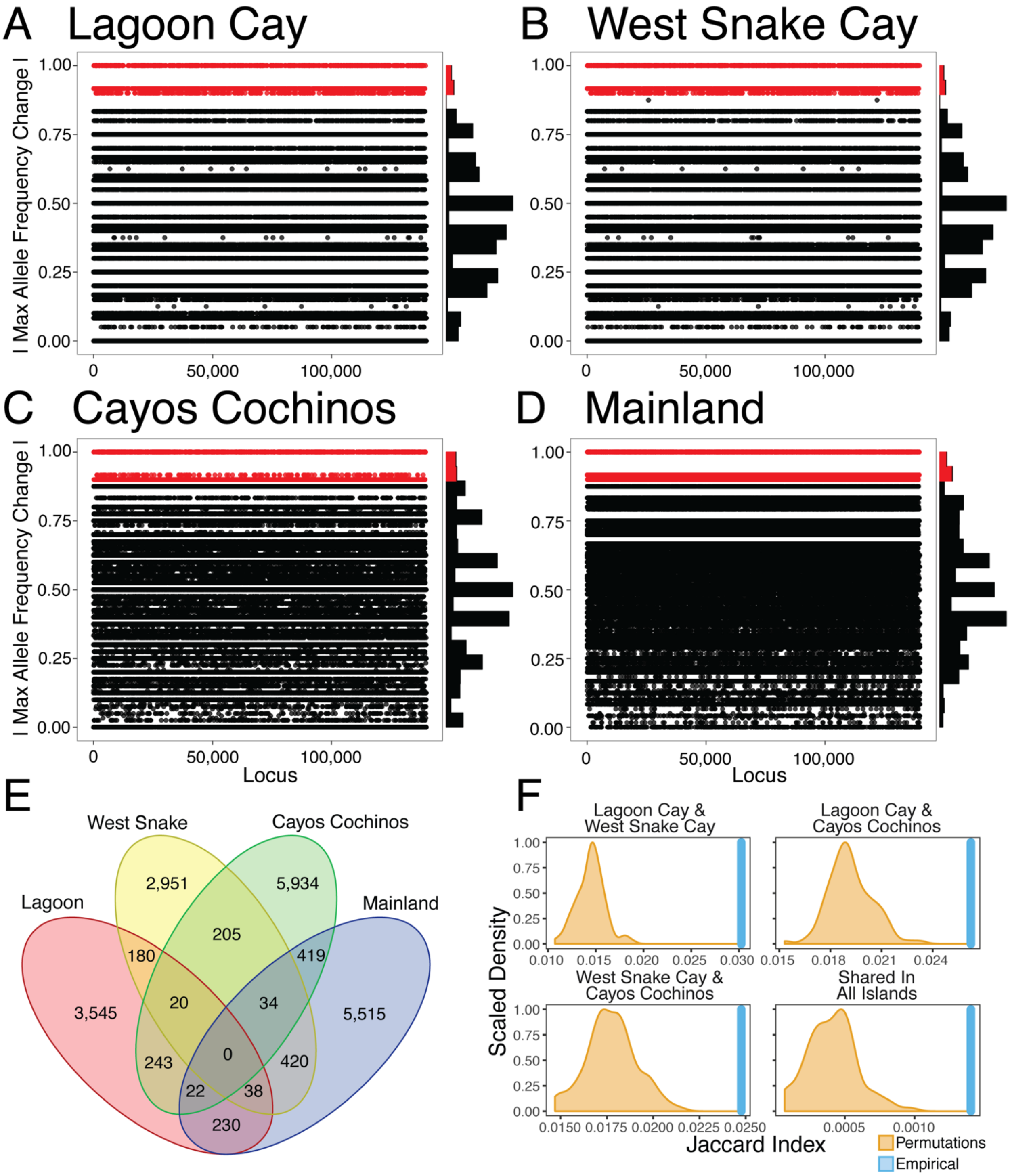


Figure 4. Extreme fluctuations in allele frequencies in island populations are shared between islands. Manhattan plots of the maximum allele frequency fluctuation in genome-wide 10-kb windows between island-mainland populations pairs – Lagoon Cay (**A**), West Snake Cay (**B**), and Cayos Cochinos (**C**) – and between the two mainland populations (**D**). Windows with an allele frequency fluctuation of 0.9

or greater are indicated in red, and the marginal distribution of allele frequency fluctuations is displayed as a histogram along the y-axis. Genomic windows are shared across island comparisons (**E**) and the degree of parallel allele frequency fluctuations is higher than expected based on 100 random data permutations (**F**).

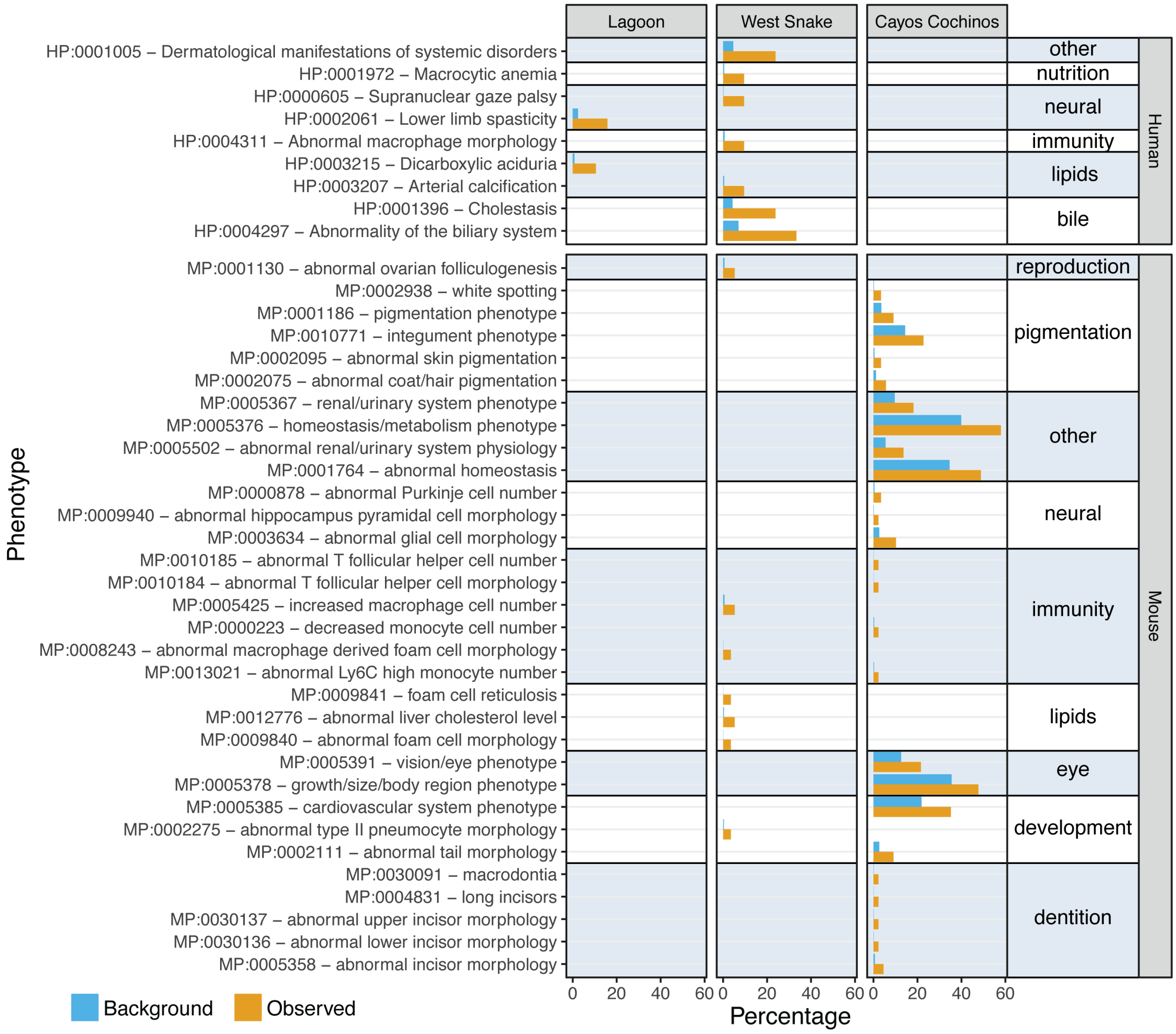


Figure 5. Genes with phenotypically-relevant coding variation within regions of individual island extreme allelic fluctuation are enriched for phenotypes connected to island boa traits. Paired bar plots showing the percentage of genes linked to mouse and human phenotypes in the entire protein-coding background versus the empirical data for phenotypes with a significant FDR p-value of 0.05 or lower. Panels are arranged into grids to distinguish mouse and human phenotypes and to partition island populations. Few phenotypes are enriched in multiple islands and several enriched phenotypes have logical connections to island boa phenotypes (see main text).

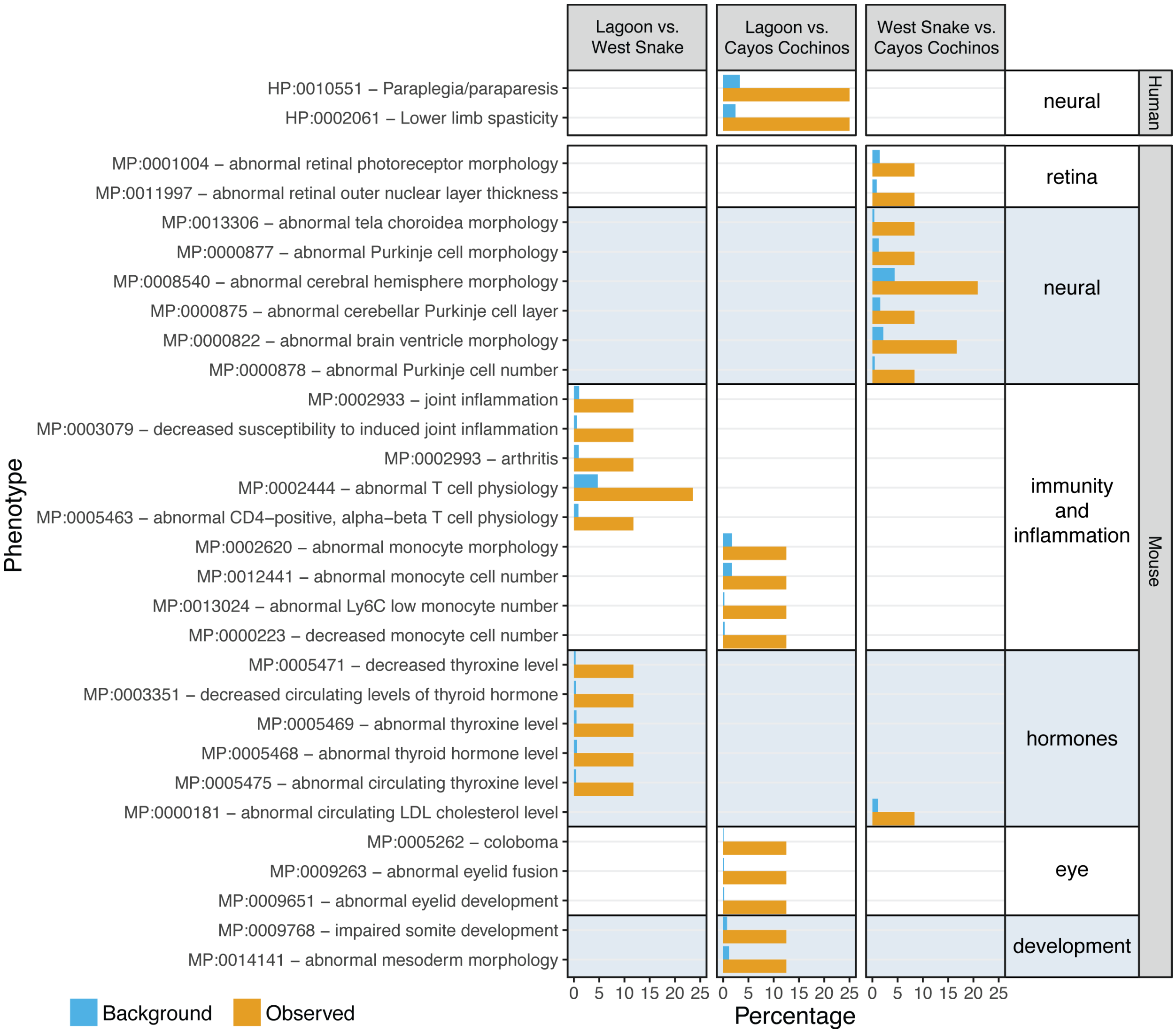


Figure 6. Clusters of related phenotypes with connections to island boa traits arise from genes with phenotypically-relevant coding variation shared between island population pairs. Paired bar plots showing the percentage of genes linked to mouse and human phenotypes in the entire protein-coding background versus the empirical data for phenotypes with a significant FDR p-value of 0.05 or lower. Panels are arranged into grids to distinguish mouse and human phenotypes and to partition each pairwise island-island comparison. Clusters of related phenotypes are observed within certain island comparisons, suggesting convergent genetic evolution may play a role in convergent island phenotypes.

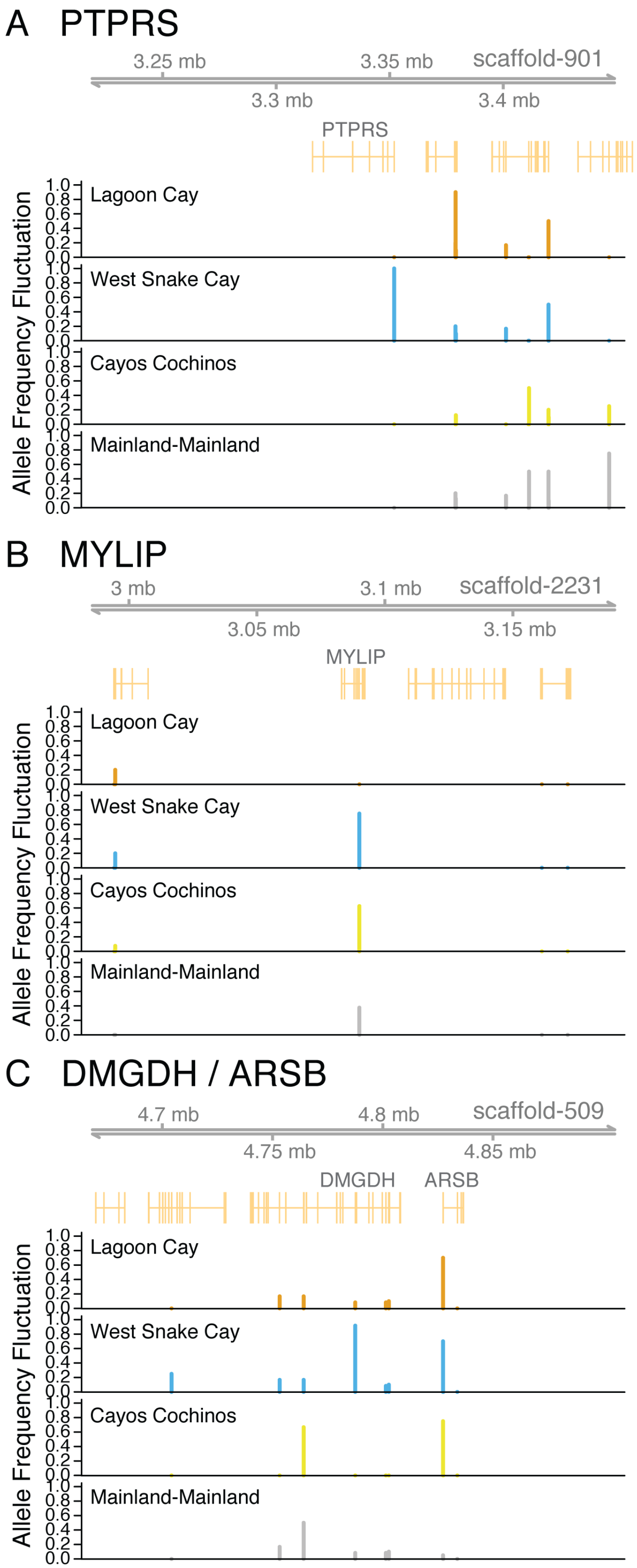

A PTPRS
3.25 mb
3.35 mb
scaffold-901
3.3 mb
3.4 mb
PTPRS
Allele Frequency Fluctuation
1.0
0.8
0.6
0.4
0.2
0.0
Lagoon Cay
West Snake Cay
Cayos Cochinos
Mainland-Mainland
B MYLIP
3 mb
3.1 mb
scaffold-2231
3.05 mb
3.15 mb
MYLIP
Allele Frequency Fluctuation
Lagoon Cay
West Snake Cay
Cayos Cochinos
Mainland-Mainland
C DMGDH / ARSB
4.7 mb
4.8 mb
scaffold-509
4.75 mb
4.85 mb
DMGDH
ARSB
Allele Frequency Fluctuation
Lagoon Cay
West Snake Cay
Cayos Cochinos
Mainland-Mainland

Figure 7. Genomic context of coding coding variation around candidate genes putatively underlying island traits in all three islands. Each panel represents the allele frequency fluctuations in coding variants within the 200-kb region surrounding (A) PTPRS, (B) MYLIP, and (C) DMGDH and ARSB. Respective gene models are indicated for each panel and each track represents a population. A variant each in PTPRS and DMGDH shows high allele frequency fluctuation in West Snake Cay while a variant in MYLIP shows high allele frequency fluctuation in West Snake Cay and Cayos Cochinos (with modest variation in the mainland). Only in ARSB is there apparent high allele frequency fluctuation across all island populations that is absent in the mainland-mainland comparison.

SUPPLEMENTARY MATERIAL

Supplementary Table 1. Details of the datasets produced in this study, including population assignments, locality information, and geographic coordinates for each sample. Results of population genetic model comparison using the two-dimensional allele frequency spectrum (2D-AFS) between Lagoon and West Snake Island populations. The best-fit model and parameters are in bold. Visual comparison of the 2D-AFS for the data and the best-fit model is provided in Figure 2.

Dataset	Sample	Clade	Population	Locality	Latitude	Longitude
RADseq	Boco011	Belize	Mainland 2	Cayo province, Belize	17.15428	-88.67733
RADseq	Boco012	Belize	Mainland 1	Belize City area, Belize province, Belize	17.50428	-88.19586
RADseq	Boco014	Belize	Lagoon Cay	Lagoon Cay	16.191679	-88.571827
RADseq	Boco015	Belize	West Snake Cay	West Snake Cay	16.191679	-88.571827
RADseq	Boco016	Belize	West Snake Cay	West Snake Cay	16.191679	-88.571827
RADseq	Boco017	Belize	Mainland 1	Belize City area, Belize province, Belize	17.49572	-88.22369
RADseq	Boco018	Belize	West Snake Cay	West Snake Cay	16.191679	-88.571827
RADseq	Boco019	Belize	West Snake Cay	West Snake Cay	16.191679	-88.571827
RADseq	Boco020	Belize	Lagoon Cay	Lagoon Cay	16.631967	-88.20894
RADseq	Boco021	Belize	Lagoon Cay	Lagoon Cay	16.631967	-88.20894
RADseq	Boco022	Belize	Lagoon Cay	Lagoon Cay	16.631967	-88.20894
RADseq	Boco023	Belize	Lagoon Cay	Lagoon Cay	16.631967	-88.20894
RADseq	Boco024	Belize	West Snake Cay	West Snake Cay	16.191679	-88.571827
RADseq	Boco025	Belize	West Snake Cay	West Snake Cay	16.191679	-88.571827
RADseq	Boco026	Belize	Lagoon Cay	Lagoon Cay	16.631967	-88.20894
RADseq	Boco027	Belize	Mainland 1	Belize City area, Belize province, Belize	17.535829	-88.235735
RADseq	Boco028	Belize	Mainland 1	Belize City area, Belize province, Belize	17.516032	-88.199182
RADseq	Boco030	Belize	West Snake Cay	West Snake Cay	16.191679	-88.571827
RADseq	Boco032	Belize	Lagoon Cay	Lagoon Cay	16.631967	-88.20894
RADseq	Boco044	Belize	Mainland 2	Zacapa province, Guatemala	N/A	N/A
RADseq	Boco047	Belize	Mainland 2	Bera Verapaz, Guatemala	N/A	N/A

RADseq	Boco051	Belize	Mainland 2	Petén province, Guatemala	17.3025	-89.63444
RADseq	Boco074	Honduras	Mainland	Pure bred descendent of lineage from Nicaragua	N/A	N/A
RADseq	Boco075	Honduras	Mainland	Pure bred descendent of lineage from Nicaragua	N/A	N/A
RADseq	Boco080	Honduras	Mainland	Pure bred descendent of lineage from El Salvador	N/A	N/A
RADseq	Boco090	Honduras	Mainland	Pure bred descendent of lineage from Costa Rica	N/A	N/A
RADseq	Boco106	Honduras	Mainland	La Ceiba area, Atlántida province, Honduras	15.727422	-86.729469
RADseq	Boco108	Honduras	Mainland	La Ceiba area, Atlántida province, Honduras	15.693739	-86.901671
RADseq	Boco109	Honduras	Cayos Cochinos Menor	Cayos Cochinos Menor	15.955966	-86.501206
RADseq	Boco113	Honduras	Cayos Cochinos Menor	Cayos Cochinos Menor	15.955966	-86.501206
RADseq	Boco115	Honduras	Cayos Cochinos Menor	Cayos Cochinos Menor	15.955966	-86.501206
RADseq	Boco118	Honduras	Cayos Cochinos Menor	Cayos Cochinos Menor	15.955966	-86.501206
RADseq	Boco123	Honduras	Cayos Cochinos Menor	Cayos Cochinos Menor	15.955966	-86.501206
RADseq	Boco126	Honduras	Cayos Cochinos Menor	Cayos Cochinos Menor	15.955966	-86.501206
RADseq	Boco128	Honduras	Cayos Cochinos Major	Cayos Cochinos Major	15.955966	-86.501206
RADseq	Boco130	Honduras	Cayos Cochinos Major	Cayos Cochinos Major	15.973197	-86.475547
RADseq	Boco132	Honduras	Cayos Cochinos Major	Cayos Cochinos Major	15.973197	-86.475547
RADseq	Boco137	Honduras	Cayos Cochinos Major	Cayos Cochinos Major	15.973197	-86.475547
RADseq	Boco138	Honduras	Cayos Cochinos Major	Cayos Cochinos Major	15.973197	-86.475547
RADseq	Boco146	Honduras	Cayos Cochinos Major	Cayos Cochinos Major	15.973197	-86.475547
RADseq	Boco150	Honduras	Cayos Cochinos Major	Cayos Cochinos Major	15.973197	-86.475547
RADseq	Boco151	Honduras	Cayos Cochinos Major	Cayos Cochinos Major	15.973197	-86.475547
RADseq	Boco153	Honduras	Cayos Cochinos Major	Cayos Cochinos Major	15.973197	-86.475547
WGS	Boco011	Belize	Mainland 2	Cayo province, Belize	17.15428	-88.67733
WGS	Boco012	Belize	Mainland	Belize City area, Belize province, Belize	17.50428	-88.19586
WGS	Boco017	Belize	Mainland	Belize City area, Belize province, Belize	17.49572	-88.22369
WGS	Boco019	Belize	West Snake Cay	West Snake Cay	16.191679	-88.571827
WGS	Boco022	Belize	Lagoon Cay	Lagoon Cay	16.631967	-88.20894
WGS	Boco025	Belize	West Snake Cay	West Snake Cay	16.191679	-88.571827
WGS	Boco026	Belize	Lagoon Cay	Lagoon Cay	16.631967	-88.20894

WGS	Boco027	Belize	Mainland 1	Belize City area, Belize province, Belize	17.535829	-88.235735
WGS	Boco028	Belize	Mainland 1	Belize City area, Belize province, Belize	17.516032	-88.199182
WGS	Boco051	Belize	Mainland 2	Petén province, Guatemala	17.3025	-89.63444
WGS	Boco074	Honduras	Mainland	Pure bred descendent of lineage from Nicaragua	N/A	N/A
WGS	Boco075	Honduras	Mainland	Pure bred descendent of lineage from Nicaragua	N/A	N/A
WGS	Boco090	Honduras	Mainland	Pure bred descendent of lineage from Costa Rica	N/A	N/A
WGS	Boco106	Honduras	Mainland	La Ceiba area, Atlántida province, Honduras	15.727422	-86.729469
WGS	Boco108	Honduras	Mainland	La Ceiba area, Atlántida province, Honduras	15.693739	-86.901671
WGS	Boco114	Honduras	Cayos Cochinos Menor	Cayos Cochinos Menor	15.955966	-86.501206
WGS	Boco125	Honduras	Cayos Cochinos Menor	Cayos Cochinos Menor	15.955966	-86.501206
WGS	Boco138	Honduras	Cayos Cochinos Major	Cayos Cochinos Major	15.973197	-86.475547
WGS	Boco147	Honduras	Cayos Cochinos Major	Cayos Cochinos Major	15.973197	-86.475547

Abbreviations are as follows: WGS, whole genome sequencing.

Supplementary Table 2. Results of population genetic model comparison using the two-dimensional allele frequency spectrum (2D-AFS) between Lagoon and West Snake Island populations. The best-fit model and parameters are in bold. Visual comparison of the 2D-AFS for the data and the best-fit model is provided in Figure 2.

Model	AIC	ΔAIC	RL	w_i	log-lik	params	theta	nu1	nu2	m12	m21	T1	T2
Divergence and asymmetrical secondary contact	1302.8	330.62	1	0.0	-645.4	6	877.48	0.2217	0.528	3.3882	2.9581	1.6147	0.3388
Divergence and symmetrical secondary contact	1367.86	395.68	0.66	0.0	-678.93	5	213.49	1.3346	2.06	0.5575	0.5575	9.9498	3.0775
Divergence with ancient asymmetrical migration	1767.62	795.44	0.00	0.0	-877.81	6	1119.52	0.0318	0.7737	3.2436	0.7082	0.0029	0.0103
Divergence with ancient symmetrical migration	3677.46	2705.28	0.00	0.0	-1833.73	5	168.67	5.1862	4.9005	16.8551	16.8551	3.3831	0.8368
Divergence with asymmetric migration	993.06	20.88	0.00	0.0	-491.53	5	1263.61	0.0113	0.0243	0.8695	0.6932	0.0039	-
Divergence with no migration	**972.18**	**0**	**0.00**	**1.0**	**-483.09**	**3**	**1289.55**	**0.0107**	**0.0188**	**-**	**-**	**0.0037**	**-**
Divergence with symmetric migration	1036.98	64.8	0.00	0.0	-514.49	4	1251.9	0.0323	0.0479	0.8478	0.8478	0.0094	-
No divergence model	6610.02	5637.84	0.00	0.0	-3302.01	0	1006.94	-	-	-	-	-	-

Abbreviations are as follows: AIC, Akaike information criterion; RL, relative likelihood; w_i, Akaike Weight; params, number of parameters in model; theta, $4N_{ref}\mu L$; nu1, effective population size of the Lagoon population; nu2, effective population size of the West Snake population; m12, migration rate from West Snake to Lagoon; m21, migration rate from Lagoon to West Snake; T1, scaled time between population split and the present or T2, the scaled time of secondary contact or isolation interval.

