2017

GENOMIC PERSPECTIVES ON AMPHIBIAN EVOLUTION ACROSS MULTIPLE PHYLOGENETIC SCALES

Paul Michael Hime
University of Kentucky, paul.hime@uky.edu
Digital Object Identifier: https://doi.org/10.13023/ETD.2017.284

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Paul Michael Hime, Student

Dr. David W. Weisrock, Major Professor

Dr. David F. Westneat, Director of Graduate Studies

GENOMIC PERSPECTIVES ON AMPHIBIAN EVOLUTION ACROSS MULTIPLE PHYLOGENETIC SCALES

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the College of Arts and Sciences at the
University of Kentucky

By

Paul Michael Hime

Lexington, Kentucky

Director: Dr. David W. Weisrock, Professor of Biology

Lexington, Kentucky

2017

ABSTRACT OF DISSERTATION

GENOMIC PERSPECTIVES ON AMPHIBIAN EVOLUTION ACROSS MULTIPLE PHYLOGENETIC SCALES

Genomes provide windows into the evolutionary histories of species. The recent accessibility of genome-scale data in non-model organisms and the proliferation of powerful statistical models are now providing unprecedented opportunities to uncover evolutionary relationships and to test hypotheses about the processes that generate and maintain biodiversity. This dissertation work reveals shallow-scale species boundaries and population genetic structure in two imperiled groups of salamanders and demonstrates that the number and information content of genomic regions used in species delimitation exert strong effects on the resulting inferences. Genome scans are employed to test hypotheses about the mechanisms of genetic sex determination in cryptobranchid salamanders, suggesting a conserved system of female heterogamety in this group. At much deeper scales, phylogenetic analyses of hundreds of protein-coding genes across all major amphibian lineages are employed to reveal the backbone topology and evolutionary timescales of the amphibian tree of life, suggesting a new set of hypotheses for relationships among extant amphibians. Yet, genomic data on their own are no panacea for the thorniest questions in evolutionary biology, and this work also demonstrates the power of a model testing framework to dissect support for different phylogenetic and population genetic hypotheses across different regions of the genome.

KEYWORDS: Population Genomics, Molecular Evolution, Species Delimitation, Gene Tree - Species Tree Discordance, Lissamphibia

Paul M. Hime

24 July, 2017

GENOMIC PERSPECTIVES ON AMPHIBIAN EVOLUTION ACROSS MULTIPLE PHYLOGENETIC SCALES

By

Paul Michael Hime

Dr. David W. Weisrock

Director of Dissertation

Dr. David F. Westneat

Director of Graduate Studies

24 July, 2017

DEDICATION

This work is dedicated to my parents, Michael Stanley Hime and the late Elizabeth Ann Davis Hime, and to my brother, Keith Isaac Hime, for instilling in me a deep curiosity about the natural world; to my children, Charlotte Elizabeth Hime, Genevieve Nicole Hime, and Natalie Andrea Hime, with whom I am honored to share my curiosity about the natural world; and especially to my wife, Nicole Meredith Ratner Hime, without whose unwavering love, patience, and support none of this would have been remotely possible.

ACKNOWLEDGMENTS

First and foremost, incalculable gratitude and praise are due to my incredible wife, Nicole Hime, who has made immense personal and professional sacrifices for me to pursue this graduate degree. In addition to being a phenomenally dedicated stay-at-home mother with our three young children, holding down multiple jobs from home to help support our family, and temporarily deferring many of her own aspirations, Nicole has been a constant source of inspiration, positivity, and love for me.

I sincerely thank my Ph.D. advisor, Dr. David W. Weisrock, for taking a big chance on me and for believing in a zookeeper who talked a big game about trying to use genomics to understand amphibian evolution. I also owe an immense debt of gratitude to my undergraduate advisor, Dr. Jonathan B. Losos, for his support and guidance throughout the years. I am also indebted to my doctoral committee members, Dr.'s Catherine R. Linnen, David F. Westneat, and Christopher L. Schardl at the University of Kentucky, and to my external examiner, Dr. Scott V. Edwards of Harvard University, for their shepherding, encouragement, and pointed critiques. I am also grateful for influential interactions during my undergraduate training with Richard E. Glor, D. Luke Mahler, Thomas J. Sanger, Joshua S. Reece, and Matthew E. Gifford, who each went out of their way to encourage and challenge me at critical moments. All of these scholars embody the characteristics to which I aspire as a scientist and I would not be where I am today without them.

I also wish to thank the current and past members of the Weisrock-Linnen "SuperLab" at the University of Kentucky for setting the bar really high, and for their support and camaraderie over these past six years. Jeramiah J. Smith and his lab at the University of Kentucky also deserve many thanks for their mentorship, computational

hand-holding, and for granting access to large numbers of CPU nodes on short notice. Additionally, I have been very fortunate to interact with and learn from the researchers at the UK Center for Computational Sciences, in particular, Mr. Vikram Gazula. When I entered graduate school I was a complete novice at computation, and the members of the UK CCS have been patient teachers and mentors while I learned to leverage supercomputing to address questions in the biological sciences. I also thank J. Alexander, J. Bridgeman, S. Clarke, A. A. Corea, W. J. Evans, K. Fareed, R. Glasper, H. Hancock, J. Haynes, T. Delvon Jones, T. S. Monk, J. Owens, E. R. Powell, J. Scofield, E. E. Spalding, M. Tyner, L. White III, and V. L. Wooten for constant inspiration.

Although this dissertation is primarily my own work, four of these six chapters (Chapters 2-5) involve key collaborations with numerous collaborators (beyond my advisor, Dr. David W. Weisrock) whom I must acknowledge. Without the contributions of these individuals, none of these projects would have been possible in their present forms. These four chapters also benefitted greatly from the contributions of many individuals and institutions who provided tissue samples, genetic data, locality information, logistical support, and intellectual input. I also thank the University of Kentucky Information Technology Department and Center for Computational Sciences for computing time on the Lipscomb High Performance Computing Cluster. All work was conducted in accordance with applicable institutional guidelines for animal welfare under IACUC protocols SLZ-2009-04 and SLZ-2010-07 to myself, UKY-2012-0952 to Dr. David W. Weisrock, and UKY-2013-1073 to Dr. Steven J. Price.

In Chapter 2, I thank my collaborators Scott Hotaling, Richard Grewelle, Eric O'Neill, Brad Shaffer, and Randal Voss. I also thank Brian and Shonna Storz for their

efforts in the field to sample *A. ordinarium* across its range. This study was greatly improved through discussions with participants of a Species Delimitation Workshop at the National Institute of Mathematical and Biological Synthesis, and through additional discussions with Laura Kubatko and Tara Pelletier. The design and implementation of PHRAPL analyses were greatly aided through discussions with PHRAPL workshop participants at the Ohio State University, and Bryan Carstens and Brian O'Meara. I thank the University of Kentucky Center for Computational Sciences and Lipscomb High Performance Computing Cluster, as well as Jeramiah Smith, for access to supercomputing resources. I also thank Sebastian Voitel for providing a photograph of *A. ordinarium.* I also thank Bryan Carstens and two anonymous reviewers for valuable comments that improved the final manuscript at *Molecular Ecology* (2016).

In Chapter 3, I thank my collaborators Shem Unger, Steve Price, Jeff Briggler, Michael Freake, Amy McMillan, Lori Williams, Andrea Drayer, Mary Foley, Dale McGinnity, John Groves, and Emily Lemmon. For collecting and sharing crucial tissues, I am most grateful to Eric Chapman, Robin Foster, Sean Graham, Joe Greathouse, Kirsten Hecht, Obed Hernandez-Gomez, Bill Hopkins, Kelly Irwin, Cathy Bodinof Jachowski, Steve Kimble, Greg Lipps, Peter Potokas, Brad Shaffer, and Rod Williams. I also thank Mary Duncan, Jeff Ettling, Jane Merkle, Sarah O'Brien, Chawna Schuette, and Mark Wanner at the St. Louis Zoo; Katharine Hope at the National Zoo; Randy Junge at the Columbus Zoo; Diane Barber at the Fort Worth Zoo; Freeland Dunker at the California Academy of Sciences Steinhart Aquarium; Sheena Feist at the Mississippi Museum of Natural Science; James Godwin at the Auburn University Museum of Natural History; and

the late Marcy Sieggreen at the Detroit Zoo for providing tissue samples from their collections.

In Chapter 4, I thank my collaborators Jeff Briggler and Josh Reece. I am very grateful to the St. Louis Zoological Park (Mary Duncan, Jeffrey Ettling, Randy Junge, Jane Merkle, Sarah O'Brian, Amanda Salb, Chawna Schuette, Mark Wanner, Martha Weber), Rod Williams at Purdue University, Freeland Dunker at the California Academy of Sciences, Katharine Hope at the National Zoo, Diane Barber at the Fort Worth Zoo, and Kelly Irwin at the Arkansas Game and Fish Commission for invaluable access to tissue samples and/or necropsy reports. I also thank the Missouri Department of Conservation and the Indiana Department of Natural Resources for permission to utilize samples. This study benefitted from discussions with Schyler Nunziata, Jeramiah Smith, Catherine Linnen, Tony Gamble, Todd Castoe, Amy McMillan, Rod Williams, and Allan Larson.

In Chapter 5, I thank my collaborators Alan Lemmon, Emily Lemmon, Elizabeth Scott-Prendini, Chris Raxworthy, Jeremy Brown, Bob Thomson, Michelle Kortyna, Pedro Peloso, Brice Noonan, Scott Keogh, Steve Donnellan, Justin Kratovil, Alex Pyron, Krushnamegh Kunte, Sandeep Das, Nikhil Gaitonde, Santiago Ron, Jim Labisko, Rachel Mueller, and David Green. I also thank the following institutions and individuals for providing access to critical tissues samples: American Museum of Natural History (Darrel Frost, David Kizirian, Julie Feinstein); California Academy of Sciences (David Blackburn, Jens Vindum); Florida Museum of Natural History (Pamela Soltis); Kansas State Natural History Museum (Rafe Brown, Linda Trueb, Andrew Campbell); Louisiana State Museum of Natural History (Robb Brumfield, Donna Dittmann); Museum of Comparative Zoology (Jim Hanken, José Rosado, Breda Zimkus); Museum of Vertebrate Zoology (Jim McGuire,

Carol Spencer, Ted Papenfuss, Marvalee Wake, Sima Bouzid); Museum Victoria (Jane Melville, Joanna Sumner); National Museum of Natural History (Kevin De Queiroz, Addison Wynn); South African National Biodiversity Institute (Zoe Davids); Saint Louis Zoological Park (Jeffrey Ettling, Mark Wanner, Randall Junge); University of Michigan Museum of Natural History (Ronald A. Nussbaum, Gregory Schneider); Yale Peabody Museum (Gregory Watkins-Colwell); as well as J.J. Apodaca, Alan Channing, Becky Chong, Guarino Colli, Tyler Frye, S. Blair Hedges, Elizabeth Jockusch, Jarrett Johnson, Christopher McNamara, Eric O'Neill, Todd Pierson, Steve Richards, and Kelly Zamudio.

This research was supported financially by grants from: the Gertrude F. Ribble Endowment at the University of Kentucky, the Cryptobranchid Interest Group, the Society of Systematic Biologists Graduate Research Award, the Kentucky Academy of Sciences, the Kentucky Science and Engineering Foundation, the St. Louis Zoo's WildCare Institute, the Fresno Chaffee Zoo, Doctoral Dissertation Improvement Grant from the National Science Foundation (DEB-1601586), and research grants to my advisor from the National Science Foundation (DEB-0949532 and DEB-1355000). During my time at the University of Kentucky, I have been generously supported by a Graduate Research Fellowship from the National Science Foundation (award 3048109801) and by a Blue Waters Graduate Research Fellowship, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

Introduction.

ABSTRACT

Organisms' genomes reflect their evolutionary history and can be used to estimate the relationships among species. In this dissertation I investigate amphibian evolution across multiple phylogenetic scales, from the early stages of speciation in Mexican and North American aquatic salamanders (Chapters 2 and 3, respectively), to the evolution of sex-linked genes across a deeply divergent family of salamanders (Chapter 4), to the divergences among and within the major family-level amphibian orders (Chapter 5). In these chapters, I address fundamental questions about the sources, magnitude, and downstream effects of varying, and sometimes conflicting, phylogenetic signals from across the nuclear genome. Each of these four empirical chapters seeks to test hypotheses about aspects of evolutionary biology in particular organismal systems, and each chapter brings some sort of "genomic" data to bear on these questions. In three of these four chapters, these genomic resources were developed from scratch specifically for the taxa and questions at hand, requiring non-trivial amounts of effort to optimize and deploy these new systems of data collection. Yet, these data are merely tools with which to investigate pressing applied and basic evolutionary questions in non-model species, and beyond the organismal foci of some of these chapters, the more general and unifying themes of this

body of work revolve around issues of model adequacy in phylogenetics and the quantification of information content for different regions of the genome.

INTRODUCTION

This is a tremendously exciting time to be an evolutionary biologist. Advances in high throughput sequencing for non-model organisms (e.g., Lemmon & Lemmon 2013; Peterson *et al*. 2012) and emerging statistical models for phylogenetic reconstruction (e.g., Mirarab & Warnow 2015; Stamatakis 2014) and species delimitation (Rannala & Yang 2013) are providing unprecedented opportunities to unlock the mysteries of the tree of life. Yet, as systematists push these boundaries, it is becoming increasingly apparent that genomic data and standard analyses, on their own, are no panacea for the thorniest problems in evolutionary biology (e.g. Brown *et al*. 2016; Reddy *et al*. 2016). Accordingly, my broad research aims are to unify the fields of genomics and phylogenetics and to work to transform this deluge of data into novel evolutionary insights about how species form and how genomes evolve.

One of the most tantalizing features of life on Earth is that all organisms trace their origins back to a single common ancestor nearly four billion years ago (Baum *et al*. 2016). Yet today, life has diversified into a panoply of tens of millions of species (Mora *et al*. 2011). Accurately reconstructing these evolutionary relationships is the primary aim of phylogenetics, and such insights may inform nearly all areas of modern biology (e.g., Barraclough & Nee 2001; Castro-Nallar *et al*. 2012; Glor 2010; Tong *et al*. 2015). The genomics revolution is now radically altering the historically data-limited field of

molecular phylogenetics; genome-scale data are now almost trivial to generate for any organism (e.g., Faircloth *et al*. 2012; Lemmon *et al*. 2012; Lemmon & Lemmon 2013). Yet, phylogenetic inferences are only as accurate as their underlying model assumptions are appropriate, and as the complexity of genomic data sets grows, so also does the need to rigorously assess the information content landscape across data sets, as well as to scrutinize potentially hidden sources of spurious support or conflicting signals.

My doctoral dissertation research seeks to reconstruct the evolutionary relationships among organisms to better understand the origins and maintenance of this biodiversity, and to test hypotheses about the processes that generate the patterns of biodiversity we observe in the natural world. All life forms each possess genetic material which orchestrates the ways in which they are constructed and which also evolves through time, along the way, documenting many of the evolutionary events in the lifespans of species and populations. Historically, the DNA sequence data from which phylogeneticists reconstruct evolutionary relationships has represented only a relatively small portion of the genome. In fact, until recently, mitochondrial DNA (mtDNA) comprised the vast majority of genetic loci collected for phylogenetics and phylogeography (Avise *et al*. 1987). However, we now know that different regions of the genome may, under fairly common conditions, be expected to each tell different versions of this shared evolutionary history (Maddison 1997). The long-standing reliance in phylogenetics on a small sampling of the genome may not accurately reflect the true evolutionary origins of organisms. More genetic data should, in theory, produce more reliable estimates of evolutionary relationships as well as more credible downstream inferences (Holder & Lewis 2003). Yet, DNA sequences have traditionally been very difficult to obtain *en masse*. Recently, advances in genome

sequencing technologies (e.g., Bentley *et al*. 2008) have opened up exciting new avenues for phylogeneticists to survey broad swaths of the genome and to untangle some of the most difficult branches in the Tree of Life.

Fundamentally, we need two things to answer evolutionary questions: sufficiently informative data, and adequate models to describe those data (Lewis 2016; Yang & Rannala 2012). The data themselves are not of primary interest, but rather it is the information which those data bring to bear on the question at hand (Akaike 1974; Shannon 1948), and the ability our models to leverage that information towards our questions which is what we seek (Burnham & Anderson 2003). "Data" do not categorically equate to information (in the sense of anything that decreases our uncertainty about the question at hand), except in the context of an adequate model. The quest to answer questions about speciation, or genome evolution, or macroevolutionary patterns of diversification, very much hinges upon quantifying and comparing the information content of available data, and then using those data to discriminate among a set of competing models based on the difference in information content between them. The multispecies coalescent model provides a powerful framework within which to test hypotheses at both shallow (Fujita *et al*. 2012) and deeper (Edwards *et al*. 2016) scales.

Historically, the vast majority of phylogenetic studies have been conducted using concatenation, a method by which multiple genomic loci are consolidated into a "supermatrix" and analyzed under either a single model of molecular evolution or under partitioning schemes which account for substitutional heterogeneity across the concatenated alignment. But with the realization that different regions of the genome are expected to be discordant from each other (Degnan & Rosenberg 2006; Maddison 1997),

the validity of concatenation-based approaches in particular situations where gene tree - species tree discordance is likely (such as for rapid divergences and/or large effective population sizes) was called into question (e.g., Kubatko & Degnan 2007).

Although so-called "shortcut" coalescent methods which attempt to estimate species trees from collections of inferred gene trees (e.g., Chaudhary *et al*. 2014; Liu & Pearl 2007; Liu *et al* 2010; Mirarab & Warnow 2015) have been critiqued (Springer & Gatesy 2016) for model violations or for not accounting for the full range of causes underlying gene tree variation beyond incomplete lineage sorting, these approaches offer powerful means by which to efficiently investigate phylogeny in systems with large numbers of species and large numbers of genetic loci (Edwards *et al*. 2016).

Even three decades ago Thorne *et al*. (1991) recognized that systematists' abilities to generate sequence data were rapidly exceeding their abilities to appropriately analyze those data (and this was at a time when sequencing single loci was state-of-the-art). This was true even in that bygone single-gene era, and is especially salient now that hundreds or thousands of loci can be readily sequenced in as many taxa for (relatively) modest investments of money and time (Glenn 2011). There is an inherent temptation to assume that any murky portions of phylogenies, any recalcitrant branches, and any parameter estimates with huge variances will be "resolved" by taking a genomic perspective. Many studies have demonstrated the increased resolution which can come with increasing numbers of loci for population genetics and phylogenetic reconstruction. One expects noise and murkiness under coalescent models, even with perfectly known gene trees with no error whatsoever, there will be cases where different genes have different genealogical histories from each other, and from the species tree (Maddison 1997). This was a wake-

up call that using gene trees as a proxy for species trees may not always be wise, if the aim is to understand the species tree itself (Edwards 2009). Interests shifted towards estimating species trees, and this is a primary occupation of the majority of phylogeneticists today. But others (e.g. Hahn & Nakhleh 2015) have discussed the importance of considering not only the relationships implied as the branching order among lineages of organisms in the species tree, but also the gene-level histories as well. As phylogeneticists continue to push the boundaries of increased taxonomic sampling and the numbers of loci sequenced for those taxa, the need to rigorously assess the dependence of inferences on the underlying data and models will take on increased importance.

OVERVIEW OF EMPIRICAL CHAPTERS

In the first empirical chapter of my dissertation, I address whether the information content and the number of loci used in species delimitation studies influence the potential to detect and validate shallow-scale divergence between populations and species. Using a group of endangered Mexican stream salamanders (*Ambystoma ordinarium*) as a test case, my co-authors and I demonstrated that the balance between sampling large numbers of individuals for a few loci versus sampling fewer individuals for greater numbers of loci should be tipped in a direction dictated by the expected levels of divergence in the focal group. We further show that while a few, relatively high-information loci can resolve deeper evolutionary histories just as well as large numbers of less informative loci, the latter case may be more informative about more recent divergence events. This study also addresses pressing applied conservation questions in this endemic and endangered

salamander, and suggests that a cryptic species exists in the western portion of the current range.

A second major component of my dissertation research has grown out of conservation genetic research I initiated while at the St. Louis Zoo. I worked to design a battery of genomic resources for the endangered hellbender salamander (*Cryptobranchus*) which have led to several revolutionary insights into cryptobranchid salamander evolution and conservation. I designed and optimized a reduced representation genome sequencing protocol which has yielded ~75,000 anonymous loci (~605,000 single-nucleotide polymorphisms) across tens of millions of years of divergence with very low rates of missing data. By applying recent advances in Bayesian species delimitation and coalescent approaches for species tree estimation, this work has revealed unexpectedly high levels of microendemism in this freshwater salamander and has identified several deeply divergent, cryptic species for which my co-authors and I are in the process of working to validate and formally describe. Results from this work support recognition of several reproductively isolated species of hellbenders which are broadly aligned with the major continental watersheds of eastern North America. These species boundaries have significant implications for regional and range-wide conservation and management of *Cryptobranchus*. For instance, several states actually host more than one species of hellbender, implying that current management strategies aligned along state boundaries, and not watershed boundaries, may warrant reconsideration.

In a third portion of my dissertation, I leveraged the genomic markers that I developed to study species boundaries in hellbenders to identify putatively sex-linked regions of the genome and to test hypotheses about the genetic sex determination system

in this group. I developed the first genetic sex diagnostic for a non-model salamander by using reduced representation genome sequencing across known males and females to identify putatively sex-linked loci which I validated by PCR. This work confirmed that this salamander family has a conserved system of female-heterogametic sex determination and allowed the development of a universally effective PCR-based assay for sex in several species of conservation concern. These discoveries have significant implications for cryptobranchid conservation because, previously, it was very difficult to reliably distinguish males from females in the wild or in captive assurance populations due to delayed sexual maturity and a narrow annual time window of morphological distinctiveness. This genomic research also informs ongoing theoretical work into the dynamics of vertebrate sex chromosome evolution and provides an intriguing system wherein the W-linked chromosomal regions I identified appear to have been conserved over ~60 million years of independent evolution, in contrast to the rapid degradation of sex-limited chromosomes observed in other taxa. This project is nearing the manuscript submission stage, and I am aiming to distribute a draft to co-authors for review in the near future. This method has already been put to use in an independent research laboratory which has confirmed that it is highly effective and which is already using the genetic sex assay described in this chapter to inform *in situ* and *ex situ* conservation initiatives with hellbenders.

In the final component of my dissertation, I led a large-scale collaborative initiative to resolve the contentious relationships among extant amphibians. My co-authors and I sequenced genomes and transcriptomes across a dozen representative amphibians to design an amphibian-specific sequence capture system that is effective across this entire vertebrate

class. Sampling ~300 nuclear exons and complete mitochondrial genomes for nearly 300 species of amphibians (representing 97% of families and over 50% of amphibian genera), we have produced comprehensive species tree hypotheses for extant amphibians and identified extensive gene-tree/species-tree conflict throughout even the deepest branches of the amphibian phylogeny. These results clarify several murky portions of the amphibian tree and are providing fresh insights into the timescale of amphibian evolution. But more broadly, this study addresses fundamental questions about the sources and magnitude of phylogenetic signal across large multi-gene data sets. I advocate that an information theoretic framework may help systematists to parse informative signal from noise in this new era of too much data and potentially insufficient models.

CHAPTER TWO

The influence of locus number and information content on species delimitation: an empirical test case in an endangered Mexican salamander

ABSTRACT

Perhaps the most important recent advance in species delimitation has been the development of model-based approaches to objectively diagnose species diversity from genetic data. Additionally, the growing accessibility of next-generation sequence datasets provides powerful insights into genome-wide patterns of divergence during speciation. However, applying complex models to large datasets is time consuming and computationally costly, requiring careful consideration of the influence of both individual and population sampling, as well as the number and informativeness of loci on species delimitation conclusions. Here, we investigated how locus number and information content affect species delimitation results for an endangered Mexican salamander species, *Ambystoma ordinarium*. We compared results for an eight-locus, 137-individual dataset and an 89-locus, seven-individual dataset. For both datasets, we used species discovery methods to define delimitation models and species validation methods to rigorously test these hypotheses. We also used integrated demographic model selection tools to choose among delimitation models, while accounting for gene flow. Our results indicate that while cryptic lineages may be delimited with relatively few loci, sampling larger numbers of loci may be required to ensure that enough informative loci are available to accurately identify

and validate shallow-scale divergences. These analyses highlight the importance of striking a balance between dense sampling of loci and individuals, particularly in shallowly-diverged lineages. They also confirm the presence of a currently unrecognized, endangered species in the western part of *A. ordinarium*'s range.

INTRODUCTION

An important recent advance in molecular systematics has been the development of refined evolutionary models and new analytical approaches for delimiting species using multi-locus DNA sequence data (Fujita *et al.* 2012). One of the most important decisions required by these new species delimitation methods involves the trade-off between the numbers of loci and individuals that should be sampled to accurately identify cryptic species lineages. Increased sampling of loci provides more precise and accurate inferences of population-level parameters (Harris *et al.* 2014), while increased sampling of individuals more completely captures variation within and between populations. Historically, phylogeographic investigations have relied on sampling genetic data from relatively small numbers of loci, often from many individuals (e.g., hundreds). However, population genetic theory demonstrates that variation among gene histories can be enormous, often dwarfing variation among individuals, suggesting that increased gene sampling is key to accurate inferences (Irwin 2002). Now, advances in next-generation sequencing (NGS) are providing the opportunity to sample large numbers of loci (e.g., hundreds or thousands), although pragmatic (cost) considerations may limit the number of individuals that can be sampled. In practice, this results in a spectrum of potential data sets, ranging from small

numbers of loci sampled from many individuals (e.g., Hotaling *et al.* 2016) to NGS-scale datasets sampled from fewer specimens (e.g., Rittmeyer & Austin 2015; Smith *et al.* 2014). To date, few studies have examined the influence of gene and individual sampling density on the detection of lineage divergence events. Similarly, an exploration of the information content of loci versus the number of loci needed to detect lineage divergence events would be informative in guiding future delimitation studies that use NGS-scale data.

Equally important to the accurate delimitation of species from multi-locus datasets is the use of analyses that account for the population-level forces structuring genetic variation within and between populations. Methods testing species-divergence hypotheses using coalescent models have recently grown to include information-theoretic (Ence & Carstens 2011), Bayesian (Grummer *et al.* 2014; Yang & Rannala 2010), and approximate Bayesian (Camargo *et al.* 2012) frameworks. While these methods have the ability to explicitly incorporate population size and divergence time parameters into species delimitation tests, they do not allow for gene flow and its potential influence in structuring genetic variation. Alternatively, a number of methods explicitly estimate gene flow as a model parameter, permitting researchers to gauge the extent to which hypothesized lineages exchange genes. However, current methods either assess gene flow independently of divergence (e.g., Migrate-n; Beerli & Felsenstein 2001), or as part of a model that includes a divergence time parameter but does not assess alternative models lacking a history of divergence, which may offer closer fits to the data (e.g., IMa2; Hey 2010). Furthermore, extreme values of population genetic parameters (divergence, population size and migration rates) can all exert an influence of expected patterns of gene coalescence and affect species delimitation results (Rannala 2015). For example, small effective population

sizes could make populations appear more diverged than they are. Consequently, species delimitation studies that aim to incorporate all of these parameters into decisions about lineage divergence must use multiple methods to estimate different, and partially overlapping, sets of population genetic parameters relevant to the process of speciation (e.g., Jonsson *et al.* 2014; Reid *et al.* 2014).

Here, we use an eight nuclear locus data set (referred to as the "8L" data set) sampled from a large number of individuals (n = 137) and an 89 nuclear locus data set (referred to as the "89L" data set) sampled from a small number of individuals (n = 7) to empirically quantify the effects of gene and sampling density on species delimitation outcomes. Both data sets were generated from populations across the range of a narrowly-distributed stream-dwelling salamander species, *Ambystoma ordinarium.* A member of the tiger salamander complex (Shaffer and McKnight, 1996), *A. ordinarium* has previously been diagnosed as a genealogically exclusive species based on patterns of monophyly in reconstructed gene trees from the 8L data set (Weisrock *et al.* 2006). The species has a small, ~120 x 20 km range across high elevations (>2200 m) of the Mesa Central in the western Trans Mexican Volcanic Belt (TMVB), a region characterized by rugged terrain created through a history of volcanism and tectonic uplift (Figure 2.1). Furthermore, its range includes streams that flow into multiple independent drainage systems (Anderson & Worthington 1971), providing an opportunity for hydrologically mediated lineage divergence in allopatry and potential speciation, despite a small geographic scale.

Using these two different data sets, we explore patterns of lineage divergence within *A. ordinarium*, and in the process, address the impact of locus sampling and individual sampling on the delimitation of cryptically diverging lineages. We specifically

investigate the impact of both the number and information content of loci on our results by using random subsamples of our data and by using increasing numbers of loci ordered by their phylogenetic information content. As part of this process, we also implement a novel approach towards species delimitation using a recently developed species-tree reconstruction method, SVDQuartets (Chifman & Kubatko 2014). This method is restricted to a coalescent model that does not currently parameterize gene flow, but which scales well to multi-locus sequence data. Drawing on this species delimitation work, we then extend our species delimitation tests through model selection across models that simultaneously consider divergence and gene flow.

METHODS AND MATERIALS

8L sequence data

The 8L data used here are those published in Weisrock *et al.* (2006), with the exception that the mtDNA data were excluded from all analyses given the clear signatures of mitochondrial DNA (mtDNA) introgression involving *A. ordinarium* and other *A. tigrinum* complex species (Weisrock *et al.* 2006). Analyses were focused on phased Sanger DNA sequence data from eight nuclear loci (*col1a1*, *dlx3*, *ctg1506*, *ctg1908*, *g1d6*, *g1f1*, *g1c12*, and *g3d7*) generated from 217 paedomorphic or young larval specimens of *A. ordinarium* sampled from 20 geographic localities distributed across the known species range (Figure 2.1; Table 2.1). We used this 217-individual data set for all population structure analyses. For all subsequent species delimitation tests, we reduced the total 8L

sequence data set down to 137 individuals that contained no missing data for all eight loci. Sequence data from a single *A. tigrinum melanostictum* individual was used as an outgroup. Measures of genetic variation within *A. ordinarium*, including numbers of haplotypes, substitution and indel variation, and estimates of heterozygosity were previously reported in Weisrock *et al.* (2006), and are included in Table 2.2. Nuclear DNA sequence data for all 217 *A. ordinarium* individuals are available in GenBank (*col1a1*: DQ252580-252937; *ctg1506*, DQ252938-253365; *ctg1908*, DQ254388-254797; *dlx3*, DQ248436-248859; *g1c12*, DQ254798-255197; *g1d6*; DQ255356-255783; *g1f1*, DQ253450-253837; *g3d7*, DQ253924-254301).

89L data

The *A. ordinarium* NGS-scale data used here are a subset of the loci published in O'Neill *et al.* (2013) as part of a larger study of the *A. tigrinum* complex. These data were generated via parallel tagged amplicon sequencing (PTAS) on a Roche 454 sequencing platform. Full details of marker selection, field sampling, and sequence generation can be found in O'Neill *et al.* (2013). For this study, phased sequence data from seven *A. ordinarium* individuals and one *A. tigrinum melanostictum* individual were extracted from this larger data set. The *A. ordinarium* individuals represent samples from seven different localities broadly covering the geographic range. Of the 95 total loci in the original study, 81 contained one or more variable sites and had data present for at least five *A. ordinarium* individuals and the *A. t. melanostictum* outgroup. We combined these 81 loci with the 8L

loci (using the same set of individuals present in the PTAS data) to produce the 89L data set used here (details of the 81 additional loci are provided in Table 2.3).

Species hypothesis generation - population structure

Genetically-based species delimitation is typically a multipart process (Carstens *et al.* 2013), beginning with a discovery phase aimed at identifying hypothetical lineages and assigning sampled individuals to these lineages. Inaccuracy in this 'discovery phase' can lead to biased downstream results (Edwards & Knowles 2014; Olave *et al.* 2014; but see Zhang *et al.* 2014). Therefore, we utilized several discovery approaches to identify consensus patterns of population structure. For the full 217-individual 8L data set, we used the program Structure (Pritchard *et al.* 2000) to estimate the relative assignments of individuals into K populations (allowing for admixture) for all K from 1 to 20. We used plots of the log probability of the data [ln Pr(X|K)] and ΔK (Evanno *et al.* 2005) to determine the optimal number of population clusters and to detect hierarchical structure, respectively. We also used the program Structurama (Huelsenbeck *et al.* 2011), treating K as a random variable, although results were largely concordant with those from Structure. Finally, we also used the program SplitsTree (Huson & Bryant 2006) to assess the degree to which haplotypes were shared among versus within putative clusters. Additional details of Structure, and full details for the Structurama and SplitsTree analyses are provided in Figure 2.4 and Figure 2.6, respectively.

Delimitation hypothesis testing - SpedeSTEM

Using lineage hypotheses for *A. ordinarium* from the discovery phase, we tested a two-lineage hypothesis, which splits *A. ordinarium* into western and eastern lineages, and a four-lineage hypothesis, which further subdivides the western clade into two lineages (WE1 and WE2) and the eastern clade into two lineages (EA1 and EA2). The two-lineage model roughly corresponds to separate stream drainages on either side of a mountain ridge, and although the four-lineage model does not correspond to any known geographic barriers, its utility is supported by results of the discovery phase. We tested these hypotheses using an information-theoretic model selection approach implemented in SpedeSTEM v0.9.5 (Ence & Carstens 2011). SpedeSTEM uses point estimates of gene trees, but considers all or many of the possible lumpings and splittings of hypothesized lineages.

Attempts to perform SpedeSTEM analyses on the 89L data presented computational challenges; therefore, we focused our analyses on the 8L data. For both lineage hypotheses, we performed tip subsampling using two, three, five, and 10 alleles from each lineage in reconstructed gene trees, with 1000 replicates each. For the two-lineage hypothesis, it was also computationally feasible to sample 25 tips from each lineage for 100 replicates. Likelihood scores were averaged across replicates and Akaike information criterion (AIC) scores and relative model probabilities were calculated for each species tree model.

Delimitation hypothesis testing - BPP

We also used the program BPP3 (Yang & Rannala 2010) to test our two- and four-lineage species delimitation hypotheses. Although this popular method no longer requires an *a priori* guide tree, we chose to utilize the fixed tree topology (Figure 2.1) which was strongly supported by our discovery approaches. BPP uses reversible-jump Markov chain Monte Carlo sampling to compare nested species models by collapsing or failing to collapse nodes in the user-specified guide tree. In doing so, BPP calculates posterior probabilities (PPs) of those nodes in the guide tree and the relative model probability of competing delimitation models.

We used BPP v3.0 (Yang & Rannala 2014) and BPP v3.1, respectively, to analyze the 8L and 89L data sets using nine combinations of priors for θ and τ, corresponding to small (Gamma distribution set with $\alpha = 2$, $\beta = 1000$), medium (2, 100) or large (2, 10) population sizes, and shallow (2, 1000), intermediate (2, 100), or deep (2, 10) divergence times. In addition, we performed a set of analyses using priors that reflected empirical estimations of θ (3, 1250) and τ (25, 1149), which were most similar to the small population size and shallow divergence time prior.

Delimitation hypothesis testing - SVDQuartets

We applied a recently developed species tree reconstruction method using the program SVDQuartets (Chifman & Kubatko 2014) implemented in PAUP v4.0a146 (Swofford 2015). For these analyses, we analyzed our 89L data as a 43-locus subset (referred to as the "43L" data set), which contained complete data sampling across all seven *A. ordinarium* individuals and the *A. t. melanostictum* outgroup. We analyzed our 43L data

two ways, each taking a different approach for assessing support for our species delimitation hypotheses. First, we estimated a "lineage tree" following Chifman and Kubatko (2014), where tips in the tree represent the random pairing of gene copies across loci for a diploid individual. Here, SVDQuartets can provide support for species divergence without *a priori* identification of species hypotheses; branches separating populations that are part of the same species are not expected to be reconstructed with high branch support in the lineage tree. This approach may be viewed as a species discovery approach. Secondly, we performed analyses using a five-tip species tree model that corresponded to putative species (WE1, WE2, EA1, and EA2) within *A. ordinarium,* and an *A. t. melanostictum* outgroup. We treated these analyses as a species validation test, where placement of tips into their expected clades with high bootstrap support is interpreted as evidence for species-level entities. For our 8L data, we performed SVDQuartets analyses solely within a species tree framework, as analysis of a lineage tree using 137 individuals proved computationally intractable. For all SVDQuartets analyses, we performed exhaustive sampling of all possible quartets (every combination of four tips was examined). Branch support for the inferred trees was estimated using 100 non-parametric bootstrap replicates.

Investigating the role of data scale and content on species delimitation

Using our 89L data, we explored the degree to which the amount of sequence data influenced the results of coalescent-based species delimitation. We focused these analyses on BPP and also explored the effects of the number of sampled loci and their phylogenetic

information content on delimitation inferences. To address the influence of the number of loci, we generated nine data sets varying in the number of sampled loci by increments of 10 (10, 20, 30, etc.) up to 89 loci. For each subsampling increment, we generated 10 replicate jackknifed data sets, each using random locus sampling (89L data set analyzed only once). To examine the influence of information content, we ranked the 89 loci in order of their parsimony-informative sites (based on *A. ordinarium* and the *A. t. melanostictum* outgroup). We again generated a series of nine data set sizes (10 through 89) that increased (starting with the most informative loci) in the number of sampled loci by increments of 10. Due to the ordered nature of these data sets, only a single round of analysis was performed for each data set.

We applied these two data assembly strategies to three different sets of species hypotheses within *A. ordinarium*: (1) testing the split between western and eastern lineages, (2) testing the split between two western lineages (WE1 and WE2), and (3) testing the split between two eastern lineages (EA1 and EA2). All analyses were performed using empirically estimated priors and the same run conditions mentioned above.

Phylogeographic model selection and parameter estimation – PHRAPL

We extended our species delimitation tests to include model-selection based phylogeographic inference implemented in PHRAPL (O'Meara *et al.* 2015). PHRAPL employs a heuristic exploration of model space to define a set of the most plausible models given a set of gene trees estimated from multilocus sequence data. Using empirically estimated gene trees and a maximum number of free model parameters, PHRAPL uses

approximate likelihoods to infer the model (or models) best supported by the data. This method uses simulated gene trees [generated in the program ms (Hudson 2002)] and compares the simulated gene tree topologies to those from empirical data. Over many replicates, the number of exact matches can be used to calculate likelihoods across a wide range of pre-defined models. AIC is then used to rank models and to identify the best-fitting model. This method has a natural fit to the process of species delimitation as it allows for the assessment of divergence models that also include migration parameters, providing a more complete assessment of the divergence history of a set of hypothesized species.

For both the 8L and 43L data sets, we analyzed five possible models for a two-lineage scenario with one or two free parameters (Figure 2.2). Model parameters included divergence time (τ), and rates and direction of migration (m). Bidirectional m were constrained to be symmetric. For each of the 8L loci, we partitioned individuals into western and eastern lineages and randomly sampled six gene copies per lineage across 50 replicates, increasing the probability of sampling each gene copy at least once. We defined all five models corresponding to a two-lineage, two parameter scenario with one free parameter for m and τ, respectively. Following O'Meara *et al.* (2015), we conducted grid searches across τ and/or m reflecting arbitrary (but realistic) values (τ = 0.3, 0.58, 1.11, 2.12, 4.07, 7.81, 15.0; m = 0.10, 0.22, 0.46, 1.00, 2.15, 4.64, 10.0).

For each combination of parameters for each model, we simulated 100,000 balanced 12-tip gene trees in ms and compared the topologies of the observed (subsampled) empirical gene trees to the simulated gene trees. We sought exact topological matches with the caveat that the labeling of individuals drawn from the same population was arbitrary. We defined the approximate likelihood of a given model with a given set of parameter

values to be equal to the number of matches between the empirical and simulated trees divided by the number of replicates. Log likelihoods of models were summed across loci and an AIC score was defined as -2 x ln(L(model$_i$|Data)) + 2K, where K is the number of free parameters in a given model. We computed model likelihoods for each model and final model selection was performed by ranking models by increasing AIC and observing the plot of ΔAIC across models. PHRAPL analysis of the 43L data followed the same approach described above, except that five gene copies from both the western and eastern lineage were sampled, and fewer replicates (five) were needed to ensure that all gene copies were sampled from all gene trees at least once.

For all PHRAPL analyses, we also explored models including five or fewer free parameters which allowed for changes in effective population size and asymmetric migration. However, PHRAPL was unable to discriminate among this larger set of models as evidenced by low ΔAIC (< 0.25) values, likely due to an insufficient number of loci and/or variable sites per locus, and these more complex models were not considered further. To generate model averaged estimates of parameters for each data set, we calculated the likelihood-weighted arithmetic mean of each parameter across all models using the CalculateModelAverages function in PHRAPL.

Demographic model selection and parameter estimation – Migrate-n

We used Migrate-n v3.6 (Beerli 2006) to estimate gene flow under a coalescent framework for a range of two-population models using the 8L data as limited individual sampling in the 89L data set precluded its use. We tested: (1) a 'panmixia' model treating

eastern and western lineages as a single population, (2) a two-population model with bidirectional gene flow, (3) a two-population model with unidirectional gene flow from the western lineage into the eastern lineage, and (4) a two- population model with no migration (Figure 2.3). Initial parameter values were calculated using F_{ST} and we employed model averaging to estimate migration rate (m) and θ.

RESULTS

Data summary

The 8L data set contained a total of 4,176 nucleotide sites. Including the *A. t. melanostictum* outgroup, the number of parsimony-informative sites (PIS) across loci ranged from 4 to 25, with a mean of 14.9. Within *A. ordinarium* the number of PIS across loci ranged from 3 to 25, with a mean of 13.9. The 81 PTAS loci from O'Neill *et al.* (2013) contained a total of 20,006 nucleotide sites. Including the *A. t. melanostictum* outgroup, PIS across loci ranged from 0 to 12, with a mean of 4.06. Within *A. ordinarium* PIS across loci ranged from 0 to 7 with a mean of 1.07 (Tables 2.2 and 2.3). Maximum likelihood gene trees for the 8L are included in Figure 2.5.

Population structure and hypothesis generation

Analysis of the 8L data using Structure resulted in a Δ*K* that supported a $K = 2$ model separating western and eastern populations of *A. ordinarium* with low levels of

admixture (Figure 2.1). Separate analyses on each of these groups identified a $K = 2$ level of population structure within each (hereafter referred to as WE1 and WE2 across western populations and EA1 and EA2 across eastern populations). In some cases, further Structure analysis within these groups suggested additional population structure, but with high degrees of admixture, and these clusters were not explored further. Population structure results were not method dependent and additional population clustering results for Structure, Structurama (Figure 2.5), and SplitsTree (Figure 2.6) are provided.

Delimitation hypothesis testing - SpedeSTEM

For the two-lineage model, SpedeSTEM analysis of the 8L data only supported divergence between western and eastern lineages when 25 alleles were sampled. Under a four-lineage model, significant divergence was detected when sampling five or ten gene copies (Table 2.4); however, divergence was restricted to the splitting of WE2 from all other hypothesized lineages (WE1, EA1, and EA2). In addition, we tested models that fixed divergence between western and eastern lineages, and then tested for divergence within either the western (WE1 and WE2) or eastern lineage (EA1 and EA2). In both cases, SpedeSTEM supported models that lacked divergence within western and eastern lineages (Table 2.5).

Delimitation hypothesis testing - BPP

BPP analysis of the 8L data produced strong support (PPs = 1.0) for divergence between western and eastern lineages across all combinations of priors for Θ and τ (Figure 2.7A). There was no difference in delimitation results between algorithms 0 and 1. Analyses using randomized tip labeling produced low posterior support for divergence between the western and eastern lineage, indicating that results were not biased by our choice of priors (Figure 2.7B). Support for divergence between WE1 and WE2 varied across prior combinations. Small and intermediate population size priors produced strong support for divergence (PPs = 1.0), regardless of divergence time prior, while larger population size priors weakly supported divergence between WE1 and WE2. Divergence between EA1 and EA2 received weak support across all prior combinations (Figure 2.7A).

BPP analysis of the 89L data set produced PPs = 1.0 for the split between the western and eastern lineages under all prior combinations (Figure 2.7C). Randomized tip labeling generally yielded low PP support for the western-eastern split (Figure 2.7D). For intermediate and large population size priors, posterior support varied for the EA1-EA2 and WE1-WE2 splits (Fig 3C). Across replicates, support ranged from as little as PP = 0 to PP = 1 for these divergence events. Prior combinations featuring small population size (including our empirical-based priors) produced PPs close to one for both the EA1-EA2 and WE1-WE2 splits. Under all prior combinations, randomized tip labeling produced PPs close to 0 for these splits.

Delimitation hypothesis testing - SVDQuartets

SVDQuartets analysis of the 43L data resulted in a lineage tree with a strongly supported split between the western and eastern lineages (Figure 2.8A), with each forming a separate clade of haplotypes with high bootstrap support (> 99%). The WE1 and WE2 splits were similarly well-supported in the lineage tree, with bootstrap values of 92.6% and 99.4%, respectively. In contrast, the EA1 and EA2 groups were not resolved as reciprocally monophyletic (Figure 2.8A). Branches within the eastern lineage generally received low bootstrap support. SVDQuartets analyses of the 8L and 43L data using a species tree framework generated a tree supporting the split between the western and eastern lineages with high levels of bootstrap support (Fig 4B; 8L: 96% and 99%, respectively; 43L: 100% for both lineages).

Influence of data scale and content

Using BPP, we achieved strong PP support for divergence between the western and eastern lineage with as few as 10 of our 89L loci (Figure 2.9A) with minimal variation across replicates. When sampling 20 or more loci, support for this divergence received PPs = 1.0 across all replicates. For more shallow divergences, a greater effect of locus sampling was detected. Strong support for divergence between WE1 and WE2 was detected with as few as 30 loci (Figure 2.9B); however, not all replicates provided strong support for this split, with at least one 30-locus replicate yielding PP = 0.05. This large difference in maximum and minimum posterior support persisted with increasing numbers of sampled loci, with a mean PP $\geq$ 0.95 achieved with 80 loci (minimum PP = 0.83). Similar results were obtained for analyses of the EA1-EA2 divergence (Figure 2.9C). High levels of

posterior support for divergence were produced with as few as 10 sampled loci (maximum PP = 0.94); however, large differences in PPs were detected across replicates, a pattern observed for most levels of locus subsampling. A total of 70 loci were required to produce a mean PP ≥ 0.95 (minimum PP = 0.83). Overall, for these shallower divergences, variance in support across replicates decreased with greater locus sampling.

Analysis of 89L loci with the highest number of PIS had a strong effect on support for the more shallow divergence events. Whereas at least 80 randomly sampled 89L loci were needed to produce mean PP ≥ 0.95 for the WE1-WE2 split, only 40 of the most informative 89L loci were needed to produce a similar level of support (Figure 2.9D). As few as 30 of the most informative 89L loci produced a PP > 0.95 for the WE1-WE2 split. A similar pattern was observed for the EA1-EA2 split, which required as many as 70 randomly sampled 89L loci to produce a mean PP ≥ 0.95, but which required only 50 of the most informative 89L loci to produce the same level of support (Figure 2.9D). For the western-eastern split, as few as 10 of the most informative 89L loci produced posterior support of 1.0.

Phylogeographic model selection and parameter estimation – PHRAPL

PHRAPL analysis of the 8L data resulted in the greatest support for model 4 (Figure 2.2), which specified divergence between the western and eastern lineages, along with gene flow from the eastern lineage into the western lineage (Table 2.6; model probability = 0.67). A model treating *A. ordinarium* as a single lineage received the next highest support (ΔAIC = 1.77, model probability = 0.28), while support for distinct western and eastern

lineages with no gene flow was lowest (ΔAIC = 35.21, model probability = 1.51 x 10^{-8}). Model averaged parameter estimates suggest a relatively deep divergence with low-level gene flow from western populations into eastern populations and relatively high post-divergence gene flow from eastern lineages into western lineages (Table 2.6).

PHRAPL analysis of the 43L data indicated the greatest support for a model of divergence between the western and eastern lineage with no gene flow (Table 2.6; model probability = 0.63). The next best-supported model was one of no divergence between eastern and western lineages (ΔAIC = 3.24, model probability = 0.12). Model averaged parameter estimates suggest a relatively deep divergence with near zero post-divergence gene flow in either direction between western and eastern populations (Table 2.6).

Demographic model selection and parameter estimation – Migrate-n

Migrate-n analysis of the 8L data best supported a bidirectional migration model between the western and eastern lineages (Table 2.7; model 2, model probability > 0.99). The next best model included unidirectional gene flow from the eastern lineage into the western lineage; however, this model received a very low probability (log Bayes factor ≥ 81.6, model probability = 1.9 x 10^{-18}). A model combining the western and eastern populations into a single population (model 1) received the lowest model probability (3.5 x 10^{-105}).

Estimates of the number of migrants per generation were significantly skewed towards migration from the western into the eastern lineage (Table 2.8), with a mean N_m

of 0.44 (95% confidence interval = 0-1.29) in this direction versus a mean N_m of 0.16 (95% confidence interval = 0-0.77) for the opposite direction.

DISCUSSION

Data sampling in species delimitation

Species delimitation is in a state of transition in terms of the data used for analysis, with systematists facing important choices regarding the numbers of loci and individuals to sample. Recent studies have investigated the role of locus number and gene copy sampling in the performance of genetically-based species delimitation methods (e.g., Camargo *et al.* 2012; Hird *et al.* 2010). However, no empirical study has compared the influence of locus number and information content on species delimitation results. One general conclusion from this study is that both small and large data sets have the potential to resolve cryptic species boundaries between recently diverged species. Across data sets, species discovery methods (e.g., Structure) highlight the same candidate species, and species validation tests (e.g., BPP, PHRAPL) provide similarly strong support for the western-eastern divergence event within *A. ordinarium*. These results are, in part, encouraging for the broader molecular systematic community, suggesting that large-scale data sets, for example those generated with NGS methods, may not always be necessary for the delimitation of morphologically cryptic species. This may be particularly true for older and well-differentiated species, where species delimitation is expected to be straightforward (Shaffer & Thomson 2007). Within *A. ordinarium*, the split between the

western and eastern lineage was recovered with strong posterior support in all BPP analyses of the 8L data and in every subsampled 10-locus 89L data set (Figure 2.7A). Furthermore, given that all subsampled 10-locus 89L data sets resulted in strong support for the western-eastern split, and that these loci were drawn from a pool of loci with a wide range of variability, for smaller data sets it may not be as important to make highly informed choices about which loci to use in the delimitation of deeper divergence events.

In contrast, our investigations of information content (i.e., PIS sites) show that all loci are not equal in their ability to recover signatures of shallower divergence events. In the case of the WE1-WE2 and EA1-EA2 splits, while at least one 30-locus data set provided strong support for these divergences, other 30-locus data sets did not (Figure 2.9B-C). While this discrepancy was also observed in increasingly larger data sets, support variance declined as locus number increased. There is a wide range of phylogenetic information across our 89L loci, with the 20 most-informative loci having approximately the same total number of PISs (n = 214) as the next 50 (n = 184; Figure 2.9D), and this range in information content is likely driving the large swing in support for these more shallow divergences when randomly selecting loci. For example, the mean posterior support for the WE1-WE2 split from a randomly sampled 30-locus data set was ~0.4 (Figure 2.9B), while the 30 most-informative loci produced posterior support of 0.94 (Figure 2.9D). Similar to conclusions derived from studies focused on factors influencing species tree reconstruction at shallow tree depths (Harris *et al.* 2014; Huang *et al.* 2010; Knowles *et al.* 2012; Lanier *et al.* 2014), our results indicate that the phylogenetic information content of loci is a primary factor in the delimitation of species separated by shallow depths of divergence.

Overall, our results provide a mixed message to the systematist considering how to generate data for a species delimitation study. A small number of loci may be sufficient to both discover and validate many cryptic species, allowing researchers with the ability to generate relatively small data sets to continue the identification of new lineages and taxa. However, this work shows that many shallowly diverged species may go undiscovered, or not pass statistical validation via coalescent tests, when using these smaller data sets. While some cryptic species may only require a small number of loci to be detected, this is impossible to know *a priori*, and as a result, systematists are most likely to be assured of clarifying the boundary between cryptic species and structured populations when analyzing large multilocus data sets. With the growing accessibility of genome-scale data sets for phylogeography and species delimitation (e.g., Lemmon & Lemmon 2012; Smith *et al.* 2014), an increasing number of studies are likely to include sufficient numbers of loci for drawing boundaries between intraspecific and interspecific variation. However, in the case of salamanders, large genome size (often > 30 Gb) may preclude genome-scale data generation for species delimitation using standard NGS approaches, though recent advances (e.g., McCartney-Melstead *et al.* 2016) may reduce this bottleneck in the future.

Phylogeographic model selection and species delimitation

To date, most coalescent-based species delimitation studies have been restricted to the parameterization of N_e and divergence time in models meant to capture the history of populations. Yet, data that include signatures of gene flow are likely to have impacts on likelihood calculations, with potentially important ramifications for accurate species

delimitation. A few studies (Jackson & Austin 2012; Leache 2009; Ruane *et al.* 2014) have examined the impacts of gene flow on species tree reconstruction at the phylogeographic level, demonstrating the importance of considering the effects of introgression and low-level gene flow. Here, we extended these efforts by using a model selection approach in PHRAPL to consider gene flow as a parameter in our species validation tests. In doing so, we rejected the hypothesis that *A. ordinarium* represents a single lineage. Our 8L and 43L data sets contrasted, however, in support for a history of gene flow between lineages, with the 8L data favoring a model that included unidirectional gene flow and the 43L data favoring a model of no gene flow. This difference may be related to the much greater individual sampling of the 8L data, which may have included individuals bearing the signatures of gene flow. Indeed, the 8L data, but not the 43L data, contained sampling from localities 7 and 8 of the eastern lineage, both of which contain individuals with signatures of admixture within the western lineage (Figure 2.1B). In addition, the 8L loci are longer and more variable, on average, than the 81 PTAS loci, which may make them more informative in detecting historical, low-level gene flow. Finally, it is worth noting that estimates of gene flow between these two lineages are low, with Migrate-n estimates of Nm much less than one (Table 2.8). In any case, the consideration of both gene flow and divergence using a model-selection approach are concordant with results from other analyses in supporting divergence between the western and eastern lineages.

We note, however, that our model selection approach faced substantial limitations in the complexity of models and the numbers of free parameters that could be considered. For example, we were unable to confidently perform analyses with additional parameters that allowed for changing N_e through time or asymmetric migration rates. Similarly,

expanding PHRAPL analyses to account for models with three or four lineages, which also required an expanded set of free parameters, proved challenging. While our data appear to be informative in recovering the deeper western-eastern divergence event while accounting for gene flow, testing more complex models that take these additional parameters into account would likely require increased sampling of both individuals and loci to produce credible estimates of parameters such as migration rates and effective population sizes, and accordingly, to distinguish more complex models of gene flow and divergence from each other.

Species boundaries

Collectively, our genetic results strongly support eastern and western populations of *A. ordinarium* as independently evolving population-level lineages and we diagnose these as distinct species (see Appendix 1). Population structure results strongly delineate these as separate clusters with limited evidence for admixture, and the 8L SVDQuartets lineage tree reconstructs the western and eastern populations as two strongly supported clades. Coalescent-based tests using BPP validate this hypothesis: divergence between western and eastern populations was strongly supported with both our 8L and 89L data across all explored prior combinations and across different subsets of loci. Beyond genetic evidence, these two species also occur in separate headwater systems of southward-flowing streams in the TMVB (Figure 2.1).

Within both diagnosed species, support for additional levels of lineage divergence varied markedly across analyses, and we refrain from diagnosing additional population-

level lineages within the western and eastern species. This conclusion is principally derived from the inconsistent patterns of support in BPP validation analyses for both data sets and varied prior combinations (Figure 2.7). This was particularly true for divergence within the eastern lineage, which was poorly supported under all priors in analyses of the 8L data, and only received high posterior support in analyses of the 89L data that featured small population size priors. Similar patterns of inconsistent support in BPP validation tests were also seen within the western lineage. In addition, a subset of SpedeSTEM results split WE2 while lumping all other hypothesized lineages, in contrast to all species discovery results, which found a clear division between eastern and western populations. This latter result could indicate an inapplicability of SpedeSTEM to this particular study (Camargo *et al.* 2012; Carstens *et al.* 2013). However, while population structure is clearly evident within the eastern and western lineages, given the lack of consistent support for the delimitation of additional lineages, and with an aim to not promote taxonomic instability (e.g., Turtle Taxonomy Working Group 2007), we do not describe additional species-level taxa within the western and eastern lineages.

Evidence for lineage divergence within A. ordinarium

Though we do not have a divergence time estimate for the split between the western and eastern lineages, we expect that it does not coincide with the geological evolution of the TMVB. The majority of tectonic and volcanic activity producing the TMVB occurred in the mid-Miocene, approximately 7-11 million years ago (Ferrari *et al.* 1999; Ferrari *et al.* 2000), likely predating the common ancestor of the entire *A. tigrinum* species complex

(Shaffer & McKnight 1996). Phylogeographic studies of birds (McCormack *et al.* 2008), lizards (Zarza *et al.* 2008), and toads (Mulcahy *et al.* 2006) support the role of the TMVB in Pliocene-Pleistocene species divergence, a time of active, but less extreme uplift (Ferrusquía-Villafranca & González-Guzmán 2005). Tectonic and volcanic activity in the TMVB also substantially changed its hydrology over time (Israde-Alcantara & Garduno-Monroy 1999), leading to the divergence of fish species (Doadrio & Dominguez 2004; Dominguez-Dominguez *et al.* 2008; Hulsey *et al.* 2004; Mateos *et al.* 2002; Schönhuth & Doadrio 2003). Many of these species divergences have been dated to the Pliocene with the most recent described divergence among them occurring 0.6-0.8 Ma between two allopatric species of the genus *Allotoca* found in Lakes Patzcuaro and Zirahuen (Dominguez-Dominguez *et al.* 2006).

The Late Pleistocene in central Mexico was characterized by cooler and drier conditions (Metcalfe *et al.* 2000) and palynological data indicate an absence of pine forest across upper elevations (>2500 m) of much of central Mexico at this time (Lozano-Garcia & Vazquez-Selem 2005). In contrast, pollen studies of Lago Patzcuaro, which is at a lower elevation than contemporary *A. ordinarium* populations (~2000 m), reveal stable pine forest over the last 48,000 years, indicating that lower elevations of central Mexico were not as strongly impacted by drier conditions (Bradbury 2000). Given this environmental history, it is possible that *A. ordinarium* populations, which are facultative in their ability to metamorphose, tracked the movement of available pine forest into lower elevations during the late Pleistocene and into the Holocene, and that these distributional shifts, perhaps into refugia representative of the current drainage basins occupied by *A. ordinarium*, initiated lineage divergence. The strong signature of recent mtDNA

introgression between the western lineage of *A. ordinarium* and the Lago Patzcuaro endemic paedomorph *A. dumerilii* (Weisrock *et al.* 2006), combined with their current allopatric distribution, further supports this lower elevation refugia hypothesis.

Conservation implications

We anticipate that recognition of an additional, cryptic species within the endangered and range-restricted *A. ordinarium* will have immediate conservation implications for this group of ambystomatid salamanders. *Ambystoma ordinarium* already has an IUCN "Endangered" listing due to its limited distribution and disappearing forest habitat (Shaffer et al. 2004). When a single endangered species is recognized as two, each by necessity has an even more restricted range, and must be considered even more fragile and threatened. As an example in ambystomatid salamander conservation, the discovery and recognition of two cryptic species (*A. cingulatum* and *A. bishopi*) within the former *A. cingulatum* species (Pauly *et al.* 2007), was quickly adopted by the U.S. Fish and Wildlife Service, with both species upgraded to endangered status. As such, the recognition of cryptic and recently diverged species using the methods outlined here may be especially important beyond salamanders in the conservation of biodiversity in recently derived, endangered taxa.

Table 2.1. Sampling information for 8L and 43L/89L data sets.

Locality	8L Structure sampling	8L BPP sampling	43L/89L sampling	Latitude	Longitude	Description
1	15	7	0	19.3700000	-101.3825000	Small stream in Cruz de Plato, ~11 km W (by road) Villa Madero, 0.3 km W of paved road
2	14	8	1	19.3694444	-101.3813889	Small stream in Cruz de Plato, ~11 km W (by road) Villa Madero, 0.3 km W of paved road
3	9	7	1	19.3016667	-101.5150000	Spring-fed stream, in town of El Pedregoso, ~3.5–4 km W of Patzcuaro-Taucomaro Hwy
4	11	7	1	19.3077778	-101.4677778	Large stream passing under paved road, 10.2 km (by road), E of San Gregorio
5	10	7	0	19.6172222	-101.1241667	10 km SSE (straight line) of Morelia, S of San Miguel del Monte, in N flowing creek
6	16	11	1	19.5872222	-101.1286111	12.5 km SSE (straight line) of Morelia, in SW flowing stream
7	7	5	0	19.5327778	-101.1516667	18.75 km S (straight line) of Morelia, in S flowing creek
8	9	6	0	19.5590000	-101.1627000	17.5 km S (straight line) of Morelia, in town of Las Palomas, in WNW flowing stream
9	9	7	0	19.5597222	-101.1372222	~16.75 km S (straight line) of Morelia, in E flowing stream
10	10	9	0	19.6480556	-101.0158333	Small stream, 12.6 km E (by road), then 2.4 km S (by road) of Morelia at town of Pino Real
11	10	5	0	19.6680000	-100.8660000	SW flowing stream, S of Hwy. 15, 0.4 km W of San Jose Lagunillas between Morelia and Ciudad Hidalgo
12	10	6	0	19.7540000	-100.7480000	Small S flowing stream, 12 km S of Hwy. 126 and 51 intersection, then 2.6 km E of Hwy 51 on dirt road
13	10	6	0	19.6719444	-100.7400000	Small stream ~0.2 km N of intersection of Hwy. 51 and 15, where it crosses under Hwy. 51
14	10	8	0	19.6677778	-100.7022222	Small E flowing stream, 4.7 km E of intersection of Hwy 15 and 51
15	10	9	1	19.6666667	-100.6833333	Small S flowing stream, 10.7 km E of intersection of Hwy 51 and 15
16	10	5	1	19.5483333	-100.6183333	14.7 km S (by road) of Ciudad Hidalgo in small stream just E of road
17	15	5	0	19.5700000	-100.6161111	12.6 km S (by road) Ciudad Hidalgo, in E flowing stream
18	12	3	0	19.6160000	-100.6170000	7.8 km S (by road) Ciudad Hidalgo, in main N flowing stream
19	10	9	1	19.6658333	-101.0052778	Small N flowing stream 23 km E (by road) of Morelia (Jose Maria Morelos Parque Nacional)
20	10	6	0	19.5094444	-100.7544444	Small SW flowing stream 1.7 km S (by road) of San Antonio Villalongin

Table 2.2. Summary statistics for all loci included in the 8L data set.