Supplementary Table 3. Results of population genetic model comparison using the two-dimensional allele frequency spectrum (2D-AFS) between Cayos Cochinos island populations. The best-fit model and parameters are in bold. Visual comparison of the 2D-AFS for the data and the best-fit model is provided in Figure 2C.

Model	AIC	ΔAIC	RL	w_i	log-lik	params	theta	nu1	nu2	m12	m21	T1	T2
Divergence and asymmetrical secondary contact	**552.32**	**0**	**1.000**	**0.943**	**-270.16**	6	**243.45**	**0.2758**	**0.5228**	**5.9056**	**2.2127**	**9.9413**	**0.5139**
Divergence and symmetrical secondary contact	559.36	7.04	0.030	0.028	-274.68	5	561.09	0.0815	0.0947	17.2321	8.4756	0.2673	-
Divergence with ancient asymmetrical migration	949.02	396.7	0.000	0.000	-468.51	6	656.79	0.1147	0.8616	15.8427	1.2036	0.4086	0.0003
Divergence with ancient symmetrical migration	755.26	202.94	0.000	0.000	-372.63	5	827.17	0.2048	0.2572	8.9725	0.1864	0	-
Divergence with asymmetric migration	688.12	135.8	0.000	0.000	-339.06	5	751.61	0.0251	0.0322	0.3502	19.8903	0.005	
Divergence with no migration	692.92	140.6	0.000	0.000	-343.46	**3**	744.57	0.01	0.0139	0.0019	**-**	**-**	**-**
Divergence with symmetric migration	559.26	6.94	0.031	0.029	-275.63	4	1022.33	0.0664	0.075	19.9883	0.0557	-	-
No divergence model	2443.06	1890.74	0.000	0.000	-1218.53	0	629.27	-	-	-	-	-	-

Abbreviations are as follows: AIC, Akaike information criterion; RL, relative likelihood; w_i, Akaike Weight; params, number of parameters in model; theta, $4N_{ref}\mu L$; nu1, effective population size of the Menor population; nu2, effective population size of the Major population; m12, migration rate from Major to Menor; m21, migration rate from Menor to MajorM12; T1, scaled time between population split and the present or T2, the scaled time of secondary contact or isolation interval.

Supplementary Table 4. Results of GppFst PPS analysis expectations of allelic differentiation based on simulations and corresponding findings based on the empirical datasets.

Population Comparison	Mean F_{ST}	Median F_{ST}	95% CI F_{ST}	Maximum F_{ST}	97.5% F_{ST} Threshold	Expected Loci (upper 2.5% tail)	Observed Loci (upper 2.5% tail	Binomial test P-value
Lagoon v. Mainland Belize	0.07	0.04	-0.20 – 0.35	1.0	0.50	71.8	149	6.08e-10
West Snake v. Mainland Belize	0.03	0.005	-0.18 – 0.26	0.76	0.35	150.3	175	0.025
Menor v. Mainland Honduras	0.12	0.08	-0.23 – 0.48	0.92	0.75	13.5	42	3.68e-10
Major v. Mainland Honduras	0.12	0.08	-0.24 – 0.48	1.0	0.65	34.7	75	1.60e-9
Mainland Belize v. Mainland Honduras	0.09	0.04	-0.25 – 0.43	1.0	0.55	20.8	27	0.42

Abbreviations are as follows: CI, confidence interval

Supplementary Table 5. Statistically enriched phenotypes from genes with regions of extreme (≥0.9) allele frequency fluctuation in the Lagoon Cay population following stringent filtering.

Phenotype Source	Phenotype Name	Phenotype ID	Percent genes with term observed	Percent genes with term in background	Fisher's Exact Test P-value	FDR P-value
Human	Lower limb spasticity	HP:0002061	15.79% (3/19)	2.41% (88/3644)	0.011	0.088
Human	Dicarboxylic aciduria	HP:0003215	10.53% (2/19)	0.77% (28/3644)	0.01	0.088

Abbreviations are as follows: FDR, False Discovery Rate

Supplementary Table 6. Statistically enriched phenotypes from genes with regions of extreme (≥0.9) allele frequency fluctuation in the West Snake Cay population following stringent filtering.

Phenotype Source	Phenotype Name	Phenotype ID	Percent genes with term observed	Percent genes with term in background	Fisher's Exact Test P-value	FDR P-value
Mouse	abnormal ovarian folliculogenesis	MP:0001130	5.26% (3/57)	0.64% (61/9598)	0.006	0.072
Mouse	abnormal type II pneumocyte morphology	MP:0002275	3.51% (2/57)	0.35% (34/9598)	0.019	0.051
Mouse	increased macrophage cell number	MP:0005425	5.26% (3/57)	0.61% (59/9598)	0.006	0.072
Mouse	abnormal macrophage derived foam cell morphology	MP:0008243	3.51% (2/57)	0.05% (5/9598)	0.0007056	0.016
Mouse	abnormal foam cell morphology	MP:0009840	3.51% (2/57)	0.06% (6/9598)	0.0009372	0.098
Mouse	foam cell reticulosis	MP:0009841	3.51% (2/57)	0.05% (5/9598)	0.0007056	0.038
Mouse	abnormal liver cholesterol level	MP:0012776	5.26% (3/57)	0.32% (31/9598)	0.001	0.035
Human	Supranuclear gaze palsy	HP:0000605	9.52% (2/21)	0.14% (5/3642)	0.0006462	0.042
Human	Dermatological manifestations of systemic disorders	HP:0001005	23.81% (5/21)	4.61% (168/3642)	0.002	0.043
Human	Cholestasis	HP:0001396	23.81% (5/21)	4.28% (156/3642)	0.002	0.043
Human	Macrocytic anemia	HP:0001972	9.52% (2/21)	0.47% (17/3642)	0.005	0.078
Human	Arterial calcification	HP:0003207	9.52% (2/21)	0.49% (18/3642)	0.006	0.078
Human	Abnormality of the biliary system	HP:0004297	33.33% (7/21)	7.0% (255/3642)	0.00043	0.036
Human	Abnormal macrophage morphology	HP:0004311	9.52% (2/21)	0.66% (24/3642)	0.009	0.098

Abbreviations are as follows: FDR, False Discovery Rate

Supplementary Table 7. Statistically enriched phenotypes from genes with regions of extreme (≥0.9) allele frequency fluctuation in the Cayos Cochinos population following stringent filtering.

Phenotype Source	Phenotype Name	Phenotype ID	Percent genes with term observed	Percent genes with term in background	Fisher's Exact Test P-value	FDR P-value
Mouse	decreased monocyte cell number	MP:0000223	2.27% (2/88)	0.29% (28/9567)	0.03	0.1
Mouse	abnormal Purkinje cell number	MP:0000878	3.41% (3/88)	0.44% (42/9567)	0.008	0.01
Mouse	pigmentation phenotype	MP:0001186	9.09% (8/88)	3.57% (342/9567)	0.014	0.065
Mouse	abnormal homeostasis	MP:0001764	48.86% (43/88)	34.6% (3310/9567)	0.004	0.063
Mouse	abnormal coat/hair pigmentation	MP:0002075	5.68% (5/88)	1.11% (106/9567)	0.003	0.063
Mouse	abnormal skin pigmentation	MP:0002095	3.41% (3/88)	0.41% (39/9567)	0.006	0.076
Mouse	abnormal tail morphology	MP:0002111	9.09% (8/88)	2.64% (253/9567)	0.003	0.063
Mouse	white spotting	MP:0002938	3.41% (3/88)	0.21% (20/9567)	0.001	0.089
Mouse	abnormal glial cell morphology	MP:0003634	10.23% (9/88)	2.64% (253/9567)	0.0006113	0.089
Mouse	long incisors	MP:0004831	2.27% (2/88)	0.05% (5/9567)	0.002	0.024
Mouse	abnormal incisor morphology	MP:0005358	4.55% (4/88)	0.63% (60/9567)	0.003	0.024
Mouse	renal/urinary system phenotype	MP:0005367	18.18% (16/88)	9.66% (924/9567)	0.01	0.065
Mouse	homeostasis/metabolism phenotype	MP:0005376	57.95% (51/88)	39.91% (3818/9567)	0.0005005	0.014
Mouse	growth/size/body region phenotype	MP:0005378	47.73% (42/88)	35.53% (3399/9567)	0.013	0.065
Mouse	cardiovascular system phenotype	MP:0005385	35.23% (31/88)	21.86% (2091/9567)	0.003	0.042
Mouse	vision/eye phenotype	MP:0005391	21.59% (19/88)	12.57% (1203/9567)	0.013	0.065
Mouse	abnormal renal/urinary system physiology	MP:0005502	13.64% (12/88)	5.48% (524/9567)	0.003	0.063
Mouse	abnormal hippocampus pyramidal cell morphology	MP:0009940	2.27% (2/88)	0.14% (13/9567)	0.008	0.01
Mouse	abnormal T follicular helper cell morphology	MP:0010184	2.27% (2/88)	0.06% (6/9567)	0.002	0.006
Mouse	abnormal T follicular helper cell number	MP:0010185	2.27% (2/88)	0.06% (6/9567)	0.002	0.024

Mouse	integument phenotype	MP:0010771	22.73% (20/88)	14.4% (1378/9567)	0.025	0.1
Mouse	abnormal Ly6C high monocyte number	MP:0013021	2.27% (2/88)	0.22% (21/9567)	0.018	0.068
Mouse	macrodontia	MP:0030091	2.27% (2/88)	0.08% (8/9567)	0.004	0.024
Mouse	abnormal lower incisor morphology	MP:0030136	2.27% (2/88)	0.11% (11/9567)	0.006	0.03
Mouse	abnormal upper incisor morphology	MP:0030137	2.27% (2/88)	0.09% (9/9567)	0.004	0.009

Abbreviations are as follows: FDR, False Discovery Rate

Supplementary Table 8. Statistically enriched phenotypes from genes with regions of extreme (≥0.9) allele frequency fluctuation in both the Lagoon Cay and West Snake Cay population following stringent filtering.

Phenotype Source	Phenotype Name	Phenotype ID	Percent genes with term observed	Percent genes with term in background	Fisher's Exact Test P-value	FDR P-value
Mouse	abnormal T cell physiology	MP:0002444	23.53% (4/17)	4.7% (453/9638)	0.007	0.028
Mouse	joint inflammation	MP:0002933	11.76% (2/17)	1.04% (100/9638)	0.014	0.077
Mouse	arthritis	MP:0002993	11.76% (2/17)	0.94% (91/9638)	0.011	0.035
Mouse	decreased susceptibility to induced joint inflammation	MP:0003079	11.76% (2/17)	0.53% (51/9638)	0.004	0.021
Mouse	decreased circulating levels of thyroid hormone	MP:0003351	11.76% (2/17)	0.37% (36/9638)	0.002	0.022
Mouse	decreased susceptibility to autoimmune disorder	MP:0005351	11.76% (2/17)	1.53% (147/9638)	0.028	0.098
Mouse	abnormal CD4-positive, alpha-beta T cell physiology	MP:0005463	11.76% (2/17)	0.89% (86/9638)	0.01	0.07
Mouse	abnormal thyroid hormone level	MP:0005468	11.76% (2/17)	0.58% (56/9638)	0.005	0.072
Mouse	abnormal thyroxine level	MP:0005469	11.76% (2/17)	0.5% (48/9638)	0.003	0.022
Mouse	decreased thyroxine level	MP:0005471	11.76% (2/17)	0.33% (32/9638)	0.002	0.021
Mouse	abnormal circulating thyroxine level	MP:0005475	11.76% (2/17)	0.43% (41/9638)	0.003	0.021

Abbreviations are as follows: FDR, False Discovery Rate

Supplementary Table 9. Statistically enriched phenotypes from genes with regions of extreme (≥0.9) allele frequency fluctuation in both the Lagoon Cay and Cayos Cochinos population following stringent filtering.

Phenotype Source	Phenotype Name	Phenotype ID	Percent genes with term observed	Percent genes with term in background	Fisher's Exact Test P-value	FDR P-value
Mouse	decreased monocyte cell number	MP:0000223	12.5% (2/16)	0.29% (28/9639)	0.001	0.001
Mouse	abnormal monocyte morphology	MP:0002620	12.5% (2/16)	1.72% (166/9639)	0.031	0.052
Mouse	coloboma	MP:0005262	12.5% (2/16)	0.09% (9/9639)	0.0001404	0.004
Mouse	abnormal eyelid fusion	MP:0009263	12.5% (2/16)	0.16% (15/9639)	0.0003451	0.005
Mouse	abnormal eyelid development	MP:0009651	12.5% (2/16)	0.18% (17/9639)	0.0004331	0.006
Mouse	impaired somite development	MP:0009768	12.5% (2/16)	0.77% (74/9639)	0.007	0.049
Mouse	abnormal monocyte cell number	MP:0012441	12.5% (2/16)	1.7% (164/9639)	0.03	0.03
Mouse	abnormal Ly6C low monocyte number	MP:0013024	12.5% (2/16)	0.23% (22/9639)	0.0006957	0.003
Mouse	abnormal mesoderm morphology	MP:0014141	12.5% (2/16)	1.14% (110/9639)	0.014	0.065
Human	Lower limb spasticity	HP:0002061	25.0% (2/8)	2.44% (89/3655)	0.016	0.081
Human	Paraplegia/paraparesis	HP:0010551	25.0% (2/8)	3.28% (120/3655)	0.027	0.081

Abbreviations are as follows: FDR, False Discovery Rate

Supplementary Table 10. Statistically enriched phenotypes from genes with regions of extreme (≥0.9) allele frequency fluctuation in both the West Snake Cay and Cayos Cochinos population following stringent filtering.

Phenotype Source	Phenotype Name	Phenotype ID	Percent genes with term observed	Percent genes with term in background	Fisher's Exact Test P-value	FDR P-value
Mouse	abnormal circulating LDL cholesterol level	MP:0000181	8.33% (2/24)	1.08% (104/9631)	0.028	0.091
Mouse	abnormal brain ventricle morphology	MP:0000822	16.67% (4/24)	2.15% (207/9631)	0.002	0.082
Mouse	abnormal cerebellar Purkinje cell layer	MP:0000875	8.33% (2/24)	1.53% (147/9631)	0.052	0.091
Mouse	abnormal Purkinje cell morphology	MP:0000877	8.33% (2/24)	1.23% (118/9631)	0.035	0.07
Mouse	abnormal Purkinje cell number	MP:0000878	8.33% (2/24)	0.45% (43/9631)	0.005	0.005
Mouse	abnormal retinal photoreceptor morphology	MP:0001004	8.33% (2/24)	1.45% (140/9631)	0.048	0.091
Mouse	abnormal cerebral hemisphere morphology	MP:0008540	20.83% (5/24)	4.38% (422/9631)	0.003	0.054
Mouse	abnormal retinal outer nuclear layer thickness	MP:0011997	8.33% (2/24)	0.85% (82/9631)	0.018	0.091
Mouse	abnormal tela choroidea morphology	MP:0013306	8.33% (2/24)	0.37% (36/9631)	0.004	0.054

Abbreviations are as follows: FDR, False Discovery Rate

Supplementary Table 11. Statistically enriched phenotypes from genes with regions of extreme (≥0.9) allele frequency fluctuation in the Lagoon Cay, West Snake Cay, and Cayos Cochinos population following less stringent filtering.

Phenotype Source	Phenotype Name	Phenotype ID	Percent genes with term observed	Percent genes with term in background	Fisher's Exact Test P-value	FDR P-value
Mouse	abnormal litter size	MP:0001933	22.73% (5/22)	2.75% (265/9633)	0.0002936	0.012
Mouse	short facial bone	MP:0030384	9.09% (2/22)	0.81% (78/9633)	0.014	0.098
Human	Hypermetropia	HP:0000540	25.0% (2/8)	1.01% (37/3655)	0.003	0.095
Human	Abnormality of the optic nerve	HP:0000587	50.0% (4/8)	14.75% (539/3655)	0.02	0.095
Human	Abnormality of the ribs	HP:0000772	37.5% (3/8)	4.38% (160/3655)	0.004	0.076
Human	Missing ribs	HP:0000921	25.0% (2/8)	0.47% (17/3655)	0.0007007	0.015
Human	Interphalangeal joint contracture of finger	HP:0001220	25.0% (2/8)	4.19% (153/3655)	0.042	0.066
Human	Slender finger	HP:0001238	25.0% (2/8)	2.24% (82/3655)	0.013	0.091
Human	Abnormal mitral valve morphology	HP:0001633	25.0% (2/8)	1.86% (68/3655)	0.009	0.09
Human	Conotruncal defect	HP:0001710	25.0% (2/8)	3.06% (112/3655)	0.024	0.095
Human	Abnormality of pelvic girdle bone morphology	HP:0002644	50.0% (4/8)	9.66% (353/3655)	0.005	0.095
Human	Bowing of the legs	HP:0002979	25.0% (2/8)	3.34% (122/3655)	0.028	0.083
Human	Abnormal anterior segment morphology	HP:0004328	75.0% (6/8)	24.02% (878/3655)	0.003	0.095
Human	Reduced bone mineral density	HP:0004349	37.5% (3/8)	8.34% (305/3655)	0.024	0.095
Human	Neuroblastic tumors	HP:0004376	25.0% (2/8)	0.88% (32/3655)	0.002	0.076
Human	Aortic aneurysm	HP:0004942	25.0% (2/8)	1.5% (55/3655)	0.006	0.076
Human	Abnormality of phalangeal joints of the hand	HP:0006261	25.0% (2/8)	4.35% (159/3655)	0.045	0.085
Human	Bowing of the long bones	HP:0006487	25.0% (2/8)	3.67% (134/3655)	0.033	0.085
Human	Lipid accumulation in hepatocytes	HP:0006561	25.0% (2/8)	2.85% (104/3655)	0.021	0.095
Human	Aplasia/Hypoplasia of the ribs	HP:0006712	25.0% (2/8)	2.13% (78/3655)	0.012	0.091
Human	Iris hypopigmentation	HP:0007730	25.0% (2/8)	0.52% (19/3655)	0.0008587	0.015

Human	Corneal opacity	HP:0007957	25.0% (2/8)	3.09% (113/3655)	0.024	0.095
Human	Abnormal iris pigmentation	HP:0008034	25.0% (2/8)	2.87% (105/3655)	0.021	0.095
Human	Joint contracture of the hand	HP:0009473	25.0% (2/8)	4.27% (156/3655)	0.044	0.085
Human	Dilatation of the renal pelvis	HP:0010946	25.0% (2/8)	3.78% (138/3655)	0.035	0.085
Human	Functional abnormality of the middle ear	HP:0011452	50.0% (4/8)	6.05% (221/3655)	0.0007976	0.076
Human	Abnormality of corneal thickness	HP:0011486	25.0% (2/8)	2.68% (98/3655)	0.019	0.095
Human	Camptodactyly	HP:0012385	25.0% (2/8)	4.3% (157/3655)	0.044	0.085
Human	Thoracic aortic aneurysm	HP:0012727	25.0% (2/8)	1.45% (53/3655)	0.006	0.076
Human	Flexion contracture of finger	HP:0012785	25.0% (2/8)	4.27% (156/3655)	0.044	0.083
Human	Abnormality of the optic disc	HP:0012795	50.0% (4/8)	12.23% (447/3655)	0.011	0.063
Human	Flexion contracture of digit	HP:0030044	37.5% (3/8)	6.02% (220/3655)	0.01	0.09
Human	Neuroectodermal neoplasm	HP:0030061	25.0% (2/8)	1.37% (50/3655)	0.005	0.095
Human	Primitive neuroectodermal tumor	HP:0030065	25.0% (2/8)	0.93% (34/3655)	0.003	0.076
Human	Peripheral primitive neuroectodermal neoplasm	HP:0030067	25.0% (2/8)	0.88% (32/3655)	0.002	0.023
Human	Asymmetric growth	HP:0100555	25.0% (2/8)	0.77% (28/3655)	0.002	0.084
Human	Decreased corneal thickness	HP:0100689	25.0% (2/8)	2.63% (96/3655)	0.018	0.068
Human	Increased corneal curvature	HP:0100692	25.0% (2/8)	2.63% (96/3655)	0.018	0.068
Human	Long fingers	HP:0100807	25.0% (2/8)	3.99% (146/3655)	0.039	0.085

Abbreviations are as follows: FDR, False Discovery Rate

CITATIONS

Adams, R. H., Schield, D. R., Card, D. C., Blackmon, H., & Castoe, T. A. (2017). GppFst: genomic posterior predictive simulations of FST and dXY for identifying outlier loci from population genomic data. *Bioinformatics*, *33*(9), 1414–1415. https://doi.org/10.1093/bioinformatics/btw795

Adams, R. H., Schield, D. R., Card, D. C., Corbin, A., Castoe, T. A., & Stegle, O. (2018). ThetaMater: Bayesian estimation of population size parameter θ from genomic data. *Bioinformatics*, *34*(6), 1072–1073. https://doi.org/10.1093/bioinformatics/btx733

Adeva-Andany, M., Souto-Adeva, G., Ameneiros-Rodríguez, E., Fernández-Fernández, C., Donapetry-García, C., & Domínguez-Montero, A. (2018). Insulin resistance and glycine metabolism in humans. *Amino Acids*, *50*(1), 11–27. https://doi.org/10.1007/s00726-017-2508-0

Adler, G. H., & Levins, R. (1994). The Island Syndrome in Rodent Populations. *The Quarterly Review of Biology*, *69*(4), 473–490.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amit, T., Hertz, P., Ish-Shalom, S., Lotan, R., Luboshitzki, R., Youdim, M. B., & Hochberg, Z. (1991). Effects of hypo or hyper-thyroidism on growth hormone-binding protein. *Clinical Endocrinology*, *35*(2), 159–162. https://doi.org/10.1111/j.1365-2265.1991.tb03515.x

Azevedo, A., Schwartz, I. V., Kalakun, L., Brustolin, S., Burin, M. G., Beheregaray, A. P. C., … Giugliani, R. (2004). Clinical and biochemical study of 28 patients with mucopolysaccharidosis type VI. *Clinical Genetics*, *66*(3), 208–213. https://doi.org/10.1111/j.1399-0004.2004.00277.x

Babu, E., Ananth, S., Veeranan-Karmegam, R., Coothankandaswamy, V., Smith, S. B., Boettger, T., … Martin, P. M. (2011). Transport via SLC5A8 (SMCT1) Is Obligatory for 2-Oxothiazolidine-4-Carboxylate to Enhance Glutathione Production in Retinal Pigment Epithelial Cells. *Investigative Ophthalmology & Visual Science*, *52*(8), 5749–5757. https://doi.org/10.1167/iovs.10-6825

Baker, J., Liu, J.-P., Robertson, E. J., & Efstratiadis, A. (1993). Role of insulin-like growth factors in embryonic and postnatal growth. *Cell*, *75*(1), 73–82. https://doi.org/10.1016/S0092-8674(05)80085-6

Batt, J., Asa, S., Fladd, C., & Rotin, D. (2002). Pituitary, Pancreatic and Gut Neuroendocrine Defects in Protein Tyrosine Phosphatase- Sigma-Deficient Mice. *Molecular Endocrinology*, *16*(1), 155–169. https://doi.org/10.1210/mend.16.1.0756

Becker, N. S., Verdu, P., Georges, M., Duquesnoy, P., Froment, A., Amselem, S., … Heyer, E. (2013). The role of *GHR* and *IGF1* genes in the genetic determination of African pygmies' short stature. *European Journal of Human Genetics*, *21*(6), 653–658. https://doi.org/10.1038/ejhg.2012.223

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Bergey, C. M., Lopez, M., Harrison, G. F., Patin, E., Cohen, J., Quintana-Murci, L., … Perry, G. H. (2018). Polygenic adaptation and convergent evolution across both growth and cardiac genetic pathways in African and Asian rainforest hunter-gatherers. *BioRxiv*, 300574. https://doi.org/10.1101/300574

Bhattacharyya, S., Feferman, L., Tobacman, J. K., Bhattacharyya, S., Feferman, L., & Tobacman, J. K. (2017). Chondroitin sulfatases differentially regulate Wnt signaling in prostate stem cells through effects on SHP2, phospho-ERK1/2, and Dickkopf Wnt signaling pathway inhibitor (DKK3). *Oncotarget*, *8*(59), 100242–100260. https://doi.org/10.18632/oncotarget.22152

Boback, S. M. (2005). Natural History and Conservation of Island Boas (*Boa constrictor*) in Belize. *Copeia*, *2005*(4), 879–884. https://doi.org/10.1643/0045-8511(2005)005[0879:NHACOI]2.0.CO;2

Boback, S. M. (2006). A Morphometric Comparison of Island and Mainland Boas (*Boa constrictor*) in Belize. *Copeia*, *2006*(2), 261–267. https://doi.org/10.1643/0045-8511(2006)6[261:AMCOIA]2.0.CO;2

Boback, S. M., & Carpenter, D. M. (2007). Body size and head shape of island *Boa constrictor* in Belize: environmental versus genetic contributions. In R. W. Henderson & R. Powell (Eds.), *Biology of the Boas and Pythons* (pp. 102–117). Eagle Mountain, UT: Eagle Mountain Publishing.