Locus	Source	Length (bp)	Variable Sites ord. + mel.	PI Sites ord. + mel.	Variable Sites ord. + mel.	PI Sites ord. + mel.	Per Locus Φ (ord. only)	Per Site Φ (ord. only)	π (ord. only)	Unique Haplotypes
COL1A1	Weisrock *et al.* 2006	705	12	12	5	5	0.9434	0.001342	0.001204	4
CTG1506	Weisrock *et al.* 2006	277	8	8	4	4	0.9434	0.003406	0.00484	3
CTG1908	Weisrock *et al.* 2006	503	23	22	2	1	0.3311	0.000664	0.000334	2
DLX3	Weisrock *et al.* 2006	150	4	3	3	3	0.6289	0.004193	0.002711	3
G1C12	Weisrock *et al.* 2006	1059	25	23	9	6	1.5723	0.001606	0.001987	4
G1D6	Weisrock *et al.* 2006	309	7	6	4	2	0.6289	0.002042	0.00289	3
G1F1	Weisrock *et al.* 2006	403	15	12	9	9	2.2012	0.005476	0.009103	3
G3D7	Weisrock *et al.* 2006	774	25	25	13	13	3.1445	0.004255	0.003926	8

Table 2.3. Summary statistics of the 89 loci included in the 43L and 89L data sets.

Locus	Source	Length (bp)	PI Sites*	PI Sites †	Per-Locus Φ†	π†	Num. *A. ordinarium*	Unique Haplotypes	*Ambystoma* Linkage Group	Primer 1	Primer 2
COL1A1	Weisrock *et al*. 2006	703	9	2	0.943400	0.001204	7	4	11	CACCGAAGC CTCCCAAAA CATCAC	GAGCCCTTCC ATCTTAGTCGT
CTG1506	Weisrock *et al*. 2006	277	7	3	0.943400	0.004840	7	3	11	AGGATATCC GCTCAGAAA TATGAAG	CTGACCACTTG CAAAACTTAC TACCT
CTG1908	Weisrock *et al*. 2006	503	8	0	0.331100	0.000334	6	2	-	CTCATGACT TAATTGCTG TTCTTCG	ATAACCATTCT GAGGTTTTGA GTTG
DLX3	Weisrock *et al*. 2006	150	1	1	0.628900	0.002711	7	3	11	GGCGAGGC GCACCTCTC CAACTGGTG A	AGGCTCCCAC CTTCTGAGTTG GGAAGG
G1C12	Weisrock *et al*. 2006	1057	29	5	1.572300	0.001987	7	4	2	CCCAAATCC AGGAGTTCA AA	CAAGGCAGCC AAATTATCGT
G1D6	Weisrock *et al*. 2006	309	6	2	0.628900	0.002890	7	3	6	CAGCGTGCC CACCCGATA GAA	TCCCAAAAAG TAAAATGTGC AAAGAAAA
G1F1	Weisrock *et al*. 2006	403	10	7	2.201200	0.009103	7	3	-	TTAGTTTGG GTGCAGACA GGA	GGTGCTCAAC AACAAATCAA CT
G3D7	Weisrock *et al*. 2006	774	21	5	3.144500	0.003926	7	8	-	TCCTTTTCC CCAGTTTGT TG	TATGAAACCC TGCTCCTTGG
CTG355	O'Neill *et al*. 2013	173	2	0	0.000000	0.000000	7	1	3	GTGAAGTCA GTGATGAAA GTCCATGT	CTAGGATACC AGTGGGAGAG TGTAAT
E10A7	O'Neill *et al*. 2013	172	6	0	0.331100	0.000975	6	2	4	AATCCAGCC TAATCCCTA AAGATAAT	CAAACTCTTCA AAACCTATCTC CTTC
E10C11	O'Neill *et al*. 2013	215	1	0	0.331100	0.000786	6	2	4	CAGGAGGA CTGGATCTT CTGG	GTGAGTACAA GCACTTTGGA AGTTAG
E10C5	O'Neill *et al*. 2013	300	2	0	0.331100	0.000921	6	2	2	GAAGGACTT GTTATTCAG GCATATTT	ACGTTTATACA AAGAACTAAA CGCCT
E10C6	O'Neill *et al*. 2013	264	2	0	1.655700	0.003230	6	4	2	TGATAGCTC TTAAAAGAA ACCAGACA	GTAGCTCAAA ATCCATGACA GTAAGA
E11G6	O'Neill *et al*. 2013	216	3	0	0.000000	0.000000	6	1	4	ATGATGATT GAACAAAC AGACACTTT	AAGCAATTAA AACAGTAAAG AAGGGA
E12A3	O'Neill *et al*. 2013	211	5	2	0.662300	0.004753	6	2	3	GCTGGATTG AAACTTCTC TAGTCTCT	CCACCAACTA CTACAATCAA ATCATC
E12A4	O'Neill *et al*. 2013	194	4	1	0.314500	0.002719	7	2	9	AGAACCTGG AGTCTTACA GTACAACA	TACACTTGTTT TGCAGTTAAT AAGGC
E12A9	O'Neill *et al*. 2013	212	1	0	0.000000	0.000000	6	1	8	GGAATGCAT GGATTAAGG ATTTATAC	CTAAACAAAT GTTGTAGGGG AATTTT
E12C11	O'Neill *et al*. 2013	201	2	0	0.000000	0.000000	7	1	14	CCAGCTGTT AAAGTAAA GAAGGAAG T	GTTTTAAAAAT TTCATAAGGC AGCTC
E12C3	O'Neill *et al*. 2013	349	1	1	0.331100	0.000894	6	2	10	TAACGAAA GATGAAGA AAAACAACT G	CATAATTATTT GTAACCGTTG ACGAC
E12C6	O'Neill *et al*. 2013	320	12	4	1.060500	0.002614	5	4	1	AAACTGCAA CATAATGAA AGCCTAC	GAGAGTAGAG AGGTATTTAG GCAACC
E12C7	O'Neill *et al*. 2013	149	1	0	0.000000	0.000000	7	1	4	AGACATTCC TTTAAGAGA TTACTCGG	CCCTTTGTAAA ATAATTCCAA GAAAA
E12E3	O'Neill *et al*. 2013	302	0	0	0.662300	0.001134	6	3	12	GACTGAGG ATCATTTGT TTGTTAATG	GACTTCAGTTT CAAAGTACGT ATCCA
E12G1	O'Neill *et al*. 2013	369	10	0	0.000000	0.000000	7	1	5	CACTGTCAA AACATTGTT AGTTGATG	CTATGACGGTT TACAGCAGTG ACTAT
E12G12	O'Neill *et al*. 2013	134	1	3	0.628900	0.001992	7	3	6	ACGAGATG ACCAACTAT AGGAATGAT	GTAGTATCTTC GTCCTCGTGAT CTTG
E12G2	O'Neill *et al*. 2013	239	4	2	0.331100	0.002184	6	2	6	AGTTATGCA TTGGTTCTT ATGTTCAC	AAACAAAGGA ATGTTTTGAAT GACTT
E12G5	O'Neill *et al*. 2013	371	11	7	1.413900	0.005034	5	5	10	CCTATTCCA CTGCAAGAG TAGTTACA	TTTGAAAATAT TTATTGTACAG GCCA
E13A3	O'Neill *et al*. 2013	262	3	0	0.000000	0.000000	7	1	7	AACATGCTT CTTTTTATG CTTCTTTT	TTACCTTAAAA CACTTATGCCA GATG
E13A6	O'Neill *et al*. 2013	304	1	0	0.314500	0.000521	7	2	13	TGTTTAGGT ATTCTATGC CACTCTTG	ATCTCTAACCT TAAGTGCAAA CCAGT
E13C1	O'Neill *et al*. 2013	215	2	0	0.000000	0.000000	7	1	7	GTGTATGTA ACTTTCCTC AGAGTCCA	ACAGTAGCAC CCTTAGTTAAG CAAAT
E13C7	O'Neill *et al*. 2013	298	9	0	1.257800	0.003572	7	4	6	CAATGTGTA TGAAAGCTG GATGTAAT	CAGAATAGAG CCCTGAAAGT AGAAAG
E13E2	O'Neill *et al*. 2013	419	2	3	0.000000	0.000000	6	1	4	GCATTTTGA GCAGTTATT GTTTTAGT	ACTTAAAATC CCAAGTTCAG AAGCTA
E14A2	O'Neill *et al*. 2013	218	5	4	0.000000	0.000000	7	1	6	CGTTGGTTA ACAGTAACC TCACTAAA	TGCTGTAGGA TTCTCTACTAC AGGTG
E14A8	O'Neill *et al*. 2013	322	4	1	0.314500	0.001460	7	2	8	TAGTTTTTG TAAATGCTT TGATCCAG	AGTAATTTACC CGTTGACGAA AATAC
E14E10	O'Neill *et al*. 2013	177	0	1	0.000000	0.000000	7	1	5	TGAGGACTT CATCTTACA CTCTGAAC	TATATAGCTGC GAGACCACAA AATAC
E14E2	O'Neill *et al*. 2013	208	1	0	0.000000	0.000000	7	1	5	ATCCCGTAT ATCATTCTA AACCATGT	AAAAATATCC CCAACTAATTT CAGTG
E14E3	O'Neill *et al*. 2013	160	3	1	0.314500	0.001648	7	2	6	TGTCTATAA AGCTGGATT TGTTTGTC	ATGATGTACC CTCCACATTAC ACTTA
E14G10	O'Neill *et al*. 2013	363	5	5	0.353500	0.004040	5	2	4	TGGATTAGA ATAAGAGC ATTCAACTG	ACTTTCACAAC AACTAATTTGC TGAC
E14G11	O'Neill *et al*. 2013	284	7	1	0.353500	0.001891	5	2	2	CTTGTTTTG TCTGTTCTT GATGATCT	CCTAATTTCTT CAGGTGATCT GTGTA
E15A2	O'Neill *et al*. 2013	431	9	1	0.993400	0.002596	6	4	2	TAAATCTTT CGCTAATAT CTCCCAGT	TTCACACATTT CAATATATTCC GTCT
E15E12	O'Neill *et al*. 2013	341	8	3	0.993400	0.002867	6	5	5	TTTATGACT GTTGCTGTT TCCTATTC	GGGAAGAGAT TTATTTACAGA AGCTG

Table 2.3 (continued). Summary statistics of the 89 loci included in the 43L and 89L data sets.

E15E2	O'Neill *et al*. 2013	185	4	2	0.662300	0.005270	6	2	3	AATGCCTGC TTAAAGCTA GAACTTAG	CTTTAGACTCCT CCTTTAGCCTGT AA
E15G5	O'Neill *et al*. 2013	432	4	4	0.314500	0.000545	7	2	3	TAAACAGG AATGTACAA GGCACTAAC	GATCCTTCTTCAT AGAATGTCAACA A
E16A12	O'Neill *et al*. 2013	388	6	4	1.655700	0.003725	6	6	1	TGGTATGAT ATTTTGGTG TTCTCCTA	CAGTTGCATTAA AAATGTAGAGGA AA
E16A9	O'Neill *et al*. 2013	287	3	0	0.353500	0.001655	5	2	4	TACCTCTAT TTTTATTCA TGTGCTGC	TCTTCTGGACTA GAACAATGAATC TC
E16C7	O'Neill *et al*. 2013	370	7	3	1.060500	0.003545	5	3	5	GACAGGAG AATGAGTGA GTTACAAAA	AGAAGTGTTTCA ACAGCATTATAT CG
E16G9	O'Neill *et al*. 2013	154	5	3	2.515600	0.010531	7	4	4	CATCATGGG CATATTTTA CTACAAAC	AACTTTATTCGC TGAAACAGTCAG
E17A2	O'Neill *et al*. 2013	225	4	1	0.331100	0.001365	6	2	7	CTATAACAC GTCAATGTC CCAATATC	GGACTTGGATAA AATTTGTCTTTG AT
E17G3	O'Neill *et al*. 2013	211	1	1	0.314500	0.001256	7	2	3	CTAAATCCT AACCATCAC CCTACAAT	CTAGAACTAGGC ATATTGGCCTAA AC
E18C3	O'Neill *et al*. 2013	202	5	1	1.257800	0.006641	7	3	4	GTATTTATA AAGATTTTT GGAGCCGT	CTCATGAGTAGT CAATTGAAGGGG
E18C7	O'Neill *et al*. 2013	229	4	0	1.257800	0.005316	7	5	3	GTCCTGACT AGTTCTTAC CCTGAAAG	GTGTTTGCTATA CTGAGTCCAAAG AA
E18C8	O'Neill *et al*. 2013	154	1	1	0.662300	0.002179	6	3	3	TCAACAAAT TAAACTCCA CTCTGAAC	GATAATAGAATG CTCAATGACTGC AT
E19C7	O'Neill *et al*. 2013	204	3	1	0.707000	0.003573	5	2	1	CAACACTGA TCTTTCTTC ATTCTCTC	TGGTGTTTAAGG TCTTCCTATCTTC T
E19E12	O'Neill *et al*. 2013	246	3	0	0.628900	0.002736	7	3	11	ACATGAATG AAAAGATTA AGGGAAAC	TAGGTAAATGGA ATGTGGATACTG AA
E19E7	O'Neill *et al*. 2013	205	2	0	0.331100	0.000821	6	2	2	ATTGAATTC ACACGTATC CTAAAACA	GTAAATCCTTTC CCCCTTCACTTA AA
E20A8	O'Neill *et al*. 2013	252	1	1	0.000000	0.000000	7	1	5	AAGATGAG GATACCATT GATGTGTT	GCAGAGAAATAA TATTGGTTTTAG CA
E20A9	O'Neill *et al*. 2013	180	3	0	0.000000	0.000000	7	1	2	TTCAATATT TGGGGACA ATGTAGATA	CAAATAACAGAA CAGGTTTTCTTC AA
E20C1	O'Neill *et al*. 2013	216	3	0	1.257800	0.002774	7	3	4	CTAATCAAC TTCATCAAG CAGCAC	GTTGATGTTTAA TTGTGCGAATTT AG
E20C2	O'Neill *et al*. 2013	388	3	0	0.331100	0.001122	6	2	10	ATCATGTTA ATAGTGTAT GTGCGGTT	ATTTACACAGAT TCTGCAGTACAA GG
E20E5	O'Neill *et al*. 2013	340	7	1	0.662300	0.002145	6	3	14	GTTGTTATC ACAAACTAG AGCGGTTA	ATTGTACATCGT ACTATGGTGGTC TC
E20G5	O'Neill *et al*. 2013	331	2	1	0.943400	0.001323	7	2	9	GTGAAGTCT GAAGACTGT GACCTTAG	AAGGCACTCAAA AAGCAAATAAAA TA
E20G6	O'Neill *et al*. 2013	258	7	2	0.943400	0.004670	7	4	2	GTTTGTACA TGGGTGAAG AGTAAGTC	TTAAACAAAGTG TTTGCAAGGTAG AG
E22A7	O'Neill *et al*. 2013	365	2	0	0.353500	0.000592	5	2	1	CAGAGCTGC CGAAGTTGA C	GTCCCCAAACCC CTGTCAT
E23C6	O'Neill *et al*. 2013	187	2	1	0.314500	0.002926	7	2	4	TAAGACCTG CTTCACGTT TTGCTAC	CAAGTAAAGGGT TTCTCTTGTTAAG G
E23G1	O'Neill *et al*. 2013	174	3	0	0.000000	0.000000	7	1	9	GATGAAAAT GAACAACTA AAACAGGA	GACGAAAATCAA TGTCCATTCTACT A
E23G7	O'Neill *et al*. 2013	172	3	2	0.314500	0.002108	7	2	4	GACAATATG ATAAAGAC AGTGATGGC	ACAAGTGATACA CAAAATGTGGAA AT
E24A12	O'Neill *et al*. 2013	380	10	0	1.324600	0.004397	6	3	4	TACTACTGT CCTCACAAC ACATGAAC	AAACAGCTGCAG ATATGTTAAACA AG
E24A6	O'Neill *et al*. 2013	383	12	1	0.662300	0.001599	6	3	10	CAGTATCGT TTAACAGGG CCAG	GTTACTAACCAA TCAAACAGCAAG AA
E24C10	O'Neill *et al*. 2013	355	4	4	0.000000	0.000000	7	1	1	GTGTTTCGA ACTCCTTAT TTCTCATT	GATTTGGATACA GTTGATTTTACG AA
E24E9	O'Neill *et al*. 2013	368	9	0	0.353500	0.000565	5	2	1	AACTGCTGC TAAGACTGC AAATC	GGAATAATCAAC ACAACTAGTCTC CA
E24G3	O'Neill *et al*. 2013	364	1	1	0.000000	0.000000	6	1	6	AAGATGCA GTGCTGTCA AAATATCTA	CAGCCCTACATA AAACCACCAATA
E26G9	O'Neill *et al*. 2013	156	3	0	0.000000	0.000000	6	1	7	TAACTGACT TGACTAACC CCACTATG	GTCCATTGTACA AAGCCTCTATTA AA
E5E6	O'Neill *et al*. 2013	152	3	0	0.353500	0.002339	5	2	3	CTTCTTAAT CATCATATC CTGGCAGT	TGTCTAGGTATG ACGTTGTTTTTCT C
E6A11	O'Neill *et al*. 2013	154	2	0	0.314500	0.000934	7	2	7	ACACTTCCA AACTAAGG AAGAAAGT C	ACACTTCCAAAC TAAGGAAGAAA GTC
E6C2	O'Neill *et al*. 2013	131	1	1	0.000000	0.000000	7	1	3	GTCTAATCA GCTCCGAAA CAATAAAT	GCAATACGTACT GTGACAAAAGAA AT
E6C8	O'Neill *et al*. 2013	220	1	0	0.314500	0.002465	7	2	3	CAAGTAGCT AAATCCTAC ATCAAGCA	TAATTGGAAATT GCAGTCTGAGTT AG
E6E11	O'Neill *et al*. 2013	271	2	1	0.331100	0.000634	6	2	11	AAGAGAAG TTCCTAGAT GAGTTGGAG	TGAAGAGAGAAC TCAAAGTGTCTG AT
E6E7	O'Neill *et al*. 2013	256	10	0	0.662300	0.002782	6	3	11	GGATAGATA CCATGAATC CATTGAG	GTTGGTGCACTA CCTGGAGTAAG
E7A5	O'Neill *et al*. 2013	264	2	2	0.000000	0.000000	5	1	14	CCTCTGTTG GTTAAAGTC TAGTGACC	GTGATATTCCTC TAAAGGGTCCAT AA
E7C12	O'Neill *et al*. 2013	176	2	0	0.000000	0.000000	6	1	1	CACATTCTA AGAAGCTG GTTTTCAAT	ATTCCTGTTTGTA TGCACTGTCAAC
E7G10	O'Neill *et al*. 2013	219	2	0	0.000000	0.000000	6	1	8	AAACATTTC ATTTATTTC TACCTGGG	TAAGCTTTCCTC AAGATCTTTACA GC
E7G8	O'Neill *et al*. 2013	169	2	1	0.314500	0.003516	7	2	1	GAGTATTGT TTGAAACTG GGTAGACA	GGTACAACCTAG TTCAGGTTCTTA GG
E8A10	O'Neill *et al*. 2013	127	2	2	0.943400	0.005278	7	3	4	AAAGTTTTC TTTTTAAGT TGCCAAAA	CATAATTTCTAC ATGATTTATGCG CT

Table 2.3 (continued). Summary statistics of the 89 loci included in the 43L and 89L data sets.

E8E1	O'Neill *et al*. 2013	188	4	2	0.628900	0.005611	7	2	4	AACACAAG GAAAAATG AAGAGTCCT A	TTCAGAAAGTCC AACGTTTATTAG TG
E8G11	O'Neill *et al*. 2013	155	4	0	0.331100	0.001089	6	2	9	AAATCACAA GTGGATGTT TACGTTC	CAAGACTGTAAG TTTAGTGCAACA CA
E8G8	O'Neill *et al*. 2013	187	0	0	0.000000	0.000000	7	1	13	CAAGGTTTT TGAAACATG CTCTT	ACTAGGTTAGAG AAAAACTAGCGC AC
E9A7	O'Neill *et al*. 2013	231	2	2	0.000000	0.000000	7	1	9	AAATTCAGT GAAAAGAG ACCGATG	AAATTAGCAAAG GCAGAAGAATTA AA
E9C4	O'Neill *et al*. 2013	211	2	0	0.000000	0.000000	7	1	1	GTGGTTATT TGTAACATT TCGTTGAC	AATTACATTTGG GCTTCTCAATTT AC
E9E4	O'Neill *et al*. 2013	123	3	1	0.628900	0.002361	7	2	12	GAAGATGCT TATGACATG AGGAAAG	AAACGTTTTTCA TCTGAAATGTTA GG
MGF	O'Neill *et al*. 2013	191	7	0	0.662300	0.003466	6	3	9	ACCTCCCAA GTGACTACA GTATATC	ACCTCCCACTCA AACAGCTTC

* Calculated across *A*. *ordinarium* and *A*. *melanostictum*
† Calculated across *A*. *ordinarium* only

Table 2.4. Results of SpedeSTEM analyses under two- and four-lineage scenarios. Underscores in the delimitation result indicate grouping; commas indicate divergence. ω_i indicates the relative model probability.

Gene copies sampled	Replicates	ω_i	Delimitation result
Two-lineage			
2	1000	0.72	One species (WE_EA)
3	1000	0.72	One species (WE_EA)
5	1000	0.69	One species (WE_EA)
10	1000	0.66	One species (WE_EA)
25	100*	0.62	Two species (WE, EA)
Four-lineage			
2	1000	0.22	One species (WE1_WE2_EA1_EA2)
3	1000	0.20	One species (WE1_WE2_EA1_EA2)
5	1000	1.00	Two species (WE1_EA1_EA2, WE2)
10	1000	1.00	Two species (WE1_EA1_EA2, WE2)

*Fewer replicates were performed due to computational limits.

Table 2.5. Results of spedeSTEM analyses under two three-lineage scenarios: (A) the eastern populations are fixed as a lineage and the split between WE1 and WE2 is tested, and (B) the western populations are fixed as a single lineage and the split between EA1 and EA2 is tested. Underscores in the delimitation results indicate grouping, commas indicate divergence events.

Gene copies sampled	Replicates	ω_i	Delimitation result
A. Eastern fixed, WE1 and WE2 tested			
2	1000	0.73	No divergence (WE1_WE2, eastern)
3	1000	0.73	No divergence (WE1_WE2, eastern)
5	1000	0.72	No divergence (WE1_WE2, eastern)
10	1000	0.71	No divergence (WE1_WE2, eastern)
B. Western fixed, EA1 and EA2 tested			
2	1000	0.73	No divergence (EA1_EA2, western)
3	1000	0.73	No divergence (EA1_EA2, western)
5	1000	0.73	No divergence (EA1_EA2, western)
10	1000	0.72	No divergence (EA1_EA2, western)

Table 2.6. Results of PHRAPL analyses of the 8L and 43L data for five two-lineage, two-parameter phylogeographic models.

Model	AIC	Log likelihood	ΔAIC	ωAIC	Rank	$\tau_{WE\text{-}EA}$	$m_{WE\text{->}EA}$	$m_{EA\text{->}WE}$
8L						4.10	0.72	3.81
1	354.22	−176.11	1.77	0.28	2			
2	357.49	−176.74	5.04	0.05	3			
3	363.86	−179.93	11.42	2.22×10^{-3}	4			
4	352.45	−174.22	–	0.67	1			
5	387.66	−192.83	35.21	1.51×10^{-8}	5			
43L					5.33	0.03	0.03	
1	2069.24	−1033.62	3.24	0.12	2			
2	2071.22	−1033.61	5.23	0.05	5			
3	2069.87	−1032.94	3.88	0.09	4			
4	2069.52	−1032.76	3.52	0.11	3			
5	2065.99	−1032.00	–	0.63	1			

Table 2.7. Model descriptions and selection results for a range of two-species migration models tested in MIGRATE-N. BAS: Bezier approximation score (log marginal likelihood) for all loci. LBF: log Bayes factor. LBFs and model probabilities calculated following Beerli and Palczewski (2010) and Kass and Raftery (1995).

Model	Description	BAS	LBF	Probability	Choice
1	Panmixia	−7184.5	481.0	3.5×10^{-105}	4
2	Full migration	−6944.0	–	1	1
3	Unidirectional: western into eastern	−7007.4	126.7	3.1×10^{-28}	3
4	Unidirectional: eastern into western	−6984.8	81.6	1.9×10^{-18}	2

Table 2.8. Rate of migration (M), direction, θ (mutation-scaled effective population size), and m (number of immigrants per generation) for the best-fit model (Model 2) between eastern and western *A. ordinarium* lineages estimated using Migrate-n. All numerical values listed are the mean estimate with 95% confidence intervals in parentheses. Provided θ values are for the lineage receiving migrants.

M	Direction	θ	Nm
813.9 (0 – 1600)	western into eastern	2.2×10^{-3} ($1.1 \times 10^{-3} – 3.2 \times 10^{-3}$)	0.44 (0 – 1.29)
1155.6 (0 – 2200)	eastern into western	5.4×10^{-4} (0 – 1.4×10^{-3})	0.16 (0 – 0.77)

Figure 2.1. Geographic sampling of *Ambystoma ordinarium* populations used in this study and patterns of population genetic structure. (A) Localities 1-20 represent sampling for the 8L data set. Stars denote sampling localities used in the 89L data set. Locality numbers match those of Weisrock et al. (2006). Dashed outlines enclose population genetic clusters identified from Structure analyses at $K = 2$ (western and eastern), while solid colored areas represent hierarchical population genetic structure identified within the western and eastern clusters. Nodes on the hypothesized species tree topology represent the putative divergence events tested in this study. Arrows denote distinct southward-flowing drainages of streams in the western and eastern portions of the range. (B) $K = 2$ is best supported in analyses of the full 8L data set, but when western and eastern populations were analyzed separately, each had a ΔK favoring a $K = 2$. Photograph of *A. ordinarium* courtesy of Sebastian Voitel.

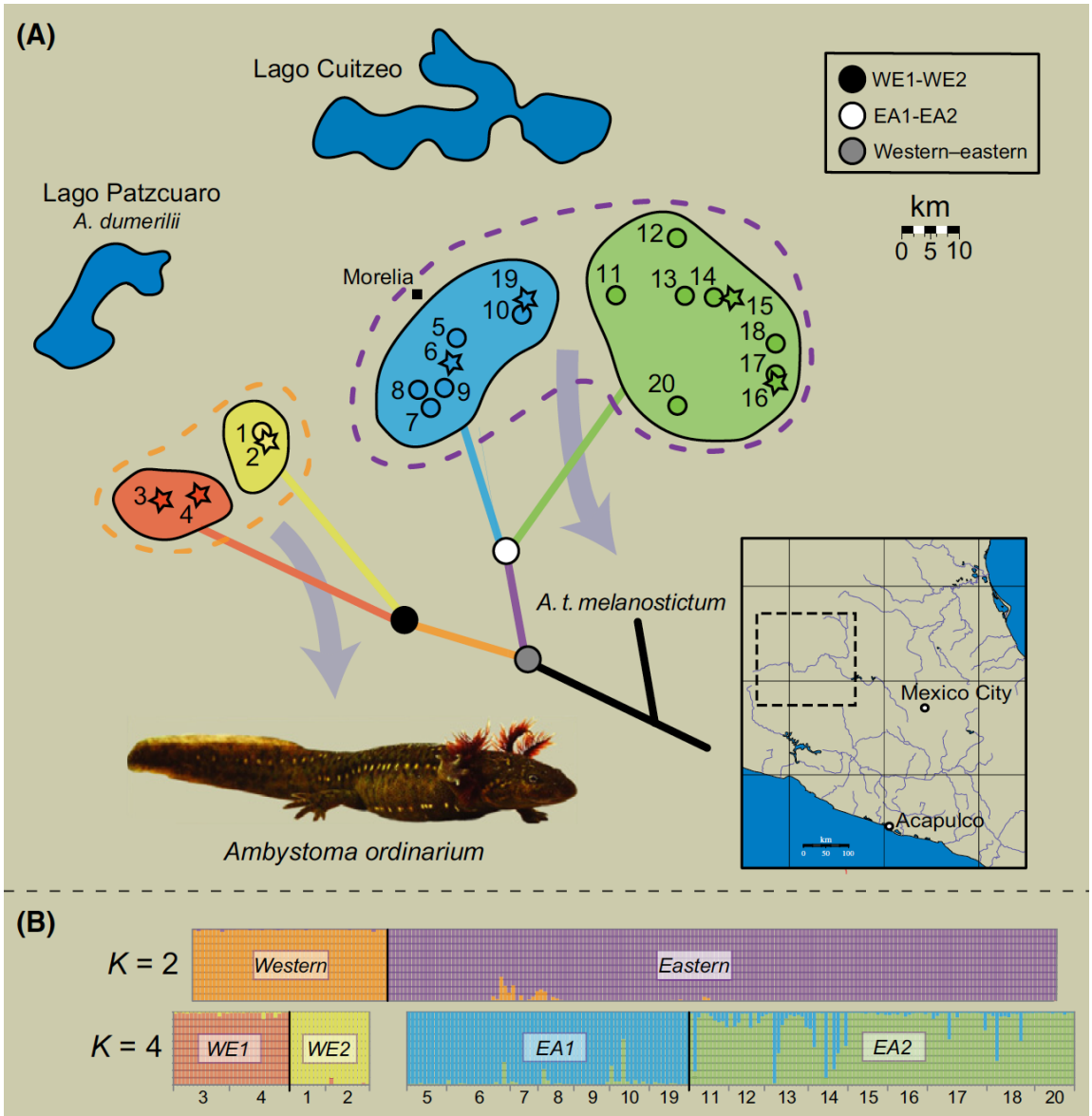


Figure 2.2. Demographic and phylogeographic models for the western and eastern *A. ordinarium* lineages tested in PHRAPL analyses. Horizontal arrows indicate gene flow between lineages.

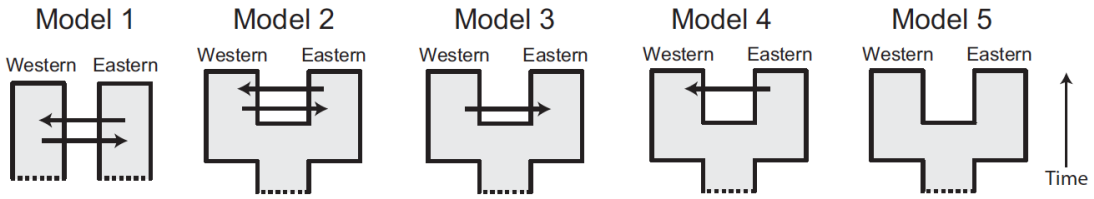


Figure 2.3. Demographic and phylogeographic models for the western and eastern *A. ordinarium* lineages tested in MIGRATE-N analyses.

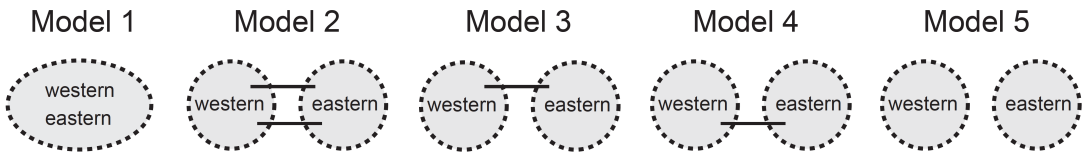


Figure 2.4. Additional Structure results for non-hierarchical K = 3 and K = 4, and for optimal K (3) from STRUCTURAMA treating K as a random variable.

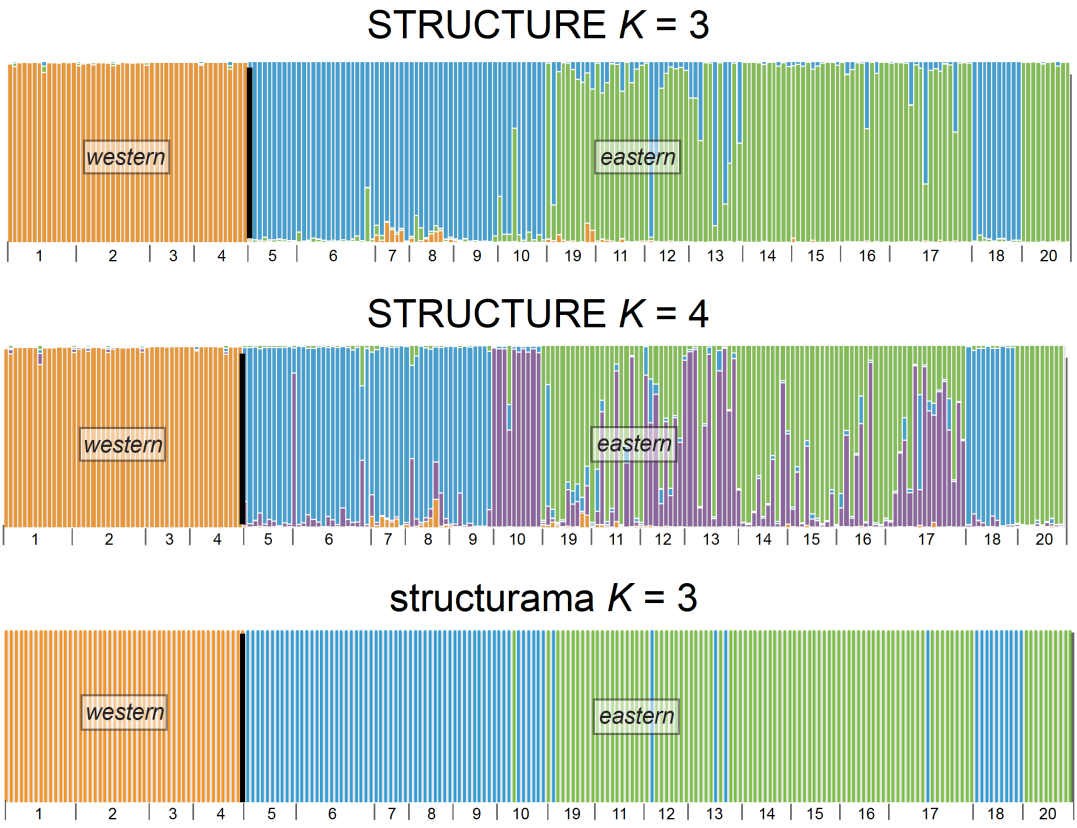
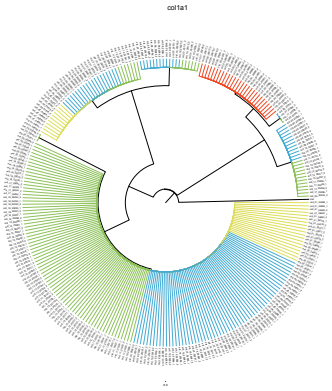
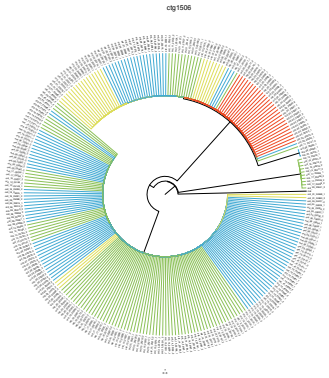


Figure 2.5. Gene trees estimated for 8L loci. Colors correspond to Figure 2.1.

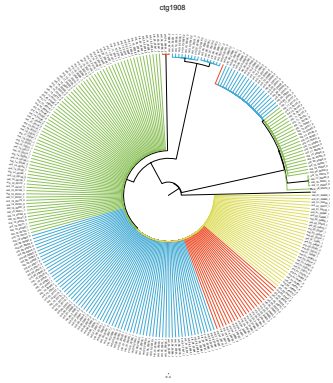
A. *col1a1*



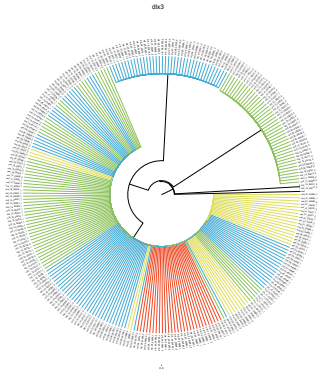
B. *ctg1506*



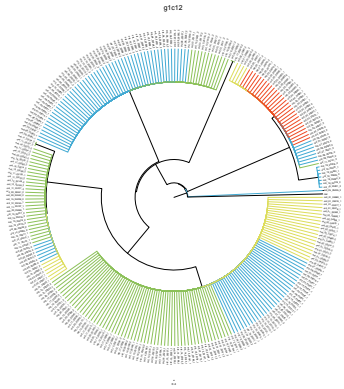
C. *ctg1908*



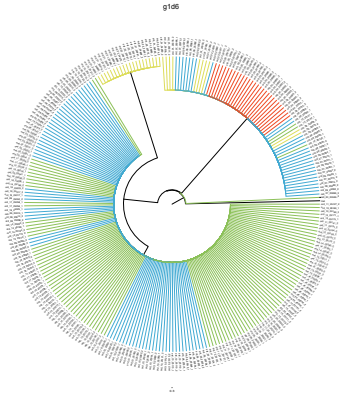
D. *col1a1*



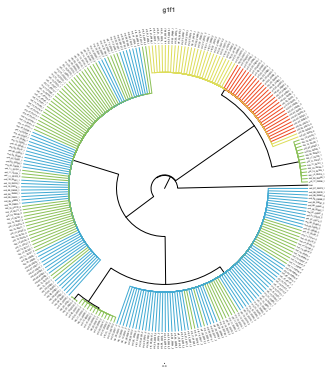
E. *ctg1506*



F. *ctg1908*



G. *g1f1*



H. *g3d7*

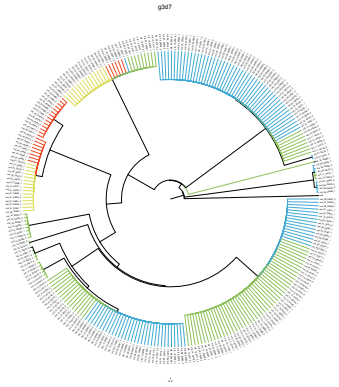


Figure 2.6. Generalized multilocus haplotype networks for *A. ordinarium* inferred in SplitsTree for the (A) 8L and (B) 89L data sets. Uncorrected P-distances were plotted using convex hull representation.

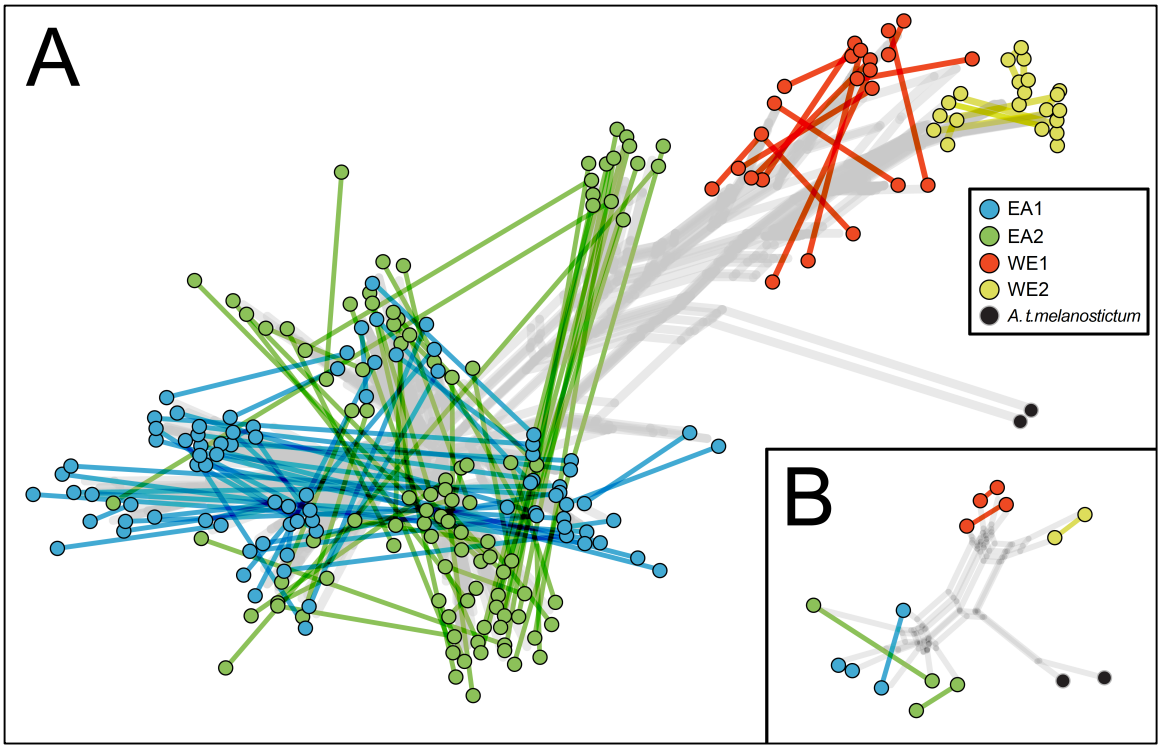


Figure 2.7. Results from BPP analyses of the 8L and 89L data sets. Circle coloration corresponds with hypothesized divergences in Figure 2.1. The x-axis is labeled with two-letter designations for prior combinations of Θ and τ, with Θ designated as large (L), intermediate (I), or small (S) and τ designated as deep (D), intermediate (I), or shallow (S). Results are also presented for empirically (EM) derived priors. In total, ten different combinations of priors for were tested and mean posterior probability and standard error is reported for 10 replicates per prior for the 89L data and 20 replicates per prior for the 8L data. Plots are color coded by nodes in the hypothesized species topology shown in Figure 2.1. Figure panels correspond to (A) 8L data, tips assigned to hypothesized species, (B) 8L data, tips randomly assigned, (C) 89L data, tips assigned to hypothesized species, (D) 89L data, tips randomly assigned.

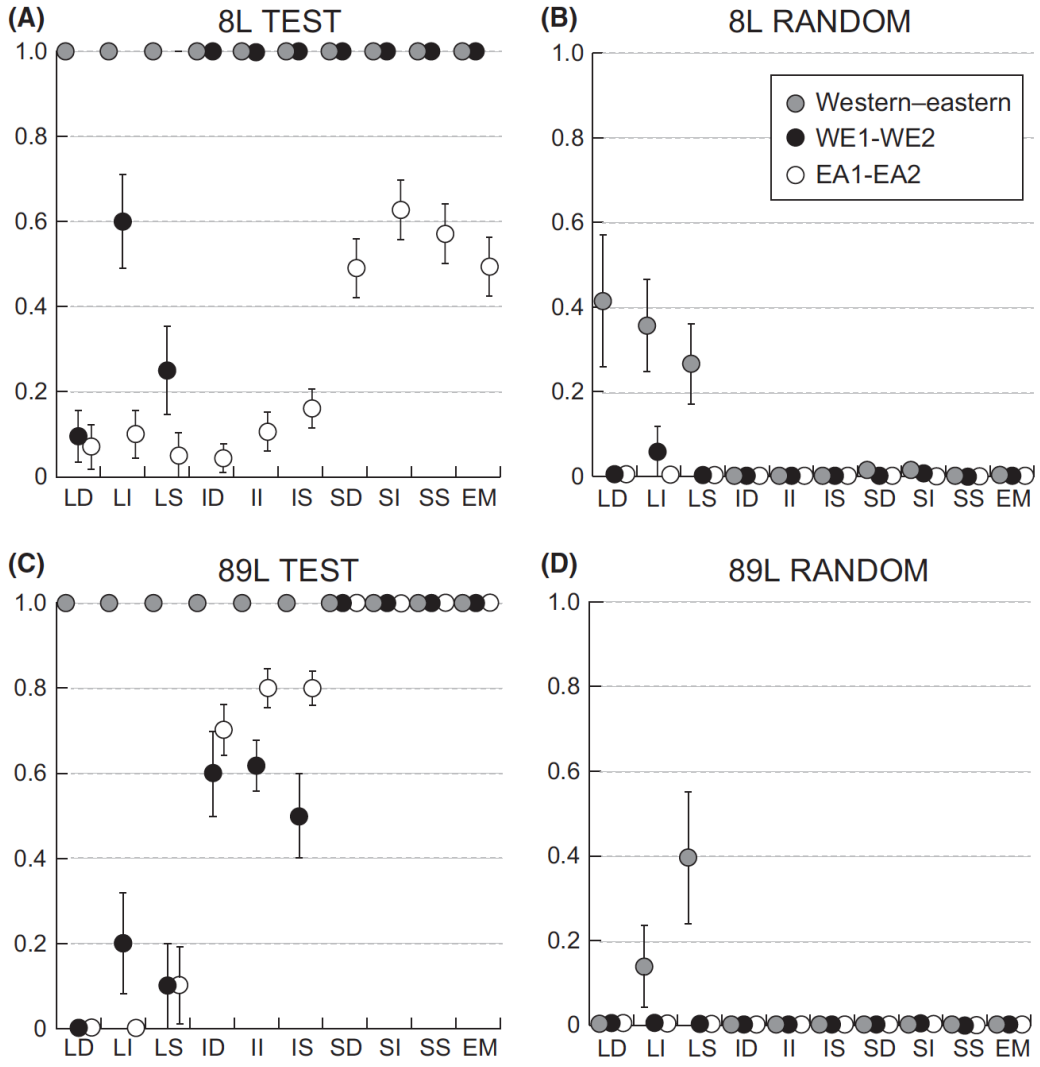


Figure 2.8. Relationships among *A. ordinarium* inferred with SVDQuartets. a) "Lineage tree" for 43L data set using exhaustive sampling of quartets over 1,000 bootstrap replicates. b) Species tree inferred under a four-lineage constraint for the 8L and 43L data sets using exhaustive quartet sampling and 100 bootstrap replicates. Branch support values to the left of the backslash are for the 8L data set and those to the right are for the 43L data set. Colors correspond to Figure 2.1.

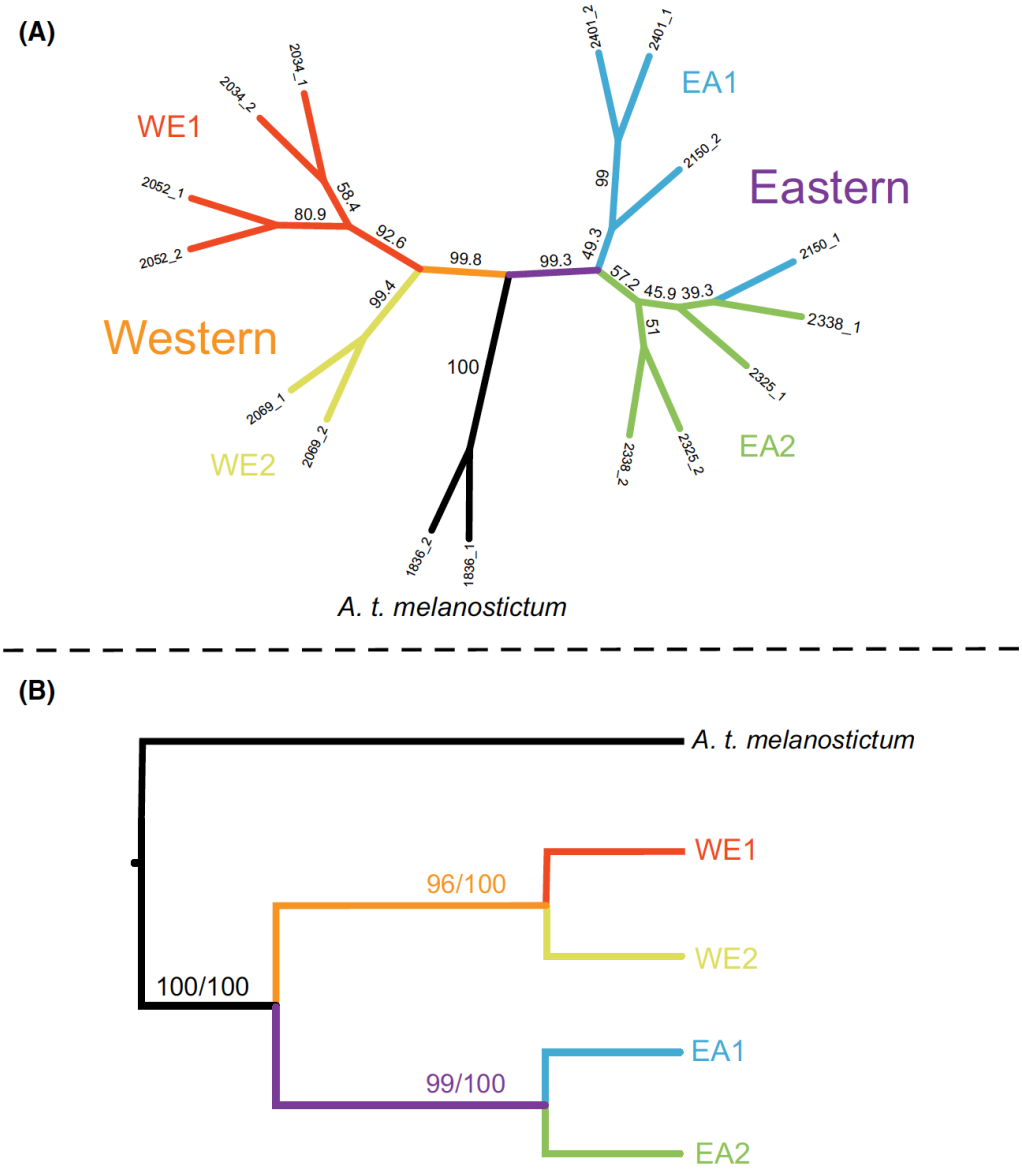
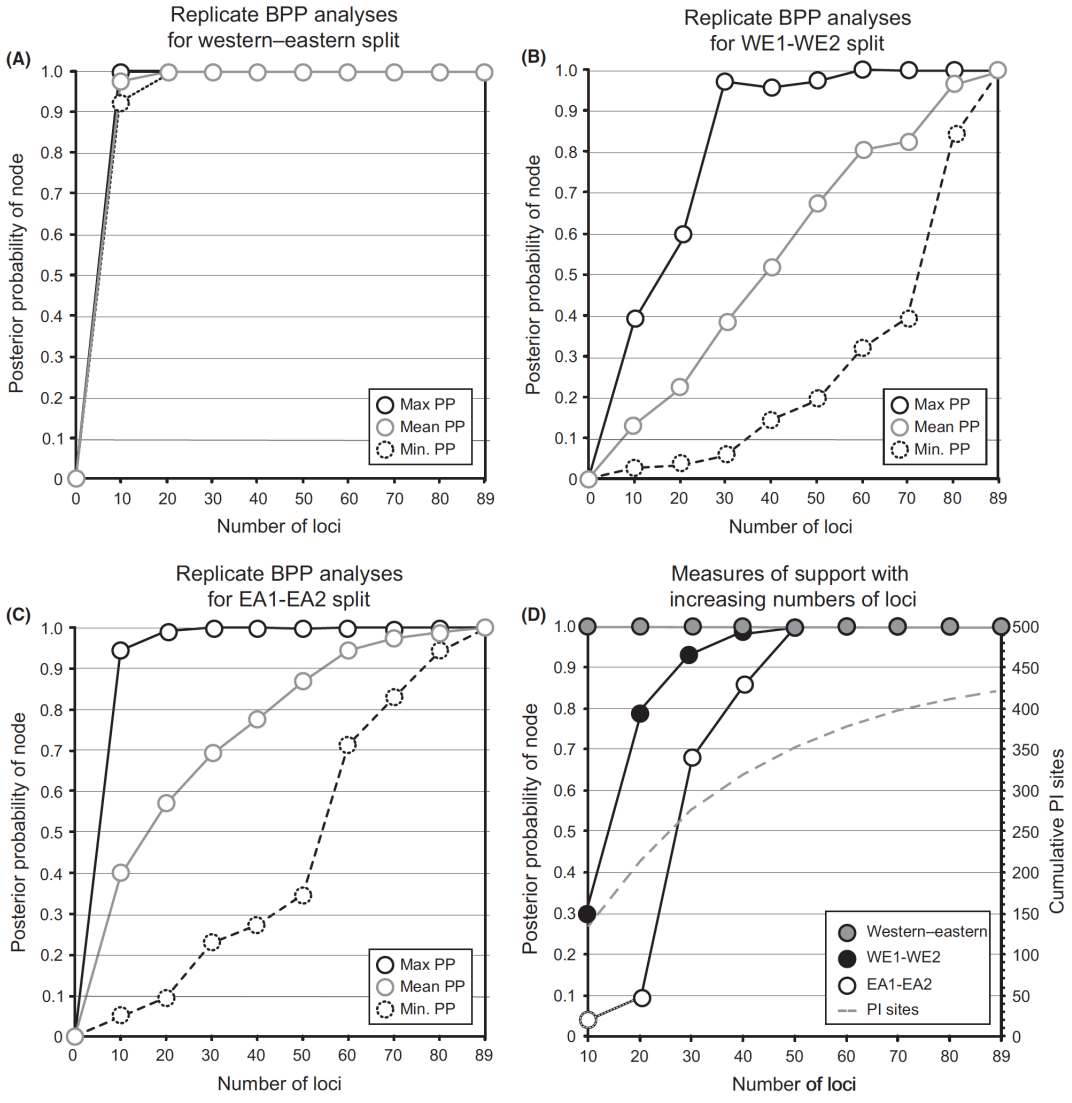


Figure 2.9. Effects of locus subsampling on BPP node support for 89L *A. ordinarium* data sets. Randomly selected loci were sampled without replacement in increments of 10 (i.e., 10, 20, 30 ... 89) across 10 independent replicates. The maximum, mean, and minimum posterior probability (PP) for nodes in the guide tree are shown for (A) western-eastern, (B) WE1-WE2, and (C) EA1-EA2. All runs used the empirically estimated priors described in the text. (D) When loci are rank ordered by number of parsimony-informative (PI) sites and analyzed in multiples of 10 loci, support for eastern-western split is unanimous and support for EA1-EA2 and WE1-WE2 both reach PP = 1.0 by 50 loci.



CHAPTER THREE

Cryptic branches in the *Cryptobranchus* tree: Genomic data reveal an underestimation of North American aquatic salamander diversity

ABSTRACT

Identifying the demographic and historical forces which have shaped contemporary patterns of biodiversity is a primary aim of phylogeographic studies. Central to these efforts, is the desire to delimit species-level entities objectively across organismal groups. Perception biases have potentially skewed species description towards taxa with conspicuous species boundaries, whereas a large number of drab, relatively poorly studied taxa await detailed studies of species limits. We investigate range-wide phylogenetic relationships and putative species boundaries in the imperiled North American hellbender salamander (genus *Cryptobranchus*), integrating comprehensive geographic sampling and dense sampling of the nuclear genome in a model-based statistical framework. Our results suggest that *Cryptobranchus* contains as many as five deeply divergent, cryptic lineages which are broadly aligned with the major continental watersheds of eastern North America. We demonstrate that rates of effective gene flow between these lineages are up to four orders of magnitude lower than rates of gene flow within lineages, and that these lineages share genealogical exclusivity across the genome. These findings have significant implications for delimiting species with genomic data, and imply that freshwater diversity in temperate environments may be underestimated for cryptic taxa. This work also has

applied conservation implications for hellbenders in that, regardless of the true number of hellbender species, given current trends, a model of zero *Cryptobranchus* species may be impossible to reject in the foreseeable future.

INTRODUCTION

Hellbender salamanders (genus *Cryptobranchus*) are large, obligately aquatic amphibians with a wide-ranging historical distribution across central and eastern North America (Nickerson & Mays 1973). Hellbenders were historically common across most rivers and creeks in: the lower Osage and Missouri River drainages in Missouri; the upper White and Black Rivers in the Ozarks of Missouri and Arkansas; the entire Tennessee River drainage in Tennessee, Alabama, Mississippi, Georgia, North Carolina, and Virginia; the entire Ohio/Allegheny River drainage in Kentucky, Illinois, Indiana, Ohio, West Virginia, Pennsylvania, New York, and Maryland; the entire Kanawha/New River drainage in North Carolina, West Virginia, and Virginia; and the western portions of the Susquehanna River drainage in Pennsylvania and New York. Many hellbender populations have crashed in recent decades (Wheeler *et al*. 2003; Pitt *et al*. 2017), sometimes by as much as nearly 90%, and understanding the genetic relationships among the remaining populations is paramount for conservation of these enigmatic salamanders.

Despite its broad geographic distribution, one single species, *Cryptobranchus alleganiensis*, was described over 200 years ago (Sonnini de Manoncourt and Latreille, 1801). Grobman (1943) described a new hellbender species, *Cryptobranchus bishopi*, from the Current River in the southern Ozarks in Missouri. Distinguished by its smaller body

size, heavily blotched dorsal patterning, and a disjunct geographic distribution in the southward flowing tributaries of the White and Black Rivers in southern Missouri and northern Arkansas, the species status of the Ozark hellbender has long been of interest to systematists and herpetologists (e.g., Firschein 1951; Routman *et al*. 1994; Crowhurst *et al*. 2011). Another isolated group of hellbender populations west of the Mississippi River in the northward flowing tributaries of the Osage and Missouri Rivers in central Missouri has traditionally been classified with all eastern hellbender populations as *C*. *alleganiensis*. *C*. *bishopi* was reclassified as a subspecies of *C*. *alleganiensis* by Dundee & Dundee (1965) based on an argument that an allopatric distribution alone or in combination with a handful of morphological differences were insufficient to warrant species status. The two species were synonymized under *C. alleganiensis*, and two subspecies are currently recognized, *C*. *a*. *bishopi* and *C*. *a*. *alleganiensis*, the "Ozark" and "eastern" hellbenders, respectively.

Previous genetic studies have hinted at the potential for, perhaps extensive, cryptic diversity in *Cryptobranchus*, but to date there has been no comprehensive effort to obtain the large amounts of genomic data from many individuals across the complex geographic distribution needed to rigorously test species boundaries in this group (Fujita *et al*. 2012). This work is timely and pressing because hellbenders are rapidly declining in many parts of their historical range (Pitt *et al*. 2017; Wheeler *et al*. 2003) and *in-situ* and *ex-situ* conservation and management efforts are trying to curb the losses. Assessing the species status of this group is paramount to their conservation because the actionable consequences of there being a single species versus multiple species are very different. It is critically important to get this right, and to avoid under-splitting (which would miss distinct species and potentially doom them) or over-splitting (which would complicate conservation

strategies and squander limited financial resources on conserving trivially diverged subsets of the same species as distinct) (Carstens *et al*. 2013). Though agnostic with regard to the number of hellbender species, our analyses are aimed at evaluating the evidentiary support across large swaths of the genome under a wide range of competing demographic and phylogeographic models to arrive at a subset of models which best capture the information content from our data with respect to hellbender species boundaries.

Early allozyme work (Merkle *et al*. 1977; Shaffer & Breden 1989) seemed to show a pattern of very low genetic variability at the levels of individuals and populations. The earliest sequence-based genetic research on hellbenders was conducted by Routman and Templeton (1994), using mitochondrial DNA (mtDNA) restriction fragment length polymorphisms to estimate haplotype networks among different populations and to probe the trans-Beringian distribution of *Cryptobranchus* and its sister lineage *Andrias*. *Cryptobranchus* are obligately tied to riverine systems, and much of the differentiation between populations was expected to reflect the structure of the river networks in which they reside. However, some of the early population genetic work in hellbenders hinted at ancient connections between the Kanawha River in West Virginia, North Carolina, and Virginia and the southward flowing rivers of the southern Ozarks in Missouri and Arkansas (Routman and Templeton 1994), a result which has been corroborated with mtDNA sequencing (Sabatino and Routman 2009) and more recently with nuclear microsatellite data (Tonione *et al*. 2011). In additional studies (Unger *et al*. 2013; Unger *et al*. 2016), the Kanawha/New River, Tennessee River, Ohio River drainages were each shown to be divergent from each other, but the topologies of the underlying phylogeny were unclear. In another important contribution, Crowhurst *et al*. (2011) used microsatellite markers to

detect extensive population genetic structure between Missouri/Mississippi River populations and Ozark populations in Missouri. This work also revealed significant structuring between White River and Black River populations within Ozarks. Despite the potential for the isolated nature of hellbender populations to lead to differentiation between populations, genetic variation within hellbender populations suggests panmictic metapopulation dynamics with virtually no genetic structuring between different sites in that watershed (e.g., Feist *et al* 2014). Large rivers may act as barriers to dispersal between smaller tributaries, and may serve as isolating barriers to gene flow.

METHODS AND MATERIALS

Geographic sampling of individuals

We aimed for a balance between the number of geographic sites sampled (constrained by population status) the and number of individuals sampled per site. For analyses presented here, we sampled 93 individual hellbenders from 39 sites across eight states, representing all major watersheds across the range and fine-scale sampling within many drainages (color-filled circles in Figure 3.1). We worked to sample all of the major watersheds and as many tributaries of these as possible. This was not a trivial task, and over 13 years of sampling in 16 states, the authors and others expended nearly 12,000 person-hours of survey effort. We obtained tissue samples from nearly every HUC-8 or greater USGS watershed division in which *Cryptobranchus* is thought to persist. Sampling at historical sites at greater levels of spatial resolution yielded low, in many cases abysmal,

rates of capture. Animals were predominately captured by hand or net during skin diving and snorkel surveys, although a few were recovered from anglers. Tissue or blood samples were collected for DNA extraction. For tissue samples, a small biopsy was taken from the tip of the tail and stored in 95% ethanol. For blood samples, 0.1 - 0.5 mL of blood (depending on the size of the animal) was collected via venipuncture of the ventral caudal vein and stored in standard lysis buffer using non-heparinized syringes. All animal manipulations were conducted in accordance with relevant IACUC guidelines for animal welfare and under scientific collection permits for each respective state. We also sampled two Chinese giant salamanders (*Andrias davidianus*) and two Japanese giant salamanders (*A. japonicus*) to serve as comparative outgroups. These individuals were from the collections of the St. Louis Zoo, the California Academy of Sciences Steinhart Aquarium, and the National Zoo. Although little specific locality information was available for these *Andrias* individuals, their classifications as either Chinese or Japanese giant salamanders were unambiguous.

Generating genome-wide genetic markers

Genomic DNA was extracted using Qiagen DNeasy spin column extraction kits and was quantitated with a Qubit™ fluorescence spectrophotometer (Thermo Fisher Scientific). We initially sought to develop genomic resources in *Cryptobranchus* by using a targeted sequence capture approach to sequence approximately 400 phylogenetically conserved nuclear exons identified in a multi-tissue transcriptome assembly which we generated for the hellbender. We designed custom DNA capture probes tiled across these

target loci and used a modification of the method of Lemmon *et al*. 2012 to sequence these gene regions in multiplex. In an initial trial, we sequenced six *Cryptobranchus* (two individuals each from the White River in Missouri, the Elk River in Tennessee, and Tionesta Creek in Pennsylvania) and one of each species of *Andrias*. Although we achieved favorable recovery of loci (91%, 312 out of 343) and low rates of missing data across these eight test individuals (Figure 3.2A), the recovered loci were not particularly variable, especially within *Cryptobranchus*. The target loci were on average approximately 1,300 base pairs (bp) in length. Within *Cryptobranchus*, nearly half of loci were invariant, and the remainder had on average approximately 2.1 variable nucleotide positions. As expected, genetic variation between *Cryptobranchus* and *Andrias* was greater, with nearly all loci being phylogenetically informative and with an average of roughly 9 variable nucleotide positions (Figure 3.2B). An inherent constraint of this method is that we were limited to multiplexing 12-24 individuals on a single Illumina HiSeq lane to achieve sufficient depth of coverage, and given that we were essentially generating SNP data for a relatively small number of loci (relative to the potential number of loci from a reduced representation approach in a large genome), we pursued alternative protocols.

We developed genomic resources *de novo* with double digest restriction site-associated DNA sequencing (ddRAD) (Peterson *et al*. 2012). From a methodological perspective, the large genome size in *Cryptobranchus* (approximately 55 GB) (Gregory 2017) presents both challenges and opportunities for high-throughput genomic data collection. But by carefully optimizing the combination of restriction enzymes and the size selection window of retained double digested fragments, we were able to successfully design a reduced representation protocol for hellbenders that balanced the number of

expected loci sequenced with the expected coverage of loci and the number of individuals which could be multiplexed during a sequencing run. The number of loci generated with a ddRAD approach depends on the specific combination of enzymes chosen (and on the base composition and relative rarity of their recognition sequences), the size range of fragments selected, and the density of the empirical fragment distribution. We performed a series of test restriction enzyme digestions for several potential enzyme combinations to estimate the potential number of loci that would be recovered and to select a pair of enzymes which would optimize the balance between the number of fragments and the number of individuals which could be multiplexed while retaining sufficient sequencing coverage. For each of six potential combinations of restriction enzymes, we performed a double digestion and each of the single digestions separately for two test individuals (from the White River in Missouri). Digests were performed with enzymes from New England Biolabs according to manufacturer's recommended reaction conditions.

After digestion for four hours at manufacturer recommended enzyme pair-specific temperatures (the enzymes were not heat inactivated), we cleaned the resulting digestion products with Agencourt Ampure XP beads, and visualized the resulting fragment distributions on an Agilent 2100 Bioanalyzer with high-sensitivity DNA chips (Agilent Technologies). For each restriction enzyme combination, we used the method of Peterson *et al*. (2012, Supplemental Materials) and empirical genome size estimates for *Cryptobranchus* to estimate the number of double digested fragments within three different size ranges (300 ± 30 bp, 400 ± 40 bp, 500 ± 50 bp). Based on these estimates, it became apparent that most enzyme combinations and most size selection windows resulted in many hundreds of thousands or even millions of potential ddRAD loci. In order to permit more

efficient multiplexing of individuals while still allowing for sufficiently high expected depth of coverage of loci, we chose the restriction enzyme combination and size selection window with the smallest number of expected loci per individual (*EcoRI*/*SphI* with size selection from 450 - 550 bp). Under these conditions, we expected the ddRAD approach to yield approximately 300,000 loci per individual, and we reasoned that with an 85% on-target rate of sequencing, we would be able to multiplex 10-12 individuals on a single Illumina HiSeq 2500 lane while still achieving greater than 30X mean depth of coverage across loci. However, we note that these expectations do not account for repetitive elements in the *Cryptobranchus* genome, which because of their potential high copy number and the potential difficulties resolving these paralogous sequences, will likely cause our estimates of locus number to differ from empirical observations. We prepared a test ddRAD library (as detailed below) consisting of these two individual hellbenders and sequenced this on a half lane of Illumina HiSeq. We sought to overshoot in terms of sequencing depth so that we could estimate the effects of different levels of multiplexing on locus recovery and coverage. The results were favorable, and we adopted this ddRAD protocol going forward.

Because of the large hellbender genome size, we increased the amount of input genomic DNA of each individual for restriction enzyme digestion to 3.00 μg (from 50 - 100 ng in the original protocol), reasoning that higher input DNA quantity would result in more accurate sequence determination because fewer PCR cycles (which inherently introduce base errors) would be required to obtain sufficient quantities of prepared ddRAD libraries. We used a dual index combinatorial multiplexing strategy, identifying individuals by unique combinations of 5 bp inline barcodes and 6 bp Illumina indices. We pooled four sets of five individuals after individual restriction enzyme digestion and adapter ligation,

and then we size selected each set of five individuals in its own well of a Pippin Prep (Sage Sciences) cartridge. In practice, size selection of *in situ* fragments in the 442 - 558 bp range was performed using the "tight" collection protocol with an actual size selection window setting of 518-634 bp (576 bp +/- 10%, to account for the combined 76 bp lengths of the Illumina adapters which were ligated to the ends of all fragments).

Size-selected products were pooled into sets of ten to twelve total individuals, bead cleaned, and amplified by PCR for 8 cycles with a high-fidelity polymerase (NEB Phusion), as in Peterson *et al*. 2012. This low-cycle PCR step was aimed at avoiding low-complexity molecular bottlenecks and reducing PCR duplicates and enzymatic polymerase errors in the resulting amplicons. Final bead cleanup steps were performed with Dynabeads and then Agencourt AmPure XP beads, and the completed ddRAD libraries were quantified on a Qubit fluorescence spectrophotometer and visualized with the Agilent Bioanalyzer 2100 fragment analyzer. The unique combination of index, in-line barcode, and sequencing lane allowed us to trace raw reads back to individual hellbenders after sequencing the genomic libraries in multiplex. We prepared batches of multiples of up to 12 individuals at a time representing a total of 203 *Cryptobranchus* individuals and four *Andrias*. Resulting libraries were sequenced on 20 full Illumina HiSeq2500 lanes in Rapid Run mode with paired-end 150 bp reads (utilizing onboard cluster generation). A 10% *PhiX* DNA spike-in was used to increase nucleotide diversity and produce more optimal clonal cluster generation (reads from the *PhiX* spike-in are automatically removed by the sequencing center). Illumina sequencing was performed at the Florida State University School of Medicine Core Facility.

Locus assembly and characterization

The particular library preparation protocol that we employed results in strand-specific loci because our PCR primers for fragment amplification effectively selected for only those fragments with *SphI* at the 5' end and *EcoRI* at the 3' end. The total lengths of the fragments which we sequenced to generate our ddRAD loci exceeds the combined length of both of the 150 bp paired-end reads. Each fragment is essentially represented by loci comprising 150 bp of 5' sequence and 150 bp of 3' sequence at the flanks, with a central un-sequenced region of unknown length (fragments originated from a fragment distribution centered around 500 ± ~50 bp). To account for this feature of our particular combination of size-selection window and read lengths, and in an effort to retain information from both the R1 and R2 read pairs, we used custom bash scripts to concatenate reads from the 5' ends of fragments (R1 of an Illumina read pair) with the reverse complement of reads from the 3' ends of fragments (R2 of an Illumina read pair), recapitulating the original orientation in the genome.

Although this "stitching" procedure unites noncontiguous genetic regions by not accounting for the internal un-sequenced regions, this approach retains the provenance between these stitched flanking regions in R1 and R2, in contrast to methods that treat 5' and 3' fragments as separate loci or which simply exclude half of the read data. Loci should all be greater than 400 bp in length, so none should have overlap between 5' and 3' fragments. The process_radtags function in stacks (v1.29, Catchen *et al*. 2013) was used to demultiplex the raw, stitched reads by individual, allowing for one nucleotide mismatch in the observed barcodes from the reference list (all barcodes used were two or more

substitutions away from each other in substitution space). Stitched reads were only retained if they contained the appropriate restriction enzyme cut sites at both ends and also had a mean Phred quality score greater than 20 over all 45 bp sliding window intervals along their total length. These parameters amounted to the following settings for the stacks process_radtags algorithm: --renz_1 sphI --renz_2 ecoRI -c -q -r -D -w 0.15 -s 20 --barcode_dist_1 2. A preliminary analysis of SNP variation across sites in loci reconstructed from these data suggested that the 3' ends of the Illumina reads contained significantly elevated levels of polymorphism, likely due to increased rates of sequencing errors towards the ends of R1 and R2 which were not removed under our filtering parameters. To avoid introducing thousands of known erroneous variable sites into downstream analyses, a 2 bp region from the 3' end of each raw read in a pair was removed prior to read stitching (Figure 3.3). These 3'-truncated, stitched reads were then used to assemble loci.

The stacks assembly pipeline (v1.29) was used to assemble unique loci and to make preliminary haplotype calls for each individual (ustacks); to assemble a locus catalog for all individuals (cstacks) denoting which loci are shared by which individuals; to find catalog matches for each individual (sstacks); and to call haplotypes across all individuals (genotypes). Stitched, demultiplexed, and filtered reads were assembled for each individual in parallel for six combinations of assembly parameter settings. We attempted to consider multiple expectations for the range of nucleotide variation between alleles at a given locus (ustacks -M = 4, 10), for the range of sequencing coverage across individuals (ustacks -m = 3, 10), and for variation between alleles across the set of individuals (cstacks -n = 0, 16, 32). We used sstacks to match individual loci back to the full catalog, and we reconstructed haplotypes across all loci for all 93 individuals with genotypes (-r 1 -m 3). Exploring these

twelve combinations of assembly parameters for ustacks and cstacks, we aimed to choose parameter settings which would optimize the recovery of putatively orthologous, single-copy regions of the genome from our assembled loci. After a moderate exploration of various assembly parameters, we arrived at a set of parameter settings for ustacks (-m 10 -M 16 -N 16 -H) and cstacks (-n 16) which seemed to optimize the number of shared putatively orthologous loci within and between populations, while limiting the number of loci which could be rejected as orthologous on the basis of coverage, zygosity, or patterns of missing data. We explored several thresholds of missing data in the stacks genotypes algorithm and ultimately produced a data set of 74,084 loci which were present in at least 72 of the 93 *Cryptobranchus* individuals. Of these loci, 71,734 were variable.

Population genetic structure

We took a population genetic approach to understanding genetic structure across the distribution of hellbenders. We used linear discriminant analysis of principal components (Jombart *et al*. 2010) to summarize genetic variation across the hellbender genome and to visualize genetic differentiation among lineages. These analyses were performed in the Adegenet R package for the set of 93 hellbenders from 39 different localities. We first calculated the posterior probability of varying numbers of population clusters from 1 to 40, inclusive, and used the Bayesian Information Criterion (BIC) to select a number of clusters (K) that minimized the information loss associated with describing our data with that model. A preferred model for the number of population genetic clusters was identified from the first inflection point of a plot of K versus ΔBIC. An optimal number

of principal components (PCs) (n = 3) were selected to transform the data from all loci, and we then selected the smallest number of linear discriminant functions which were capable of describing more than 80% of the observed variance in allele frequencies across population clusters. This discriminant analysis of principal components (DAPC) allowed us to visualize the clustering of individual hellbenders in genetic variation space.

F_{ST} across the hellbender genome

We also calculated the fixation index, F_{ST}, across the set of 71,734 loci under the five-cluster population assignments identified through the DAPC analysis. We performed these analyses in the Adegenet R package (Jombart & Ahmed 2011) and then visualized the results as a discretized distribution of 100 increments.

Haplotype network analysis and species tree estimation

We estimated phylogenetic relationships among sampled individuals using two different methods. First, a generalized multilocus haplotype network was estimated for *Cryptobranchus* in SplitsTree (Huson *et al*. 2008) using the set of 71,734 variable loci. This approach makes no assumptions about how individuals are binned into populations and represents all possible historical connections between individuals. We also estimated a species tree in SVDQuartets (Chifman and Kubatko 2012) for all *Cryptobranchus* and *Andrias* individuals using the full set of 74,084 loci. Here we forced individuals from each of our 39 sampling sites to cluster together, but imposed no constraint on the relationships

among river lineages. This approach obviates the need to individually estimate gene trees for every locus. Instead, for each of the 45,697,312 unique groupings of four individuals (quartets), we compared the ranks of flattening matrices for each of the three possible quartet topologies, selecting the one with a rank of at least 10 as the correct configuration. Each inferred correct quartet was then amalgamated into an estimate of the species tree using the. Branch support was measured using 100 nonparametric bootstrap replicates and these bootstrap trees were summarized with a maximum clade credibility tree in Dendropy-4.0.0 (Sukumaran & Holder 2010).

Geographic patterns of genetic differentiation

We investigated hierarchical correlations between pairwise genetic distance and pairwise geographic distance for all of our *Cryptobranchus* samples and for our *Andrias* samples. Pairwise genetic distances were calculated in PAUP4.0a152 (Swofford 2015) as Jukes-Cantor distances which account for back mutation. Because we sequenced both alleles for each locus in every individual, the four pairwise distances between the two alleles (A and B) of two diploid individuals (1 and 2) (e.g. A1-A2, A1-B2, A2-B1, B1-B2) were averaged for each pairwise contrast of two individuals. Because cryptobranchid salamanders are obligately aquatic and incapable of long distance over-land dispersal, we calculated all pairwise geographic distances as minimum resistance distances along stream courses. For cases where two samples were not in the same watershed (e.g., comparisons between Atlantic-draining Susquehanna River sites and Mississippi River-draining watersheds), the shortest over-land distance connecting the disparate watersheds was

calculated. We first calculated the average, maximum, and minimum pairwise genetic distances within and between *Cryptobranchus* and *Andrias*.

Spatial patterns of genetic diversity

To understand the spatial distribution of genetic diversity across the range of the hellbender, we analyzed a set of 8,606 bi-allelic loci in EEMS (Petkova *et al*. 2016). We constructed two different polygons around the geographic distribution of *Cryptobranchus*. The first polygon roughly followed a convex hull around the general outline of the hellbender range and was overlaid with a regular triangular grid representing 750 different demes. The second polygon was more complex and delineated the hellbender distribution into regions representing the five lineages identified from the other analyses. This more complex polygon was overlaid with a regular triangular grid representing 3,000 different demes (higher grid density was required to individually demarcate and fill the geographic distribution of each lineage). For each habitat outline, we ran a series of Markov chain Monte Carlo simulations in EEMS to obtain estimates of the posterior mean genetic diversity (heterozygosity) across demes in each model.

Spatial patterns of gene flow

Similarly to the estimates of genetic diversity across the landscape, we also estimated effective rates of migration across the landscape using EEMS. Here, the surfaces estimated are effective migration rates, not absolute migration rates.

Topological concordance between phylogeny and river networks

We used the cophylo.plot() function in the phytools R package (Revell 2012) to compare the topologies of the estimated relationships among populations of hellbenders and the contemporary network of river connectivity across the hellbender's distribution. 10,000 random trees were simulated as a null distribution for a model of no association between the river network and the phylogeny. A P-value was calculated for the rejection of this null hypothesis. We note that it is unclear whether a null expectation of no correlation between river network and phylogeny is necessarily an appropriate null.

Coalescent species delimitation

We used BPP v3.2 (Yang & Rannala 2010) to test different species delimitation models based on a fixed species tree topology estimated above in SVDQuartets. For the BPP analyses, we immediately recognized the need to either down-sample the number of loci under consideration if we to use all 186 alleles from all 93 individuals (settling on a data set of 150 loci), or to reduce the number of individuals under consideration. After some experimentation, we arrived at a set of 35 individuals representing the major lineages identified in the population genetic and phylogenetic analyses, sampled for 23,724 loci with no missing data across these 35 individuals. We used a fixed species tree topology representing the backbone of the SVDQuartets topology and we explored prior settings for ancestral effective population sizes (θ) and divergence times in coalescent units (τ),

selecting Gamma distributions with $\alpha = 2$ and β = either 10, 100, or 1000 for both parameters. We settled on $\alpha = 2$ and $\beta = 1000$ for the priors on θ and τ. Ten replicate BPP analyses were conducted using an MCMC chain with 500,000 generations of burnin, and then sampling every 50 generations for a total of 10,000 samples in order to estimate the posterior distribution of species delimitations and model parameters for the fixed five-lineage species tree. Posterior probabilities of splits at nodes in the guide tree were averaged across the ten replicates to obtain estimates that a given node in the guide tree subtends lineages which represent distinct species under the multispecies coalescent model (Fujita *et al*. 2012; but see Sukamaran & Knowles 2017).

RESULTS

Locus assembly and characterization

We developed a novel set of genetic markers spread throughout the *Cryptobranchus* genome using double digest restriction site-associated DNA sequencing (Peterson *et al*. 2012). We assembled a data set of n = 74,084 unique genomic loci (approximately 21 million base pairs) per individual. This data set contained relatively low amounts of missing data (approximately 9%). Across all 93 *Cryptobranchus* individuals, n = 71,734 loci were variable. Within hellbenders, on average, each locus contained 4.8 variable nucleotide positions. When considering the *Andrias* outgroups as well, there were a total of 605,033 variable sites across the 74,084 loci. Each genetic marker was sampled from

both diploid chromosomes of each individual, allowing an assessment of heterozygosity at the levels of individuals and populations.

Population genetic structure

Using the set of 71,734 variable loci identified from stacks, a plot of K versus ΔBIC had an inflection point at K = 5, and we retained five clusters for further analyses. We selected a set of 50 principal components PCs of genetic variation representing 92% of the observed variance in allele frequencies among samples. These 50 PCs were then summarized by three linear discriminant (LD) functions retaining 87.5% of observed variance. Plotting individual hellbenders in LD-space clearly shows the separation of the five lineages along the three LD axes (Figure 3.4). Notably, the Little River individual appears distinct from other Tennessee River populations. In contrast to the species tree topologies, the Green River individual clusters with the remainder of Ohio, Allegheny, and Susquehanna River populations in discriminant analysis of principal components. It is also notable that the separation between these five clusters is significantly larger than separation within any single cluster. These five population genetic clusters are also evident from a plot of pairwise genetic distances between individuals (Figure 3.5).

F_{ST} across the Cryptobranchus *genome*

We calculated the fixation index, F_{ST}, across the set of 71,734 loci under the five-cluster population assignments identified through the population genetic clustering

analysis (Figure 3.6). This distribution shows that genetic differentiation is pronounced in *Cryptobranchus* and that there is great variation across the genome in the degree of differentiation. For example, roughly half of all loci have F_{ST} values above 0.33, and approximately 17% of loci have F_{ST} values above 0.5 (F_{ST} greater than 0.25 are typically interpreted as significant genetic differentiation between populations). These results suggest that although there is still some ongoing gene flow within the five putative hellbender species, genetic drift and the fixation of alternative alleles across populations is a strong population genetic process acting in this group. Additionally, the right tail of the F_{ST} distribution between 0.75 and 1.0 could seem to suggest that some loci may be under strong selection between the different lineages, however with so many loci sampled, a more likely explanation is that this pattern reflects the vagaries of genetic drift across the genome.

Haplotype network analysis and species tree estimation

The topologies estimated by these two independent methods are highly concordant and provide new insights into the evolutionary history of *Cryptobranchus*. Both methods support at least five separate lineages of hellbenders, broadly in line with the major watersheds of eastern North America. The haplotype network identified by SplitsTree (Figure 3.7) shows clear clustering of individuals by watershed. The haplotype network and the species tree both support five primary lineages consisting of populations in: 1) the White and Black River drainages in the Ozarks, 2) Kanawha and New River drainages, 3) Tennessee River drainages, 4) Ohio, Allegheny, and Susquehanna River drainages, and 5)

Missouri, Mississippi, and Green River drainages. These lineages are also recapitulated by the SVDQuartets tree (Figure 3.8). Rooting the species tree with the outgroup *Andrias* provides an important context to understand lineage boundaries in this group. With the root position of *Andrias* estimated by this method, our results suggest that the earliest divergence in the ancestors of extant hellbender populations occurred between ancestors of the Ozark lineages and ancestors of all other populations. The next-deepest divergence event took place between the ancestors of the New/Kanawha River populations and ancestors of the Tennessee River, Mississippi/Missouri River, and Ohio River populations. A more recent divergence between Mississippi/Missouri River and Ohio River populations appears to have taken place.

Several patterns emerge from these estimated relationships. Our results reveal that the Atlantic-draining Susquehanna River drainage populations in New York and Pennsylvania are very closely related to populations throughout the Ohio and Allegheny River drainages, likely reflecting one or multiple recent colonization events into the Susquehanna River watershed. Ohio and Allegheny River populations exhibit limited differentiation from each other, although the Licking River populations from Kentucky display some differentiation from other Ohio River populations. Interestingly, the single individual from the Green River in Kentucky clusters with the populations from the Missouri and Mississippi River drainages in the SVDQuartets tree with 100% bootstrap support. The species tree topology of the Tennessee River populations roughly matches the topology of the river network, with the exception of Little River individuals, which appear deeply divergent from other Tennessee River populations. Although previous studies have suggested affinities between Kanawha/New River populations and populations from the

Ozarks, our results demonstrate that these two lineages diverged prior to the other divergences in this genus and that these two lineages are deeply divergent from each other.

Geographic patterns of genetic differentiation

In general, genetic distances within *Cryptobranchus* were not as great as within either species of *Andrias*, or between *Andrias davidianus* and *Andrias japonicus* (Figure 3.9, inset). Not surprisingly, genetic variation between *Cryptobranchus* and *Andrias* was roughly one order of magnitude greater than genetic variation within *Cryptobranchus*, reflecting the deep divergence between these two genera. Genetic variation between individuals within each of our 39 sites and between alleles for individual hellbenders (within individual variation) is represented as triangles in Figure 3.8 and are all clustered x = 0 km. Comparisons between individuals from within each of our five putative lineages are represented as X's in Figure 3.9, whereas comparisons between individuals from different clusters are represented as circles. Although there is no meaningful threshold for what level of genetic differentiation between lineages marks the boundary between intraspecific genetic variation and interspecific genetic variation, it is clear from Figure 3.8 that significantly more genetic variation exists between the five putative lineages of hellbenders than within.

Spatial patterns of genetic diversity

The EEMS genetic diversity results from the two different habitat polygons are largely concordant (Figure 3.10) and point to several regions of high genetic diversity (e.g., the Ozarks, the central Tennessee River drainages in the southern Appalachian Mountains, the Kanawha/New River drainages, and the Allegheny River drainages), and several regions which are relatively depauperate of genetic variation (e.g., the lower and upper Tennessee River drainages, the middle Ohio River drainages, and especially the Missouri/Mississippi drainages in central Missouri). These results suggest that there are several hotspots of genetic variation in the hellbender range, but also highlight several regions of concern. In the context of the five lineages identified here, the Kanawha, Ozark, Ohio, and Tennessee lineages appear to each be centers of moderate to high genetic diversity, while the Mississippi/Missouri lineage appears to be a region of extremely low genetic diversity.

Spatial patterns of gene flow

Similarly to the estimates of the spatial distribution of genetic diversity, we also used EEMS to estimate the spatial distribution of gene flow between demes across the landscape (Figure 3.11). Again, results are largely concordant between the two habitat polygons. It is notable that the regions of lowest gene flow estimated in the 750 deme analysis correspond to the regions excluded between watersheds in the 3,000 deme analysis, and this suggests that these regions do indeed correspond to barriers to gene flow between lineages of hellbenders in different watersheds. Results are plotted on a Log_{10} scale, so the darkest brown regions have effective rates of migration that are approximately

10,000 times lower than the darkest blue regions. The main barriers to gene flow (dark brown color in Figure 3.11) correspond to the boundaries between the Missouri/Mississippi lineage and the Ozark lineage, between the Ohio lineage and the Tennessee lineage, between the Tennessee lineage and the Kanawha/New lineage, and between the Ohio lineage and the Kanawha/New lineage. Interestingly, these results also suggest that migration rates between the adjacent, but hydrologically disconnected, Allegheny and Susquehanna River drainages are more than an order of magnitude higher than expected under an isolation-by-distance model, in line with the possible scenario of one or multiple recent introductions from the Allegheny into the Susquehanna hinted at by the species tree topology. Although the EEMS method has been shown to be robust to patchy and uneven sampling of individuals and to uneven numbers of individuals per deme, remote or isolated samples can have a marginal effect on the estimation of local features (though this is unlikely to mislead the detection of strong barriers).

Topological concordance between phylogeny and river networks

The spatial arrangement of genetic hellbender lineages across the landscape reveals that there is significant disconnect between the contemporary distribution of and the evolutionary relationships among hellbender lineages at broad scales (Figure 3.12). We expected that the contemporary distribution of these aquatic salamanders would correlate with river networks under a model of isolation-by-distance. The co-phylogenetic plot of the species tree topology versus the hierarchical river network topology (Figure 3.13) suggests that there is substantial discordance between these topologies, potentially

reflecting a complex pattern of diversification of different river populations relative to the contemporary drainages in which they are found. We attempted to reject the hypothesis of no correlation between contemporary watershed topology and the river-level lineage topology. The observed Robinson-Foulds (Robinson & Foulds 1981) distance between these two topologies was 34 (out of a maximum value of 66), which lies well outside of a distribution of random topologies generated from 10,000 simulations (P = 0.00099) (Figure 3.14). But perhaps a more informative question is whether the contemporary river network or the paleodrainage network at the time of hellbender divergence are more correlated with the topology of the hellbender phylogeny. While the courses of many of the more fine-scale hydrologic features of eastern North America are difficult to trace in the past, the timing of more broad-scale features such as the Mississippi River, Mississippi Embayment, the Teays River are better known and provide opportunities to test historical riverine connections between contemporarily disconnected hellbender populations. Yet, under the assumption that the appropriate null distribution that implies an expectation of perfect correspondence between the river network and the lineage network lies at the left side of Figure 3.14.

Coalescent species delimitation

Although further analyses may be required to rigorously assess species boundaries under the multispecies coalescent model implemented in BPP, preliminary results across the range of prior settings tested indicate that of the five lineages identified in our other analyses, the Ozark, Kanawha River, and Tennessee River lineages are consistently

recovered as distinct species with posterior probabilities greater than 0.75 (Figure 3.15). Not surprisingly, individuals in *Andrias* and *Cryptobranchus* are found to be different species. Additionally, the Ozark, Kanawha, and Tennessee lineages appear strongly supported as separate species. The putative divergence event between Ohio River and Missouri River populations receives lower posterior support.

DISCUSSION

Genome-scale data generation in a 55 gigabase genome

Our results indicate that reduced representation genome sequencing approaches, when implemented thoughtfully, may be viable options to provide rich information about demography and phylogenetic history in non-model organisms, in spite of massive genome sizes. While our ability to multiplex individuals at both the library preparation and sequencing stages was limited by the large genome, this reduction in efficiency in sampling individuals was largely offset by great returns in terms of numbers of loci. Similar optimization techniques to those used here may be applicable in other non-model organisms with large and complex genomes. It is also notable that of the 74,084 loci which we identified as present in at least 72 of 93 individuals and having credible zygosity and coverage, 71,734 were variable in the global sampling of *Cryptobranchus*. This observation that 96.8% of the genomic regions which we sampled had segregating polymorphic sites was surprising in light of the early allozyme studies (Merkle *et al*. 1977; Shaffer & Breden 1989) in which had suggested that *Cryptobranchus*, and paedomorphic

salamanders in general, had lower levels of nuclear genetic variation than other vertebrates. However, because our results show that the Big Piney and Gasconade Rivers where these studies were conducted have extremely low genetic variation compared to the rest of the range, this may have been more of an artifact of the specific populations examined rather than reflecting the underlying genetic diversity in *Cryptobranchus*. The observed level of variation in our set of loci could potentially reflect pervasive sequencing errors in our read data or mis-assembly of loci, although we attempted to mitigate both of these potential sources of error by examining patterns of SNP variation across sites in our loci and truncating reads accordingly (Figure 3.3), and by exploring a range of assembly parameters and their impacts on patterns of variation, respectively. Genomic data for phylogenetics, demographic studies, and species delimitation are only as useful as the information content which they bring to bear on the specific questions at hand. We did not know *a priori* the extent to which these specific anonymous loci would be variable, but we did have expectations that we would recover hundreds of thousands of loci for each individual. It would appear that the levels of standing genetic variation in *Cryptobranchus* are higher than indicated by previous studies. These results hint at the possibility that anonymous genomic data may be particularly informative in organisms with large genomes, potentially offsetting some of the upfront challenges required to optimize ddRAD, or similar, markers in these challenging taxa.

Factors influencing diversification in hellbenders

Our results suggest that allopatric isolation and divergence may account for a substantial amount of the differentiation between hellbender lineages. Our estimates of gene flow between geographically proximate populations indicate that at least four of these five lineages are effectively reproductively isolated from each other. Our results also demonstrate that ancient watershed architecture may explain some of the deepest divergences within hellbenders. But perhaps a more informative question is whether the contemporary river network or the paleodrainage network at the time of hellbender divergence are more correlated with the topology of the hellbender phylogeny. While the courses of many of the more fine-scale hydrologic features of eastern North America are difficult to trace in the past, the timing of more broad-scale features such as the Mississippi River, Mississippi Embayment, the Teays River are better known (e.g., Galloway *et al*. 2011) and provide opportunities to test historical riverine connections between contemporarily disconnected hellbender populations. Ancient phylogeographic connections between populations which were once connected by the Teays River have been proposed in other aquatic salamander species (Kozak *et al*. 2006), and this pattern may also explain the topology toward the base of the *Cryptobranchus* phylogeny.

Putative hellbender species boundaries

Taken together, our results reveal that there is substantial population genetic and phylogenetic structure within *Cryptobranchus*, corroborating previous studies. However, the genomic scale of our data set, along with comprehensive sampling across the geographic distribution of hellbenders allows us to reveal the evolutionary history of this

group in much greater detail. Because different regions of the genome may each have their own specific evolutionary histories and may each provide different, sometimes conflicting, interpretations, our dense sampling of approximately 75,000 genomic regions allows us to take this variation in phylogenetic signal into account when estimating phylogenetic and demographic parameters of interest.

The species tree analyses clearly point to five evolutionarily distinct lineages of hellbenders which are highly supported and are each reciprocally monophyletic with respect to each other. With the exception of the placement of the Green River individual from Kentucky, these five lineages are recapitulated in an independent discriminant analysis of genetic variation. Pairwise F_{ST} estimates across the genome indicate that large proportions of the genome have differentiated between these five lineages, and suggest that the forces of natural selection may also be implicated in driving divergence in hellbenders, in addition to genetic drift. Analysis of pairwise genetic divergence in the context of geographic (in-stream) distance highlights that much of the genetic variation in hellbenders occurs between the five lineages we have identified and also suggests that, despite large ranges of geographic distance within these five lineages, genetic distances are roughly constant across inter-lineage comparisons. This suggests that these lineages are each genetically cohesive and that variation within lineages is much lower than variation between lineages. (The Ozark and Tennessee River lineages are exceptions here, and it may be that additional population genetic structure within these lineages is leading to that outlier pattern). Our analysis of the spatial distribution of genetic variation suggests that four of the five lineages represent hotspots of genetic diversity, but that the fifth (Mississippi/Missouri lineage) has severely reduced genetic variation. Our analysis of the

spatial distribution of migration rates suggests that significant barriers to gene flow exist between the five lineages we have identified, leading to reductions in gene flow of up to four orders of magnitude between lineages compared to rates of gene flow within lineages. And preliminary coalescent species delimitation analyses appear to support the species status of at least four of these five lineages of hellbenders.

Conservation implications

Although more work remains to be done to test whether these different lineages represent distinct species, our analyses to date imply that species diversity in *Cryptobranchus* has been underestimated. Our tentative, most conservative estimate is that *Cryptobranchus* contains at least five distinct species which are each on their own evolutionary trajectories, which are effectively reproductively isolated and no longer exchange genes, and which each have smaller effective and census population sizes than current estimates for hellbenders as a single species. The Ozark and Kanawha species appear to still retain large amounts of genetic variation, while the Ohio and Tennessee species appear to have patches of high and lower genetic variation, and the Mississippi/Missouri species appears to be very genetically depauperate. A hypothetical species delimitation model based on these results is depicted in Figure 3.16. These results imply that several states actually host multiple species of hellbenders, and that management strategies based on state, and not watershed, boundaries may warrant reconsideration. Nonetheless, given current population trends, regardless of the true number of hellbender

species, a model of zero hellbender species may be impossible to reject within the foreseeable future.

Table 3.1. Geographic sampling of *Cryptobranchus* individuals (coordinates have been omitted to safeguard sensitive population information).

STATE	SITE NAME	INDIVIDUALS	HUC_2 USGS WATERSHED
AL	Flint River	1	Tennessee
AR	Eleven Point River	2	Arkansas-White-Red
GA	Cooper Creek	2	Tennessee
GA	Fightingtown Creek	2	Tennessee
GA	Helton Creek	1	Tennessee
GA	Hiwassee River	2	Tennessee
GA	Nottley River	1	Tennessee
GA	Rock Creek	2	Tennessee
GA	Swallow Creek	1	Tennessee
GA	Tumbling Creek	3	Tennessee
IN	Blue River	4	Ohio
KY	Green River	1	Ohio
KY	Kinniconick Creek	1	Ohio
KY	Licking River (NFTC)	7	Ohio
MO	Big Piney River	4	Missouri
MO	Current River	3	Arkansas-White-Red
MO	Eleven Point River	2	Arkansas-White-Red
MO	Gasconade River	4	Missouri
MO	Meramec River	3	Upper_Mississippi
MO	Niangua River	3	Missouri
MO	North Fork White	10	Arkansas-White-Red
MS	Bear Creek	1	Tennessee
NC	Avery Creek	3	Tennessee
NC	Brasstown Creek	3	Tennessee
NC	Cane River	3	Tennessee
NC	Cartoogechaye Creek	1	Tennessee
NC	Cullasaja Creek	2	Tennessee
NC	Davidson River	1	Tennessee
NC	Deep Creek	2	Tennessee
NC	East Fork French Broad	3	Tennessee
NC	Fires Creek	3	Tennessee
NC	Hanging Dog Creek	2	Tennessee
NC	Looking Glass Creek	2	Tennessee
NC	North Fork French Broad	3	Tennessee
NC	North Fork Mills River	3	Tennessee
NC	Oconaluftee River	1	Tennessee
NC	Shooting Creek	3	Tennessee
NC	Shuler Creek	1	Tennessee
NC	Snowbird Creek	3	Tennessee
NC	South Fork Mills River	3	Tennessee
NC	South Fork New River	3	Ohio
NC	Tuckasegee River	3	Tennessee
NC	Tusquitee Creek	1	Tennessee
NC	Valley River	2	Tennessee

Table 3.1 (continued).

NC	Wayah Creek	2	Tennessee
NC	West Fork French Broad River	2	Tennessee
NY	Iscua Creek	2	Ohio
NY	Olean River	1	Ohio
NY	Oswayo Creek	2	Ohio
NY	Susquehanna River	1	Mid_Atlantic
OH	Captina Creek	2	Ohio
OH	Cross Creek	2	Ohio
OH	Kokosing River	1	Ohio
OH	West Fork Little Beaver Creek	2	Ohio
PA	Bear Creek (Lehigh River)	1	Mid_Atlantic
PA	Clarion River	1	Ohio
PA	French Creek	3	Ohio
PA	Little Mahoning Creek	3	Ohio
PA	Loyalsock Creek	1	Mid_Atlantic
PA	Tionesta Creek	3	Ohio
PA	Tubmill Creek (Conemaugh River)	3	Ohio
PA	West Fork Susquehanna River	6	Mid_Atlantic
TN	Beaverdam Creek (Holston River)	1	Tennessee
TN	Big Richland Creek (Lower Tennessee River)	2	Tennessee
TN	Big Swan Creek (Duck River)	1	Tennessee
TN	Buffalo River	2	Tennessee
TN	Clinch River	1	Tennessee
TN	Doe River (Watuga River)	4	Tennessee
TN	Duck River	2	Tennessee
TN	Factory Creek	2	Tennessee
TN	Little River	2	Tennessee
TN	Powell River	1	Tennessee
TN	Roaring River	3	Ohio
TN	Rough Creek (Ocoee River)	2	Tennessee
TN	Tellico Creek (Little Tennessee River)	3	Tennessee
TN	Watuga River	3	Tennessee
TN	White Oak Creek (Duck River)	1	Tennessee
VA	Holston River (South Fork)	3	Tennessee
VA	New River	3	Ohio
WV	Back Fork Elk River	3	Ohio
WV	Buffalo Creek	2	Ohio
WV	East Fork Greenbriar	1	Ohio
WV	Holly River	1	Ohio
WV	Middle Island Creek	1	Ohio
WV	Shavers Fork Cheat River	2	Ohio
WV	South Fork Hughes	1	Ohio
WV	Wheeling Creek	2	Ohio
WV	Williams River	1	Ohio

Figure 3.1. Geographic sampling of *Cryptobranchus* individuals. Points are color coded by major watershed.

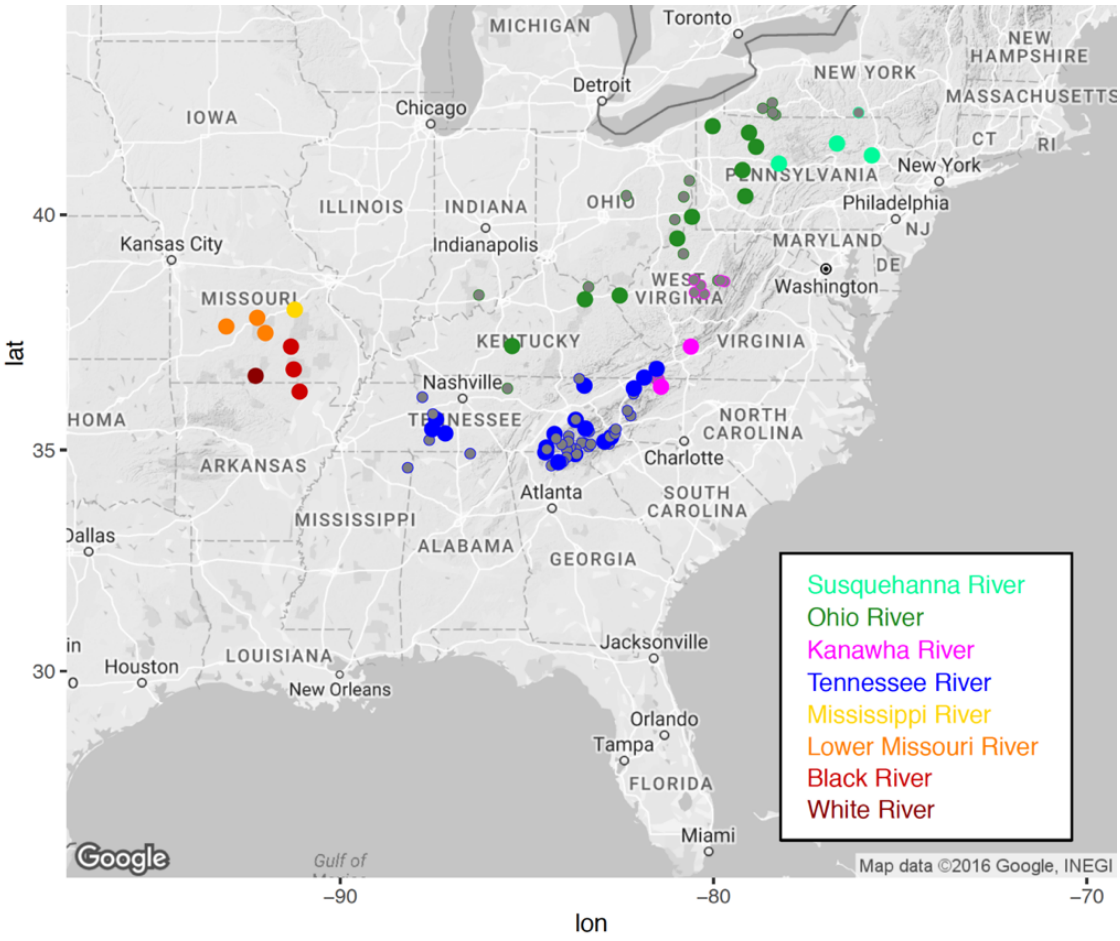
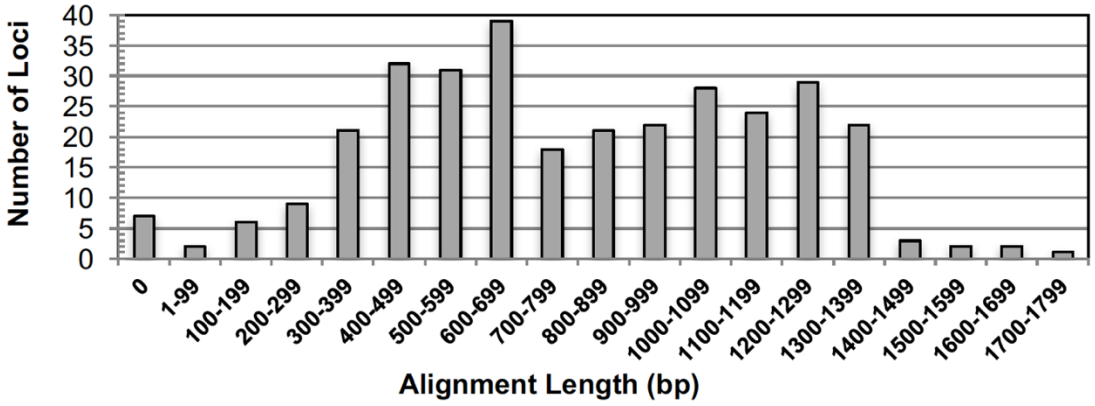


Figure 3.2. Summary of exons targeted by sequence capture in cryptobranchid salamanders.

A. Targeted sequence capture is an effective method for generating sequence data in cryptobranchid salamanders. Approximately 91% (312 / 343) of target loci are recovered across multiple *Cryptobranchus* and *Andrias* individuals.



B. However, these loci are relatively uninformative about more shallow-scale relationships within hellbenders. Across 319 loci (with a mean length of over 1,300 bp) only 506 parsimony-informative sites are present across three of the most deeply divergent hellbender lineages.

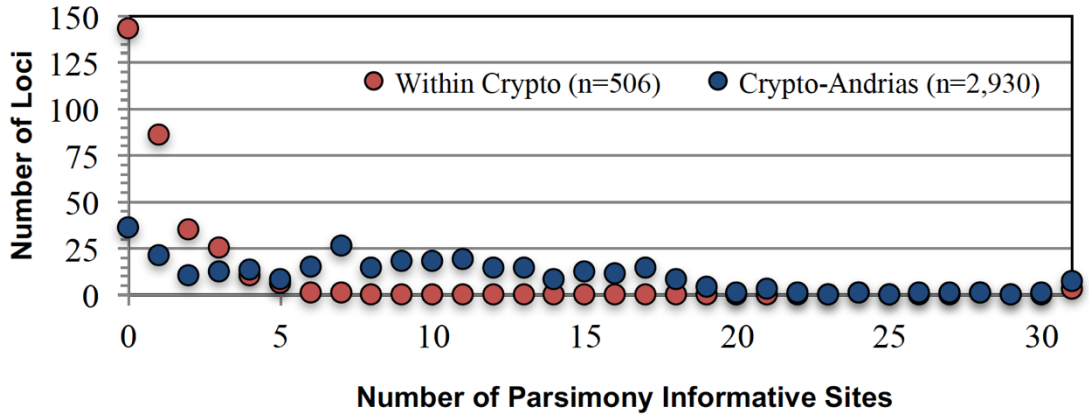


Figure 3.3. Distribution of single nucleotide polymorphisms (SNPs) across sites in ddRAD loci for comparisons across 93 *Cryptobranchus* (blue), and comparisons across 93 *Cryptobranchus* and four *Andrias* (red). A schematic representation of a stitched ddRAD locus is provided at the top of the figure for context. The first six and last four bases in each ddRAD locus represent portions of the *SphI* and *EcoRI* recognition sequences, respectively. As expected, these short, flanking regions are invariant. The last two bases of the 3' regions of Illumina read 1 and read 2 had significantly elevated SNP proportions, likely due to sequencing errors, and these regions were removed from the final stitched loci to avoid analyzing potentially spurious SNPs.

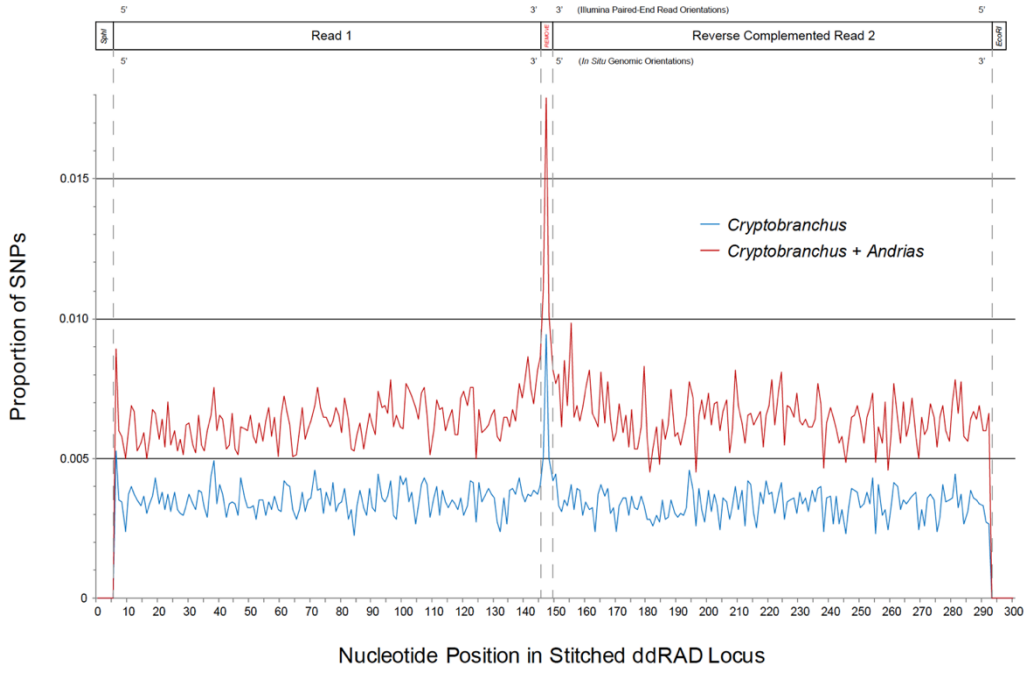


Figure 3.4. Discriminant analysis of genetic variation in *Cryptobranchus* reveals five distinct genetic clusters.

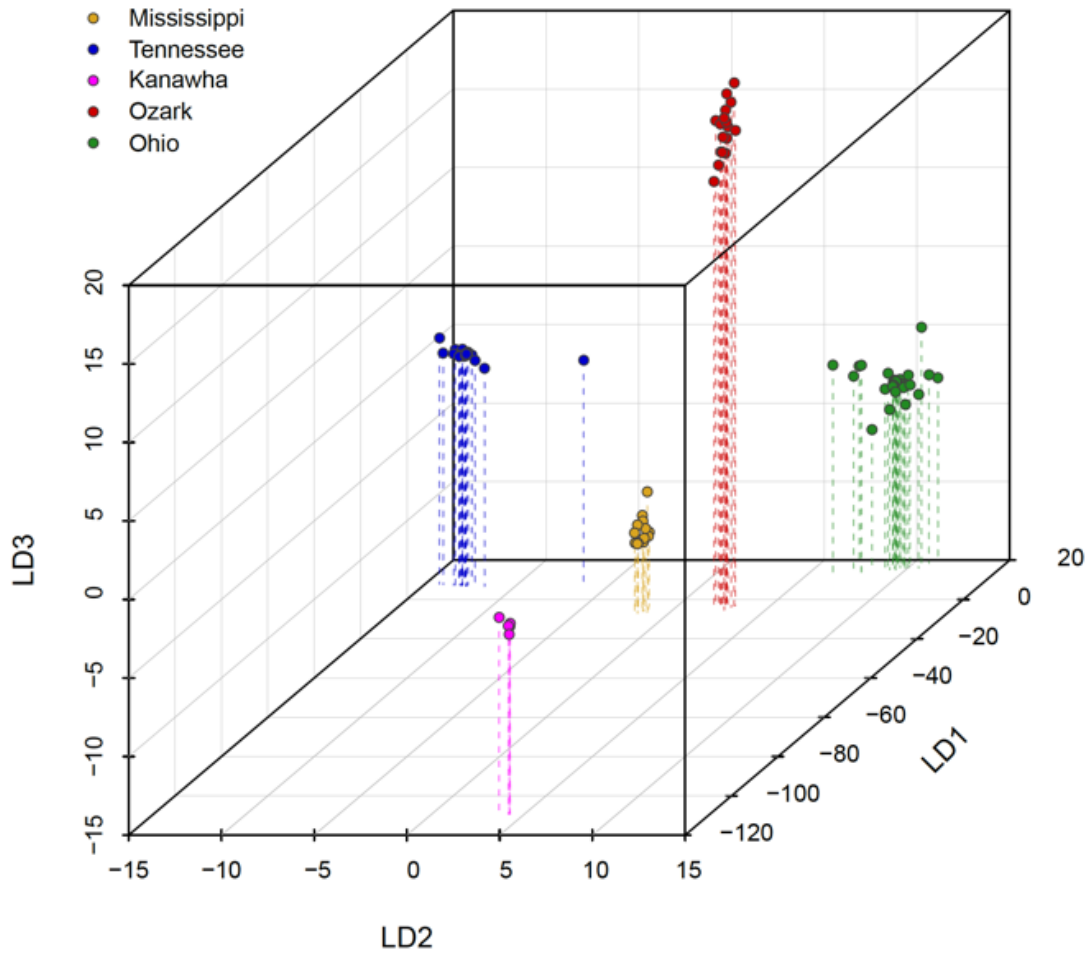


Figure 3.5. Pairwise genetic distances inferred between 93 *Cryptobranchus* individuals. The five black squares along the diagonal highlight genetic variation within each of the five putative species. Genetic differentiation between these lineages is much greater than within lineages. The Green River (KY) individual has been included with the Ohio River clade, in line with the DAPC analyses (and in contrast to the species tree analyses). Note that there is additional population genetic differentiation within the Tennessee River lineage and within the Ozark lineage.

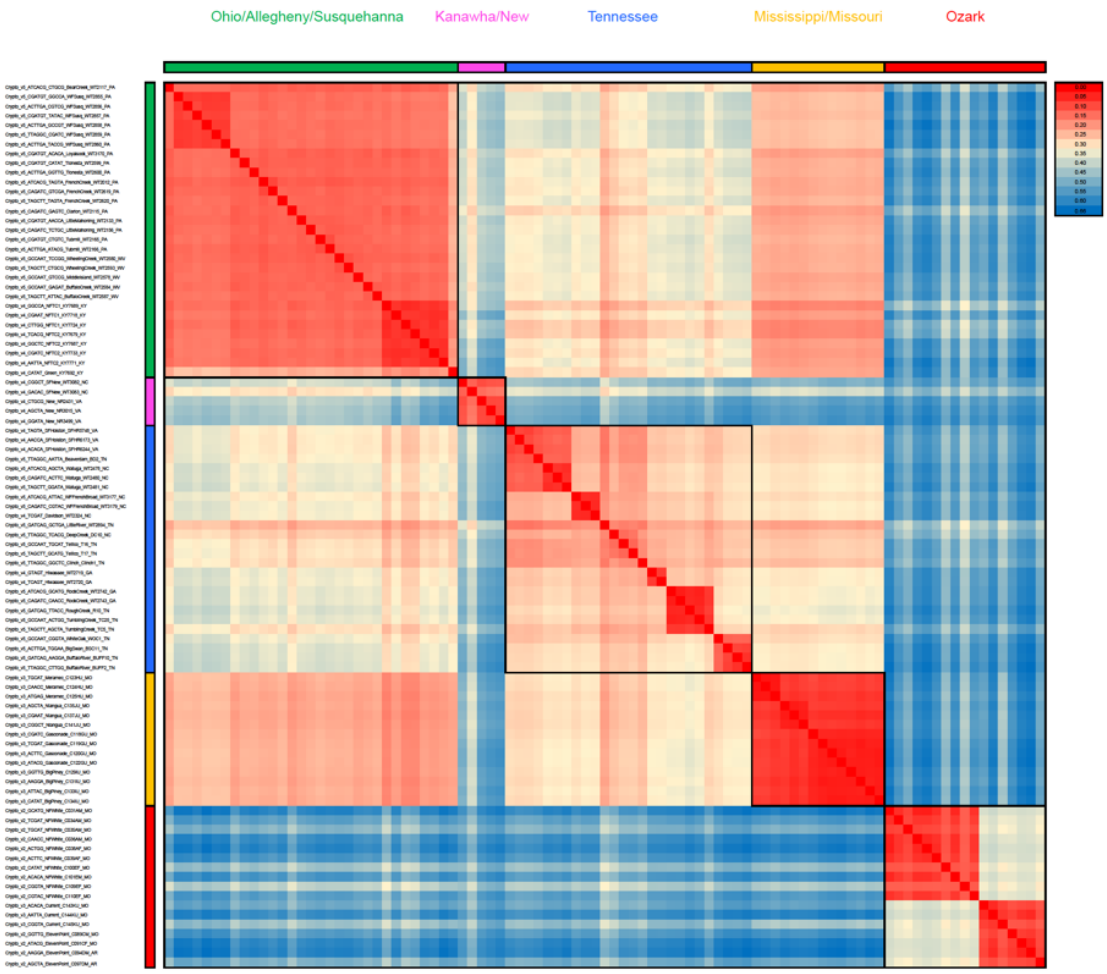


Figure 3.6. Genome-wide distribution of F_{ST} across 71,734 loci. F_{ST} values close to zero indicate no differentiation between populations, whereas F_{ST} values close to one suggest a larger degree of differentiation between lineages. Per-locus F_{ST} values were calculated in the Adegenet R package.

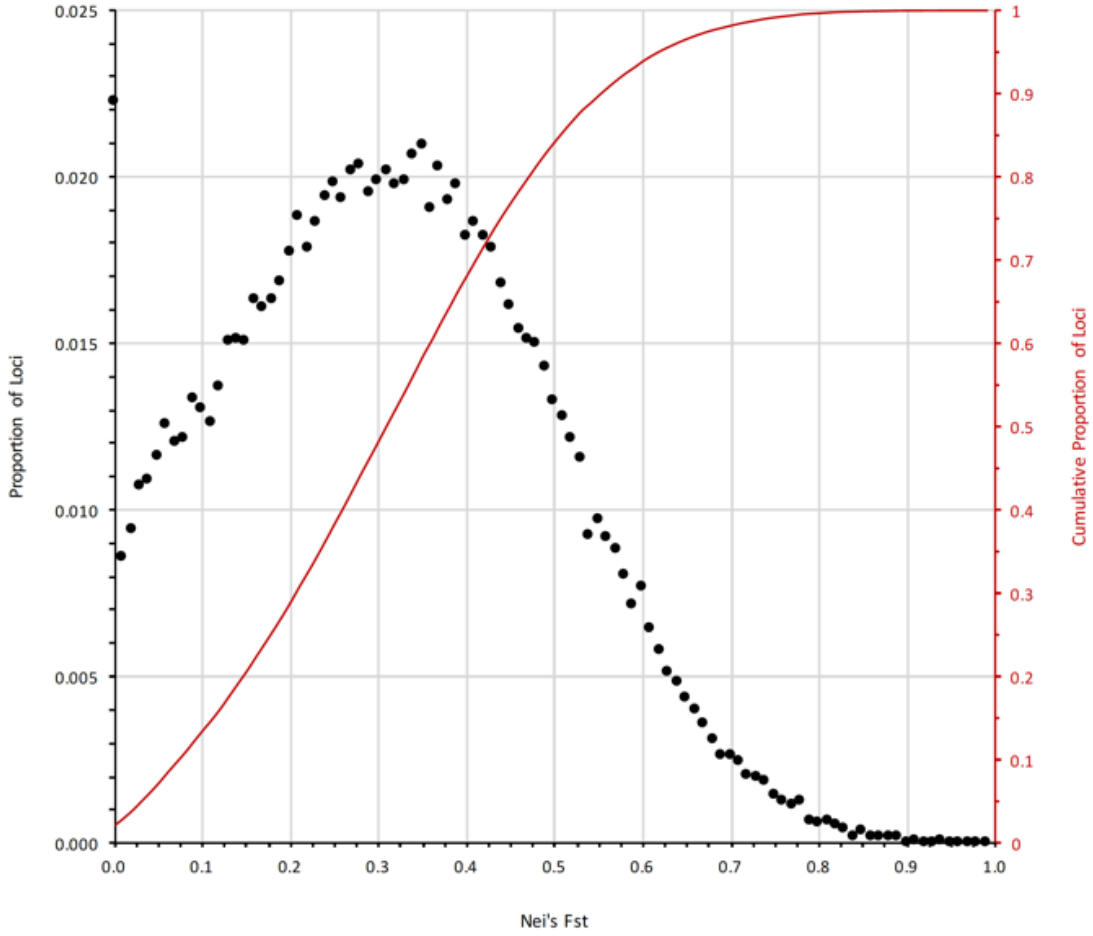


Figure 3.7. SplitsTree neighbor-joining multilocus haplotype network with convex-hull representation.

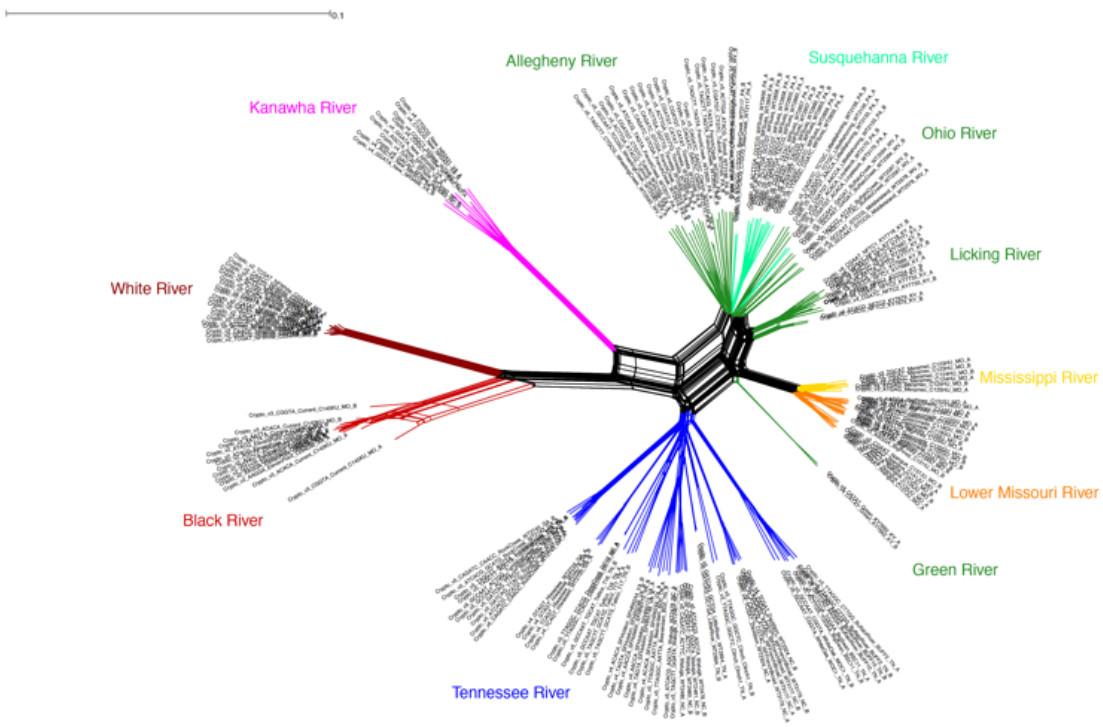


Figure 3.8. SVDQuartets species tree for 34 lineages of *Cryptobranchus*. The tree has been rooted on the branch leading to *Andrias*. Numbered branches indicate bootstrap support over 100 replicates, and numbered nodes denote the hierarchical validation scheme implemented in BPP.

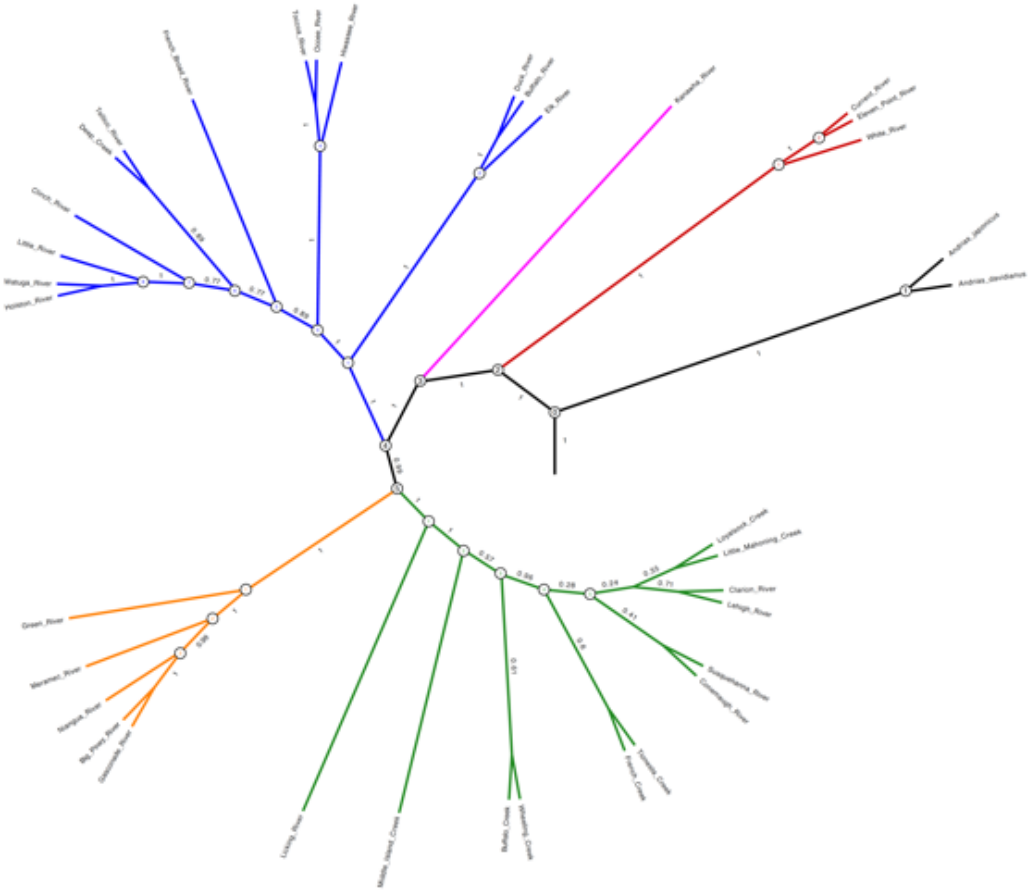


Figure 3.9. Relationship between genetic distance and geographic distance in *Cryptobranchus*. Pairwise corrected genetic distances (allele-averaged Jukes-Cantor distances) are plotted against minimum resistance geographic distance along stream courses between pairs of sampling points. Triangles represent genetic differentiation within the 39 sites sample, and within individuals. X's represent contrasts between individuals from within each of the five putative hellbender species. Circles represent contrasts between individuals from different putative species. The inset plot shows the distribution of genetic distances within *Cryptobranchus* in the context of genetic distances within *Andrias* and between *Cryptobranchus* and *Andrias*.