Boback, S. M., & Guyer, C. (2003). Empirical Evidence for an Optimal Body Size in Snakes. *Evolution*, *57*(2), 345–451. https://doi.org/10.1111/j.0014-3820.2003.tb00268.x

Boback, S. M., & Siefferman, L. M. (2010). Variation in Color and Color Change in Island and Mainland Boas (*Boa constrictor*). *Journal of Herpetology*, *44*(4), 506–515. https://doi.org/10.1670/09-026.1

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bouwman, A. C., Daetwyler, H. D., Chamberlain, A. J., Ponce, C. H., Sargolzaei, M., Schenkel, F. S., … Hayes, B. J. (2018). Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature Genetics*, *50*(3), 362–367. https://doi.org/10.1038/s41588-018-0056-5

Boyko, A. R., Quignon, P., Li, L., Schoenebeck, J. J., Degenhardt, J. D., Lohmueller, K. E., … Ostrander, E. A. (2010). A Simple Genetic Architecture Underlies Morphological Variation in Dogs. *PLOS Biology*, *8*(8), e1000451. https://doi.org/10.1371/journal.pbio.1000451

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., … Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1), 1–31. https://doi.org/10.1186/2047-217X-2-10

Brugmann, S. A., Powder, K. E., Young, N. M., Goodnough, L. H., Hahn, S. M., James, A. W., … Lovett, M. (2010). Comparative gene expression analysis of avian embryonic facial structures reveals new candidates for human craniofacial disorders. *Human Molecular Genetics*, *19*(5), 920–930. https://doi.org/10.1093/hmg/ddp559

Brugmann, Samantha A., Goodnough, L. H., Gregorieff, A., Leucht, P., Berge, D. ten, Fuerer, C., … Helms, J. A. (2007). Wnt signaling mediates regional specification in the vertebrate face. *Development*, *134*(18), 3283–3295. https://doi.org/10.1242/dev.005132

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Molecular Biology and Evolution*, *29*(8), 1917–1932. https://doi.org/10.1093/molbev/mss086

Card, D. C., Schield, D. R., Adams, R. H., Corbin, A. B., Perry, B. W., Andrew, A. L., … Castoe, T. A. (2016). Phylogeographic and population genetic analyses reveal multiple species of *Boa* and independent origins of insular dwarfism. *Molecular Phylogenetics and Evolution*, *102*, 104–116. https://doi.org/10.1016/j.ympev.2016.05.034

Carneiro, M., Hu, D., Archer, J., Feng, C., Afonso, S., Chen, C., … Andersson, L. (2017). Dwarfism and Altered Craniofacial Development in Rabbits Is Caused by a 12.1 kb Deletion at the HMGA2 Locus. *Genetics*, *205*(2), 955–965. https://doi.org/10.1534/genetics.116.196667

Castoe, T. A., Koning, A. P. J. de, Kim, H.-M., Gu, W., Noonan, B. P., Naylor, G., … Pollock, D. D. (2009). Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences*, *106*(22), 8986–8991. https://doi.org/10.1073/pnas.0900233106

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: Building and Genotyping Loci *De Novo* From Short-Read Sequences. *G3: Genes, Genomes, Genetics*, *1*(3), 171–182. https://doi.org/10.1534/g3.111.000240

Catchen, J. M., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Choi, H., Ryu, K.-Y., Roh, J., & Bae, J. (2018). Effect of radioactive iodine-induced hypothyroidism on longitudinal bone growth during puberty in immature female rats. *Experimental Animals*, 18–0013. https://doi.org/10.1538/expanim.18-0013

Choi, Y. (2012). A Fast Computation of Pairwise Sequence Alignment Scores Between a Protein and a Set of Single-locus Variants of Another Protein. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine* (pp. 414–417). New York, NY, USA: ACM. https://doi.org/10.1145/2382936.2382989

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLOS ONE*, *7*(10), e46688. https://doi.org/10.1371/journal.pone.0046688

Chung, J., Zhang, X., Collins, B., Sper, R. B., Gleason, K., Simpson, S., … Piedrahita, J. A. (2018). High mobility group A2 (HMGA2) deficiency in pigs leads to dwarfism, abnormal fetal resource allocation, and cryptorchidism. *Proceedings of the National Academy of Sciences*, *115*(21), 5420–5425. https://doi.org/10.1073/pnas.1721630115

Covas, R. (2012). Evolution of reproductive life histories in island birds worldwide. *Proceedings of the Royal Society of London B: Biological Sciences*, *279*(1733), 1531–1537. https://doi.org/10.1098/rspb.2011.1785

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491. https://doi.org/10.1038/ng.806

Diamond, J. M. (1972). Biogeographic Kinetics: Estimation of Relaxation Times for Avifaunas of Southwest Pacific Islands. *Proceedings of the National Academy of Sciences*, *69*(11), 3199–3203. https://doi.org/10.1073/pnas.69.11.3199

Divry, P., David, M., Gregersen, N., Kølvraa, S., Christensen, E., Collet, J. P., … Cotte, J. (1983). Dicarboxylic aciduria due to medium chain actyl CoA dehydrogenase defect. A cause of hypoglycemia in childhood. *Acta Paediatrica*, *72*(6), 943–949. https://doi.org/10.1111/j.1651-2227.1983.tb09849.x

D'Mello, S. A. N., Finlay, G. J., Baguley, B. C., & Askarian-Amiri, M. E. (2016). Signaling Pathways in Melanogenesis. *International Journal of Molecular Sciences*, *17*(7), 1144. https://doi.org/10.3390/ijms17071144

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*, 214. https://doi.org/10.1186/1471-2148-7-214

Duran, M., Klerk, J. B. C. D., Wadman, S. K., Bruinvis, L., & Ketting, D. (1984). The Differential Diagnosis of Dicarboxylic Aciduria. In *Organic Acidurias* (pp. 48–51). Springer, Dordrecht. https://doi.org/10.1007/978-94-009-5612-4_11

Elchebly, M., Wagner, J., Kennedy, T. E., Lanctôt, C., Michaliszyn, E., Itié, A., … Tremblay, M. L. (1999). Neuroendocrine dysplasia in mice lacking protein tyrosine phosphatase σ. *Nature Genetics*, *21*(3), 330–333. https://doi.org/10.1038/6859

Fink, T., Tiplady, K., Lopdell, T., Johnson, T., Snell, R. G., Spelman, R. J., … Littlejohn, M. D. (2017). Functional confirmation of *PLAG1* as the candidate causative gene underlying major pleiotropic effects on body weight and milk characteristics. *Scientific Reports*, *7*, 44793. https://doi.org/10.1038/srep44793

Foster, J. B. (1964). Evolution of Mammals on Islands. *Nature*, *202*(4929), 234–235. https://doi.org/10.1038/202234a0

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., & Nielsen, R. (2011). Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *PLOS Genetics*, *7*(11), e1002355. https://doi.org/10.1371/journal.pgen.1002355

Gallant, J. R., Traeger, L. L., Volkening, J. D., Moffett, H., Chen, P.-H., Novina, C. D., … Sussman, M. R. (2014). Genomic basis for the convergent evolution of electric organs. *Science*, *344*(6191), 1522–1525. https://doi.org/10.1126/science.1254432

Gischler, E. (2014). Quaternary reef response to sea-level and environmental change in the western Atlantic. *Sedimentology*, *62*(2), 429–465. https://doi.org/10.1111/sed.12174

Global Lipids Genetics Consortium, Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., … Abecasis, G. R. (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, *45*(11), 1274–1283. https://doi.org/10.1038/ng.2797

Gordon, A., & Hannon, G. (2010). *Fastx-toolkit.*

Gordon, P. R., & Gilchrest, B. A. (1989). Human Melanogenesis is Stimulated by Diacylglycerol. *Journal of Investigative Dermatology*, *93*(5), 700–702. https://doi.org/10.1111/1523-1747.ep12319900

Grant, P. R., & Grant, B. R. (2006). Evolution of Character Displacement in Darwin's Finches. *Science*, *313*(5784), 224–226. https://doi.org/10.1126/science.1128374

Gray, M. M., Parmenter, M. D., Hogan, C. A., Ford, I., Cuthbert, R. J., Ryan, P. G., … Payseur, B. A. (2015). Genetics of Rapid and Extreme Size Evolution in Island Mice. *Genetics*, *201*(1), 213–228. https://doi.org/10.1534/genetics.115.177790

Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., … Stefansson, K. (2008). Many sequence variants affecting diversity of adult human height. *Nature Genetics*, *40*(5), 609–615. https://doi.org/10.1038/ng.122

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*, *5*(10), e1000695. https://doi.org/10.1371/journal.pgen.1000695

Hale, A. J., ter Steege, E., & den Hertog, J. (2017). Recent advances in understanding the role of protein-tyrosine phosphatases in development and disease. *Developmental Biology*, *428*(2), 283–292. https://doi.org/10.1016/j.ydbio.2017.03.023

Henderson, R. W., & Binder, M. H. (1980). The Ecology and Behavior of Vine Snakes (*Ahaetulla, Oxybelis, Thelotornis, Uromacer*): A Review. *Mil. Pub. Mus. Cont. Biol. Geol.*, *1*, 1–38.

Henderson, R. W., Waller, T., Micucci, P., Puorto, G., & Bourgeois, R. W. (1995). Ecological correlates and patterns in the distribution of Neotropical boines (Serpentes: Boidae): a preliminary assessment. *Herpetological Natural History*, *3*, 1.

Hoekstra, H. E., Hirschmann, R. J., Bundey, R. A., Insel, P. A., & Crossland, J. P. (2006). A Single Amino Acid Mutation Contributes to Adaptive Beach Mouse Color Pattern. *Science*, *313*(5783), 101–104. https://doi.org/10.1126/science.1126121

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010). Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLOS Genetics*, *6*(2), e1000862. https://doi.org/10.1371/journal.pgen.1000862

Hurst, L. D., & Smith, N. G. C. (1999). Do essential genes evolve slowly? *Current Biology*, *9*(14), 747–750. https://doi.org/10.1016/S0960-9822(99)80334-0

Hynková, I., Starostová, Z., & Frynta, D. (2009). Mitochondrial DNA Variation Reveals Recent Evolutionary History of Main *Boa constrictor* Clades. *Zoological Science*, *26*(9), 623–631. https://doi.org/10.2108/zsj.26.623

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944

Kawano, Y., Kitaoka, M., Hamada, Y., Walker, M. M., Waxman, J., & Kypta, R. M. (2006). Regulation of prostate cell growth and morphogenesis by Dickkopf-3. *Oncogene*, *25*(49), 6528–6537. https://doi.org/10.1038/sj.onc.1209661

Kemper, K. E., Visscher, P. M., & Goddard, M. E. (2012). Genetic architecture of body size in mammals. *Genome Biology*, *13*, 244. https://doi.org/10.1186/gb-2012-13-4-244

King, R. B. (1987). Color Pattern Polymorphism in the Lake Erie Water Snake, *Nerodia sipedon insularum*. *Evolution*, *41*(2), 241–255. https://doi.org/10.1111/j.1558-5646.1987.tb05794.x

Köhler, M., & Moyà-Solà, S. (2009). Physiological and life history strategies of a fossil large mammal in a resource-limited environment. *Proceedings of the National Academy of Sciences*, *106*(48), 20354–20358. https://doi.org/10.1073/pnas.0813385106

Kurosaka, H., Iulianella, A., Williams, T., & Trainor, P. A. (2014). Disrupting hedgehog and WNT signaling interactions promotes cleft lip pathogenesis. *The Journal of Clinical Investigation*, *124*(4), 1660–1671. https://doi.org/10.1172/JCI72688

Larter, M., Dunbar-Wallis, A., Berardi, A. E., Smith, S. D., & Purugganan, M. (2018). Convergent Evolution at the Pathway Level: Predictable Regulatory Changes during Flower Color Transitions. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msy117

Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I., Sanna, S., … Hirschhorn, J. N. (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature Genetics*, *40*(5), 584–591. https://doi.org/10.1038/ng.125

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lillywhite, H. B., & Henderson, R. W. (2002). Behavioral and functional ecology of arboreal snakes. In J. T. Collins & R. A. Seigel (Eds.), *Snakes: Ecology and Behavior* (pp. 1–48). Caldwell, N.J: The Blackburn Press.

Liu, L., & Yu, L. (2010). Phybase: an R package for species tree analysis. *Bioinformatics*, *26*(7), 962–963. https://doi.org/10.1093/bioinformatics/btq062

Logan, C. Y., & Nusse, R. (2004). The Wnt Signaling Pathway in Development and Disease. *Annual Review of Cell and Developmental Biology*, *20*(1), 781–810. https://doi.org/10.1146/annurev.cellbio.20.010403.113126

Lomolino, M. V., Geer, A. A. van der, Lyras, G. A., Palombo, M. R., Sax, D. F., & Rozzi, R. (2013). Of mice and mammoths: generality and antiquity of the island rule. *Journal of Biogeography*, *40*(8), 1427–1439. https://doi.org/10.1111/jbi.12096

Lomolino, M. V., Riddle, B. R., Whittaker, R. J., & Brown, J. H. (2010). *Biogeography* (4th ed.). Sunderland, MA: Sinauer Associates.

Lomolino, M. V., Sax, D. F., Palombo, M. R., & Geer, A. A. van der. (2012). Of mice and mammoths: evaluations of causal explanations for body size evolution in insular mammals. *Journal of Biogeography*, *39*(5), 842–854. https://doi.org/10.1111/j.1365-2699.2011.02656.x

Losos, J. B., Warheitt, K. I., & Schoener, T. W. (1997). Adaptive differentiation following experimental island colonization in *Anolis* lizards. *Nature*, *387*(6628), 70–73. https://doi.org/10.1038/387070a0

Magnusson, M., Wang, T. J., Clish, C., Engström, G., Nilsson, P., Gerszten, R. E., & Melander, O. (2015). Dimethylglycine Deficiency and the Development of Diabetes. *Diabetes*, *64*(8), 3010–3016. https://doi.org/10.2337/db14-1863

Mazzullo, S. J. (2006). Late Pliocene to Holocene platform evolution in northern Belize, and comparison with coeval deposits in southern Belize and the Bahamas. *Sedimentology*, *53*(5), 1015–1047. https://doi.org/10.1111/j.1365-3091.2006.00800.x

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., … Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*, 122. https://doi.org/10.1186/s13059-016-0974-4

McTaggart, S. J., Obbard, D. J., Conlon, C., & Little, T. J. (2012). Immune genes undergo more adaptive evolution than non-immune system genes in *Daphnia pulex*. *BMC Evolutionary Biology*, *12*, 63. https://doi.org/10.1186/1471-2148-12-63

Nachman, M. W., Hoekstra, H. E., & D'Agostino, S. L. (2003). The genetic basis of adaptive melanism in pocket mice. *Proceedings of the National Academy of Sciences*, *100*(9), 5268–5273. https://doi.org/10.1073/pnas.0431157100

Nelder, J. A., & Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, *7*(4), 308–313. https://doi.org/10.1093/comjnl/7.4.308

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, *18*(3), 375–402. https://doi.org/10.1111/j.1365-294X.2008.03946.x

Obbard, D. J., Welch, J. J., Kim, K.-W., & Jiggins, F. M. (2009). Quantifying Adaptive Evolution in the *Drosophila* Immune System. *PLOS Genetics*, *5*(10), e1000698. https://doi.org/10.1371/journal.pgen.1000698

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, *26*(3), 419–420. https://doi.org/10.1093/bioinformatics/btp696

Parmenter, M. D., Gray, M. M., Hogan, C. A., Ford, I. N., Broman, K. W., Vinyard, C. J., & Payseur, B. A. (2016). Genetics of Skeletal Evolution in Unusually Large Mice from Gough Island. *Genetics*, *204*(4), 1559–1572. https://doi.org/10.1534/genetics.116.193805

Parsons, K. J., Taylor, A. T., Powder, K. E., & Albertson, R. C. (2014). Wnt signalling underlies the evolution of new phenotypes and craniofacial variability in Lake Malawi cichlids. *Nature Communications*, *5*, 3629. https://doi.org/10.1038/ncomms4629

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for *De Novo* SNP Discovery and Genotyping in Model and Non-Model Species. *PLOS ONE*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., … Scherer, S. W. (2014). Convergence of Genes and Cellular Pathways Dysregulated in Autism Spectrum Disorders. *The American Journal of Human Genetics*, *94*(5), 677–694. https://doi.org/10.1016/j.ajhg.2014.03.018

Porras, L. (1999). Island boa constrictors (Boa constrictor). *Reptiles*, *7*, 48–61.

Portik, D. M., Leaché, A. D., Rivera, D., Barej, M. F., Burger, M., Hirschfeld, M., … Fujita, M. K. (2017). Evaluating mechanisms of diversification in a Guineo-Congolian tropical forest frog using demographic model selection. *Molecular Ecology*, *26*(19), 5245–5263. https://doi.org/10.1111/mec.14266

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Reed, R. N., Boback, S. M., Montgomery, C. E., Green, S., Stevens, Z., & Watson, D. (2007). Ecology and conservation of an exploited insular population of *Boa constrictor* (Squamata: Boidae) on the Cayos Cochinos, Honduras. In R. W. Henderson & R. Powell (Eds.), *Biology of the Boas and Pythons* (pp. 289–403). Eagle Mountain, UT: Eagle Mountain Publishing.

Reynolds, R. G., Niemiller, M. L., & Revell, L. J. (2014). Toward a Tree-of-Life for the boas and pythons: Multilocus species-level phylogeny with unprecedented taxon sampling. *Molecular Phylogenetics and Evolution*, *71*, 201–213. https://doi.org/10.1016/j.ympev.2013.11.011

Rhead, W. J., Amendt, B. A., Fritchman, K. S., & Felts, S. J. (1983). Dicarboxylic aciduria: deficient [1-14C]octanoate oxidation and medium-chain acyl-CoA dehydrogenase in fibroblasts. *Science*, *221*(4605), 73–75. https://doi.org/10.1126/science.6857268

Root, A. W., Shulman, D., Root, J., & Diamond, F. (1986). The interrelationships of thyroid and growth hormones: effect of growth hormone releasing hormone in hypo- and hyperthyroid male rats. *Acta Endocrinologica*, *113*(4 Suppl), S367–S375. https://doi.org/10.1530/acta.0.112S367

Rosenblum, E. B., Hoekstra, H. E., & Nachman, M. W. (2007). Adaptive Reptile Color Variation and the Evolution of the MC1R Gene. *Evolution*, *58*(8), 1794–1808. https://doi.org/10.1111/j.0014-3820.2004.tb00462.x

Rosenblum, E. B., Römpler, H., Schöneberg, T., & Hoekstra, H. E. (2010). Molecular and functional basis of phenotypic convergence in white lizards at White Sands. *Proceedings of the National Academy of Sciences*, *107*(5), 2113–2117. https://doi.org/10.1073/pnas.0911042107

Schlenke, T. A., & Begun, D. J. (2003). Natural Selection Drives Drosophila Immune System Evolution. *Genetics*, *164*(4), 1471–1480.

Schmidt, C., & Patel, K. (2005). Wnts and the neural crest. *Anatomy and Embryology*, *209*(5), 349–355. https://doi.org/10.1007/s00429-005-0459-9

Sedlazeck, F. J., Rescheneder, P., & von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, *29*(21), 2790–2791. https://doi.org/10.1093/bioinformatics/btt468

Seehausen, O. (2006). African cichlid fish: a model system in adaptive radiation research. *Proceedings of the Royal Society of London B: Biological Sciences*, *273*(1597), 1987–1998. https://doi.org/10.1098/rspb.2006.3539

Shine, R. (1983). Arboreality in Snakes: Ecology of the Australian Elapid Genus *Hoplocephalus*. *Copeia*, *1983*(1), 198–205. https://doi.org/10.2307/1444714

Smith, C. L., Blake, J. A., Kadin, J. A., Richardson, J. E., & Bult, C. J. (2018). Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Research*, *46*(D1), D836–D842. https://doi.org/10.1093/nar/gkx1006

Soy, J., Leivar, P., González-Schain, N., Martín, G., Diaz, C., Sentandreu, M., … Monte, E. (2016). Molecular convergence of clock and photosensory pathways through PIF3–TOC1 interaction and co-occupancy of target promoters. *Proceedings of the National Academy of Sciences*, *113*(17), 4870–4875. https://doi.org/10.1073/pnas.1603745113

Steiner, C. C., Weber, J. N., & Hoekstra, H. E. (2007). Adaptive Variation in Beach Mice Produced by Two Interacting Pigmentation Genes. *PLOS Biology*, *5*(9), e219. https://doi.org/10.1371/journal.pbio.0050219

Stewart, K., Uetani, N., Hendriks, W., Tremblay, M. L., & Bouchard, M. (2013). Inactivation of LAR family phosphatase genes Ptprs and Ptprf causes craniofacial malformations resembling Pierre-Robin sequence. *Development*, *140*(16), 3413–3422. https://doi.org/10.1242/dev.094532

Suárez-Atilano, M., Burbrink, F., & Vázquez-Domínguez, E. (2014). Phylogeographical structure within *Boa constrictor imperator* across the lowlands and mountains of Central America and Mexico. *Journal of Biogeography*, *41*(12), 2371–2384. https://doi.org/10.1111/jbi.12372

Surakka, I., Horikoshi, M., Mägi, R., Sarin, A.-P., Mahajan, A., Lagou, V., … Consortium, E. (2015). The impact of low-frequency and rare variants on lipid levels. *Nature Genetics*, *47*(6), 589–597. https://doi.org/10.1038/ng.3300

Sutter, N. B., Bustamante, C. D., Chase, K., Gray, M. M., Zhao, K., Zhu, L., … Ostrander, E. A. (2007). A Single IGF1 Allele Is a Major Determinant of Small Size in Dogs. *Science*, *316*(5821), 112–115. https://doi.org/10.1126/science.1137045

Ueno, K., Hirata, H., Shahryari, V., Deng, G., Tanaka, Y., Tabatabai, Z. L., … Dahiya, R. (2013). microRNA-183 is an oncogene targeting Dkk-3 and SMAD4 in prostate cancer. *British Journal of Cancer*, *108*(8), 1659–1667. https://doi.org/10.1038/bjc.2013.125

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc. https://doi.org/10.1002/0471250953.bi1110s43

Veeck, J., & Dahl, E. (2012). Targeting the Wnt pathway in cancer: The emerging role of Dickkopf-3. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1825*(1), 18–28. https://doi.org/10.1016/j.bbcan.2011.09.003

Visscher, P. M. (2008). Sizing up human height variation. *Nature Genetics*, *40*(5), 489–490. https://doi.org/10.1038/ng0508-489

Wang, P., Margolis, C., Lin, G., Buza, E. L., Quick, S., Raj, K., … Giger, U. (2018). Mucopolysaccharidosis Type VI in a Great Dane Caused by a Nonsense Mutation in the ARSB Gene. *Veterinary Pathology*, *55*(2), 286–293. https://doi.org/10.1177/0300985817732115

Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., … Frayling, T. M. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics*, *40*(5), 575–583. https://doi.org/10.1038/ng.121

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Weissglas-Volkov, D., Calkin, A. C., Tusie-Luna, T., Sinsheimer, J. S., Zelcer, N., Riba, L., … Pajukanta, P. (2011). The N342S MYLIP polymorphism is associated with high total cholesterol and increased LDL receptor degradation in humans. *The Journal of Clinical Investigation*, *121*(8), 3062–3071. https://doi.org/10.1172/JCI45504

Weng, M.-P., & Liao, B.-Y. (2017). modPhEA: model organism Phenotype Enrichment Analysis of eukaryotic gene sets. *Bioinformatics*, *33*(21), 3505–3507. https://doi.org/10.1093/bioinformatics/btx426

Woods, K. A., Camacho-Hübner, C., Barter, D., Clark, A. J. L., & Savage, M. O. (1997). Insulin-like growth factor I gene deletion causing intrauterine growth retardation and severe short stature. *Acta Paediatrica*, *86*(S423), 39–45. https://doi.org/10.1111/j.1651-2227.1997.tb18367.x

Woods, Katie A., Camacho-Hübner, C., Savage, M. O., & Clark, A. J. L. (1996). Intrauterine Growth Retardation and Postnatal Growth Failure Associated with Deletion of the Insulin-Like Growth Factor I Gene. *New England Journal of Medicine*, *335*(18), 1363–1367. https://doi.org/10.1056/NEJM199610313351805

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. https://doi.org/10.1038/ng.608

Zhang, B., Kirov, S., & Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*, *33*(suppl_2), W741–W748. https://doi.org/10.1093/nar/gki475

Zoledziewska, M., Sidore, C., Chiang, C. W. K., Sanna, S., Mulas, A., Steri, M., … Cucca, F. (2015). Height-reducing variants and selection for short stature in Sardinia. *Nature Genetics*, *47*(11), 1352–1356. https://doi.org/10.1038/ng.3403

Chapter 6.

Novel ecological and climatic conditions drive rapid adaptation in invasive Florida Burmese pythons

Daren C. Card[1], Blair W. Perry[1], Richard H. Adams[1], Drew R. Schield[1], Acacia S. Young[1], Audra L. Andrew[1], Tereza Jezkova[2], Giulia I.M. Pasquesi[1], Nicole R. Hales[1], Matthew R. Walsh[1], Michael R. Rochford[3], Frank J. Mazzotti[3], Kristen M. Hart[4], Margaret E. Hunter[5], and Todd A. Castoe[1,*]

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

[2] Department of Biology, Miami University, Oxford, OH, 45056, USA

[3] Fort Lauderdale Research & Education Center, University of Florida, Fort Lauderdale, FL, 33314, USA

[4] United States Geological Survey, Wetland and Aquatic Research Center, Davie, FL, 33314, USA

[5] United States Geological Survey, Wetland and Aquatic Research Center, Gainesville, FL, 32653, USA

ABSTRACT

Invasive species provide powerful *in situ* experimental systems for studying evolution in response to selective pressures in novel habitats. While research has shown that phenotypic evolution can occur rapidly in nature, few examples exist of genome-wide adaptation on short 'ecological' timescales. Burmese pythons (*Python molurus bivittatus*) have become a successful and impactful invasive species in Florida over the last 30 years despite major freeze events that caused high python mortality. We sampled Florida pythons before and after a major freeze event in 2010 and found evidence for positive selection in genomic regions enriched for genes associated with thermosensation, behavior, and physiology. Several of these genes are linked to regenerative organ growth, an adaptive response that modulates organ size and function with feeding and fasting in pythons. Independent histological and functional genomic datasets provide additional layers of support for a contemporary shift in invasive Burmese python physiology. In the Florida population, a shift towards maintaining an active digestive system may be driven by the fitness benefits of maintaining higher metabolic rates and body temperature during freeze events. Our results suggest that a synergistic interaction between ecological and climatic selection pressures have driven adaptation in Florida Burmese pythons, demonstrating the often-overlooked potential of rapid adaptation to influence the success of invasive species.

INTRODUCTION

The most striking examples of evolution involve rapid phenotypic adaptation in natural populations (Grant & Grant, 2002; Losos, Warheitt, & Schoener, 1997), but few studies have linked genomic change to phenotypic evolution occurring over a small number of generations (though see Campbell-Staton et al., [2017], Epstein et al., [2016], and Reid et al., [2016]). Invasive species are valuable models for understanding such links because they are often subjected to strong selective pressures due to the novelty of environmental conditions they face in non-native environments (Reznick & Ghalambor, 2001; Schoener, 2011), and their success may depend more heavily on adaptability than on physiological plasticity (Lee, 2002).

Among the most widely known and impactful invasive species in the United States is the Burmese python (*Python molurus bivittatus*; Engeman, Jacobson, Avery, & Meshaka, 2011; Willson, Dorcas, & Snow, 2011). The Burmese python is a large constricting snake native to southeast Asia (Barker & Barker, 2008) that has received substantial attention due to their recent and highly successful invasive colonization of south Florida (Engeman et al., 2011; Willson et al., 2011). Burmese pythons were first discovered in Florida in the early 1980's (Meshaka, Loftus, & Steiner, 2000), and were considered established by the mid-1990's (Collins, Freeman, & Snow, 2008; Snow, Brien, Cherkiss, Wilkins, & Mazzotti, 2007). This population is thought to have originated from the release of pet pythons, including a catastrophic release event resulting from the destruction of an animal import facility during Hurricane Andrew in 1992 (Willson et al., 2011). The ecological impact resulting from predation on endangered species by pythons within Florida's Everglades National Park (ENP) is extensive, and the economic impact is estimated to be at least $83,892 per snake per year (Smith, Sementelli, Meshaka, & Engeman, 2007). These snakes prey upon many bird and mammal species endemic to ENP and listed under the US Endangered Species Act (Dove, Snow, Rochford, & Mazzotti, 2011; Reed, 2005; Snow et al., 2007) and have been implicated in recent massive declines in small mammal populations (Dorcas et al., 2012).

Several lines of evidence suggest that invasive Florida Burmese pythons may be under substantial selection pressures. First, invasive Burmese pythons reside at the margin of climatically suitable habitat within the United States (Pyron, Burbrink, & Guiher, 2008) and several studies have found high cold-induced mortality in Burmese pythons relocated to more temperate areas north of Florida (Avery et al., 2010; Dorcas, Willson, & Gibbons, 2011; Jacobson et al., 2012). Moreover, acute climatic events, including rapid shifts in temperature, also periodically impact South Florida. For example, 50-90% mortality was documented in South Florida python populations during a freeze event in January 2010 (Mazzotti et al., 2011). Collectively, this suggests that the more temperate environment in Florida (compared to tropical Southeast Asia) imposes strong selection pressures on the invasive Burmese python population.

In addition to being ill suited to the sub-tropical climates of Florida, the invasive Burmese python population has experienced a fundamental shift in prey ecology. The ecology and physiology of Burmese pythons has been strongly shaped by the monsoonal ecosystems of their native Southeast Asia, where they experience major seasonal shifts in prey availability. Indeed, Burmese pythons represent an important and unique model system for studying extreme physiological regulation (Secor, 2008; Secor & Diamond, 1995, 1998). These snakes have adapted to enduring long periods of fasting (based on their native ecology) by massively upregulating and downregulating their metabolism and their organ size and function between meals to conserve energy during long fasts associated with their native ecology (Secor, 2008; Secor & Diamond, 1995, 1998). For example, the python heart, intestine, liver, and kidneys can increase 40-100% in mass, and their metabolism can increase up to 40-fold, all within 48 hours of feeding (Secor, 2008; Secor & Diamond, 1995, 1998). Accordingly, Burmese pythons are presumably poorly adapted to the year-round prey availability typical in South Florida. However, the establishment and expansion of invasive Florida pythons has coincided with dramatic reductions in small mammal populations (Dorcas et al., 2012), indicating a potential ecological shift due to more consistent prey availability in comparison to monsoonal Southeast Asia. The expansion of this population in an ecosystem so different from its native range therefore raises the question of whether rapid evolution and adaptation may have played a role in the success of this invasive species.

Given the success and rapid proliferation of the invasive Burmese python population, especially in the face of strong ecological selection pressures, we were interested to test for evidence of rapid evolution (i.e., allele frequency fluctuations) and selection-driven adaptation. Further, we were motivated to determine if putatively selected genomic loci are associated with physiological traits linked to the novel climatic and ecological pressures present in Florida. To address these aims, we collected and analyzed multiple complementary datasets, including ecological, genomic, transcriptomic, and morphological data, and integrated the results of genomic scans, differential expression analysis, and histological analyses to test for corroborative evidence of rapid adaptation in the invasive Florida Burmese python population.

MATERIALS & METHODS

Overview of sample collection

Ninety-seven Burmese python (*Python molurus bivittatus*) samples were collected from South Florida as part of ongoing conservation efforts by state and federal agencies. This study was carried out in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health and the Animal Welfare Act. The protocol was approved by the US Geological Survey, Wetland and Aquatic Research Center Institutional Animal Care and Use Committee (IACUC; Permit Number: USGS/SESC 2013–04). Additionally, samples were collected under the National Park Service (NPS; Everglades) Permits EVER-2007-SCI-001 and EVER-2009-SCI-001. Samples collected by researchers at the University of Florida were also collected under an approved IACUC protocol (Study #201408432). All efforts were made to minimize distress during handling and no snakes were euthanized for the purposes of the study. These samples were obtained during two general time periods: (1) N = 48 samples from 19 May, 2003 to 17 June, 2009 and (2) N = 49 samples from 30 October, 2012 to 6 December, 2013 (Supplementary Figure 1). Most sampling was separated by only seven years. These time points are on both sides of an extreme freeze event that occurred in January 2010, and we refer to them as pre-freeze and post-freeze, respectively. Supplementary Table 1 contains complete information for all samples used in this study.

Estimates of habitat suitability in the United States

We used ecological niche modeling (ENM) to reconstruct the climatic niche of the Burmese python based on climatic variables associated with its native range and to project the suitable invasive range in the United States (Elith et al., 2006). For occurrence data, we used a total of 90 georeferenced localities throughout the species native range. The climatic niche was derived from 11 bioclimatic variables (Bio2, Bio3, Bio7, Bio10, Bio11, Bio14, Bio15, Bio16, Bio17, Bio18, Bio19) from the WorldClim dataset v. 1.4 (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005) with resolution of 30 seconds (~1km). These 11 variables minimized the amount of redundant climatic information due to correlation, and were selected from the original set of 19 bioclimatic variables after performing pairwise correlation tests between variables and identifying variables that exhibited a correlation coefficient of 0.8 or greater. From pairs of highly correlated variables, we selected seasonal variables over monthly and yearly averages. Climatic niche models were constructed using the program MAXENT v. 3.3.3k (Phillips, Anderson, & Schapire, 2006). We used the default parameters in MAXENT: 500 maximum iterations, convergence threshold of 0.00001, regularization multiplier of 1, and 10 000 background points. We first ran five model replicates using climatic layers spanning the entire world. We also ran a second set of 10 models, where we constrained climatic layers to the areas of interest (i.e., southern half of Asia and southeastern USA). The two sets of models were very similar and therefore we only present the average model from the first set of models. We visualized this model in ArcGIS v. 10.3 using three logistic probability thresholds: (1) minimum training presence threshold (i.e., the lowest logistic probability inferred in the native range); (2) equal training sensitivity and specificity logistic threshold; and (3) 10th percentile training presence logistic threshold (90% of samples in the native range have a logistic probability equal or higher than this threshold).

RADseq library generation and sequencing

We isolated total genomic DNA from tissue following the manufacturer’s protocol for the DNeasy Extraction Kit (Qiagen Inc.) or using Phenol:Chloroform:Isoamyl Alcohol extractions. All extractions were quantified using Qubit broad-range DNA assays (Thermo Fisher Scientific) following the manufacturer’s instructions. For samples that contained amounts of DNA too low to be confidently used for preparing restriction-site associated DNA sequencing (RADseq) libraries, we performed whole-genome amplification (WGA) using Phi29 DNA polymerase and a random 10mer primer (5’ – NNNNNNNNNN – 3’) using a GenomiPhi kit (GE Life Sciences).

Previous work has confirmed that WGA does not preferentially amplify certain genomic regions over others (Blair, Campbell, & Yoder, 2015), and thus does not bias population genetic inference.

We used a modified version of the Peterson *et al*. (2012) protocol to prepare double digest RADseq libraries for the 48 pre-freeze and 49 post-freeze samples targeting approximately 20 000 loci throughout the genome. Genomic DNA was digested simultaneously with rare (*Sbf*I; 8bp) and common (*Sau*3AI; 4bp) cutting restriction enzymes. To allow for hierarchical pooling and multiplexing of samples, barcoded Illumina adapter oligonucleotides were ligated to the ends of digested DNA. Following adapter ligation, samples were pooled in sets of 8, and these pools were size selected for a range of 430-600 bp using a Blue Pippin (Sage Science). After size selection, samples were PCR-amplified with pool-specific indexed primers, and amplification products were further pooled into a single sample based on molarity calculations from analysis on a Bioanalyzer (Agilent Technologies) using a DNA 7500 chip. The final pooled library was sequenced using 100 bp paired-end reads on an Illumina HiSeq 2000 lane.

Read processing and genotyping

Raw Illumina sequence data were filtered to remove PCR clones using the clone_filter tool from the Stacks v. 1.35 analysis pipeline (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011; Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013). Samples were parsed using the process_radtags tool from Stacks, using the rescue feature to keep reads with restriction sites or barcodes that are separated by two or less nucleotides from expected sequences, and Trimmomatic v. 0.33 (Bolger, Lohse, & Usadel, 2014) was used to quality filter the resulting data LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:36. We used the

RADcap (Hoffberg et al., 2016) software pipeline to map reads and infer genotypes based on the Genome Analysis Toolkit (GATK) best-practices guidelines (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013). We mapped the quality-trimmed reads to the Burmese python genome (Castoe et al., 2013) using the mem algorithm in BWA v. 0.7.12-r1039 (Li & Durbin, 2009) with default mapping settings and shorter split hits marked as secondary (i.e., '-M' option). SAMtools v. 1.2 (Li, 2011; Li et al., 2009) and Picard v. 1.106 were used to process mapping files for each sample and merge mappings for downstream analyses. GATK was used to perform realignment around indels, with a minimum number of reads at a locus of 4 and a minimum LOD score of 3. Single nucleotide polymorphisms (SNPs) and indels were called separately using the UnifiedGenotyper tool in GATK, with standard minimum confidence for variant calling set to 30 for both SNPs and indels and standard minimum confidence for variant emitting set to 30 for SNPs and 10 for indels. Only SNPs were kept for subsequent analyses, and were filtered using GATK as follows: (1) SNPs within 5bp of indels were excluded; (2) clustered SNPs within a 10bp window were excluded; (3) SNPs with at least four reads with a zero mapping quality (MQ0 >= 4) or a proportion of greater than 0.10 reads with zero mapping quality (MQ0/DP > 0.10) were excluded; and (4) SNPs with an overall quality score below 30 (QUAL < 30), a quality-by-depth score below 2 (QD < 2), a read depth below 5 (DP < 5), and a genotype quality score below 20 (GQ < 20) were excluded. We used VCFtools v. 0.1.14 (Danecek et al., 2011) to subsequently filter away singleton SNPs (due to high probability of sequencing or genotyping error) and to exclude SNPs that were missing in more than half the samples in both the pre-freeze and post-freeze populations. This filtering resulted in a dataset containing 1 021 variants and custom Python and R scripts were used to format the dataset for downstream analyses.

Genotyping transcriptomic data

To better understand the amount of standing genetic variation that could be directly acted upon by selection in the invasive Burmese python population, we examined coding variation using mRNAseq data from 10 samples collected in South Florida in January of 2016 (see Supplementary Table 1 for details of sampling). Cross sections of small intestine were preserved in RNAlater and stored at -80°C. Total RNA was extracted using Trizol Reagent (Invitrogen) and Illumina mRNAseq libraries were constructed using an Illumina TruSeq RNAseq kit with poly-A selection, RNA fragmentation, cDNA synthesis, and adapter ligation. We quantified completed libraries using a BioAnalyzer (Agilent Technologies), pooled libraries in equal molar ratios, and sequenced the combined library using a single lane on an Illumina HiSeq2000.

We used the GATK Best-Practices guidelines for genotyping RNAseq data to analyze these data. Raw RNAseq reads were quality filtered with Trimmomatic using the same parameters as above. Briefly, two rounds of mapping (i.e., two-pass methodology) to the Burmese python genome were performed using the splice-aware RNAseq mapper STAR 2.5.2b (Dobin et al., 2013), followed by mapping quality control using SAMtools and Picard and variant calling using HaplotypeCaller. We filtered variants using BCFtools and the following conditions: (1) SNPs within 3bp of indels were excluded; (2) clustered indels within a 10bp window were excluded; (3) variants not passing a set of standard hard filters (see http://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set) were excluded; and (4) variants with an overall quality score below 30 (QUAL < 30), a per sample read depth below 5 (DP < 5), a total read depth above 1 000 or below 200, and with greater than two alleles were excluded. We used custom scripts to quantify the degree of genetic variation present in transcripts, including synonymous and nonsynonymous polymorphisms, for the Florida population in 2016.

Analyses of population structure

Given that the invasive Burmese python population was established from pet-trade snakes originating from various native range regions and populations, we were interested in using genetic data to estimate how many source populations comprise the invasive population. We used the LEA package (v. 1.0; Frichot, François, & O'Meara, 2015; Frichot, Mathieu, Trouillon, Bouchard, & François, 2014) in the R statistical environment (v. 3.3.1; R Core Team, 2017) to estimate the number of ancestral populations, commonly referred to as *K*, which in this case should correspond to the number of source populations given the relatively recent introduction of pythons to Florida. This analysis was conducted using the non-negative fractorization algorithm (snmf function), with 10 replicates for each *K* value between one and 10. We used the cross-entropy criterion to determine the value of *K* most supported by the genotype data and visualized the resulting admixture or ancestry coefficients with so-called "Structure" plots.

Inferring and visualizing the between-time site frequency spectrum

The two-dimensional site frequency spectrum (2D-SFS) offers an intuitive way of visualizing the density in minor allele frequencies and how they shift between pre-freeze and post-freeze populations. We used δaδi (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009) to calculate and visualize the folded 2D-SFS after projecting down to a sample size of 45 for each population. We created two 2D-SFS matrices by inverting the placement of each population time point site frequency spectrum on the x- or y-axis, effectively creating two transposed 2D-SFS, which we used to calculate linear Poisson residuals between the time points and visualize the change in the 2D-SFS between pre-freeze and post-freeze populations.

Multivariate scans for signatures of selection

We used a custom Python script to calculate six metrics to evaluate the degree of allele frequency fluctuation between the pre-freeze and post-freeze populations: (1) the absolute value of allele frequency change |ΔAF|; (2) population allelic differentiation based upon Weir & Cockerham (1984; F_{ST}); (3) the absolute differentiation in genetic diversity (D_{XY}); (4) the fluctuation in nucleotide diversity (pre-freeze – post-freeze; ΔPi); (5) the difference in heterozygosity between populations (pre-freeze – post-freeze; ΔHet) and (6) the fluctuation in Tajima's D statistic (pre-freeze – post-freeze; ΔTajD; Tajima, 1989)). Combining information from two or more summary statistics provides increased power to detect loci under natural selection (Evangelou & Ioannidis, 2013; François, Martins, Caye, & Schoville, 2016; Grossman et al., 2010; Lotterhos et al., 2017; Ma et al., 2015; Imtiaz A. S. Randhawa, Khatkar, Thomson, & Raadsma, 2015; Imtiaz Ahmed Sajid Randhawa, Khatkar, Thomson, & Raadsma, 2014; Utsunomiya et al., 2013), and thus we employed a multivariate outlier approach to identify genetic loci with strong signatures of natural selection based on the six univariate statistics. MINOTAUR (Verity et al., 2017) was used to estimate the Mahalanobis multivariate distance (Mahalanobis, 1936) based on the six statistics. Since individual univariate statistics are likely to be correlated, we used a covariance matrix to correct distances for these interactions. The top 2.5% of Mahalanobis distance measures were taken to indicate variants under putative selection, and this threshold reflected a natural break in the Mahalanobis multivariate distribution and in associated bivariate plots between pairwise univariate statistics (Supplementary Figure 4).

Inferring selection coefficients from temporal population genetic fluctuations

We used ApproxWF (Ferrer-Admetlla, Leuenberger, Jensen, & Wegmann, 2016), to estimate selection coefficients (*s*) based on the fluctuation in allele frequencies between the two sampled time points. For each variant, allele frequencies for each time point were extracted from the VCF and run through the ApproxWF MCMC for 1 000 000 million iterations, with posterior estimates sampled every 10 iterations. We confirmed proper MCMC parameter value mixing using R and plotted the median posterior estimate of *s* for each variant. Variants with a 95% high posterior density of *s* greater than or less than 0 (no selection) were inferred to be under selection. There was a high degree of correspondence between putatively selected variants between the multivariate outlier analysis and the estimated selection coefficients (Figure 2).

Permutation and simulations tests for rejecting neutral evolution

We used permutation tests to test the null hypothesis that the allele frequency estimates are not significantly different between the two populations. For each permutation, we used the observed genotypes for all 97 samples and randomly assigned each (without replacement) to the two population groups with sample sizes equal to our empirical sampling (N = 48 and N = 49 samples in pre-freeze and post-freeze populations, respectively). We then calculated the absolute value of the allele frequency difference between the population groups for each permutated dataset. We ran 1 000 permutations and compared the results of this distribution to our empirical measures of allele frequency change.

We also conducted forward-time simulations to more directly address whether fluctuations in population genetic statistics were beyond what would be expected under genetic drift. Our simulations were performed as follows: (1) We approximated the starting “sample” allele frequencies by sampling (with replacement) from the distribution of observed empirical frequencies for each of the 812 unlinked biallelic variants. (2) We simulated ending population frequencies (i.e., post-freeze) under the Wright-Fisher model of genetic drift (again, sampling with replacement) for each variant while using the observed "sample" allele frequencies to approximate the starting (i.e., pre-freeze) population frequencies. (3) Finally, we sampled from these simulated ending population frequencies to obtain a set of ending "sample" allele frequencies. We used the empirical locus lengths and sample sizes (i.e., the number of sampled genotypes) from each of the 812 variants in our simulations (empirical sample sizes we used for both the starting and ending "sample" frequencies). We used a conservative minimum effective population of 500 individuals, given that this number of snakes has been captured in 2009 (the year before the freeze event) was 496. To account for higher effective population sizes, we also ran additional simulations with values of 1 000, 10 000, and 100 000. We also varied the number of generations between the two sampling points from either one, two, or three generations, which encompass the full range of generations possible within this time period (i.e., average generation time for pythons three years; Willson et al., 2011). We simulated a total of 1 000 replicate datasets under each combination of effective population size and number of generations, and used these datasets to obtain a null distribution of allele frequency change (i.e., simulated under drift alone) to compare with our empirical observations. Overall, we found that effective population size and the number of generations of drift had little qualitative impact on the results. For further analyses we used simulations that conservatively assumed an effective population size of 500 individuals and 2 generations of genetic drift between the pre- and post-freeze population samples. Higher densities of loci in the empirical dataset versus the simulated datasets at more extreme values of population genetic statistics provide evidence for selection.

Using synteny with the Boa constrictor *genome to identify genomic regions in the python*

The Burmese python reference genome suffers from relatively low contiguity, and some genome scan outliers were on small scaffolds or were located near the ends of scaffolds. To attempt to overcome this shortcoming and enable more meaningful analyses of the genomic context of genome scan outliers, we used the highly contiguous *Boa constrictor* reference genome (Bradnam et al., 2013) to arrange and orient Burmese python genome scaffolds. We used the Chromosemble tool from Satsuma (Grabherr et al., 2010), with default parameters, to map the Burmese python genome scaffolds onto the *Boa constrictor* genome. We filtered the resulting alignments to identify mapping anchors of perfect alignment that were 25 bp or greater and inferred homology and Burmese python scaffold placement in cases where 10 or more anchors were present with logical spacing and orientation. In many cases, this allowed us to manually expand the genomic regions surrounding genome scan outliers.

Estimating gene expression for Florida pythons

Given findings from the selection scans, we were interested in comparing patterns of small intestine gene expression from invasive pythons in Florida with previous estimates of expression patterns from controlled experiments involving commercial trade pythons (Andrew et al., 2015, 2017; Castoe et al., 2013). These experiments leveraged replicate sampling of captive Burmese pythons taken at the following controlled time points: fasted (30 days since last meal), 1-day post-feeding (1DPF), 4DPF, and 10DPF. We downloaded the raw read data from these previous studies from the NCBI SRA database. These data were combined with newly-generated small intestine RNAseq data from a subset of seven pythons from the invasive Florida population collected in 2016 (also discussed above when quantifying coding variation). Due to permitting constraints, we were unable to carry out a well-controlled experiment akin to that presented in previous studies (Andrew et al., 2015, 2017; Castoe et al., 2013). However, we were able to leverage the known and well-defined, cyclical pattern of digestive physiology and gene expression following feeding to ensure that snakes were strategically fasted prior to sacrifice. Burmese pythons reach their peak digestive physiological state at 1-2 days post feeding and by four days they are starting to revert to a fasted state. Therefore, snakes that contained no meal item in the gut and that were in captivity without access to food for at least eight days were used for this experiment. We found that these expectations were upheld, as overall gene expression in these seven pythons closely resembled a fasted state in animals from previous well-controlled experiments, and we feel that this design is justified for roughly deciphering the digestive tract physiology in fasted modern Florida Burmese pythons. Further information about how these samples were collected and how the RNAseq data was generated are described in the "Genotyping transcriptomic data" section above. Raw RNAseq reads were quality filtered using Trimmomatic and mapped to the annotated transcript set of the Burmese python genome using bwa mem (as outlined in our analysis of the RADseq dataset). Raw expression counts for each reference transcript were normalized alongside existing counts from small intestine experiments (Andrew et al., 2015, 2017; Castoe et al., 2013) using the TMM method (Robinson & Oshlack, 2010) in edgeR (McCarthy, Chen, & Smyth, 2012; Robinson, McCarthy, & Smyth, 2010). We estimated significant changes in gene expression between the fasted invasive python sampling and each of these experimental time points using pairwise exact tests calculated in edgeR with subsequent independent hypothesis weighting (IHW), which used weighted Benjamini and Hochberg procedure to limit the false discovery rate (FDR; Ignatiadis, Klaus, Zaugg, & Huber, 2016).

To facilitate the analysis of gene expression in the context of canonical pathways, GO terms, and mouse knockout phenotypes, we used reciprocal and one-way best BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) searches against the *Anolis*, Chicken, and Human gene sets to infer orthology in cases where gene symbols were not available from the NCBI annotation of the Burmese python genome. Gene symbol identifiers were successfully assigned to 21 450 of 26 853 python transcripts. Genes identified as significantly differentially expressed (IHW FDR < 0.1) in pairwise comparisons between fasted Florida and fasted experimental animals were analyzed using Core Analysis in Ingenuity Pathway Analysis (IPA; Qiagen) to infer differential activity of canonical pathways and upstream regulatory interactions. Annotated genes located on scaffolds that contained putative targets of selection were analyzed for GO term and KEGG Pathway enrichment using the Web-based Gene Set Analysis Toolkit (WebGestalt 2017; Zhang, Kirov, & Snoddy, 2005) and using ClueGo v. 2.2.6 (Bindea et al., 2009) implemented in Cytoscape v. 3.3.0 (Shannon et al., 2003), with ontologies/pathways from GO, KEGG, and WikiPathways, a GO Tree Interval of 3 to 15, GO Term Fusion enabled, a Kappa score of 0.5, and Benjamani-Hochberg p-value correction. We also evaluated mouse knockout phenotype enrichment using Mammalian Phenotype Enrichment Analysis (MamPhEA; Weng & Liao, 2010), with manual *post hoc* clustering of similar phenotypes, which were visualized using a wordcloud constructed using the wordcloud2 v. 0.2.0 package in R. We used the GenometriCorr R package (Favorov et al., 2012) to test for spatial autocorrelation between differentially expressed transcripts and genome scan outliers using the Jaccard index of overlap. Previous studies describe the physiological and gene expression changes that underlie regenerative organ growth, and the experimental design for the study that originally derived the comparative RNAseq data used in these analyses (Andrew et al., 2015, 2017).

Histological analyses of organ morphology

Burmese pythons experience extreme and rapid changes in the morphology of digestive organs when transitioning between a dormant fasted state and an actively digesting state, and we were interested in comparing the morphological state of samples from fasted invasive pythons from Florida to that of experimental animals in carefully controlled fasted and fed states. Cross-sections from the anterior third of the small intestine from the seven invasive python samples from Florida taken in January of 2016 were fixed in reptilian Ringer's-buffered 10% formalin solution. All samples were embedded in paraffin, cross-sectioned (6 μm), and stained with hematoxylin and eosin on glass slides. Existing paraffin blocks from 3 replicate animals each from controlled fasted, 3DPF, and 10DPF time points were also obtained, cross-section, and stained in the same manner. Samples were viewed with a Zeiss Axio Imager A1 light microscope linked to a computer with image analysis software Zen Imaging software. For each cross-section, we measured enterocyte height and width, and calculated enterocyte volume using the formula for a cylinder. Samples of heart, liver, and kidney tissue were also taken from the 10 invasive python samples and from 3 replicate samples from fasted, 3DPF, and 10DPF time points, and were prepared as above. For these three tissue types, we counted the number of visible nuclei per field of view at 10 random points in the section. These measurements were taken at a magnification of 630x using the software ImageJ2/Fiji (Schindelin et al., 2012; Schindelin, Rueden, Hiner, & Eliceiri, 2015). Nuclei per field negatively correlates with cell size and serves as a proxy for that measurement. Measurements of cell sizes for all four organs for the three experimental time points and for fasted samples from the invasive Florida population were compared using an ANOVA with *post-hoc* Tukey's Honest Significant Difference tests of pairwise comparisons, all implemented in R.

To enable transmission electron microscopy of intestinal microvilli, small samples of intestinal tissue were fixed in 2.5% glutaraldehyde. Samples were processed following previous work (Lignot, Helmstetter, & Secor, 2005), with post-fixation in 1% osmium tetroxide, dehydration in a graded series of ethanol, and Spurr resin embedding. Ultra-thin sections (ca. 90 nm) were placed on copper grids and stained with uranyl acetate and lead citrate. We used a Jeol 1200 EX electron microscope to examine the sections and photographed four to five areas of microvillus at a magnification X7,500. The lengths and widths of 5-10 microvilli were measured, selecting only those microvilli cut along the central plane of their long axis.