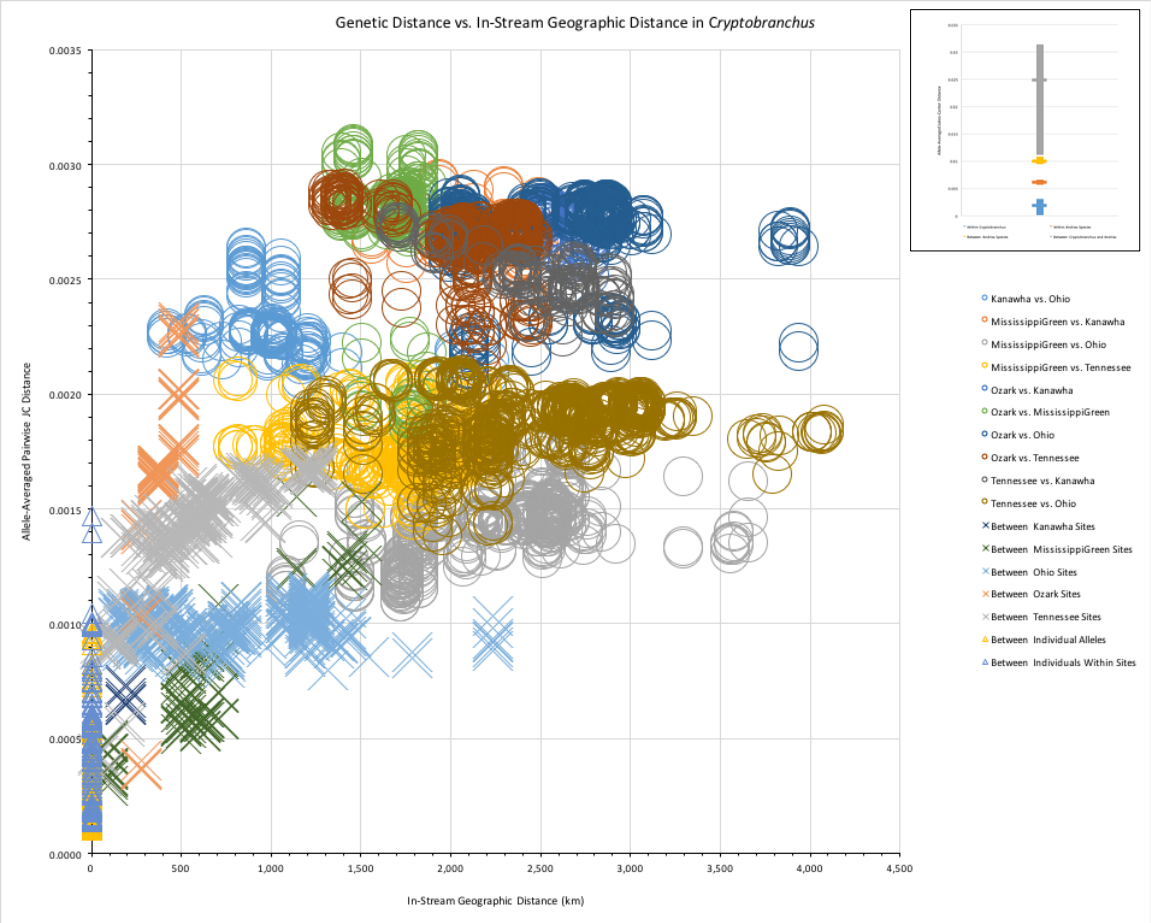
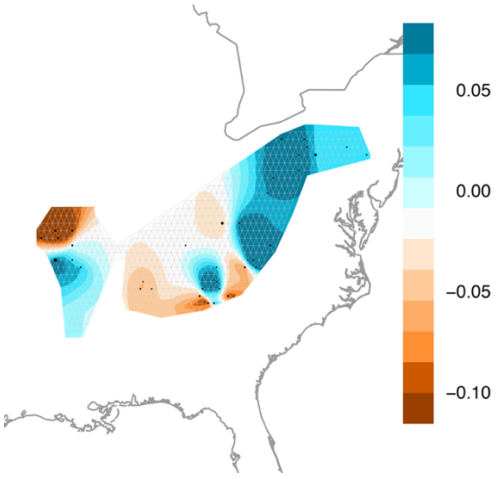


Figure 3.10. Spatial distribution of genetic diversity in *Cryptobranchus* inferred in EEMS. Estimates are presented for (A) a moderate complexity habitat polygon of 750 demes and (B) for a more complex habitat polygon of 3,000 demes. Results from 16 independent runs were averaged and visualized with rEEMSplots. The genetic diversity contours are plotted on a Log_{10} scale.

A. 750-deme grid, simple outline.



B. 3,000-deme grid, complex outline.

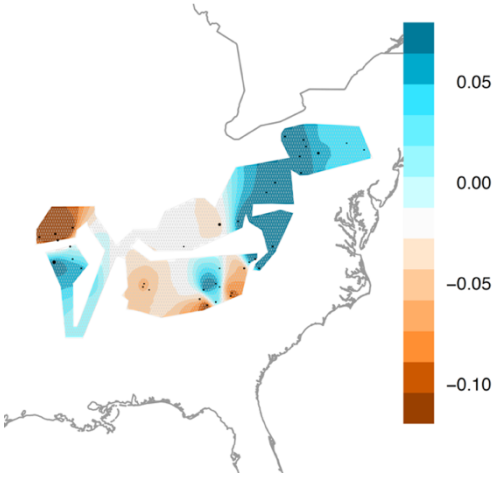
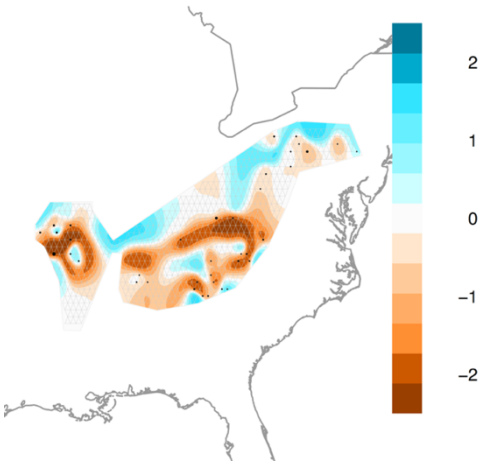


Figure 3.11. Spatial distribution of gene flow in *Cryptobranchus* inferred in EEMS. Estimates are presented for (A) a moderate complexity habitat polygon of 750 demes and for (B) a more complex habitat polygon of 3,000 demes. Results from 16 independent runs were averaged and visualized in rEEMSplots. The migration rate contours are plotted on a Log_{10} scale, and can be thought of as measuring deviations from a pure isolation-by-distance model of intraspecific gene flow.

A. 750-deme grid, simple outline.



B. 3,000-deme grid, complex outline.

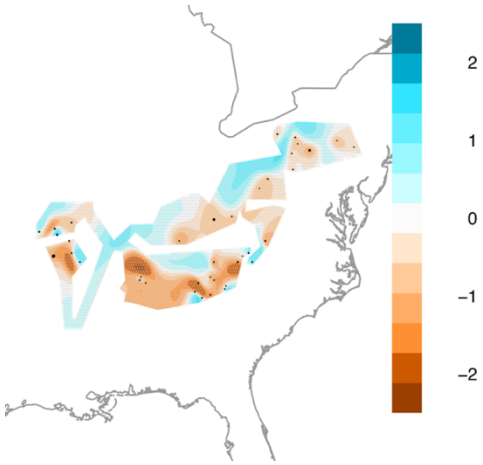


Figure 3.12. Phylogeographic patterns in *Cryptobranchus* suggest a complex relationship between the contemporary geographic distribution of hellbender lineages and the underlying evolutionary relationships in this genus. The geographically proximate hellbender populations south and north of the Ozark Plateau (red and yellow points, respectively) are evolutionarily quite distant from each other, with the former recovered as sister to all other populations and the latter placed in a clade with Ohio and Susquehanna River populations. Additionally, the Kanawha River lineage is recovered as sister to a clade containing the geographically adjacent Ohio River and Tennessee River populations.

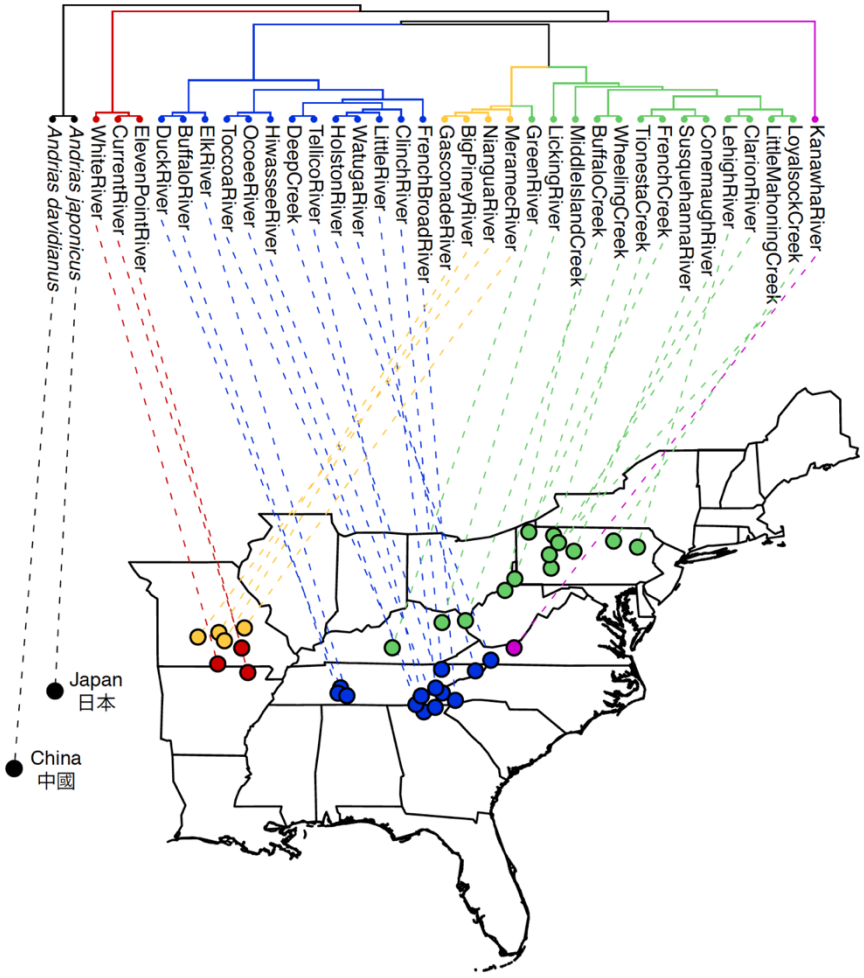


Figure 3.13. Co-phylogenetic plot relating river-level hellbender lineages to the river network.

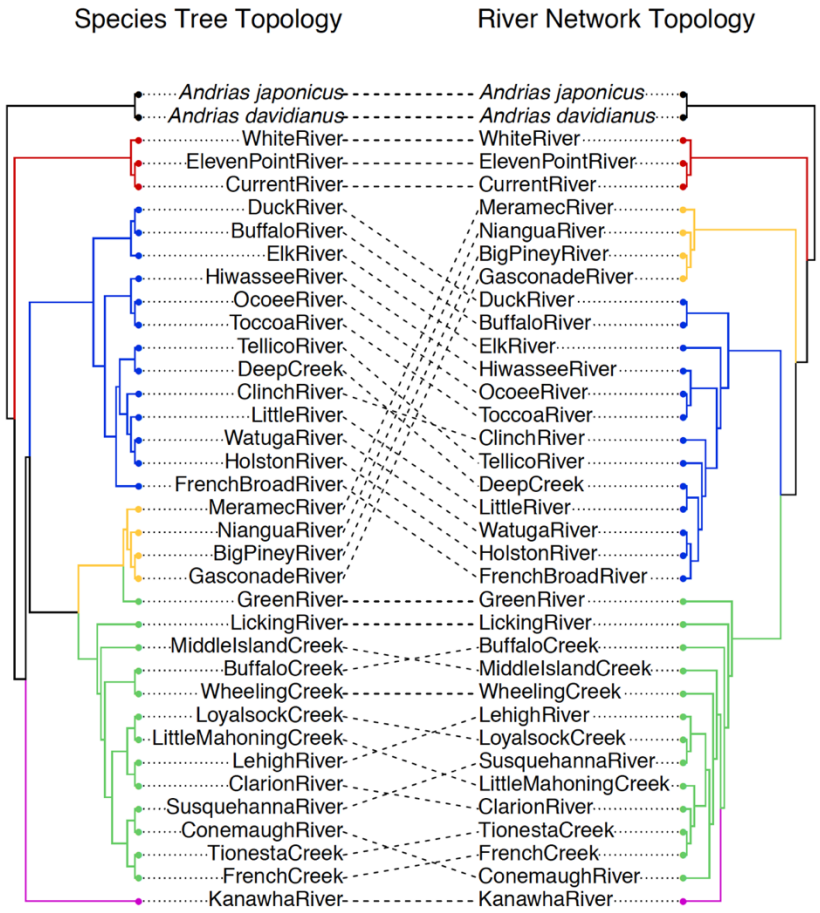


Figure 3.14. Co-phylogenetic test of correlation between river network topology and river-level lineage topology. The observed Robinson-Foulds distance between the hellbender phylogeny and the river networks lies well outside of a null distribution of values for random tree topologies, rejecting the hypothesis of no correlation between river network connectivity and phylogenetic relationships between lineages (P = 0.00099).

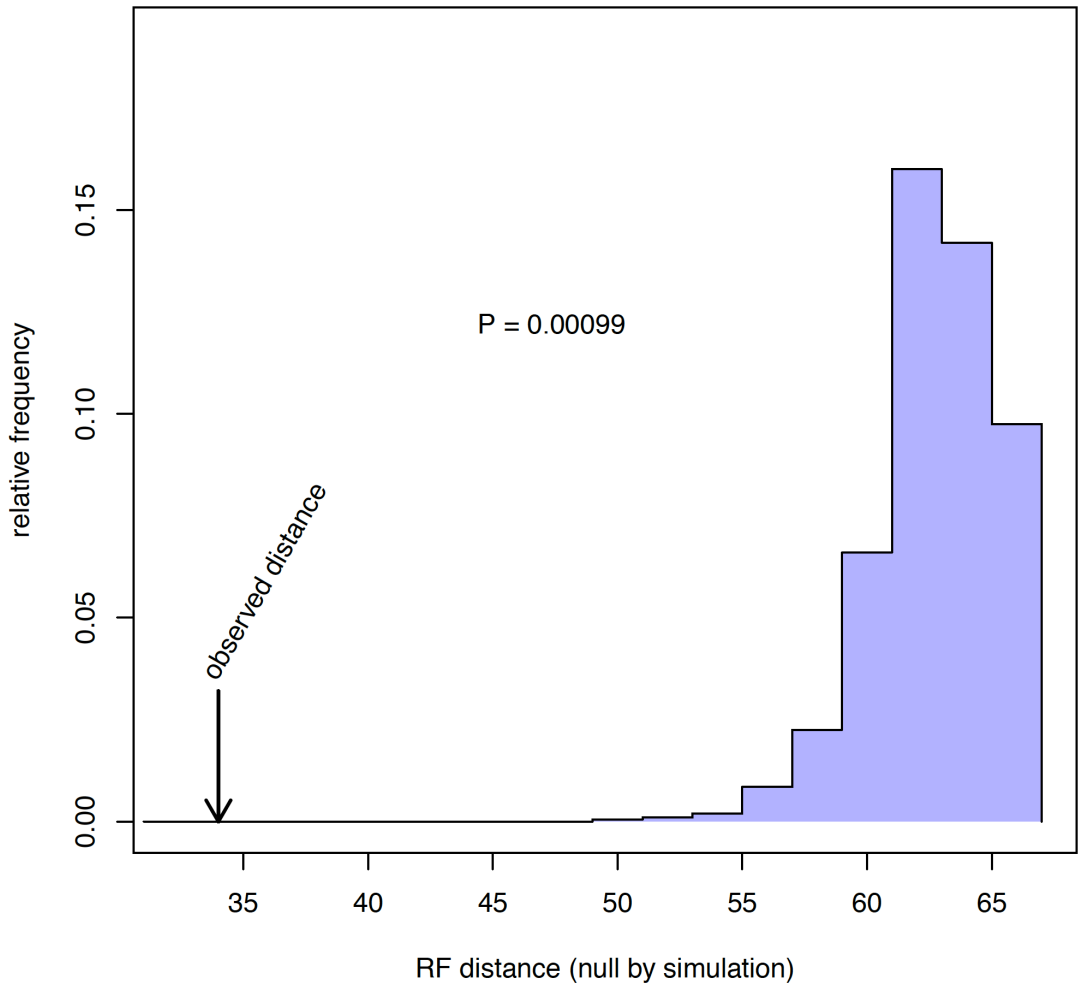


Figure 3.15. Species delimitation results for *Cryptobranchus* in BPP. Values at nodes represent the average posterior probability across ten replicates that a particular bifurcation is present in the species tree. Not surprisingly, individuals in *Andrias* and *Cryptobranchus* are found to be different species. Additionally, the Ozark, Kanawha, and Tennessee lineages appear strongly supported as separate species. The putative divergence event between Ohio River and Missouri River populations receives lower posterior support.

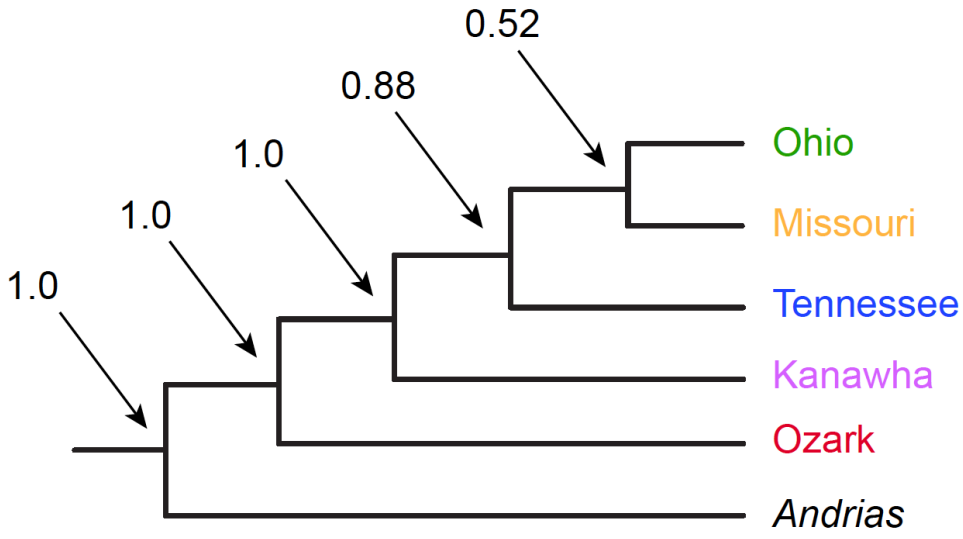
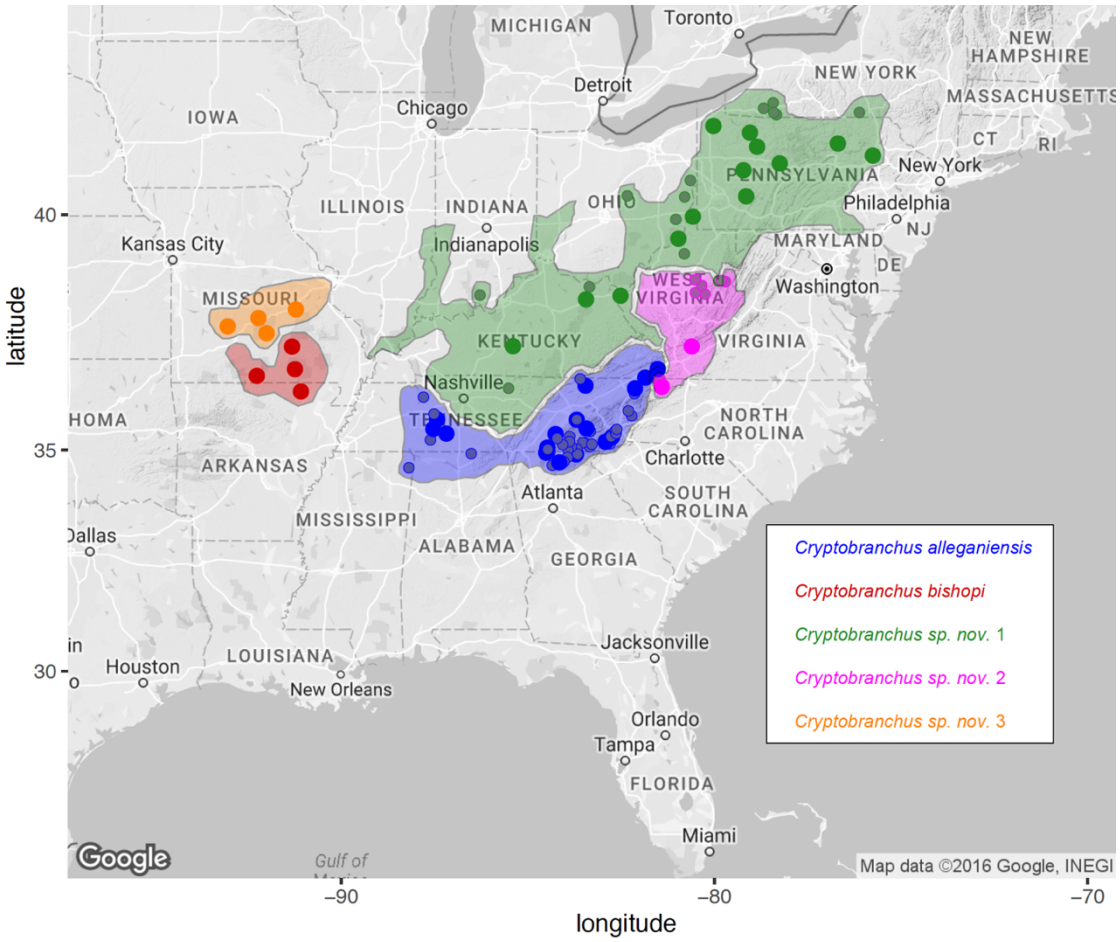


Figure 3.16. Putative species boundaries in *Cryptobranchus*. Because the type locality for *Cryptobranchus alleganiensis* is described from within the Tennessee River drainage, this lineage would retain the original species epithet. The lineage from the Ozarks would be elevated from *C. a. bishopi* to *C. bishopi*. The remaining three lineages will require additional literature review to assess whether previous authors have applied valid names which would take precedence, or whether new names could be created for the species descriptions to be valid under the International Code of Zoological Nomenclature.



CHAPTER FOUR

Genome scans reveal a conserved system of female heterogamety across the deeply divergent salamander family Cryptobranchidae

ABSTRACT

Recent investigations have revealed that both the mechanisms of genetic sex determination and the lability of these systems vary widely across vertebrates. Yet, much progress remains to be made in understanding systems of genetic sex determination in non-model organisms, especially those with homomorphic sex chromosomes and/or large genomes. We used reduced representation genome sequencing to investigate genetic sex determination in the salamander family Cryptobranchidae (genera *Cryptobranchus* and *Andrias*), which typifies both of these inherent difficulties. We sequenced hundreds of thousands of anonymous genomic regions in a panel of known-sex cryptobranchids and characterized patterns of presence/absence, inferred zygosity, and depth of coverage across these loci. These results allowed us to test the alternative hypotheses of either male- or female-heterogamety, demonstrating that all recognized species of this family possess a ZZ/ZW system of female heterogamety which has likely been conserved over approximately 60 million years of evolution. Additionally, we report a highly reliable and non-invasive PCR-based assay for sex diagnosis in *Cryptobranchus* and *Andrias* which has utility for research and conservation efforts with these endangered salamanders. These results have significant implications for cryptobranchid conservation because, previously,

it was very difficult to reliably distinguish males from females in the wild or in captive assurance populations due to delayed sexual maturity and a narrow annual time window of morphological distinctiveness. This approach to characterize the mode of genetic sex determination and to identify and interrogate putative sex-linked genomic regions in non-model taxa holds potential to inform basic and applied studies of demography, population biology, and chromosome evolution in a wide range of species.

INTRODUCTION

The existence of discrete sexes and has ultimately played a major role in generating and maintaining much of the genetic variation out of which natural selection and genetic drift have shaped the diversity of life (Charlesworth & Mank, 2010). The phenotype of sex plays is relevant to numerous areas of organismal biology and sex determination in vertebrates involves a complicated cascade of different agents, and may include genetic factors (Smith *et al*. 2009) or environmental factors (Gallego-García & Páez 2016; Santoyo-Brito *et al*. 2017), or an interplay of both (Matsumoto *et al*. 2013). Among genetic sex determination systems in vertebrates, either male- (XX/XY) or female-heterogamety (ZZ/ZW) predominate (Ezaz *et al.* 2006). In the case of homomorphic sex chromosomes, it can often be difficult to distinguish female- from male-heterogamety using traditional genetic tools. Genome-scale data are now available to inform the search for sex-linked genetic regions in non-model taxa, by allowing access to vast numbers of genetic markers which may happen to be in linkage with sex. Bioinformatic evaluation of patterns of presence and absence, inferred zygosity, and depth of sequencing coverage in known-sex

individuals can be used to identify putative sex-linked loci and to test alternative hypotheses of female- or male-heterogamety in non-model taxa (Gamble 2016; Gamble & Zarkower 2014), and these approaches hold great potential for accelerating our understanding of sex determination systems (e.g., Gamble *et al*. 2015; Montiel *et al*. 2017; Rovatsos & Kratochvíl 2017; Smith & Voss 2009).

Mammals and birds are among the best-studied clades with respect to the genomic underpinnings of sex determination (Charlesworth & Mank 2010). All eutherian mammals have an XY/XX system of male heterogamety. Birds, in contrast, have a ZZ/ZW system of female heterogamety (Smith *et al*. 2009). Whether or not the *Gallus* W chromosome is homologous to the *Homo* Y and whether the *Gallus* Z is homologous to the *Homo* X remain controversial among sex chromosome researchers (Ezaz *et al.* 2016). However, it is generally assumed that several independent origins of different sex determination systems have occurred between the deepest animal lineages. Data from additional vertebrate taxa, especially amphibians, would provide important contrasts to inform ongoing debate over sex chromosome synteny among highly divergent lineages. Amphibian sex determination systems are in general more labile and more poorly understood than mammalian or avian systems. Among frogs, multiple transitions to and from male- and female-heterogametic sex determination appear to have taken place (Schmid & Steinlein 2001; Nakamura 2009). Likewise in salamanders, XY and ZW systems are spread throughout the phylogeny (Sessions 2008). Little is known about the actual composition of sex-specific chromosomal regions in salamanders beyond the near-model axolotl (*Ambystoma mexicanum*). Smith & Voss 2009 and Keinath *et al*. 2017 both provided some insights into sex chromosome

evolution in the family Ambystomatidae, which is approximately 250 million years divergent from Cryptobranchidae (Roelants *et al*. 2007).

Given the variation in male- and female-heterogamety across salamanders (Sessions 2008), it is reasonable to expect that sex determination involves a genetic component (but see Nakamura 2013), and to our knowledge, no study has proposed a purely temperature-dependent mechanism of salamander sex determination. Although the types of genomic regions implicated in sex determination share homology over deep evolutionary timescales (Charlesworth & Mank 2010; Ezaz *et al*. 2016; Gamble & Zarkower 2012; Graves & Peichel 2010), transitions between XY and ZW systems are widespread across more shallow-scale lineages (Furman & Evans 2016; Gamble *et al*. 2015; Stöck *et al*. 2011; Stöck *et al*. 2013, but see also Rovatsos *et al*. 2015). Transitions from homomorphic to heteromorphic sex chromosomes (and the converse) are also known (e.g., Rodrigues *et al*. 2014). A correlation between the presence of ZW sex chromosome systems and limited sexual dimorphism (as is the case in hellbenders) has recently been proposed (Adkins-Reagan & Reeve 2014). Genetic linkage map construction from genome-wide markers is a method which has been used to identify sex-linked regions (e.g., Cano *et al*. 2011, Keinath *et al*. 2015), but because access to a set of siblings and their parents was not possible in the case of hellbenders, we focused on alternative methods.

Few genomic resources existed for cryptobranchid salamanders at the inception of this study (but see Che *et al*. 2014, Fan *et* al. 2015, Qi *et al*. 2016), and so we sought to develop both transcriptomic and reduced representation genomic libraries for *Cryptobranchus de novo*. We used reduced representation genome sequencing (Peterson *et al*. 2012) to investigate the mode of sex determination in the imperiled North American

hellbender salamander (genus *Cryptobranchus*), a non-model species with a 55 Gb nuclear genome (Gregory 2017). We tested the alternative hypotheses of female- or male-heterogamety in hellbenders, and in their closest extant relatives the Asian giant salamanders (genus *Andrias*). These two genera are both assumed to have a ZW system of female-heterogametic sex determination, although the putative sex chromosomes are possibly homomorphic, making it difficult to identify specific sex-linked loci. These large, obligately aquatic salamanders were historically widespread across streams and rivers in eastern and central North America (Nickerson & Mays 1973). Wild populations have been in sharp decline across their range for the past several decades (Pitt *et al*. 2017; Wheeler *et al*. 2003), and today, numerous *in situ* and *ex situ* conservation and management efforts are underway to attempt to stabilize wild populations and to establish captive breeding populations for eventual re-release. Traditionally, determining sex ratios or population demographic parameters for wild or captive hellbender populations has been very difficult due to delayed sexual maturity (4-7 years) and a narrow annual time window of morphological distinctiveness during the breeding season when males express a swollen cloaca (Nickerson & Mays 1973). Across most age classes and most times of year, morphological sex diagnosis has limited utility in hellbenders.

Ultrasound examination of gonads has been used to determine sex in adult hellbenders, but this technique is subject to individual interpretation and may not be effective outside of the reproductive season. Laproscopy may potentially reveal sex in adult animals regardless of the time of year (Roth & Obringer 2003), but this technique can be highly invasive and cannot be used on smaller or wild individuals. Previous work has used serum calcium level differences (Nickerson & Mays 1973) to discern females from males,

but there are numerous advantages to a genetic sex assay such as effectiveness across all age classes and the ability to analyze banked tissue samples. Assuming that the sex chromosomes could be differentiated reliably by visual means, cytogenetic techniques such as karyotype analysis could potentially be used to diagnose sex in hellbenders. However, these techniques require access to fresh material for tissue culture (generally difficult for endangered species) and are notoriously low-throughput. Additionally, it is unclear whether karyotypic differences can reliably distinguish the two sexes of hellbenders (Morescalchi *et al*. 1977 implies homomorphic ZW, but Zhu *et al*. 2002 implies heteromorphic XY). A genetic sex assay for hellbenders, similar to those already widely employed in avian taxa (Ellegren 1996), would ameliorate many of the inherent limitations of alternative techniques. Yet, it was first necessary to resolve whether *Cryptobranchus* has a ZW or XY system of genetic sex determination, and then to identify specific sex-limited loci (W- or Y-linked) from which we could design a PCR-based assay. Were an eventual assay effective in the related (and also endangered) *Andrias* salamanders, this would add additional impact to such an assay. A PCR-based genetic sex diagnostic would have significant importance for cryptobranchid salamander conservation specifically, and more generally for the study of amphibian sex determination evolution. In salamanders, ZW and XY systems abound across the phylogeny (Sessions 2008). However, nearly all research on sex determination in salamanders has relied on cytogenetic techniques such as karyotyping and C-banding (e.g. Sessions *et al*. 2016; Sessions *et al*. 1982), and even these methods can be unreliable if the sex chromosomes are not strongly differentiated.

METHODS AND MATERIALS

Initial misadventures searching for sex-linked loci

We sought to identify and sequence sex-linked regions of the *Cryptobranchus* genome and to exploit these regions to develop a PCR-based assay for sex. Our first attempts naively focused on designing degenerate PCR primers from known sex-linked gene regions in the deeply divergent salamander *Ambystoma*, the even more divergent frog *Xenopus*, and the even more deeply divergent fish *Danio*. This candidate locus approach quickly proved fruitless, likely due to these taxa having independent sex determination systems and/or very divergent nucleotide sequences across orthologous sex determining loci. We next attempted to use amplified fragment length polymorphism (AFLP) markers (Vos *et al*. 1995) as a method for anonymous interrogation of genomic regions in known-sex individuals. Although thousands of AFLP markers were successfully generated and scored across a panel of 20 known-sex hellbenders, these markers in total only reflected a very small portion of the massive hellbender genome. These AFLPs were methodologically challenging to generate, moderately low-throughput, and ultimately failed to detect putative sex-linked genetic markers. The limited differentiation between sex chromosomes in cryptobranchids meant that it would be necessary to screen a large number of markers in order to identify the relatively small region of difference between the genomes of males and females. The very large genome size further complicated matters because the sex-specific regions were effectively diluted by autosomal and pseudoautosomal regions. Reasoning that protein-coding sex-linked loci may be expressed differentially in ovary and testis tissues, we also performed transcriptome sequencing of ovary, testis, and somatic

tissues (obtained opportunistically during necropsy at the St. Louis Zoo) to identify putative sex-linked genes. Transcriptome sequencing identified a very large set of putative sex-linked contigs which were expressed uniquely in only one gonad type (54,831 ovary-specific contigs and 345,146 testis-specific contigs). These numbers of candidate markers were far too large to effectively screen with PCR. We screened a subset of these candidate loci with sequence similarity to known sex-linked regions in *Homo*, *Gallus*, *Xenopus*, or *Ambystoma*, but ultimately failed to detect any sex-linked loci. Seeking a method that would provide much greater numbers of candidate markers than the AFLP-based approach, but which would be able to be filtered to include a much smaller list of candidates than the transcriptome-based markers (based on data from multiple, known-sex individuals), we adopted a double digestion restriction site-associated DNA sequencing protocol (ddRAD, Peterson *et al*. 2012). At the time of this work, no RAD markers had been developed for salamanders, so we set out to develop these *de novo*.

Collection of individuals and DNA extraction

We obtained tissue or blood samples from known-sex *Cryptobranchus* from our own field collections and from captive individuals at the St. Louis Zoo. We also obtained blood samples from *Andrias davidianus* and *A*. *japonicus* from the St. Louis Zoo, the California Academy of Sciences Steinhart Aquarium, and the Smithsonian National Aquarium. Twenty known-sex *Cryptobranchus* individuals were included from two separate tributaries of the White River in Missouri and Arkansas (nine females and 11 males) to serve as reference individuals for ddRAD. All of these reference individuals were

sexed definitively either by necropsy, observation of gametes, or gonadal histology. Candidate markers identified from this set of 20 reference individuals were further screened in six different individuals from the White River drainage in Missouri, two individuals from the Blue River in Indiana, and two *Andrias davidianus*. An additional 23 individual hellbenders of known-sex (but unknown to the investigators) from the Gasconade, Big Piney, Niangua, Meramec, and Current Rivers across Missouri were used to conduct a series of blind trials of candidate loci passing initial screening steps. Finally, retained candidate loci were screened in a panel of 18 known- or suspected-sex hellbenders from three sites in Kentucky. Table 4.1 provides details about these sampled individuals. High molecular weight genomic DNA (gDNA) was extracted from all individuals using Qiagen DNeasy column kits, quantified on a Qubit fluorescence spectrophotometer, and confirmed to be intact by 2% agarose gel electrophoresis.

DdRAD library construction and high-throughput sequencing

Nuclear genomes in the salamander family Cryptobranchidae are enormous (~55 Gbp, Gregory 2017), and this presents several methodological challenges for reduced representation sequencing which we attempted to mitigate through a careful process of empirical test restriction enzyme digestions and bioinformatic estimation of the numbers of fragments expected per individual when using different library preparation protocols. We performed initial explorations with several restriction enzyme combinations and estimated the numbers of unique loci which might be generated under a ddRAD protocol (detailed in Figure 4.1). We tested four 5' enzymes against three 3' enzymes (as in Peterson

et al. 2012) for a total of 12 possible enzyme combinations. The enzymes which we tested varied in both the lengths and base compositions of the recognition sequences, and we selected these combinations in an attempt to generate a wide range of numbers of loci from which to choose a suitable number for multiplexed sequencing of individuals. For each enzyme combination, we generated single digestion products for both enzymes individually and the double digestion products from both enzymes in combination. These test digests were performed for two individual *Cryptobranchus* from the White River drainage in Missouri.

We quantified the resulting fragment length distributions with an Agilent Bioanalyzer 2100 high-sensitivity DNA system, and calculated the estimated number of sequence-able fragments for different combinations of restriction enzymes and size selection windows. Using these empirical fragment distributions, we estimated the number of unique genomic regions targeted by each enzyme pair at size selection windows of 300 ± 30 bp, 400 ± 40 bp, and 500 ± 50 bp, following the methods of Peterson *et al*. (2012; Supplemental Materials). Based on these empirical fragment distribution tests in both individuals, we selected *EcoRI* (3') and *SphI* (5') with a fragment size selection window of 450 - 550 bp for downstream library preparations. Of all 48 enzyme-by-size selection window combinations considered, these library preparation parameters were estimated to yield approximately 350,000 unique fragments per individual (significantly fewer loci than any of the other potential combinations). We also found that it was necessary to increase the amount of input gDNA in our restriction enzyme digestions from the recommended 50 ng (Peterson *et al*. 2012) to 3 μg per individual in order to retain a sufficiently large quantity of post-bead-cleaned product to perform the adapter ligation steps.

We used a dual index combinatorial multiplexing strategy, identifying individuals by unique combinations of 5 bp inline barcodes and 6 bp Illumina indices. We pooled four sets of five individuals after individual restriction enzyme digestion and adapter ligation, and then we size selected each set of five individuals in its own well of a Pippin Prep (Sage Sciences) cartridge. In practice, size selection of *in situ* fragments in the 442 - 558 bp range was performed using the "tight" collection protocol with an actual size selection window setting of 518-634 bp (to account for the lengths of the Illumina adapters which were ligated to the ends of all fragments). These size-selected products were then pooled into two sets of ten total individuals, bead cleaned, and amplified by PCR for 8 cycles with a high-fidelity polymerase (New England Biolabs Phusion), as in Peterson *et al*. 2012. This low-cycle PCR step was aimed at avoiding low-complexity molecular bottlenecks and reducing PCR duplicates and enzymatic polymerase errors in the resulting amplicons. Final bead cleanup steps were performed with Thermo Fisher Dynabeads and then Agencourt AmPure XP beads, and the completed ddRAD libraries were quantified on a Thermo Fisher Qubit fluorescence spectrophotometer and visualized with the Agilent Bioanalyzer 2100 fragment analyzer. Resulting libraries were sequenced on two Illumina HiSeq2500 lanes in Rapid Run mode with paired-end 150 bp reads (utilizing C-bot cluster generation). A 10% *PhiX* DNA spike-in was used to increase nucleotide diversity and produce more optimal clonal cluster generation (reads from the *PhiX* spike-in are automatically removed by the sequencing center). Illumina sequencing was performed at the Florida State University School of Medicine Core Facility.

Locus assembly and characterization

The particular library preparation protocol that we employed results in strand-specific loci because our PCR primers for fragment amplification effectively selected for only those fragments with *SphI* at the 5' end and *EcoRI* at the 3' end. The total lengths of the fragments which we sequenced to generate our ddRAD loci exceeds the combined length of both of the 150 bp paired-end reads. Each fragment is essentially represented by loci comprising 150 bp of 5' sequence and 150 bp of 3' sequence at the flanks, with a central un-sequenced region of unknown length (fragments originated from a fragment distribution centered around 500 ± 50 bp). To account for this feature of our particular combination of size-selection window and read lengths, and in an effort to retain information from both the read one (R1) and read 2 (R2) read pairs, we used custom bash scripts to concatenate reads from the 5' ends of fragments (R1 of an Illumina read pair) with the reverse complement of reads from the 3' ends of fragments (R2 of an Illumina read pair), recapitulating the original orientation in the genome. Although this "stitching" procedure unites noncontiguous genetic regions by not accounting for the internal un-sequenced regions, this approach retains the provenance between these stitched flanking regions in R1 and R2, in contrast to methods that treat 5' and 3' fragments as separate loci or which simply exclude half of the read data. Loci should all be greater than 400 bp in length, so none should have overlap between 5' and 3' fragments. The process_radtags function in stacks v1.29 (Catchen *et al*. 2013) was used to demultiplex the raw, stitched reads by individual, allowing for one nucleotide mismatch in the observed barcodes from the reference list (all barcodes used were two or more substitutions away from each other in substitution space). Stitched reads were only retained if they contained the appropriate

restriction enzyme cut sites at both ends and also had a mean Phred quality score greater than 20 over all 45 bp sliding window intervals along their total length. These parameters amounted to the following settings for the stacks process_radtags algorithm: --renz_1 sphI --renz_2 ecoRI -c -q -r -D -w 0.15 -s 20 --barcode_dist_1 2.

The stacks assembly pipeline was used to assemble unique loci and to make preliminary haplotype calls for each individual (ustacks); to assemble a locus catalog for all individuals (cstacks) denoting which loci are shared by which individuals; to find catalog matches for each individual (sstacks); and to call haplotypes across all individuals (genotypes). Stitched, demultiplexed, and filtered reads were assembled for each individual in parallel for six combinations of assembly parameter settings. We attempted to consider multiple expectations for the range of nucleotide variation between alleles at a given locus (ustacks -M = 4, 10), for the range of sequencing coverage across individuals (ustacks -m = 3, 10), and for variation between alleles across the set of individuals (cstacks -n = 0, 16). We used sstacks to match individual loci back to the full catalog, and we reconstructed haplotypes across all loci for all 20 individuals with genotypes (-r 1 -m 3). Exploring these six combinations of assembly parameters for ustacks and cstacks, we aimed to choose parameter settings which would optimize the recovery of putatively orthologous, single-copy regions of the genome from our assembled loci.

Identification of candidate sex-linked loci

We sought to exclude from consideration any locus in which we lacked confidence of proper assembly. The large size, complexity, and repetitiveness of the hellbender

genome all contribute to the assembly in stacks of some loci which have greater than two haplotypes in some individual(s), often representing low-confidence SNPs being called in outlier loci of extremely high coverage. Many of these loci with inferred ploidy level greater than two have high sequence similarity to known transposable elements (TEs) in the *Cryptobranchus* genome and these outliers appear to represent cases of multiple copies of slightly divergent TEs chaining together during the assembly process in stacks. In the most extreme cases, some loci were assembled with read coverage >5,000 times above the global mean coverage for all loci. Other, less severe cases of confounded ploidy appeared to result from SNP calling errors due to the low stringency coverage thresholds which were required to enable detection of poorly sequenced genomic regions. Any locus with more than two haplotypes in any of the 20 reference individuals was excluded from further analyses.

Because there was also uncertainty about whether cryptobranchid salamanders have a ZW or XY sex determination system, we conducted analyses agnostically for both scenarios, testing a specific set of hypotheses based on expected patterns of genetic variation in males and females under each alternative model. Converse expectations exist for patterns of presence and absence, patterns of individual zygosity, and relative depths of coverage across loci for male heterogamety (XY) versus female heterogamety (ZW). We aimed to evaluate the evidentiary support for these competing models of sex determination system by quantifying these attributes in anonymous loci across the hellbender genome. Briefly, sex-limited loci (non-pseudoautosomal Y- or W-linked loci) are expected to only ever be present in one sex (males or females, respectively) and to never be present in the opposite sex. By quantifying patterns of presence and absence across shared loci, we first

identified a set of putative male-specific loci and a set of putative female-specific loci based on presence in all 11 or all nine individuals, respectively. Next, we refined this set of candidates by only considering sex-specific loci which were only ever homozygous in every individual. Because the heterogametic sex only has one copy of the sex-limited chromosome, all loci in the non-recombining (heteromorphic) regions are hemizygous (appearing homozygous in the absence of information regarding their sex-linkage). The depth of read coverage of loci should also be informative about sex-linkage because one expects sex-specific loci to have roughly half the depth of coverage of autosomal or pseudoautosomal regions. Based on a combination of these criteria, we identified a set of potentially sex-specific loci for both males and females, and then sought to design PCR primers for these candidates and to attempt to validate these loci in a set of known-sex hellbenders.

PCR primer design and validation of candidate loci

These loci exist in the genome as fragments with a distribution of lengths centered around 500 bp, of which we sequenced 150 bp on the 5' and 3' ends, respectively. To produce reasonable estimates of product sizes and annealing temperatures for these loci when designing PCR primers, we artificially inserted a 200 bp tract of N characters between the R1 and reverse complemented R2 sequences in the fasta file prior to primer design. Oligonucleotide primers were designed for each candidate locus (seven putative Y-linked, 35 putative W-linked, and an 18S rRNA positive control) in BatchPrimer3 (You *et al*. 2008), using default parameters except for: primer length (minimum 23 bp, optimum

30 bp, maximum 33 bp), maximum difference in melting temperature (T_m) between forward and reverse primers (5 C), and optimal amplicon fragment length (minimum 375 bp, optimum 500 bp, max 550 bp). We used an optimal primer length setting of 30 bp (range 23-33 bp) to ensure that the primers would be sufficiently long to have a high probability of non-random binding within the complex hellbender genome (e.g., any 21-mer sequence has a moderate chance of occurring in a 55 Gb genome, but any 30-mer sequence has a substantially smaller chance of occurring at random). Because putative Y- or W-linked loci should only be present in one sex (males and females, respectively), diagnostic PCR reactions for the putative sex-limited loci which produce no bands on an agarose gel could either indicate that the individual in question does not have the sex-limited chromosome (males for ZW or females for XY), or that a PCR failure has led to a lack of amplification.

As a positive control to ensure that PCR reactions were successful and to validate that any non-amplifying putative sex-limited loci are genuine, we also designed primers for a 756 bp fragment of the nuclear-encoded 18S ribosomal RNA subunit. This 18S fragment was designed from a complete rDNA subunit for *Cryptobranchus* which we assembled from raw transcriptome data. The 18S positive control is used to confirm that the absence of a band on a gel was not due to PCR failure and likely resulted because a given sequence was absent in the genome of that individual. The presence of a band for a given individual at a given candidate locus was taken as evidence for successful amplification, while the absence of a band was seen as failure for a given pair of primers to amplify in the genome, so long as the positive controls successfully amplified for that individual.

PCR reactions were assembled using locus-specific master mixes of all components except for gDNA in order to standardize conditions across individuals and loci for a given trial. PCR reactions were carried out in 20 μl reactions (200 μM dNTPs, 0.5 μM forward and reverse primers, 109 ng gDNA, 0.4 U New England Biolabs Phusion DNA polymerase) with a "hotstart" initial denaturation of 98 C for 30 seconds, followed by 40 cycles of 98 C denaturation for 10 seconds, 64 C annealing for 20 seconds, and 72 C extension for 30 seconds with a final 72 C extension for 10 minutes. PCR products were stained with 2X EZ-Vision I dye and visualized on 1.3% agarose gels run for 45 minutes at 110 volts.

As long as they were not excluded at a previous step, all candidate loci were subjected to successive rounds of PCR validation in increasing numbers of known-sex individuals. First, an initial PCR validation step was performed in one male and one female hellbender. Loci passing this two-individual test were screened in a broader 10-individual panel consisting of novel *Cryptobranchus* and *Andrias* individuals. Loci passing this panel were then tested in a blind trial with a panel of 23 known-sex individuals, and were also subjected to *post hoc* validation in additional *Andrias davidianus* and *Andrias japonicus* individuals. We first sought to test the alternative hypotheses that *Cryptobranchus* has a ZW or XY system of genetic sex determination. We attempted to amplify candidate sex-limited loci in two known-sex individuals (one male and one female hellbender from the White River System in Missouri) which were not included as part of the ddRAD sequencing. Loci which co-amplified in both sexes were not considered further, while loci with sex-specific amplification were retained for further screening in a larger panel of known-sex individuals. Any locus passing this broader panel was subjected to a blind trial

in a geographically diverse set of 18 hellbenders to verify efficacy across multiple populations. Because in the absence of environmental influences on sex, ZW and XY systems are mutually exclusive, we expected that either all of the male-specific candidates or all of the female-specific candidates would co-amplify in both sexes, enabling us to reject one of these alternative hypotheses about the mode of sex determination in cryptobranchid salamanders.

RESULTS

DNA extraction, high-throughput sequencing, and demultiplexing

All individuals yielded sufficient quantities of high quality gDNA for ddRAD library construction (72 - 768 ng/μl per individual) and gDNA was confirmed to be intact as evidenced by high molecular weight bands on agarose gels. In total across all 20 reference individuals, we obtained 163,104,028 pairs of 150 bp in length. After initial demultiplexing, quality filtering, and restriction enzyme cut site verification and truncation, we retained 113,835,666 read pairs totaling 33,809,192,802 bp (after trimming in-line adapter sequences). On average, males and females had roughly equal numbers of retained reads per individual, but there was significant variation in the numbers of reads per individual. The 9 females had on average 5,925,697 raw read pairs per individual (range 1,671,803 - 10,897,888 reads) and the 11 males had on average 5,500,399 raw read pairs per individual (range 2,980,522 - 8,934,375 reads). This is depicted on the horizontal axis of Figure 4.2.

Locus assembly and characterization

We explored a range of stacks assembly parameter settings and selected a combination of settings for ustacks (-m 3 -M 4 -N 10) and for cstacks (-n 0) which we reasoned should be effective for detecting putatively sex-linked loci and for distinguishing these candidates from loci with no evidence of sex-linkage. We chose to allow a minimum depth of coverage of three reads to form primary stacks in ustacks. We enforced relatively strict pairwise matching between reads forming a locus for an individual, allowing only up to four pairwise mismatches between stacks of reads, and no more than 10 pairwise differences between secondary reads.

As with the numbers of retained reads per individuals, there was also significant variation in the number of ustacks loci assembled for each individual, with a general positive (but asymptotic) correlation between the number of input reads and the number of ustacks loci (Figure 4.2). Females on average had slightly fewer loci (283,416 loci per individual vs. 305,379 loci per individual in males), and also had a wider range of numbers of loci than males (147,468 - 387,597 loci for females vs. 216,458 - 366,678 for males). A logarithmic distribution best fit these data, with high coefficients of correlation for the groups of males ($y = 1.1539 * \ln(x) + 1.1632$, $R^2 = 0.9169$) and females ($y = 1.2995 * \ln(x) + 0.7338$, $R^2 = 0.9929$). The cstacks, sstacks, and genotypes pipeline was used to generate a matrix of loci which are present or absent across all individuals, as well as characterizing SNP variation among the set of individuals and loci.

Across all ustacks loci for all 20 individuals, we assembled a catalog of 2,441,226 unique loci in cstacks. Not all loci were present for all individuals, and in fact, all individuals had loci which were shared with every possible combination of other individuals. This complex situation probably largely reflects a lack of saturation of locus sampling due to uneven sequencing coverage across individuals (some loci were not recovered in some individuals because of lower sequencing output). However, some of this variation in overlap of loci across individuals may reflect cryptic patterns of sex-linkage. For each individual, we quantified the numbers of loci which were shared in all possible numbers of individuals from one to 20, inclusive (Figure 4.3). Based on the haplotype calls made in sstacks, we also characterized patterns of SNP variation across loci. We discarded any locus for which any individual had more than two alleles. Among the set of retained loci with either one or two alleles, we identified sets of loci which were either always homozygous in all males or all females and either homozygous or heterozygous females or males (putative pseudoautosomal Y- or pseudoautosomal W-linked loci, respectively), and sets of loci which were always homozygous in either all males or all females and absent in females or males (putative sex-specific Y- or W-linked loci).

Identification of candidate sex-linked loci

Figure 4.4 shows that comparisons involving greater numbers of individuals significantly refine the successive sets of putative sex-linked loci. When comparing only a few individuals of each sex, many loci appear to be present uniquely in one sex and absent in the other sex, and this pattern holds for both males and females. But as greater numbers

of each sex are compared, the numbers of putatively sex-specific loci drop precipitously. A power distribution best fit these data, with high coefficients of correlation for the groups of males ($y = 188133 * x^{-3.818}$, $R^2 = 0.8485$) and females ($y = 53119 * x^{-2.953}$, $R^2 = 0.9374$). After comparing all nine female and 11 male hellbenders, we retained a set of 7 loci present in all males and absent in all females (putatively Y-linked) and a set of 100 loci present in all females and absent in all males (putatively W-linked). To reduce the number of loci in the PCR screening steps, we excluded any sex-specific candidates for which any individual of the putatively heterogametic sex had more than one haplotype identified. Because *Cryptobranchus* is diploid, one expects that sex-limited loci should only ever have one haplotype in a given individual (though more haplotypes may be present in the population) because these loci are effectively hemizygous in the heterogametic sex. We further filtered these candidate loci according to their uniqueness in the total set of assembled loci, excluding any loci which had a blastn sequence similarity hit greater than 85% to any other locus present in the alternative sex. This procedure was expected to partially account for the possibility that our assembly parameter settings in stacks may have led to divergent allele copies for an individual being assembled as separate loci. From our initial set of eight male-specific and 100 female-specific candidates identified from the presence/absence analysis, these additional filtering steps yielded 35 putative female-specific loci and (the same set of) eight putatively male-specific loci. We aimed to design oligonucleotide primers for these loci and to use PCR validation to test whether any of these candidate loci could be rejected as sex-specific by PCR (non-sex-specific candidates should co-amplify in both males and females, whereas genuine sex-specific candidates should only amplify

by PCR in one sex). This procedure for bioinformatic sex-linked locus identification and PCR-based validation is outlined in Figure 4.5.

PCR primer design and validation of candidate loci

We attempted to design PCR primers for the eight putative male-specific loci, the 35 putative female-specific loci, and one nuclear 18S ribosomal DNA fragment which would serve as a positive control. Oligonucleotide primer sequences were designed in BatchPrimer3 (You *et al*. 2008) using custom parameter settings. We successfully designed primers for all 45 loci of interest that met these specifications. Primer details are provided in Table 4.2.

After our initial two-individual tests, we were left with a set of zero putatively male-specific loci (all of these loci co-amplified in both sexes) and 12 putatively female-specific loci. After screening in the 10-individual panel, two putatively W-linked loci remained which consistently amplified in all females and never in any males. Figure 4.6 shows the agarose gels from one of the W-linked markers and the 18S positive control in the 10-individual panel. Next, we conducted a blind trial with 23 known-sex hellbenders, screening each individual for the two W-linked markers and the 18S positive control. The sexes of these 23 individuals were known to JTB, but not to PHM prior to the blind trials. PCR results were sent to JTB and compared to known sexes. Both W-linked markers successfully amplified with strong, crisp bands around ~500 bp for all 12 known females, and failed to amplify in all 11 known males. The 18S marker successfully amplified in all 23 individuals.

Finally, we conducted additional PCR validation on a set of 18 known- or suspected-sex hellbenders from a divergent river system in the Licking and Green Rivers in Kentucky, as well as for eight additional *Andrias japonicus* and four additional *Andrias davidianus*. Both of the retained W-linked markers appeared to be effective across all hellbender populations examined and across multiple individuals from both species of *Andrias*. Based on these PCR validation tests, we reject the hypothesis of male-heterogamety in the family Cryptobranchidae. These two putatively W-linked markers appear to robustly amplify in all of the female *Cryptobranchus* and *Andrias* tested, but appear to never amplify in any of the males tested, suggesting that these W-linked markers were likely sex-linked in the common ancestor of *Cryptobranchus* and *Andrias* at least 60 million years ago (Zhang & Wake 2009) when these two genera likely diverged.

DISCUSSION

From a methodological standpoint, this study demonstrates the power of reduced representation genome scans for identifying sex-linked genes in non-model organisms in the absence of pre-existing genetic resources. Although the genomic revolution is now permeating all areas of biology, its application in amphibians (and specifically in salamanders) has lagged behind because of challenges posed by massive genome sizes. These results suggest that genome size may no longer be a limiting factor in generating informative genome-scale data to answer evolutionary questions in salamanders. Our results highlight the importance of sampling multiple known-sex individuals in order to winnow down the number of candidate sex-specific loci which must be screened by PCR

(an expensive step, relative to filtering data). This system also in a large genome with homomorphic sex chromosomes

Conservation implications

From an applied conservation perspective, the identification of sex-linked genes and the development of a simple PCR-based assay for sex in these imperiled salamanders provides unprecedented opportunities to direct conservation efforts and to understand aspects of hellbender demography, natural history, and reproduction that have previously remained inaccessible. It may now be possible to, for instance, accurately determine sex ratios in wild populations, conduct captive breeding and repatriation projects with full knowledge of sex for all individuals and all age classes, assess sex ratios within individual clutches of eggs, and begin to assess the potential effects of environmental chemical pollutants on reproductive health of wild populations. Although these W-linked markers appear to be robust across all *Cryptobranchus* populations tested (spanning most of the major lineages across the geographic distribution) and in both species of *Andrias*, it is possible that nucleotide substitutions in the regions where we designed our primers could potentially lead to locus dropout. In this case, one would fail to amplify the W-linked fragments in females, leading to incorrect inferences that those individuals were male. Locus dropout during diagnostic PCR would be difficult to detect because the 18S positive control would still be expected to amplify (and 18S is under much tighter functional constraint than many genomic regions, making it less likely that mutations in those priming sites would occur). However, it is very unlikely that parallel mutations would alter primer

binding regions in two independent loci simultaneously. Accordingly, performing sex diagnosis with both of the W-linked loci identified here alleviates concerns about locus dropout due to PCR primer binding site mutations. The converse situation, in which the W-linked loci would spuriously amplify in males seems very unlikely. We conclude that performing genetic sex diagnosis in cryptobranchid salamanders using the two W-linked loci and the 18S positive control reported here is a robust assay, with type I and type II error rates close to zero (though type I error rates are expected to be larger than type II error rates).

Implications for understanding sex determination in salamanders

More generally, this work has provided important baseline information about the sex determination system in an early-diverging salamander lineage. Together, the failure of all putatively male-specific markers in the PCR panels to amplify and the sex-specific amplification of several putatively female-specific markers provide evidence in support of the hypothesis that cryptobranchid salamanders possess a ZW sex determination. That two of these candidate markers are effective in both species of *Andrias* further suggests that a conserved ZW sex determination system was present in the most recent common ancestor of *Cryptobranchus* and *Andrias*. The observation that the same two pairs of W-linked primers are effective in both genera (which diverged approximately 60 million years ago) suggests that rates of substitution in these loci are very low, possibly due to stabilizing selection on these loci or loci in linkage with them.

Practical considerations for investigating sex determination systems

The power of reduced representation genome scans to detect sex-specific genomic regions depends on interactions between several factors. More markers and/or more densely spaced markers may be required for taxa with homomorphic sex chromosomes, relative to taxa with heteromorphic sex chromosomes, because the relative proportion of the genome which is unique to either males or females is smaller in the former case. The absolute size of the genome will also influence the numbers of loci needed to detect sex-linkage, with larger genomes requiring greater numbers of markers. Also, the depth of sequencing coverage across loci and the uniformity of coverage across individuals impact researchers' ability to detect sex-specific loci, with greater sequencing coverage being expected to reduce false positive hits. Other genomic attributes (e.g., genome size, genome complexity, base composition, etc.) and methodological considerations (e.g., which restriction enzymes and size selection windows to test) will dictate the range of possible numbers of loci among which a researcher may choose. There is also an inherent trade-off between the number of loci generated by a particular ddRAD protocol and the number of different individuals which can be multiplexed together for sequencing (sampling greater numbers of loci would result in fewer individuals per sequencing lane to retain the same levels of per-locus coverage). Researchers should carefully consider how many loci and how many individuals they expect to be sufficient to detect sex-specific loci in their particular organismal system, based on any available information about the degree of heteromorphy between sex chromosomes and absolute genome size. In the absence of definitive knowledge about whether a particular taxon exhibits male- or female-

heterogamety, both scenarios can be evaluated in parallel in an attempt to reject one of these alternative hypotheses.

Comparing multiple, known-sex individuals is also an important aspect of sex-specific locus detection, and contrasts drawn from greater numbers of individuals reduce the numbers of putative sex-linked loci which must be screened by PCR in downstream steps. In our case with hellbenders, the number of candidate loci identified by analyzing a single representative of each sex decreased by roughly an order of magnitude each time that we doubled the number of each sex (comparing two, four, or eight individuals of each sex). In more challenging cases where sequencing coverage is particularly uneven across individuals, relaxing the requirement that loci are present for all individuals of a given sex (but still requiring absence in all individuals of the opposite sex) could potentially lead to greater detection of putatively sex-linked loci.

The ability to assemble loci at appropriately deep coverage to confidently call SNPs is a pressing concern for phylogeographic or demographic studies employing reduced representation sequencing approaches. However, in our study the primary aim was characterize patterns of presence and absence across loci relative to sex and to identify sex-specific genetic regions (Y- or W-linked). Although our assembly parameters may lead to some formation of low-coverage loci, this lenient assembly strategy accommodates variation in sequencing output across samples, in that even if a locus is poorly sampled from a given individual in terms of read depth, that locus may still be detected if present in the genome of a particular individual. Secondary inferences about putative patterns of zygosity (which can help refine lists of candidate loci which have been identified from presence/absence) may be more susceptible to assembly errors arising from our relaxed

ustacks settings. Though in these cases, it is more likely that SNP variation at a locus would be underestimated, not overestimated, leading to a reduced ability to reject candidate Y- or W-linked loci on the basis zygosity.

As a cautionary note, we stress that individuals' sexes must be known with absolute certainty, especially for species with large genomes which will have large numbers of loci under consideration. When sequencing coverage across individuals is uneven, comparisons of loci between individuals with mis-assigned sexes can result in spurious lists of putatively sex-linked loci (false positives). We discovered this type of error early in our study when we scrutinized the patterns of presence/absence across what we initially believed to be 10 males and 10 females. One "female" had patterns of missing loci which made us suspect that it may in fact be a male. We gained access to necropsy records from this individual and confirmed that it was indeed a male which had been mis-recorded as female at the time of death due to a database transposition. Until that point, our bioinformatic pipelines had identified a set of seemingly plausible candidate loci which were present in all 10 males and all 10 "females", highlighting the potential for erroneous sex calls to produce spurious lists of candidate sex-linked loci. In our specific case, all 20 reference individuals had some loci which were shared with nearly every possible combination of the other 19 individuals, meaning that even if one randomly assigned sexes to samples, it could be possible to recover lists of candidate loci which appear to be present only in one sex and absent in the other, underscoring that definitive identification of sex for all reference individuals is crucial for generating credible lists of candidate sex-linked loci.