RESULTS AND DISCUSSION

Climatic and Feeding Data Indicate Ecological Shifts in Invasive Burmese Pythons

Burmese python physiology is highly adapted to monsoonal Southeast Asian ecosystems with major seasonal shifts in prey availability that lead to these snakes enduring long periods of fasting (Secor, 2008; Secor & Diamond, 1995, 1998). In Florida, however, invasive pythons have constant access to prey and thus feed year-round (Dorcas et al., 2012). Our analyses, based on five years (2003 – 2008) of necropsy data from the Florida population of Burmese pythons (Florida pythons, hereafter), indicated that an annual average of 94% of captured snakes contained a meal (97% in wet season and 91% in dry season; Figure 1A). These data indicate that Florida pythons are constantly feeding year-round, which represents a major ecological shift from the "feast-famine" feeding patterns associated with their native range.

We also found that invasive Florida pythons experience climatic conditions that are distinct from their native range. Our ecological niche models agreed with previous estimates (Pyron et al., 2008) that this population persists at the margin of the predicted climatic suitability of this species (Figure 1B). Further supporting this inference, Burmese pythons exhibit high mortality (50% or higher) when relocated to more temperate U.S. locations (Avery et al., 2010; Dorcas et al., 2011) and during freeze events in South Florida (Mazzotti et al., 2011). We hypothesized that these novel ecological factors – more extreme cold climatic events and consistent prey availability – have acted as strong selective catalysts to drive the evolution of Florida pythons.

Genomic Evidence for Rapid Evolution Driven by Natural Selection in Florida Pythons

To test for evidence of evolution and selection on genetic variation through time, we generated genomic data (using ddRADseq; Peterson et al., 2012) from Florida pythons collected before and after a January 2010 freeze event that occurred in South Florida, which is known to have caused high python mortality (50-90%; see Supplementary Figure 1 for temporal ranges of sampling; Mazzotti et al., 2011). Genomic variation measured at 23,041 nuclear loci from 97 Florida pythons (48 sampled before and 49 after the freeze event; Figure 1B; Supplementary Figure 1; Supplementary Table 1) indicates that the Florida population was likely derived from 2-3 distinct source populations (Figure 2A). These findings align with importation records indicating three main native range sources (Engeman et al., 2011) and expand on previous microsatellite analyses (Collins et al., 2008) to provide genomic evidence for a panmictic invasive population. Among all samples, we found that 3.5% of sequenced loci contained multiple alleles in the Florida population. To assess genetic variation more likely to be phenotypically relevant, we estimated variants in exons of 3,664 expressed intestinal genes (18% of annotated python genes) using 10 Florida Burmese pythons collected in 2016. This analysis identified 2,197 total variants, including 638 nonsynonymous variants (~29%), suggesting that despite a likely bottleneck during colonization, the Florida population contains substantial standing variation available for adaptation via natural selection to act upon.

To test whether selection on standing genetic variation is leading to temporal fluctuations in allele frequencies, we compared genomic variation between the pre- and post-freeze populations. Evidence for rapid genomic evolution and adaptation in the Florida python population through time were evident in the two-dimensional allele frequency spectrum and from genome-wide population genetic diversity statistics (Figure 2B; Supplementary Figure 2). The allelic fluctuation between our empirical pre- versus post-freeze Florida population samples was also significantly different than random population assignments of samples, indicating evolution across our temporal samples (Supplementary Figure 3). Using forward time simulations across a range of plausible demographic scenarios, we found that the largest empirical allele frequency changes are unlikely to have occurred due to neutral genetic drift alone (Supplementary Figures 4-5). Collectively, these results provide strong support for rapid evolution of the Florida python population, and the role of selection at a subset of genomic regions.

To further test for evidence of locus-specific signatures of selection, we used multiple genomic-scan approaches to survey our genome-wide variant dataset. We identified evidence of temporal genetic differentiation driven by selection at several loci by summarizing six population genetic statistics using a multivariate composite measure (Mahalanobis distance) that identified multivariate outliers (Figure 3A; Supplementary Figures 6-7; Lotterhos et al., 2017; Verity et al., 2017). We also conducted an independent estimation of locus-specific selection coefficients (*s*) based on temporal allele frequency changes (Ferrer-Admetlla et al., 2016) that identified many of these same genomic regions as evolving under directional selection (Figure 3B). These two approaches together implicated 12 candidate genomic regions as likely influenced by positive selection between pre- and post-freeze event pythons. Collectively, population genomic data provide compelling evidence for genome-wide evolution (i.e., allele frequency change) and for evolution driven by natural selection at a subset of genomic regions.

Rapid Adaptive Evolution Targeted Genes Related to Ecological Shifts

Given strong evidence for selection-driven evolution, we were motivated to identify the potential functional targets of selection. We used the Burmese python genome annotation to identify genes that are genetically linked to the 12 candidate genomic regions inferred to be under selection (Supplementary Figure 8; Castoe et al., 2013). We used alignments to the more contiguous *Boa constrictor* genome (Bradnam et al., 2013) to identify adjacent syntenic scaffolds in the python (see Supplemental Methods). 78 genes were identified within the 12 putatively selected genomic regions, and functional annotations for these genes demonstrated striking relevance to physiological features that were *a priori* predicted to be relevant to the novel ecological conditions of Florida. Analyses of associated Mouse Knockout (MKO) phenotypes identified four prominent clusters of phenotypes: sensory perception and responsiveness, thermosensation and hypothermia responsiveness, learning and behavior, and organ form and function (Figure 3C). Gene ontology (GO) analyses also indicated enrichment for genes linked to cell division, organ growth and development (including calcium signaling), reproduction, immunity and responses to stress, and to neuronal function and behavior (Figure 3D; Supplementary Figure 9).

Because Burmese pythons are known for their ability to undergo extreme organ growth upon feeding, we cross-referenced genes in these 12 regions with genes relevant to regenerative organ growth (Andrew et al., 2015, 2017) and found several genes in key organ growth regulation pathways. Multiple genes were involved in calcium-mediated signaling, which plays a central role in organ hypertrophy, including *PLEK*, *CHP2*, and, importantly, *PPP3R1*, which encodes a regulatory subunit of calcineurin (a key regulator of cardiac hypertrophy), and *PLCE1* (a regulator of processes including cell growth and differentiation). This gene set also included *PITX2*, a regulator of abdominal development, and a long non-coding RNA with homology to *PTEN* – a gene that functions in the mTOR growth pathway that is central to modulating post-feeding organ growth in pythons (Figure 4; Andrew et al., 2015, 2017). Overall, genomic data broadly correspond with ecological and climatic data in implicating strong selection on traits related thermal tolerance as well as feeding ecology/physiology. Furthermore, these multiple lines of evidence that implicate changes in feeding physiology raise the intriguing question of whether Florida pythons have adapted to alter their dynamic physiology to a more consistently active state based on increased prey availability in South Florida.

Histological and Functional Genomic Data Implicate Phenotypes Linked to Ecological Pressures and Correspond with Putative Genes Under Selection

We conducted a second set of experiments to identify whether any evidence outside of genomic allele frequency changes might corroborate (or refute) our inference that rapid adaptation has occurred that may have altered physiological regulation in the invasive Florida pythons. We tested for evidence that modern Florida pythons possess a more up-regulated fasting physiological state by comparing gene expression and histological data on organ cell sizes between fasted Florida pythons captured in 2016 and captive bred laboratory descendants of imported pythons while fasting and at various post-feeding time points. Specifically, we tested if fasted post-freeze Florida pythons had substantially different cellular and transcriptomic states compared to fasted laboratory pythons – a pattern that is predicted by our inferences from genomic and ecological data. While we acknowledge that this experiment was not ideally controlled (e.g., common garden design) due to permitting and regulatory constraints, it did allow us to test for phenotypic evidence that might support the hypothesis of a shift in Florida python physiology. We found that patterns of gene expression in seven fasted post-freeze Florida pythons resembled fasted laboratory pythons (Supplementary Figure 10), yet were distinct in several key features. Importantly, when comparing fasted post-freeze pythons and laboratory pythons, we found that an excess of differentially expressed genes were located in putatively selected genomic regions based on our analyses of population genomic data ($p < 0.01$). Additionally, five of the six genes identified by our analyses of allele frequency changes, and highlighted above as being important in organ growth, showed differentially expressed transcripts between fasted post-freeze pythons and laboratory pythons (IHW FDR < 0.1). We also found that gene expression interpreted in the context of pathways known to mediate regenerative growth in Burmese pythons (Andrew et al., 2015, 2017) indicates that fasted post-freeze pythons exhibit pathway and upstream pathway regulatory molecule states that are intermediate between the fasted and fed states in laboratory pythons (Figure 5A). Lastly, we examined cell sizes from four organs in fasted post-freeze pythons and found that they more closely resemble actively digesting laboratory pythons more so than fasted laboratory pythons (Figure 5B-C). While the transcriptome and histological data alone do not provide definitive proof of adaptation, it is notable that transcriptome and histological results support the independent predictions from population genomic and ecological data – that post-freeze Florida pythons may exhibit a unique and more consistently upregulated physiology.

Conclusions and Synthesis

Overall, our results provide evidence for rapid evolution by natural selection in invasive Florida pythons, together with multiple lines of evidence that adaptation may be linked to freeze-tolerance and a shift in feeding physiology. Our ecological data provide compelling evidence for a massive shift in feeding ecology occurring in invasive Burmese pythons since their introduction to Florida, and field mortality estimates together with our ecological niche models indicate Florida pythons exist at the margins of their thermal tolerance. Our genomic data demonstrate that evolution (allele frequency change through time) has occurred, and that a subset of genomic regions exhibit hallmarks of natural selection. Interestingly, these regions are enriched for genes related to thermal tolerance, behavior, and physiological phenotypes. Finally, independent gene expression and histological data provide an intriguing added layer of support for a shift in Florida python feeding physiology, which implicates many of the same key genes identified by the population genomic data. These results collectively support the hypothesis that new ecological pressures in Florida, such as a more temperate climate and more consistent prey availability, have driven adaptation by favoring the maintenance of a physiologically active state and enhanced thermoregulatory responsiveness.

A compelling question remains of whether behavioral changes, thermal tolerance, and shifts in digestion physiology are linked, and future *in situ* and common-garden experiments would be valuable to test for these phenotypic differences and discern connections between these putative adaptations. Moreover, the relative contributions of longer term, consistent selection pressures versus acute, strong natural selection (e.g., rare freeze events) remain unclear from our analyses. It is possible that rapid adaptation in invasive Florida Burmese pythons may be the result of synergistic interactions between consistent ecological pressures, such as shifts in food availability, and acute climatic pressures associated with periodic freeze events. Fasting laboratory Burmese pythons have among the lowest vertebrate basal metabolic rates, yet upon feeding experience extreme organ growth that coincides with the highest increase in metabolic rate in vertebrates (40-fold) (Secor, 2008; Secor & Diamond, 1995, 1998). Positron emission tomography (PET) scans of fasted versus fed laboratory pythons highlight that the hypermetabolic state of pythons with up-regulated organ systems results in increased body temperature (Secor, 2008; Secor & Diamond, 1995, 1998), which would make physiologically up-regulated, hypermetabolic pythons – due either to having recently fed or due to heritable variation in their degree of post-feeding downregulation – resistant to freezing, and may explain how the high mortality 2010 freeze event could have catalyzed adaptive evolution. *In situ* adaptation of Burmese pythons to the South Florida environment has broad ecosystem-scale ramifications for persistence and expansion of this impactful invasive species. This and other examples (Phillips, Brown, Webb, & Shine, 2006) also demonstrate the surprising evolutionary potential of invasive species, and the importance of accounting for adaptation in predicting the outcomes of biological invasions.

ACKNOWLEDGMENTS

This project was supported by UTA startup funds to T.A.C., an NSF DDIG award to D.C.C. and T.A.C. (DEB-1501747), and a Society for the Study of Evolution award to D.C.C. Stephen Secor provided histological samples for comparative purposes, and James Kennedy and Heather Perry assisted with microscopy. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the United States Government.

DATA AVAILABILITY

Raw Illumina sequencing data have been accessioned at the NCBI SRA. See Supplementary Table 1 for accession information.

AUTHOR CONTRIBUTIONS

D.C.C. and T.A.C. designed the experiment. D.C.C., B.W.P., R.H.A., M.R.R., F.J.M., K.M.H., M.E.H., and T.A.C. were involved in sample acquisition. D.C.C., B.W.P., R.H.A., D.R.S., A.S.Y., A.L.A., T.J., G.I.M.P., N.R.H., M.R.W., and T.A.C. contributed to data analysis and interpretation. D.C.C. and T.A.C. wrote the manuscript. All authors read and approved the final manuscript.

FIGURES

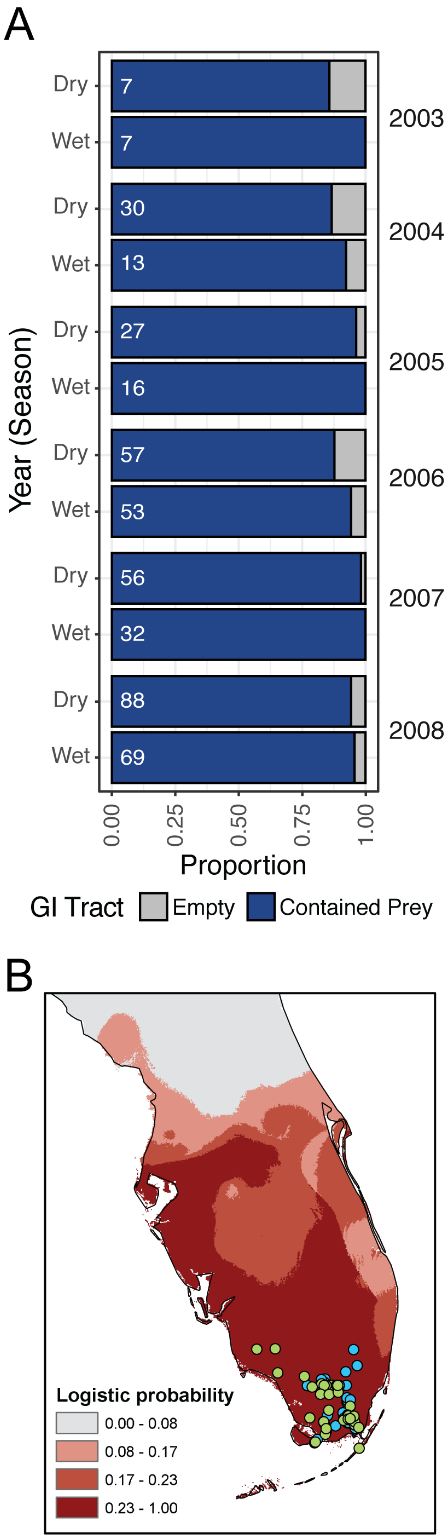

A
Year (Season)
Dry
Wet
2003
Dry
Wet
2004
Dry
Wet
2005
Dry
Wet
2006
Dry
Wet
2007
Dry
Wet
2008
7
7
30
13
27
16
57
53
56
32
88
69
0.00
0.25
0.50
0.75
1.00
Proportion
GI Tract
Empty
Contained Prey
B
Logistic probability
0.00 - 0.08
0.08 - 0.17
0.17 - 0.23
0.23 - 1.00

Figure 1. Evidence of novel ecological conditions for invasive Burmese pythons in South Florida. (**A**) Temporal analyses of the proportion of captured pythons containing a food item. White numbers within the bars indicate sample sizes. (**B**) A map of the sampling used for this work from pre-freeze (N = 48; green points) and post-freeze (N = 49; blue points) populations and habitat suitability estimates based on the ecological niche modeling of native-range Burmese pythons.

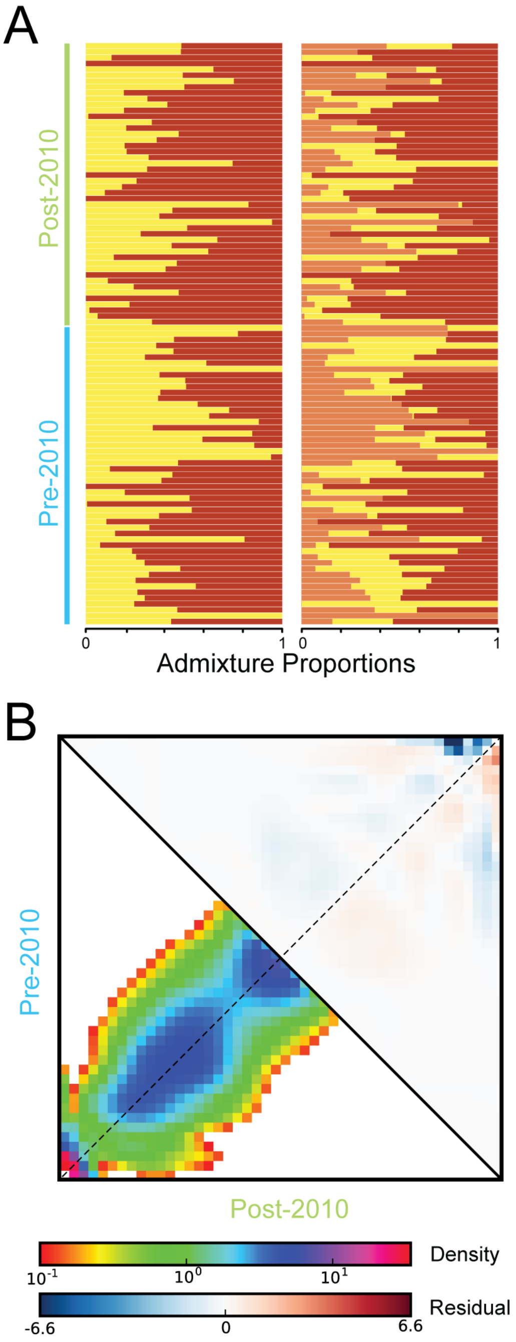


Figure 2. Genomic evidence for mixed ancestry and genetic evolution in the invasive Burmese python population. (**A**) Structure plot showing the admixture proportions for *K*=2 and *K*=3 source populations. (**B**) Allele frequency shifts in the Florida population illustrated by the 2D site allele frequency spectrum (below solid diagonal) and the residual change in allele spectrum density between the two time points (above solid diagonal). Each axis represents the distribution of minor allele frequencies for variant loci at the time point, which was projected down to a sample size of 45.

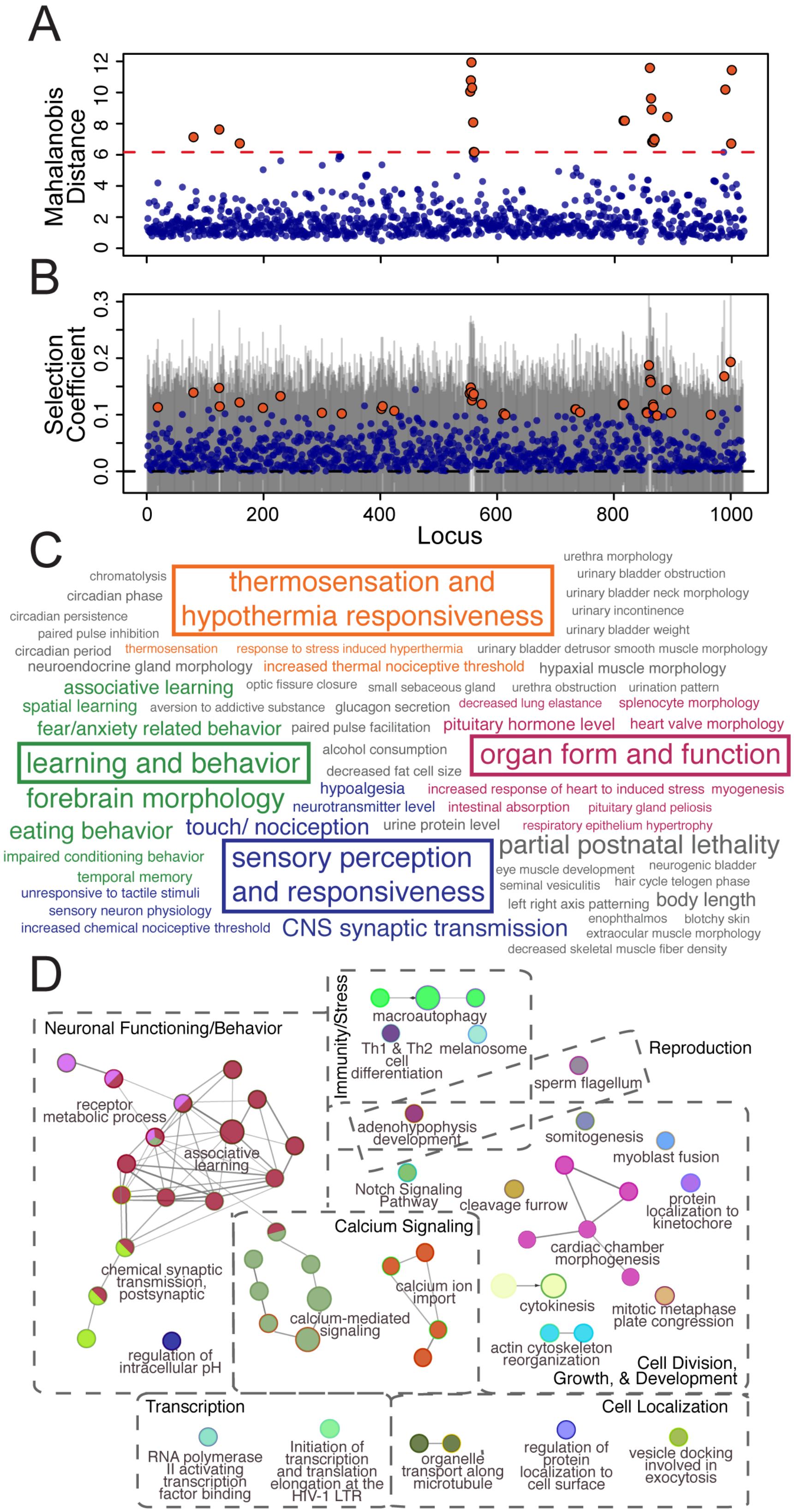

A
Mahalanobis Distance
0
2
4
6
8
10
12
B
Selection Coefficient
0.0
0.1
0.2
0.3
0
200
400
600
800
1000
Locus
C
chromatolysis
circadian phase
circadian persistence
paired pulse inhibition
circadian period
thermosensation and hypothermia responsiveness
thermosensation
response to stress induced hyperthermia
increased thermal nociceptive threshold
neuroendocrine gland morphology
urethra morphology
urinary bladder obstruction
urinary bladder neck morphology
urinary incontinence
urinary bladder weight
urinary bladder detrusor smooth muscle morphology
hypaxial muscle morphology
associative learning
optic fissure closure
small sebaceous gland
urethra obstruction
urination pattern
spatial learning
aversion to addictive substance
glucagon secretion
decreased lung elastance
splenocyte morphology
fear/anxiety related behavior
paired pulse facilitation
pituitary hormone level
heart valve morphology
learning and behavior
alcohol consumption
decreased fat cell size
organ form and function
hypoalgesia
increased response of heart to induced stress
myogenesis
forebrain morphology
neurotransmitter level
intestinal absorption
pituitary gland peliosis
eating behavior
touch/ nociception
urine protein level
respiratory epithelium hypertrophy
partial postnatal lethality
impaired conditioning behavior
sensory perception and responsiveness
eye muscle development
neurogenic bladder
temporal memory
seminal vesiculitis
hair cycle telogen phase
unresponsive to tactile stimuli
left right axis patterning
body length
sensory neuron physiology
enophthalmos
blotchy skin
increased chemical nociceptive threshold
CNS synaptic transmission
extraocular muscle morphology
decreased skeletal muscle fiber density
D
Neuronal Functioning/Behavior
Immunity/Stress
macroautophagy
Th1 & Th2 cell differentiation
melanosome
Reproduction
sperm flagellum
receptor metabolic process
associative learning
adenohypophysis development
somitogenesis
myoblast fusion
Notch Signaling Pathway
cleavage furrow
protein localization to kinetochore
Calcium Signaling
cardiac chamber morphogenesis
chemical synaptic transmission, postsynaptic
calcium ion import
calcium-mediated signaling
cytokinesis
mitotic metaphase plate congression
regulation of intracellular pH
actin cytoskeleton reorganization
Cell Division, Growth, & Development
Transcription
Cell Localization
RNA polymerase II activating transcription factor binding
Initiation of transcription and translation elongation at the HIV-1 LTR
organelle transport along microtubule
regulation of protein localization to cell surface
vesicle docking involved in exocytosis

Figure 3. Genome-wide shifts in population genetic variation indicate selection in genes related to reproduction, behavior, and regenerative organ growth. (A) Manhattan plot of multivariate Mahalanobis distance across variants with points above the 97.5% quantile (red broken line) indicated in red. **(B)** Manhattan plot of selection coefficients for each genome-wide variant. Gray lines represent the 95% high posterior density for each point (truncated at 0 for visualization). Red points have a 95% high posterior density (HPD) that falls entirely above 0, indicating selection. **(C)** Word cloud of MKO phenotypes associated with genes in regions with genomic outliers, clustered by color into broader physiological categories. **(D)** Enriched GO networks differentiated by color and clustered into broader physiological categories.