We expected the large (55 Gb) hellbender genome to pose significant challenges, not only for generating genomic data, but also for our ability to screen and validate putative

sex-specific loci. In comparing multiple combinations of restriction enzymes and size selection windows, we learned that nearly all combinations would have produced far too many loci (as many as four million loci per individual) to achieve adequate multiplexing of individuals. The *SphI*/*EcoRI* enzyme pair, with size selection between 450 and 550 bp, was expected to produce approximately 350,000 loci per individual, very much in line with the empirical numbers of ustacks loci which we assembled in the highest-coverage individuals. Based on an estimated genome size of 55 Gb for the hellbender, this suggests that approximately 0.32% of the hellbender genome is sampled by our ddRAD loci. If these loci were spread evenly across all 30 haploid chromosomes (which they almost certainly are not), this would roughly equate to one ddRAD locus being present roughly every 157 kb in the genome. The fact that only two loci, out of over two million total catalog loci which we screened, were successfully validated as W-linked suggests that the region of heteromorphy between the *Cryptobranchus* W and Z sex chromosomes is small, relative to the entire hellbender genome. Future efforts to isolate and sequence larger genomic fragments flanking these sex-linked regions (for instance, using long-range inverse PCR and shotgun sequencing) may help to better characterize these chromosomal regions. Additionally, the upcoming development of a genetic linkage map for *Cryptobranchus* (from analysis of ddRAD data in known-sex F1 and parental individuals with pedigree information) is expected to be extremely informative about the degree of heteromorphy between the Z and W chromosomes.

CONCLUSIONS

Using ddRAD genome scans in known-sex hellbenders, we developed the first genetic sex diagnostic for a non-model salamander. We also demonstrated that the W-linked chromosomal regions we identified are conserved and sex-linked across divergent populations of *Cryptobranchus* from the White, Missouri, and Ohio River drainages, as well as in both species of *Andrias*. Our results allow us to reject a hypothesis of male-heterogamety and are consistent with an ancient, conserved system of female-heterogametic sex determination in the salamander family Cryptobranchidae. This work has also allowed the development of a universally effective PCR-based assay for sex in several species of conservation concern. The W-linked loci described here may enable new and important research and conservation directions for hellbender and giant salamanders. These methods for interrogating genetic sex determination systems in non-model taxa are also broadly applicable in other species and may hold great promise for testing hypotheses about sex chromosome evolution in poorly characterized organisms.

Table 4.1. *Cryptobranchus* and *Andrias* individuals examined.

ID	Taxon	Origin	Sex	Taxon Set
C038AF	*Cryptobranchus alleganiensis*	White River, MO	Female	Reference
C039AF	*Cryptobranchus alleganiensis*	White River, MO	Female	Reference
C091CF	*Cryptobranchus alleganiensis*	Eleven Point River, MO	Female	Reference
C092DF	*Cryptobranchus alleganiensis*	Eleven Point River, AR	Female	Reference
C093DF	*Cryptobranchus alleganiensis*	Eleven Point River, AR	Female	Reference
C100EF	*Cryptobranchus alleganiensis*	White River, MO	Female	Reference
C104EF	*Cryptobranchus alleganiensis*	White River, MO	Female	Reference
C109EF	*Cryptobranchus alleganiensis*	White River, MO	Female	Reference
C110EF	*Cryptobranchus alleganiensis*	White River, MO	Female	Reference
C031AM	*Cryptobranchus alleganiensis*	White River, MO	Male	Reference
C032AM	*Cryptobranchus alleganiensis*	White River, MO	Male	Reference
C033AM	*Cryptobranchus alleganiensis*	White River, MO	Male	Reference
C034AM	*Cryptobranchus alleganiensis*	White River, MO	Male	Reference
C035AM	*Cryptobranchus alleganiensis*	White River, MO	Male	Reference
C036AM	*Cryptobranchus alleganiensis*	White River, MO	Male	Reference
C089CM	*Cryptobranchus alleganiensis*	Eleven Point River, MO	Male	Reference
C094DM	*Cryptobranchus alleganiensis*	Eleven Point River, AR	Male	Reference
C097DM	*Cryptobranchus alleganiensis*	Eleven Point River, AR	Male	Reference
C101EM	*Cryptobranchus alleganiensis*	White River, MO	Male	Reference
C108EM	*Cryptobranchus alleganiensis*	White River, MO	Male	Reference
C120GU	*Cryptobranchus alleganiensis*	Gasconade River, MO	Male	Blind Trial
C121GU	*Cryptobranchus alleganiensis*	Gasconade River, MO	Male	Blind Trial
C122GU	*Cryptobranchus alleganiensis*	Gasconade River, MO	Male	Blind Trial
C123HU	*Cryptobranchus alleganiensis*	Meramec River, MO	Male	Blind Trial
C124HU	*Cryptobranchus alleganiensis*	Meramec River, MO	Male	Blind Trial
C125HU	*Cryptobranchus alleganiensis*	Meramec River, MO	Male	Blind Trial
C126HU	*Cryptobranchus alleganiensis*	Meramec River, MO	Female	Blind Trial
C127HU	*Cryptobranchus alleganiensis*	Meramec River, MO	Female	Blind Trial
C128HU	*Cryptobranchus alleganiensis*	Meramec River, MO	Male	Blind Trial
C131IU	*Cryptobranchus alleganiensis*	Big Piney River, MO	Female	Blind Trial
C132IU	*Cryptobranchus alleganiensis*	Big Piney River, MO	Female	Blind Trial
C133IU	*Cryptobranchus alleganiensis*	Big Piney River, MO	Female	Blind Trial
C136JU	*Cryptobranchus alleganiensis*	Niangua River, MO	Male	Blind Trial
C137JU	*Cryptobranchus alleganiensis*	Niangua River, MO	Male	Blind Trial
C138JU	*Cryptobranchus alleganiensis*	Niangua River, MO	Male	Blind Trial
C139JU	*Cryptobranchus alleganiensis*	Niangua River, MO	Female	Blind Trial
C140JU	*Cryptobranchus alleganiensis*	Niangua River, MO	Female	Blind Trial
C143KU	*Cryptobranchus alleganiensis*	Current River, MO	Female	Blind Trial
C144KU	*Cryptobranchus alleganiensis*	Current River, MO	Female	Blind Trial
C145KU	*Cryptobranchus alleganiensis*	Current River, MO	Male	Blind Trial
C146KU	*Cryptobranchus alleganiensis*	Current River, MO	Female	Blind Trial
C147KU	*Cryptobranchus alleganiensis*	Current River, MO	Female	Blind Trial

Table 4.1 (continued). *Cryptobranchus* and *Andrias* individuals examined.

C148KU	*Cryptobranchus alleganiensis*	Current River, MO	Female	Blind Trial
C36AM	*Cryptobranchus alleganiensis*	White River, MO	Male	2-Indiv Test
C37AF	*Cryptobranchus alleganiensis*	White River, MO	Female	2-Indiv Test
C36AM	*Cryptobranchus alleganiensis*	White River, MO	Male	10-Indiv Test
C37AF	*Cryptobranchus alleganiensis*	White River, MO	Female	10-Indiv Test
C106EM	*Cryptobranchus alleganiensis*	White River, MO	Male	10-Indiv Test
C39AF	*Cryptobranchus alleganiensis*	White River, MO	Female	10-Indiv Test
C90CM	*Cryptobranchus alleganiensis*	Eleven Point River, MO	Male	10-Indiv Test
C92DF	*Cryptobranchus alleganiensis*	Eleven Point River, MO	Female	10-Indiv Test
C52BM	*Cryptobranchus alleganiensis*	Blue River, IN	Male	10-Indiv Test
C57BF	*Cryptobranchus alleganiensis*	Blue River, IN	Female	10-Indiv Test
AD03	*Andrias davidianus*	St. Louis Zoo	Female	10-Indiv Test
AD01	*Andrias davidianus*	California Academy of Sciences	Male	10-Indiv Test
KINN1	*Cryptobranchus alleganiensis*	Kinniconick Creek, KY	Male	KY Test
NT1	*Cryptobranchus alleganiensis*	Licking River, KY	Male	KY Test
NT2	*Cryptobranchus alleganiensis*	Licking River, KY	Male	KY Test
NT3	*Cryptobranchus alleganiensis*	Licking River, KY	Male	KY Test
7718	*Cryptobranchus alleganiensis*	Licking River, KY	Male	KY Test
7689	*Cryptobranchus alleganiensis*	Licking River, KY	Male	KY Test
7771	*Cryptobranchus alleganiensis*	Licking River, KY	Male	KY Test
7687	*Cryptobranchus alleganiensis*	Licking River, KY	Male	KY Test
7761	*Cryptobranchus alleganiensis*	Licking River, KY	Female	KY Test
7686	*Cryptobranchus alleganiensis*	Licking River, KY	Female	KY Test
7724	*Cryptobranchus alleganiensis*	Licking River, KY	Female	KY Test
7742	*Cryptobranchus alleganiensis*	Licking River, KY	Female	KY Test
7732	*Cryptobranchus alleganiensis*	Licking River, KY	Female	KY Test
7733	*Cryptobranchus alleganiensis*	Licking River, KY	Female	KY Test
7678	*Cryptobranchus alleganiensis*	Licking River, KY	Female	KY Test
7679	*Cryptobranchus alleganiensis*	Licking River, KY	Female	KY Test
7698	*Cryptobranchus alleganiensis*	Licking River, KY	Female	KY Test
GREEN1	*Cryptobranchus alleganiensis*	Green River, KY	Male	KY Test
AD01	*Andrias davidianus*	California Academy of Sciences	Male	*Andrias* Test
AD02	*Andrias davidianus*	California Academy of Sciences	Male	*Andrias* Test
AD03	*Andrias davidianus*	Saint Louis Zoo	Female	*Andrias* Test
AD04	*Andrias davidianus*	California Academy of Sciences	Male	*Andrias* Test
AJ01	*Andrias japonicus*	National Zoo	Male	*Andrias* Test
AJ02	*Andrias japonicus*	National Zoo	Male	*Andrias* Test
AJ03	*Andrias japonicus*	National Zoo	Male	*Andrias* Test
AJ04	*Andrias japonicus*	National Zoo	Male	*Andrias* Test
AJ05	*Andrias japonicus*	National Zoo	Female	*Andrias* Test
AJ06	*Andrias japonicus*	National Zoo	Female	*Andrias* Test
AJ07	*Andrias japonicus*	National Zoo	Female	*Andrias* Test
AJ08	*Andrias japonicus*	National Zoo	Female	*Andrias* Test

Table 4.2. PCR primer information and primer validation results.

Candidate Type	Locus ID	Primer Orientation	Start Position in Locus	Primer Length (bp)	T_m (C)	Primer Sequence (5'-3')	Amplicon Length (bp)	Annealing Temperature (C)	Two-Individual Test	Ten-Individual Test
Male-Specific	1449	Forward	28	27	60.80	ATAATGGGAAAATTCCAACTACAAATC	473	64	Fail	Fail
Male-Specific	1449	Reverse	500	29	60.77	AATTCTGATAGGAAAATGTTAATCCAAAT				
Male-Specific	7849	Forward	71	24	67.62	CCCACCTCAGTGCTTCTTCTGTCC	404	63	Fail	Fail
Male-Specific	7849	Reverse	474	25	60.20	CTGAGAGGGAAAACTTACAGTTCAA				
Male-Specific	70276	Forward	6	31	60.89	GCTTTAGGGTATTTGTAATAGAAAATGTCTT	465	63	Fail	Fail
Male-Specific	70276	Reverse	470	28	59.76	GAGTTAATGTGTGTGTATTTGTCTGTGT				
Male-Specific	105687	Forward	23	28	66.17	CTTTATATTTGGCGTACGGCTATCATCC	390	69	Fail	Fail
Male-Specific	105687	Reverse	412	27	66.47	AGGCTTTAAGAGGGACACATGGAAAAC				
Male-Specific	142307	Forward	76	26	68.62	CTCAGTGTTTCTTCTGTCCGGGCTTT	396	63	Fail	Fail
Male-Specific	142307	Reverse	471	25	60.30	AGAGGGAAAACTTACAGTTCAAACC				
Male-Specific	143182	Forward	51	32	62.15	TTAAAAATCTAATCCCCATACTGCTAAAATAC	412	64	Fail	Fail
Male-Specific	143182	Reverse	462	30	60.68	GCTATCATTAAGGTAAGTGTTATTGAACCT				
Male-Specific	166076	Forward	54	27	65.45	ATCCCCAGAAGTGCCTATAAAACACAG	352	65	Fail	Fail
Male-Specific	166076	Reverse	405	32	62.34	TACAGTTGTAAATGTTGGTATTACTCAAGACA				
Male-Specific	211821	Forward	67	28	64.53	CTTGCTTCAGCTGTAGGTATGTGCTAAC	422	65	Fail	Fail
Male-Specific	211821	Reverse	488	31	61.75	GTTAACTCTGTTTCTTCCTCTGCTAGTAAAC				
Female-Specific	1024220	Forward	84	32	59.39	TTTTAAAACTAGCATATAGTCATAGCTTTCTT	383	62	Pass	Pass
Female-Specific	1024220	Reverse	466	31	65.32	AGAAGATCCGAAACAAGGAAACTTAAAATCT				
Female-Specific	1026674	Forward	20	30	61.47	GTCAGATTTACATGACATAAGAAGGAGAAT	430	64	Pass	Fail
Female-Specific	1026674	Reverse	449	28	61.11	CACTGTTACAGATGAATGTGTGTACTTG				
Female-Specific	1028723	Forward	110	27	62.86	CTATCCCATACCACCGTATGTAGTGAC	374	63	Fail	Fail
Female-Specific	1028723	Reverse	483	32	60.06	GAATATTGGTGTACAAACTATACCATACTAGG				
Female-Specific	1029082	Forward	82	26	60.53	GATGAAAACCGAATGAATAGAAAAAG	378	64	Pass	Fail
Female-Specific	1029082	Reverse	459	30	62.18	GTCTGCGTTATTGTAGACTGCTTTACTAAG				
Female-Specific	1029288	Forward	79	31	68.97	AAACAGACATTGTGTCAGCTTTCCAATTGAT	380	64	Fail	Fail
Female-Specific	1029288	Reverse	458	29	61.20	ACTGAGTCTAACTCAATCCCTAAAAATGT				
Female-Specific	1031163	Forward	50	26	59.71	CAGGTAAGAAAAGCTAAAAACAACCT	326	63	Pass	Fail

Table 4.2 (continued). PCR primer information and primer validation results.

Female-Specific	1031163	Reverse	375	26	68.82	TCCCTGGGACCCTTTTAAACCTTCAG				
Female-Specific	1032633	Forward	98	30	65.41	GTACTGGGACAGAATGAGAACCAGAGTTAC	341	64	Fail	Fail
Female-Specific	1032633	Reverse	438	32	60.52	GAGACTAATAGCAATACTATGAAGTAGGGTCT				
Female-Specific	1036594	Forward	80	31	62.10	CTTTAATAACAGATGTTACGATTACCCAACT	393	63	Pass	Fail
Female-Specific	1036594	Reverse	472	32	60.13	TAGAATCAGTAGAAAAATTCAAGAGAGAACTA				
Female-Specific	1041601	Forward	23	30	61.37	ATAAACATCACTTTTTGGTTTTACTGAGTC	427	64	Fail	Fail
Female-Specific	1041601	Reverse	449	31	63.83	ATTAATTAGTGAAAATCCTCAGCGATTAGTG				
Female-Specific	1050159	Forward	116	27	59.78	GCTCAAATTAAGAAGTTCCTTTGTAGA	378	63	Fail	Fail
Female-Specific	1050159	Reverse	493	26	67.04	GAGTTGAATCATTGGCCTGGATCTTC				
Female-Specific	1053606	Forward	67	28	60.62	GAGATCAAATAACAGGGCATATTTTAAC	408	63	Pass	Fail
Female-Specific	1053606	Reverse	474	32	59.72	TTTAATGATTTAGAGTTGTTTACAATACAGTG				
Female-Specific	1054505	Forward	37	27	67.99	AAGTATTTGCTGCGGAAGGCTTTCTCT	375	63	Fail	Fail
Female-Specific	1054505	Reverse	411	27	59.77	CAAAAGAAAATGGGACTAAAAACATAG				
Female-Specific	1054621	Forward	44	31	64.97	CTAGGGGTTTTCTTTATCCCTATCTGGTTAC	410	64	Fail	Fail
Female-Specific	1054621	Reverse	453	26	61.08	ATCTCCAATCTGTGAGATACCTGAAC				
Female-Specific	1055964	Forward	21	29	61.00	CAGAGCTCAGATAGTTCAGTAACAAAGTT	437	63	Fail	Fail
Female-Specific	1055964	Reverse	457	26	59.76	CTGCTCAGGTCATTAGTTATCTTGAC				
Female-Specific	1059944	Forward	15	32	60.25	ACTTAACTGATAGATATAGAAAAAGCTCCAGT	433	62	Pass	Fail
Female-Specific	1059944	Reverse	447	27	59.42	AGGTCTAGAAAAATGATACAGGATGAC				
Female-Specific	1062379	Forward	42	26	66.36	CTGTGGTAATTCTGCTGGGAATGTGT	352	69	Fail	Fail
Female-Specific	1062379	Reverse	393	26	66.83	GGCAAAGCTATATTTTGCTGCCTCAC				
Female-Specific	1070303	Forward	72	24	66.46	GTCACGCCACACACCTTTCTCTTC	429	65	Fail	Fail
Female-Specific	1070303	Reverse	500	31	61.55	AATTCACAATTTAAGTGACATGCTATAAAAA				
Female-Specific	1076166	Forward	63	31	66.24	ATAAATACACACACGCTTAGCATTGCAGTTA	438	65	Fail	Fail
Female-Specific	1076166	Reverse	500	27	62.12	AATTCTGCCTTGGTTAGTAGTTCCTCT				
Female-Specific	1077146	Forward	18	24	67.30	GGAGAACTCTAACGCCCACACAGG	474	67	Fail	Fail
Female-Specific	1077146	Reverse	491	25	63.91	AGTGTTTCACACCTCCCTTTTGAAG				
Female-Specific	1080512	Forward	98	28	63.35	AATCCTAAGGAGGATTCAACTAAGCAAG	372	65	Fail	Fail
Female-Specific	1080512	Reverse	469	32	61.57	TATATGCTGTTATTATGTTTTGGAACTCAGTA				

Table 4.2 (continued). PCR primer information and primer validation results.

Female-Specific	1080569	Forward	9	26	64.05	GACACCTGGAG CTTTCCTTATAT GCT	413	64	Fail	Fail
Female-Specific	1080569	Reverse	421	32	61.31	CTTGTTAATGA CTTACAATGTA CTTTTGTGTT				
Female-Specific	1086769	Forward	1	26	65.06	CATGCTAGGAG TTACGGGATTT CAAG	448	64	Fail	Fail
Female-Specific	1086769	Reverse	448	26	61.43	AGAGCTACGAG TGGTATATGCT CAAG				
Female-Specific	1092737	Forward	112	32	65.01	CTAGCTTCAAA AGTGAGTCATA GCCATAAGAT	378	67	Fail	Fail
Female-Specific	1092737	Reverse	489	29	64.44	ATTCTTGGCCTT TCTATGTAACT GGTTCT				
Female-Specific	1098439	Forward	70	26	60.23	ACTTTATGGTTG CTTTCTCTGTCT CT	396	63	Pass	Fail
Female-Specific	1098439	Reverse	465	32	61.50	AAGAACAATGT CAGGAGATAAA CAGTAGTAGT				
Female-Specific	1098757	Forward	121	26	63.53	GCAGTACTTGG GAGACCTGTCT ATTG	380	67	Fail	Fail
Female-Specific	1098757	Reverse	500	26	66.98	AATTCGTGGTG CTGTCCTCTACC CTA				
Female-Specific	1102805	Forward	82	31	62.84	AATGCACACAT CTTTTTCACATA CATTATTA	419	65	Pass	Pass
Female-Specific	1102805	Reverse	500	31	62.09	AATTCAGTAAA TTTTAAACAAA CAGGATCAC				
Female-Specific	1103907	Forward	118	26	59.62	GATAACGAGAA AGCCTTTGATTC TAT	382	63	Fail	Fail
Female-Specific	1103907	Reverse	499	32	61.83	ATTCAGTGATT GTATTAAGTAT ATCTGGGAGA				
Female-Specific	1106277	Forward	111	28	59.95	GTTTCTCTTTAC TTTTGTACTGGG ACTT	367	63	Fail	Fail
Female-Specific	1106277	Reverse	477	26	67.16	GAGAGAATCAT GGAGGTGGATT GGTC				
Female-Specific	1106395	Forward	28	26	60.49	GAACTAACTCA AGGAATGACCC ATAC	429	63	Fail	Fail
Female-Specific	1106395	Reverse	456	27	62.64	AAGTGTAAGTC GTGCTGCAAAG TTAAT				
Female-Specific	1110384	Forward	93	32	60.55	GAAACTACATA TATTCAGTGAG CTTCAGTAAC	395	64	Pass	Fail
Female-Specific	1110384	Reverse	487	31	61.35	CACATACATAC ACACTCATCCTT TTATAGTG				
Female-Specific	1120030	Forward	41	26	66.32	ATCTGCTCCAT GTACAGTGCTC GAAT	424	66	Pass	Fail
Female-Specific	1120030	Reverse	464	29	63.12	TTTTCTCAGAC ATGGTGATTCT CTTAAC				
Female-Specific	1120855	Forward	60	32	68.00	CTAGACTAGCC TCTTCCCTGTCC TCTTTCTCT	422	64	Fail	Fail
Female-Specific	1120855	Reverse	481	28	61.41	ACTCACCTGAT TTAAGTAGCTA CACACC				
Female-Specific	1123226	Forward	54	28	60.91	CAGAGATGTCA GAAAAGAAAAG AGAAAT	361	64	Fail	Fail
Female-Specific	1123226	Reverse	414	32	62.00	CTCATAACAGA AATTGTATAAA TGGAGAAGAG				
Female-Specific	1123747	Forward	100	26	64.42	GGACGTTAGAA AGATGGACAAG GAAG	284	67	Fail	Fail
Female-Specific	1123747	Reverse	383	32	64.32	AGTTAAGGATC TCCTTCCAGCTA AGAGTCTAT				
Female-Specific	1130179	Forward	19	24	68.78	CCAGCAGTATT CCCCAGCGTCT CT	425	65	Pass	Fail

Table 4.2 (continued). PCR primer information and primer validation results.

Female-Specific	1130179	Reverse	443	27	61.73	CTGACTGGTTTTGGAAGAATTTAGAAC				
Positive Control	18S_comp35_c1_seq1	Forward	514	30	60.57	GTAATTGGAATGAGTACACTTTAAATCCTT	756	64	Fail	Fail
Positive Control	18S_comp35_c1_seq1	Reverse	1269	30	64.37	GAGAAAGAGCTATCAATCTGTCAATCCTTT				

Figure 4.1. Overview of ddRAD protocol. Genomic DNA for each individual is double restriction digested and size selected (A), individual-specific adapters are ligated (B), samples are pooled and sequenced in multiplex (C), resulting reads are demultiplexed by individual (D), *de novo* assembly of loci is performed, and presence/absence and coverage are calculated across loci (E), loci from individuals are compared (F), and haplotypes are called for variable loci (G)

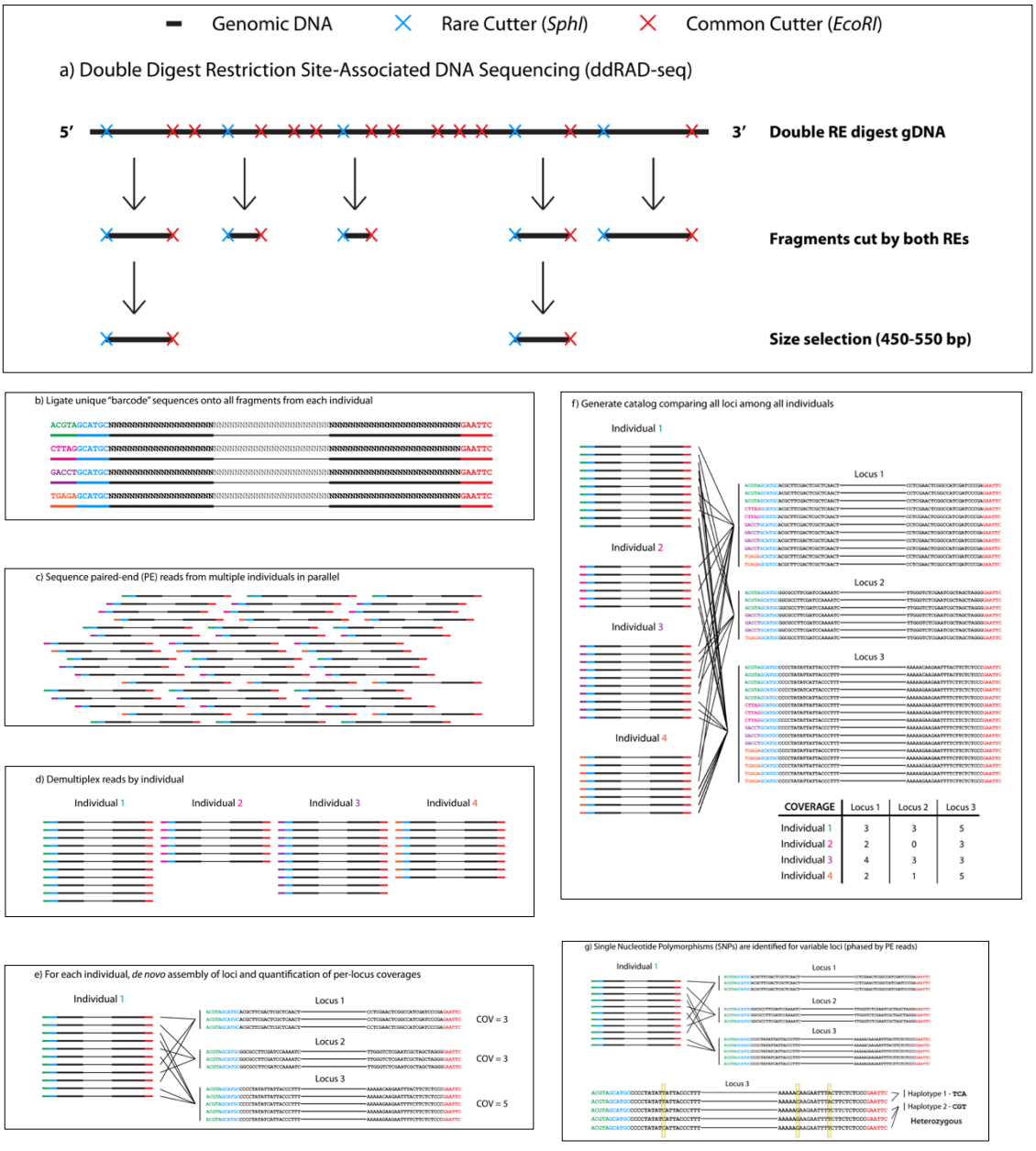


Figure 4.2. Relationship between number of input reads and number of output ustacks loci across nine female and 11 male hellbenders (blue and orange points, respectively). Trend lines represent logarithmic regressions.

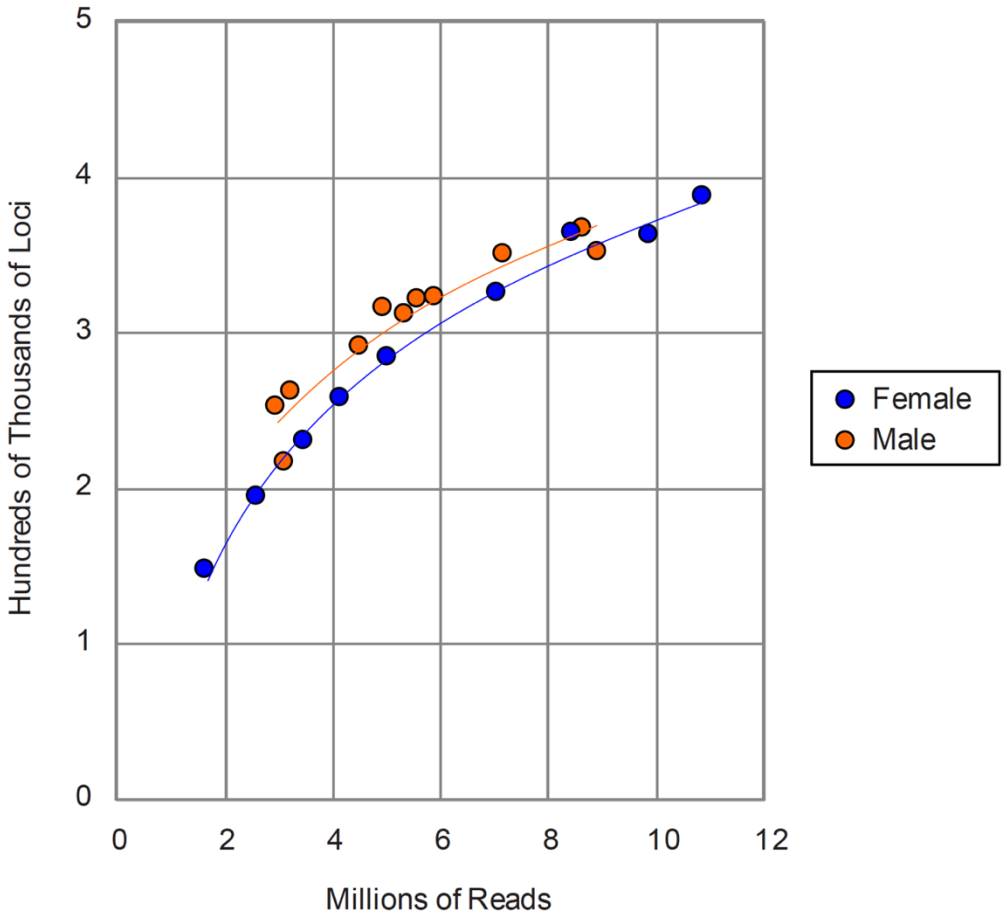


Figure 4.3. Shared ddRAD ustacks loci across 20 *Cryptobranchus* individuals. Females and males are indicated by blue or orange brackets, respectively. For each individual, sets of loci are color-coded according to the number of individuals sharing particular loci. For instance, magenta loci at the bottom of each individual plot are present in all 20 individuals, while the red loci near the top of each individual plot are present in only that particular individual. Qualitatively, each individual has roughly the same profile of locus overlap with other individuals, although not surprisingly, individuals with greater sequencing coverage have not only greater numbers of loci, but also greater numbers of unique loci shared with no other individuals.

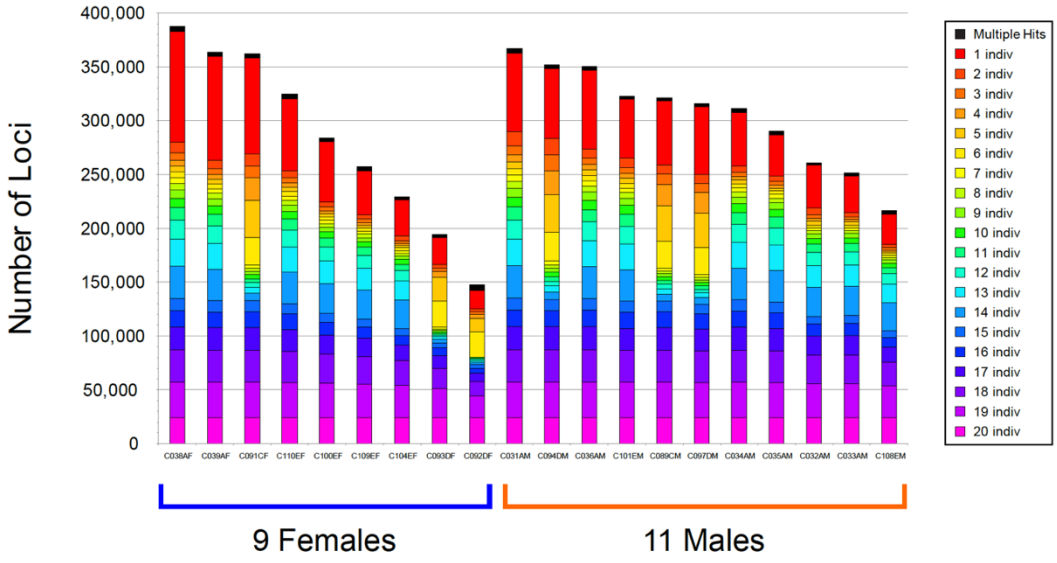


Figure 4.4. Comparisons involving greater numbers of individuals of each sex help to refine the sets of candidate sex-linked loci. Each point represents the number of putative sex-specific loci based on presence and absence patters across random draws of individuals for different numbers of each sex. Females and males are color-coded in blue and orange, respectively. When comparing only a few individuals of each sex, many loci appear to be present uniquely in one sex and absent in the other sex, and this pattern holds for both males and females. But as greater numbers of each sex are compared, the numbers of putatively sex-specific loci drop precipitously. After comparing all nine female and 11 male hellbenders, we retained a set of 8 loci present in all males and absent in all females (putatively Y-linked) and a set of 100 loci present in all females and absent in all males (putatively W-linked). Trend lines represent power regressions.

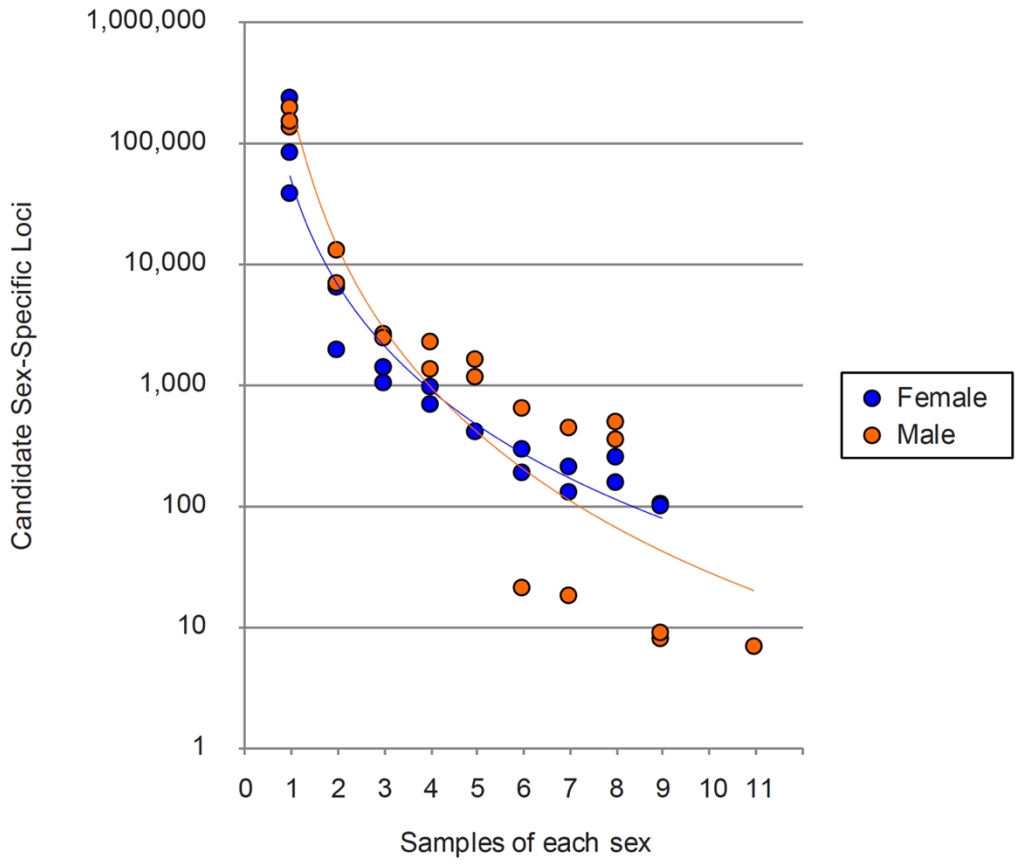


Figure 4.5. Pipeline for identifying putative sex-linked loci and for testing the competing hypotheses of female-heterogametic (ZW) versus male-heterogametic (XY) sex determination in Cryptobranchidae. A. Patterns of presence/absence, inferred zygosity, and depth of coverage were calculated across 2,441,226 cstacks loci and used to generate and refine lists of candidate sex-specific loci based on expectations under either a ZW or XY system. B. PCR was used to validate candidate loci by testing for amplification in known-sex individuals from increasingly divergent populations and species.

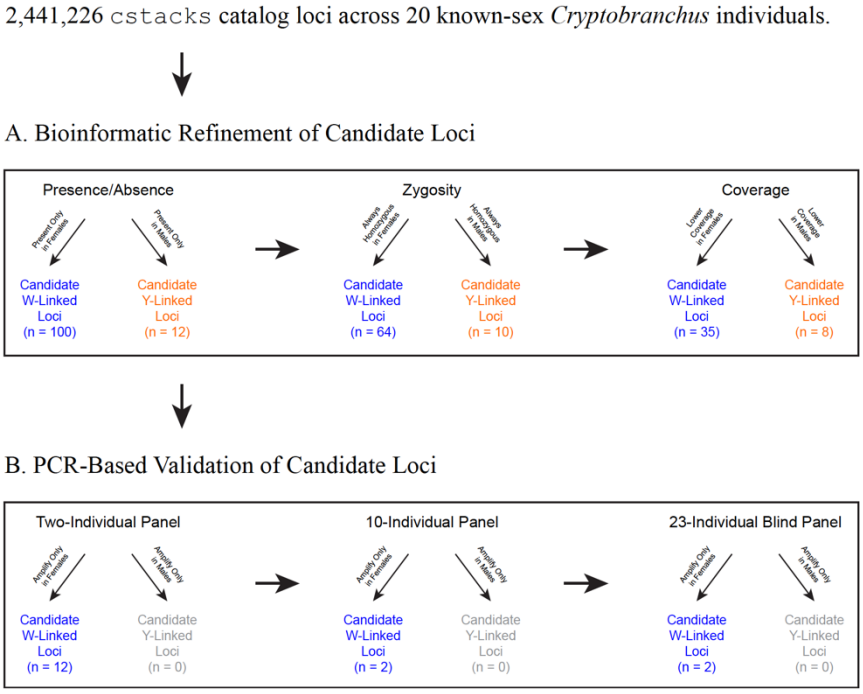
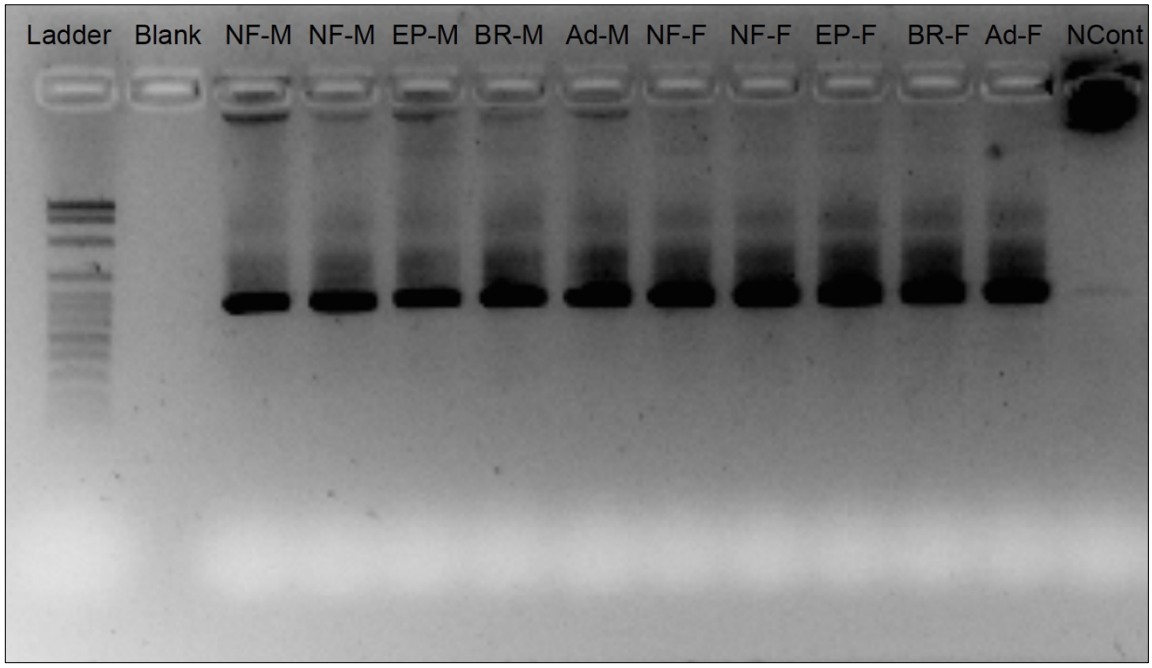
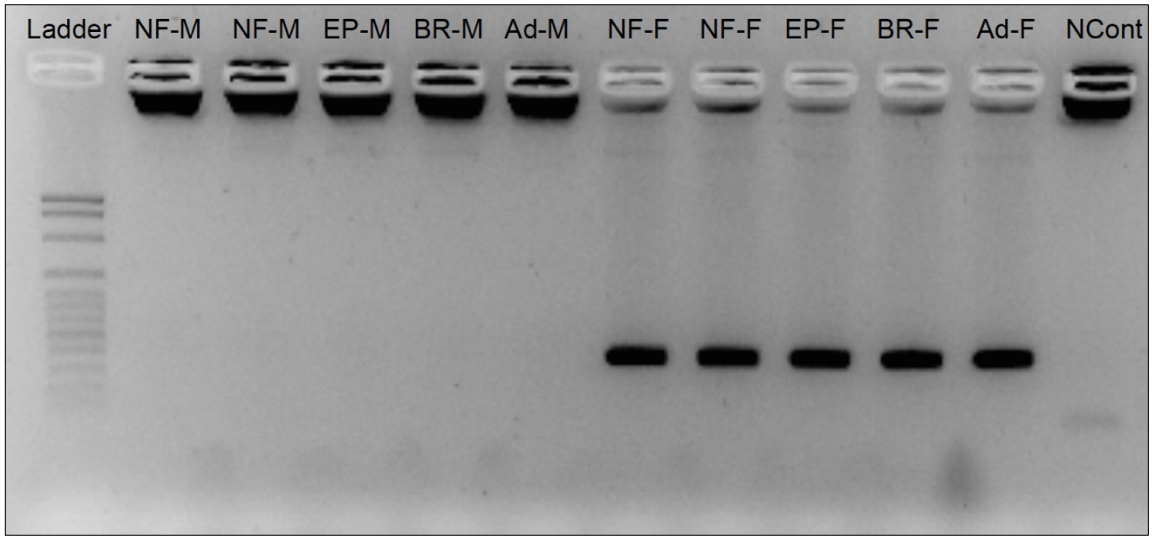


Figure 4.6. Example of genetic sex assay in *Cryptobranchus* and *Andrias*. M = male, F = female, NF = North Fork White River, EP = Eleven Point River, BR = Blue River, Ad = *Andrias davidianus*, NCont = negative PCR control.

A. 18S rDNA positive control.



B. W-linked locus 1024220.



CHAPTER FIVE

Genomic perspectives on the amphibian tree of life

ABSTRACT

Despite extensive investigation, relationships among extant lissamphibians remain murky, especially at the deep, inter-ordinal branches. In order to provide a genomic perspective on amphibian evolution, we developed an amphibian-specific sequence capture system targeting hundreds of conserved exons which is effective across this entire vertebrate class. Sampling 220 nuclear exons and complete mitochondrial genomes for 296 species of amphibians (representing 97% of families and greater than 50% of amphibian genera), we produced comprehensive species tree hypotheses for extant amphibians and identified extensive gene tree - species tree conflict throughout even the deepest branches of the amphibian phylogeny. We perform locus-by-locus interrogation of alternative topological hypotheses for inter-ordinal lissamphibian relationships, as well as for models of a non-monophyletic Amphibia. We find that phylogenetic signal deep in the amphibian tree varies greatly across loci, but in a manner that is not inconsistent with incomplete lineage sorting in the ancient populations of the ancestors of modern amphibians. Our results overwhelmingly support a sister relationship between frogs and salamanders, the Batrachia hypothesis. Multiple analyses (RAxML concatenated, Astral, MulRF) appear to converge on a small set of topological hypotheses for the relationships among extant amphibians. These results clarify several contentious portions of the amphibian tree, and in conjunction

with a set of vetted fossil calibrations, provide fresh insights into the timescales of amphibian diversification. But more importantly, this study provides insights into the sources, magnitude, and heterogeneity across the genome in large, phylogenomic data sets.

INTRODUCTION

Extant amphibians represent one of the most diverse and imperiled classes of vertebrates, with over 7,500 species (http://amphibiaweb.org). The ancient ancestors of lissamphibians were the first vertebrates to leave the water and to evolve a terrestrial existence hundreds of millions of years ago (Anderson 2008). Relationships among the three orders have remained contentious (Chen *et al*. 2015; Feller & Hedges 1998; Fong *et al*. 2012; Frost *et al*. 2008; Larson & Wilson 1989; Roelants *et al*. 2007), with support generated for several competing hypotheses of inter-ordinal relationships by different studies utilizing different sets of genetic and morphological characters and employing different methods.

Several broad-scale studies (e.g., Pyron & Wiens 2011; Roelants *et al*. 2007) have significantly advanced our understanding of evolutionary relationships among dee shallow amphibian lineages and have included very broad taxonomic sampling, but these have mainly relied upon organellar loci or small numbers of nuclear loci, potentially not reflecting the overall phylogenetic signal from across the nuclear genome. Several more recent studies have explored support for amphibian relationships using genome-scale data, including a very recent study sampling 95 nuclear exons for 156 species of frogs (Feng *et al*. 2017), but to date no study has combined dense sampling from the nuclear genome with

comprehensive taxon sampling at the levels of families and subfamilies across the three orders of amphibians.

Amphibian diversity comprises ~7,500 species, ~550 genera, ~80 families, and three orders (http://amphibiaweb.org). Although the monophyly of each order is practically indisputable (Anderson 2008; Pyron & Wiens 2011), the relationships among the three orders have remained murky. Three possible topologies exist for a monophyletic Amphibia. The Batrachia hypothesis surmises that frogs and salamanders are each other's closest relatives, while the Procera hypothesis (Feller & Hedges 1996) posits a sister relationship between salamanders and caecilians, and the Acauda hypothesis (named here) proposes that frogs and caecilians are monophyletic. Although different regions of the genome may potentially support any of these topologies, it is assumed that the species tree is only consistent with a single topology for interordinal amphibian relationships, and so the question of the relationships between frogs, salamanders, and caecilians becomes a model selection problem amenable to testing with data from across the genome. The multispecies coalescent model provides a useful framework in which to test questions about deep phylogenetic relationships among amphibians (Edwards *et al*. 2016).

The Batrachia and Procera hypotheses have received the most support in previous studies (Roelants *et al*. 2007; Wiens 2011), although recent genome-scale analyses with small numbers of taxa (Chen *et al*. 2015; Fong *et al*. 2012) have found some loci which support all three of these hypotheses assuming a monophyletic Amphibia. These studies have also found that some loci appear to support relationships where amphibians are non-monophyletic with respect to amniotes, and so here we consider all of these alternative topological hypotheses for inter-ordinal relationships among extant amphibians. There are

15 possible topologies relating frogs, salamanders, caecilians, amniotes, and the outgroup *Latimeria*. Three of these hypotheses imply amphibian monophyly, whereas the remaining 12 suggest non-monophyly of amphibians (Figure 5.1).

We also examine the extent to which genomic data sets may enable us to disentangle the deepest branches of the amphibian tree of life and to distinguish among loci which have retained informative phylogenetic signal across hundreds of millions of years of evolution from loci which primarily contain noise at the deepest timescales. Several studies have begun to explore whether gene-tree/species-tree discordance observed at deep phylogenetic timescales is genuine, or whether it may be the artifact of violations of the underlying molecular evolutionary models and/or may reflect an erosion of phylogenetic signal on these deepest branches (Fong *et al*. 2012). Simulation work suggests that, given certain combinations of rapid cladogenesis and/or large effective ancestral population sizes, deep coalescence may be expected on the order of tens or even millions of years of divergence (Oliver 2013). While many previous studies on amphibian relationships have either included broad taxon sampling and sparse genetic sampling (e.g., Pyron & Wiens 2011) or dense genetic sampling and sparse taxon sampling (e.g. Chen *et al*. 2015), ideally, studies would sample hundreds of species in order to overcome the potential effects of long branch attraction (Bergsten 2005), as well as hundreds of distinct genomic regions in order to overcome the potential effects of conflicting phylogenetic signal across different regions of the genome (Maddison 1997). It has also been proposed that the types of loci examined (coding versus non-coding) may exert a stronger influence on which phylogenetic topologies are supported by the data than the number of taxa sampled (Reddy *et al*. 2017). The proximate aim of our study is to generate genome-scale hypotheses for the

relationships and divergence times among extant amphibians. But, the more general aim is to examine the nature of signal and support in large phylogenomic data sets and to assess support for competing hypotheses about deep amphibian relationships.

METHODS AND MATERIALS

Taxon sampling

We assembled tissues and/or genomic DNA for a set of 310 amphibian species broadly covering family- and subfamily-level diversity and performed targeted sequence capture using our amphibian-specific probe set. Of this initial set of taxa, 296 produced useable sequence data, as assessed by having at least 30% of loci. Including eight of the 13 "model" taxa used in the probe kit design, our ingroup taxa consisted of 15 caecilians, 42 salamanders, and 239 frogs (Table 5.1). Within each of the three amphibian orders, we attempted to sample representatives from each recognized family, and from multiple sub-families in the case of particularly diverse families. We sought to sample taxa in rough proportion to the species richness of their respective families, but we were also constrained by the availability of tissues and by the quality of the genomic DNA available for sequencing. To guide our choices of taxa, we consulted previously published phylogenies and, where possible, we attempted to include similar taxonomic coverage of the different amphibian families. We especially attempted to sample deeply divergent lineages or taxa which would effectively break up long branches deep in the amphibian phylogeny or which would potentially provide resolution of recalcitrant nodes. Along these lines, capturing

large numbers of loci from the deeply divergent salamander *Siren intermedia* proved difficult because of the combination of an extremely large genome and deep divergence from any of the model probe taxa. Because of the importance of placing *Siren* for understanding the salamander phylogeny, we sequenced a multi-tissue transcriptome for this species and mined orthologs from this assembly to include with our alignments.

The choice of outgroups can have important downstream implications for phylogenetic inference (Wilberg 2015) and so we attempted to include multiple outgroups in order to avoid potential long branch attraction and to better estimate model parameters, including divergence times, for deep branches. We included four amniote outgroups (*Anolis carolinensis*, *Chrysemys picta*, *Gallus gallus*, and *Homo sapiens*) and the coelacanth (*Latimeria chalumnae*). For these five outgroup species we mined available genomic resources from GenBank (Taxon ID's: 28377, 8478, 9031, 9606, and 7897, respectively) and performed blastn searches with the NCBI BLAST algorithm (Johnson *et al*. 2008) to identify putative orthologs for each of these outgroup taxa.

Designing an amphibian-specific gene capture system

We sought to develop a targeted sequence capture system that would be effective across the vertebrate class Amphibia. Targeted sequencing by probe-based capture is becoming a popular and efficient method to obtain genome-scale data in non-model organisms (Faircloth *et al*. 2012; Lemmon *et al*. 2012), yet much of the power of these methods derive from having probe sequences designed specifically from moderately close relatives of the focal taxa. We designed a probe set which targets amphibian-specific

orthologs of a subset of the loci developed by Lemmon *et al*. (2012) and is designed from a diverse array of representatives of each of the three amphibian orders. We mined the publicly available genome sequence for the model frog *Silurana* (*Xenopus*) *tropicalis* (Hellsten *et al*. 2010) and complete transcriptomes for the salamanders *Ambystoma mexicanum* (Wu *et al*. 2013) and *Notophthalmus viridescens* (Abdullayev *et al*. 2013). To increase taxon representation in our probe design, we also developed and mined genomic resources *de novo* for six additional frogs (*Ascaphus montanus*, *Gastrophryne carolinensis*, *Mixophes schevilli*, *Pseudacris feriarum*, *Pseudacris nigrita*, and *Rana sphenocephala*), one salamander (*Desmognathus fuscus*), and one caecilian (*Ichthyophis bannanicus*), as well as transcriptomic resources for two additional salamanders (*Cryptobranchus alleganiensis* and *Ensatina eschscholtzii*). For each of these 13 amphibian taxa, we attempted to identify putative orthologs to a subset of 366 of the original 512 anchored hybrid enrichment loci. Although not all of the 366 target loci were identified in all 13 model taxa, each locus was represented by on average 11.1 model taxa.

Next, we designed a set of 120-mer DNA probes tiled across each of these loci for each of the 4,061 locus-by-model-taxon combinations. The tiling density of probes over target regions ranged from 1.0 to 2.0. Adjacent probe alignments to a given locus were reverse complemented in order to increase capture efficiency by capturing from both the heavy and light strands of the genomic DNA. Each locus consists of an evolutionarily conserved core region flanked by more variable regions on either side. Our probes for each model taxon covered these core regions and also extended into the flanks in order to increase the lengths of captured loci across diverse taxa. Across all 13 model taxa and 366 target loci, the region covered by our probes was ~1,090 bp per locus on average. In

practice, longer assemblies are generated from this type of data because the use of paired-end sequencing allows for the extension of sequenced regions beyond the core conserved regions covered by the probes. This set of 57,750 unique 120-mer probes was synthesized by Agilent Technologies.

Genomic library preparation and high-throughput sequencing

Genomic DNA was extracted using a standard silica column protocol. We performed 2% agarose gel electrophoresis for each sample to confirm that the gDNA was intact (degraded samples were subjected to less fragmentation during library preparation). Based on these gel results, each sample was fragmented using a Covaris sonicator to a mean fragment size of 300 - 500 bp. Individual samples were uniquely barcoded by ligation of dual index oligonucleotides and samples were pooled together in batches of 8-12 for multiplexed target capture. Capture reactions were carried out as in Lemmon *et al*. (2012). The enriched, captured products were amplified by low-cycle PCR with high fidelity polymerase. Resulting genomic libraries were bead-cleaned and pooled for sequencing (12-24 caecilians/frogs per Illumina HiSeq2500 lane, 24-60 frogs per lane). In total, the resulting libraries were sequenced across 14 lanes of an Illumina HiSeq2500 platform. Illumina sequencing was performed at the Florida State University Medical Center.

Nuclear locus assembly and characterization

Loci were assembled using a semi-reference guided strategy of partial kmer matching and extension alignment from raw paired-end Illumina reads (Prum *et al*. 2015; Rokyta *et al*. 2012). We compared results from assembling all three orders together versus assembling each order individually. Order-specific assemblies tended to recover slightly more loci (because the expectations of the orthology filters were generally better met). We performed reference-guided assembly for each order separately to generate sets of supercontigs representing groups of potentially orthologous loci. Some supercontigs represented groups of paralogous gene copies, and we quantified the number of potential ortholog sets for each supercontig for each of the three amphibian orders. Although it may have been possible to parse these groups of paralogs into presumptive orthologs, we took the conservative approach of excluding any locus for which a supercontig had two or more potential gene copies in any of the three orders. This ortholog filtering scheme resulted in a set of 220 putatively single-copy nuclear loci for which all five outgroup taxa were also sampled. Of this set of 220 loci, 194 contained representatives of each of the three amphibian orders.

MtDNA assembly and sample vetting

In order to verify the identities of samples and their placements in the resulting phylogenies, we exploited an inherent inefficiency of targeted sequence capture and were able to successfully assemble complete and partial mitochondrial genomes for nearly all taxa from the off-target bycatch reads. These mitochondrial fragments served as integrated "barcodes" with which we could verify the integrity of our taxon identifiers and ferret out

any potential cases of swapped tubes, mislabeled tissues, or misidentified taxa. Raw read data were assembled *de novo* with trinityrnaseq v2.0.3 (Grabherr *et al*. 2011) and the resulting assemblies were mined for mtDNA regions (detailed in supplemental methods). This procedure identified six instances of apparent pairwise transposition of samples which were also manifest as aberrant placements in preliminary gene trees (e.g., cases where placements of two well-established taxa appeared transposed).

Locus phasing

We expected to recover both allele copies for diploid individuals (and multiple allele copies for ploidy levels greater than 2) and we used a set of custom scripts to phase the resulting sequences within loci. By utilizing the paired-end nature of our Illumina reads and by taking into account the empirical fragment length distributions of the original sequencing libraries, we were able to establish phase for variable sites in our loci. In the case of ambiguities in phase, nucleotides were randomly resolved to one of their potential states. Although we generated phased data for every individual, we retained only one randomly chosen gene copy for each individual at each locus in order to greatly reduce the computational burdens of downstream analyses.

Multi-sequence alignment and reading frame determination

Multi-sequence alignment was performed in a nested procedure. We first performed four separate alignments for frogs, salamanders, caecilians, and the amniote and

Latimeria outgroups in MAFFT v7.221 (Katoh & Standley 2013) with the L-INS-i parameter settings. These sub-alignments were then combined using the MAFFT --merge function. Because the assembled loci represented a mixture of genuine and reverse complemented orientations, we used the known orientations of the outgroup taxa to report all aligned loci in their native orientations.

Preliminary examination of gene trees for these alignments also revealed that some taxa had very unexpected placements in a handful of gene trees (e.g., a salamander placed within frogs, or a caecilian nested within amniotes). These taxa were typically characterized by very long branch lengths in these gene trees and further scrutiny revealed that in nearly every case, large numbers of ambiguous or missing sites were apparently driving this pattern. To clean up these alignments, we implemented a taxon-filtering procedure for each locus which culled any taxa with greater than 85% missing and/or undetermined sites across an alignment or which had a terminal gene tree branch length greater than 5 times the average branch length for that tree. This filtering procedure removed less than 1% of the taxon-by-locus combinations, but greatly improved the consistency of estimated gene trees.

Next, culled alignments were examined by eye in order to correct obvious misalignment issues (e.g., large gaps anchored by a single leading nucleotide), and to establish reading frames across protein-coding portions of each locus. Between zero and two base pairs were trimmed from the ends of each alignment such that the first nucleotide represented the first codon position and the last nucleotide represented the third codon position of each locus. Alignment corrections were performed in Geneious R8 (Kearse *et al*. 2012) with hydrophobicity display enabled for translated amino acid sequences. In

many cases, manual alignment correction was significantly improved in the context of conserved polarity across codon sites in the alignments, especially around gaps where there can be multiple possible alignment resolutions.

From the potential set of 366 loci targeted by our probe set, we assembled on average 312 loci per individual greater than 250 bp. However, because we were especially interested in the deepest branches of the amphibian tree, we excluded any loci which were missing any of the five outgroup taxa. This reduced the number of retained loci to 253. We next excluded any loci with greater than 50% missing taxa (Hosner *et al*. 2016), which brought the number of retained loci to 220. This set of 220 loci was used to estimate the gene trees and species trees for amphibians and to conduct divergence time estimation (Table 5.2). Of this set of retained loci, 194 contained representatives of all three amphibian orders, and this was the data set that we used for testing the inter-ordinal amphibian topology.

Gene tree estimation

These loci are all protein-coding, and as such, a features-based partitioning strategy based on codon positions is a reasonable strategy for identifying and modeling variation in the underlying evolutionary dynamics of different portions of loci. For each locus, we used PartitionFinder2 (Lanfear *et al*. 2016) to simultaneously select among possible partitioning schemes and general models of molecular evolution using the greedy search algorithm (Lanfear *et al*. 2012). We also explored optimal partitioning schemes for the concatenated alignment in two different ways. As a first pass, we compared partitioning schemes based

on aggregate first, second, and third codon positions across all loci using the greedy algorithm. However, the greedy search is computationally prohibitive for complex partitioning schemes on large alignments. We used the rcluster algorithm in PartitionFinder2 to conduct a heuristic search of the model space for 660 maximum potential partitions from 220 loci partitioned by codon position (Lanfear *et al*. 2014). These best partitioning schemes for each locus were used to parameterize downstream gene tree estimation, and the best partitioning scheme for the concatenated alignment was used to inform concatenated maximum likelihood phylogeny estimation and the divergence time analyses.

Maximum likelihood (ML) estimates of each individual gene tree were obtained in RAxML v8 (Stamatakis 2014) for nucleotide models under several different topological constraint schemes. We estimated parameters for a GTR+Γ model (the best fit model for every locus). For each analysis, 500 rapid bootstrap analyses were conducted followed by a series of 20 slow ML optimization steps before the full ML analysis. For each locus, we performed 18 separate RAxML analyses. The first three were: a completely unconstrained analysis, an analysis enforcing intra-ordinal monophyly of amphibians and amniotes only, and a constraint enforcing a monophyletic Amphibia with monophyletic orders. Additionally, we performed analyses for each of the 15 backbone constraints representing the 15 possible tree topologies relating frogs, salamanders, caecilians, amniotes, and the outgroup *Latimeria* (Figure 5.1). In all constraints, intra-ordinal amphibian monophyly and the monophyly of amniotes were both enforced. In the 15 models for deep tetrapod relationships, no constraints were imposed on relationships within the amphibian orders or within amniotes.

We used the Robinson-Foulds (RF) distance (Robinson & Foulds 1981) as a metric to quantify discordance between gene trees, between gene trees and species trees, and between species trees estimated in different ways. We used the RFdist() function in Phangorn v2.1.1 (Schliep 2011) to calculate RF distances between pairs of trees. For comparisons of species trees to other species trees, this was relatively straightforward because these trees all had the same number of tips. However, when comparing gene trees to other gene trees or when comparing gene trees to species trees, the numbers of taxa in each pairwise contrast may differ due to some gene trees missing some taxa. In these cases, we used custom scripts to leverage the drop.tip() function in the Ape v4.0 R package (Paradis *et al*. 2004) to prune each tree to contain only those tips that were shared in common. In these cases, we also calculated the normalized RF distance between trees, where the raw RF distance is divided by the maximum RF distance: (2 * [n -3]) for n taxa.

Species tree estimation

We employed several methods to identify sets of phylogenetic hypotheses for the topology of the amphibian species tree. As a first pass, we concatenated all 220 loci together into a single alignment with 291,921 characters for 301 taxa (296 amphibians, four amniotes, and the *Latimeria* outgroup). We analyzed this concatenated data set in two distinct ways. First, we generated maximum likelihood estimates of the topology and branch lengths for amphibians in RAxML for three different partitioning schemes to account for variation in rates across different groups of sites. We considered data sets with one partition (an un-partitioned analysis), with three partitions delimited by aggregate first,

second, or third codon positions, or with a set of 76 partitions identified through a heuristic search in Partitionfinder2. Each of these three analyses were also repeated for three additional concatenated data sets for which particular sequences had been trimmed from alignments on the basis of missing alignment sites at thresholds for either 50%, 75%, or 90% of sites being present for any taxon. The missing-data-culled alignments were generated to investigate the effects of short or gappy sequences on the stability of placements in the anuran tree, as discussed below. In total, 12 concatenated RAxML analyses were conducted for the combination of three partitioning schemes and four data filtering schemes. These RAxML analyses also included 500 rapid bootstrap replicates, followed by 20 slow ML optimization steps before the full ML analysis. For these concatenated RAxML analyses, the only constraint we enforced on the topology was that *Latimeria* was the outgroup.

To account for different coalescent histories between loci, we also explored two different "shortcut" methods for estimating species trees from collections of estimated gene trees. We first performed species tree estimation in Astral2 v4.10.0 (Mirarab & Warnow 2015) from the 500 bootstrap replicates and the best ML tree for each of the 220 loci for the unconstrained gene tree analysis. We employed site- and gene-level multilocus bootstrapping and conducted 250 Astral bootstrap replicates. Although Astral2 is technically not a coalescent method, it is statistically consistent with the coalescent model for large numbers of loci. We also estimated a species tree topology which would minimize the composite Robinson-Foulds distance between all input gene trees and the species tree using the program MulRF (Chaudhary *et al*. 2014). For both of these methods, analyses

were run using unconstrained RAxML gene trees estimated for each of the four filtering schemes for missing sites.