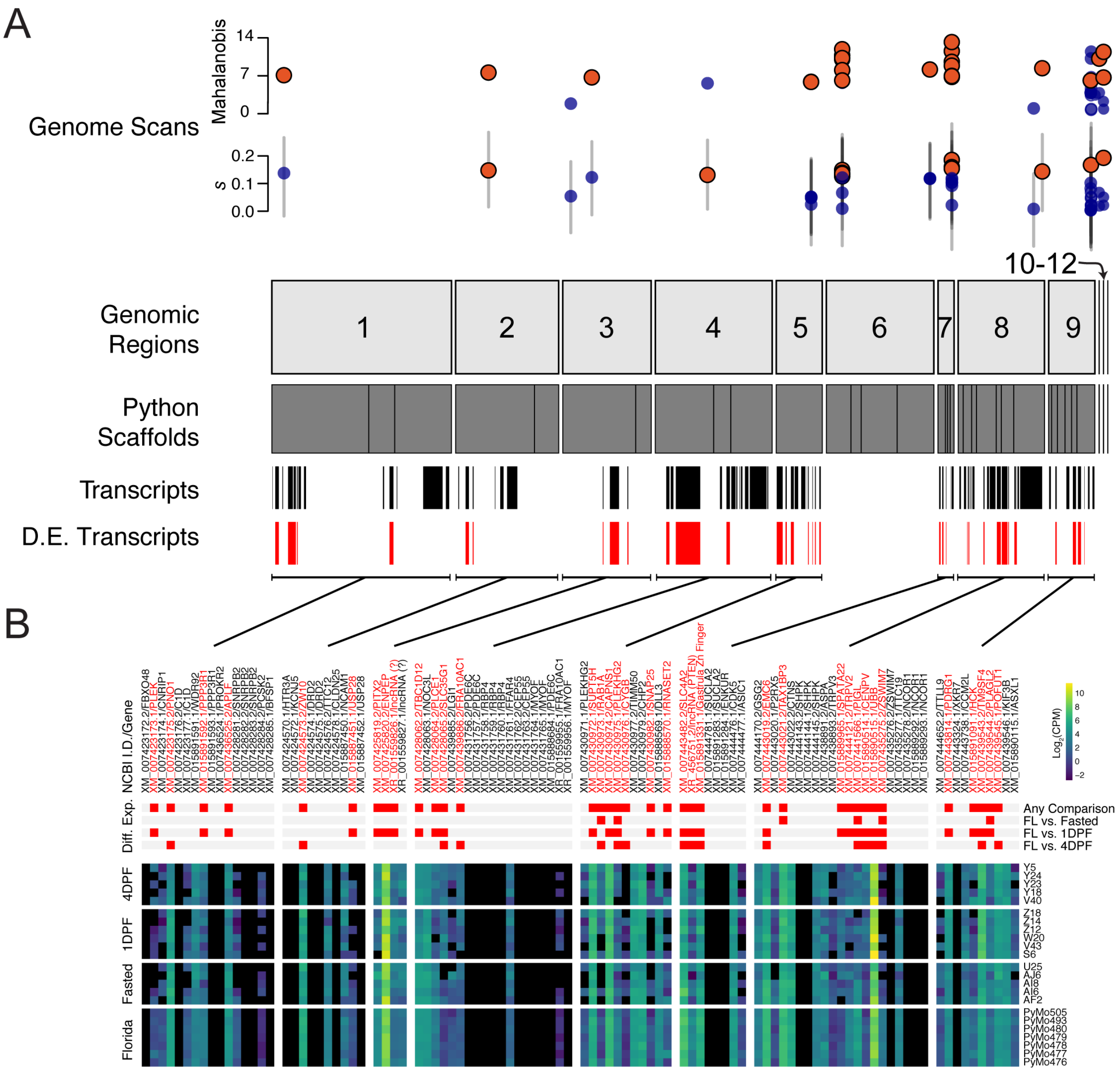


Figure 4. Natural selection modulates differential expression and is associated with genes known to play a role in regenerative organ growth. (A) Broader context of Mahalanobis distances and selection coefficients in syntenic genomic regions with selected variants, and associated Burmese python genome scaffolds, annotated transcripts, and significantly differentially expressed transcripts. Region 6 contained no annotated genes. **(B)** Significant pairwise differential expression comparisons (red in top heatmap and red transcript labels) and normalized expression heatmap for fasted post-freeze Florida pythons and laboratory pythons in fasted and post-feeding morphological states. FL = invasive Florida python; lncRNA (?) = long non-coding RNA with unknown homology; DPF = days post fed; CPM = counts per million; D.E. = differentially expressed.

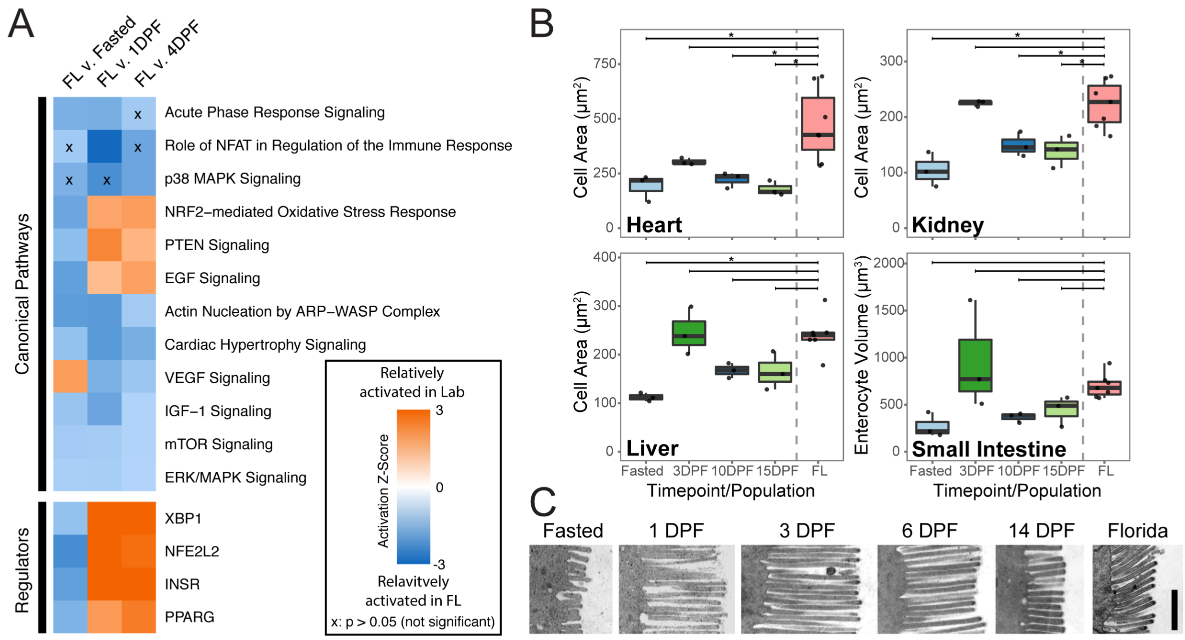
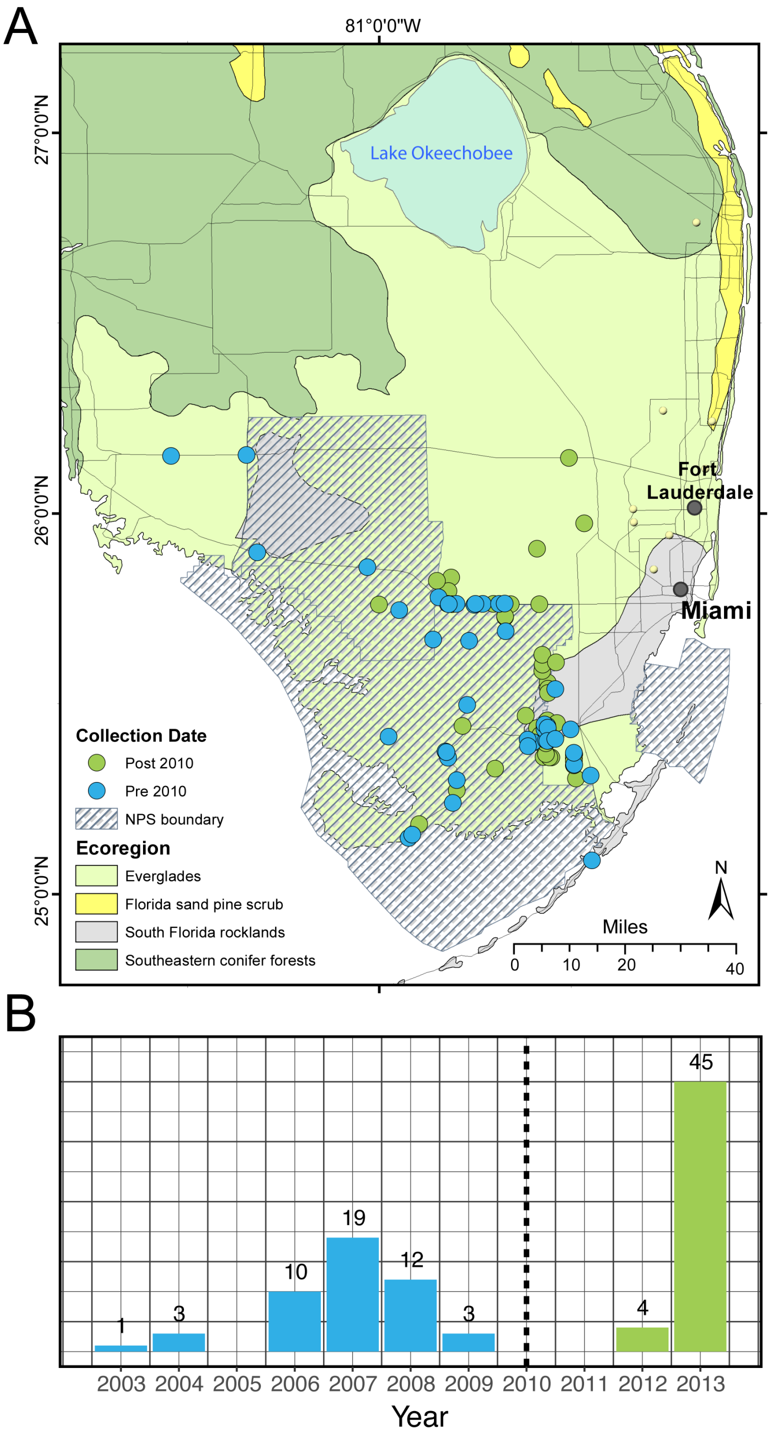
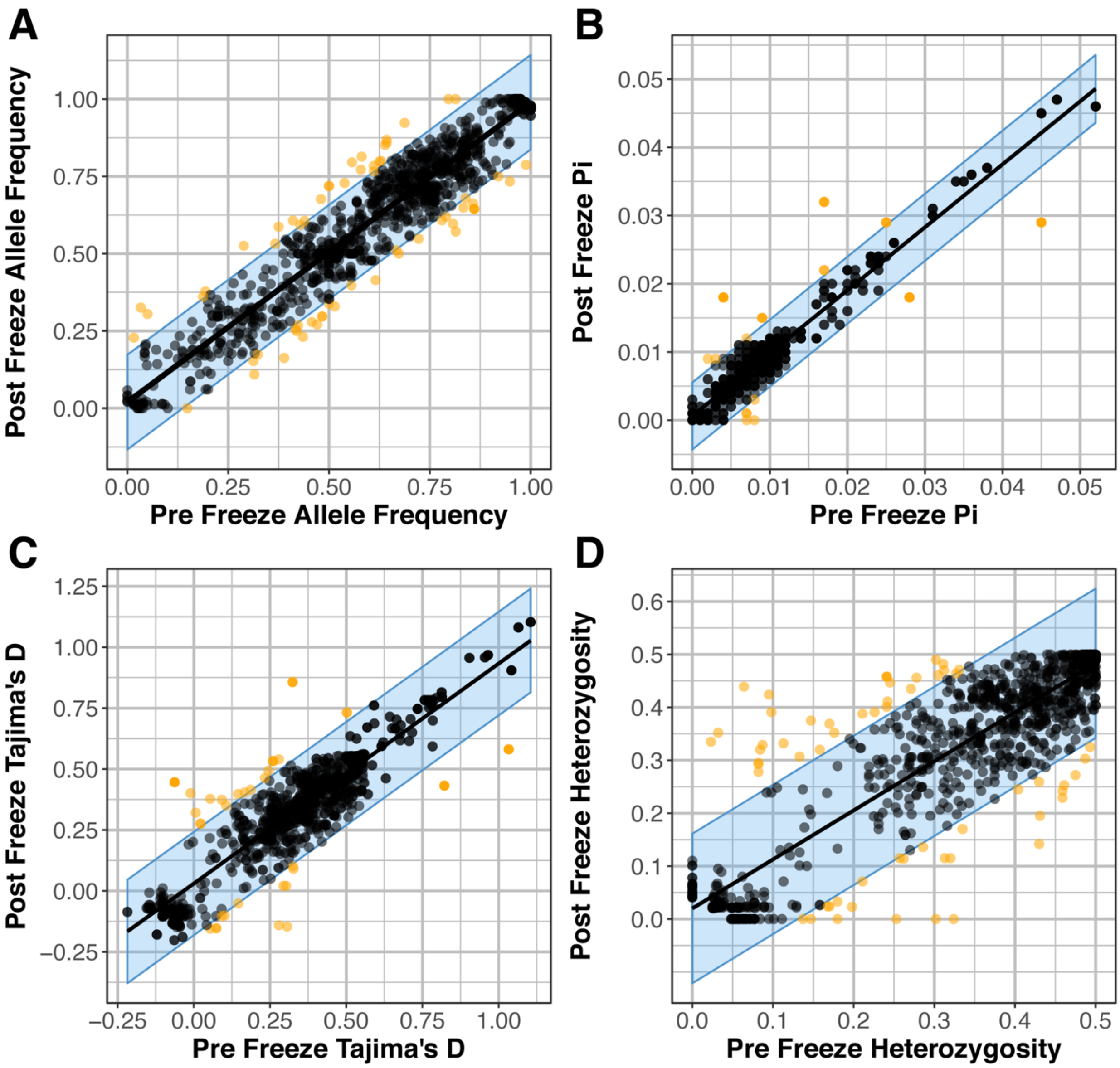


Figure 5. Cellular and anatomic evidence for unique, up-regulated fasted physiological states in adapted Florida pythons. (**A**) Relative activation states for canonical pathways and upstream regulatory molecules previously shown to be important in python regenerative organ growth based on gene expression data. Pairwise comparisons represent relative activation between fasted post-freeze Florida pythons and laboratory pythons in fasted and post-feeding morphological states. (**B**) Boxplots showing cell size measurements between laboratory pythons in fasted and post-feeding morphological states and in fasted post-freeze Florida pythons for four organs. Horizontal bars indicate pairwise comparisons between the measurements from the fasted invasive Florida pythons and respective treatments from the laboratory pythons, with an asterisk indicating a statistically significant difference (Tukey's HSD p-value ≤ 0.05). (**C**) Example electron micrographs of proximal intestinal microvilli at several key time points during the normal feeding cycle in laboratory pythons, and in a post-freeze fasted Florida python. Scale bar = 1 µM. FL = invasive Florida python; DPF = days post fed.

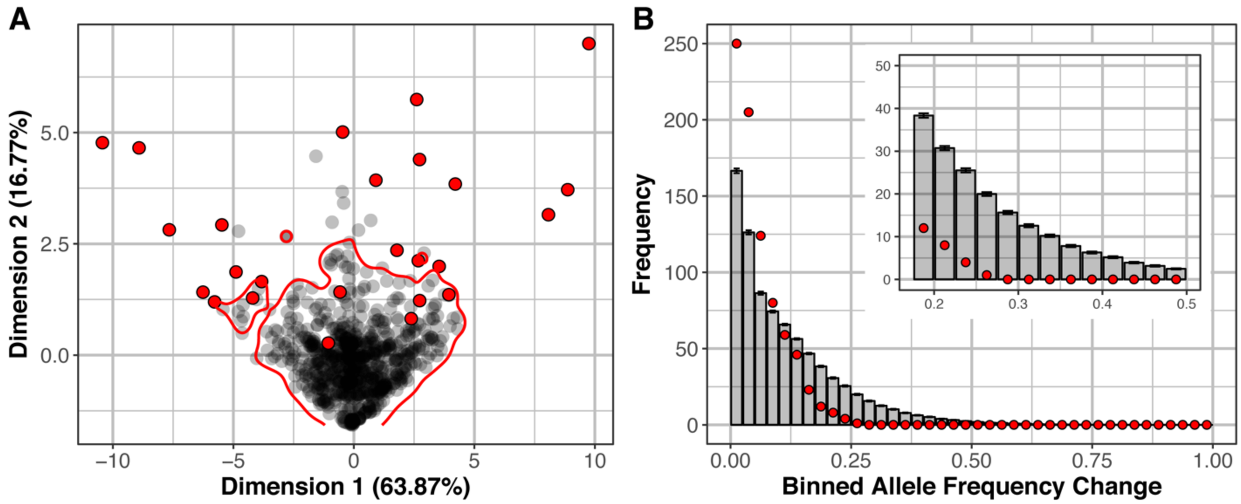
SUPPLEMENTARY MATERIAL



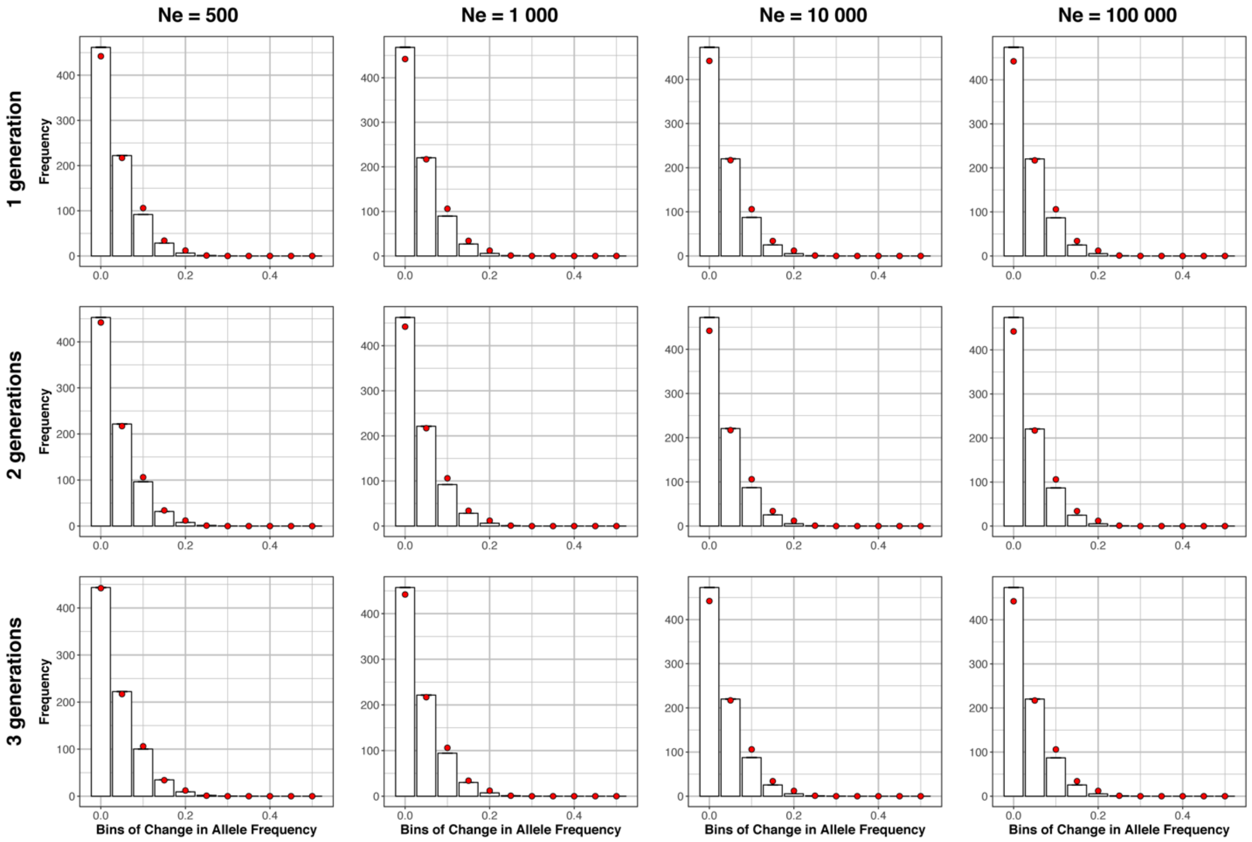
Supplementary Figure 1. Geographical and temporal patterns of population sampling. (**A**) Map outlining the locations where invasive pythons were sampled as part of this work. (**B**) Histogram of years where samples from the invasive population were collected.



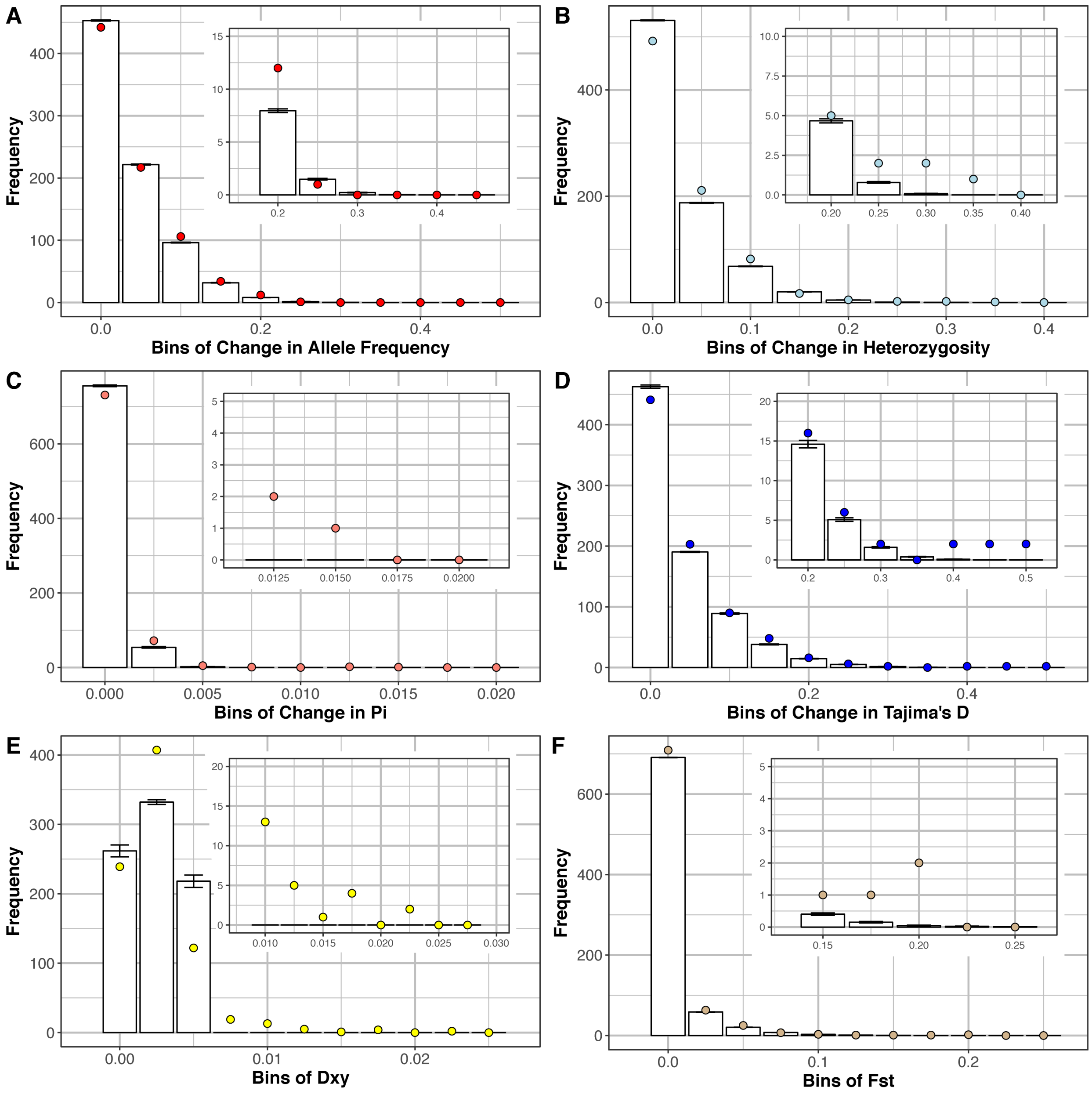
Supplementary Figure 2. Population genetic fluctuations between pre- and post-freeze populations. Comparisons of four univariate intrapopulation genetic statistics between the Pre-Freeze and Post-Freeze populations: (**A**) allele frequency; (**B**) nucleotide diversity; (**C**) Tajima's D; and (**D**) heterozygosity. Each point represents an estimate from a variant site. The least-squares line indicates the high amount of correlation between these measures at both time points. The 95% confidence interval of this correlation is shown in blue. Those points falling outside of that interval, which represent high fluctuations in these parameter estimates, are indicated in orange.



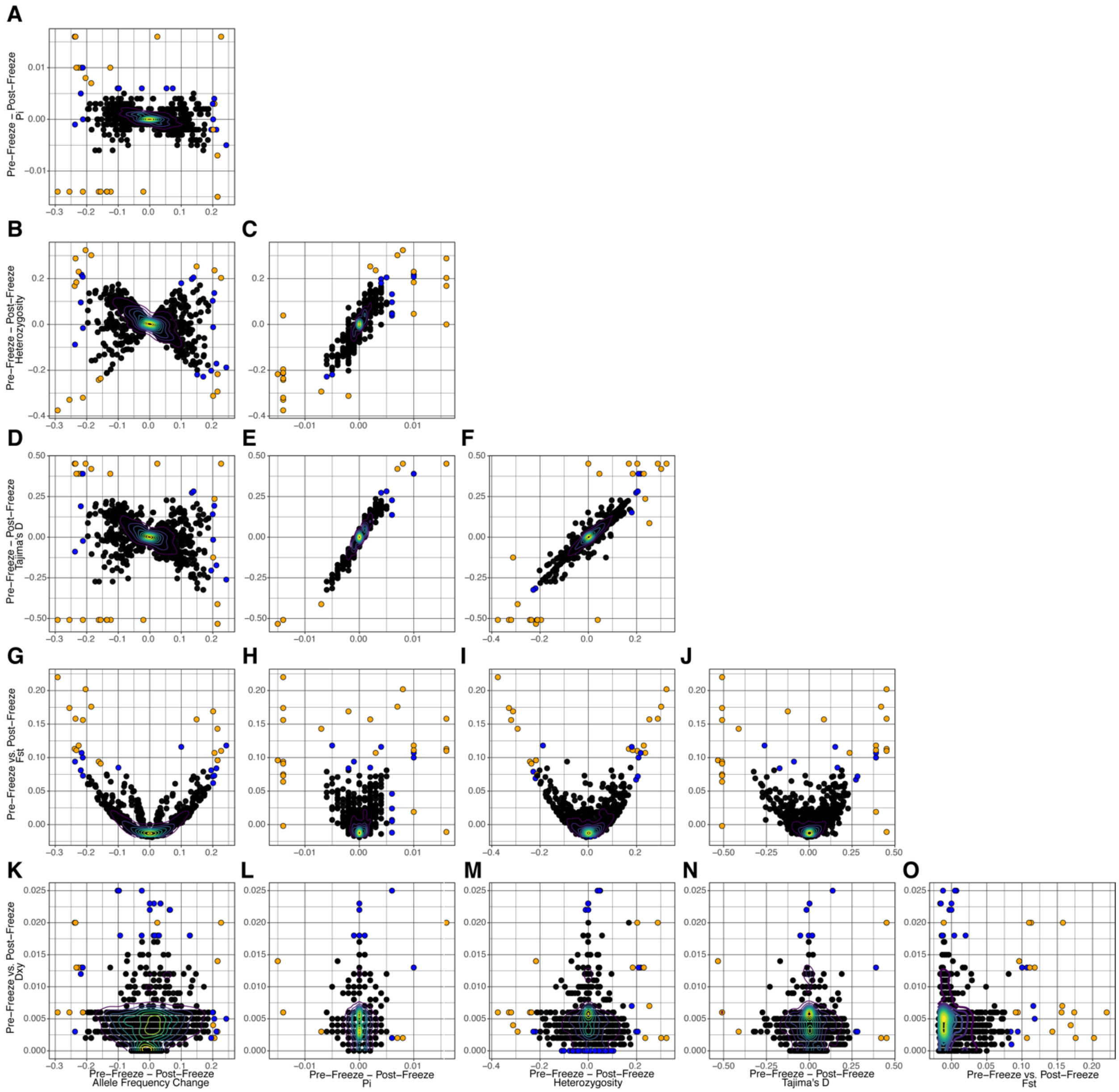
Supplementary Figure 3. Analysis of empirical genomic variation data support inferences of evolution and selection. (**A**) Principle component analysis of six univariate population genetic statistics for each locus demonstrating outlying nature of loci identified using Mahalanobis distance and selection coefficients as 'genome scan outliers'. Each point represents a locus, with gray points representing loci not identified as "genome scan outliers' and red points representing loci identified as outliers by these approaches. The 95% quantile of the distribution of all loci is shown with the red contour line. (**B**) Comparison of allele frequency fluctuations between empirical temporal population sampling and permutations with random population assignments of individuals are significantly different. Shown are the mean (bars) and 95% confidence intervals (error bars) of allele frequency change based on 1 000 permutations, and the empirical distribution of allele frequency change between pre- and post-freeze populations is shown as red-colored points.