Support across loci for inter-ordinal relationships

We initially sought to quantify support for competing inter-ordinal models for amphibian relationships. We quantified the proportion of individual gene tree bootstrap replicates that support each of the three monophyletic amphibian models using custom scripts based on the is.monophyletic() function in the Ape to bin bootstrap replicate gene trees by support for the different inter-ordinal topologies. We then rank ordered loci supporting each topology by decreasing levels of support as measured by the number of bootstrap replicates supporting each competing model. At the extremes of each category of genes, a few genes provide either very decisive, or nearly equivocal support for the best supported model, while many genes appear to provide intermediate levels of support for any particular model. These bootstrap topology analyses are informative about the direction of support for competing topological models, but being essentially bounded between zero and one, they may obscure the magnitude of support. For ML estimates of gene trees, the Akaike information criterion (AIC) (Akaike 1974) is used to compute relative model support and select the best model, with ΔAIC reflecting the strength of support against alternative hypotheses. Although it is not necessarily clear how best to count parameters in competing models, or how best to calculate the corrected AIC (AIC_C), the ΔAIC estimates provide an unbounded metric of support against alternative models.

Here, the magnitude of the differences in the likelihoods between competing models drives the patterns of ΔAIC across loci regardless of how parameters are counted in the gene trees.

To examine support across loci for the inter-ordinal amphibian relationships, we also applied a recently proposed method of gene genealogy interrogation (GGI) (Arcila *et al*. 2017) to provide a statistical test of support for different competing topological models on a locus-by-locus basis. We tested the ability to reject alternative topological hypotheses for the 15 possible inter-ordinal models by performing an approximately unbiased (AU) test (Shimodaira 2002) for each locus. These topological tests were performed three ways: once using all 15 constrained RAxML trees, again using all 15 constrained trees plus the unconstrained tree, and a third time considering only the three constraint trees which enforce the monophyly of amphibians. For each GGI analysis, we scored which of the possible topologies was identified as the best topology on the basis of its log likelihood (-ln(L)) for each locus. We then plotted the cumulative number of loci supporting each alternative topological hypothesis, rank ordered by decreasing statistical significance of the AU P-value. AU tests were performed in a development version of PAUP* v4a.151 (Swofford 2015).

Support across loci for neobatrachian relationships

In the course of comparing the various species trees which we had estimated, we observed strange placements of the neobatrachian frog *Nasikabatrachus*. Pursuing this further, we identified a set of branches deep in the anuran phylogeny which were recovered differentially between concatenated and multi-locus methods, and which differ from

previously proposed relationships. Our desire to thoroughly explore these unexpected results motivated us to extend the constrained, locus-by-locus tests of topology which we used to investigate inter-ordinal relationships to scrutinize support for this contentious portion of the frog tree.

Divergence time estimation

Divergence times were estimated in the MCMCTree program in the PAML package v4.9e (Yang 2007) using a set of 25 fossil calibrations, 19 of which were recently used by Feng *et al*. (2017) (Table 5.3). These fossil calibration points cover many of the deep branches within tetrapods and within the amphibian orders (Figure 5.2). We started by estimating the substitution rate for each of 220 loci which were partitioned by codon position in the PAML program baseml under a GTR+Γ model of nucleotide substitution with five discretized rate categories. For each locus, the gene tree topology was fixed to that from the unconstrained RAxML gene tree analyses and the root age for the divergence between *Latimeria* and tetrapods was set to 450 MYA (Benton *et al.* 2015). Based on the average substitution rate across all loci and codon positions, a mean substitution rate across loci of 0.899 substitutions per billion years was estimated and used to parameterize a diffuse gamma Dirichlet prior on locus rates (rgene_gamma) as G(1 1.11). A concatenated alignment of 220 loci, partitioned into aggregate first, second, and third codon positions was used as input for mcmctree. 25 fossil calibration points were used to constrain nodes in the Astral species tree (estimated from a filtering scheme retaining sequences with >90% of sites present) with the 95% confidence interval of prior densities falling between the

lower and upper (soft) bounds of the estimated divergence time range, and with 2.5% prior density extending above and below these bounds. Maximum likelihood estimates (MLEs) of branch lengths were obtained by approximate likelihood and the gradient and Hessian matrices were calculated at MLEs of branch lengths with the usedata=3 option. The output of these runs were used as input for the estimation of divergence times in mcmctree with the usedata=2 option, uncorrelated rates across loci, and a GTR+Γ model of nucleotide substitution. Two independent MCMC chains were run for 100,000 generations of burnin, and subsequently sampled every 100 generations until 10,000 samples were collected. 95% confidence intervals on divergence times were calculated. The posterior mean divergence times were nearly identical between the two runs (R^2 = 0.99). We also performed divergence time analyses for the unfiltered alignment, and for the 50% and 75% sites-present data sets, and results were largely consistent among runs.

RESULTS

Taxon sampling

We obtained sufficient quantities of high quality genomic DNA from most individuals, but several key taxa were either removed from library preparation because they lacked sufficient DNA quantities for targeted enrichment, or because they did not produce significant numbers of loci. Some key taxa that had to be excluded include the deeply divergent caecilian *Rhinatrema* and the salamander *Siren*. In the latter case, very few loci were assembled, likely due to the deep divergence between *Siren* and the salamanders used

to design our probe kit, and because of the very large genome size in this genus. We obtained fresh tissues for an individual *Siren* and used RNA-seq to sequence expressed transcripts. We then mined the resulting assembly for orthologs to our set of exons, and integrated these data into the respective alignments of all other taxa. Beyond these cases, nearly all of our intended sampling was successful and we assembled a data set of 239 anurans, 42 salamanders, 15 caecilians, four amniote outgroups, and *Latimeria*.

Designing an amphibian-specific gene capture system

The hybrid exon capture system which we designed appears to be effective across this entire vertebrate class and successfully recovers large numbers of informative loci. Figure 5.3 shows that the genome size of the target taxa and the evolutionary distances from target taxa to probe taxa are two main correlates that predict capture efficiency. In large genomes, the targets of the capture probes are effectively diluted in the resulting fragment pools, driving down the rate of locus recovery. Additionally, as the divergence time between the focal species and the species from which the capture system was designed increases, locus recovery decreases because the pairwise nucleotide differences between species increases, limiting effective hybridization of probes to targets. However, our results also suggest that the effects of large genome size can be partly offset by utilizing probes designed from relatively closely related species.

Genomic library preparation and high-throughput sequencing

Most individuals were sequenced using paired-end 150 bp reads on an Illumina HiSeq2500 platform run in Rapid Run mode with onboard cluster generation. A subset of the microhylid frogs (those with PT identification codes) were sequenced on an Illumina MiSeq platform with paired end 300 bp reads (longer read lengths were required here because the insert size of these libraries was longer than other taxa). In total, we generated approximately 14 lanes of sequence data across this set of taxa.

Nuclear locus assembly and characterization

Our assembly pipeline identified a set of 392 putative loci across the three orders of amphibians. However, many of these represented potentially paralogous loci identified during the orthology determination steps. We initially attempted to confirm that each paralog from a set of presumptive homologous loci could be readily distinguished from closely related sequences. In several cases, we founds instances of potential mis-resolution of paralogous gene copies, and so with an aim to avoid potentially introducing unresolved paralogs into our analyses, we excluded from consideration any locus which had been assembled with greater than one other homolog, regardless of whether our orthology determination steps had resolved these or not. This paralog exclusion reduced the number of loci to 302. Next, when considering the amniote and *Latimeria* outgroups, we enforced that every locus should be present in all five outgroup taxa, and this brought our list of retained loci to 253. From this set, we next sought to exclude any locus with greater than 50% of taxa missing, which brought out final set of loci to 220. Of these, 194 loci had at least one representative of each of the tree amphibian orders, and these loci were used to

explore support for inter-ordinal relationships. Another, different set of 194 loci (not exactly the same set as the inter-ordinal loci) were also identified which contained sampling of seven main neobatrachian frog lineages, and it was this set of loci which we included for the topology testing at the base of Neobatrachia.

Across our set of 301 taxa, the numbers of loci per taxon and the proportion of sites present in each locus varied substantially and was also highly skewed across the three amphibian orders (Figure 5.4). Individual frogs, caecilians, and salamanders had on average 214, 165, and 146 loci present, and these loci had on average 97%, 75%, and 66% of sites in the alignments present, respectively. This variation in capture success by order is largely accounted for by the deep divergences between many of our sampled taxa and the handful of representatives of those two orders which were included in our probe kit, and by the smaller average size of frog genomes relative to salamanders or caecilians.

MtDNA assembly and sample vetting

We successfully assembled the raw reads from every individual which we sequenced. For roughly two thirds of taxa (n = 188), we were able to identify a single contig between 13 kbp and 19 kbp in length which appeared to constitute the complete or nearly complete mitochondrial genome. For all but eight of the remaining taxa, we were able to recover contigs of over 1 kbp with affinity to known mtDNA regions. Our blast searches resulted in the identification of six pair-wise instances of misidentified samples which we verified from blast results and then corrected in our alignments and trees.

Orthology and phasing

Our hybrid probe design where each locus is represented by probes made specifically from multiple, divergent taxa permits more effective capture and enrichment across deep divergences, but may also increase recovery of somewhat similar, potentially paralogous sequences, relative to a single-model-taxon probe kit design. To overcome this inherent limitation, we used an orthology determination pipeline to distinguish potential paralogs from each other at each locus. As described above, we decided to conservatively exclude any loci for which we had any doubt of orthology, and this reduced the 392 locus data set to a 220 locus data set.

Multi-sequence alignment and reading frame determination

We successfully aligned all taxa and corrected slight breaks in codon triplets by examining each of the 220 alignments by eye. It was possible to determine the reading frame for all loci, and this information was used in the partitioning analyses. In order to focus our attention on regions of loci which could readily be modeled as coding sequence, we truncated the non-coding ends from all loci at the 5' and 3' ends of start and stop codons respectively, when alignments included these coding features. The total length of the concatenated data set generated from these groomed alignments was 291,219 bp. Details of loci are provided in Table 5.2.

Gene tree estimation

Unconstrained, partitioned gene tree estimates across all 220 loci were obtained and compared to assess the extent to which the topologies estimated by genes across the genome were concordant or discordant with each other. We calculated the distribution of Robinson-Foulds (RF) distances across all RAxML bootstrap replicate trees for each locus against all other loci. RF distances of zero indicate that two trees have identical topologies, and increasing RF distances indicate that trees are less similar. Because these gene trees each had (potentially) different numbers of taxa and the taxon sets may not necessarily overlap, we applied a custom script using the R commands lapply() and drop.tip() from the Ape package (Paradis 2004) to reduce all bootstrap trees for each pair of loci to a common set of taxa before calculating the distribution of all unique pairwise comparisons of bootstrap gene trees from both loci. This procedure was repeated across all 220 loci. To account for the different numbers of taxa across loci in these comparisons (which result in different maximum values for RF), we normalized these RF density plots so that all loci could be plotted on a common axis (thin blue lines in Figure 5.5). These results indicate that there is either substantial discordance in topology between nearly all loci, or that there is substantial noise in the signal from these gene trees, or some combination of both factors.

Species tree estimation

Similar RF comparisons were also carried out between the bootstrap trees from each gene tree and the bootstrap trees from our preferred Astral species tree topology. These results (thin red lines in Figure 5.5) also show that there is substantial discordance

between individual gene trees and the estimates of the species tree. However, as discussed below, the discordance between bootstrap replicates of the Astral species tree is relatively low, suggesting that the Astral analysis arrives at a set of species tree estimates which are highly concordant with each other, despite extensive discordance between gene trees and between gene trees and the resulting species tree.

Prior to scrutinizing patterns of missing sites in our loci, Astral recovered a strange topology with *Nasikabatrachus* placed as sister to Microhylidae (Figure 5.6), in contrast to all previous molecular studies (e.g., Biju & Boussuyt 2003). After filtering sequences with high proportions of missing sites (as described above), we re-estimated the species tree with Astral, RAxML, and MulRF, and we compared the pairwise RF distances between these new topologies. Overall, all of the missing-sites data filtering schemes produced species trees which placed *Nasikabatrachus* sister to *Sooglossus*. The resulting set of species trees had normalized RF distances that ranged from 0.000 - 0.058, and the majority of the differences in these trees involved minor switches of more shallow taxa. We selected the topology from the Astral analysis with at least 50% of all sites present in all sequences as our point estimate of the species tree in downstream analyses, as depicted in Figure 5.7. The topologies of the Astral and MulRF species trees and the RAxML concatenated trees were highly concordant with each other as gauged by RF distances (normalized RF ranged from 0.000 - 0.058). The topologies of the RAxML concatenated trees varied little across different data filtering schemes (unfiltered, 50%, 75%, or 90% of sites present) and across partitioning schemes (unpartitioned, partitioned by aggregate codon position, or partitioned by 76 clusters identified in rcluster analyses). However, the MulRF and RAxML trees estimated from unfiltered data sets differed from the unfiltered Astral topology in placing

Nasikabatrachus sister to *Sooglossus*, while Astral recovered *Nasikabatrachus* sister to a clade containing all microhylids. Bootstrap support across branches in the RAxML concatenated trees were generally much higher than in the Astral trees (MulRF does not provide the option of bootstrapping), which may reflect that the Astral bootstrap values are capturing heterogeneity across sites and loci, leading to lower support for branches leading to taxa with increased missing data.

Support across loci for inter-ordinal relationships

By numerical tally, nearly half of all 194 loci (n = 98) with all three orders present have RAxML "best" gene trees that support Batrachia, followed by 51 and 45 gene trees which support Procera and Acauda, respectively. In Figure 5.8, the lower panel depicts the extreme variation observed across loci in the proportion of bootstrap gene trees which support each of the two other topologies from their "best" topology. The upper panel in Figure 5.8 shows that the magnitude of support against these rejected topological models for each locus also vary substantially, and that qualitatively, there is some concordance between the strength of support for the preferred topology and with the proportion of bootstrap replicates (essentially the proportion of sites in the alignment) which support different models.

We also adopted a recently proposed method to perform model selection among all possible sets of constrained trees on a gene-by-gene basis: gene genealogy interrogation (GGI) (Arcila *et al*. 2017). Approximately unbiased (AU) tests of topology (Shimodaira 2002) are conducted for constrained ML gene trees for each of the 15 possible inter-ordinal

topologies (allowing the placement of amniotes relative to the amphibian orders to be tested). Rank-ordered P-values are plotted for sets of genes supporting each competing topology. The GGI results (Figure 5.9) suggest that while the topology of the species tree is likely to be consistent coalescent expectations under the Batrachia hypothesis, substantial discordance exists across the genome in terms of which of the 15 possible topological models are supported by which genes. The majority of strong signal coming from the nuclear genome is in support of Batrachia. Although the Acauda and Procera hypotheses receive non-trivial proportions of support, numerically the Batrachia hypothesis is unable to be rejected.

Support across loci for deep neobatrachian relationships

Neobatrachia is recovered as monophyletic in all analyses, and all methods support *Heleophryne* as sister to all other neobatrachians. The remaining neobatrachian lineages form two clades: (Hyloidea + Myobatrachidae + *Calyptocephalella*) and (*Sooglossus* + *Nasikabatrachus* + Microhylidae + Afrobatrachia + Ranoidea). These groupings are consistent between ASTRAL and RAxML, and are in line with previous studies. Within the latter clade, relationships among five main lineages differ between the ASTRAL, MulRF, and concatenated RAxML trees, differ from previous studies with respect to the placement of *Nasikabatrachus* (Figure 5.10). For this more complex case with five focal taxa, the probability that anomalous gene trees may numerically dominate the sampled set of gene trees is non-zero. Further scrutiny here revealed that short and gappy sequences in alignments appear be driving the placement of *Nasikabatrachus* as sister to *Oreophryne* at

the gene tree levels, instead of the typical placement as sister to *Sooglossus* (Figure 5.11). 98 gene trees supported *Nasikabatrachus*+*Oreophryne*, while slightly fewer (n = 94) supported the canonical *Nasikabatrachus*+*Sooglossus* relationship (two loci supported neither arrangement). Yet, when alignments were filtered to exclude individuals having less than 50%, 75%, or 90% of sites present for any locus, the Astral and MulRF trees supported the neobatrachian topology of the concatenated analyses. For example, the unfiltered Astral tree (Figure 5.12) and the unfiltered MulRF tree (Figure 5.13) both support *Nasikabatrachus* sister to Microhylidae. In contrast, the unfiltered RAxML tree supports *Nasikabatrachus*+*Sooglossus* (Figure 5.14), as do the trees generated from alignments filtered for missing sites, depicted by trees for data sets containing at least 90% of sites in all taxa for Astral (Figure 5.15), MulRF (Figure 5.16), and RAxML (Figure 5.17).

Divergence time estimation

Our results suggest a much more recent origin of neobatrachian frogs than suggested by most recent studies (Roelants *et al*. 2007; Wiens 2011, but see Feng *et al*. 2017). Figure 5.18 depicts the divergence times estimated in mcmctree, with 95% confidence intervals highlighted in gray. We estimate that the *Latimeria*-Tetrapoda split took place in the Silurian or early Devonian, that the Amniota-Lissamphibia split happened in the late Devonian or early Carboniferous, that the ancestors of caecilians diverged from the ancestors of frogs and salamanders in the late Carboniferous or early Permian, and that frogs split from salamanders in the middle Permian. Within frogs, our results suggest that

Leiopelmatoidea split from all other frogs in the late Triassic or early Jurassic, that *Leiopelma* and *Ascaphus* diverged in the early Jurassic, that Neobatrachia split from Pelobatoidea in the middle Jurassic, that *Heleophryne* split from all other neobatrachians in the late Jurassic or early Cretaceous, and that the splits between Hyloidea and Ranoidea, between Microhylidae and (Natatanura+Afrobatrachia), and between Natatanura and Afrobatrachia all took place roughly contemporaneously in the middle Cretaceous. Hyloidea, Microhylidae, and Natatanura all show evidence for marked upticks in lineage diversification rates around the Cretaceous-Tertiary boundary (~65 MYA). This surprising finding corroborates very recent work by Feng *et al*. (2017) which also supported these more contemporary divergence times in these specious groups of frogs. Our results suggest more recent divergences between the Cryptobranchoidea and the Salamandroidea than suggested by other work (Roelants *et al*. 2007; Vietes *et al*. 2011; Zhang & D.B. Wake 2009), with dates placed at the late Jurassic. However, our divergence times with Plethodontidae are more in line with the work of Shen *et al*. (2015). Our divergence times in caecilians are markedly younger than most other studies (Roelants *et al*. 2007; San Mauro *et al*. 2004; Zhang & M.H. Wake 2009). This effect may be due to the topology which we recover placing *Rhinatrema* sister to *Ichthyophis*.

DISCUSSION

Inter-ordinal amphibian relationships

Previous work has supported either the Batrachia (e.g., Roelants *et al*. 2007) or Procera (e.g., Feller & Hedges 1998) hypotheses for relationships among the extant amphibian orders, although few studies have found unanimous support. More recent work sampling much greater numbers of loci has demonstrated that all three possible inter-ordinal topologies are observed in empirical gene trees (Shen *et al*. 2015), and that some gene trees even recover a paraphyletic Amphibia with respect to amniotes (Fong *et al*. 2012). These observations are not inconsistent with the notion that the topology of the species tree is in line with the Batrachia hypotheses, but that either a historical demographic signal of incomplete lineage sorting has persisted over hundreds of millions of years, the phylogenetic signal at the base of the amphibian tree has eroded out of many loci, or there is homoplasy affecting the resolution of these gene trees which support non-Batrachia (or at least non-monophyletic Amphibia) and a lack of informative signal for these deepest divergences in amphibians. All three of these possibilities may also be overlaid on top of each other by different loci. Although there substantial is variation across loci, our results strongly support the Batrachia hypothesis (frogs+salamanders) for inter-ordinal amphibian relationships. That the distribution of inter-ordinal models supported by the 194 gene tree bootstrap replicates varies so greatly across loci suggests that even within some loci, the direction of support for these deepest amphibian relationships varies across sites. This also suggests that were one to sample only a few loci at random, their numerical distribution across possible deep models could lead to strongly supported, spurious support for an alternative inter-ordinal model. The results from the GGI analysis strongly suggest that Batrachia is favored for the relationships among orders, but that many loci may lack definitive signal this far back in the phylogeny. The magnitudes of the ΔAIC values in the

model comparison analysis for deep relationships also suggest that while some loci very strongly reject the alternative topological models, many loci contain either very little relative information content about these deep portions of the tree, or are completely agnostic with respect to inter-ordinal relationships. Overall, we reject the hypotheses of a paraphyletic Amphibia in favor of inter-ordinal relationships consistent with the Batrachia hypothesis.

Although at least one simulation study (e.g., Oliver 2013) has demonstrated the potential for incomplete lineage sorting (ILS) to affect deep branches and to potentially lead to genuine gene tree - species tree discordance deep in the past, we are skeptical that this is the case in amphibians. Nearly 12 million years separate the 95% credible intervals for the divergences between caecilians and batrachians (frogs and salamanders), and it would seem particularly unlikely for ILS to generate discordance in gene trees over such a long period of time. Additionally, the GGI results suggest that nearly all loci cannot confidently discriminate between the alternative topologies for inter-ordinal relationships. However, the observation that roughly half of all loci support Batrachia while the remainder of loci are roughly evenly split between support for Procera or Acauda is not inconsistent with expectations under a coalescent model of genetic drift in finite populations. Still, work by Edwards *et al*. (2004) and Poe & Chubb (2004) suggest that incomplete lineage sorting may be a factor driving the observation of discordance across genes even at deep timescales.

Relationships among caecilians

The topology of the caecilian tree which we recover is largely consistent with previous investigations using nuclear and mtDNA markers. Although our taxon sampling for caecilians was somewhat limited compared to salamanders or frogs, we nonetheless sampled nine of the ten recognized families of caecilians and recover a very similar family-level topology to previous studies. One notable exception here is the placement of *Ichthyophis*. Most previous work has suggested that the earliest divergence in the ancestors of living caecilians was between the common ancestor of (*Rhinatrema* + *Epicrionops*) and the common ancestor of all other extant caecilians (San Mauro *et al*. 2004, Zhang & M.H. Wake 2009). In contrast, all of our species tree analyses and nearly all gene trees support a clade containing (*Ichthyophis* + *Epicrionops*) as sister to all other caecilians. This perplexing result might be an artifact of long branch attraction between these two deeply divergent families which resulted because the *Rhinatrema* individual which we had selected to break the long branch at the base of caecilians had to be excluded from analyses because of low locus recovery. Our results corroborate the monophyly of the families Dermophiidae and Siphonopidae, which are recovered as sister to each other, and the family Indotyphlidae is recovered as the sister to this clade. Caeciliidae and Typhlonectidae are recovered as sister clades, and together they form the sister clade to the (Dermophiidae+Siphonopidae+Indotyphlidae) lineage. Scolecomorphidae is the sister lineage to the aforementioned families.

Relationships among salamanders

Our estimated topology for the salamander phylogeny is also largely concordant with previous investigations. We successfully sampled all ten of the salamander families and were able to include multiple representatives for most of these families. Our results unequivocally support an initial divergence event in crown salamanders between the Cryptobranchoidea (Cryptobranchidae + Hynobiidae) and the Salamandroidea (all other salamanders). At nearly the same time within the Salamandroidea, the divergence between the Sirenidae and all other salamandroid salamanders likely took place. Our study corroborates other work (e.g. Roelants *et al*. 2007) which has suggested that Sirenidae diverged after the split with Cryptobranchoidea but before the divergences among other members of Salamandroidea. After the branch leading to Sirenidae, the next divergence within Salamandroidea was between a lineage comprising the Ambystomatoidea (Ambystomatidae+Dicamptodontidae) and Salamandridae and a lineage comprising successive divergence events between the Proteidae, Rhyacotritonidae, Amphiumidae, and Plethodontidae. Within the diverse Plethodontidae, we recover monophyly of both recognized subfamilies, the Hemidactyliinae (*Batrachoseps*, *Bolitoglossa*, *Eurycea*, *Gyrinophilus*, *Hemidactylium*, *Nyctanolis*, and *Pseudotriton*) and the Plethodontinae (*Aneides*, *Desmognathus*, *Karsenia*, *Phaeognathus*, *Plethodon*).

Relationships among frogs

The anuran portion of our phylogeny is the region with the greatest topological discordance between methods and data sets, and which most strikingly disagrees with

previous studies. Overall, the deepest branches of the frog tree and the most shallow-scale relationships are highly concordant between our different data sets and methods.

Among the early-branching frog lineages (Archeobatrachia), our results corroborate previous investigations (Roelants *et al*. 2007; Zhang *et al*. 2013). The earliest branching lineages (superfamilies, by some authors) of frogs which form monophyletic groups to the exclusion of all other anurans are, from root to tip, the Leiopelmatoidea (Leiopelmatidae + Ascaphidae), the Bombinatoroidea (Bombinatoridae + Discoglossidae + Alytidae), the Pipoidea (Rhinophrynidae + Pipidae), and a clade containing (Scaphiopodidae + Pelodytidae + Pelobatidae + Megophryidae). All other frogs are classified in the Neobatrachia, a globally distributed clade of frogs containing over two thirds of extant amphibian species diversity. Consistently between methods and data sets, and in accord with previous work, we find that *Heleophryne* is recovered as sister to all other neobatrachians. The remaining neobatrachians form two reciprocally monophyletic clades. The earliest divergence event in the first of these clades is the split between (Calyptocephalellidae + Myobatrachidae) and the Hyloidea (Nobleobatrachia) superfamily. The second neobatrachian clade distal to *Heleophryne* is composed broadly of the Microhyloidea, the Afrobatrachia (Hemisotidae + Hyperoliidae + Brevicepitidae + Arthroleptidae/Astylosternidae), and the Ranoidea (Natatanura). The placements of two deeply divergent, obscure frogs, *Sooglossus* from the Seychelles and *Nasikabatrachus* from India initially differed between our multi-locus species tree analyses (MulRF and Astral) and concatenated ML analyses. After we discovered the tendency of the short *Nasikabatrachus* loci to (spuriously) cluster with loci from the New Guinea endemic microhylid *Oreophryne*, we began to question the surprising result from Astral which

strongly supported *Nasikabatrachus* as sister to Microhylidae. *Nasikabatrachus* and *Sooglossus* have traditionally been recovered as sister taxa which themselves are sister to (Microhylidae + Afrobatrachia + Ranoidea) (e.g., Pyron & Wiens 2011). But, the Astral tree built from loci without outlier taxa filtered for missing sites appeared to strongly support *Nasikabatrachus* as sister to Microhylidae. But, when we filtered out sequences for all taxa which had greater than 50% of sites missing, the apparent support for this surprising result vanished and we recovered the canonical *Nasikabatrachus* + *Sooglossus* configuration. However, the unconventional recovery of Microhylidae + Afrobatrachia by every method and every data is significant and stands in contrast to all previously proposed topologies with the very recent exception of Feng *et al*. 2017. Within the hyperdiverse Microhylidae, our results clarify relationships among the subfamilies, and are largely in line with results from de Sá *et al*. (2012) and a more recent genomic study from which we drew many of our microhylid samples (Peloso *et al*. 2016).

Although our study largely recapitulates the topologies for amphibian relationships established by numerous other studies, this work presents an important dissection of signal and support across the amphibian genome, demonstrating the potential for studies utilizing smaller numbers of loci to potentially be misled because of stochastic sampling effects. Without the genomic perspectives developed in our study, it would not have been possible to assess the nature of support across the genome. And our work also helps to clarify some of the murky neobatrachian branches which have been contentious in previous investigations. An aspect of our study where the genomic scale of our data were particularly informative (and simultaneously burdensome) was in estimating parameters such as branch

lengths and divergence times, where sampling across the substitutional variance present in the genome may provide more accurate estimates.

Amphibian diversification through time

Our study demonstrates the utility of large phylogenomic data sets for estimating divergence times across ancient lineages and the relatively small confidence intervals which we recover for nodes in our tree likely stem from the large amount of data with which we were able to perform parameter estimation. Our use of multiple fossil calibrations throughout the deeper portions of the phylogeny may also account for the relatively small variance in our estimates. The three amphibian orders appear to have diverged from each other by the end of the Permian, and to have persisted through the Permian-Jurassic mass extinction event. The Jurassic and Cretaceous saw the origins of most of the higher-order salamander and frog lineages, while caecilian diversification appears to have occurred much later than suggested by several previous studies (Roelants *et al*. 2007, Wiens *et al*. 2011). The mass extinction event at the end of the Cretaceous appears to coincide with increases in diversification in three major frog lineages (Hyloidea, Microhylidae, and Natatanura) (similar to Feng *et al*. 2017) and possible the diversification of the major caecilian lineages.

CONCLUSIONS

Although the empirical aims of this study were to bring nuclear genomic data to bear on the questions of inter-ordinal and deep intraordinal amphibian relationships and to establish a timescale in which to understand patterns of amphibian diversification, we also address pressing and timely questions about the sources, magnitude, and heterogeneity of phylogenetic signal in phylogenomic data sets. In terms of amphibian evolution, our study corroborates many of the previously conducted phylogenetic studies, but does add some clarity to relationships within Neobatrachia, in line with the only other nuclear phylogenomic study across all frogs (Feng *et al*. 2017). However, importantly, these results also indicate that support for competing phylogenetic hypotheses can vary substantially across different exonic regions of the genome, indicating that large numbers of loci (hundreds of more) may be required to adequately account for stochastic coalescent processes which can generate gene tree - species tree discordance, or to overcome noisy loci whose phylogenetic information content may have eroded over time or been altered by selective forces. Our results also highlight some of the potentially latent sources of systematic error in phylogenomics, as evidenced by our findings that patterns of missing sites within loci can drive spurious support for incorrect phylogenetic hypotheses, and that these erroneous placements at the level of gene trees can propagate up to the level of species trees to provide positively misleading, strong support for an inaccurate result. Despite the inherent discordance across nuclear gene trees, we arrive at a set of credible topologies for the backbone structure of the amphibian (species) tree of life. The amphibian-specific exon capture system reported here provides a rich suite of nuclear loci for conducting phylogenomic studies across this entire vertebrate class. The current iteration of this probe set is extensible in light of newly available genomic resources in amphibians, and future

versions can now be targeted to specific amphibian clades along the way to eventually generating a species-level phylogeny for all extant amphibians.

Table 5.1. Taxon sampling.

ID	Institution/Collector	Specimen ID	Order	Family	Genus	Species
I4372	MVZ	188060	Anura	Alsodidae	*Alsodes*	*gargola*
I4373	MVZ	231914	Anura	Alytidae	*Alytes*	*obstetricans*
I12044	SR	QCAZA44783	Anura	Aromobatidae	*Allobates*	*insperatus*
I12045	SR	QCAZA56305	Anura	Aromobatidae	*Allobates*	*insperatus*
I7557	ESP	R1020	Anura	Arthroleptidae	*Arthroleptis*	*variablis*
I7559	ESP	R846	Anura	Arthroleptidae	*Arthroleptis*	*wahlbergi*
I6478	MCZ	A139626	Anura	Arthroleptidae	*Cardioglossa*	*leucomystax*
I4375	CAS	168499	Anura	Arthroleptidae	*Leptopelis*	*parkeri*
I6485	MCZ	A137988	Anura	Arthroleptidae	*Schoutedenella*	*sylvatica*
I13520	REF	AscMon	Anura	Ascaphidae	*Ascaphus*	*montanus*
I6477	MCZ	A136805	Anura	Astylosternidae	*Astylosternus*	*diadematus*
I7720	ESP	R306	Anura	Astylosternidae	*Leptopelis*	*vermiculatus*
I4442	AMCC	122836	Anura	Astylosternidae	*Leptydactylodon*	*bicolor*
I4376	AMCC	122837	Anura	Astylosternidae	*Nyctibates*	*corrugatus*
I6483	MCZ	A139709	Anura	Astylosternidae	*Scotobleps*	*gabonicus*
I6437	MCZ	A136806	Anura	Astylosternidae	*Trichobatrachus*	*robustus*
I4377	MVZ	164828	Anura	Batrachylidae	*Batrachyla*	*taeniata*
I8555	CAS	242112	Anura	Bombinatoridae	*Bombina*	*microdeladigitora*
I8556	CFBHT	55	Anura	Brachycephalidae	*Brachycephalus*	*ephippium*
I4380	USNM	533994	Anura	Brachycephalidae	*Ischnocnema*	*ramagii*
I4432	AMCC	105557	Anura	Brevicipitidae	*Breviceps*	*macrops*
I4382	MCZ	138534	Anura	Brevicipitidae	*Callulina*	*kisiwamsitu*
I6474	CAS	168560	Anura	Brevicipitidae	*Probreviceps*	*macrodactylus*
I6476	MCZ	A140276	Anura	Bufonidae	*Amietophrynus*	*camerunensis*
I4383	ECM	4908	Anura	Bufonidae	*Anaxyrus*	*terrestris*
I4429	YPM	13738	Anura	Bufonidae	*Ansonia*	*longidigita*
I4430	YPM	13728	Anura	Bufonidae	*Atelopus*	*hoogmoedi*
I4433	AMCC	105533	Anura	Bufonidae	*Capensibufo*	*rosei*
I6464	MVZ	239399	Anura	Bufonidae	*Leptophryne*	*borbonica*
I12500	PMH	2014	Anura	Bufonidae	*Melanophryniscus*	*stelzneri*
I6481	MCZ	A139634	Anura	Bufonidae	*Nectophryne*	*batesii*
I7581	ESP	R690	Anura	Bufonidae	*Poyntonophrynus*	*damaranus*
I6467	MVZ	231697	Anura	Bufonidae	*Rhamphophryne*	*macrorhina*
I8576	LSUMNS	15190	Anura	Bufonidae	*Rhinella*	*marinus*
I4384	PMH	CAL1	Anura	Calyptocephalellidae	*Calyptocephalella*	*gayi*
I8557	LSUMNS	16979	Anura	Centrolenidae	*Centrolene*	*prosoblepon*
I8558	LSUMNS	17409	Anura	Centrolenidae	*Cochranella*	*adenocheira*
I4386	AMCC	118359	Anura	Centrolenidae	*Hyalinobatrachium*	*fleischmanni*
I4431	AMCC	125449	Anura	Ceratobatrachidae	*Batrachylodes*	*vertebralis*
I6472	CPM	2014	Anura	Ceratobatrachidae	*Ceratobatrachus*	*guentheri*
I6418	AMCC	125415	Anura	Ceratobatrachidae	*Discodeles*	*bufoniformis*

Table 5.1. Taxon sampling (continued).

I4387	CAS	237845	Anura	Ceratobatrachidae	*Platymantis*	*pelewensis*
I4388	MVZ	247561	Anura	Ceratophryidae	*Ceratophrys*	*cornuta*
I4434	AMCC	125581	Anura	Ceratophryidae	*Chacophrys*	*pierottii*
I4441	YPM	13120	Anura	Ceratophryidae	*Lepidobatrachus*	*laevis*
I4195	SBH	268267	Anura	Ceuthomantidae	*Ceuthomantis*	*smaragdinus*
I4390	MVZ	253198	Anura	Conrauidae	*Conraua*	*crassipes*
I4391	USNM	534194	Anura	Craugastoridae	*Craugastor*	*noblei*
I4371	Cab	381	Anura	Cycloramphidae	*Cycloramphus*	*cavagua*
I8559	LSUMNS	16955	Anura	Dendrobatidae	*Colostethus*	*caeruleodactylus*
I6449	ITF	2014	Anura	Dendrobatidae	*Dendrobates*	*leucomelas*
I8563	LSUMNS	13667	Anura	Dendrobatidae	*Epipedobates*	*femoralis*
I4393	CAS	231821	Anura	Dendrobatidae	*Mannophryne*	*trinitatus*
I4446	YPM	13066	Anura	Dendrobatidae	*Phyllobates*	*vittatus*
I6424	AMCC	106520	Anura	Dicroglossidae	*Chaparana*	*delacouri*
I6450	CAS	243255	Anura	Dicroglossidae	*Euphlyctis*	*cyanophlyctis*
I6419	AMCC	144930	Anura	Dicroglossidae	*Fejervarya*	*limnocharis*
I6452	CAS	241469	Anura	Dicroglossidae	*Hoplobatrachus*	*rugulosus*
I7654	ESP	R059	Anura	Dicroglossidae	*Ingerana*	*sp_nov_2*
I4394	CAS	221360	Anura	Dicroglossidae	*Limnonectes*	*kuhlii*
I7668	ESP	R057	Anura	Dicroglossidae	*Limnonectes*	*limborgii*
I8219	ESP	R180	Anura	Dicroglossidae	*Nanorana*	*bourreti*
I6465	MVZ	231208	Anura	Dicroglossidae	*Nanorana*	*pleskei*
I4395	CAS	239527	Anura	Dicroglossidae	*Occidozyga*	*lima*
I6417	AMCC	144942	Anura	Dicroglossidae	*Quasipaa*	*verrucospinosa*
I4397	MVZ	235689	Anura	Discoglossidae	*Discoglossus*	*pictus*
I4398	EMO	1	Anura	Eleutherodactylidae	*Eleutherodactylus*	*coqui*
I8578	LSUMNS	21241	Anura	Eleutherodactylidae	*Syrrhophus*	*cystignathoides*
I4399	SANBI	1954	Anura	Heleophrynidae	*Heleophryne*	*purcelli*
I4400	BPN	1286	Anura	Hemiphractidae	*Stefania*	*evansi*
I8225	ESP	R012	Anura	Hemisotidae	*Hemisus*	*guineensis*
I4401	MVZ	249304	Anura	Hemisotidae	*Hemisus*	*marmoratus*
I8560	SR	CHUNB64717	Anura	Hylidae	*Corythomantis*	*greeningei*
I4157	SR	QCAZA48552	Anura	Hylidae	*Cruziohyla*	*calcarifer*
I6462	MVZ	264263	Anura	Hylidae	*Dendropsophus*	*microcephalus*
I4160	SR	QCAZA51852	Anura	Hylidae	*Hyloscirtus*	*palmeri*
I4439	YPM	10666	Anura	Hylidae	*Hypsiboas*	*crepitans*
I8568	PMH	Litoria2014	Anura	Hylidae	*Litoria*	*caerulea*
I8569	LSUMNS	9884	Anura	Hylidae	*Litoria*	*thesaurensis*
I4169	SR	QCAZA53552	Anura	Hylidae	*Nyctimantis*	*rugiceps*
I6482	MCZ	A148702	Anura	Hylidae	*Osteopilus*	*dominicensis*
I4158	SR	QCAZA48818	Anura	Hylidae	*Phyllomedusa*	*vaillantii*
I6431	AMCC	117944	Anura	Hylidae	*Plectrohyla*	*matudai*

Table 5.1. Taxon sampling (continued).

I13521	REF	PseFer	Anura	Hylidae	*Pseudacris*	*feriarum*
I13522	REF	PseNig	Anura	Hylidae	*Pseudacris*	*nigrita*
I6468	MVZ	257781	Anura	Hylidae	*Scinax*	*staufferi*
I4447	YPM	14191	Anura	Hylidae	*Smilisca*	*fodiens*
I6442	CAS	245062	Anura	Hylidae	*Sphaenorhynchus*	*lacteus*
I6471	MVZ	247548	Anura	Hylidae	*Trachycephalus*	*coriaceus*
I6427	AMCC	125603	Anura	Hylidae	*Triprion*	*petasatus*
I4411	KZ	1713	Anura	Hylodidae	*Hylodes*	*phyllodes*
I6475	MCZ	A139760	Anura	Hyperoliidae	*Afrixalus*	*fulvovittatus*
I4428	AMCC	125880	Anura	Hyperoliidae	*Alexteroon*	*obstetricans*
I7701	ESP	R1139	Anura	Hyperoliidae	*Cryptothylax*	*greshoffi*
I7704	ESP	R1129	Anura	Hyperoliidae	*Heterixalus*	*luteostriatus*
I4403	MCZ	136920	Anura	Hyperoliidae	*Hyperolius*	*guttulatus*
I7707	ESP	R843	Anura	Hyperoliidae	*Kassina*	*senegalensis*
I7708	ESP	R1195	Anura	Hyperoliidae	*Opisthothylax*	*immaculatus*
I6457	AMCC	124754	Anura	Hyperoliidae	*Phlyctimantis*	*leonardi*
I7713	ESP	R838	Anura	Hyperoliidae	*Semnodactylus*	*wealii*
I4448	DMG	5134	Anura	Leiopelmatidae	*Leiopelma*	*hochstetteri*
I4405	CAS	245125	Anura	Leptodactylidae	*Leptodactylus*	*fuscus*
I8567	LSUMNS	15432	Anura	Leptodactylidae	*Lithodytes*	*lineatus*
I6458	CAS	231794	Anura	Leptodactylidae	*Physalaemus*	*pustulosus*
I6441	MVZ	264270	Anura	Leptodactylidae	*Physalaemus*	*pustulosus*
I4406	MVZ	231766	Anura	Leptodactylidae	*Pleurodema*	*bibroni*
I6422	MVZ	238723	Anura	Mantellidae	*Aglyptodactylus*	*madagascariensis*
I8229	CJR/ESP	R928	Anura	Mantellidae	*Boophis*	*albipunctatus*
I4407	MVZ	238732	Anura	Mantellidae	*Boophis*	*pyrrhus*
I8233	CJR/ESP	R942	Anura	Mantellidae	*Gephyromantis*	*ambohitra*
I8244	CJR/ESP	R971	Anura	Mantellidae	*Guibemantis*	*pulcher*
I7730	CJR/ESP	R930	Anura	Mantellidae	*Mantella*	*betsileo*
I7734	CJR/ESP	R969	Anura	Mantellidae	*Mantidactylus*	*lugabris*
I4408	MVZ	226277	Anura	Megophryidae	*Brachytarsophrys*	*feae*
I6429	AMCC	106397	Anura	Megophryidae	*Leptobrachium*	*chapaense*
I6425	AMCC	106489	Anura	Megophryidae	*Leptolalax*	*bourreti*
I4409	CAS	240922	Anura	Megophryidae	*Megophrys*	*glandulosa*
I6416	AMCC	144796	Anura	Megophryidae	*Ophryophryne*	*hansi*
I6473	CAS	234295	Anura	Megophryidae	*Scutiger*	*gongshanensis*
I4410	CAS	220433	Anura	Micrixalidae	*Micrixalus*	*borealis*
I10391	ROM	44169	Anura	Microhylidae	*Adelastes*	*hylonomos*
I13334	PLVP	PT425	Anura	Microhylidae	*Albericus*	*exclamitans*
I13335	PLVP	PT321	Anura	Microhylidae	*Altigius*	*alios*
I13336	PLVP	PT359	Anura	Microhylidae	*Anodonthyla*	*nigrigularis*
I13337	PLVP	PT281	Anura	Microhylidae	*Arcovomer*	*sp*

Table 5.1. Taxon sampling (continued).

I13338	PLVP	PT439	Anura	Microhylidae	*Barygenys*	*nana*
I4419	CAS	236077	Anura	Microhylidae	*Calluella*	*guttulata*
I13339	PLVP	PT164	Anura	Microhylidae	*Calluella*	*yunnanensis*
I13340	PLVP	PT440	Anura	Microhylidae	*Callulops*	*personatus*
I10392	FMNH	231112	Anura	Microhylidae	*Chaperina*	*fusca*
I13341	PLVP	PT441	Anura	Microhylidae	*Choerophryne*	*proboscidea*
I13342	PLVP	PT448	Anura	Microhylidae	*Cophixalus*	*balbus*
I13343	PLVP	PT428	Anura	Microhylidae	*Copiula*	*oxyrhina*
I8562	LSUMNS	17434	Anura	Microhylidae	*Ctenophryne*	*geayi*
I13344	PLVP	PT332	Anura	Microhylidae	*Dasypops*	*schirchi*
I4435	AMCC	125588	Anura	Microhylidae	*Dermatonotus*	*muelleri*
I6463	MVZ	238744	Anura	Microhylidae	*Dyscophus*	*guineti*
I13345	PLVP	PT059	Anura	Microhylidae	*Elachistocleis*	*helianneae*
I13523	REF	GasCar	Anura	Microhylidae	*Gastrophryne*	*carolinensis*
I13346	PLVP	PT452	Anura	Microhylidae	*Genyophryne*	*thomsoni*
I6451	CAS	234799	Anura	Microhylidae	*Glyphoglossus*	*molossus*
I13347	PLVP	PT043	Anura	Microhylidae	*Hamptophryne*	*boliviana*
I13348	PLVP	PT424	Anura	Microhylidae	*Hylophorbus*	*rainerguntheri*
I13349	PLVP	PT284	Anura	Microhylidae	*Hyophryne*	*histrio*
I13350	PLVP	PT168	Anura	Microhylidae	*Kalophrynus*	*interlineatus1*
I8566	CAS	247917	Anura	Microhylidae	*Kalophrynus*	*pleurostigma*
I4440	YPM	13065	Anura	Microhylidae	*Kaloula*	*pulchra*
I13351	PLVP	PT507	Anura	Microhylidae	*Metamagnusia*	*slateri*
I13352	PLVP	PT236	Anura	Microhylidae	*Metaphrynella*	*sundana*
I6454	CAS	233947	Anura	Microhylidae	*Microhyla*	*ornata*
I6455	CAS	247906	Anura	Microhylidae	*Micryletta*	*inornata*
I13353	PLVP	PT340	Anura	Microhylidae	*Myersiella*	*sp*
I10393	ABTC	50092	Anura	Microhylidae	*Oreophryne*	*brachypus*
I13354	PLVP	PT459	Anura	Microhylidae	*Otophryne*	*robusta*
I13355	PLVP	PT455	Anura	Microhylidae	*Oxydactyla*	*alpestris*
I7739	ESP	R1330	Anura	Microhylidae	*Phrynomantis*	*annectens*
I13356	PLVP	PT287	Anura	Microhylidae	*Phrynomantis*	*bifasciatus*
I6436	AMCC	103335	Anura	Microhylidae	*Platypelis*	*occultans*
I6435	AMCC	128714	Anura	Microhylidae	*Plethodontohyla*	*notosticta*
I7740	ESP	R1208	Anura	Microhylidae	*Ramanella*	*variegata*
I13357	PLVP	PT312	Anura	Microhylidae	*Scaphiophryne*	*brevis*
I13358	PLVP	PT273	Anura	Microhylidae	*Stereocyclops*	*incrassatus*
I6430	AMCC	103414	Anura	Microhylidae	*Stumpffia*	*grandis*
I13359	PLVP	PT265	Anura	Microhylidae	*Stumpffia*	*roseifemoralis*
I13360	PLVP	PT271	Anura	Microhylidae	*Synapturanus*	*salseri2*
I13361	PLVP	PT198	Anura	Microhylidae	*Syncope*	*carvalhoi*
I13362	PLVP	PT454	Anura	Microhylidae	*Xenobatrachus*	*fuscigula*

Table 5.1. Taxon sampling (continued).

I8561	MV	18153	Anura	Myobatrachidae	*Crinia*	*signifera*
I8564	MV	21476	Anura	Myobatrachidae	*Geocrinia*	*victoriana*
I6486	SAMAR	66870	Anura	Myobatrachidae	*Lymnodynastes*	*dumerilli*
I13524	REF	MixSch	Anura	Myobatrachidae	*Mixophyes*	*schevilli*
I8570	MV	21528	Anura	Myobatrachidae	*Neobatrachus*	*sudelli*
I9034	JSK/SCD	70661	Anura	Myobatrachidae	*Notaden*	*nichollsi*
I8571	MV	21479	Anura	Myobatrachidae	*Paracrinia*	*haswelli*
I10935	NCBS	AI442	Anura	Nasikabatrachidae	*Nasikabatrachus*	*sahyadrensis*
I10934	NCBS	AG004	Anura	Nyctibatrachidae	*Nyctibatrachus*	*petraeus*
I4415	CAS	230053	Anura	Odontobatrachidae	*Odontobatrachus*	*natator*
I4412	MVZ	145208	Anura	Odontophrynidae	*Odontophrynus*	*occidentalis*
I4413	MVZ	234650	Anura	Pelobatidae	*Pelobates*	*syriacus*
I4414	MVZ	186009	Anura	Pelodytidae	*Pelodytes*	*ibericus*
I6439	MCZ	A139541	Anura	Petropedetidae	*Petropedetes*	*parkeri*
I4416	CAS	218893	Anura	Phrynobatrachidae	*Phrynobatrachus*	*leveleve*
I6443	MCZ	A136791	Anura	Phrynobatrachidae	*Phrynodon*	*sandersoni*
I6453	PMH	2014	Anura	Pipidae	*Hymenochirus*	*boettgeri*
I6444	MVZ	247511	Anura	Pipidae	*Pipa*	*pipa*
I8572	LSUMNS	12511	Anura	Pseudidae	*Pseudis*	*paradoxa*
I7783	ESP	R1068	Anura	Ptychadenidae	*Ptychadena*	*mascareniensis*
I4418	CAS	219251	Anura	Ptychadenidae	*Ptychadena*	*newtoni*
I6438	AMCC	105559	Anura	Pyxicephalidae	*Arthroleptella*	*bicolor*
I6428	AMCC	106956	Anura	Pyxicephalidae	*Arthroleptides*	*martiensseni*
I7794	ESP	R527	Anura	Pyxicephalidae	*Aubria*	*subsigillata*
I8191	ESP	R371	Anura	Pyxicephalidae	*Cacosternum*	*albiventer*
I6461	MVZ	226261	Anura	Pyxicephalidae	*Cacosternum*	*boettgeri*
I8199	ESP	R363	Anura	Pyxicephalidae	*Cacosternum*	*platys*
I8205	ESP	R569	Anura	Pyxicephalidae	*Natalobatrachus*	*bonebergi*
I7801	ESP	R725	Anura	Pyxicephalidae	*Pyxicephalus*	*adspersus*
I6433	AMCC	105565	Anura	Pyxicephalidae	*Strongylopus*	*bonaespei*
I7822	ESP	R831	Anura	Pyxicephalidae	*Strongylopus*	*fasciatus*
I7827	ESP	R410	Anura	Pyxicephalidae	*Tomopterna*	*cryptotis*
I6446	CAS	242607	Anura	Ranidae	*Amolops*	*medogensis*
I7849	ESP	R185	Anura	Ranidae	*Babina*	*chapaensis*
I6420	AMCC	138323	Anura	Ranidae	*Huia*	*nasica*
I7868	ESP	R1144	Anura	Ranidae	*Sylvirana*	*nigrovittata*
I6480	YPM	13741	Anura	Ranidae	*Hylarana*	*picturata*
I8573	LSUMNS	255	Anura	Ranidae	*Limnonectes*	*limnocharis*
I8574	LSUMNS	17589	Anura	Ranidae	*Rana*	*palmipes*
I7875	ESP	R1141	Anura	Ranidae	*Rana*	*pipiens*
I13526	REF	LitSph	Anura	Ranidae	*Rana*	*sphenocephala*
I7882	ESP	R1162	Anura	Ranidae	*Meristogenys*	*orphnocnemis*

Table 5.1. Taxon sampling (continued).

I6466	MVZ	258265	Anura	Ranidae	*Odorrana*	*banaorum*
I8575	LSUMNS	10459	Anura	Ranidae	*Papurana*	*papua*
I7897	ESP	R153	Anura	Ranidae	*Pelophylax*	*ridibunda*
I6460	CAS	234711	Anura	Ranidae	*Pterorana*	*khare*
I10411	CAS	202097	Anura	Ranidae	*Amietia*	*cf_tenuiplicata*
I7908	ESP	R1168	Anura	Ranidae	*Sanguirana*	*sanguinea*
I7910	ESP	R1164	Anura	Ranidae	*Staurois*	*natator*
I7915	ESP	R107	Anura	Ranixalidae	*Indirana*	*leithi*
I7917	ESP	R1145	Anura	Micrixalidae	*Micrixalus*	*sp*
I6440	MVZ	241442	Anura	Rhacophoridae	*Buergeria*	*oxycephalus*
I7925	ESP	R538	Anura	Rhacophoridae	*Chiromantis*	*xerampelina*
I7927	ESP	R233	Anura	Rhacophoridae	*Feihyla*	*palpebralis*
I7929	ESP	R1149	Anura	Rhacophoridae	*Gorhixalus*	*hosii*
I7935	ESP	R1112	Anura	Rhacophoridae	*Kurixalus*	*appendiculatus*
I7946	ESP	R075	Anura	Rhacophoridae	*Nyctixalus*	*pictus*
I6456	CAS	233160	Anura	Rhacophoridae	*Philautus*	*parvulus*
I6459	CAS	241141	Anura	Rhacophoridae	*Polypedates*	*leucomystax*
I7961	ESP/CJR	R1120	Anura	Rhacophoridae	*Raorchestes*	*gryllus*
I7967	ESP	R241	Anura	Rhacophoridae	*Rhacophorus*	*pardalis*
I4421	CAS	224676	Anura	Rhacophoridae	*Rhacophorus*	*rhodopus*
I6470	MVZ	225131	Anura	Rhacophoridae	*Theloderma*	*corticale*
I4422	MVZ	164829	Anura	Rhinodermatidae	*Rhinoderma*	*darwinii*
I4423	MVZ	164756	Anura	Rhinophrynidae	*Rhinophrynus*	*dorsalis*
I4424	CAS	229217	Anura	Scaphiopodidae	*Scaphiopus*	*couchii*
I6469	MVZ	145187	Anura	Scaphiopodidae	*Spea*	*hammondii*
I9326	CR04	Labisko	Anura	Sooglossidae	*Sooglossus*	*sechellensis*
I4445	AMCC	107352	Anura	Strabomantidae	*Phrynopus*	*sp*
I4426	USNM	268942	Anura	Strabomantidae	*Pristimantis*	*ridens*
I4427	KU	290640	Anura	Telamatobiidae	*Telmatobius*	*niger*
I4349	DWW	1781	Caudata	Ambystomatidae	*Ambystoma*	*mexicanum*
I3541	RB01	OP4	Caudata	Ambystomatidae	*Ambystoma*	*opacum*
I3538	RB09	T23	Caudata	Ambystomatidae	*Ambystoma*	*talpoidium*
I3544	JK02	Tig	Caudata	Ambystomatidae	*Ambystoma*	*tigrinum*
I4351	MVZ	232868a	Caudata	Amphiumidae	*Amphiuma*	*tridactylum*
I3702	PMH	AD03	Caudata	Cryptobranchidae	*Andrias*	*davidianus*
I3703	PMH	AJ12	Caudata	Cryptobranchidae	*Andrias*	*japonicus*
I3709	PMH	BC16	Caudata	Cryptobranchidae	*Cryptobranchus*	*alleganiensis*
I3704	PMH	C37AF	Caudata	Cryptobranchidae	*Cryptobranchus*	*alleganiensis*
I3707	PMH	ELK13	Caudata	Cryptobranchidae	*Cryptobranchus*	*alleganiensis*
I4356	DWW	2567	Caudata	Dicamptodontidae	*Dicamptodon*	*copei*
I3700	DWW	379	Caudata	Hynobiidae	*Batrachuperus*	*persicus*
I3542	YPM	9865	Caudata	Hynobiidae	*Hynobius*	*nigrescens*

Table 5.1. Taxon sampling (continued).

I3539	YPM	10577	Caudata	Hynobiidae	*Pachyhynobius*	*shangchengensis*
I3701	DWW	392	Caudata	Hynobiidae	*Salamandrella*	*keyserlingii*
I3710	RLM/MVZ	CSU01	Caudata	Plethodontidae	*Aneides*	*flavipunctatus*
I11148	ELJ	1554	Caudata	Plethodontidae	*Batrachoseps*	*nigriventris*
I4358	AMCC	118113	Caudata	Plethodontidae	*Bolitoglossa*	*riletti*
I3715	JDK	JK03	Caudata	Plethodontidae	*Desmognathus*	*fuscus*
I3716	JDK	JK08	Caudata	Plethodontidae	*Desmognathus*	*quadromaculatus*
I3711	JDK	JK07	Caudata	Plethodontidae	*Desmognathus*	*wrighti*
I9327	JRJ	2012	Caudata	Plethodontidae	*Eurycea*	*lucifuga*
I9336	PMH	15APR2013	Caudata	Plethodontidae	*Gyrinophilus*	*porphyriticus*
I12499	TPierson	TPierson3	Caudata	Plethodontidae	*Hemidactylium*	*scutatum*
I12498	MVZ	247157	Caudata	Plethodontidae	*Karsenia*	*koreana*
I12496	MVZ	263972	Caudata	Plethodontidae	*Nyctanolis*	*pernix*
I4359	JJA	P82	Caudata	Plethodontidae	*Phaeognathus*	*hubrichti*
I3717	JDK	JK09	Caudata	Plethodontidae	*Plethodon*	*jordani*
I12497	PMH	PR02	Caudata	Plethodontidae	*Pseudotriton*	*ruber*
I3535	PMH	7759	Caudata	Proteidae	*Necturus*	*maculosus*
I4362	MVZ	244076	Caudata	Proteidae	*Proteus*	*anguinus*
I3536	LSUMNS	H11333	Caudata	Rhyacotritonidae	*Rhyacotriton*	*olympicus*
I9330	TP	TP24749	Caudata	Salamandridae	*Cynops*	*ensicauda*
I9337	TP	TP26195	Caudata	Salamandridae	*Echinotriton*	*chinhaiensis*
I9338	TP	TP27066	Caudata	Salamandridae	*Neurergus*	*crocatus*
I3534	LSUMNS	H11856	Caudata	Salamandridae	*Notophthalmus*	*viridescens*
I9339	TP	TP24839	Caudata	Salamandridae	*Paramesotriton*	*hongkongensis*
I9331	TP	TP25088	Caudata	Salamandridae	*Salamandra*	*salamandra*
I9332	TP	s7539	Caudata	Salamandridae	*Salamandrina*	*terdigitata*
I9340	TP	TP26609	Caudata	Salamandridae	*Triturus*	*vulgaris*
I9333	TP	TP25555	Caudata	Salamandridae	*Tylototriton*	*kweichowensis*
I13533	REF	SirInt	Caudata	Sirenidae	*Siren*	*intermedia*
I4337	BPN	1499	Gymnophiona	Caeciliidae	*Caecilia*	*tentaculata*
I6479	SLZ	971026	Gymnophiona	Dermophiidae	*Dermophis*	*mexicanus*
I4436	YPM	13118	Gymnophiona	Dermophiidae	*Geotrypetes*	*seraphini*
I4338	MVZ	228795	Gymnophiona	Dermophiidae	*Gymnopis*	*multiplicata*
I4339	CAS	218738	Gymnophiona	Dermophiidae	*Schistometopum*	*thomense*
I4340	MVZ	179505	Gymnophiona	Herpelidae	*Boulengerula*	*taitana*
I4437	YPM	13116	Gymnophiona	Herpelidae	*Herpele*	*squalostoma*
I13518	REF	IchBan	Gymnophiona	Ichthyophiidae	*Ichthyophis*	*bannanicus*
I4342	MVZ	258024	Gymnophiona	Indotyphlidae	*Grandisonia*	*alternans*
I4343	MVZ	265495	Gymnophiona	Rhinatrematidae	*Epicrionops*	*petersi*
I4345	AMCC	117706	Gymnophiona	Scolecomorphidae	*Crotaphatrema*	*tchabalmbaboensis*
I8577	CAS	168812	Gymnophiona	Scolecomorphidae	*Scolecomorphus*	*vittatum*
I4346	BPN	Ga169	Gymnophiona	Siphonopidae	*Microcaecilia*	*sp*

Table 5.1. Taxon sampling (continued).

I4347	MVZ	162592	Gymnophiona	Siphonopidae	*Siphonops*	*annulatus*
I4348	MVZ	179733	Gymnophiona	Typhlonectidae	*Typhlonectes*	*natans*
N/A	GENBANK	TAXID_28377			*Anolis*	*carolinensis*
N/A	GENBANK	TAXID_8478			*Chrysemys*	*picta*
N/A	GENBANK	TAXID_9031			*Gallus*	*gallus*
N/A	GENBANK	TAXID_9606			*Homo*	*sapiens*
N/A	GENBANK	TAXID_7897			*Latimeria*	*chalumnae*
N/A	GENBANK	TAXID_8364			*Xenopus*	*tropicalis*

Museum, specimen, and individual acronyms are: ABTC, Australian Biological Tissue Collection; AMCC, Ambrose Monell Cryogenic Collection, The American Museum of Natural History; BPN, Brice P. Noonan; CAS, California Academy of Sciences; CFBHT, Kelly Zamudio; CJR, Christopher J. Raxworthy; CPM, Christopher P. McNamara; CR04, Jim Labisko; Cab, Kelly Zamudio; DMG, David M. Green; DWW, David W. Weisrock; ECM, Emily C. Moriarty-Lemmon; ELJ, Elizabeth L. Jockusch; EMO, Eric M. O'Neill; ESP, Elizabeth Scott-Prendini; FMNH, Florida Museum of Natural History; ITF, I. Tyler Frye; JDK, Justin D. Kratovil; JJA, J. J. Apodaca; JRJ, Jarrett R. Johnson; JSK, J. Scott Keogh; KU, University of Kansas Museum of Natural History; KZ, Kelly Zamudio; LSUMNS, Louisiana State University Museum of Natural Science; MCZ, Museum of Comparative Zoology; MV, Museum Victoria; MVZ, Museum of Vertebrate Zoology; NCBS, National Centre for Biological Sciences, India; PLVP, Pedro L. V. Peloso; PMH, Paul M. Hime; RB, Schyler Nunziata; RLM/MVZ, Rachel L. Mueller; ROM, Royal Ontario Museum; SAMAR, South Australian Museum; SANBI, South African National Biodiversity Institute; SBH, S. Blair Hedges; SLZ, St. Louis Zoo; SR, Santiago Ron; TP, Ted Pappenfus; TPierson, Todd Pierson; USNM, Smithsonian National Museum of Natural History; YPM, Yale Peabody Museum.

Table 5.2. Details of 220 nuclear loci.

Locus ID	Taxa	Characters	Codons	Concatenated Alignment Start	Concatenated Alignment End	Missing Taxa	Details
1	297	1494	498	1	1494	4	
4	283	1503	501	1495	2997	18	
5	257	1440	480	2998	4437	44	No Salamanders
10	260	759	253	4438	5196	41	
11	258	1023	341	5197	6219	43	
13	295	846	282	6220	7065	6	
14	239	1674	558	7066	8739	62	No Salamanders
15	286	1356	452	8740	10095	15	
16	287	1110	370	10096	11205	14	
17	222	1332	444	11206	12537	79	No Salamanders or Caecilians
20	276	1440	480	12538	13977	25	
28	289	1182	394	13978	15159	12	
30	283	804	268	15160	15963	18	
31	293	1608	536	15964	17571	8	
34	296	858	286	17572	18429	5	
35	230	765	255	18430	19194	71	No Salamanders or Caecilians
36	282	1314	438	19195	20508	19	
38	279	1110	370	20509	21618	22	
41	296	1656	552	21619	23274	5	
45	276	1593	531	23275	24867	25	
46	291	1290	430	24868	26157	10	
47	232	1410	470	26158	27567	69	No Salamanders or Caecilians
48	230	1416	472	27568	28983	71	No Salamanders or Caecilians
49	285	1425	475	28984	30408	16	
53	226	1392	464	30409	31800	75	No Salamanders or Caecilians
54	265	1455	485	31801	33255	36	
55	240	1371	457	33256	34626	61	No Salamanders
56	294	1668	556	34627	36294	7	
57	277	1104	368	36295	37398	24	
59	299	1254	418	37399	38652	2	
61	293	1587	529	38653	40239	8	
62	287	1263	421	40240	41502	14	
63	223	519	173	41503	42021	78	No Salamanders or Caecilians
65	297	1557	519	42022	43578	4	

Table 5.2. Details of 220 nuclear loci (continued).

69	227	588	196	43579	44166	74	
78	296	1284	428	44167	45450	5	
80	273	633	211	45451	46083	28	
82	286	1155	385	46084	47238	15	
86	295	1263	421	47239	48501	6	
88	251	552	184	48502	49053	50	
92	293	1341	447	49054	50394	8	
93	294	1203	401	50395	51597	7	
95	292	1707	569	51598	53304	9	
97	291	1569	523	53305	54873	10	
99	279	1362	454	54874	56235	22	
100	294	1419	473	56236	57654	7	
102	264	1047	349	57655	58701	37	
105	275	1053	351	58702	59754	26	
107	281	564	188	59755	60318	20	
109	228	951	317	60319	61269	73	No Salamanders
110	282	1506	502	61270	62775	19	
112	296	1047	349	62776	63822	5	
113	269	732	244	63823	64554	32	
115	296	1737	579	64555	66291	5	
116	274	1536	512	66292	67827	27	
118	281	1182	394	67828	69009	20	
121	297	1518	506	69010	70527	4	
122	292	1548	516	70528	72075	9	
123	287	2058	686	72076	74133	14	
124	279	903	301	74134	75036	22	
125	296	1782	594	75037	76818	5	
126	288	1437	479	76819	78255	13	
127	275	1842	614	78256	80097	26	
130	288	1683	561	80098	81780	13	
132	262	804	268	81781	82584	39	
135	260	1092	364	82585	83676	41	
136	296	1485	495	83677	85161	5	
137	298	1287	429	85162	86448	3	
138	291	1710	570	86449	88158	10	
141	274	912	304	88159	89070	27	
144	283	1113	371	89071	90183	18	
146	294	1401	467	90184	91584	7	
147	226	993	331	91585	92577	75	
149	275	885	295	92578	93462	26	
151	290	1557	519	93463	95019	11	
152	248	1377	459	95020	96396	53	No Caecilians

Table 5.2. Details of 220 nuclear loci (continued).

153	273	1395	465	96397	97791	28	
154	296	1464	488	97792	99255	5	
155	284	1635	545	99256	100890	17	
156	268	1194	398	100891	102084	33	
159	278	1137	379	102085	103221	23	
160	272	1392	464	103222	104613	29	
161	269	741	247	104614	105354	32	
162	285	1413	471	105355	106767	16	
163	295	318	106	106768	107085	6	
164	294	1644	548	107086	108729	7	
165	286	1365	455	108730	110094	15	
166	283	1575	525	110095	111669	18	
169	293	930	310	111670	112599	8	
172	263	1599	533	112600	114198	38	
173	254	1533	511	114199	115731	47	
174	272	1137	379	115732	116868	29	
175	292	810	270	116869	117678	9	
177	255	1521	507	117679	119199	46	No Salamanders
179	278	1704	568	119200	120903	23	
182	296	1722	574	120904	122625	5	
183	277	1098	366	122626	123723	24	
184	290	1311	437	123724	125034	11	
187	266	960	320	125035	125994	35	
191	299	1428	476	125995	127422	2	
192	299	1662	554	127423	129084	2	
193	288	1614	538	129085	130698	13	
194	285	1710	570	130699	132408	16	
196	288	1749	583	132409	134157	13	
197	261	1596	532	134158	135753	40	
198	292	1629	543	135754	137382	9	
199	252	1158	386	137383	138540	49	No Salamanders
200	267	1278	426	138541	139818	34	
201	298	1581	527	139819	141399	3	
202	298	1761	587	141400	143160	3	
203	287	1020	340	143161	144180	14	
204	249	1158	386	144181	145338	52	
208	280	1467	489	145339	146805	21	
209	291	1314	438	146806	148119	10	
210	291	1584	528	148120	149703	10	
211	292	1545	515	149704	151248	9	
212	285	1089	363	151249	152337	16	
214	262	1374	458	152338	153711	39	

Table 5.2. Details of 220 nuclear loci (continued).

216	220	1485	495	153712	155196	81	No Salamanders or Caecilians
217	237	768	256	155197	155964	64	No Salamanders
218	291	921	307	155965	156885	10	
219	296	1686	562	156886	158571	5	
220	285	825	275	158572	159396	16	
222	226	1581	527	159397	160977	75	No Salamanders or Caecilians
224	287	1593	531	160978	162570	14	
225	280	1884	628	162571	164454	21	
226	275	1986	662	164455	166440	26	
227	254	1467	489	166441	167907	47	No Salamanders
229	237	1197	399	167908	169104	64	No Salamanders or Caecilians
230	262	819	273	169105	169923	39	
231	233	1194	398	169924	171117	68	No Salamanders or Caecilians
234	278	1983	661	171118	173100	23	
239	295	1626	542	173101	174726	6	
240	292	1773	591	174727	176499	9	
241	277	1260	420	176500	177759	24	
242	289	855	285	177760	178614	12	
243	285	1050	350	178615	179664	16	
244	268	738	246	179665	180402	33	
245	288	1806	602	180403	182208	13	
246	296	1623	541	182209	183831	5	
248	294	1107	369	183832	184938	7	
249	209	399	133	184939	185337	92	No Salamanders or Caecilians
251	284	1284	428	185338	186621	17	
252	267	1653	551	186622	188274	34	
253	282	957	319	188275	189231	19	
254	244	1128	376	189232	190359	57	
255	290	1434	478	190360	191793	11	
258	286	1647	549	191794	193440	15	
262	217	1563	521	193441	195003	84	No Salamanders or Caecilians
264	287	1098	366	195004	196101	14	
265	268	780	260	196102	196881	33	
267	269	1026	342	196882	197907	32	
268	298	1500	500	197908	199407	3	
269	285	1542	514	199408	200949	16	

Table 5.2. Details of 220 nuclear loci (continued).

271	251	999	333	200950	201948	50	
272	285	1185	395	201949	203133	16	
274	268	840	280	203134	203973	33	
275	260	1119	373	203974	205092	41	
278	287	1560	520	205093	206652	14	
279	271	1485	495	206653	208137	30	
280	300	1119	373	208138	209256	1	
281	284	1383	461	209257	210639	17	
282	281	1629	543	210640	212268	20	
284	286	1611	537	212269	213879	15	
285	259	1761	587	213880	215640	42	
287	241	876	292	215641	216516	60	No Salamanders
288	288	1647	549	216517	218163	13	
290	243	1347	449	218164	219510	58	No Salamanders
291	287	1560	520	219511	221070	14	
293	262	705	235	221071	221775	39	
294	293	1677	559	221776	223452	8	
296	292	1785	595	223453	225237	9	
297	280	948	316	225238	226185	21	
299	274	1413	471	226186	227598	27	
304	287	1839	613	227599	229437	14	
305	216	426	142	229438	229863	85	No Salamanders or Caecilians
306	294	1602	534	229864	231465	7	
307	278	717	239	231466	232182	23	
309	253	585	195	232183	232767	48	
310	285	1503	501	232768	234270	16	
311	291	1536	512	234271	235806	10	
312	296	1689	563	235807	237495	5	
317	280	1275	425	237496	238770	21	
320	261	1560	520	238771	240330	40	
321	299	1572	524	240331	241902	2	
324	265	1110	370	241903	243012	36	
325	297	1572	524	243013	244584	4	
327	277	1848	616	244585	246432	24	
328	282	1653	551	246433	248085	19	
329	288	1761	587	248086	249846	13	
331	300	1602	534	249847	251448	1	
334	284	1224	408	251449	252672	17	
335	299	1452	484	252673	254124	2	
336	230	894	298	254125	255018	71	No Salamanders or Caecilians

Table 5.2. Details of 220 nuclear loci (continued).

337	279	1062	354	255019	256080	22	
339	294	1938	646	256081	258018	7	
340	274	897	299	258019	258915	27	
343	294	1197	399	258916	260112	7	
345	274	969	323	260113	261081	27	
346	291	1701	567	261082	262782	10	
347	296	1653	551	262783	264435	5	
348	227	1056	352	264436	265491	74	
349	284	1773	591	265492	267264	17	
350	293	1419	473	267265	268683	8	
353	294	1323	441	268684	270006	7	
354	268	1458	486	270007	271464	33	
355	294	1902	634	271465	273366	7	
358	294	1524	508	273367	274890	7	
359	277	1110	370	274891	276000	24	
360	287	1362	454	276001	277362	14	
362	235	1092	364	277363	278454	66	No Salamanders or Caecilians
367	284	960	320	278455	279414	17	
368	296	1533	511	279415	280947	5	
369	275	1101	367	280948	282048	26	
371	257	1491	497	282049	283539	44	
372	282	1674	558	283540	285213	19	
375	279	1629	543	285214	286842	22	
376	298	1767	589	286843	288609	3	
378	278	1482	494	288610	290091	23	
379	278	1830	610	290092	291921	23	

Table 5.3. Fossil calibrations for divergence time analyses.

Calibration	**Node**	**Fossils**	**Minimum (MYA)**	**Maximum (MYA)**	**Source (from Feng *et al*. 2017, except for *)**
1	Osteichthyes	*Guiyu oneiros*	420.7	444.9	* Benton *et al*. (2015); Zhu *et al*. (2009)
2	Tetrapoda	*Lethiscus stocki*	337.0	351.0	Benton *et al*. (2015)
3	Amniota	*Hylonomus lyelli*	318.0	332.9	Benton *et al*. (2015)
4	Diapsida	*Protorosaurus*	255.9	295.9	* Benton *et al*. (2015)
5	Lissamphibia	*Gerobatrachus hottoni*	270.6	337.0	Anderson *et al*. (2008); Anderson (2008)
6	Batrachia	*Triadobatrachus massinoti*	252.0	272.8	Cannatella (2015); Benton *et al*. (2015)
7	Caudata	*Iridotriton hechti*	146.8	252.0	Evans *et al*. (2005)
8	Gymnophiona	*Apodops pricei*	56.0	252.0	* Estes and Wake (1972)
9	Anura	*Liaobatrachus zhaoi*	129.7	252.0	Chang *et al*. (2009)
10	Amphiumidae + Plethodontidae	*Proamphiuma cretacea*	65.5	148.1	* Gardner (2003)
11	*Ambystoma* + *Dicamptodon*	*Dicamptodon antiquus*	55.8	148.1	* Naylor and Fox (1993)
12	Proteidae	*Necturus krausei*	56.8	148.1	* Naylor (1978)
13	Cryptobranchoidea	*Chunerpeton tianyiensis*	161.2	252.0	Gao and Shubin (2003)
14	Alytoidea	*Iberobatrachus angelae*	125.0	252.0	Gomez *et al*. (2016)
15	Pipanura	*Rhadinosteus parvus*	148.1	252.0	Cannatella (2015)
16	Pipoidea	*Neusibatrachus wilferti*	127.2	252.0	Gomez *et al*. (2016)
17	Pipidae	*Pachycentra taqueti*	83.6	148.1	Cannatella (2015)
18	Pelobatoidea	*Elkobatrachus brocki*	46.1	148.1	Henrici and Haynes (2006)
19	Pelodytes + (Pelobatidae + Megophryidae)	*Miopelodytes gilmorei*	38.9	148.1	Henrici and Haynes (2006)
20	Pelobatidae + Megophryidae	*Macropelobates osborni*	28.1	148.1	Cohen *et al*. (2013)
21	Acosmanura	*Eurycephalella alcinae*	113.0	252.0	Baez (2009)
22	Neobatrachia	*Beelzebufo ampinga*	66.0	148.1	Rogers *et al*. (2013)
23	Myobatrachoidea	*Calyptocephalella pichileufensis*	47.5	148.1	Gomez *et al*. (2011)
24	Ranoidea	*Thamastosaurus gezei*	33.9	148.1	Rage and Roček (2007)
25	Ptychadenidae + Phrynobatrachidae	Ptychadenidae fossil	25.0	148.1	Blackburn *et al*. (2015)

Figure 5.1. The 15 possible models for relationships among extant amphibian orders. Frogs, salamanders, and caecilians are either monophyletic (models 1-3) or non-monophyletic (models 4-15) with respect to amniotes. Tips are labeled as: Anura = frogs, Cauda = salamanders, Gymno = caecilians, Amniota = amniotes. *Latimeria*, the coelacanth, is assumed to be the sister taxon to (Amphibia + Amniota).

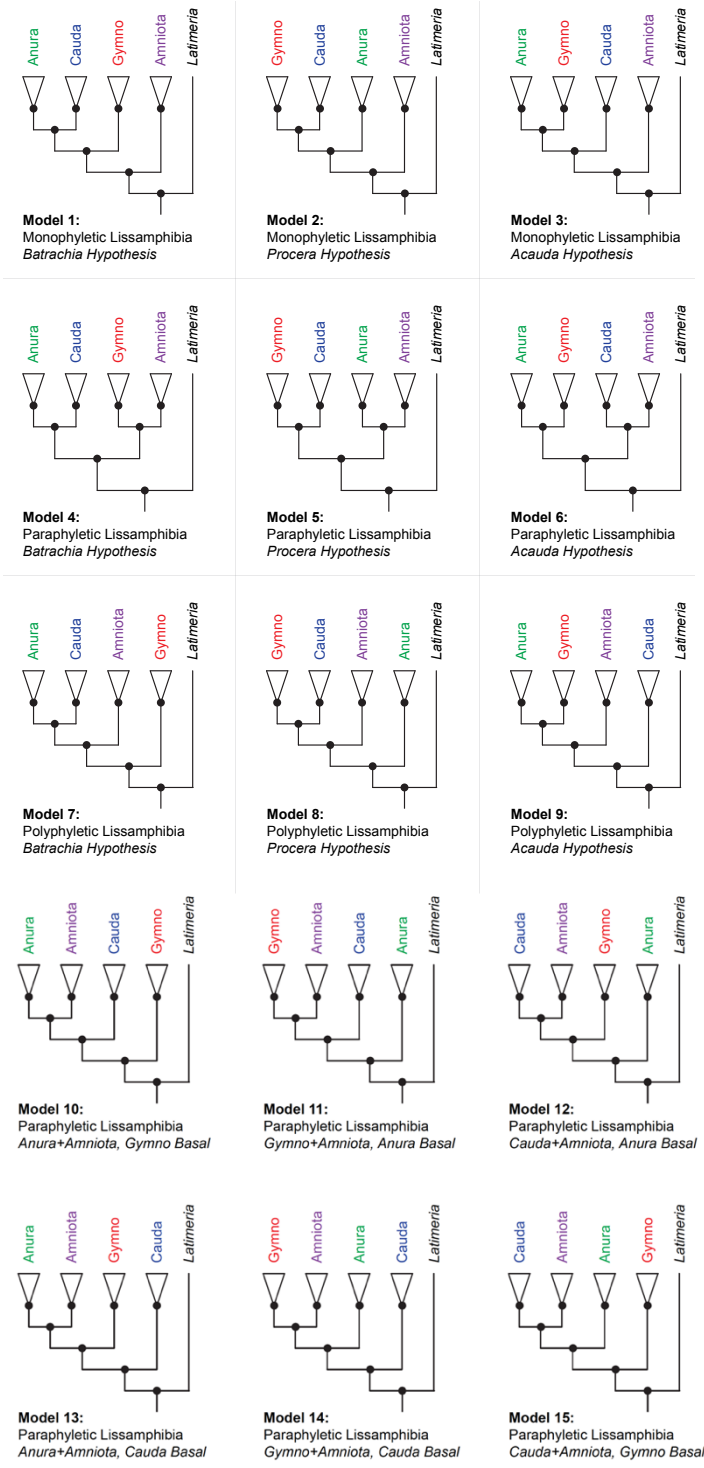


Figure 5.2. Backbone amphibian phylogeny depicting the fossil calibration points in Table 5.3.

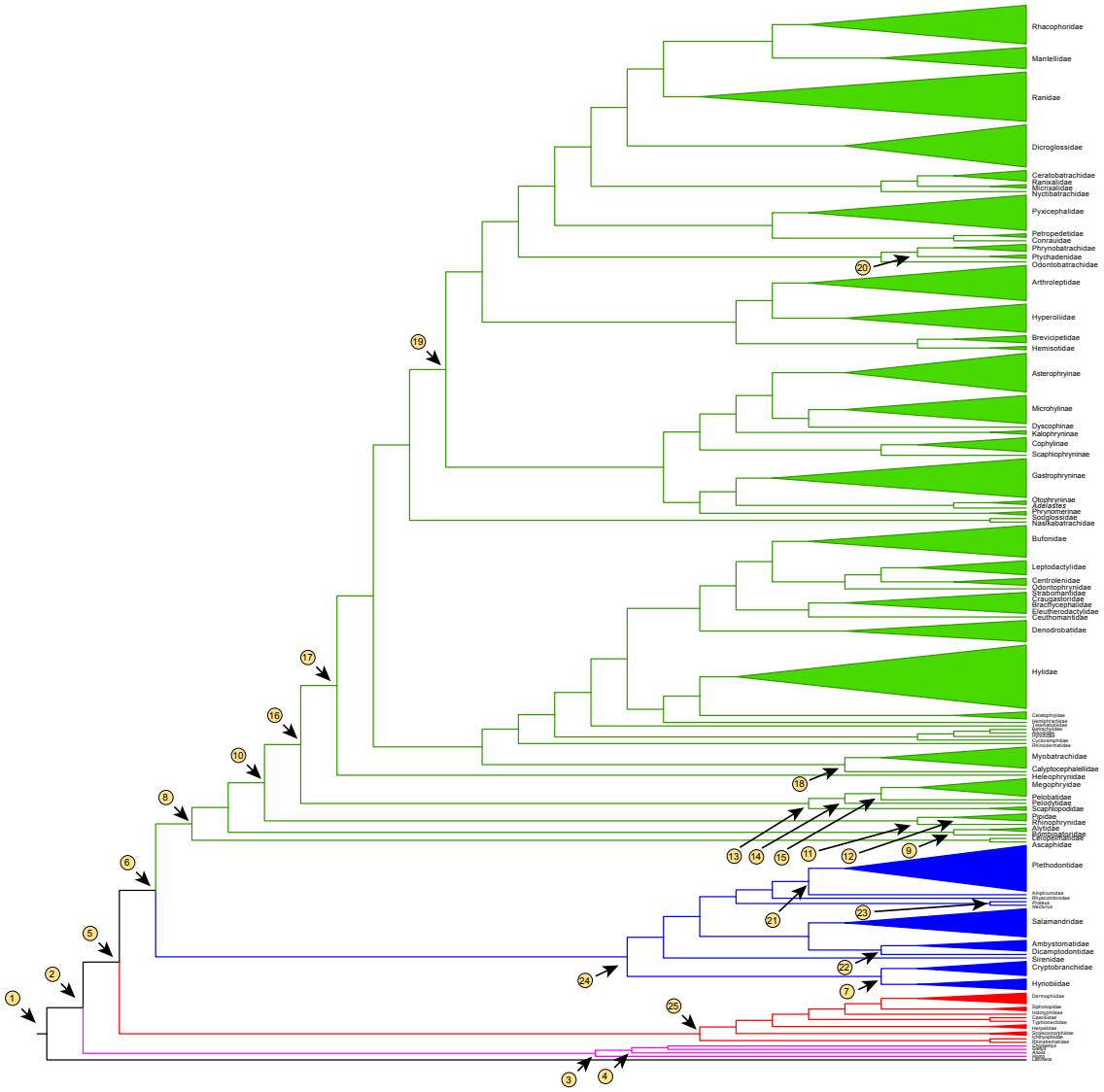
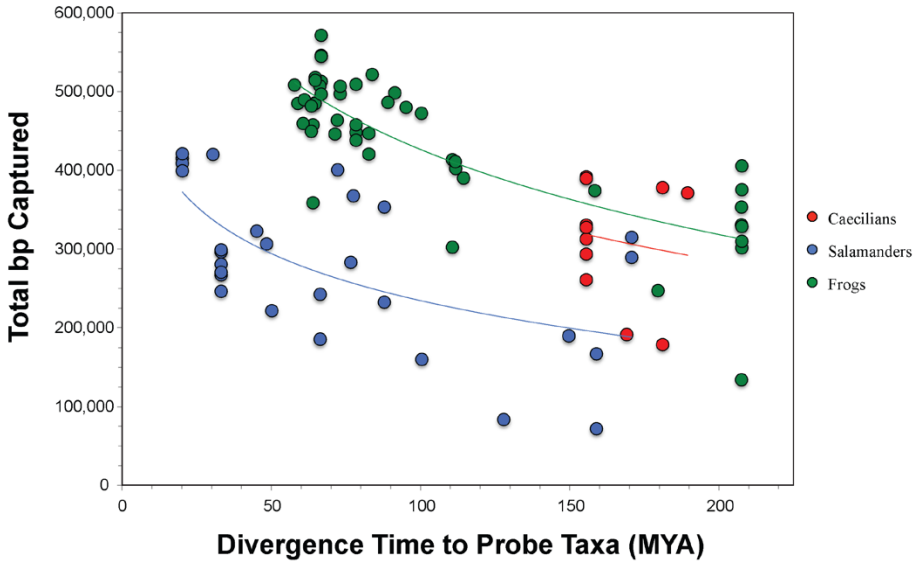


Figure 5.3. Correlates of targeted sequence enrichment and capture across amphibians.

A. Capture works better over more recent divergences.



B) Capture works better in small genomes.

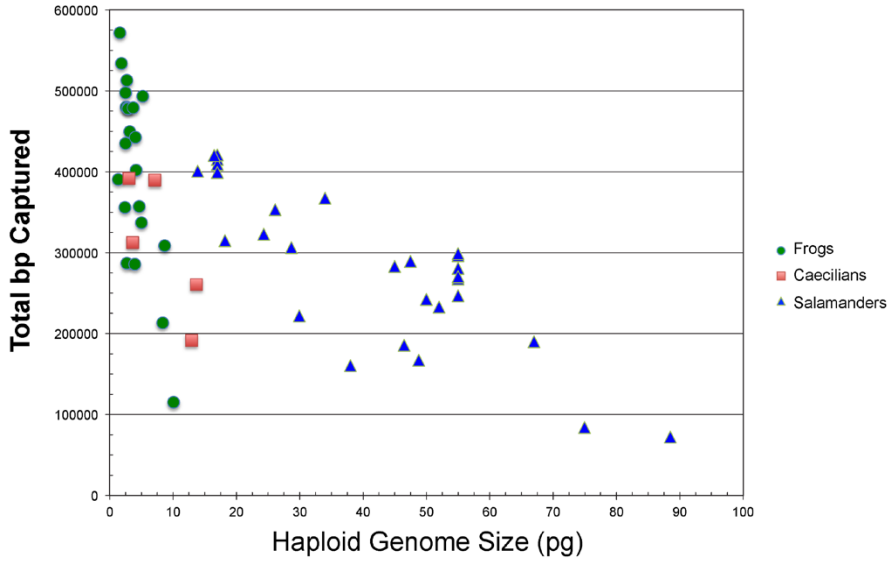


Figure 5.4. Patterns of missing loci and missing sites across 301 individuals. Points represent individuals and frogs, salamanders, caecilians, amniotes, and *Latimeria* are color coded in green, blue, red, magenta, and black, respectively.

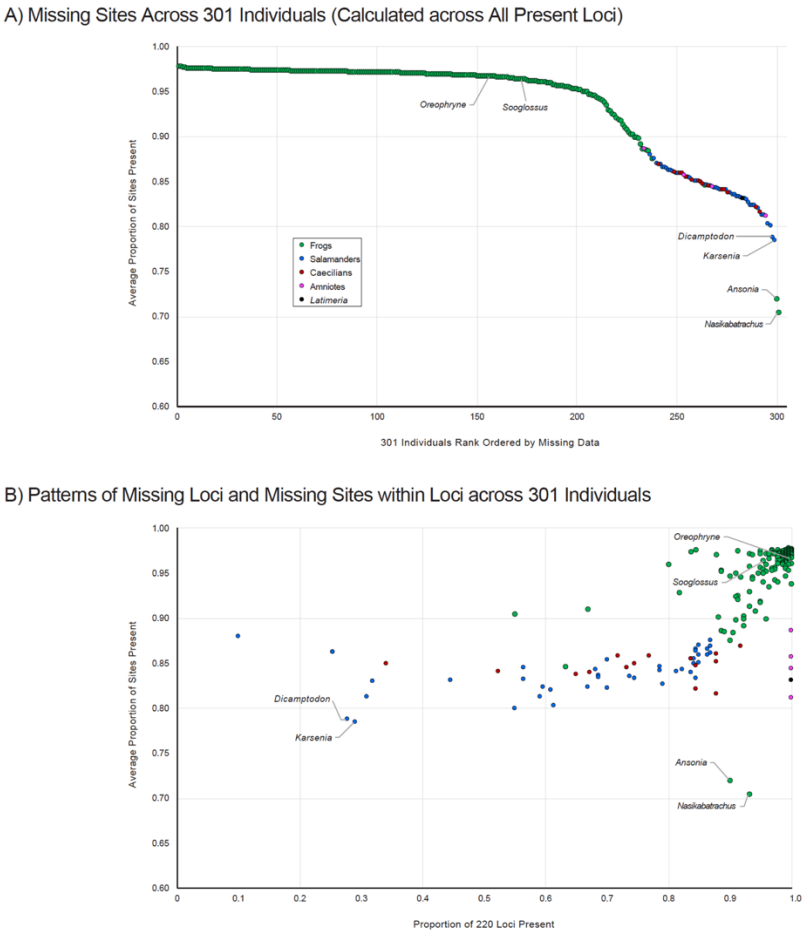


Figure 5.5. Distributions of Robinson-Foulds distances for species tree bootstraps (black), species trees versus gene trees (red), and gene trees versus gene trees (blue). Although there is substantial discordance between gene trees and between gene trees and the species tree, Astral still arrives at a relatively concordant set of species tree topologies.

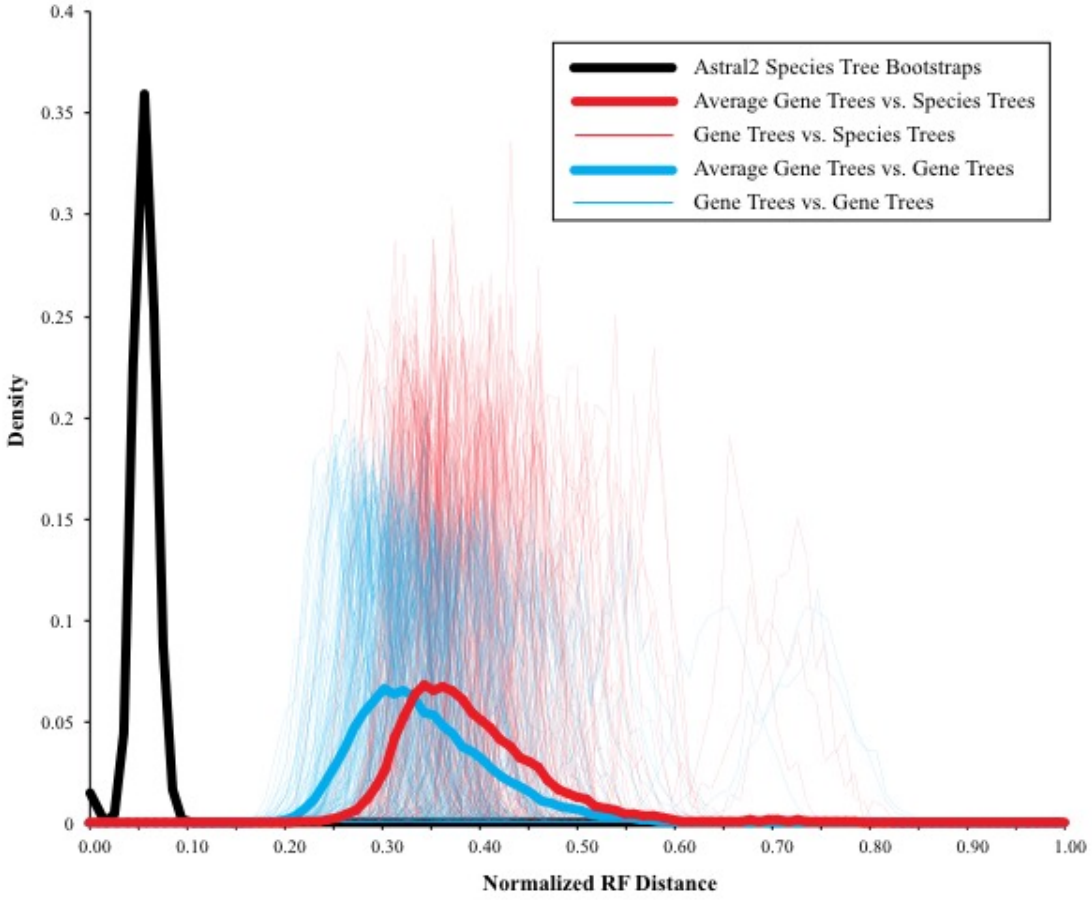


Figure 5.6. Backbone ASTRAL topology of major family-level amphibian lineages for the alignments which had not been filtered for loci with high proportions of missing sites. The orange box highlights a set of branches deep in the frog phylogeny (Neobatrachia) which are recovered differentially between ASTRAL and RAxML analyses, and which differ from previous phylogenetic studies of amphibian relationships. RF distance between ASTRAL and RAxML trees is 24/596 (0.04). This topology is recovered when using alignments not filtered for missing sites.

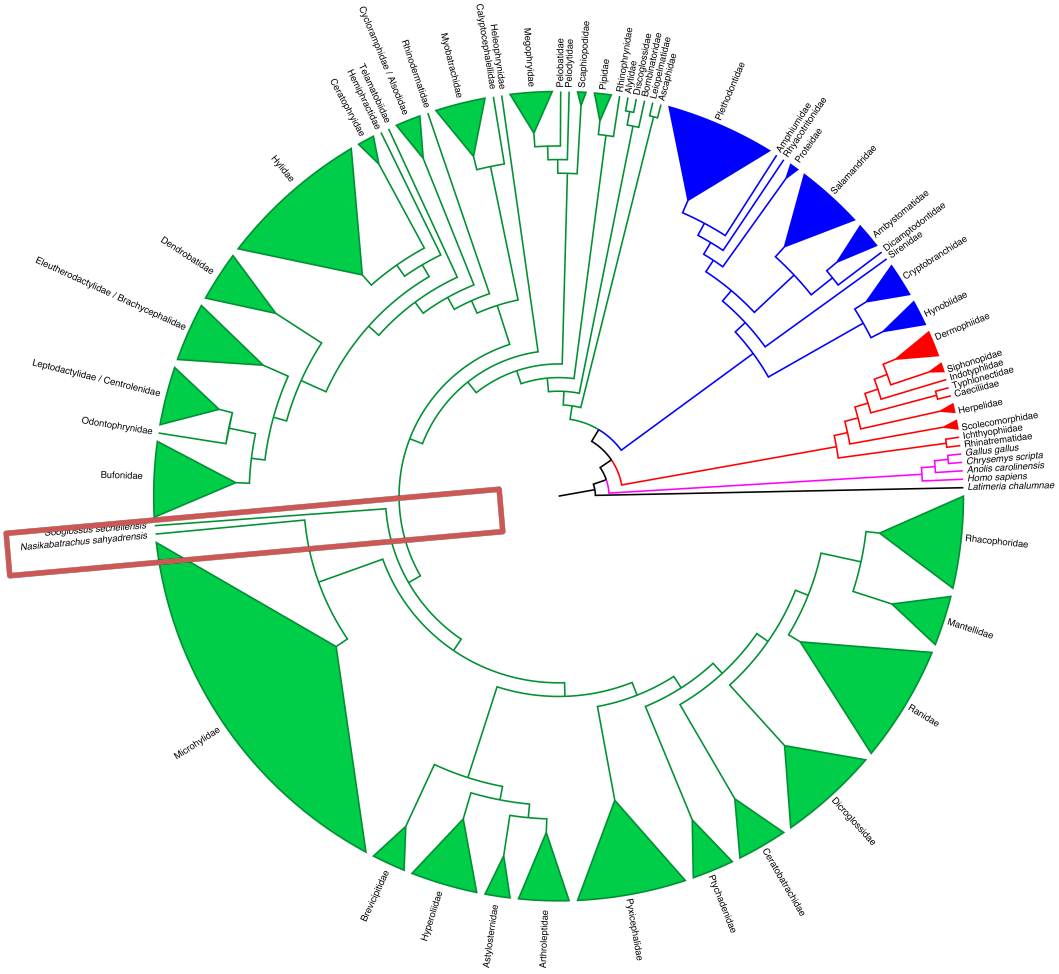


Figure 5.7. Astral topology from the alignments filtered for greater than 50% present sites.

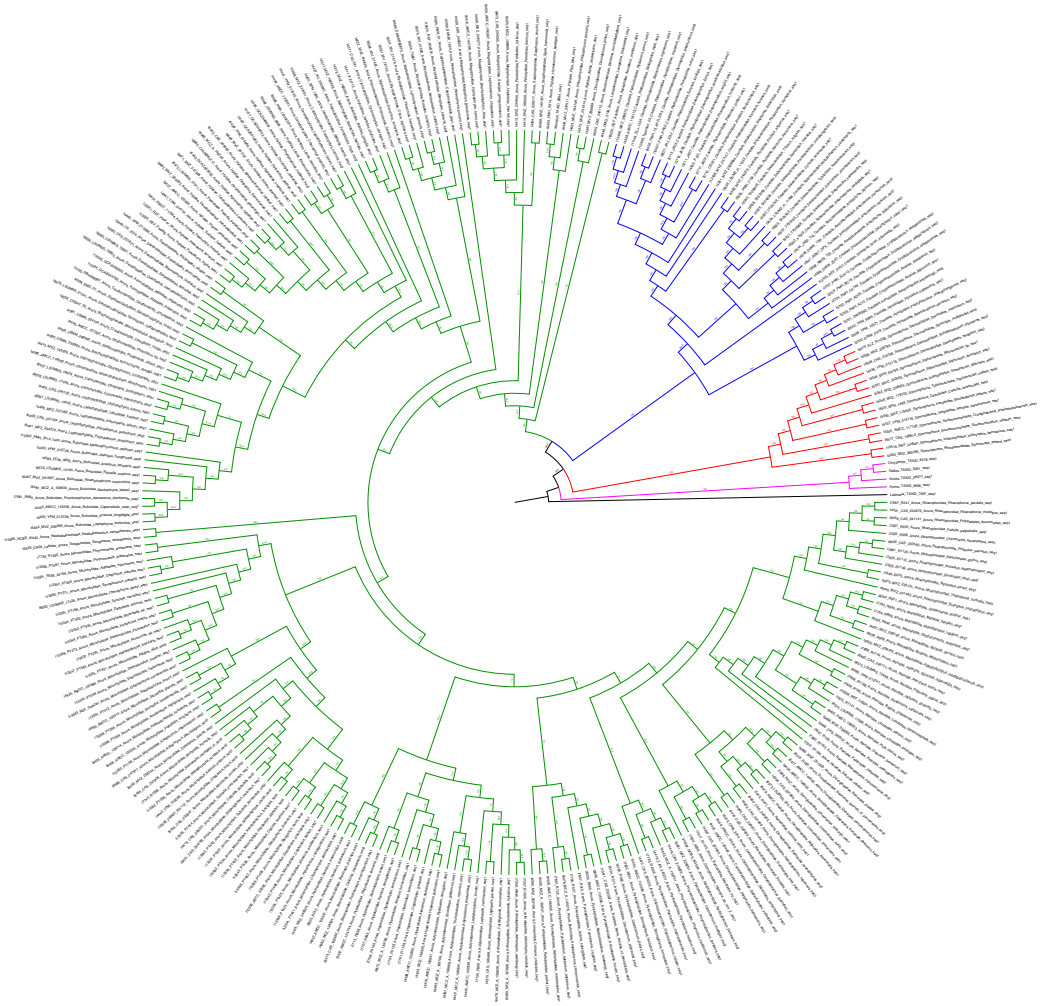


Figure 5.8. AIC-based approach to quantify the magnitude and direction of support for inter-ordinal amphibian relationships, assuming a monophyletic Amphibia. The bottom depicts the proportion of bootstrap replicates supporting each of the three possible topologies (color coded as noted) along the vertical axis for 194 genes binned by which model is supported overall along the horizontal axis. ΔAIC in the top plot measures the magnitude of support against rejected models.

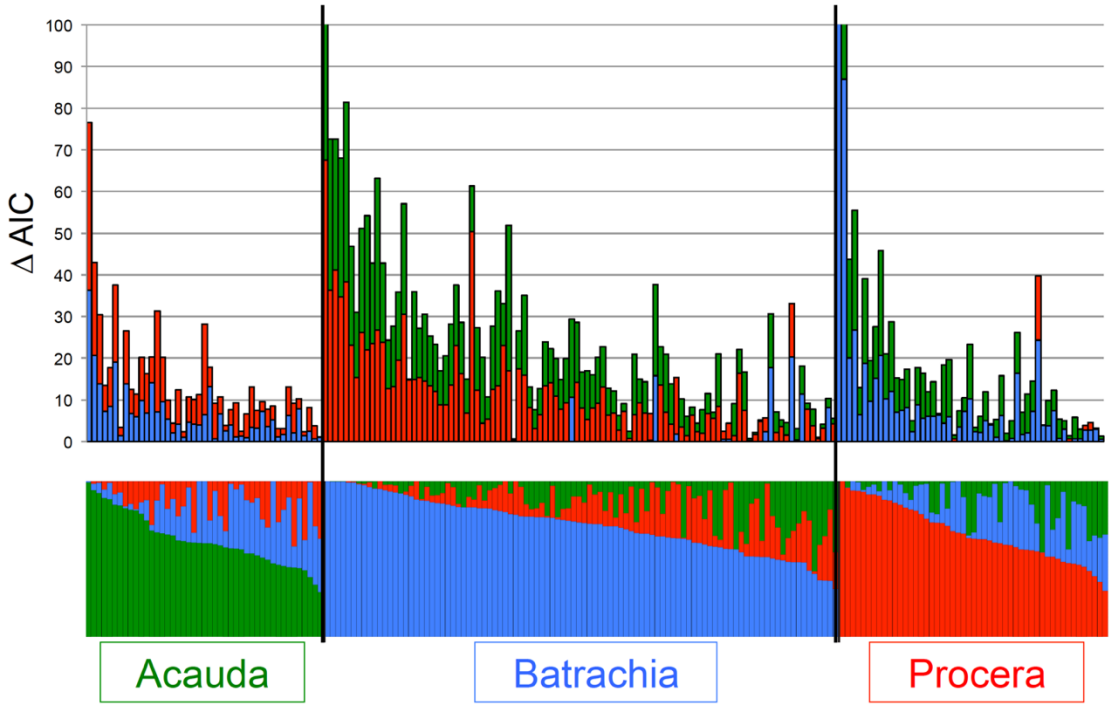
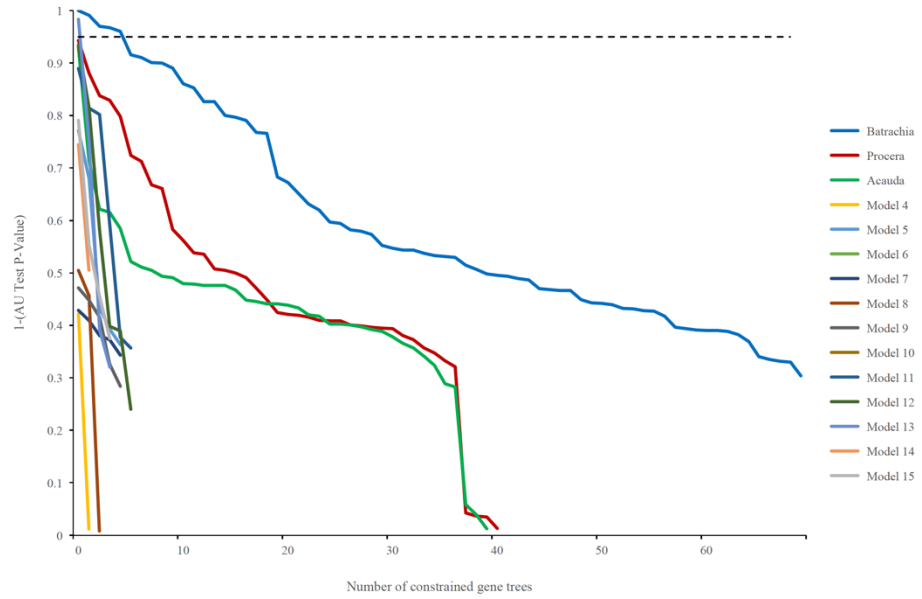


Figure 5.9. Gene genealogy interrogation (GGI) of constrained gene tree topologies for the 15 possible topologies relating frogs, salamanders, caecilians, and amniotes. Approximately unbiased tests of topology were conducted for each gene using the set of 15 best RAxML gene trees from constrained ML searches for all 15 possible topologies. Groups of loci supporting each competing topology are plotted by rank-ordered AU test P-values. The dashed line represents the 0.05 significance threshold for the approximately unbiased (AU) test. In the upper panel, it is clear that most genes support one of the three monophyletic Amphibia models, although a small number of genes support each of the alternative non-monophyletic models. In the lower panel, genes supporting the twelve non-monophyletic models are binned together for clarity.



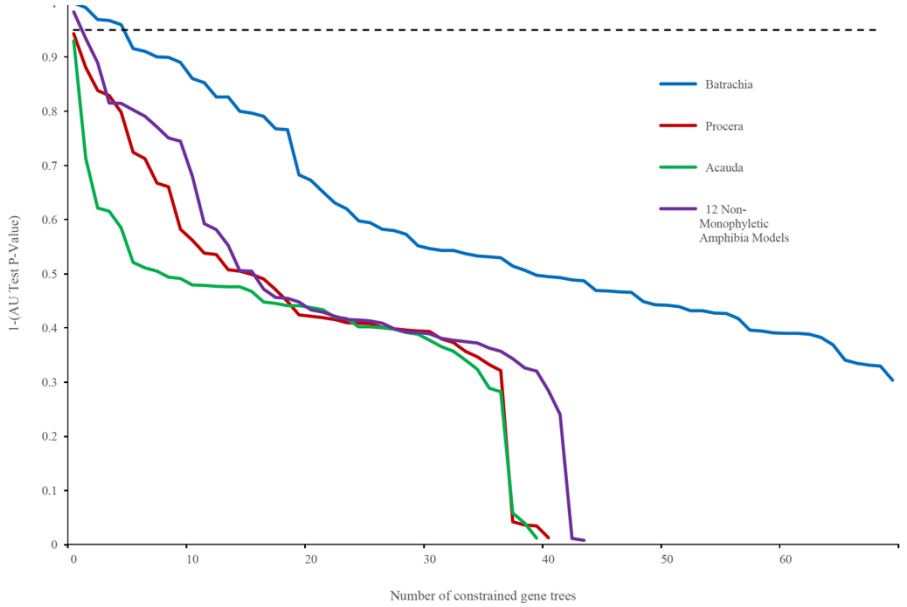


Figure 5.10. Conflicting neobatrachian relationships are inferred (and strongly supported) by different tree reconstruction methods. Additionally, both the ASTRAL and RAxML trees support a novel placement of Afrobatrachia as sister to Ranoidea (traditionally Afrobatrachia is found as sister to Microhylidae). RAxML places *Nasikabatrachus* sister to *Sooglossus* (the canonical placement), in contrast to ASTRAL. Branch labels are nonparametric bootstrap percentages over 500 replicates.

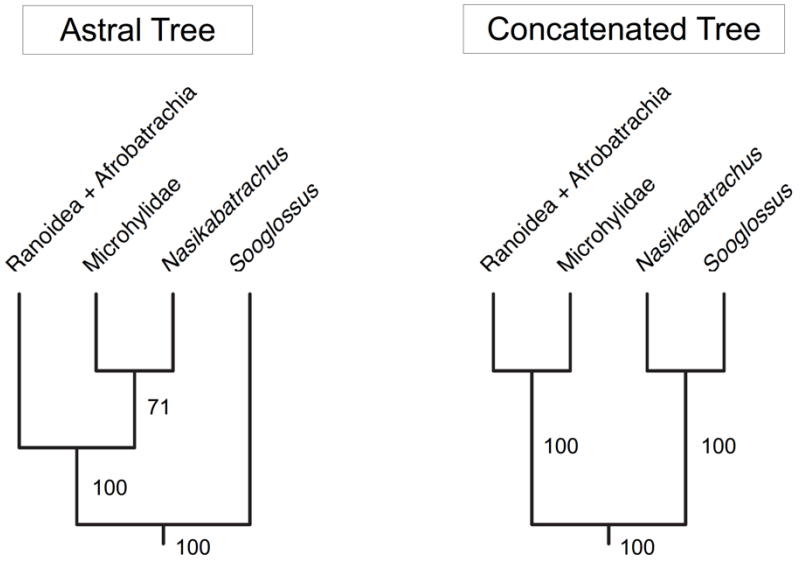


Figure 5.11. Short, gappy sequences for *Nasikabatrachus* drive gene tree support for (*Nasikabatrachus* + *Oreophryne*). The lengths of ungapped *Nasikabatrachus* sequences are plotted against total alignment lengths for 194 loci. Points (loci) are color coded by which alternative placement of *Nasikabatrachus* is supported. Overall, 94 loci support the canonical (*Nasikabatrachus* + *Sooglossus*) arrangement, while 98 loci support (*Nasikabatrachus* + *Oreophryne*), and two loci support some other topology. Most loci supporting (*Nasikabatrachus* + *Oreophryne*) have significantly more missing data (sites) for *Nasikabatrachus* than do the loci supporting (*Nasikabatrachus* + *Sooglossus*). Ungapped locus lengths for *Oreophryne* and *Sooglossus* are very close to the overall alignment lengths for nearly all loci.

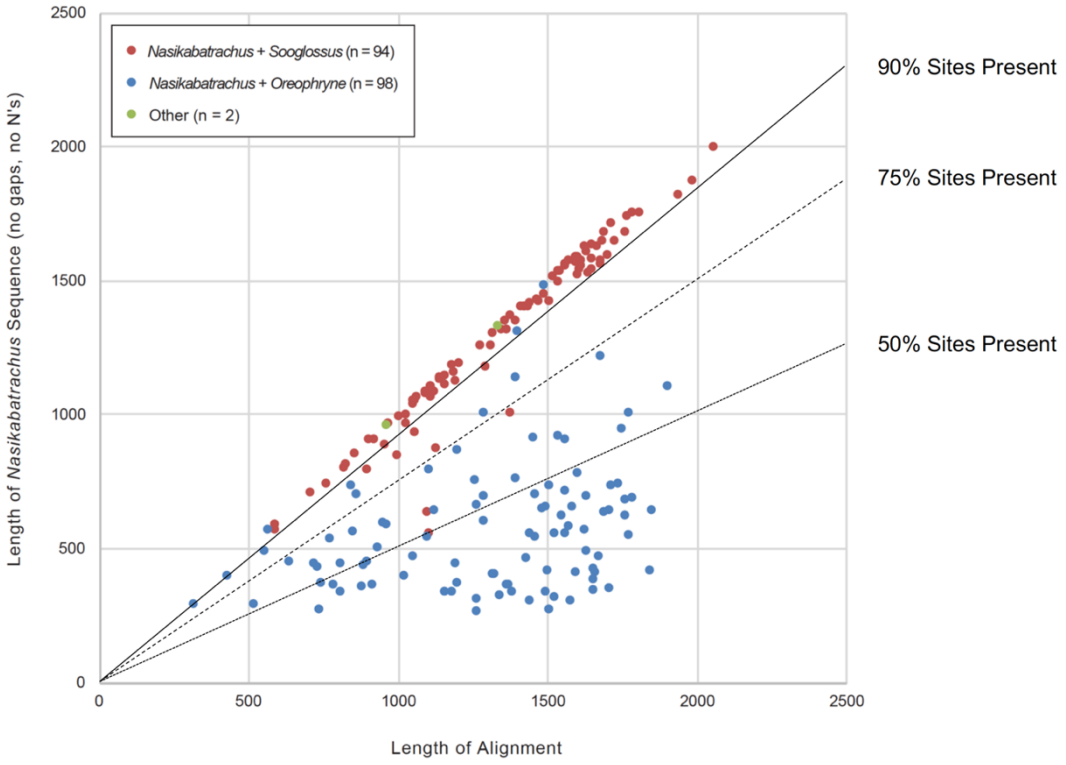


Figure 5.12. Unfiltered Astral tree.

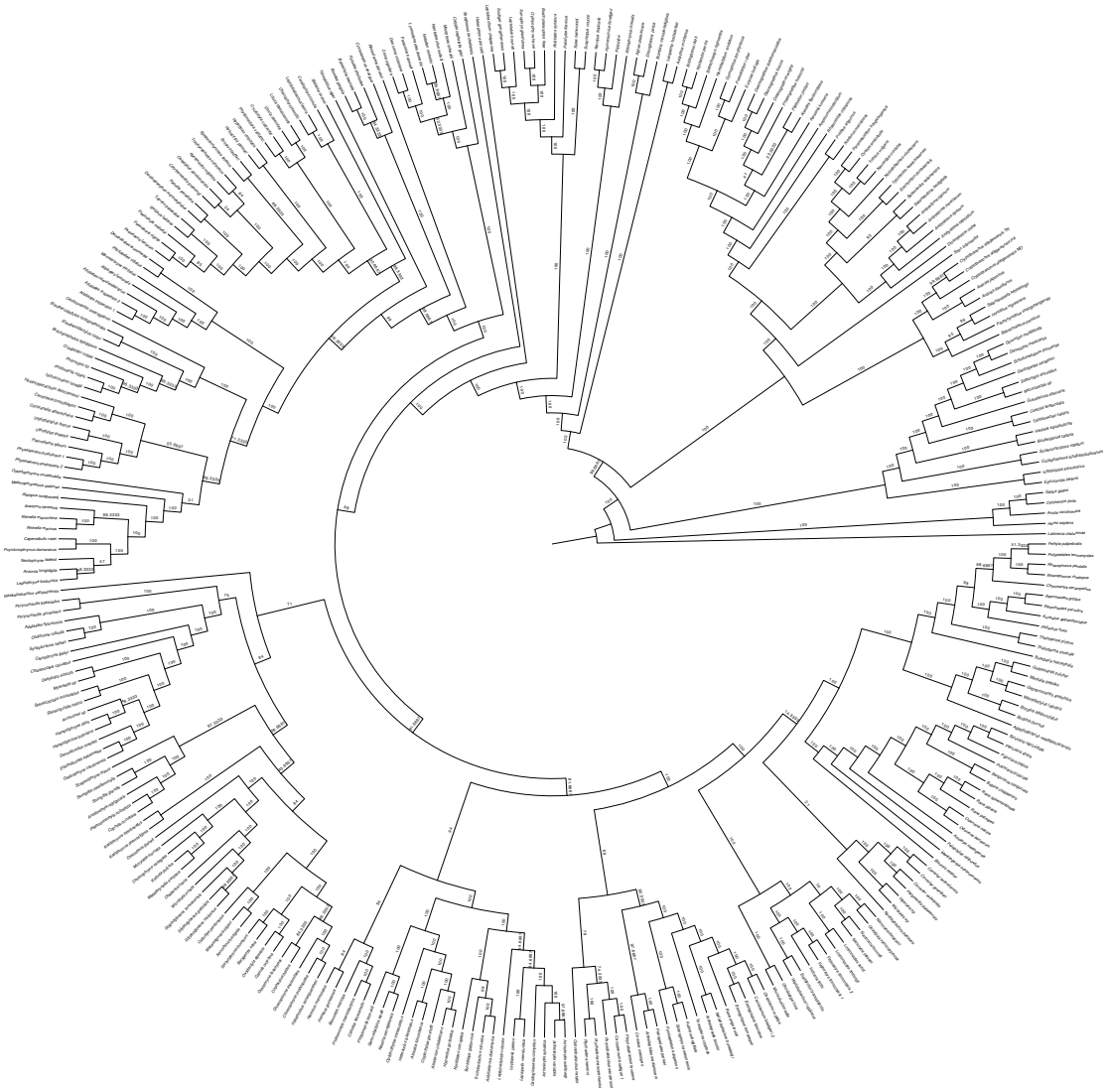


Figure 5.13. Unfiltered MulRF tree.

Figure 5.14. Unfiltered RAxML tree.

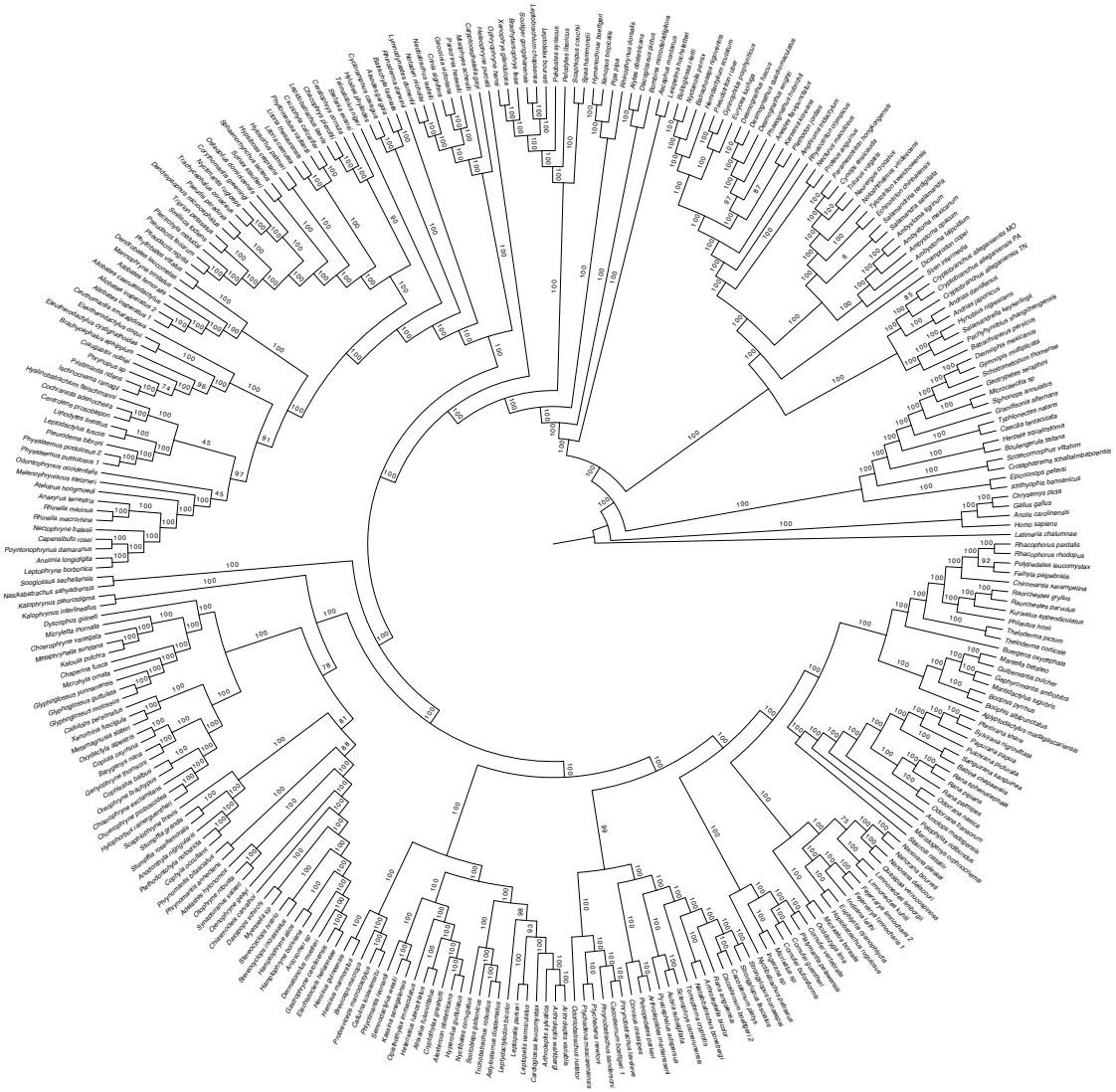


Figure 5.15. 90% present sites Astral tree.

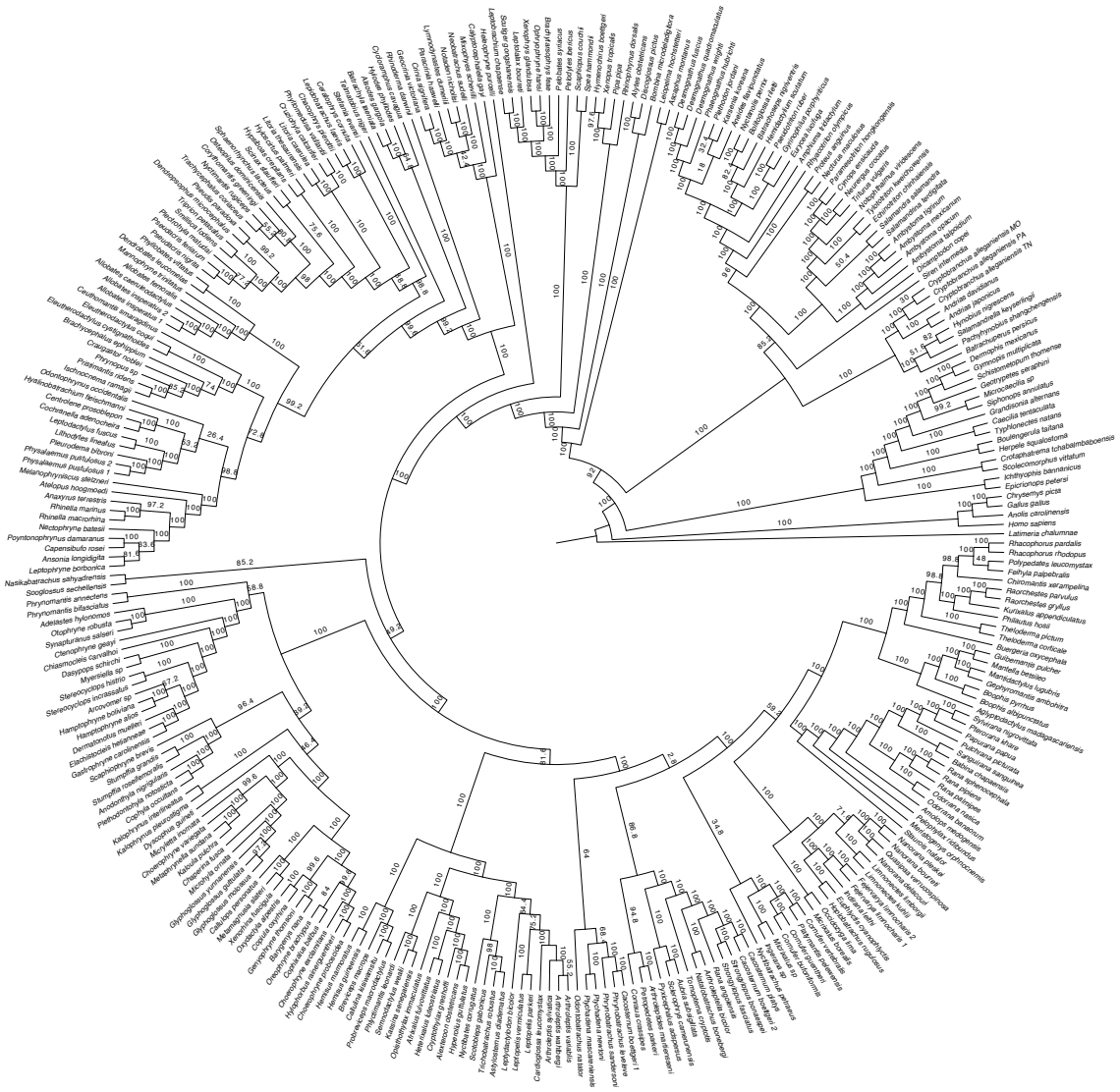


Figure 5.16. 90% present sites MulRF tree.

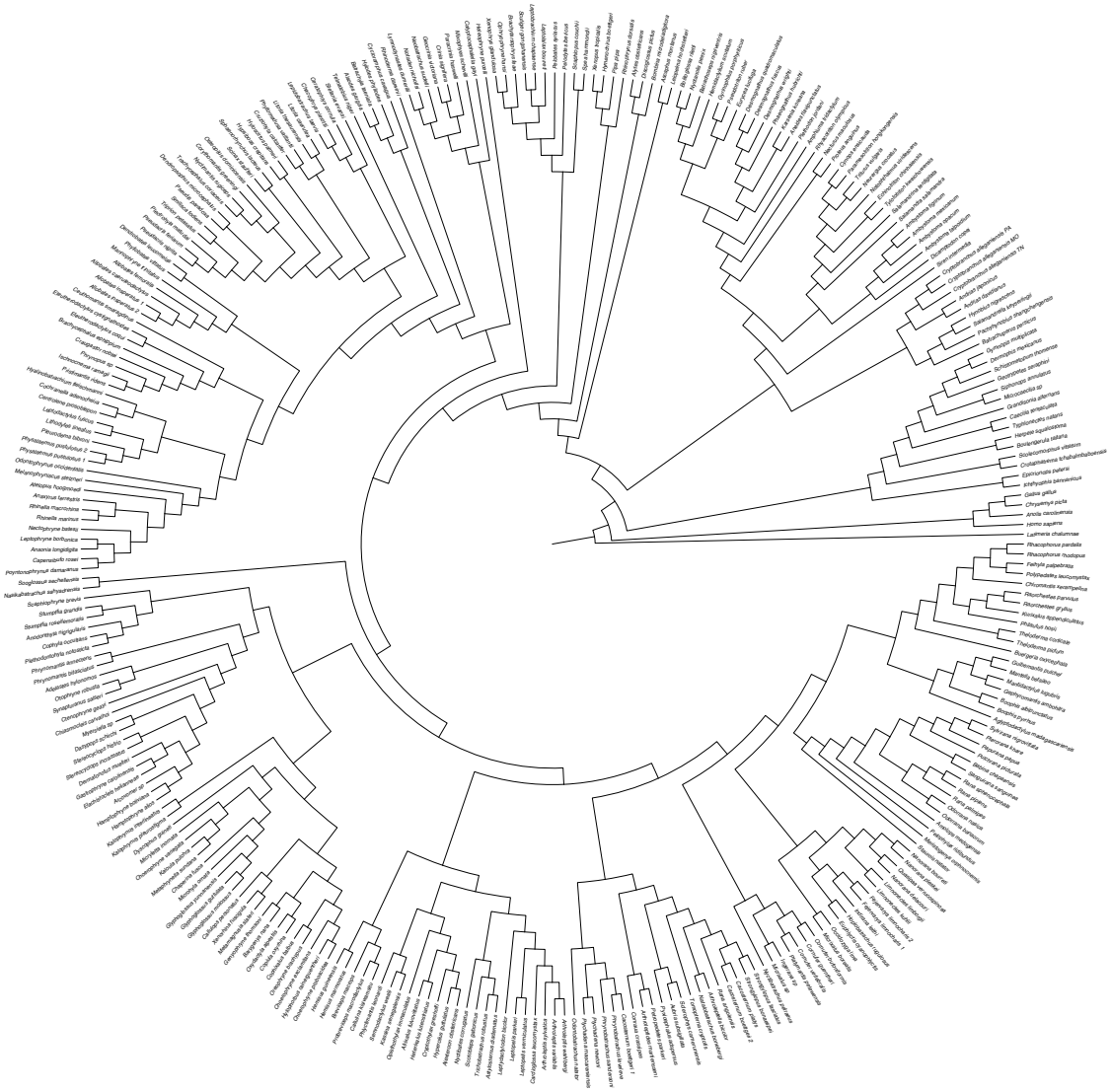


Figure 5.17. 90% present sites RAxML tree.