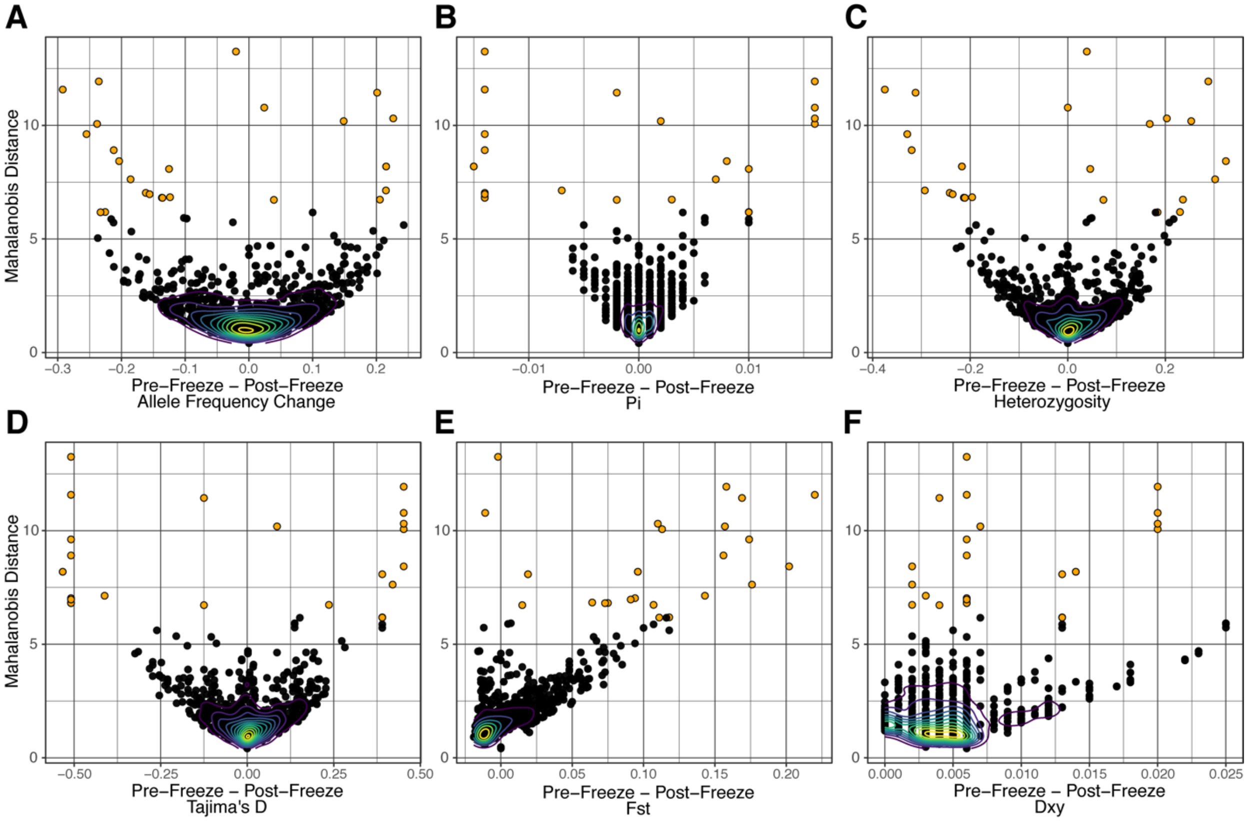
Supplementary Figure 4. Simulations of neutral genetic drift are robust to different demographic assumptions describing the Florida Burmese python population. Both the assumed effective population size (Ne; columns) and the number of generations of drift between pre- and post-freeze populations (rows) were varied. Ne values of 500, 1 000, 10 000, and 100 000 cover the likely range of actual Ne in the Florida Burmese python population. Typical average generation times in Burmese pythons are three years, so 1-3 generations of genetic drift encompasses all variability in this estimate. For further investigations, we used simulations assuming Ne of 500 and 2 generations of genetic drift.



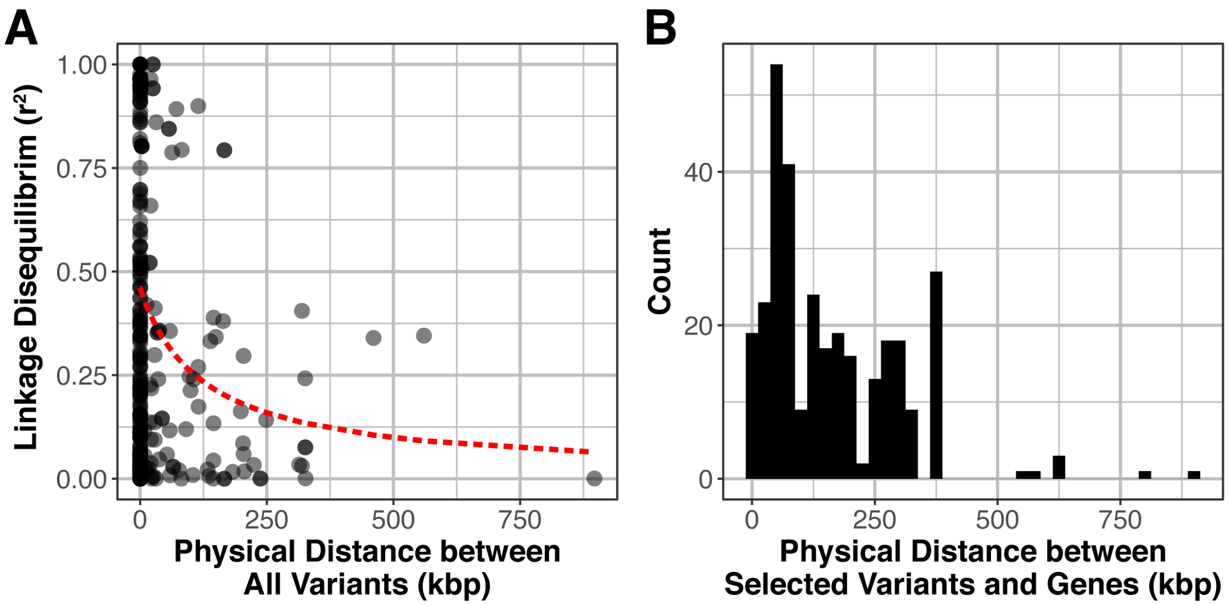
Supplementary Figure 5. Simulations indicate an excess of variants with high allelic differentiation based on several population genetic statistics. Histograms of means with 95% confidence intervals from 1 000 simulated datasets for six univariate population genetic statistics: **(A)** change in allele frequency between time points; **(B)** change in heterozygosity between time points; **(C)** change in nucleotide diversity between time points; **(D)** change in Tajima's D; **(E)** D_{XY} between time points; and **(F)** F_{ST} between time points. Colored points represent the counts of empirical values within each bin for respective univariate population genetic statistics. Inset plots show frequency histograms and data points for the more extreme bins of allelic differentiation.



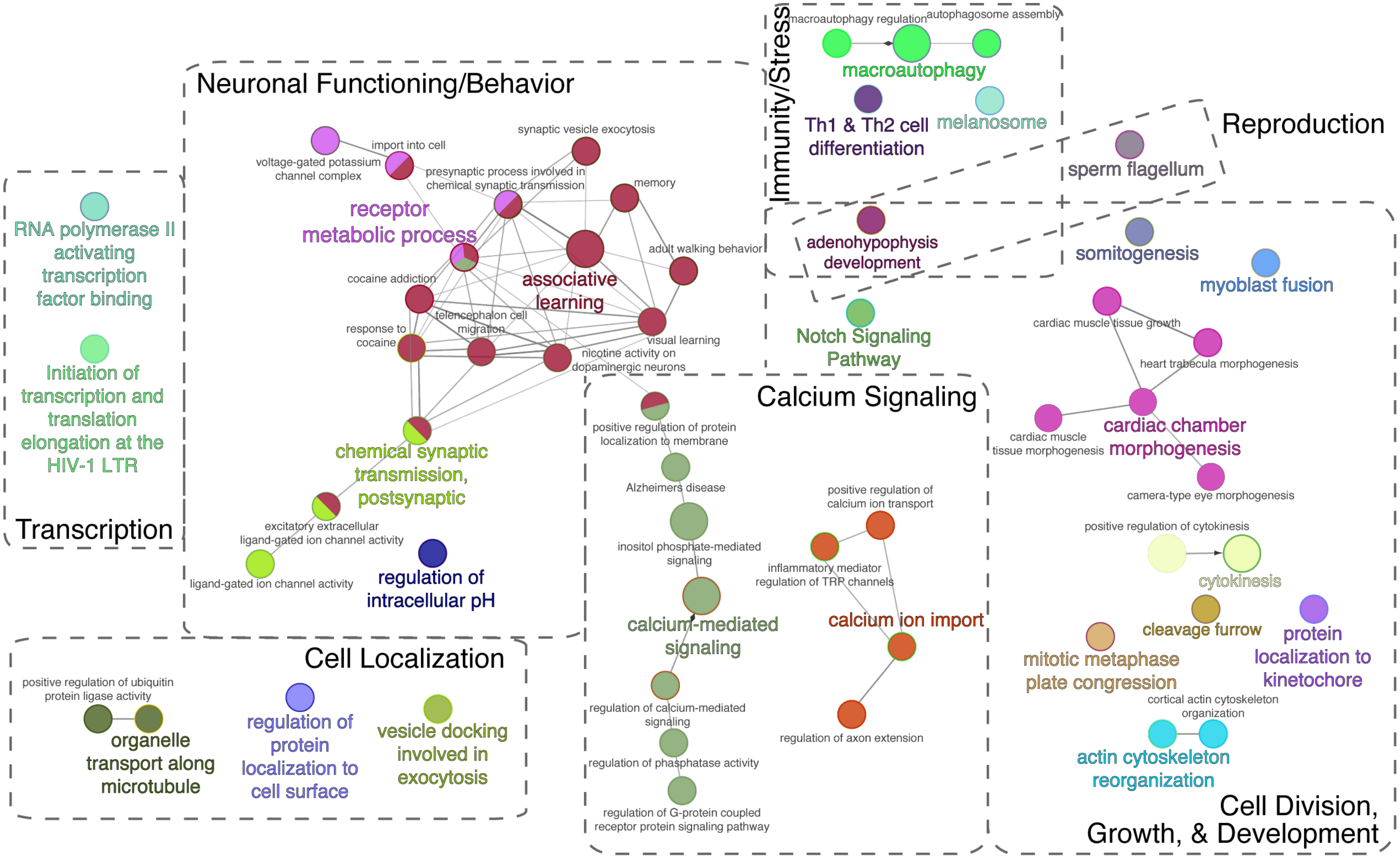
Supplementary Figure 6. Distributions of univariate population genetic statistics are correlated but identify subsets of distinct outlier variants. Plots comparing all univariate population genetic statistics: **(A)** change in nucleotide diversity vs. allele frequency change; **(B)** change in heterozygosity vs. allele frequency change; **(C)** change in heterozygosity vs. change in nucleotide diversity; **(D)** difference in Tajima's D vs. allele frequency change; **(E)** difference in Tajima's D vs. change in nucleotide diversity; **(F)** difference in Tajima's D vs. change in heterozygosity; **(G)** F_{ST} vs. allele frequency change; **(H)** F_{ST} vs. change in nucleotide diversity; **(I)** F_{ST} vs. change in heterozygosity; **(J)** F_{ST} vs. difference in Tajima's D; **(K)** D_{XY} vs. allele frequency change; **(L)** D_{XY} vs. change in nucleotide diversity; **(M)** D_{XY} vs. change in heterozygosity; **(N)** D_{XY} vs. difference in Tajima's D; and **(O)** D_{XY} vs. F_{ST}. Each point represents an estimate from a variant site. The relative density of points is indicated by the isolines that range from yellow (high density) to blue (low density). Blue points indicate variants where either univariate statistic falls above the 97.5% quantile in its respective distribution. Points with a Mahalanobis distance greater than the 97.5% quantile are indicated in orange.



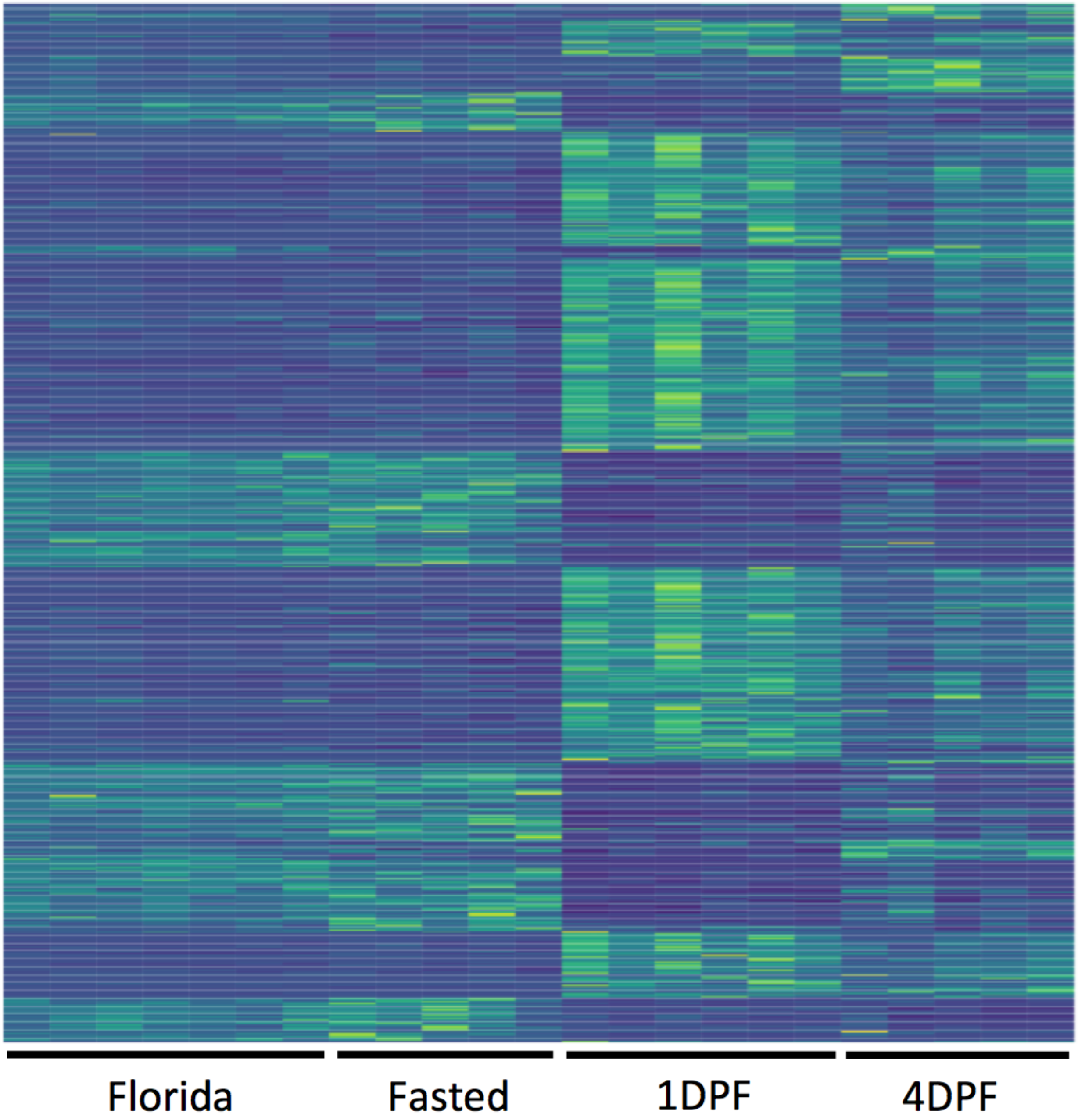
Supplementary Figure 7. Multivariate Mahalanobis distance identifies both shared and distinct sets of outlier variants based on each univariate statistic. Bivariate plots comparing each univariate population genetic statistic with the composite multivariate Mahalanobis distance that they contribute to: **(A)** allele frequency fluctuation; **(B)** difference in nucleotide diversity; **(C)** difference in heterozygosity; **(D)** difference in Tajima's D; **(E)** F_{ST}; and **(F)** D_{XY}. Each point represents an estimate from a variant site. The reletive density of points is indicated by the isolines that range from light blue (high density) to dark blue (low density). Points with a Mahalanobis distance greater than the 97.5% quantile are indicated in orange.



Supplementary Figure 8. Persistent long-range linkage disequilibrium spans physical distance between putatively selected variants and annotated genes. Identification of genes linked to putatively selected variants. (A) Patterns of linkage disequilibrium decay across all pairwise comparisons between variant sites genome-wide indicate that linkage disequilibrium extend up to 1 Mb. **(B)** Distribution of pairwise distances between putatively selected variants and genes within syntenic regions of the Burmese python genome. Patterns of linkage disequilibrium decay and the relatively low contiguity of the Burmese python genome assembly led us to further investigate all genes contained within the 12 identified syntenic regions with putatively selected variants.


Neuronal Functioning/Behavior
voltage-gated potassium channel complex
import into cell
synaptic vesicle exocytosis
presynaptic process involved in chemical synaptic transmission
memory
receptor metabolic process
adult walking behavior
cocaine addiction
associative learning
telencephalon cell migration
response to cocaine
visual learning
nicotine activity on dopaminergic neurons
chemical synaptic transmission, postsynaptic
excitatory extracellular ligand-gated ion channel activity
ligand-gated ion channel activity
regulation of intracellular pH
RNA polymerase II activating transcription factor binding
Initiation of transcription and translation elongation at the HIV-1 LTR
Transcription
Immunity/Stress
macroautophagy regulation
autophagosome assembly
macroautophagy
Th1 & Th2 cell differentiation
melanosome
Reproduction
sperm flagellum
adenohypophysis development
somitogenesis
myoblast fusion
Notch Signaling Pathway
cardiac muscle tissue growth
heart trabecula morphogenesis
cardiac muscle tissue morphogenesis
cardiac chamber morphogenesis
camera-type eye morphogenesis
Calcium Signaling
positive regulation of protein localization to membrane
Alzheimers disease
inositol phosphate-mediated signaling
positive regulation of calcium ion transport
inflammatory mediator regulation of TRP channels
calcium-mediated signaling
calcium ion import
regulation of calcium-mediated signaling
regulation of phosphatase activity
regulation of axon extension
regulation of G-protein coupled receptor protein signaling pathway
positive regulation of cytokinesis
cytokinesis
cleavage furrow
protein localization to kinetochore
mitotic metaphase plate congression
cortical actin cytoskeleton organization
actin cytoskeleton reorganization
Cell Division, Growth, & Development
Cell Localization
positive regulation of ubiquitin protein ligase activity
organelle transport along microtubule
regulation of protein localization to cell surface
vesicle docking involved in exocytosis

Supplementary Figure 9. GO analyses find enrichment for genes related to reproduction, behavior, and regenerative organ growth. Enlarged version of Fig. 2D showing the results of GO term analysis. More specific node labels have been included to provide further context.



Supplementary Figure 10. Fasted post-freeze Florida Burmese pythons resemble normal fasted pythons in gene expression state across genes. Heatmap comparing expression between fasted samples from the Florida python population and experimental pythons at controlled time points for N = 1 118 genes that are significantly differentially expressed across the experimental time points based on a regression analysis.

Supplementary Table 1. Details of Burmese python sampling used in this study**.** DPF = Days post fed.

Sample ID	Population/ Organ	Date Collected	Location Collected	Data Type	Tissue Type – State (RNAseq)	NCBI Accession
Pymo001	Post-Freeze	10-Feb-2013	25.3593636 N, -80.5520245 W	RADseq	N/A	PENDING
Pymo002	Post-Freeze	1-Feb-2013	25.42050236 N, -80.57419218 W	RADseq	N/A	PENDING
Pymo003	Post-Freeze	7-Feb-2013	25.45586783 N, -80.56272861 W	RADseq	N/A	PENDING
Pymo004	Post-Freeze	27-Jan-2013	25.39690941 N, -80.57542826 W	RADseq	N/A	PENDING
Pymo005	Post-Freeze	1-Feb-2013	25.41866146 N, -80.58093049 W	RADseq	N/A	PENDING
Pymo006	Post-Freeze	10-Feb-2013	25.76171716 N, -80.58329367 W	RADseq	N/A	PENDING
Pymo007	Post-Freeze	10-Feb-2013	25.35934545 N, -80.55199475 W	RADseq	N/A	PENDING
Pymo008	Post-Freeze	10-Feb-2013	25.55573989 N, -80.5616799 W	RADseq	N/A	PENDING
Pymo009	Post-Freeze	7-Feb-2013	25.83128263 N, -80.81300196 W	RADseq	N/A	PENDING
Pymo010	Post-Freeze	8-Feb-2013	25.79578525 N, -80.8200003 W	RADseq	N/A	PENDING
Pymo011	Post-Freeze	6-Feb-2013	25.36027289 N, -80.57234608 W	RADseq	N/A	PENDING
Pymo012	Post-Freeze	18-Feb-2013	25.43636777 N, -80.5897005 W	RADseq	N/A	PENDING
Pymo013	Post-Freeze	18-Feb-2013	25.35919037 N, -80.55745169 W	RADseq	N/A	PENDING
Pymo014	Post-Freeze	30-Oct-2012	25.97380498 N, -80.466018 W	RADseq	N/A	PENDING
Pymo015	Post-Freeze	5-Nov-2012	25.76211981 N, -80.70341641 W	RADseq	N/A	PENDING

Pymo016	Post-Freeze	10-Mar-2013	25.30387135 N, -80.48827234 W	RADseq	N/A	PENDING
Pymo017	Post-Freeze	14-Jan-2013	25.90688079 N, -80.5891549 W	RADseq	N/A	PENDING
Pymo018	Post-Freeze	10-Mar-2013	25.72797406 N, -80.67252656 W	RADseq	N/A	PENDING
Pymo019	Post-Freeze	18-Feb-2013	25.35920846 N, -80.55746157 W	RADseq	N/A	PENDING
Pymo020	Post-Freeze	20-Feb-2013	25.36740488 N, -80.49211997 W	RADseq	N/A	PENDING
Pymo021	Post-Freeze	19-Jan-2013	25.35559579 N, -80.49290476 W	RADseq	N/A	PENDING
Pymo022	Post-Freeze	11-Jan-2013	25.36650456 N, -80.49290894 W	RADseq	N/A	PENDING
Pymo023	Post-Freeze	15-May-2013	25.5391747 N, -80.5604362 W	RADseq	N/A	PENDING
Pymo024	Post-Freeze	16-May-2013	25.6014265 N, -80.5748581 W	RADseq	N/A	PENDING
Pymo025	Post-Freeze	13-Dec-2012	25.7617847 N, -80.657715 W	RADseq	N/A	PENDING
Pymo026	Post-Freeze	6-Dec-2013	26.1457993 N, -80.5064158 W	RADseq	N/A	PENDING
Pymo027	Post-Freeze	4-Mar-2013	25.3676784 N, -80.5663563 W	RADseq	N/A	PENDING
Pymo028	Post-Freeze	18-Feb-2013	25.3382125 N, -80.493007 W	RADseq	N/A	PENDING
Pymo029	Post-Freeze	10-Dec-2012	25.3583138 N, -80.4928735 W	RADseq	N/A	PENDING
Pymo031	Post-Freeze	4-Mar-2013	25.3496722 N, -80.4930189 W	RADseq	N/A	PENDING
Pymo032	Post-Freeze	27-Mar-2013	25.7616945 N, -80.8111964 W	RADseq	N/A	PENDING
Pymo034	Post-Freeze	22-Feb-2013	25.7292746 N, -80.6726127 W	RADseq	N/A	PENDING
Pymo035	Post-Freeze	23-Apr-2013	25.5278867 N,	RADseq	N/A	PENDING