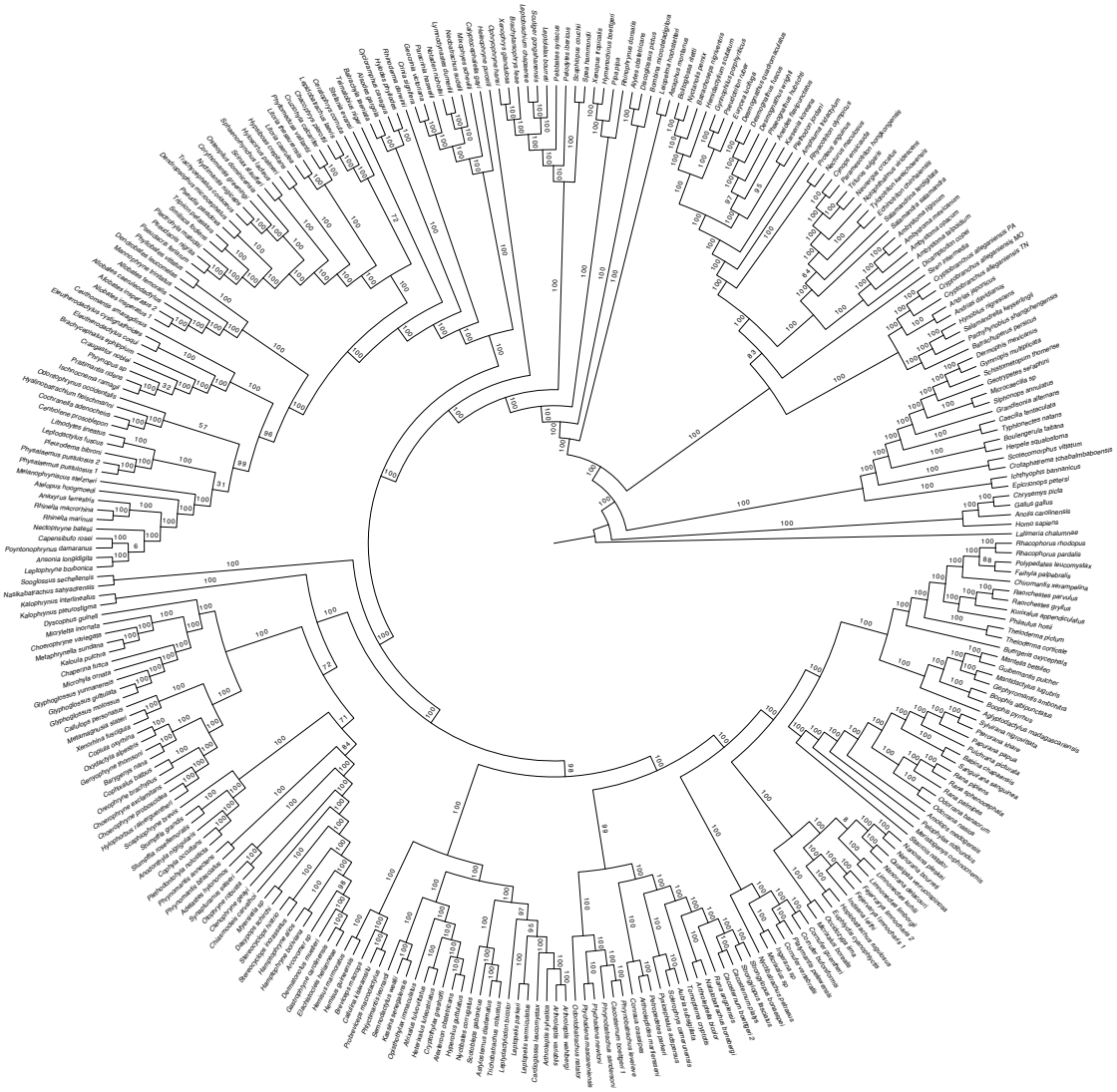
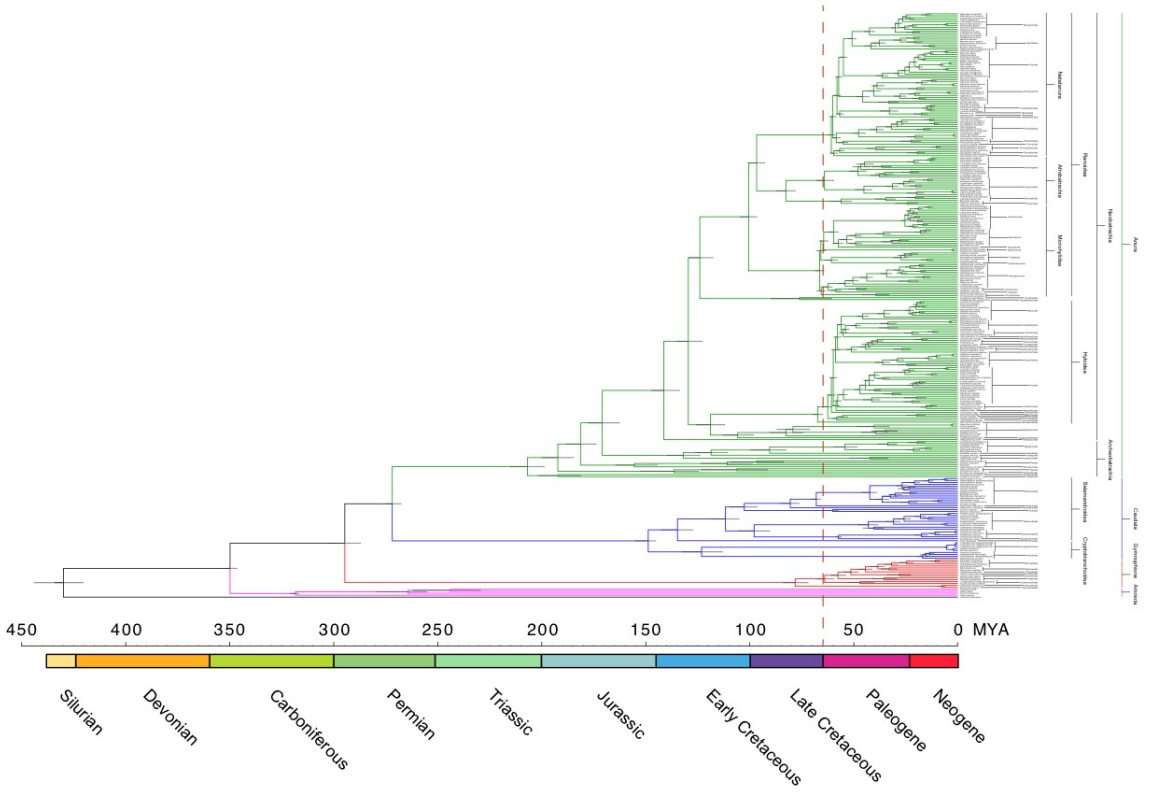


Figure 5.18. Divergence times estimated across Amphibia using 25 fossil calibrations with soft bounds. Amniotes, caecilians, salamanders, and frogs are shown in magenta, red, blue, and green respectively. Families and key subfamilies are labeled, as are higher-order clades, at right. The dashed red line indicates the Cretaceous-Tertiary boundary. 95% credible intervals for divergence times are depicted as gray bars on nodes in the tree.



CHAPTER SIX

Synthesis

ABSTRACT

In this dissertation I investigate amphibian evolution across multiple phylogenetic scales, from the early stages of speciation in Mexican and North American aquatic salamanders (Chapters 2 and 3, respectively), to the evolution of sex-linked genes across a deeply divergent family of salamanders (Chapter 4), to the divergences among and within the major family-level amphibian orders (Chapter 5). In these chapters, I address fundamental questions about the sources, magnitude, and downstream effects of varying, and sometimes conflicting, phylogenetic signals from across the nuclear genome. Each of these four empirical chapters seeks to test hypotheses about aspects of evolutionary biology in particular organismal systems, and each chapter brings some sort of genomic data to bear on these questions. In three of these four chapters, these genomic resources were developed from scratch specifically for the taxa and questions at hand, requiring non-trivial amounts of effort to optimize and deploy these new systems of data collection. Yet, these data are merely tools with which to investigate pressing applied and basic evolutionary questions in non-model species, and beyond the organismal foci of some of these chapters, the more general and unifying themes of this body of work revolve around issues of model adequacy in phylogenetics and the quantification of information content for different regions of the genome. My dissertation also probes the impending dilemma facing many

phylogeneticists in the genomic age wherein systematists can now collect data sets which overwhelm one's abilities to perform analyses with the same standards and rigor from the bygone PCR-era. Given that phylogeneticists are still in the early years of the emerging post-modern synthesis of genomics and phylogenetics, it is highly likely that the coming decade will present fantastic new opportunities to address longstanding evolutionary questions in the light of completely novel types of genomic information and new, more powerful statistical models. As today's (mainly) sub-genomic approaches give way to complete genome sequences, the importance of assessing the fit of the underlying statistical models to these data sets and the rigor with which one must scrutinize inferences will both increase markedly. Ultimately, this dissertation aims to illuminate the historical and demographic factors which have produced the rich diversity of life which is observed on Earth today.

INTRODUCTION

The field of phylogenetics has progressed substantially in terms of the amounts and types of data available for estimating evolutionary relationships. But perhaps more importantly, the field has also progressed in terms of the rigor and scalability of statistical models to explain these data and from which to address key questions in evolutionary biology. But as far as the field has come, there is still a long way to go. Many of the most powerful and most informative models simply do not accommodate today's data sets (let alone tomorrow's).

Systematists are very excited about "data" these days (e.g., Glenn 2011; Goodwin *et al*. 2016; Lemmon & Lemmon 2013). But, data by themselves are meaningless outside of the context of an appropriate and explanatory model, and "data" do not uniformly equate to "information" (Lewis *et al*. 2016). However, data and models do not exist in isolation; the availability of new types of data drives the development of new models which may better account for them (e.g., Miyamoto & Fitch 1995, Catchen *et al*. 2013). And the development of more rigorous models and simulation studies to assess their potential explanatory power can also guide the collection of appropriate types of data. Data and models go hand in hand. The supposed debate between the relative importance of better models or more data is a bit of a strawman argument. Of course evolutionary biologists want both of these desires to be met. And data do not contain information in the absence of a good model, just as much as models are pointless (debatably) if one have insufficient data to test (and hopefully reject some of) them (Burnham & Anderson 2003).

The inherent tensions between the (both reasonable) desires for more data and/or better models is nothing new in systematics. Yet now, unlike ever before, the rate of data set expansion is far outpacing the refinement of existing models and software applications to process what often amounts to orders of magnitude larger molecular data sets. Beyond that, applying many of the cutting edge advances in model-based approaches to phylogenetics or population genetics is computationally burdensome enough on its own, but add to those complications significantly more (and potentially more noisy) data, and the analytical hurdles to performing phylogenetics in the age of genomics become truly daunting. In general, systematists are constrained by trade-offs between the speed with which analyses can be executed, the degree of analytical rigor which can be expected, and

the tractability of the analyses. This tension, depicted in Figure 6.1, often dictates many decisions in phylogenetics, and not always in the direction of increasing rigor.

DISCUSSION

As the number of loci in genetic data sets balloons, the level of hands-on interaction that systematists can have with these data is on the decline. Yet, some best practices are beginning to emerge, and this dissertation research intervenes in the discussion about scrupulous handling of massive phylogenomic and population genomic data sets. First, it is critical to sample genetic markers from across the genome in order to obtain estimates of phylogeny and parameters of interest which account for variation across the genome. At shallow scales, for instance, the number and information content of the genetic markers analyzed can influence the outcomes of species delimitation studies (e.g., Hime *et al*. 2016). Additionally, support for different phylogenetic hypotheses can vary substantially across loci at deeper phylogenetic scales (e.g., Chen *et al*. 2015; Fong *et al*. 2012). Different types of genetic markers (coding or non-coding) may strongly support different, conflicting topologies (e.g., Harvey *et al*. 2016; Jarvis *et al*. 2014, Prum *et al*. 2015, Reddy *et al*. 2016). Furthermore, it is essential to scrupulously examine the robustness of support for phylogenetic hypotheses across loci (e.g., Arcila *et al*. 2017). In many cases, a few outlier genes (either misaligned or containing strange patterns of missing data) can drive strong support for incorrect inferences (Brown & Thomson 2016). Lastly, the nonparametric bootstrap has long been used to assess confidence in the branches in phylogenetic trees (Hillis & Bull 1993), although as an unbounded metric of support, these values better

measure the variance in phylogenetic signal across sites in the sequence alignments than they measure actual confidence in branches or in specific phylogenetic hypotheses (Erixon *et al*. 2003; Kumar *et al*. 2011). Unbounded metrics of support such as the Akaike information criterion (Akaike 1974) or Bayes factors (Kass & Raftery 1995) may provide additional insights into not only the direction of support for phylogenetic hypotheses, but also about the magnitude of support (e.g., Brown & Thomson 2016).

Systematists, especially those being trained today, may be tempted to consider inferences from species tree methods to be more reliable than concatenation-based approaches, in cases where the two methods disagree, because the former can better account for variation in phylogenetic information across different loci. Yet, as the example in Chapter 5 demonstrates with the placement of *Nasikabatrachus* within the frog phylogeny varying markedly between species tree methods or concatenation, cryptic systematic bias in the input gene trees for "shortcut" methods can lead to strongly supported, yet spurious, phylogenetic inferences. In this case, an artifact of missing sites in the individual gene alignments apparently drove slightly more than half of the loci examined to support *Nasikabatrachus* sister to Microhylidae in the species tree analysis, whereas the concatenated analyses recovered the "correct" topology of *Nasikabatrachus*+*Sooglossus*. Were it not for the unanimous support for the canonical placement of *Nasikabatrachus* within the frog tree, this type of error propagating from the level of gene trees up to the level of species trees would not have necessarily been apparent. This example serves to highlight the importance of scrupulously examining the strength of support for potentially novel phylogenetic hypotheses (perhaps in proportion to how extraordinary those claims may be), particularly in cases where different tree

reconstruction methods strongly support conflicting inferences. In this case, adding more data to the question of amphibian relationships actually exacerbated the problem of phylogenetic reconstruction in a few key portions of the tree (*Nasikabatrachus*), but was likely necessary for uncovering other genuinely surprising aspects of the amphibian phylogeny (such as the relationships within the three main lineages in Ranoidea). Nonetheless, overall, the inclusion of large amounts of sequence data for large numbers of amphibian taxa resulted in refined estimates of parameters of interest (such as divergence times), even though the benefit of hundreds of loci for resolving the tree topology was partly offset by topological inaccuracies in other parts of the tree under species tree methods.

In other cases though, the underlying models to be tested are relatively simple, and the limiting factor in the resolution of an evolutionary question was more a simple matter of collecting greater amounts of data. This was the case in Chapter 4 wherein it was necessary to sequence hundreds of thousands of anonymous genomic markers in order to discover a set of only two which appear to be in linkage with the sex chromosomes. Previous efforts to tackle the question of sex-linked loci and the system of heterogamety in cryptobranchid salamanders had been stymied by a lack of genetic data.

Similarly, in the case of population genetic structure and potential species boundaries within hellbender salamanders, access to large swaths of the nuclear genome appears to have provided resolution of not only the degree of structure across the geographic distribution of these aquatic salamanders, but also into the relationships among these different genetic lineages of hellbenders. The distinctiveness of the Tennessee River, Ohio River, and Ozarks populations had been hinted at by several previous studies using

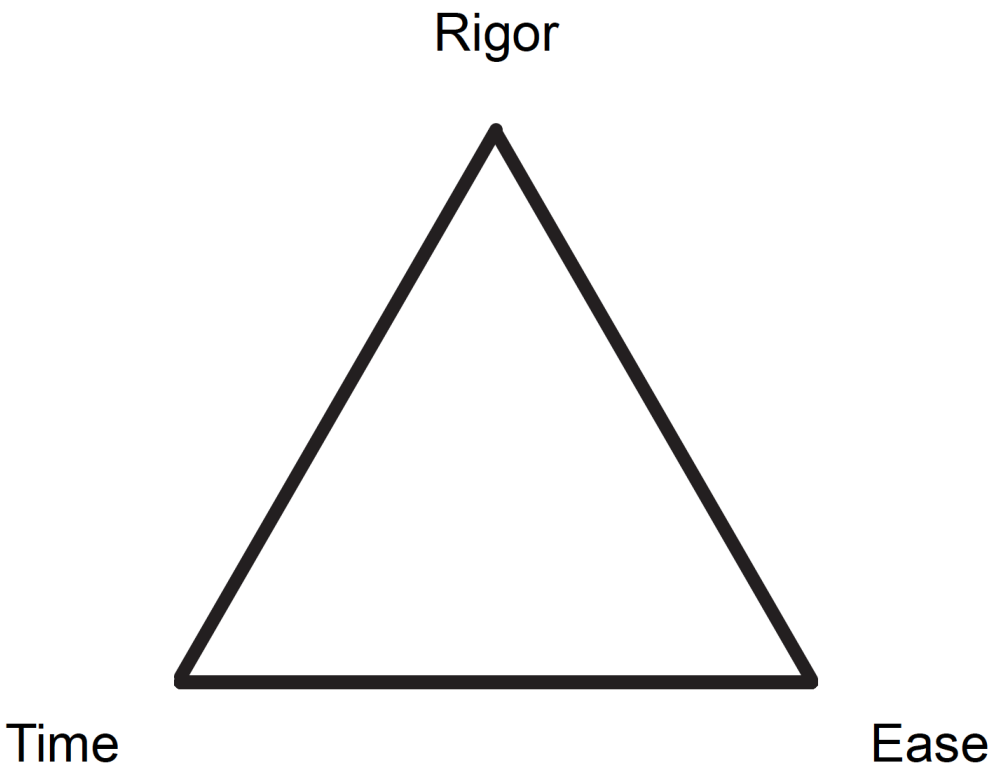
microsatellite markers (e.g., Crowhurst *et al*. 2011; Tonione *et al*. 2011; Unger *et al*. 2012), and connections had been proposed between the Kanawha River and the Ozarks populations when examining mitochondrial DNA (Routman & Templeton 1994; Sabatino & Routman 2009). However, the genome-wide markers developed here for hellbenders allow substantially higher resolution of the divisions between lineages, the phylogenetic relationships among these lineages, and the demographic parameters (such as genetic diversity, effective population sizes, and rates of effective gene flow) which are relevant to delimiting putative species boundaries and for applied conservation efforts.

The tradeoffs between sampling individuals, populations, and genetic markers is becoming less stark as phylogenetics enters the genomic era, yet, it will likely still remain an important tradeoff to consider. The research in Chapter 2 highlights that sampling greater numbers of individuals for a handful of loci provided as much if not more information about the primary divergence in a group of Mexican stream salamanders (eastern versus western populations/species) than a data set with an order of magnitude more loci of much lower individual information content. Yet, species delimitation using larger numbers of less informative loci also consistently supported two secondary divergence events (within eastern and within western populations). Clearly the scales of divergence with which a particular study is concerned will influence the tradeoffs between the sampling of sites, loci, individuals, populations, and species.

"Simply" collecting genomic data, regardless of how fashionable the latest approach, will not necessarily, on its own, resolve evolutionary questions any better than traditional types of data. In fact, just the "simple" act of increasing data set size adds non-linearly to many of the already non-trivial computation burdens and concerns about model

adequacy. To the extent that genomic data sets have great power to address key evolutionary questions, so also do evolutionary biologists have a great responsibility to analyze them scrupulously. This dissertation research may help to inform best practices for phylogenomics which may be applicable across multiple phylogenetic scales.

Figure 6.1. Some of the many tradeoffs in phylogenomics.



REFERENCES

Abdullayev, I., Kirkham, M., Björklund, Å. K., Simon, A., & Sandberg, R. (2013). A reference transcriptome and inferred proteome for the salamander *Notophthalmus viridescens*. *Experimental Cell Research*, 319(8), 1187-1197.

Adkins-Regan, E., & Reeve, H. K. (2014). Sexual dimorphism in body size and the origin of sex-determination systems. *The American Naturalist*, 183(4), 519-536.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723.

AmphibiaWeb.org. http://amphibiaweb.org (accessed 25 March, 2017).

Anderson, J. D., & Worthington, R. D. (1971). The life history of the Mexican salamander *Ambystoma ordinarium*. *Herpetologica*, 27, 165-176.

Anderson, J. S. (2008). Focal review: the origin(s) of modern amphibians. *Evolutionary Biology*, 35(4), 231-247.

Anderson, J. S., Reisz, R. R., Scott, D., Fröbisch, N. B., & Sumida, S. S. (2008). A stem batrachian from the Early Permian of Texas and the origin of frogs and salamanders. *Nature*, 453(7194), 515.

Arcila, D., Ortí, G., Vari, R., Armbruster, J.W., Stiassny, M.L., Ko, K.D., Sabaj, M.H., Lundberg, J., Revell, L.J., & Betancur-R, R. (2017). Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology & Evolution*, 1(1), 0020.

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A., & Saunders, N. C. (1987). Intraspecific phylogeography: the mitochondrial DNA

bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, 18(1), 489-522.

Báez, A. M., Moura, G. J., & Gómez, R. O. (2009). Anurans from the Lower Cretaceous Crato Formation of northeastern Brazil: implications for the early divergence of neobatrachians. *Cretaceous Research*, 30(4), 829-846.

Barraclough, T. G., & Nee, S. (2001). Phylogenetics and speciation. *Trends in Ecology & Evolution*, 16(7), 391-399.

Baum, D.A., Ané, C., Larget, B., Solís-Lemus, C., Ho, L.S.T., Boone, P., Drummond, C.P., Bontrager, M., Hunter, S.J., & Saucier, W. (2016). Statistical evidence for common ancestry: Application to primates. *Evolution*, 70(6), 1354-1363.

Beerli, P. (2006). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, 22, 341-345.

Beerli, P., & Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 4563-4568.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., & Boutell, J. M. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), 53-59.

Benton, M. J., Donoghue, P. C., Asher, R. J., Friedman, M., Near, T. J., & Vinther, J. (2015). Constraints on the timescale of animal evolutionary history. *Palaeontologia Electronica*, 18(1), 1-106.

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, 21(2), 163-193.

Biju SD & Bossuyt F. (2003). New frog family from India reveals an ancient biogeographical link with the Seychelles. *Nature*, 425(6959), 711–714.

Blackburn, D. C., Roberts, E. M., & Stevens, N. J. (2015). The earliest record of the endemic African frog family Ptychadenidae from the Oligocene Nsungwe Formation of Tanzania. *Journal of Vertebrate Paleontology*, 35(2), e907174.

Bradbury, J. (2000). Limnologic history of Lago de Patzcuaro, Michoacan, Mexico for the past 48,000 years: impacts of climate and man. *Palaeogeography, Palaeoclimatology, Palaeoecology* 163, 69-95.

Brown, J. M., & Thomson, R. C. (2016). Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Systematic Biology*, 66(4), 517-530.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Camargo, A., Morando, M., Avila, L. J., & Sites, J. W., Jr. (2012). Species delimitation with ABC and other coalescent-based methods: a test of accuracy with simulations and an empirical example with lizards of the *Liolaemus darwinii* complex (Squamata: Liolaemidae). *Evolution*, 66, 2834-2849.

Cannatella, D. (2015). *Xenopus* in space and time: Fossils, node calibrations, tip-dating, and paleobiogeography. *Cytogenetic and Genome Research*, 145(3-4), 283-301.

Cano, J. M., Li, M. H., Laurila, A., Vilkki, J., & Merilä, J. (2011). First-generation linkage map for the common frog *Rana temporaria* reveals sex-linkage group. *Heredity*, 107(6), 530.

Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. (2013). How to fail at species delimitation. *Molecular Ecology*, 22(17), 4369-4383.

Castro-Nallar, E., Pérez-Losada, M., Burton, G. F., & Crandall, K. A. (2012). The evolution of HIV: inferences using phylogenetics. *Molecular Phylogenetics and Evolution*, 62(2), 777-792.

Catchen, J., 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.

Chang, S. C., Zhang, H., Renne, P. R., & Fang, Y. (2009). High-precision $^{40}Ar/^{39}Ar$ age for the Jehol biota. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 280(1), 94-104.

Charlesworth, D., & Mank, J. E. (2010). The birds and the bees and the flowers and the trees: lessons from genetic mapping of sex determination in plants and animals. *Genetics*, 186(1), 9-31.

Chaudhary, R., Fernández-Baca, D., & Burleigh, J. G. (2014). MulRF: a software package for phylogenetic analysis using multi-copy gene trees. *Bioinformatics*, 31(3), 432-433.

Che, R., Sun, Y., Wang, R., & Xu, T. (2014). Transcriptomic analysis of endangered Chinese salamander: identification of immune, sex and reproduction-related genes and genetic markers. *PLoS One*, 9(1), e87940.

Chen, M. Y., Liang, D., & Zhang, P. (2015). Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. *Systematic Biology*, 64(6), 1104–1120.

Chifman, J., & Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics*, 30(23), 3317-3324.

Cohen, K. M., Finney, S. C., Gibbard, P. L., & Fan, J. X. (2013). The ICS international chronostratigraphic chart. *Episodes*, 36(3), 199-204.

Crowhurst, R. S., Faries, K. M., Collantes, J., Briggler, J. T., Koppelman, J. B., & Eggert, L. S. (2011). Genetic relationships of hellbenders in the Ozark highlands of Missouri and conservation implications for the Ozark subspecies (*Cryptobranchus alleganiensis bishopi*). *Conservation Genetics*, 12(3), 637-646.

de Sá, R. O., Streicher, J. W., Sekonyela, R., Forlani, M. C., Loader, S. P., Greenbaum, E., Richards, S., & Haddad, C. F. (2012). Molecular phylogeny of microhylid frogs (Anura: Microhylidae) with emphasis on relationships among New World genera. *BMC Evolutionary Biology*, 12(1), 241.

Degnan, J. H., & Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5), e68.

Doadrio, I., & Dominguez, O. (2004). Phylogenetic relationships within the fish family Goodeidae based on cytochrome *b* sequence data. *Molecular Phylogenetics and Evolution*, 31, 416-430.

Dominguez-Dominguez, O., Alda, F., de Leon, G. P. P., Garcia-Garitagoitia, J. L., & Doadrio, I. (2008). Evolutionary history of the endangered fish *Zoogoneticus quitzeoensis* (Bean, 1898) (Cyprinodontiformes: Goodeidae) using a sequential approach to phylogeography based on mitochondrial and nuclear DNA data. *BMC Evolutionary Biology*, 8, 161.

Dominguez-Dominguez, O., Doadrio, I., & de Leon, G. P. P. (2006). Historical biogeography of some river basins in central Mexico evidenced by their goodeine freshwater fishes: a preliminary hypothesis using secondary Brooks parsimony analysis. *Journal of Biogeography*, 33, 1437-1447.

Dundee, H. A., & Dundee, D. S. (1965). Observations on the systematics and ecology of *Cryptobranchus* from the Ozark Plateaus of Missouri and Arkansas. *Copeia*, 1965(3), 369-370.

Edwards, D. L., & Knowles, L. L. (2014). Species detection and individual assignment in species delimitation: can integrative data increase efficacy? *Proceedings of the Royal Society B*, 281, 20132765.

Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging?. *Evolution,* 63(1), 1-19.

Edwards, S. V., Jennings, W. B., & Shedlock, A. M. (2005). Phylogenetics of modern birds in the era of genomics. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1567), 979-992.

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D, Liu, L., & Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, 94(A), 447-462.

Ellegren, H. (1996). First gene on the avian W chromosome (CHD) provides a tag for universal sexing of non-ratite birds. *Proceedings of the Royal Society of London B: Biological Sciences*, 263(1377), 1635-1641.

Ence, D. D., & Carstens, B. C. (2011). SpedeSTEM: a rapid and accurate method for species delimitation. *Molecular Ecology Resources*, 11, 473-480.

Erixon, P., Svennblad, B., Britton, T., & Oxelman, B. (2003). Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic Biology*, 52(5), 665-673.

Estes, R., & Wake, M. H. (1972). The first fossil record of caecilian amphibians. *Nature*, 239(5369), 228-231.

Evans, S. E., Lally, C., Chure, D. C., Elder, A., & Maisano, J. A. (2005). A late Jurassic salamander (Amphibia: Caudata) from the Morrison formation of north America. *Zoological Journal of the Linnean Society*, 143(4), 599-616.

Ezaz, T., Srikulnath, K., & Graves, J. A. M. (2016). Origin of amniote sex chromosomes: an ancestral super-sex chromosome, or common requirements? *Journal of Heredity*, 108(1), 94-105.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, 61(5), 717-726.

Fan, Y., Chang, M. X., Ma, J., LaPatra, S. E., Hu, Y. W., Huang, L., Nie, P., & Zeng, L. (2015). Transcriptomic analysis of the host response to an iridovirus infection in Chinese giant salamander, *Andrias davidianus*. *Veterinary Research*, 46(1), 136.

Feist, S. M., Briggler, J. T., Koppelman, J. B., & Eggert, L. S. (2014). Within-river gene flow in the hellbender (*Cryptobranchus alleganiensis*) and implications for restorative release. *Conservation genetics*, 15(4), 953-966.

Feller, A. E., & Hedges, S. B. (1998). Molecular evidence for the early history of living amphibians. *Molecular Phylogenetics and Evolution*, 9(3), 509–516.

Feng, Y. J., Blackburn, D. C., Liang, D., Hillis, D. M., Wake, D. B., Cannatella, D. C., & Zhang, P. (2017). Phylogenomics reveals rapid, simultaneous diversification of three major clades of Gondwanan frogs at the Cretaceous–Paleogene boundary. *Proceedings of the National Academy of Sciences*, 201704632.

Ferrari, L., Conticelli, S., Vaggelli, G., Petrone, C. M., & Manetti, P. (2000). Late Miocene volcanism and intra-arc tectonics during the early development of the Trans-Mexican Volcanic Belt. *Tectonophysics*, 318, 161-185.

Ferrari, L., Lopez-Martinez, M., Aguirre-Diaz, G., & Carrasco-Nunez, G. (1999). Space-time patterns of Cenozoic arc volcanism in central Mexico: From the Sierra Madre Occidental to the Mexican Volcanic Belt. *Geology*, 27, 303-306.

Ferrusquía-Villafranca, I., & González-Guzmán, L. I. (2005). Northern Mexico's landscape, part II: the biotic setting across time. In: *Biodiversity, Ecosystems, and Conservation in Northern Mexico* (eds. Cartron J-LE, Ceballos G, Fleger RS), pp. 39-51. Oxford University Press, Oxford, U.K.

Firschein, I. L. (1951). The range of *Cryptobranchus bishopi* and remarks on the distribution of the genus *Cryptobranchus*. *The American Midland Naturalist*, 45(2), 455-459.

Fong, J. J., Brown, J. M., Fujita, M. K., & Boussau, B. (2012). A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic lissamphibia. *PLoS One*, 7(11), e48990.

Frost, D. R., Grant, T., Faivovich, J., Bain, R. H., Haas, A., Haddad, C. F., De Sá, R. O., Channing, A., Wilkinson, M., Donnellan, S. C., & Raxworthy, C. J. (2006). The amphibian tree of life. *Bulletin of the American Museum of Natural History*, 1–291.

Fujita, M. K., Leaché, A. D., Burbrink, F. T., McGuire, J. A., & Moritz, C. (2012). Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution*, 27(9), 480-488.

Furman, B. L., & Evans, B. J. (2016). Sequential turnovers of sex chromosomes in African clawed frogs (*Xenopus*) suggest some genomic regions are good at sex determination. *G3: Genes, Genomes, Genetics*, 6(11), 3625-3633.

Gallego-García, N., & Páez, V. P. (2016). Geographic variation in sex determination patterns in the river turtle *Podocnemis lewyana*: implications for global warming. *Journal of Herpetology*, 50(2), 256-262.

Galloway, W. E., Whiteaker, T. L., & Ganey-Curry, P. (2011). History of Cenozoic North American drainage basin evolution, sediment yield, and accumulation in the Gulf of Mexico basin. *Geosphere*, 7(4), 938-973.

Gamble, T. (2016). Using RAD-seq to recognize sex-specific markers and sex chromosome systems. *Molecular Ecology*, 25(10), 2114-2116.

Gamble, T., & Zarkower, D. (2012). Sex determination. *Current Biology*, 22(8), R257-R262.

Gamble, T., & Zarkower, D. (2014). Identification of sex-specific molecular markers using restriction site-associated DNA sequencing. *Molecular Ecology Resources*, 14(5), 902-913.

Gamble, T., Coryell, J., Ezaz, T., Lynch, J., Scantlebury, D. P., & Zarkower, D. (2015). Restriction site-associated DNA sequencing (RAD-seq) reveals an extraordinary number of transitions among gecko sex-determining systems. *Molecular Biology and Evolution*, 32(5), 1296-1309.

Gao, K. Q., & Shubin, N. H. (2003). Earliest known crown-group salamanders. *Nature*, 422(6930), 424.

Gardner, J. D. (2003). The fossil salamander *Proamphiuma cretacea* Estes (Caudata; Amphiumidae) and relationships within the Amphiumidae. *Journal of Vertebrate Paleontology*, 23(4), 769-782.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, 11(5), 759-769.

Glor, R. E. (2010). Phylogenetic insights on adaptive radiation. *Annual Review of Ecology, Evolution, and Systematics*, 41, 251-270.

Gómez, R. O., & Turazzini, G. F. (2016). An overview of the ilium of anurans (Lissamphibia, Salientia), with a critical appraisal of the terminology and primary homology of main ilial features. *Journal of Vertebrate Paleontology*, 36(1), e1030023.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333-351.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., & Chen, Z. (2011). Full-length

transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7), 644-652.

Graves, J. A. M., & Peichel, C. L. (2010). Are homologies in vertebrate sex determination due to shared ancestry or to limited options?. *Genome Biology*, 11(4), 205.

Gregory, T. R. (2017). Animal Genome Size Database. http://www.genomesize.com.

Grobman, A. G. (1943). Notes on salamanders with the description of a new species of *Cryptobranchus*. Occasional Papers of the Museum of Zoology, University of Michigan. University of Michigan Press, Ann Arbor, MI.

Grummer, J. A., Bryson, R. W., Jr., & Reeder, T. W. (2014). Species delimitation using Bayes factors: simulations and application to the *Sceloporus scalaris* species group (Squamata: Phrynosomatidae). *Systematic Biology*, 63, 119-133.

Hahn, M. W., & Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70(1), 7-17.

Harris, R. B., Carling, M. D., & Lovette, I. J. (2014). The influence of sampling design on species tree inference: a new relationship for the New World chickadees (Aves: *Poecile*). *Evolution*, 68, 501-513.

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., & Brumfield, R. T. (2016). Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Systematic Biology*, 65(5), 910-924.

Hellsten, U., Harland, R. M., Gilchrist, M. J., Hendrix, D., Jurka, J., Kapitonov, V., Ovcharenko, I., Putnam, N. H., Shu, S., Taher, L., & Blitz, I. L. (2010). The genome of the Western clawed frog *Xenopus tropicalis*. *Science*, 328(5978), 633-636.

Henrici, A. C., & Haynes, S. R. (2006). *Elkobatrachus brocki*, a new pelobatid (Amphibia: Anura) from the Eocene Elko Formation of Nevada. *Annals of the Carnegie Museum*, 75(1), 11-35.

Hey, J. (2010). Isolation with migration models for more than two populations. *Molecular Biology and Evolution* 27, 905-920.

Hillis, D. M., & Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42(2), 182-192.

Hime, P. M., Hotaling, S., Grewelle, R. E., O'Neill, E. M., Voss, S. R., Shaffer, H. B., & Weisrock, D. W. (2016). The influence of locus number and information content on species delimitation: an empirical test case in an endangered Mexican salamander. *Molecular Ecology*, 25(23), 5959-5974.

Hird, S., Kubatko, L., & Carstens, B. C. (2010). Rapid and accurate species tree estimation for phylogeographic investigations using replicated subsampling. *Molecular Phylogenetics and Evolution*, 57, 888-898.

Holder, M., & Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics*, 4(4), 275-284.

Hosner, P. A., Faircloth, B. C., Glenn, T. C., Braun, E. L., & Kimball, R. T. (2016). Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). *Molecular Biology and Evolution*, 33(4), 1110-1125.

Hotaling, S., Foley, M. E., Lawrence, N. M., Bocanegra, J., Blanco, M. B., Rasoloarison, R., Kappeler, P. M., Barrett, M. A., Yoder, A. D., & Weisrock, D. W. (2016). Species discovery and validation in a cryptic radiation of endangered primates:

coalescent-based species delimitation in Madagascar's mouse lemurs. *Molecular Ecology*, 25, 2029-2045.

Huang, H., He, Q., Kubatko, L. S., & Knowles, L. L. (2010). Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Systematic Biology*, 59, 573-583.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18, 337-338.

Huelsenbeck, J. P., Andolfatto, P., & Huelsenbeck, E. T. (2011). Structurama: Bayesian inference of population structure. *Evolutionary Bioinformatics*, 7, 55-59.

Hulsey, C. D., de Leon, F. J. G., Johnson, Y. S., Hendrickson, D. A., & Near, T. J. (2004). Temporal diversification of Mesoamerican cichlid fishes across a major biogeographic boundary. *Molecular Phylogenetics and Evolution*, 31, 754-764.

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies, *Molecular Biology and Evolution*, 23, 254-267.

Huson, D. H., Kloepper, T., & Bryant, D. (2008). SplitsTree 4.0: Computation of phylogenetic trees and networks. *Bioinformatics*, 14, 68-73.

Irwin, D. E. (2002). Phylogeographic breaks without geographic barriers to gene flow. *Evolution*, 12, 2383-2394.

Israde-Alcantara, I., & Garduno-Monroy, V. H. (1999). Lacustrine record in a volcanic intra-arc setting: the evolution of the Late Neogene Cuitzeo basin system (central-western Mexico, Michoacan). *Palaeogeography Palaeoclimatology Palaeoecology*, 151, 209-227.

Jackson, N. D., Austin, C. C. (2012). Inferring the evolutionary history of divergence despite gene flow in a lizard species, *Scincella lateralis* (Scincidae), composed of cryptic lineages. *Biological Journal of the Linnean Society*, 107, 192-209.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y., Faircloth, B. C., Nabholz, B., Howard, J. T., & Suh, A. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215), 1320-1331.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, 36(suppl_2), W5-W9.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27(21), 3070-3071.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, 11(1), 94.

Jónsson, H., Schubert, M., Seguin-Orlando, A., Ginolhac, A., Petersen, L., Fumagalli, M., Albrechtsen, A., Petersen, B., Korneliussen, T. S., Vilstrup, J. T., & Lear, T. (2014). Speciation with gene flow in equids despite extensive chromosomal plasticity. *Proceedings of the National Academy of Sciences*, 111(52), 18655-18660.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773-795.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772-780.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., & Thierer, T. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), 1647-1649.

Keinath, M. C., Voss, S. R., Tsonis, P. A., & Smith, J. J. (2017). A linkage map for the newt *Notophthalmus viridescens*: Insights in vertebrate genome and chromosome evolution. *Developmental Biology*, 426(2), 211-218.

Kozak, K. H., Blaine, R. A., & Larson, A. (2006). Gene lineages and eastern North American palaeodrainage basins: phylogeography and speciation in salamanders of the *Eurycea bislineata* species complex. *Molecular Ecology*, 15(1), 191-207.

Kubatko, L. S., & Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1), 17-24.

Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L., & Tamura, K. (2011). Statistics and truth in phylogenomics. *Molecular Biology and Evolution*, 29(2), 457-472.

Lanfear, R., Calcott, B., Ho, S. Y., & Guindon, S. (2012). PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, 29(6), 1695-1701.

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., & Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evolutionary Biology*, 14(1), 82.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., & Calcott, B. (2016). PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, 34(3), 772-773.

Lanier, H. C., Huang, H., & Knowles, L. L. (2014). How low can you go? The effects of mutation rate on the accuracy of species-tree estimation. *Molecular Phylogenetics and Evolution*, 70, 112-119.

Larson, A., & Wilson, A. C. (1989). Patterns of ribosomal RNA evolution in salamanders. *Molecular Biology and Evolution*, 6(2), 131–154.

Leaché, A. D. (2009). Species tree discordance traces to phylogeographic clade boundaries in North American fence lizards (*Sceloporus*). *Systematic Biology*, 58, 547-559.

Lemmon, A. R., & Lemmon, E. M. (2012). High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. *Systematic Biology*, 61, 745-761.

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, 61(5), 727-744.

Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 44, 99-121.

Lewis, P. O., Chen, M. H., Kuo, L., Lewis, L. A., Fučíková, K., Neupane, S., Wang, Y. B., & Shi, D. (2016). Estimating Bayesian phylogenetic information content. *Systematic Biology*, 65(6), 1009-1023.

Linkem, C. W., Minin, V. N., & Leaché, A. D. (2016). Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (Squamata: Scincidae). *Systematic Biology*, 65(3), 465-477.

Liu, L., & Pearl, D. K. (2007). Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology*, 56(3), 504-514.

Liu, L., Yu, L., & Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1), 302.

Lozano-Garcia, S., & Vazquez-Selem, L. (2005). A high-elevation Holocene pollen record from Iztaccihuatl volcano, central Mexico. *The Holocene*, 15, 329-338.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3), 523-536.

Mateos, M., Sanjur, O. I., & Vrijenhoek, R. C. (2002). Historical biogeography of the livebearing fish genus *Poeciliopsis* (Poeciliidae: Cyprinodontiformes). *Evolution*, 56, 972-984.

Matsumoto, Y., Buemio, A., Chu, R., Vafaee, M., & Crews, D. (2013). Epigenetic control of gonadal aromatase (*cyp19a1*) in temperature-dependent sex determination of red-eared slider turtles. *PLoS One*, 8(6), e63599.

McCartney-Melstad, E., Mount, G. G., & Shaffer, H. B. (2016). Exon capture optimization in large-genome amphibians. *Molecular Ecology Resources*, 16, 1084-1094.

McCormack, J. E., Peterson, A. T., Bonaccorso, E., & Smith, T. B. (2008). Speciation in the highlands of Mexico: genetic and phenotypic divergence in the Mexican jay (*Aphelocoma ultramarina*). *Molecular Ecology*, 17, 2505-2521.

Merkle, D. A., Guttman, S. I., & Nickerson, M. A. (1977). Genetic uniformity throughout the range of the hellbender, *Cryptobranchus alleganiensis*. *Copeia*, 3, 549-553.

Metcalfe, S. E., O'Hara, S. L., Caballero, M., & Davies, S. J. (2000). Records of Late Pleistocene-Holocene climatic change in Mexico - a review. *Quaternary Science Reviews*, 19, 699-721.

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*. 31(12), i44-i52.

Miyamoto, M. M., & Fitch, W. M. (1995). Testing the covarion hypothesis of molecular evolution. *Molecular Biology and Evolution*, 12(3), 503-513.

Montiel, E. E., Badenhorst, D., Tamplin, J., Burke, R. L., & Valenzuela, N. (2017). Discovery of the youngest sex chromosomes reveals first case of convergent co-option of ancestral autosomes in turtles. *Chromosoma*, 126(1), 105-113.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G., & Worm, B. (2011). How many species are there on Earth and in the ocean?. *PLoS Biology*, 9(8), e1001127.

Morescalchi, A., Odierna, G., & Olmo, E. (1977). Karyological relationships between the Cryptobranchid salamanders. *Cellular and Molecular Life Sciences*, 33(12), 1579-1581.

Mulcahy, D .G., Morrill, B. H., & Mendelson, J. R. (2006). Historical biogeography of lowland species of toads (*Bufo*) across the Trans-Mexican Neovolcanic Belt and the Isthmus of Tehuantepec. *Journal of Biogeography*, 33, 1889-1904.

Nakamura, M. (2009). Sex determination in amphibians. *Seminars in Cell & Developmental Biology*, 20(3), 271-282.

Nakamura, M. (2013). Is a sex-determining gene(s) necessary for sex-determination in amphibians? Steroid hormones may be the key factor. *Sexual Development*, 7(1), 104-114.

Naylor, B. G. (1978). The earliest known Necturus (Amphibia, Urodela), from the Paleocene Ravenscrag Formation of Saskatchewan. *Journal of Herpetology*, 12(4), 565-569.

Naylor, B. G., & Fox, R. C. (1993). A new ambystomatid salamander, *Dicamptodon antiquus n. sp*., from the Paleocene of Alberta, Canada. *Canadian Journal of Earth Sciences*, 30(4), 814-818.

Nickerson, M. A., & Mays, C. E. (1973). The hellbenders: North American "giant salamanders". Milwaukee Public Museum Press.

O'Meara, B. C., Jackson, N. D., Morales-Garcia, A. E., & Carstens, B. C. (2015). Phylogeographic inference using approximate likelihoods. *BioRxiv* doi: http://dx.doi.org/10.1101/025353.

O'Neill, E. M., Schwartz, R., Bullock, C. T., Williams, J. S., Shaffer, H. B., Aguilar-Miguel, X., Parra-Olea, G., & Weisrock, D. W. (2013). Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North

American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology*, 22(1), 111-129.

Olave, M., Sola, E., & Knowles, L. L. (2014). Upstream analyses create problems with DNA-based species delimitation. *Systematic Biology*, 63, 263-271.

Oliver, J. C. (2013). Microevolutionary processes generate phylogenomic discordance at ancient divergences. *Evolution*, 67(6), 1823-1830.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20, 289-290.

Pauly, G. B., Piskurek, O., & Shaffer, H. B. (2007). Phylogeographic concordance in the southeastern United States: the flatwoods salamander, *Ambystoma cingulatum*, as a test case. *Molecular Ecology*, 16, 415-429.

Peloso, P. L., Frost, D. R., Richards, S. J., Rodrigues, M. T., Donnellan, S., Matsui, M., Raxworthy, C. J., Biju, S. D., Lemmon, E. M., Lemmon, A .R., & Wheeler, W. C. (2016). The impact of anchored phylogenomics and taxon sampling on phylogenetic inference in narrow-mouthed frogs (Anura, Microhylidae). *Cladistics*, 32(2), 113-140.

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.

Petkova, D., Novembre, J., & Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, 48(1), 94-100.

Pitt, A. L., Shinskie, J. L., Tavano, J. J., Hartzell, S. M., Delahunty, T., & Spear, S. F. (2017). Decline of a giant salamander assessed with historical records,

environmental DNA and multi-scale habitat data. *Freshwater Biology*, 62(6), 967-976.

Poe, S., & Chubb, A. L. (2004). Birds in a bush: five genes indicate explosive evolution of avian orders. *Evolution*, 58(2), 404-415.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945-959.

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Lemmon, E. M., & Lemmon, A. R. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, 526(7574), 569.

Pyron, R. A. (2015). Post-molecular systematics and the future of phylogenetics. *Trends in Ecology & Evolution*, 30(7), 384-389.

Pyron, R. A., & Wiens, J. J. (2011). A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Molecular Phylogenetics and Evolution*, 61(2), 543–583.

Qi, Z., Zhang, Q., Wang, Z., Ma, T., Zhou, J., Holland, J. W., & Gao, Q. (2016). Transcriptome analysis of the endangered Chinese giant salamander (*Andrias davidianus*): Immune modulation in response to *Aeromonas hydrophila* infection. *Veterinary Immunology and Immunopathology*, 169, 85-95.

Rage, J. C., & Roček, Z. (2007). A new species of *Thaumastosaurus* (Amphibia: Anura) from the Eocene of Europe. *Journal of Vertebrate Paleontology*, 27(2), 329-336.

Rannala, B. (2015). The art and science of species delimitation. *Current Zoology*, 61, 846-853.

Rannala, B., & Yang, Z. (2013). Improved reversible jump algorithms for Bayesian species delimitation. *Genetics*, 194(1), 245-253.

Reddy, S., Kimball, R. T., Pandey, A., Hosner, P. A., Braun, M. J., Hackett, S. J., Han, K. L., Harshman, J., Huddleston, C. J., Kingston, S. and Marks, B. D. (2017). Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. *Systematic Biology*, DOI: 10.1093/sysbio/syx041.

Reid, N. M., Hird, S. M., Brown, J. M., Pelletier, T. A., McVay, J. D., Satler, J. D., & Carstens, B. C. (2013). Poor fit to the multispecies coalescent is widely detectable in empirical data. *Systematic Biology*, 63(3), 322-333.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2), 217-223.

Rittmeyer, E. N., & Austin, C. C. (2015). Combined next-generation sequencing and morphology reveal fine-scale speciation in Crocodile Skinks (Squamata: Scincidae: *Tribolonotus*). *Molecular Ecology*, 24, 466-483.

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2), 131-147.

Rodrigues, N., Merilä, J., Patrelle, C., & Perrin, N. (2014). Geographic variation in sex-chromosome differentiation in the common frog (*Rana temporaria*). *Molecular Ecology,* 23(14), 3409-3418.

Roelants, K., Gower, D. J., Wilkinson, M., Loader, S. P., Biju, S. D., Guillaume, K., Moriau, L., & Bossuyt, F. (2007). Global patterns of diversification in the history of modern amphibians. *Proceedings of the National Academy of Sciences*, 104(3), 887-892.

Rogers, R. R., Krause, D. W., Kast, S. C., Marshall, M. S., Rahantarisoa, L., Robins, C. R., & Sertich, J. J. (2013). A new, richly fossiliferous member comprised of tidal deposits in the Upper Cretaceous Maevarano Formation, northwestern Madagascar. *Cretaceous Research*, 44, 12-29.

Rokyta, D. R., Lemmon, A. R., Margres, M. J., & Aronow, K. (2012). The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). *BMC Genomics*, 13(1), 312.

Roth, T. L., & Obringer, A. R. (2003). Reproductive research and the worldwide amphibian extinction crisis. In 'Reproductive Science and Integrated Conservation'. (Eds WV Holt, AR Pickard, JC Rodger and DE Wildt.) pp. 359–374.

Routman, E., Wu, R., & Templeton, A. R. (1994). Parsimony, molecular evolution, and biogeography: the case of the North American giant salamander. *Evolution*, 48(6), 1799-1809.

Rovatsos, M., & Kratochvíl, L. (2017). Molecular sexing applicable in 4000 species of lizards and snakes? From dream to real possibility. *Methods in Ecology and Evolution*, doi:10.1111/2041-210X.12714.

Rovatsos, M., Vukić, J., Lymberakis, P., & Kratochvíl, L. (2015). Evolutionary stability of sex chromosomes in snakes. *Proceedings of the Royal Society B*, 282(1821), 20151992.

Ruane, S., Bryson, R. W., Jr., Pyron, R. A., & Burbrink, F. T. (2014). Coalescent species delimitation in milksnakes (genus *Lampropeltis*) and impacts on phylogenetic comparative analyses. *Systematic Biology*, 63, 231-250.

Sabatino, S. J., & Routman, E. J. (2009). Phylogeography and conservation genetics of the hellbender salamander (*Cryptobranchus alleganiensis*). *Conservation Genetics*, 10(5), 1235.

San Mauro, D., Gower, D. J., Oommen, O. V., Wilkinson, M., & Zardoya, R. (2004). Phylogeny of caecilian amphibians (Gymnophiona) based on complete mitochondrial genomes and nuclear *RAG1*. *Molecular Phylogenetics and Evolution*, 33(2), 413-427.

Santoyo-Brito, E., Anderson, M., & Fox, S. (2017). Incubation temperature modifies sex ratio of hatchlings in collared lizards, *Crotaphytus collaris*. *Journal of Herpetology*, 51(2), 197-201.

Schliep, K. P. (2011). Phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4), 592-593.

Schmid, M., & Steinlein, C. (2001). Sex chromosomes, sex-linked genes, and sex determination in the vertebrate class Amphibia. In Genes and Mechanisms in Vertebrate Sex Determination (pp. 143-176). Birkhäuser, Basel.

Schönhuth, S., & Doadrio, I. (2003). Phylogenetic relationships of Mexican minnows of the genus *Notropis* (Actinopterygii, Cyprinidae). *Biological Journal of the Linnean Society*, 80, 323-337.

Sessions, S. K. (2008). Evolutionary cytogenetics in salamanders. *Chromosome Research*, 16(1), 183-201.

Sessions, S. K., León, P. E., & Kezer, J. (1982). Cytogenetics of the Chinese giant salamander, *Andrias davidianus* (Blanchard): The evolutionary significance of cryptobranchoid karyotypes. *Chromosoma*, 86(3), 341-357.

Sessions, S. K., Mali, L. B., Green, D. M., Trifonov, V., & Ferguson-Smith, M. (2016). Evidence for sex chromosome turnover in proteid salamanders. *Cytogenetic and Genome Research*, 148(4), 305-313.

Shaffer, H. B., & Breden, F. (1989). The relationship between allozyme variation and life history: Non-transforming salamanders are less variable. *Copeia*, 4, 1016-1023.

Shaffer, H. B., & McKnight, M. L. (1996). The polytypic species revisited: genetic differentiation and molecular phylogenetics of the tiger salamander *Ambystoma tigrinum* (Amphibia: Caudata) complex. *Evolution*, 50, 417-433.

Shaffer, H. B., & Thomson, R. C. (2007). Delimiting species in recent radiations. *Systematic Biology*, 56, 896-906.

Shaffer, H. B., Flores-Villela, O., Parra-Olea, G., & Wake, D. B. (2004). *Ambystoma ordinarium*. In: IUCN 2009. IUCN Red List of Threatened Species. Version 2009.1

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*. 27, 379-423.

Shen, X. X., Liang, D., Chen, M. Y., Mao, R. L., Wake, D. B., & Zhang, P. (2015). Enlarged multilocus data set provides surprisingly younger time of origin for the Plethodontidae, the largest family of salamanders. *Systematic Biology*, 65(1), 66-81.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51(3), 492–508.

Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C., & Brumfield, R. T. (2014). Target capture and massively parallel sequencing of ultraconserved elements for

comparative studies at shallow evolutionary time scales. *Systematic Biology*, 63, 83-95.

Smith, C. A., Roeszler, K. N., Ohnesorg, T., Cummins, D. M., Farlie, P. G., Doran, T. J., & Sinclair, A. H. (2009). The avian Z-linked gene *DMRT1* is required for male sex determination in the chicken. *Nature*, 461(7261), 267-271.

Smith, J. J., & Voss, S. R. (2009). Amphibian sex determination: segregation and linkage analysis using members of the tiger salamander species complex (*Ambystoma mexicanum* and *A. t. tigrinum*). *Heredity*, 102(6), 542.

Sonnini de Manoncourt, C. S., & Latreille, P. A. (1801). *Histoire Naturelle des Reptiles, avec Figures dissinées d'après Nature.* Volume 4. Paris: Deterville.

Springer, M. S., & Gatesy, J. (2016). The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94, 1-33.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312-1313.

Stöck, M., Croll, D., Dumas, Z., Biollay, S., Wang, J., & Perrin, N. (2011). A cryptic heterogametic transition revealed by sex-linked DNA markers in Palearctic green toads. *Journal of Evolutionary Biology*, 24(5), 1064-1070.

Stöck, M., Savary, R., Zaborowska, A., Górecki, G., Brelsford, A., Rozenblut-Kościsty, B., Ogielska, M. and Perrin, N. (2013). Maintenance of ancestral sex chromosomes in Palearctic tree frogs: direct evidence from *Hyla orientalis*. *Sexual Development*, 7(5), 261-266.

Sukumaran, J., & Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26(12), 1569-1571.

Sukumaran, J., & Knowles, L. L. (2017). Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences*, 114(7), 1607-1612.

Swofford, D. L. (2015) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, Massachusettes.

Thorne, J. L., Kishino, H., & Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, 33(2), 114-124.

Tong, S. Y., Holden, M. T., Nickerson, E. K., Cooper, B. S., Köser, C. U., Cori, A., Jombart, T., Cauchemez, S., Fraser, C., Wuthiekanun, V., & Thaipadungpanit, J. (2015). Genome sequencing defines phylogeny and spread of methicillin-resistant *Staphylococcus aureus* in a high transmission setting. *Genome Research*, 25(1), 111-118.

Tonione, M., Johnson, J. R., & Routman, E. J. (2011). Microsatellite analysis supports mitochondrial phylogeography of the hellbender (*Cryptobranchus alleganiensis*). *Genetica*, 139(2), 209-219.

Turtle Taxonomy Working Group (2007). Turtle taxonomy: methodology, recommendations, and guidelines. *Chelonian Research Monographs* 4, 73-84.

Unger, S. D., Chapman, E. J., Regester, K. J., & Williams, R. (2016). Genetic signatures follow dendritic patterns in the eastern hellbender (*Cryptobranchus alleganiensis*). *Herpetological Conservation and Biology*, 11(1), 40-51.

Unger, S. D., Rhodes Jr, O. E., Sutton, T. M., & Williams, R. N. (2013). Population genetics of the eastern hellbender (*Cryptobranchus alleganiensis alleganiensis*) across multiple spatial scales. *PloS One*, 8(10), e74180.

Vieites, D. R., Román, S. N., Wake, M. H., & Wake, D. B. (2011). A multigenic perspective on phylogenetic relationships in the largest family of salamanders, the Plethodontidae. *Molecular Phylogenetics and Evolution*, 59(3), 623-635.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T. V. D., Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M. and Zabeau, M. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, 23(21), pp.4407-4414.

Weisrock, D. W., Shaffer, H. B., Storz, B. L., Storz, S. R., & Voss, S. R. (2006). Multiple nuclear gene sequences identify phylogenetic species boundaries in the rapidly radiating clade of Mexican ambystomatid salamanders. *Molecular Ecology*, 15, 2489-2503.

Wheeler, B. A., Prosen, E., Mathis, A., & Wilkinson, R. F. (2003). Population declines of a long-lived salamander: a 20+-year study of hellbenders, *Cryptobranchus alleganiensis*. *Biological Conservation*, 109(1), 151-156.

Wiens, J. J. (2011). Re-evolution of lost mandibular teeth in frogs after more than 200 million years, and re-evaluating Dollo's law. *Evolution*, 65(5), 1283-1296.

Wilberg, E. W. (2015). What's in an outgroup? The impact of outgroup choice on the phylogenetic position of *Thalattosuchia* (Crocodylomorpha) and the origin of Crocodyliformes. *Systematic Biology*, 64(4), 621-637.

Wu, C. H., Tsai, M. H., Ho, C. C., Chen, C. Y., & Lee, H. S. (2013). De novo transcriptome sequencing of axolotl blastema for identification of differentially expressed genes during limb regeneration. *BMC Genomics*, 14(1), 434.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8), 1586-1591.

Yang, Z., & Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, 107(20), 9264-9269.

Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5), 303-314.

Yang, Z., & Rannala, B. (2014). Unguided species delimitation using DNA sequence data from multiple Loci. *Molecular Biology and Evolution*, 31, 3125-3135.

You, F.M., Huo, N., Gu, Y.Q., Luo, M.C., Ma, Y., Hane, D., Lazo, G.R., Dvorak, J., & Anderson, O.D. (2008). BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics*, 9(1), 253.

Zarza, E., Reynoso, V. H., & Emerson, B. C. (2008). Diversification in the northern neotropics: mitochondrial and nuclear DNA phylogeography of the iguana *Ctenosaura pectinata* and related species. *Molecular Ecology*, 17, 3259-3275.

Zhang, C., Rannala, B., & Yang, Z. (2014). Bayesian species delimitation can be robust to guide-tree inference errors. *Systematic Biology*, 63, 993-1004.

Zhang, P., & Wake, D. B. (2009). Higher-level salamander relationships and divergence dates inferred from complete mitochondrial genomes. *Molecular Phylogenetics and Evolution*, 53(2), 492-508.

Zhang, P., & Wake, M. H. (2009). A mitogenomic perspective on the phylogeny and biogeography of living caecilians (Amphibia: Gymnophiona). *Molecular Phylogenetics and Evolution*, 53(2), 479-491.

Zhang, P., Liang, D., Mao, R. L., Hillis, D. M., Wake, D. B., & Cannatella, D. C. (2013). Efficient sequencing of anuran mtDNAs and a mitogenomic exploration of the

phylogeny and evolution of frogs. *Molecular Biology and Evolution*, 30(8), 1899-1915.

Zhu, B., Feng, Z., Qu, A. I., Gao, H., Zhang, Y., Sun, D., Song, W., & Saura, A. (2002). The karyotype of the caudate amphibian *Andrias davidianus*. *Hereditas*, 136(1), 85-88.

VITA

Paul Michael Hime was born in Nashville, Tennessee, the son of Michael Stanley Hime and the late Elizabeth Ann Davis Hime. After graduating from Martin Luther King Jr. Magnet School for the Health Sciences and Engineering in 2000, Paul enrolled at Washington University in St. Louis, Missouri. Under the supervision of Jonathan B. Losos, Paul conducted ecological field work in the Caribbean and North America, and earned a Bachelor of Arts degree in Biology in 2004. After graduation, he worked for four years as an animal care technician and research technician in the laboratories of Jonathan B. Losos and Ralph S. Quatrano, respectively, at Washington University. In 2007, Paul joined the St. Louis Zoo as a herpetology keeper under the supervision of Jeffrey Ettling and Mark Wanner, where he focused on amphibian conservation initiatives. In 2011, Paul joined the laboratory of David W. Weisrock at the University of Kentucky in Lexington, Kentucky to pursue a doctoral degree in evolutionary biology. During his graduate tenure, Paul was awarded graduate research fellowships from the National Science Foundation and the National Center for Supercomputing Applications, as well as a Doctoral Dissertation Improvement Grant from the National Science Foundation, for his genomic research into amphibian evolution. Paul's doctoral research focuses on computational phylogenetic approaches to study speciation and genome evolution in non-model organisms.

This dissertation was typed by the author.