			-80.5603778 W			
Pymo036	Post-Freeze	30-Jan-2013	25.3961601 N, -80.5662347 W	RADseq	N/A	PENDING
Pymo037	Post-Freeze	5-Mar-2013	25.467925 N, -80.61858333 W	RADseq	N/A	PENDING
Pymo038	Post-Freeze	4-Apr-2013	25.4183729 N, -80.5810906 W	RADseq	N/A	PENDING
Pymo039	Post-Freeze	21-Jun-2013	25.7607242 N, -81.0008775 W	RADseq	N/A	PENDING
Pymo040	Post-Freeze	22-Feb-2013	25.7292746 N, -80.6726127 W	RADseq	N/A	PENDING
Pymo041	Post-Freeze	8-Jan-2013	25.4493327 N, -80.537131 W	RADseq	N/A	PENDING
Pymo042	Post-Freeze	1-Jan-2013	25.8233791 N, -80.8500004 W	RADseq	N/A	PENDING
Pymo044	Post-Freeze	10-May-2013	25.6275806 N, -80.5757816 W	RADseq	N/A	PENDING
Pymo045	Post-Freeze	23-May-2013	25.5864373 N, -80.5751601 W	RADseq	N/A	PENDING
Pymo046	Post-Freeze	23-May-2013	25.6027449 N, -80.5748435 W	RADseq	N/A	PENDING
Pymo047	Post-Freeze	28-May-2013	25.1841438 N, -80.896416 W	RADseq	N/A	PENDING
Pymo048	Post-Freeze	27-Apr-2013	25.2723096 N, -80.7984334 W	RADseq	N/A	PENDING
Pymo049	Post-Freeze	18-May-2013	25.6290888 N, -80.5758161 W	RADseq	N/A	PENDING
Pymo050	Post-Freeze	9-May-2013	25.4416039 N, -80.7837412 W	RADseq	N/A	PENDING
Pymo051	Post-Freeze	5-May-2013	25.3293934 N, -80.6983882 W	RADseq	N/A	PENDING
Pymo052	Post-Freeze	2-Jul-2013	25.6085005 N, -80.5401649 W	RADseq	N/A	PENDING
Pymo053	Pre-Freeze	19-May-2003	25.3581395 N, -80.8212499 W	RADseq	N/A	PENDING

Pymo055	Pre-Freeze	19-Jul-2004	25.4328743 N, -80.5619862 W	RADseq	N/A	PENDING
Pymo056	Pre-Freeze	17-Nov-2004	25.4416709 N, -80.5623423 W	RADseq	N/A	PENDING
Pymo057	Pre-Freeze	17-Dec-2004	25.4426579 N, -80.5693703 W	RADseq	N/A	PENDING
Pymo063	Pre-Freeze	5-Nov-2006	25.7600886 N, -80.752025 W	RADseq	N/A	PENDING
Pymo067	Pre-Freeze	19-Aug-2007	25.3985396 N, -80.5996512 W	RADseq	N/A	PENDING
Pymo070	Pre-Freeze	13-Nov-2007	25.1473418 N, -80.924337 W	RADseq	N/A	PENDING
Pymo072	Pre-Freeze	25-Nov-2007	25.5387304 N, -80.5421533 W	RADseq	N/A	PENDING
Pymo073	Pre-Freeze	28-Nov-2007	25.4028295 N, -80.5618064 W	RADseq	N/A	PENDING
Pymo075	Pre-Freeze	14-Jan-2008	25.5387304 N, -80.5421533 W	RADseq	N/A	PENDING
Pymo077	Pre-Freeze	9-Apr-2008	25.4058376 N, -80.6136561 W	RADseq	N/A	PENDING
Pymo078	Pre-Freeze	9-May-2008	25.7621015 N, -80.731179 W	RADseq	N/A	PENDING
Pymo081	Pre-Freeze	13-Dec-2008	25.7621082 N, -80.7501759 W	RADseq	N/A	PENDING
Pymo084	Pre-Freeze	30-Jan-2009	25.7619858 N, -80.7994487 W	RADseq	N/A	PENDING
Pymo114	Pre-Freeze	21-Dec-2006	25.858325 N, -81.031975 W	RADseq	N/A	PENDING
Pymo115	Pre-Freeze	21-Jan-2007	26.15351944 N, -81.34676389 W	RADseq	N/A	PENDING
Pymo116	Pre-Freeze	15-Feb-2007	25.4972585 N, -80.7717014 W	RADseq	N/A	PENDING
Pymo117	Pre-Freeze	15-Feb-2007	25.497168 N, -80.7715722 W	RADseq	N/A	PENDING
Pymo118	Pre-Freeze	9-Mar-2007	25.3401891 N,	RADseq	N/A	PENDING

			-80.4926808 W			
Pymo119	Pre-Freeze	18-Oct-2007	25.4138927 N, -80.9757878 W	RADseq	N/A	PENDING
Pymo120	Pre-Freeze	13-Jan-2008	25.3713721 N, -80.4929383 W	RADseq	N/A	PENDING
Pymo121	Pre-Freeze	17-Apr-2008	25.7620418 N, -80.8216468 W	RADseq	N/A	PENDING
Pymo122	Pre-Freeze	4-May-2008	25.779435 N, -80.8464851 W	RADseq	N/A	PENDING
Pymo123	Pre-Freeze	30-May-2008	25.761687 N, -80.8195531 W	RADseq	N/A	PENDING
Pymo124	Pre-Freeze	16-Aug-2008	25.89740556 N, -81.31851667 W	RADseq	N/A	PENDING
Pymo125	Pre-Freeze	13-Nov-2008	25.0884853 N, -80.4464208 W	RADseq	N/A	PENDING
Pymo126	Pre-Freeze	17-Nov-2008	25.7455083 N, -80.9484307 W	RADseq	N/A	PENDING
Pymo127	Pre-Freeze	21-Dec-2008	25.6685123 N, -80.8595314 W	RADseq	N/A	PENDING
Pymo128	Pre-Freeze	14-Jan-2009	26.1507214 N, -81.5434788 W	RADseq	N/A	PENDING
Pymo129	Pre-Freeze	17-Jun-2009	25.311551 N, -80.449776 W	RADseq	N/A	PENDING
Pymo174	Pre-Freeze	25-Apr-2006	25.66504302 N, -80.76655029 W	RADseq	N/A	PENDING
Pymo175	Pre-Freeze	21-Jul-2006	25.40396139 N, -80.56587874 W	RADseq	N/A	PENDING
Pymo176	Pre-Freeze	27-Jul-2006	25.43281416 N, -80.56914669 W	RADseq	N/A	PENDING
Pymo177	Pre-Freeze	10-Aug-2006	25.44542253 N, -80.56980799 W	RADseq	N/A	PENDING
Pymo178	Pre-Freeze	10-Aug-2006	25.44542253 N, -80.56980799 W	RADseq	N/A	PENDING
Pymo179	Pre-Freeze	11-Aug-2006	25.43695477 N, -80.56157366 W	RADseq	N/A	PENDING

Pymo180	Pre-Freeze	22-Aug-2006	25.76200644 N, -80.68805959 W	RADseq	N/A	PENDING
Pymo181	Pre-Freeze	8-Nov-2006	25.23952362 N, -80.80869462 W	RADseq	N/A	PENDING
Pymo182	Pre-Freeze	2-Mar-2007	25.37492582 N, -80.82745773 W	RADseq	N/A	PENDING
Pymo183	Pre-Freeze	8-Mar-2007	25.38838908 N, -80.61304563 W	RADseq	N/A	PENDING
Pymo184	Pre-Freeze	8-Mar-2007	25.69007178 N, -80.67127488 W	RADseq	N/A	PENDING
Pymo185	Pre-Freeze	23-Mar-2007	25.37276469 N, -80.82507519 W	RADseq	N/A	PENDING
Pymo186	Pre-Freeze	3-Apr-2007	25.76279721 N, -80.67392691 W	RADseq	N/A	PENDING
Pymo187	Pre-Freeze	18-Oct-2007	25.29940167 N, -80.7983687 W	RADseq	N/A	PENDING
Pymo188	Pre-Freeze	14-Nov-2007	25.15629503 N, -80.91431977 W	RADseq	N/A	PENDING
Pymo189	Pre-Freeze	21-Nov-2007	25.43306117 N, -80.50162074 W	RADseq	N/A	PENDING
Pymo191	Pre-Freeze	28-Nov-2007	25.40247937 N, -80.56248378 W	RADseq	N/A	PENDING
Pymo192	Pre-Freeze	30-Nov-2007	25.40730194 N, -80.54183531 W	RADseq	N/A	PENDING
Pymo208	Post-Freeze (2016)	19-Jan-2016	25.77873774 N, -80.84465086 W	RNAseq	Small intestine – Fed	SRX2724380
Pymo209	Post-Freeze (2016)	20-Jan-2016	25.33463564 N, -80.4927636 W	RNAseq	Small intestine – Fasted	SRX2724379
Pymo210	Post-Freeze (2016)	20-Jan-2016	25.37749462 N, -80.4929227 W	RNAseq	Small intestine – Fasted	SRX2724378
Pymo211	Post-Freeze (2016)	20-Jan-2016	25.37833457 N, -80.49295892 W	RNAseq	Small intestine – Fasted	SRX3447894
Pymo212	Post-Freeze (2016)	20-Jan-2016	25.81181338 N, -80.43318676 W	RNAseq	Small intestine – Fasted	SRX3447895
Pymo213	Post-Freeze	20-Jan-2016	25.43146598 N,	RNAseq	Small intestine –	SRX2724377

	(2016)		-80.57443211 W		Fasted	
Pymo215	Post-Freeze (2016)	24-Jan-2016	25.55532341 N, -80.55533021 W	RNAseq	Small intestine – Fasted	SRX2724376
Pymo216	Post-Freeze (2016)	24-Jan-2016	25.55134569 N, -80.55984434 W	RNAseq	Small intestine – Fed	SRX2724375
Pymo218	Post-Freeze (2016)	21-Jan-2016	25.82958118 N, -80.84849588 W	RNAseq	Small intestine – Fasted	SRX2724374
Pymo219	Post-Freeze (2016)	25-Jan-2016	25.58127291 N, -80.52830067 W	RNAseq	Small intestine – Fed	SRX2724373
AF2	Laboratory	N/A	N/A	RNAseq	Small Intestine – 0 DPF	SRX834426
AI6-1	Laboratory	N/A	N/A	RNAseq	Small Intestine – 0 DPF	SRX834422
AI6-2	Laboratory	N/A	N/A	RNAseq	Small Intestine – 0 DPF	SRX834423
AI8	Laboratory	N/A	N/A	RNAseq	Small Intestine – 0 DPF	SRX834424
AJ6-1	Laboratory	N/A	N/A	RNAseq	Small Intestine – 0 DPF	SRX834418
AJ6-2	Laboratory	N/A	N/A	RNAseq	Small Intestine – 0 DPF	SRX834419
AJ6-3	Laboratory	N/A	N/A	RNAseq	Small Intestine – 0 DPF	SRX834420
U25	Laboratory	N/A	N/A	RNAseq	Small Intestine – 0 DPF	SRX834425
S6-1	Laboratory	N/A	N/A	RNAseq	Small Intestine – 1 DPF	SRX834445
S6-2	Laboratory	N/A	N/A	RNAseq	Small Intestine – 1 DPF	SRX834446
V43	Laboratory	N/A	N/A	RNAseq	Small Intestine – 1 DPF	SRX834447
W20	Laboratory	N/A	N/A	RNAseq	Small Intestine – 1 DPF	SRX834448
Z12-1	Laboratory	N/A	N/A	RNAseq	Small Intestine – 1 DPF	SRX2506434

Z12-2	Laboratory	N/A	N/A	RNAseq	Small Intestine – 1 DPF	SRX2506435
Z14-1	Laboratory	N/A	N/A	RNAseq	Small Intestine – 1 DPF	SRX834439
Z14-2	Laboratory	N/A	N/A	RNAseq	Small Intestine – 1 DPF	SRX834440
Z14-3	Laboratory	N/A	N/A	RNAseq	Small Intestine – 1 DPF	SRX834441
Z18	Laboratory	N/A	N/A	RNAseq	Small Intestine – 1 DPF	SRX834444
V40	Laboratory	N/A	N/A	RNAseq	Small Intestine – 4 DPF	SRX834458
Y18-1	Laboratory	N/A	N/A	RNAseq	Small Intestine – 4 DPF	SRX834451
Y18-2	Laboratory	N/A	N/A	RNAseq	Small Intestine – 4 DPF	SRX834452
Y18-3	Laboratory	N/A	N/A	RNAseq	Small Intestine – 4 DPF	SRX834453
Y23-1	Laboratory	N/A	N/A	RNAseq	Small Intestine – 4 DPF	SRX834454
Y23-2	Laboratory	N/A	N/A	RNAseq	Small Intestine – 4 DPF	SRX834455
Y24	Laboratory	N/A	N/A	RNAseq	Small Intestine – 4 DPF	SRX834456
Y5-1	Laboratory	N/A	N/A	RNAseq	Small Intestine – 4 DPF	SRX834449
Y5-2	Laboratory	N/A	N/A	RNAseq	Small Intestine – 4 DPF	SRX834450

CITATIONS

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andrew, A. L., Card, D. C., Ruggiero, R. P., Schield, D. R., Adams, R. H., Pollock, D. D., … Castoe, T. A. (2015). Rapid changes in gene expression direct rapid shifts in intestinal form and function in the Burmese python after feeding. *Physiological Genomics*, *47*(5), 147–157. https://doi.org/10.1152/physiolgenomics.00131.2014

Andrew, A. L., Perry, B. W., Card, D. C., Schield, D. R., Ruggiero, R. P., McGaugh, S. E., … Castoe, T. A. (2017). Growth and stress response mechanisms underlying post-feeding regenerative organ growth in the Burmese python. *BMC Genomics*, *18*, 338. https://doi.org/10.1186/s12864-017-3743-1

Avery, M. L., Engeman, R. M., Keacher, K. L., Humphrey, J. S., Bruce, W. E., Mathies, T. C., & Mauldin, R. E. (2010). Cold weather and the potential range of invasive Burmese pythons. *Biological Invasions*, *12*(11), 3649–3652. https://doi.org/10.1007/s10530-010-9761-4

Barker, D. G., & Barker, T. M. (2008). The Distribution of the Burmese Python, *Python molurus bivittatus*. *Bulletin of the Chicago Herpetological Society*, *43*(3), 33–38.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., … Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, *25*(8), 1091–1093. https://doi.org/10.1093/bioinformatics/btp101

Blair, C., Campbell, C. R., & Yoder, A. D. (2015). Assessing the utility of whole genome amplified DNA for next-generation molecular ecology. *Molecular Ecology Resources*, *15*(5), 1079–1090. https://doi.org/10.1111/1755-0998.12376

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., … Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1), 1–31. https://doi.org/10.1186/2047-217X-2-10

Campbell-Staton, S. C., Cheviron, Z. A., Rochette, N., Catchen, J., Losos, J. B., & Edwards, S. V. (2017). Winter storms drive rapid phenotypic, regulatory, and genomic shifts in the green anole lizard. *Science*, *357*(6350), 495–498. https://doi.org/10.1126/science.aam5512

Castoe, T. A., Koning, A. P. J. de, Hall, K. T., Card, D. C., Schield, D. R., Fujita, M. K., … Pollock, D. D. (2013). The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proceedings of the National Academy of Sciences*, *110*(51), 20645–20650. https://doi.org/10.1073/pnas.1314475110

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes, Genomes, Genetics*, *1*(3), 171–182. https://doi.org/10.1534/g3.111.000240

Catchen, J. M., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Collins, T. M., Freeman, B., & Snow, R. W. (2008). *Final report: genetic characterization of populations of the nonindigenous Burmese python in Everglades National Park* (Final report for the South Florida Water Management District) (pp. 1–30). Miami, FL: Florida International University.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491. https://doi.org/10.1038/ng.806

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dorcas, M. E., Willson, J. D., & Gibbons, J. W. (2011). Can invasive Burmese pythons inhabit temperate regions of the southeastern United States? *Biological Invasions*, *13*(4), 793–802. https://doi.org/10.1007/s10530-010-9869-6

Dorcas, M. E., Willson, J. D., Reed, R. N., Snow, R. W., Rochford, M. R., Miller, M. A., … Hart, K. M. (2012). Severe mammal declines coincide with proliferation of invasive Burmese pythons in Everglades National Park. *Proceedings of the National Academy of Sciences*, *109*(7), 2418–2422. https://doi.org/10.1073/pnas.1115226109

Dove, C. J., Snow, R. W., Rochford, M. R., & Mazzotti, F. J. (2011). Birds Consumed by the Invasive Burmese Python (*Python molurus bivittatus*) in Everglades National Park, Florida, USA. *The Wilson Journal of Ornithology*, *123*(1), 126–131. https://doi.org/10.1676/10-092.1

Elith, J., Graham, C. H., Anderson, R. P., Miroslav, D., Ferrier, S., Guisan, A., … Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Engeman, R., Jacobson, E., Avery, M. L., & Meshaka, W. E. (2011). The aggressive invasion of exotic reptiles in Florida with a focus on prominent species: A review. *Current Zoology*, *57*(5), 599–612. https://doi.org/10.1093/czoolo/57.5.599

Epstein, B., Jones, M., Hamede, R., Hendricks, S., McCallum, H., Murchison, E. P., … Storfer, A. (2016). Rapid evolutionary response to a transmissible cancer in Tasmanian devils. *Nature Communications*, *7*, 12684. https://doi.org/10.1038/ncomms12684

Evangelou, E., & Ioannidis, J. P. A. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, *14*(6), 379–389. https://doi.org/10.1038/nrg3472

Favorov, A., Mularoni, L., Cope, L. M., Medvedeva, Y., Mironov, A. A., Makeev, V. J., & Wheelan, S. J. (2012). Exploring Massive, Genome Scale Datasets with the GenometriCorr Package. *PLOS Computational Biology*, *8*(5), e1002529. https://doi.org/10.1371/journal.pcbi.1002529

Ferrer-Admetlla, A., Leuenberger, C., Jensen, J. D., & Wegmann, D. (2016). An Approximate Markov Model for the Wright–Fisher Diffusion and Its Application to Time Series Data. *Genetics*, *203*(2), 831–846. https://doi.org/10.1534/genetics.115.184598

François, O., Martins, H., Caye, K., & Schoville, S. D. (2016). Controlling false discoveries in genome scans for selection. *Molecular Ecology*, *25*(2), 454–469. https://doi.org/10.1111/mec.13513

Frichot, E., François, O., & O'Meara, B. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, *6*(8), 925–929. https://doi.org/10.1111/2041-210X.12382

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and Efficient Estimation of Individual Ancestry Coefficients. *Genetics*, *196*(4), 973–983. https://doi.org/10.1534/genetics.113.160572

Grabherr, M. G., Russell, P., Meyer, M., Mauceli, E., Alföldi, J., Di Palma, F., & Lindblad-Toh, K. (2010). Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*, *26*(9), 1145–1151. https://doi.org/10.1093/bioinformatics/btq102

Grant, P. R., & Grant, B. R. (2002). Unpredictable Evolution in a 30-Year Study of Darwin's Finches. *Science*, *296*(5568), 707–711. https://doi.org/10.1126/science.1070315

Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., … Sabeti, P. C. (2010). A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. *Science*, *327*(5967), 883–886. https://doi.org/10.1126/science.1183863

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*, *5*(10), e1000695. https://doi.org/10.1371/journal.pgen.1000695

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*(15), 1965–1978. https://doi.org/10.1002/joc.1276

Hoffberg, S. L., Kieran, T. J., Catchen, J. M., Devault, A., Faircloth, B. C., Mauricio, R., & Glenn, T. C. (2016). RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Molecular Ecology Resources*, *16*(5), 1264–1278. https://doi.org/10.1111/1755-0998.12566

Ignatiadis, N., Klaus, B., Zaugg, J. B., & Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, *13*(7), 577–580. https://doi.org/10.1038/nmeth.3885

Jacobson, E. R., Barker, D. G., Barker, T. M., Mauldin, R., Avery, M. L., Engeman, R., & Secor, S. (2012). Environmental temperatures, physiology and behavior limit the range expansion of invasive Burmese pythons in southeastern USA. *Integrative Zoology*, *7*(3), 271–285. https://doi.org/10.1111/j.1749-4877.2012.00306.x

Lee, C. E. (2002). Evolutionary genetics of invasive species. *Trends in Ecology & Evolution*, *17*(8), 386–391. https://doi.org/10.1016/S0169-5347(02)02554-5

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lignot, J.-H., Helmstetter, C., & Secor, S. M. (2005). Postprandial morphological response of the intestinal epithelium of the Burmese python (*Python molurus*). *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *141*(3), 280–291. https://doi.org/10.1016/j.cbpb.2005.05.005

Losos, J. B., Warheitt, K. I., & Schoener, T. W. (1997). Adaptive differentiation following experimental island colonization in *Anolis* lizards. *Nature*, *387*(6628), 70–73. https://doi.org/10.1038/387070a0

Lotterhos, K. E., Card, D. C., Schaal, S. M., Wang, L., Collins, C., & Verity, B. (2017). Composite measures of selection can improve the signal-to-noise ratio in genome scans. *Methods in Ecology and Evolution*, *8*(6), 717–727. https://doi.org/10.1111/2041-210X.12774

Ma, Y., Ding, X., Qanbari, S., Weigend, S., Zhang, Q., & Simianer, H. (2015). Properties of different selection signature statistics and a new strategy for combining them. *Heredity*, *115*(5), 426–436. https://doi.org/10.1038/hdy.2015.42

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *In Proceedings National Institute of Science, India*, *2*(1), 49–55.

Mazzotti, F. J., Cherkiss, M. S., Hart, K. M., Snow, R. W., Rochford, M. R., Dorcas, M. E., & Reed, R. N. (2011). Cold-induced mortality of invasive Burmese pythons in south Florida. *Biological Invasions*, *13*(1), 143–151. https://doi.org/10.1007/s10530-010-9797-5

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–4297. https://doi.org/10.1093/nar/gks042

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Meshaka, W. E., Loftus, W. F., & Steiner, T. (2000). The herpetofauna of Everglades National Park. *Florida Scientist*, *63*(2), 84–103.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for *De Novo* SNP Discovery and Genotyping in Model and Non-Model Species. *PLOS ONE*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Phillips, B. L., Brown, G. P., Webb, J. K., & Shine, R. (2006). Invasion and the evolution of speed in toads. *Nature*, *439*(7078), 803. https://doi.org/10.1038/439803a

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Pyron, R. A., Burbrink, F. T., & Guiher, T. J. (2008). Claims of Potential Expansion throughout the U.S. by Invasive Python Species Are Contradicted by Ecological Niche Models. *PLOS ONE*, *3*(8), e2931. https://doi.org/10.1371/journal.pone.0002931

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Randhawa, Imtiaz A. S., Khatkar, M. S., Thomson, P. C., & Raadsma, H. W. (2015). Composite Selection Signals for Complex Traits Exemplified Through Bovine Stature Using Multibreed Cohorts of European and African *Bos taurus*. *G3: Genes, Genomes, Genetics*, *5*(7), 1391–1401. https://doi.org/10.1534/g3.115.017772

Randhawa, Imtiaz Ahmed Sajid, Khatkar, M. S., Thomson, P. C., & Raadsma, H. W. (2014). Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genetics*, *15*, 34. https://doi.org/10.1186/1471-2156-15-34

Reed, R. N. (2005). An Ecological Risk Assessment of Nonnative Boas and Pythons as Potentially Invasive Species in the United States. *Risk Analysis*, *25*(3), 753–766. https://doi.org/10.1111/j.1539-6924.2005.00621.x

Reid, N. M., Proestou, D. A., Clark, B. W., Warren, W. C., Colbourne, J. K., Shaw, J. R., … Whitehead, A. (2016). The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science*, *354*(6317), 1305–1308. https://doi.org/10.1126/science.aah4993

Reznick, D. N., & Ghalambor, C. K. (2001). The population ecology of contemporary adaptations: what empirical studies reveal about the conditions that promote adaptive evolution. *Genetica*, *112–113*(1), 183–198. https://doi.org/10.1023/A:1013352109042

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*, R25. https://doi.org/10.1186/gb-2010-11-3-r25

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., … Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods*, *9*(7), 676–682. https://doi.org/10.1038/nmeth.2019

Schindelin, J., Rueden, C. T., Hiner, M. C., & Eliceiri, K. W. (2015). The ImageJ ecosystem: An open platform for biomedical image analysis. *Molecular Reproduction and Development*, *82*(7–8), 518–529. https://doi.org/10.1002/mrd.22489

Schoener, T. W. (2011). The Newest Synthesis: Understanding the Interplay of Evolutionary and Ecological Dynamics. *Science*, *331*(6016), 426–429. https://doi.org/10.1126/science.1193954

Secor, S. M. (2008). Digestive physiology of the Burmese python: broad regulation of integrated performance. *Journal of Experimental Biology*, *211*(24), 3767–3774. https://doi.org/10.1242/jeb.023754

Secor, S. M., & Diamond, J. (1995). Adaptive responses to feeding in Burmese pythons: pay before pumping. *Journal of Experimental Biology*, *198*(6), 1313–1325.

Secor, S. M., & Diamond, J. (1998). A vertebrate model of extreme physiological regulation. *Nature*, *395*(6703), 659–662. https://doi.org/10.1038/27131

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Smith, H. T., Sementelli, A. J., Meshaka, W. E., & Engeman, R. M. (2007). Reptilian Pathogens of the Florida Everglades: The Associated Costs of Burmese Pythons. *Endangered Species Update*, *24*(3), 63–71.

Snow, R. W., Brien, M. L., Cherkiss, M. S., Wilkins, L., & Mazzotti, F. J. (2007). Dietary habits of the Burmese python, *Python molurus bivittatus*, in Everglades National Park, Florida. *Herpetological Bulletin*, *101*, 5–7.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595.

Utsunomiya, Y. T., O'Brien, A. M. P., Sonstegard, T. S., Tassell, C. P. V., Carmo, A. S. do, Mészáros, G., … Garcia, J. F. (2013). Detecting Loci under Recent Positive Selection in

Dairy and Beef Cattle by Combining Different Genome-Wide Scan Methods. *PLOS ONE*, *8*(5), e64280. https://doi.org/10.1371/journal.pone.0064280

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc. https://doi.org/10.1002/0471250953.bi1110s43

Verity, R., Collins, C., Card, D. C., Schaal, S. M., Wang, L., & Lotterhos, K. E. (2017). minotaur: A platform for the analysis and visualization of multivariate results from genome scans with R Shiny. *Molecular Ecology Resources*, *17*(1), 33–43. https://doi.org/10.1111/1755-0998.12579

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Weng, M.-P., & Liao, B.-Y. (2010). MamPhEA: a web tool for mammalian phenotype enrichment analysis. *Bioinformatics*, *26*(17), 2212–2213. https://doi.org/10.1093/bioinformatics/btq359

Willson, J. D., Dorcas, M. E., & Snow, R. W. (2011). Identifying plausible scenarios for the establishment of invasive Burmese pythons (*Python molurus*) in Southern Florida. *Biological Invasions*, *13*(7), 1493–1504. https://doi.org/10.1007/s10530-010-9908-3

Zhang, B., Kirov, S., & Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*, *33*(suppl_2), W741–W748. https://doi.org/10.1093/nar/gki475